Sex biases in cancer and autoimmune disease incidence are strongly positively correlated with mitochondrial gene expression across human tissues

David R. Crawford, PhD, MSc[a,b]; Sanju Sinha, PhD[a]; Nishanth Ulhas Nair, PhD[a]; Bríd M. Ryan, PhD, MPH[c]; Jill S. Barnholtz-Sloan, PhD[d,e]; Stephen M. Mount, PhD[b]; Ayelet Erez, MD, PhD[f]; Ken Aldape, MD[g]; Philip E. Castle, PhD, MPH[e,h]; Padma S. Rajagopal, MD, MPH, MSc[a,i]; Chi-Ping Day, PhD[j]; Alejandro A. Schäffer, PhD[a]; Eytan Ruppin, MD, PhD[a,*]

[a] Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD

[b] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD

[c] Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD

[d] Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health, Bethesda, MD

[e] Trans-Divisional Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD

[f] Department of Biological Regulation, Weizmann Institute of Science, Rehovot, Israel

[g] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD

[h] Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, Bethesda, MD

[i] Women's Malignancies Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD

[j] Laboratory of Cancer Biology and Genetics, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD

[*]Corresponding author, eytