Cancer initiation is influenced by sex-biased tissue environment or imbalanced hormones

Feng Zhang[1], Minghua Deng[1,2,3]*

[1]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China.

[2]School of Mathematical Sciences, Peking University, Beijing, China.

[3]Center for Statistical Science, Peking University, Beijing, China.
*Corresponding author：Minghua Deng

Email: dengmh@math.pku.edu.cn

Abstract

Gender difference has long been observed in non-reproductive cancer susceptibility and prognosis. In addition to external factors, biological sex bias, such as sex chromosomes, hormones, and immune function, is thought to influence the evolutionary process of cancer initiation as a selective pressure. However, the extent and mechanism of how sex bias influences cancer initiation remain largely unknown. In this study, we show that tissue sex bias directed by gonadotropin-releasing hormone varies among different tissues and is associated with two distinct age-specific cancer incidence patterns: the parallel and nonparallel patterns. In addition, imbalances in estrogen and thyroid hormone are shown to correlate with levels of hypoxia induce factors, resulting in three phases (hypoxia, hyperoxia, and "chaotic-oxia") that exist across the majority of TCGA cancers and are linked to certain cancer subtypes, including cancers with microsatellites, CpG island methylator phenotype or hypermethylation. Our results suggest that sex-biased tissue environments and imbalanced hormones could influence the evolutionary process of cancer initiation, emphasizing the importance of maintaining hormone homeostasis for cancer prevention and providing insights for improving therapy for cancer types with hormone imbalances.

Main Text

The age-specific incidence curve of cancer can reflect the evolutionary process of cancer initiation. For carcinomas (most solid cancers), age-specific incidence curves were found approximately linear in log-log scale[1]. Following multistage model proposed by Armitage and Doll[2] explained the power law relationship, based on the assumption that carcinogenesis is a multistage process of driver mutation accumulation. The model has been modified to incorporate new knowledge by others[3-8]. Cancer initiation is influenced by selective pressure of microenvironment. During this process, driver mutations (genetic or epigenetic) are selected and accumulated. Generally, different cancer types have different mutation spectra[9,10], indicating that variety of tissue environments and cell types can influence the evolutionary trajectories of cancer initiation.

Females and males have different susceptibilities and prognoses to nearly all non-reproductive cancers. However, behavior or exposure to different mutagens can only partly explain this phenomenon[11]. Biological sex bias, such as sex chromosomes and hormones, also play roles in many aspects[12-15]. Recent studies have shown that female or male patients have a preference to mutate certain genes, to gain certain transcriptional or epigenetic signatures[16-18], and female patients prefer to accumulate more T cell receptor unrecognized mutated peptides due to their more powerful immunity[19]. These studies give evidence to the hypothesis that sex bias is a selective pressure and could influence cancer initiation. However, how sex bias influences cancer initiation is still largely unknown.

Parallel and nonparallel patterns of age-specific cancer incidence between the two sexes

Building on the hypothesis that sex bias could influence cancer initiation and the age-specific incidence curve of cancer, we examined curve patterns of major solid cancer types between females and males in selected regions or races by utilizing the CI5-X database available by the International Agency for Research on Cancer (see the Materials and Methods). Fitted age-specific incidence lines of the two sexes show roughly two patterns on double logarithmic axes: parallel and nonparallel (Fig. 1A). Parallel patterns were observed when comparing incidences between inherited and sporadic colon cancer[7], and in lung cancer incidences between smoke and non-smoke individuals[20]. These parallel patterns suggest that individuals with susceptible genetic background or those exposed to mutagens have a higher baseline risk of developing cancer. To date, nonparallel patterns have not been reported and remain a mystery.

To determine the prevalence of nonparallel patterns in major solid cancers across regions or races, we conducted a linear regression analysis and tested for age-by-sex interaction to assess the significance of the nonparallel pattern in each cohort (see the Materials and Methods). Our analysis revealed that nonparallel patterns are more frequently observed in stomach cancer, melanoma, and thyroid cancer (Fig. 1B).

Sex bias in tissue environments

One potential explanation for the prevalence of nonparallel pattern is that tissue environments of stomach, skin, and thyroid between the two sexes are more distinct compared to environments of the other tissues, and this difference could influence cancer initiation. To test this hypothesis, we evaluated sex bias of tissue environment in GTEx data and conducted a sex-biased expressed pathways analysis. In this analysis, the expression of pathway was represented by mean expression of all genes in that pathway, and we randomly selected 100 samples from each tissue to maintain consistent power (sample size) in the comparison across tissues. We tested for sex-biased expressed pathways with limma[21], repeating the process 1000 times and recording numbers of sex-biased expressed pathways in each tissue in each repetition (see the Materials and Methods). Kidney and bladder have fewer than 100 samples (Table S1) and have no significant sex-biased expressed pathway, so their results were marked as not available (NA). The results show that the adrenal gland, stomach, skin, and thyroid have the most sex-biased pathways (Fig. 2A).

To examine which pathways are more involved in tissue sex bias, we performed a retest for sex-biased expressed pathways using all the samples of each tissue, except for the bladder, which had only nine samples. Of all the pathways, only Gonadotropin-releasing hormone (GnRH) secretion pathway (hsa04929) showed significant sex bias (with a limma P value lower than 0.05) in the top three sex-biased tissue: adrenal gland, thyroid, and stomach, but not in the other tissues (the P value in skin was 0.0683, Table S2). In addition, the female-to-male log-fold changes (LogFC) of the GnRH secretion pathway in adrenal gland, stomach, and thyroid were the largest, while the skin was ranked fifth. The LogFC of the GnRH secretion pathway shows positive correlation (Pearson correlation coefficient r is 0.74) with the number of sex-biased expressed pathways across each tissue (Fig. 2B). LogFC of another GnRH-related pathway, the GnRH signaling pathway (hsa04912), also showed a positive correlation (Pearson correlation coefficient r is 0.91) with the number of sex-biased expressed pathways and has similar rank among the tissues. The adrenal gland, stomach, and thyroid had the largest fold changes, but the skin had a negative LogFC (Fig. 2C, Table S3).

Adrenal gland cancer is a rare type of cancer, and the age-specific incidence is too noisy to be fitted linearly. Therefore, it was not included in the nonparallel analysis shown in Figure 1B. However, due to the highest number of sex-biased expressed pathways and largest LogFC of the two GnRH-related pathways, we re-conducted a nonparallel pattern analysis on adrenal gland cancer. To reduce the noise of adrenal gland cancer incidence, we merged the regional cohorts into continent cohorts (see the Materials and Methods). The results revealed that adrenal cancer incidence also displayed a nonparallel pattern, and the results were significant in North American and Asian cohorts (Fig. 2D).

We observed the most prevalent nonparallel patterns in stomach cancer, melanoma, thyroid cancer, and adrenal cancer across regions or races (Fig. 1B, Fig. 2D). Correspondingly, adrenal gland, stomach, thyroid, and skin exhibited the most sex-biased expressed pathways (Fig. 2A), which were related to GnRH-related pathways. These findings suggest that sex-biased tissue environments could influence the evolutional process of cancer evolution and induce nonparallel patterns.

Sex-biased mutated genes

Since cancer initiation can be influenced by sex, it is suspected that there may exist mutated genes that show a higher likelihood of selection in one sex over the other. Previous studies attempted this work, but they relied on incomplete TCGA data[16,17] or only whole genome sequencing samples[18], which had less statistical power. Therefore, we conducted a comparative analysis of gene mutation frequency between a total of 6181 females and males across 24 non-reproductive cancers using the completed TCGA data, using both likelihood ratio test (LRT) and logistic regression to adjust for confounders (see the Materials and Methods). Our results revealed 8 genes with significant sex bias (Table 1): DNAH7, LARP4B, PGM5, and PTEN in stomach adenocarcinoma (STAD); SELP and SPAG17 in cutaneous melanoma (SKCM); RBM10 in lung adenocarcinoma (LUAD); and CASP8 in head and neck squamous cell carcinoma (HNSC).

Hotspot dominance in three sex-biased mutated genes in STAD

In STAD, three sex-biased mutated genes (PTEN, LARP4B, and PGM5) were found to display a dominance of hotspot mutations, as shown in Figure 3, painted with ProteinPaint[22]. These hotspots, Lys267ArgfsTer9 (c.800delA) in PTEN, Thr163HisfsTer47 (c.487delA) in LARP4B, and Ile98Val (c.292A>G) in PGM5, also showed a significant sex bias ($p<0.05$, Fisher exact test, Fig. 3). In TCGA data, these three hotspots were observed to be exclusively present in STAD, colon adenocarcinoma (COAD), and uterine corpus endometrial carcinoma (UCEC), except for one PGM5 hotpot in prostate adenocarcinoma (PRAD). Mutation spectra of the other sex-biased mutated genes were shown in Figure S1.

Sex-biased hotspots and sex-biased proportions of cancers with MSI

To investigate the underlying reason for the sex-biased presence of the three hotspots, we conducted a co-occurrence and exclusivity analysis for the three hotspots in STAD, COAD, and UCEC. The results suggested that the three sex-biased hotspots in STAD, COAD, and UCEC tend to co-occur (Table S4). To gain further insight, we divided the samples into two groups (HOTSPOT+ and HOTSPOT-) based on the presence or absence of at least one of the three hotspots across STAD, COAD, and UCEC. Subsequently, we conducted a differentially expressed genes (DEGs) analysis between the two groups with limma.

Four genes were found to be differentially expressed between the two groups in all three cancers: C7orf31, MLH1, EPM2AIP1, and LINC00667. MLH1, a well-known tumor suppressor gene, was found to have reduced expression in the HOTSPOT+ groups across all three cancers (Fig. S2). MLH1 deficiency can induce microsatellites (MSI), which is commonly found in colorectal, gastric, and endometrial cancer. The signatures of the three hotspots correspond to the mutation signatures of MSI cancer types (PTEN and LARP4B hotspots are typical MSI mutation signatures ID1, and the PGM5 hotspot is typical of MSI mutation signatures SBS21 or SBS26) as reported in a previous study[23]. Furthermore, all HOTSPOT+

samples were labeled with MSI subtype in TCGA data (see the Materials and Methods). The evidence suggests that the three hotspots are result of hyper-mutation in MSI cancer subtypes.

The finding suggests that the sex bias of hotspots may be induced by the sex bias of MSI cancer subtype. To investigate this further, we checked the proportions of MSI subtype in STAD and COAD, and we found that proportion of MSI subtype in females was higher than in males in both STAD and COAD. This difference was significant in STAD (Table S5). In the data of a previous study[24], the proportion of MSI subtype in females was significantly higher than in males both in gastric and colorectal cancer (Table S6).

Estrogen receptor alpha, thyroid hormone receptors and hypoxia induce factors in cancers with MSI

The question of why MSI subtype cancer is sex-biased has arisen from the aforementioned findings. To uncover the underlying mechanism, we divided the samples into two groups, MSI positive samples (MSI+) and the others (MSI-) across STAD, COAD, and UCEC. We then performed DEGs analysis for each cancer type (see the Materials and Methods).

It is well-documented that hypoxia can suppress MMR, such as by reducing the expression of MLH1[25-28]. In STAD, COAD and UCEC, we have observed a consistent phenomenon that MSI+ samples had significantly elevated HIF1A, which is the key regulator of hypoxia signaling pathways, or reduced HIF3A, which is generally considered as a repressor of hypoxia signaling[29], compared to MSI- samples. This indicates MSI+ samples had a higher level of hypoxia. Than we used the counts ratio of HIF1A to HIF3A, termed as HIF, to represent the hypoxia level of each sample. We found that HIF levels are significantly elevated in MSI+ samples than MSI- samples and have a positive correlation with MLH1 methylation levels across all three cancers (Fig. 4A).

It is also well-documented that estrogen receptors alpha (ERα, ESR1) directly regulates hypoxia induce factor 1 alpha (HIF-1α, HIF1A), indicating a crosstalk between hypoxia and estrogen signaling pathway[30]. However, in STAD, COAD and UCEC, we didn't observe significantly elevated level of ESR1 in MSI+ samples than MSI- samples. Furthermore, the expression of ESR1 only had a positive correlation with HIF levels in UCEC, but not in STAD and COAD (Fig. 4B and 5A). In addition, ESR1 has only a slight positive correlation with HIF levels in TCGA pan-cancer level and had a negative correlation with HIF level in GTEx pan-tissue level (Fig. S3).

Our aforementioned results showed that GnRH was involved in sex-biased tissue environment. GnRH is the key regulator of the hypothalamus-pituitary-gonad (HPG) axis, which is the upper cascade of estrogen and has substantial crosstalk with hypothalamus-pituitary-adrenal gland (HPA) and the hypothalamus-pituitary-thyroid (HPT) axes[31]. These connections prompted us to investigate whether other hormones or hormones receptors are involved in the sex bias of MSI subtype. Therefore, we check the expression of a set of genes which are involved in HPG, HPA and HPT axes (see the Materials and Methods).

The results showed that only one gene, thyroid hormone receptor beta (THRB), was significantly differentially expressed between the MSI+ and MSI- samples in all three cancer types (Fig. S4). Another thyroid hormone receptor, thyroid hormone receptor alpha (THRA), was significantly differentially

expressed in STAD and COAD, but not in UCEC (Fig. S4). Both THRA and THRB are important receptors of thyroid hormone (TH), so we used the summed logarithmic counts of THRA and THRB, termed as THR, to represent thyroid hormone receptor level of each sample. We found THR levels are significantly reduced in MSI+ samples than MSI- samples and have a negative correlation with HIF levels across all three cancers (Fig. 4C). The negative correlation between HIF and THR is also confirmed in TCGA pan-cancer level and in GTEx pan-tissue level (Fig. 4D).

ESR1, THR and HIF in cancers with hypermethylation or in ER+ breast cancers

Epigenetic silence of the MLH1 gene is a major cause of MSI, and it can occur in cancers with CpG island methylator phenotype (CIMP) or hypermethylation. A significant overlap between cancers with MSI and CIMP was observed in colorectal cancer[32]. This observation prompted us to investigate whether cancers with CIMP or hypermethylation exhibit similar expression signatures (ESR1, THR and HIF) to MSI+ cancers, we conducted a similar analysis in cancers that have sufficient CIMP or hypermethylation labeling in TCGA data. We categorized these cancers as HyperMeth, while the control groups were termed HypoMeth (see Materials and Methods for more details).

Based on their expression signatures, HyperMeth cancers can be divided into two subgroups: the hypoxia group and the hyperoxia group. In the hypoxia group, similarly to MSI+ cancers, all HyperMeth cancers have significant elevated levels of HIF (except for HyperMeth SKCM, where the increase was not significant), typically have reduced levels of THR, and elevated levels of ESR1 (Fig. 5A and Fig. S5). In the hyperoxia group, all HyperMeth cancers have significant reduced levels of HIF (except for HyperMeth THCA, where the reduce was not significant), and have significant elevated levels of THR or significant reduced levels of ESR1 (Fig. 5A).

Adrenocortical carcinoma (ACC) and Pheochromocytoma and Paraganglioma (PCPG) both originate from the adrenal gland. However, HyperMeth ACC and HyperMeth PCPG cancers have a reversed expression signature of THR and HIF. Specifically, ACC cancers are in hypoxia state, while HyperMeth PCPG cancers are in hyperoxia state. This suggests that they may have initiated under reversed selective pressure or evolved in opposite directions.

Breast cancer is classified into four well-characterized molecular subtypes: luminal A, luminal B, HER2-positive, and triple-negative. Estrogen plays a crucial role in the development of ER positive (ER+) breast cancers (luminal A and luminal B). To investigate whether ER+ breast invasive carcinomas (BRCA) exhibit hypoxia state and have a similar expression signature to MSI+ or hypoxia HyperMeth cancers, we compared the expression signature of ER+ BRCA to ER- (HER2-positive and triple-negative) BRCA. Unlike the results seen in MSI+ and HyperMeth cancers, we found an antagonistic expression pattern in ER+ BRCA: both ESR1 and THR are significantly elevated, while HIF level is not significantly different (Fig. 5A). However, HIF1A level was marginally close to significantly reduced, and HIF3A levels were significantly reduced (Fig. S4).

Estrogen receptor beta (ERβ, ESR2) is also involved in hypoxia signaling. However, its role is not well established, and it even exhibit opposed roles in different cancers[33]. Our results shown that ESR2 was

significantly elevated in COAD MSI+ cancers and significantly reduced in SKCM HyperMeth cancers, indicating ESR2 may play a role in cancer initiation in these cancers.

In addition, as three hotspot mutations were exclusively found in MSI+ cancers, CASP8 was typically mutated in HyperMeth HNSC, and SELP was typically mutated in HyperMeth SKCM samples, DNAH7 mutations had a higher prevalence in MSI+ STAD samples, and these results were found to be significant (Table S7), indicating a potential link between these sex-biased mutated genes and corresponding cancer subtypes.

Imbalanced estrogen and TH model of carcinogenesis

The results above have shown that sex-biased tissue environment can influence cancer initiation. ESR1 and THR are linked to cancer HIF levels and certain cancer subtypes, suggesting that they may also play a role in cancer initiation. ESR1 can directly regulate HIF1A and induce the hypoxia signaling pathway[30]. On the other hand, TH has a more complicated role in hypoxia. TH can induce HIF-1α activity, which in turn reduces TH signaling. However, TH also plays a critical role in protecting tissues from hypoxic injury in experimental and clinical practices[34]. This complicated phenomenon is also evident in our observations: THR levels are negatively correlated with HIF1A expression but positively correlated with HIF3A expression in GTEx data (Fig. S6). However, in TCGA data, the negative correlation between THR and HIF1A disappears, while the positive correlation between THR and HIF3A remains (Fig. S7). These indicate that THR is more of a hypoxia repressor than a hypoxia activator in cancers, similar to its protective role in hypoxic injury.

Both estrogen receptors (ERs) and thyroid hormone receptors (THRs) are nuclear receptors with similar DNA-binding sequences. They are able to competitively bind to each other's DNA-binding site, inhibiting transcription mediated by each other[35]. The mutual antagonistic relationship between ERs and THRs corresponds to the observations shown in Figure 5A: elevated ESR1 or reduced THR are accompanied by hypoxia, while reduced ESR1 or elevated THR are accompanied by hyperoxia.

Based on these findings, we propose an imbalanced estrogen and TH model of carcinogenesis (Fig. 5B). As previously discussed, TH and estrogen play essential roles in many aspects of biological processes and are in homeostasis in healthy conditions. However, dominance by either TH or estrogen may induce alterations in the tissue environment, such as hypoxia or hyperoxia, and alter the epigenome and transcriptome of tissue cells. In ER+ BRCA, the situation where both TH and estrogen are elevated also occurs, resulting in a different state that we term as antagonism or "chaotic-oxia" state. As a selective pressure for cancer evolution, all of these altered tissue environments can induce specific cancer subtypes, as shown in Figures 5A and 5B.

Discussion

Our study shows that cancer initiation can be influenced by sex bias in two ways: sex-biased tissue environment and hormonal imbalance. However, these two ways are not mutually exclusive and may interact with each other. GnRH, which has been shown to be related to sex-biased tissue environment in

our study, is a key regulator of the HPG axis and receives negative feedback from gonadal steroids[36]. The HPG axis has a lot of crosstalk with the HPT axis, such as crosstalk between estrogen and TH[31,35]. The HPG, HPT, and other factors together regulate the homeostasis of tissue environment. Generally, females have a higher level of estrogen than males, which may more likely lead to a hypoxic tissue environment in females and induce cancer subtypes such as MSI. As expected, we observe a higher prevalence of MSI subtypes in STAD and COAD in females, and also a high prevalence of MSI in UCEC where estrogen is believed to promote cancer progression (Table S5 and S6). On the other hand, males generally have a higher level of free triiodothyronine than females, which may more likely lead to a hyperoxic tissue environment in males and induce cancer subtypes such as hyperoxia CIMP in THCA. Although the difference is not significant due to the small sample size, we can observe more hyperoxia CIMP subtypes of THCA in males (Table S5). However, the hypoxia HyperMeth subtype in HNSC is more prevalent in males, indicating that there may be tissue-specific regulation in it. This suggests that in a sex-biased or hormonally imbalanced tissue environment, cancer precursors may have a preferred trajectory to evolve, as we observed that 6 of the 8 sex-biased mutated genes are significantly enriched in certain cancer subtypes. However, in cancers like PCPG and SKCM, the proportion of subtypes seems comparable.

In addition to influencing the direction of cancer evolution, sex-biased tissue environment and hormonal imbalance could also affect the dynamics of cancer evolution. GnRH secretion exhibits a pulsatile mode in both males and females, but in females, the frequency and amplitude of pulses are dependent on the menstrual cycle phase. Moreover, females have a unique surge mode of GnRH secretion, which occurs during the pre-ovulatory phase[36]. Sex-biased dynamics of GnRH secretion may induce sex-biased evolutionary dynamics of cancer. Hypermutation or hypermethylation, which occurs in cancers with hormonal imbalance, could lead cancer evolution into a faster trajectory and alter its dynamics, as discussed in a previous study[6]. All of these factors may contribute to the nonparallel pattern observed in cancer incidence. However, to fully understand sex-biased patterns, a more detailed understanding of sex-biased biological mechanisms is needed to develop a well-explained mathematical model.

Thyroid hormone is believed to play a role in cancer initiation. For example, females are ten times more likely than males to get hyperthyroidism and hypothyroidism, which have been linked to changes in the incidence and prognosis of many cancers[37,38]. Our study provides evidence for the emerging role of thyroid hormone in cancer. While our study focused on the balance of TH and estrogen, it is important to acknowledge that other hormones may also play a role in cancer development. For example, progestogen has been shown to be involved in luminal A breast cancer, while androgens have been linked to prostate cancer.

The sample grouping in this study is based on molecular clustering labeling, which may not fully capture the state of estrogen and TH balance. Identifying molecular patterns associated with imbalance of estrogen and TH could refine our results and reveal more sex-biased mechanisms. Additionally, levels of ERs and THRs may not directly reflect the levels of estrogen and TH. Nevertheless, our study has shown that three phases (hypoxia, hyperoxia, and chaotic-oxia) exist across the majority of cancers and are

linked to imbalance of hormones. This highlights the importance of maintaining hormone homeostasis for cancer prevention and provides insights for improving therapy for hormone imbalanced cancer types.

Materials and Methods

Epidemiology data processing

To ensure the reliability of our analysis, we selected regions with the top 10 GDP in 2005, corresponding to the CI5X database, which records incidence from 2003 to 2007. The GDP information was obtained from the World Bank database (https://databank.worldbank.org/). In the data of USA, we selected two cohorts - white and black - to cover the black race, resulting in a total of 11 cohorts of regions or races: USA black, USA white, Japan, Germany, England, China, France, Italy, Canada, Spain, and Korea. We then chose major non-reproductive cancers according to Cancer Statistics US[39]: lung (LUSC and LUAD), melanoma, colon, bladder, kidney, pancreas, thyroid, rectum, liver, stomach, brain, esophagus, oral cavity and pharynx, gallbladder, and larynx. This resulted in a total of 11 by 16 cohorts included in our analysis.

To reduce noise in the incidence of adrenal gland cancer, we merged the regional cohorts into continent cohorts. Specifically, we merged the USA black, USA white, and Canada cohorts into a North American cohort, merged Japan, China, and Korea cohorts into an Asian cohort, and merged Germany, England, France, Italy, and Spain cohorts into a Europe cohort.

We cut the age-specific incidence data from 30 to 70 years of age. Because previous studies have shown that incidence rates before the age of 30 are higher than expected due to factors such as family inheritance, exposure to carcinogens, or explosive cell division during development[1,40], and incidence rates after the age of 70 are lower than expected due to the selection of relatively unsusceptible individuals[41]. We then used the least squares method to fit straight lines separately for females and males at a log-log scale.

Nonparallel pattern test

We consider a linear equation (equation 1) as below:

$$incidence = \alpha \cdot age + \beta \cdot sex + \lambda \cdot age \cdot sex + \varepsilon$$

Denoting 0 as female and 1 as male, so the linear equation (equation 2) for male is:

$$incidence_{\text{male}} = \alpha \cdot age + \beta + \lambda \cdot age + \varepsilon$$

The linear equation (equation 3) for female is:

$$incidence_{\text{female}} = \alpha \cdot age + \varepsilon$$

When λ is 0, two lines of sexes are parallel, when λ is not 0, two lines of sexes are nonparallel. So we perform a linear regression as equation 1, and test if λ is significant not equal to 0 (significant nonparallel, Student's t-test). Then P values are adjusted by Bonferroni methods, adjusted P values which lower that 0.05 are reported as significant.

Sex-biased expressed pathways analysis

Carcinomas typically originate from solid tissues. Therefore, in the GTEX data, we selected tissues from which major solid cancers originate and excluded blood and solid tissues that are less likely to develop cancer, such as heart and fat. After this selection process, we chose 12 tissues: bladder, kidney, liver, skin, colon, stomach, pancreas, lung, thyroid, brain, esophagus, and adrenal gland. The number of samples for each tissue is shown in Table S1.

We obtained human pathway information using the R package KEGGREST[42,43], resulting in a total of 352 pathways included in our analysis.

To ensure that each tissue had equal testing power for horizontal comparison, we randomly selected 100 samples from each tissue and conducted sex-biased expressed pathways analysis. For each pathway, we used the mean logarithmic expression (log (counts+1)) of all genes in the pathway as the expression strength of the pathway. We divided the samples in each tissue into two groups based on sex and compared pathway expression strength between groups using Limma, while correcting for the factor of the position of the tissue (body site detail). As shown in Table S1, there was a large difference in sample size between tissues, and our approach allowed us to account for these differences and ensure that our analysis was statistically robust.

Our analysis was affected by poor testing power due to both the expression strength of the pathway, which can weaken the real signal of a set of specific genes in that pathway, and the small sample size resulting from our sampling approach. However, as we only made qualitative comparisons, we relaxed the significance standard and considered unadjusted P values < 0.05 to be significant. We then recorded the number of significantly different pathways in each tissue. This analysis process was repeated 1000 times, and the numbers of sex-biased expressed pathways in each tissue were displayed in Figure 2A. As the samples from bladder and kidney were less than 100, and they had no significant sex-biased expressed pathways, their final statistical results were marked as NA.

Mutation data processing for sex-biased mutated gene test

We used MC3 profile of TCGA mutation data. We excluded pediatric and blood cancers, as well as sex-specific cancers, leaving only non-reproductive adult solid tumors.

To filter out passenger mutations and reduce data noise, we performed filtering on the mutation data. Firstly, we retained mutations of the following types: nonsense, insertion, deletion, translation start site, and non-stop mutations, as these mutations would theoretically change the protein sequence. We also retained mutations predicted as "deleterious" by SIFT[44] and as "probably damaging" by PolyPhen[45]. In addition, we retained mutations at a single site where mutations occurred five or more times in the entire MC3 dataset.

To reduce potential statistical bias from extremely high mutation samples, we removed the top 5% of samples with the highest tumor mutation burden (TMB) for each cancer type. Additionally, we removed genes with a sample mutation rate below 5%, as these genes have limited biological significance and their removal can improve the power of a statistical test. For the analysis of melanoma, which is only

susceptible in white populations and typically only has nonparallel patterns in white populations, we only considered the white population.

Likelihood ratio test

Based on our preliminary analysis of sex-biased mutated genes using the Fisher exact test and the conclusions from previous work[46], we identified tumor mutation burden (TMB) as an important confounding factor for testing sex-biased mutated genes, as there are sex-biased TMB in many cancers. Failure to exclude this confounding factor can result in many false positives, particularly for super long genes. To address this issue, we used a likelihood ratio test (LRT) similar to previous work[17] to eliminate the confounding factor of TMB. The LRT also corrected for the difference in copy numbers of the X chromosome between males and females.

The background mutation rate varies between samples. Firstly, we assumed that the mutation rate of a gene between male and female samples is unbiased. Therefore, the mutation rate ratio between male and female samples should be approximately equal to the background mutation rate ratio between male and female samples. We formulated the null hypothesis that the mutation rate of a gene is not biased between male and female samples. In other words, if the null hypothesis is true, for each gene:

$$p_{female} = \lambda \cdot p_{male}$$

Where P_{female} is the mutation rate of a gene in female samples, P_{male} is the mutation rate of a gene in male samples, and λ is the background mutation rate ratio between male and female samples. We defined the mutation rate of a gene as the ratio of samples with mutations in that gene divided by the total number of female or male samples in a given cancer type. We denoted the mutation rate of a gene in female samples as Pfemale and the mutation rate of the same gene in male samples as Pmale. The background mutation rate ratio for a gene was defined as the ratio of the total number of mutations on the chromosome where the gene is located in female samples to the total number of mutations on the same chromosome in male samples.

Under the assumption that the null hypothesis is true, the likelihood function is:

$$L_0 = \max\left(\left(\lambda \times p_{male}\right)^f \times \left(1 - \lambda \times p_{male}\right)^{F-f} \times p_{male}{}^m \times \left(1 - p_{male}\right)^{M-m} \right)$$

If the null hypothesis is not true, it is assumed that P_{male} and P_{female} are not proportional to each other, and the likelihood function is:

$$L_1 = \max\left(p_{female}{}^f \times \left(1 - p_{female}\right)^{F-f} \times p_{male}{}^m \times \left(1 - p_{male}\right)^{M-m} \right)$$

The log-likelihood ratio (LLR) is then calculated as:

$$LLR = \ln\left(\frac{L_1}{L_0}\right)$$

According to Wilks' theorem, 2×LLR follows a chi-square distribution with one degree of freedom. Thus, we can calculate the P value, and the P value for each gene is adjusted using the Benjamini-Hochberg method within each cancer types.

Logistic regression analysis

To account for potential confounding factors such as age, cancer grade, smoking history, and alcohol history, we conducted logistic regression analysis with these covariates to investigate the significance of sex in gene mutations.

We used the likelihood ratio test to identify genes with adjusted P values below 0.05, which helped eliminate the confounding factor of tumor mutation burden (TMB). We also used logistic regression analysis with a significance threshold of $P < 0.05$ to identify genes that were significant after accounting for confounding factors. Genes that were identified as significant by both the likelihood ratio test and logistic regression analysis were reported as significant genes.

Test for sex biased mutated genes on Y chromosome

As females do not possess the Y chromosome, the LRT method cannot be employed for genes located on the Y chromosome. Therefore, we utilized the Fisher exact test to evaluate the significance of mutations in genes located on the Y chromosome. We adjusted the resulting P values using the Benjamini-Hochberg method to account for multiple testing. Genes with an adjusted P value below 0.05 were considered significant.

Subtype labeling in TCGA data and sample grouping

We obtained sample subtypes in TCGA from the molecular subtype profiles available on UCSC Xena[47]. We used the following sample groupings for our analysis.

For STAD and COAD, we grouped samples labeled "HM-indel" or "HM-SNV" in the MSI+ group, while samples labeled "CIN", "GS", or "EBV" were grouped in the MSI- group. These sample groupings match well with the samples labeled "msi-h" in previous study[48]. For UCEC, we grouped samples labeled "UCEC.POLE" or "UCEC.MSI" in the MSI+ group, while samples labeled "UCEC.CN_LOW" or "UCEC.CN_HIGH" were grouped in the MSI- group.

For LGG and ACC, samples labeled "CIMP-high" were grouped in the HyperMeth group, while samples labeled "CIMP-low" were grouped in the HypoMeth group. For THCA, samples labeled "CpG island methylated" were grouped in the HyperMeth group, while samples labeled "follicular" were grouped in the HypoMeth group. For SKCM and HNSC, samples labeled "CpG island-methylated" or "hyper-methylated" were grouped in the HyperMeth group, while samples labeled "normal-like" or "hypo-methylated" were

grouped in the HypoMeth group. For PCPG, samples labeled "hyper-methylated" were grouped in the HyperMeth group, while samples labeled "low-methylated" were grouped in the HypoMeth group. For LUAD, samples labeled "high" were grouped in the HyperMeth group, while samples labeled "low" were grouped in the HypoMeth group. For PRAD, samples labeled "1" were grouped in the HyperMeth group, while samples labeled "4" were grouped in the HypoMeth group.
For BRCA, samples labeled "LumA" or "LumB" were grouped in the ER+ group, while samples labeled "Her2" or "Basal" were grouped in the ER- group.

Differentially expressed genes (DEGs) analysis

We conducted DEGs analyses using limma between two groups in each cancer type with enough labeled samples (more than 50). Specifically, we compared HOTSPOT+ versus HOTSPOT- in STAD, COAD, and UCEC; MSI+ versus MSI- in STAD, COAD, and UCEC; HyperMeth versus HypoMeth in HNSC, LUAD, PRAD, ACC, PCPG, LGG, THCA, and SKCM; and ER+ versus ER- in BRCA. We corrected for covariates such as sex, age, tumor stage, race, alcohol, and smoking history during the DEG analyses.
Pseudogene THR was created by using the summed logarithmic counts of THRA and THRB, to represent thyroid hormone receptor level of each sample. Pseudogene HIF was created by using the counts ratio of HIF1A to HIF3A, to represent hypoxia level of each sample. Both THR and HIF were analyzed together with all the other genes during the DEG analyses.
We created two pseudogenes, THR and HIF, to represent thyroid hormone receptor level and hypoxia level, respectively, in each sample. Pseudogene THR was created by using the summed logarithmic counts of THRA and THRB, while pseudogene HIF was created by using the logarithmic counts of HIF1A minus the logarithmic counts of HIF3A. Both THR and HIF were analyzed together with all the other genes during the DEG analyses.
We reported the adjusted P value and log fold change (logFC) provided by limma for each gene. Specifically, genes with a limma adjusted P value lower than 0.05 and an absolute value of logFC higher than 1 were considered significant.

Sets of genes focused in expression signature

Basing on the role of GnRH on tissue sex bias and the crosstalk between HPG, HPA and HPT axes and relationship between ESR1 and hypoxia, we focused on the following sets of genes:
GnRH (GNRH1) and its receptors (GNRHR, GNRHR2), glycoprotein hormones alpha polypeptide (CGA), luteinizing hormone (LHB) and its receptors (LHCGR), follicle stimulating hormone (FSHB) and its receptor (FSHR), thyrotropin releasing hormone (TRH) and its receptor (TRHR), thyroid stimulating hormone (TSHA, TSHB) and its receptor (TSHR), corticotrophin-releasing hormone (CRH) and its receptor (CRHR1 and CRHR2), adrenocorticotropic hormone (POMC) and its receptor (MC2R), glucocorticoid receptor (NR3C1), thyroxine-binding globulin (SERPINA7), transthyretin (TTR), estrogen receptors (ESR1, ESR2), thyroid hormone receptors (THRA, THRB), testis receptor (NR2C1, NR2C2),

androgen receptor (AR), progesterone receptor (PGR), hypoxia-inducible factors (HIF1A, EPAS1, HIF3A, EGLN1, EGLN2, EGLN3, ARNT).

Acknowledgments

The authors thank Minping Qian, Hong Wu, Zexian Zeng, Ruibin Xi, Ge Gao, Qing Li, Chao Tang, Jinzhi Lei, Yihan Lin, Zemin Zhang for insightful discussions. This work was supported by the National Key Research and Development Program of China (2021YFF1200902) and the National Natural Science Foundation of China (32270689,12126305).

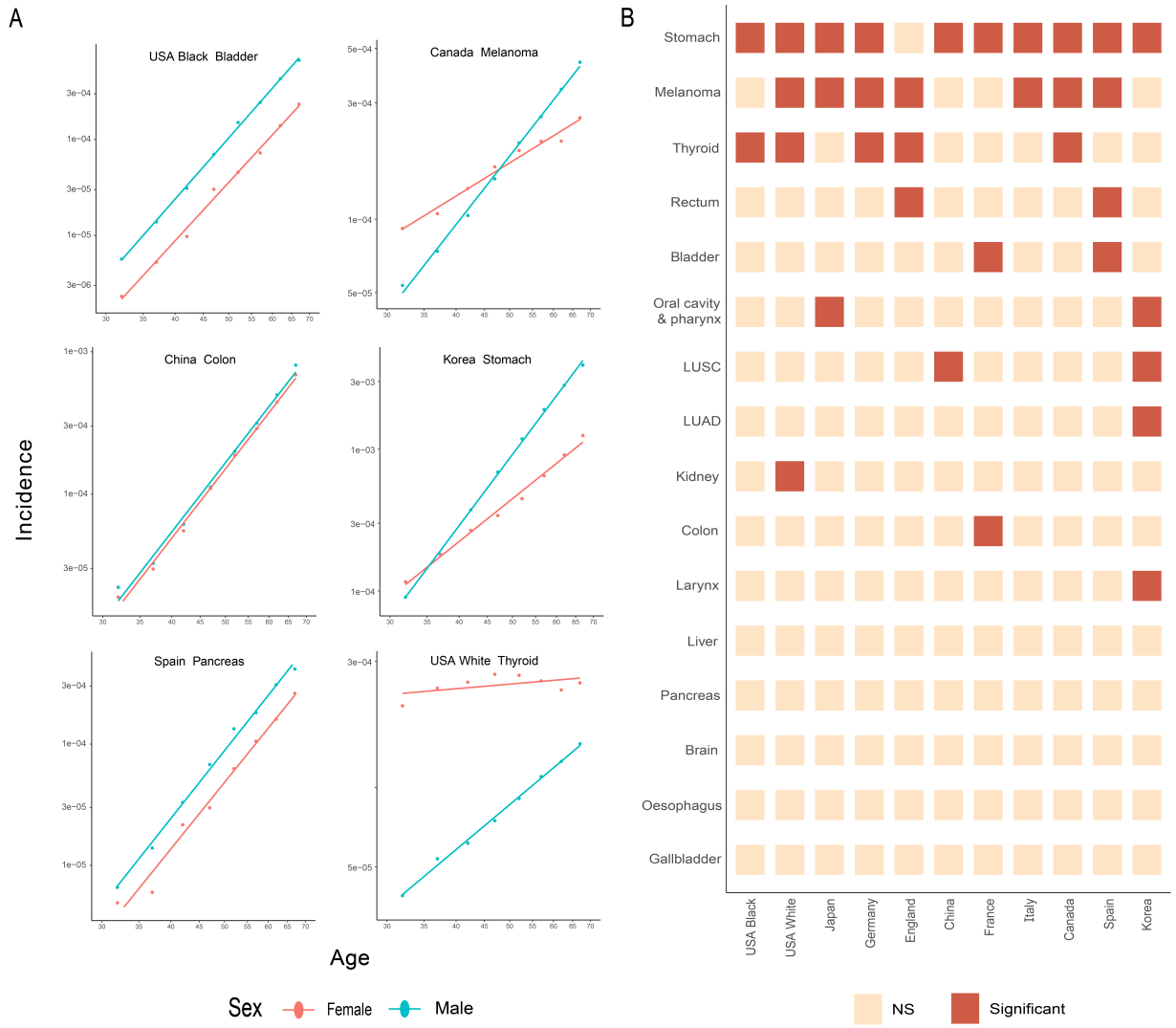
References

1 Nordling, C. O. A New Theory on the Cancer-Inducing Mechanism. *British Journal of Cancer* **7**, 68-72 (1953).
2 Armitage, P. & Doll, R. The Age Distribution of Cancer and a Multi-Stage Theory of Carcinogenesis. *British Journal of Cancer* **8**, 1-12 (1954).
3 Knudson, A. G. Mutation and Cancer - Statistical Study of Retinoblastoma. *P Natl Acad Sci USA* **68**, 820-& (1971).
4 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28, doi:10.1126/science.959840 (1976).
5 Fearon, E. R. & Vogelstein, B. A Genetic Model for Colorectal Tumorigenesis. *Cell* **61**, 759-767 (1990).
6 Michor, F., Iwasa, Y. & Nowak, M. A. Dynamics of cancer progression. *Nat Rev Cancer* **4**, 197-205, doi:10.1038/nrc1295 (2004).
7 Frank, S. A. Age-specific incidence of inherited versus sporadic cancers: A test of the multistage theory of carcinogenesis. *P Natl Acad Sci USA* **102**, 1071-1075 (2005).
8 Meza, R., Jeon, J., Moolgavkar, S. H. & Luebeck, E. G. Age-specific incidence of cancer: Phases, transitions, and biological implications. *P Natl Acad Sci USA* **105**, 16284-16289, doi:10.1073/pnas.0801151105 (2008).
9 Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546-1558 (2013).
10 Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483-1489 (2015).
11 Jackson, S. S. *et al.* Sex disparities in the incidence of 21 cancer types: Quantification of the contribution of risk factors. *Cancer-Am Cancer Soc* **128**, 3531-3540, doi:10.1002/cncr.34390 (2022).
12 Clocchiatti, A., Cora, E., Zhang, Y. & Dotto, G. P. Sexual dimorphism in cancer. *Nat Rev Cancer* **16**, 330-339 (2016).
13 Rubin, J. B. *et al.* Sex differences in cancer mechanisms. *Biol Sex Differ* **11**, doi:10.1186/s13293-020-00291-x (2020).
14 Lopes-Ramos, C. M., Quackenbush, J. & DeMeo, D. L. Genome-Wide Sex and Gender Differences in Cancer. *Front Oncol* **10**, doi:10.3389/fonc.2020.597788 (2020).
15 Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, 1331-+, doi:10.1126/science.aba3066 (2020).
16 Yuan, Y. *et al.* Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. *Cancer Cell* **29**, 711-722, doi:10.1016/j.ccell.2016.04.001 (2016).
17 Dunford, A. *et al.* Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nature Genetics* **49**, 10-16, doi:10.1038/ng.3726 (2017).
18 Li, C. H. *et al.* Sex differences in oncogenic mutational processes. *Nat Commun* **11**, 4330, doi:10.1038/s41467-020-17359-2 (2020).
19 Castro, A. *et al.* Strength of immune selection in tumors varies with sex and age. *Nature Communications* **11** (2020).
20 Doll, R. The Age Distribution of Cancer: Implications for Models of Carcinogenesis. *Journal of the Royal Statistical Society. Series A (General)* **134**, 133, doi:10.2307/2343871 (1971).
21 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
22 Zhou, X. *et al.* Exploring genomic alterations in pediatric cancer using ProteinPaint. *Cancer Res* **76** (2016).
23 Mas-Ponte, D., McCullough, M. & Supek, F. Spectrum of DNA mismatch repair failures viewed through the lens of cancer genomics and implications for therapy. *Clin Sci (Lond)* **136**, 383-404, doi:10.1042/CS20210682 (2022).
24 Akagi, K. *et al.* Real-world data on microsatellite instability status in various unresectable or metastatic solid tumors. *Cancer Science* **112**, 1105-1113, doi:10.1111/cas.14798 (2021).
25 Kaplan, A. R. & Glazer, P. M. Impact of hypoxia on DNA repair and genome integrity. *Mutagenesis* **35**, 61-67, doi:10.1093/mutage/gez019 (2020).
26 Mihaylova, V. T. *et al.* Decreased expression of the DNA mismatch repair gene Mlh1 under hypoxic stress in mammalian cells. *Mol Cell Biol* **23**, 3265-3273, doi:Doi 10.1128/Mcb.23.9.3265-3273.2003 (2003).

27 Lu, Y., Wajapeyee, N., Mitchell & Peter. Silencing of the DNA Mismatch Repair Gene MLH1 Induced by Hypoxic Stress in a Pathway Dependent on the Histone Demethylase LSD1. *Cell Reports* **8**, 501-513, doi:10.1016/j.celrep.2014.06.035 (2014).
28 Chen, H. B., Yan, Y., Davidson, T. L., Shinkai, Y. & Costa, M. Hypoxic stress induces dimethylated histone H3 lysine 9 through histone methyltransferase G9a in mammalian cells. *Cancer Res* **66**, 9009-9016, doi:10.1158/0008-5472.Can-06-0101 (2006).
29 Duan, C. Hypoxia-inducible factor 3 biology: complexities and emerging themes. *Am J Physiol Cell Physiol* **310**, C260-269, doi:10.1152/ajpcell.00315.2015 (2016).
30 Yang, J., Harris, A. L. & Davidoff, A. M. Hypoxia and Hormone-Mediated Pathways Converge at the Histone Demethylase KDM4B in Cancer. *International Journal of Molecular Sciences* **19**, doi:10.3390/ijms19010240 (2018).
31 Thambirajah, A. A. *et al.* Disruption by stealth-Interference of endocrine disrupting chemicals on hormonal crosstalk with thyroid axis function in humans and other animals. *Environ Res* **203**, doi:10.1016/j.envres.2021.111906 (2022).
32 Ogino, S. & Goel, A. Molecular classification and correlates in colorectal cancer. *J Mol Diagn* **10**, 13-27, doi:10.2353/jmoldx.2008.070082 (2008).
33 Mal, R. *et al.* Estrogen Receptor Beta (ER beta): A Ligand Activated Tumor Suppressor. *Front Oncol* **10**, doi:10.3389/fonc.2020.587386 (2020).
34 Lourbopoulos, A. I., Mourouzis, I. S., Trikas, A. G., Tseti, I. K. & Pantos, C. I. Effects of Thyroid Hormone on Tissue Hypoxia: Relevance to Sepsis Therapy. *J Clin Med* **10**, doi:10.3390/jcm10245855 (2021).
35 Ren, B. & Zhu, Y. A New Perspective on Thyroid Hormones: Crosstalk with Reproductive Hormones in Females. *Int J Mol Sci* **23**, doi:10.3390/ijms23052708 (2022).
36 Tsutsumi, R. & Webster, N. J. G. GnRH Pulsatility, the Pituitary Response and Reproductive Dysfunction. *Endocr J* **56**, 729-737, doi:DOI 10.1507/endocrj.K09E-185 (2009).
37 Taylor, P. N. *et al.* Global epidemiology of hyperthyroidism and hypothyroidism. *Nat Rev Endocrinol* **14**, 301-316, doi:10.1038/nrendo.2018.18 (2018).
38 Krashin, E., Piekielko-Witkowska, A., Ellis, M. & Ashur-Fabian, O. Thyroid Hormones and Cancer: A Comprehensive Review of Preclinical and Clinical Studies. *Front Endocrinol (Lausanne)* **10**, 59, doi:10.3389/fendo.2019.00059 (2019).
39 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *Ca-a Cancer Journal for Clinicians* **69**, 7-34 (2019).
40 Nunney, L. Commentary: The multistage model of carcinogenesis, Peto's paradox and evolution. *International Journal of Epidemiology* **45**, 649-U667, doi:10.1093/ije/dyv201 (2016).
41 Moolgavkar, S. H. Commentary: Fifty years of the multistage model: remarks on a landmark paper. *Int J Epidemiol* **33**, 1182-1183, doi:10.1093/ije/dyh288 (2004).
42 Tenenbaum, D. & Maintainer, B. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version 1.38.0. (2022).
43 Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27-30, doi:DOI 10.1093/nar/28.1.27 (2000).
44 Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).
45 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).
46 van de Haar, J. *et al.* Identifying Epistasis in Cancer Genomes: A Delicate Affair. *Cell* **177**, 1375-1383, doi:10.1016/j.cell.2019.05.005 (2019).
47 Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology* **38**, 675-678, doi:10.1038/s41587-020-0546-8 (2020).
48 Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* **8**, 15180, doi:10.1038/ncomms15180 (2017).

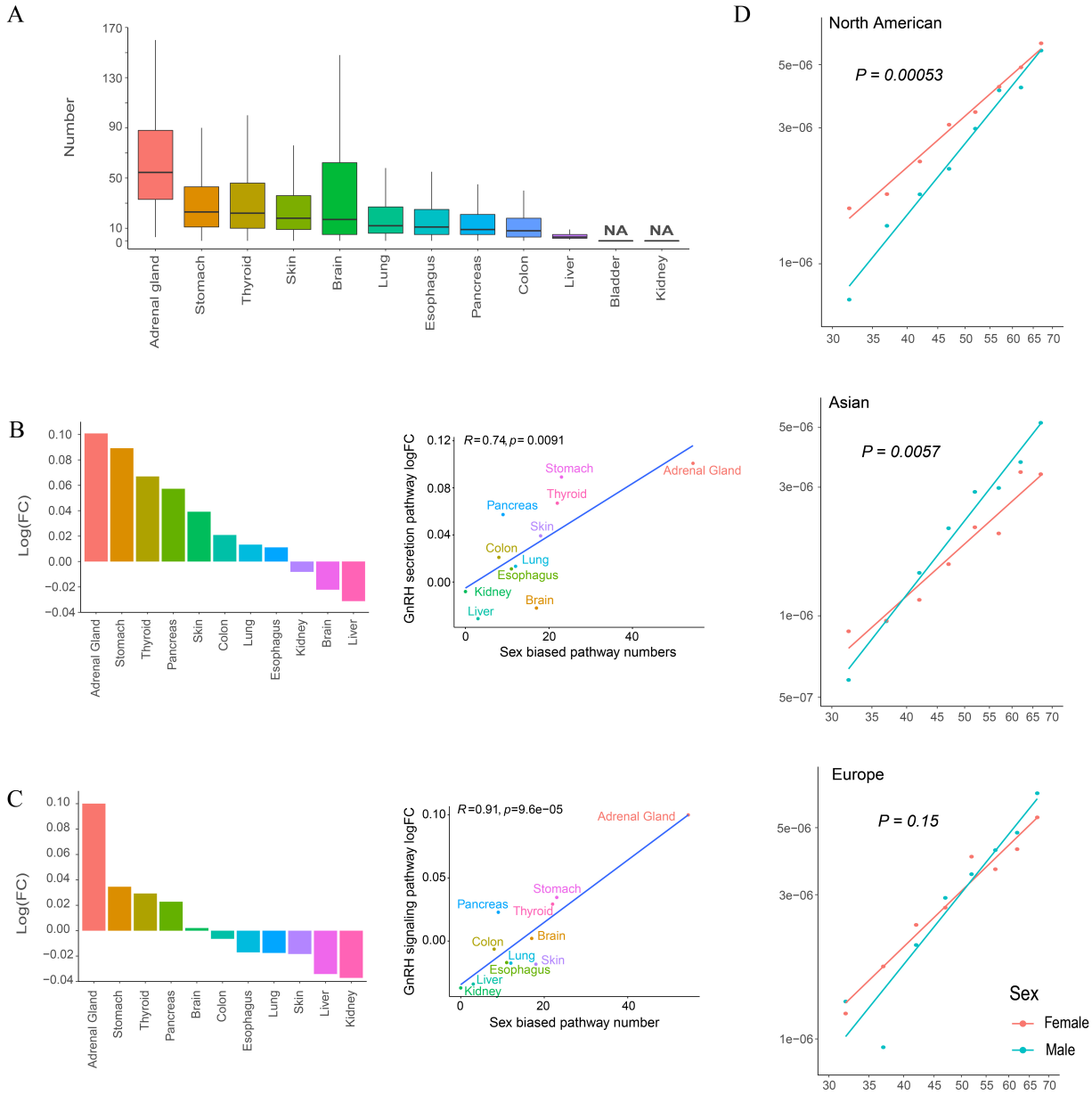
Figures and Tables

Figure 1. Curve patterns and prevalence of nonparallel patterns.



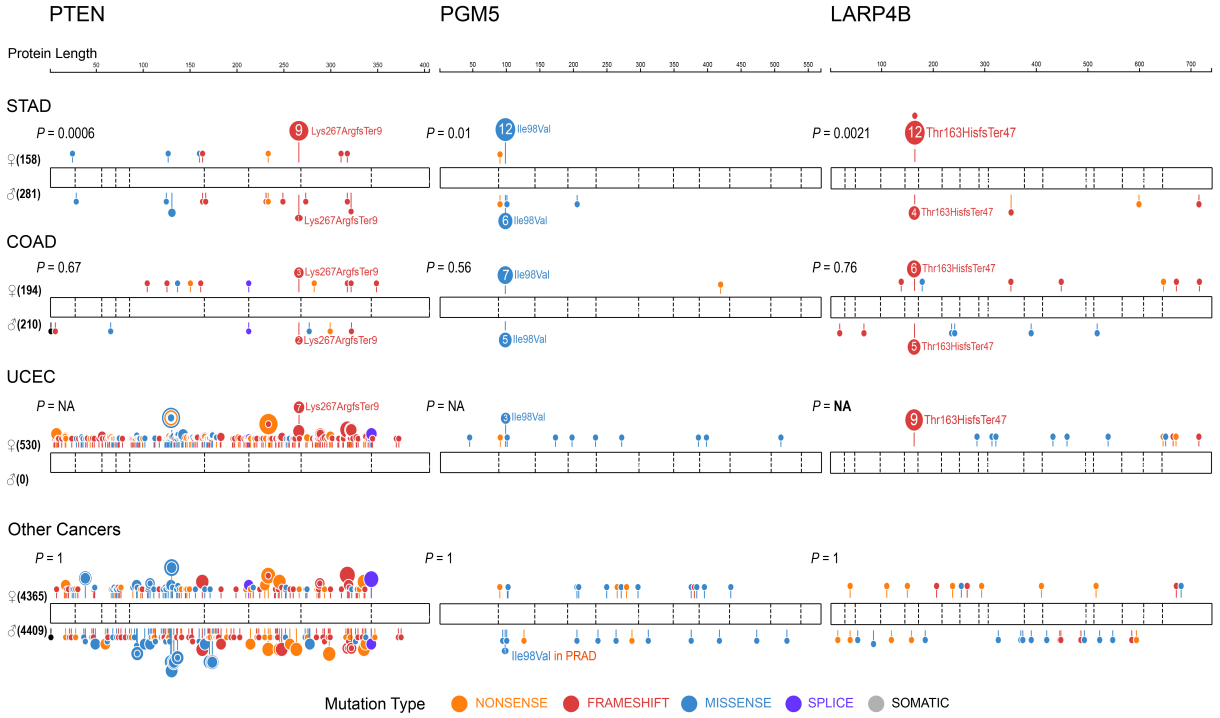
(A) Examples of parallel and nonparallel patterns. Both axes are in log scale. Parallel patterns are shown on the left; nonparallel pattern are shown on the right. Female incidences and fitted lines are colored as red; male incidences and fitted lines are colored as blue. (B) Prevalence of significant nonparallel patterns in major cancer types across regions or races. Patterns tested as significant nonparallel are colored as red, not significant (NS) nonparallel are colored as light yellow.

Figure 2. Sex bias in tissues and nonparallel patterns of adrenal gland cancer.



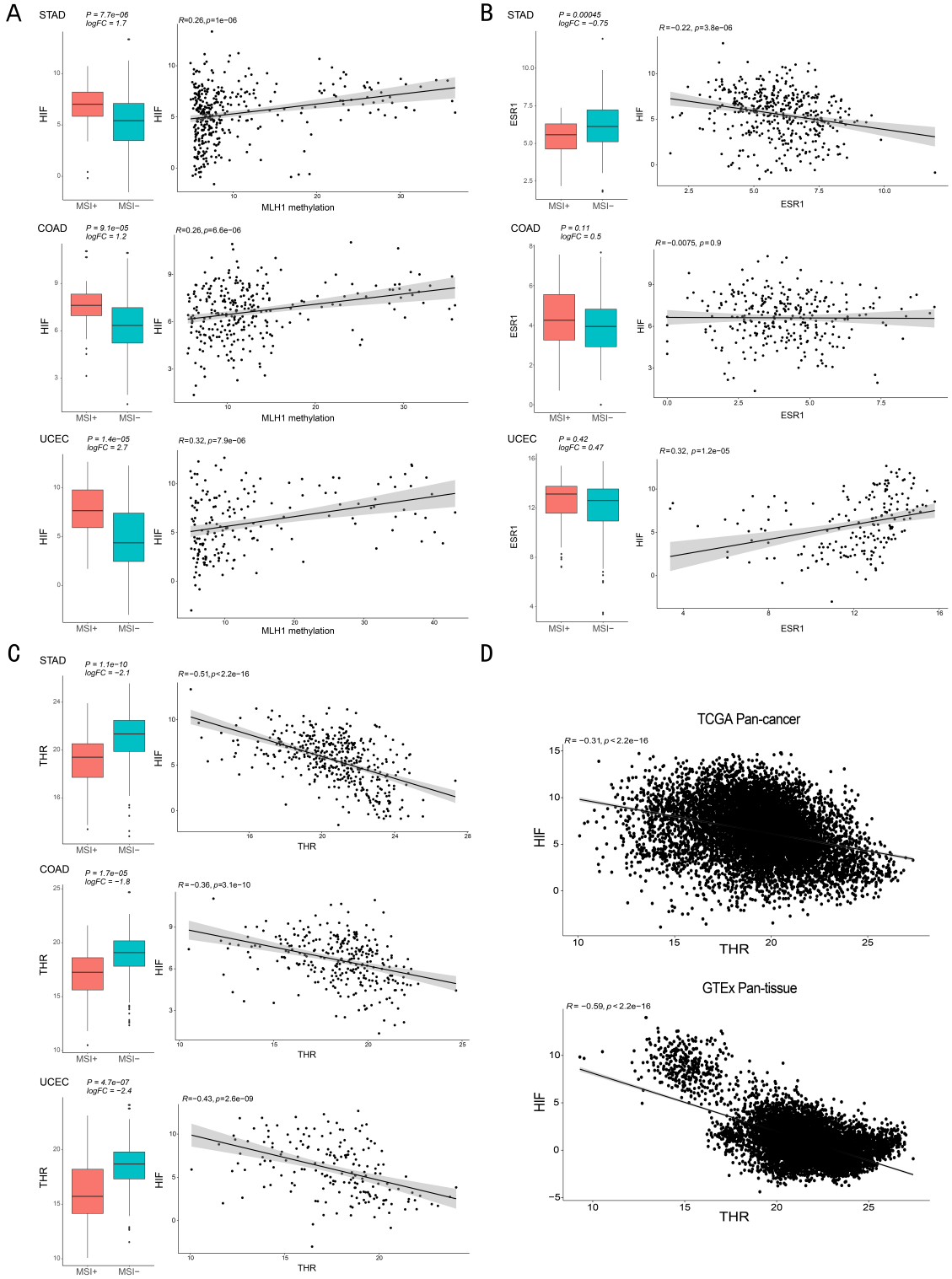
(A) Number of sex-biased expressed pathways in each tissues. The results of bladder and kidney were marked as not available (NA) because of small sizes of samples. (B and C) Female to male log-fold change and its correlation with sex-biased expressed pathway numbers in tissues of GnRH secretion pathway and GnRH signaling pathway among tissues. The results of bladder were excluded because of small size of samples. (D) Nonparallel patterns of adrenal gland cancer. Regional cohorts in one continent were merged as one continent cohort.

Fig. 3: Mutation spectra of PTEN, PGM5 and LARP4B.



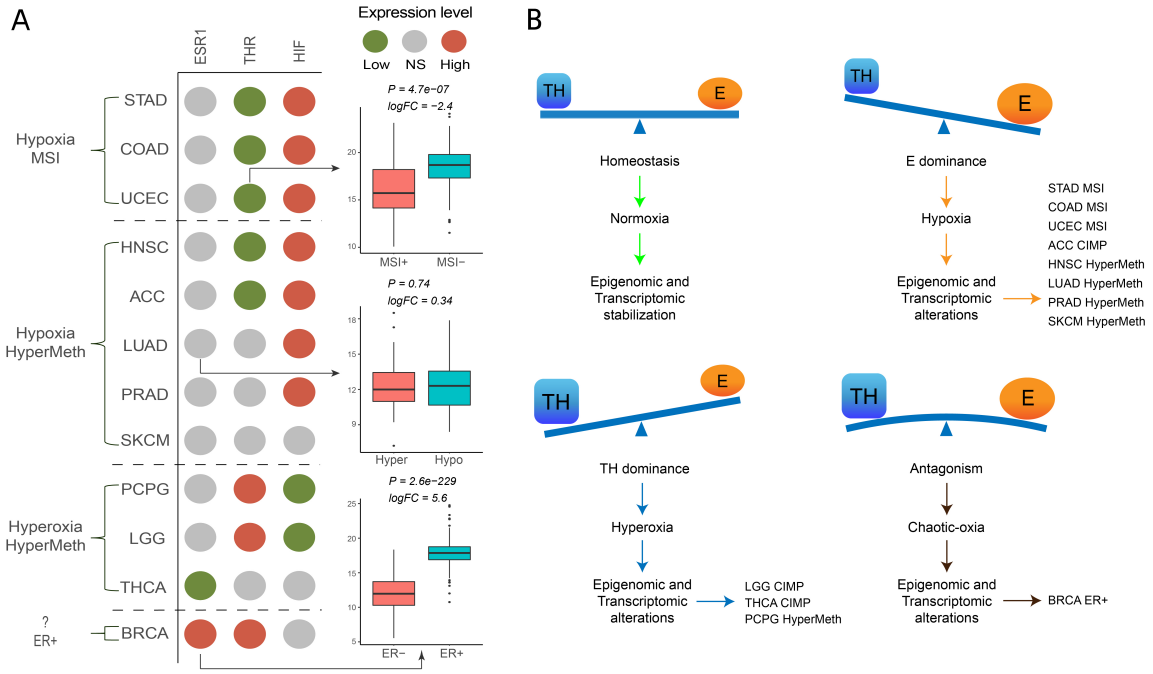
The P values above the gene body on the left side denote the significance of hotspot sex bias. UCEC is only seen in females and the P value is marked as NA. The number of female and male samples in each cancer are indicated on the left of each row. Types of mutation are labeled using different colors and are annotated on the bottom the figure.

Fig. 4: Correlation between HIF, ESR1 and THR.



(A) the comparison of HIF levels between MSI+ and MSI- samples, and the correlation between HIF and MLH1 methylation across STAD, COAD, and UCEC. (B) the comparison of ESR1 levels between MSI+ and MSI- samples, and the correlation between HIF and ESR1 across STAD, COAD, and UCEC. (C) the comparison of THR levels between MSI+ and MSI- samples, and the correlation between HIF and THR across STAD, COAD, and UCEC. (D) the correlation between HIF and THR at the pan-cancer and pan-tissue levels. In all dot plots, R represents the Pearson correlation coefficient, and p values were obtained using Student's t-test.

Fig. 5: HIF, ESR1 and THR expression signatures in TCGA cancers and Imbalanced estrogen and TH model of carcinogenesis.



(A) The comparative expression signatures of ESR1, THR and HIF were represented by colored dots. Red dots indicated significantly elevated expression, green dots indicated significantly reduced expression and grey dots indicated no significant difference in expression when comparing MSI+, HyperMeth and ER+ cancers to control cancers. Boxplots on the right provided examples to explain the dots. LGG: Brain Lower Grade Glioma. THCA: Thyroid carcinoma. (B) The relationship between estrogen (E) and TH is illustrated by a seesaw diagram. At homeostasis state, the levels of E and TH are balanced. When the balance is disrupted, either through elevated levels of E or TH, or through elevated of both hormones, epigenomic and transcriptomic alterations can occur in tissue cells. These alterations can lead to changes in the tissue environment and act as a selective pressure, inducing specific subtypes of cancer.

Table 1. Sex-biased mutated genes.

Gene	Chromosome	Cancer	Female samples +/-	Male samples +/-	Relative risk (f/m)	LRT P value
CASP8	2	HNSC	23/108	19/332	3.24	3.85e-10
RBM10	X	LUAD	9/255	18/205	0.42	3.52e-4
SELP	1	SKCM	13/149	8/249	2.58	7.71e-3
SPAG17	1		17/145	13/244	2.07	7.71e-3
DNAH7	2	STAD	13/131	11/259	2.21	4.70e-2
LARP4B	10		13/131	8/262	3.05	4.27e-2
PGM5	9		13/131	8/262	3.05	4.27e-2
PTEN	10		14/130	12/258	2.19	4.70e-2

"Female samples +/-" and "Male samples +/-" refer to the number of female or male samples with or without a sex-biased gene mutation. The "Relative risk (f/m)" represents the female-to-male relative risk of gene mutation frequency. The "LRT P value" is the P value obtained from the likelihood ratio test and adjusted by the Benjamini-Hochberg method.

Supplementary Figures

Fig. S1. Mutation spectra of the other sex-biased mutated genes in corresponding cancers.

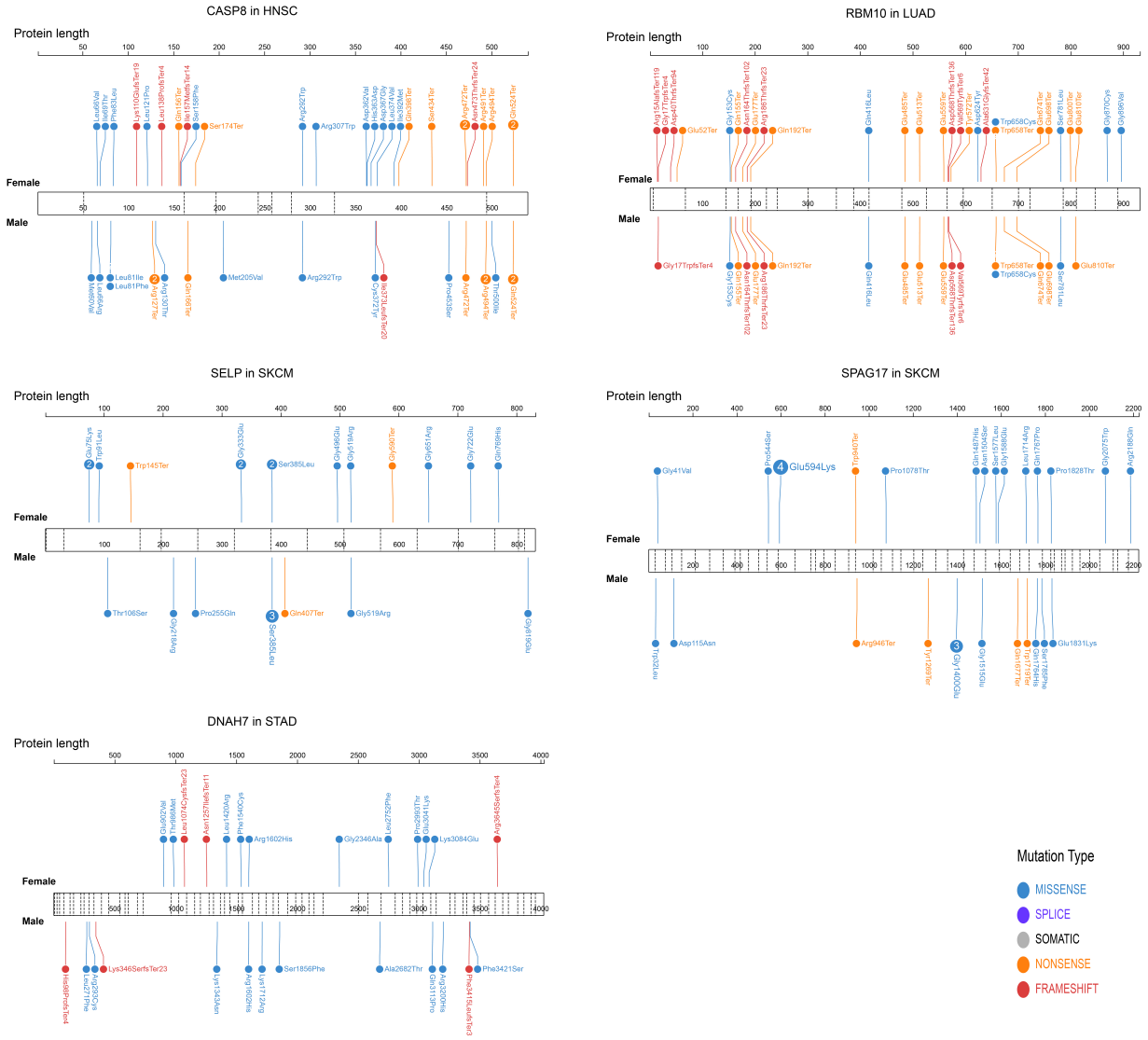
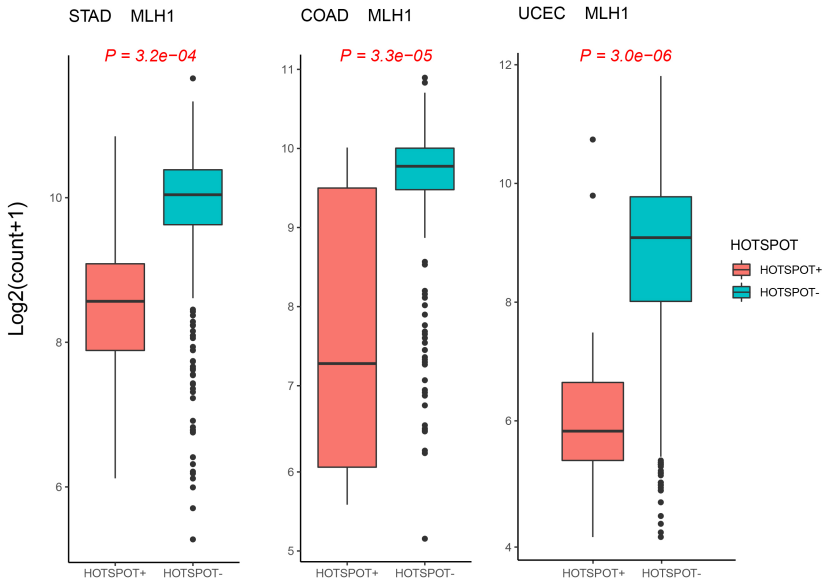
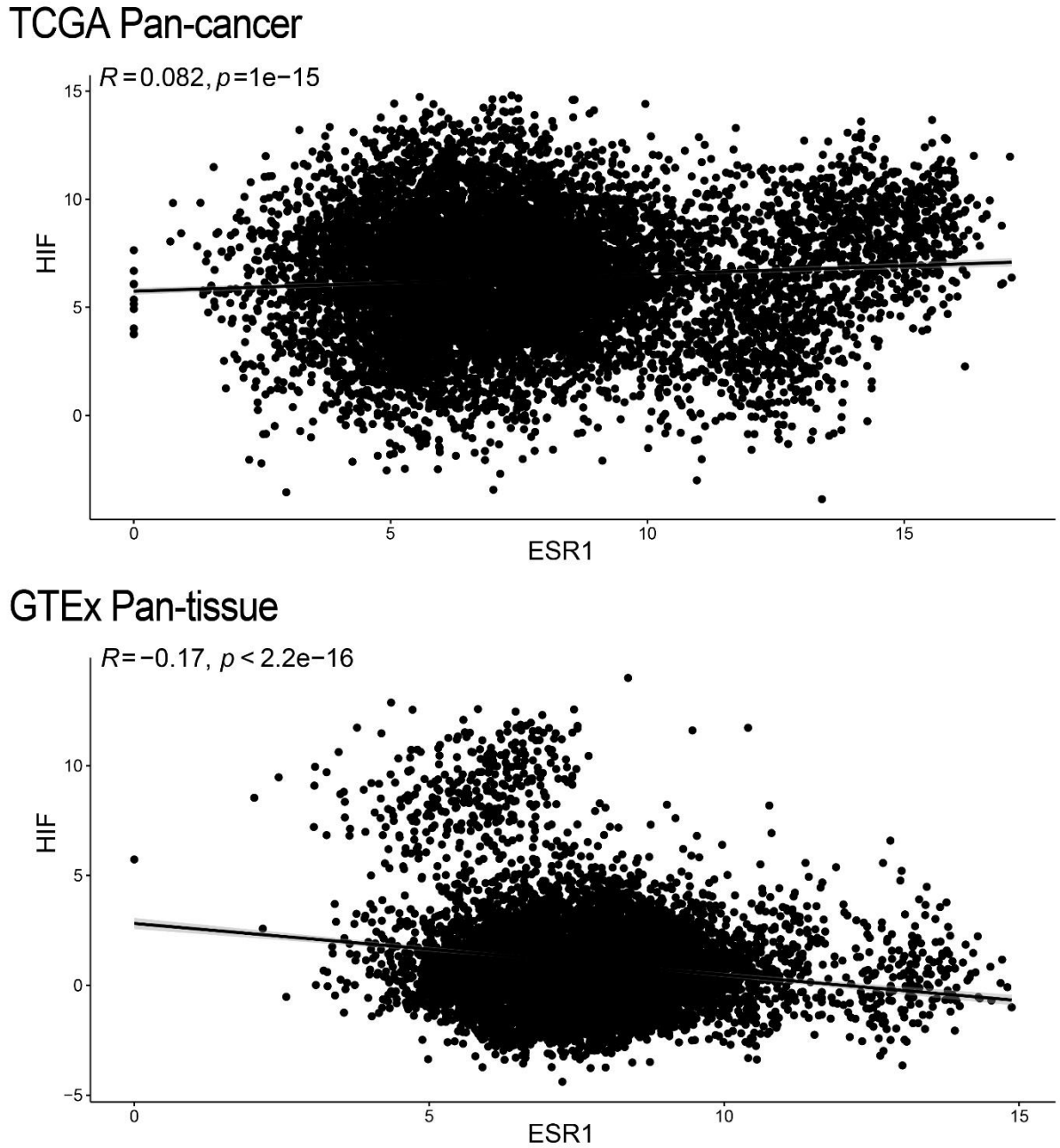


Fig. S2. MLH1 expression between HOTSPOT+ and HOTSPOT- groups.



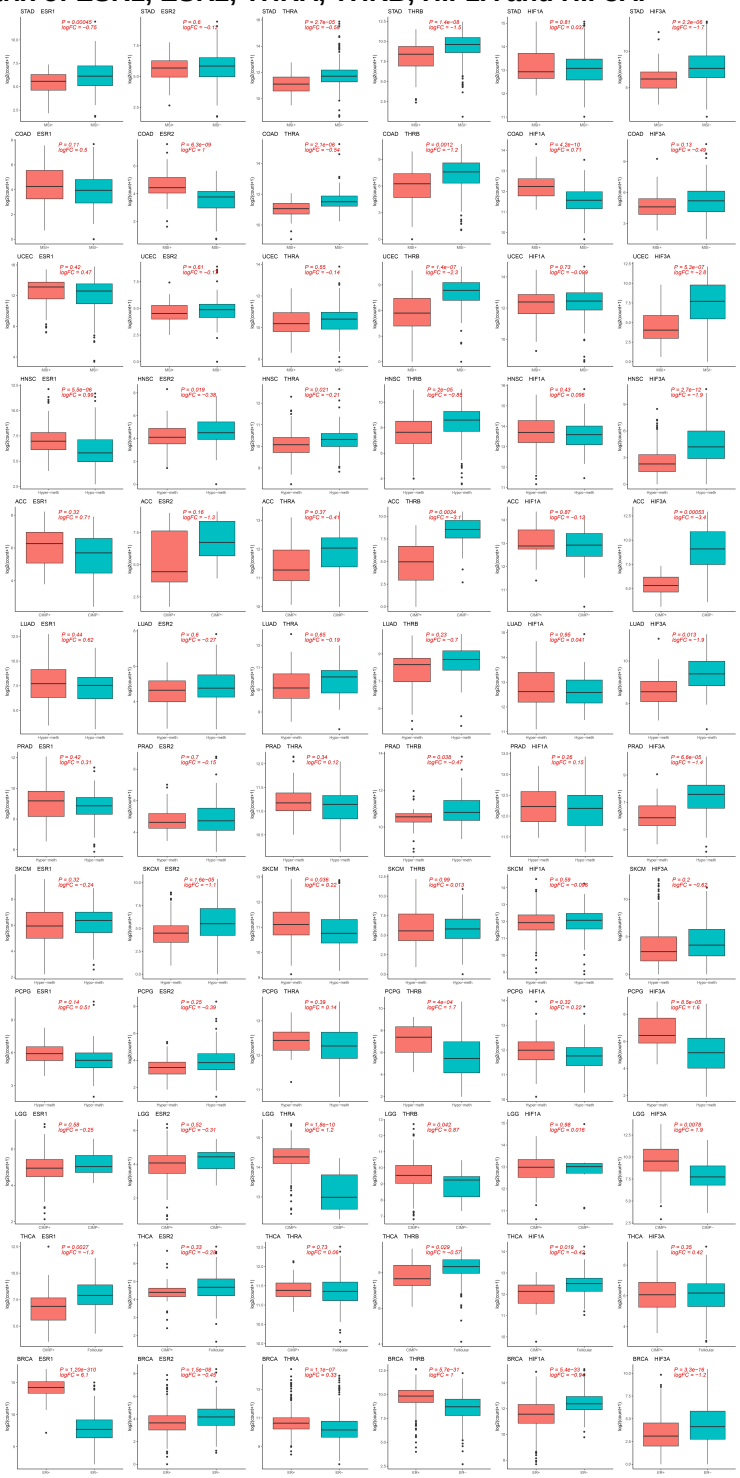
DEGs analysis were conducted between HOTSPOT+ and HOTSPOT- groups in STAD, COAD and UCEC using limma. MLH1 expression are shown in figure. P values shown are adjusted P values given by limma.

Fig. S3. Correlation between HIF and ESR1.



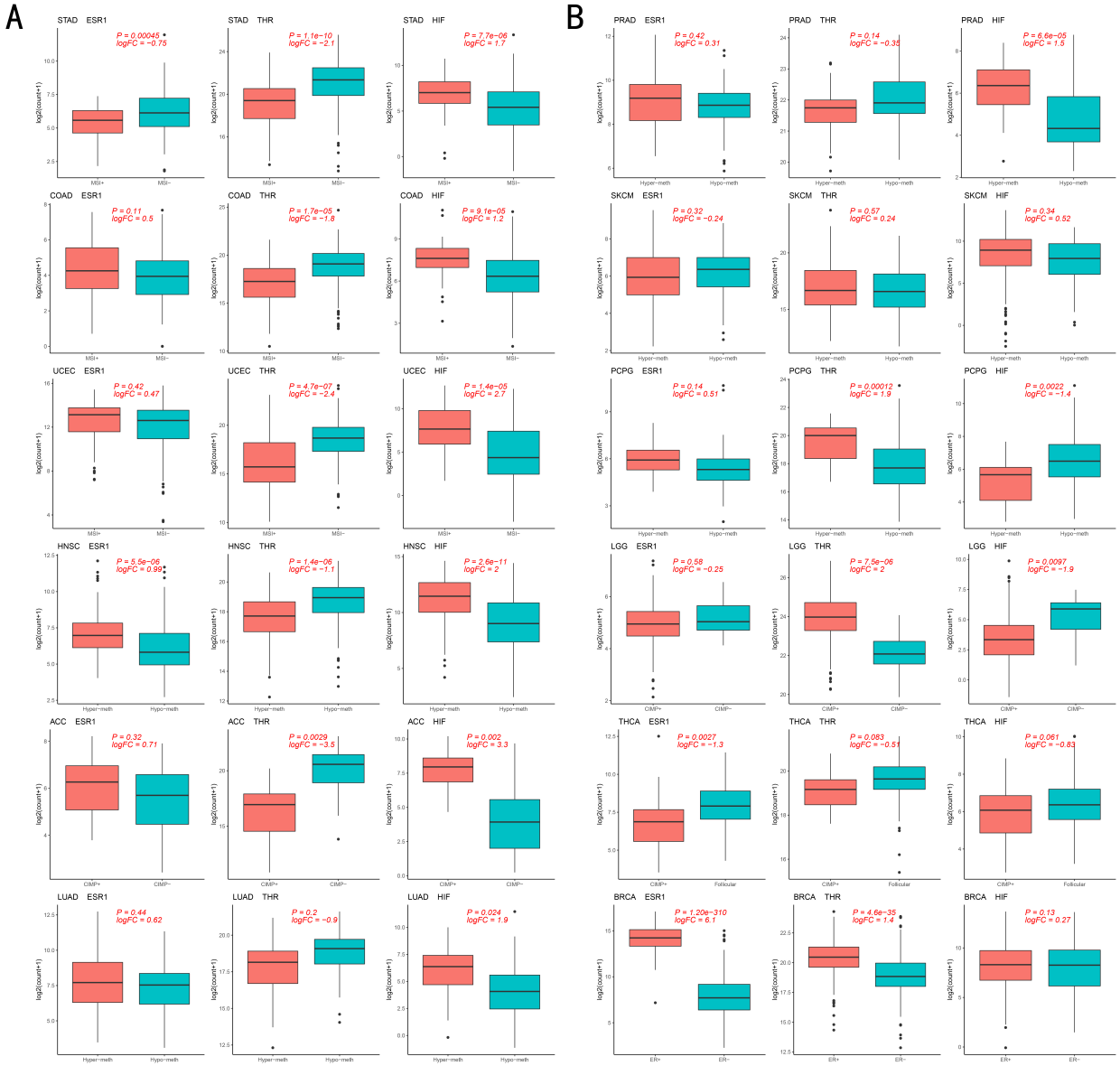
The correlation between HIF and ESR1 was shown in the TCGA pan-cancer level and in GTEx pan-tissue level. R represents the Pearson correlation coefficient, and p values were obtained using Student's t-test.

Fig. S4. Boxplot matrix of ESR1, ESR2, THRA, THRB, HIF1A and HIF3A.



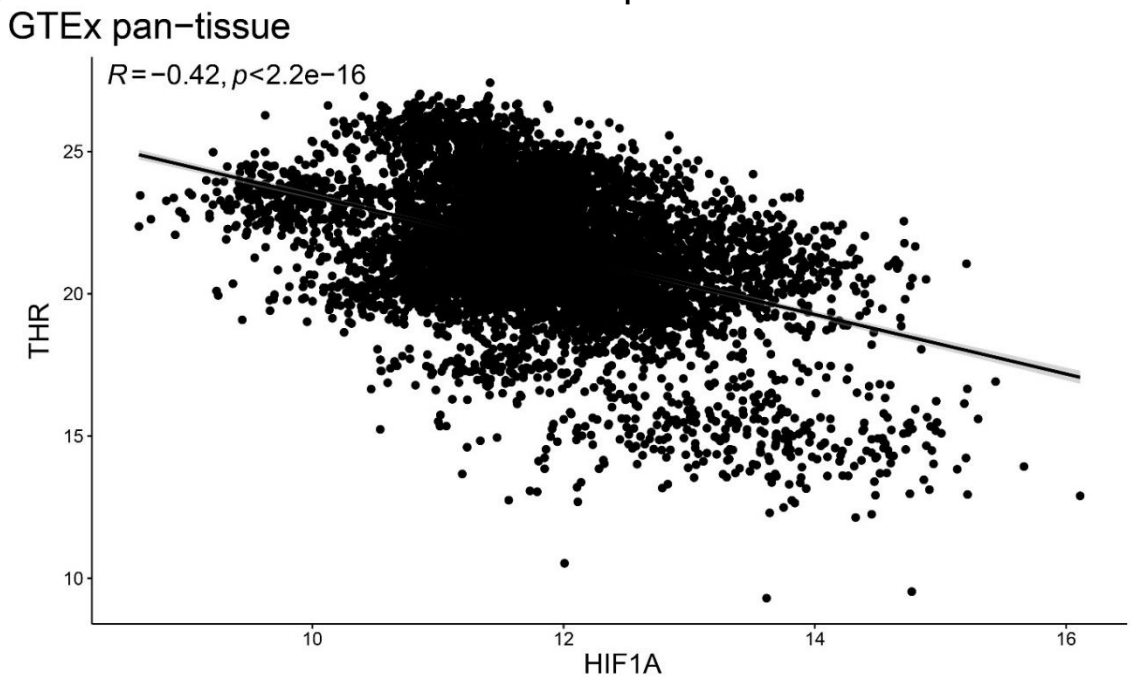
Boxplots of the six genes referred to in the main text, provide a comparison of gene expression between two groups: MSI+ versus MSI-, HyperMeth versus HypoMeth, and ER+ versus ER-. The adjusted P value and log fold change shown in figure were obtained using limma.

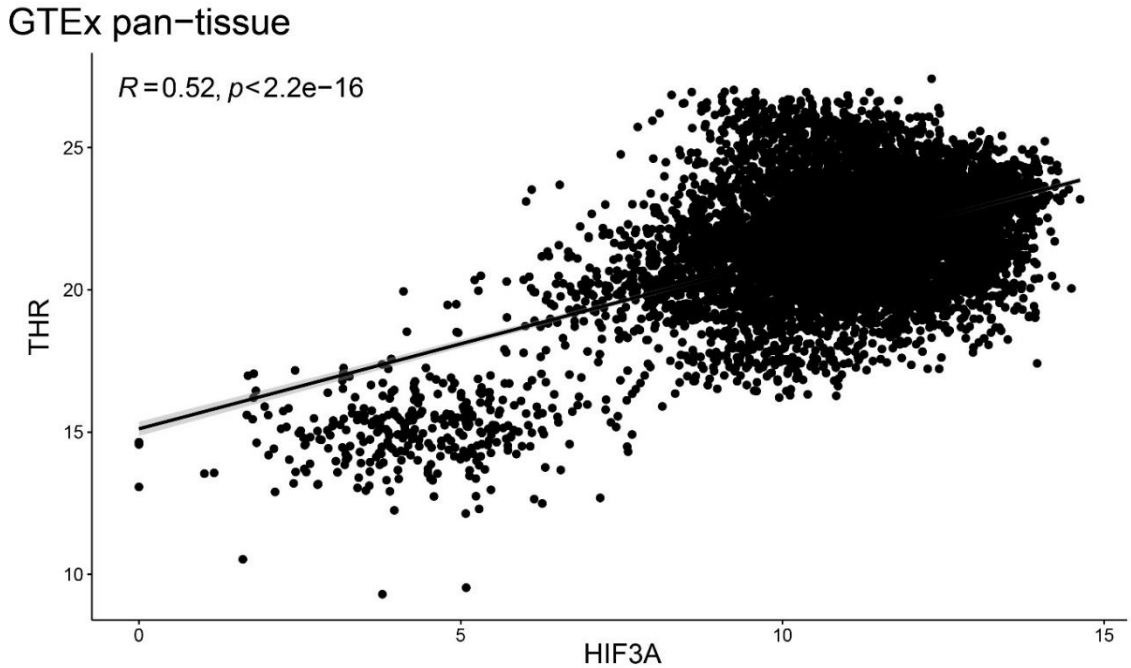
Fig. S5. Boxplot matrix corresponding to Fig. 5A.



Each boxplot was aligned with each dot in figure 5A, representing a comparison of gene expression between two groups: MSI+ versus MSI-, HyperMeth versus HypoMeth, and ER+ versus ER-. The adjusted P value and log fold change shown in figure were obtained using limma. (A) Dot plots from STAD to LUAD. (B) Dot plots from PRAD to BRCA.

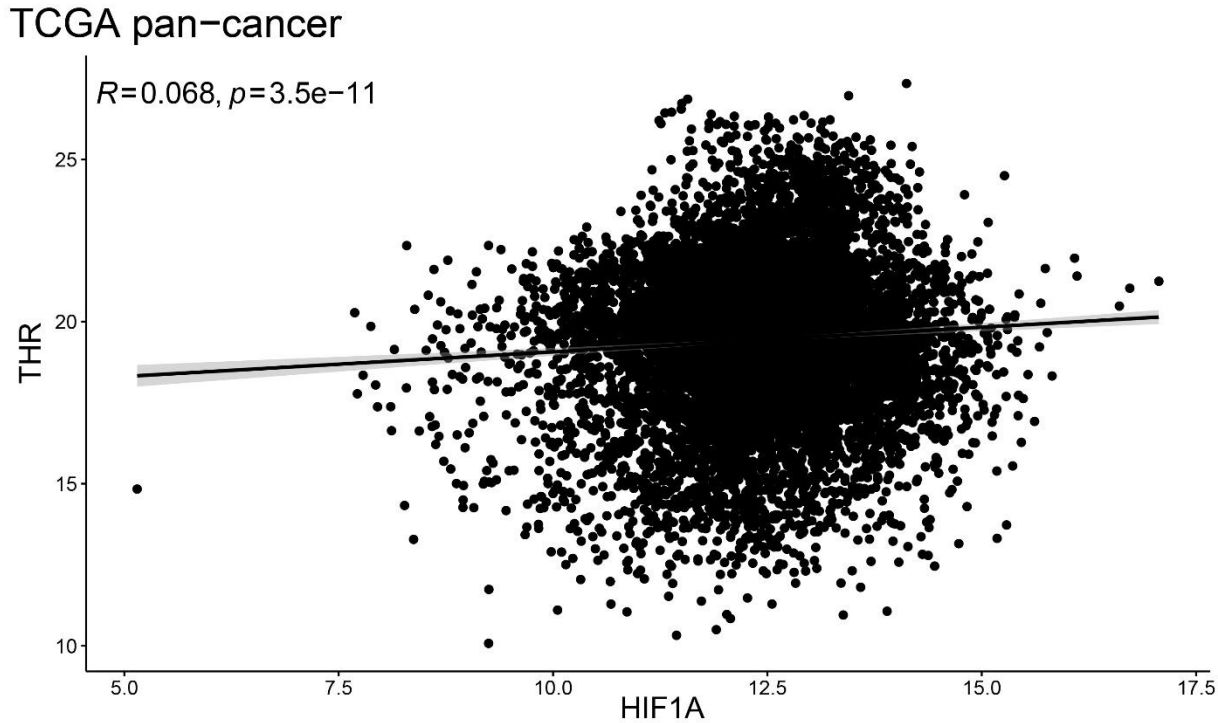
Fig. S6. Correlation of THR with HIF1A and HIF3A at pan-tissue levels.

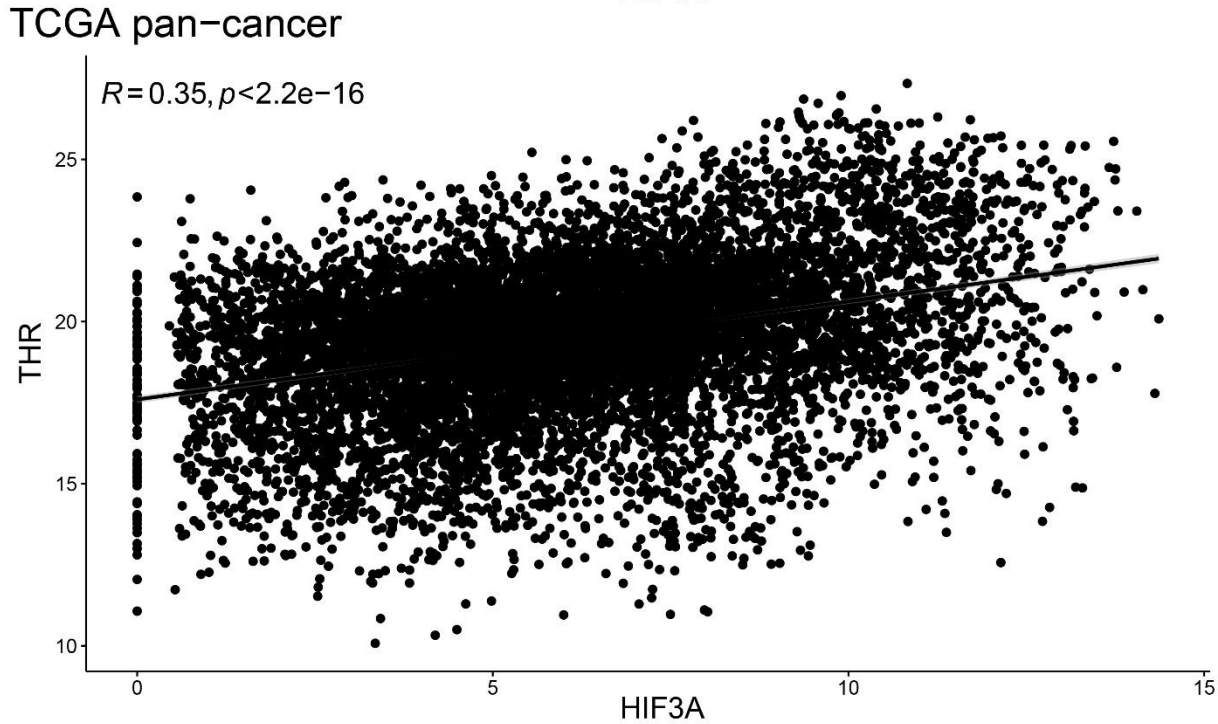




R represents the Pearson correlation coefficient, and p values were obtained using Student's t-test.

Fig. S7. Correlation of THR with HIF1A and HIF3A at pan-cancer levels.





R represents the Pearson correlation coefficient, and p values were obtained using Student's t-test.

Supplementary Tables

Table S1. Sample sizes of each tissue in the GTEx data

Tissue	Female	Male	Total
Bladder	3	6	9
Kidney	5	22	27
Liver	33	77	110
Skin	190	366	556
Colon	124	180	304
Stomach	73	99	172
Pancreas	72	93	165
Lung	100	187	287
Thyroid	105	173	278
Brain	353	783	1136
Esophagus	234	415	649
Adrenal gland	57	70	127

Table S2. Test results for sex-biased expression of GnRH secretion pathway across each tissue in the GTEx data

Tissue	P value	LogFC
Adrenal Gland	0.030899	0.100831
Stomach	0.012977	0.08919
Thyroid	0.009457	0.066973
Pancreas	0.112046	0.057273
Skin	0.068307	0.039208
Colon	0.414889	0.020896
Lung	0.565389	0.013314
Esophagus	0.533371	0.011119
Kidney	0.955948	-0.00815
Brain	0.269388	-0.02215
Liver	0.579539	-0.03118

P values shown are adjusted P values given by limma.

Table S3. Test results for sex-biased expression of GnRH signaling pathway across each tissue in the GTEx data.

Tissue	P value	LogFC
Adrenal Gland	0.002109	0.100018
Stomach	0.221583	0.034591
Thyroid	0.167189	0.029276
Pancreas	0.409804	0.022836
Brain	0.885571	0.002091
Colon	0.683874	-0.00639
Esophagus	0.196701	-0.01705
Lung	0.402423	-0.0175
Skin	0.217595	-0.01834
Liver	0.482774	-0.03411
Kidney	0.702808	-0.03714

P values shown are adjusted P values given by limma.

Table S4. Co-occurrence and exclusive test for three hotspots

Hotspot1	Hotspot2	Co-occu samples	Hotspot1 only samples	Hotspot2 only samples	No hotspot samples	P value	Direction
PTEN_267	PGM5_98	5	22	22	1162	2.2e-4	co-occu
PTEN_267	LARP4B_163	5	22	20	1164	1.5e-4	co-occu
PGM5_98	LARP4B_163	5	22	20	1164	1.5e-4	co-occu

Fisher exact test is performed on 3 pairs of the 3 hotspot mutations. All pairs trend to co-occurrence as the “Direction” column listed.

Table S5. MSI subtype proportions in TCGA.

Cancer	Sex	MSI+/ HyperMeth/ ER+ samples	Total samples	Ratio	Relative risk(f/m)	P value
STAD	Male	40	252	0.15873	1.923664	0.0013
	Female	40	131	0.305344		
COAD	Male	30	172	0.174419	1.235928	0.412533
	Female	36	167	0.215569		
UCEC	Female	124	507	0.244576	NA	NA
LGG	Male	132	135	0.977778	0.93905	0.063995
	Female	101	110	0.918182		
GBM	Male	10	17	0.588235	0.772727	0.760509
	Female	5	11	0.454545		
ACC	Male	7	16	0.4375	0.825397	0.830716
	Female	13	36	0.361111		
THCA	Male	13	42	0.309524	0.590237	0.145401
	Female	19	104	0.182692		
SKCM	Male	96	204	0.470588	0.962891	0.843433
	Female	58	128	0.453125		
PCPG	Male	46	60	0.766667	1.032609	0.893141
	Female	57	72	0.791667		
HNSC	Male	95	203	0.46798	0.702909	0.050884
	Female	25	76	0.328947		
LUAD	Male	35	56	0.625	1.134545	0.460331
	Female	39	55	0.709091		
PRAD	Male	182	333	0.546547	NA	NA
BRCA	Female	770	1043	0.738255	NA	NA

P values are given by proportion z-test.

Table S6. MSI subtype proportions in Akagi's study[24].

Cancer	Sex	MSI+ samples	Total samples	Ratio	Relative risk(f/m)	P value
Gastric	Male	60	1250	4.8%	1.63	0.013
	Female	45	574	7.8%		
Colorectal	Male	186	5676	3.3%	1.37	0.0026
	Female	185	4138	4.5%		
Endometrial	Female	234	1389	16.8%	NA	NA

Female MSI subtype proportions are higher than male's in Gastric and Colorectal cancer, and are both significant. P values are given by proportion z-test.

Table S7 Mutation distributions of sex-biased genes

Gene	Cancer	Group	Mutated samples	Wild type samples	P value
CASP8	HNSC	HyperMeth	17	142	7.97e-4
		HypoMeth	1	119	
RBM10	LUAD	HyperMeth	3	34	0.398
		HypoMeth	3	71	
SELP	SKCM	HyperMeth	17	161	0.0117
		HypoMeth	4	150	
SPAG17	SKCM	HyperMeth	10	168	0.387
		HypoMeth	13	141	
DNAH7	STAD	MSI+	12	68	3.45e-4
		MSI-	10	293	
PGM5	STAD	MSI+	18	62	1.73e-12
		MSI-	1	302	
LARP4B	STAD	MSI+	17	63	7.56e-11
		MSI-	2	301	
PTEN	STAD	MSI+	12	68	9.79e-05
		MSI-	8	295	

Mutation distributions of sex-biased genes between comparing groups were listed in table. P values are given by Fisher exact test. Hotspot mutations of PGM5, LARP4B and PTEN are exclusive occurred in MSI+ cancers as discussed in main tex.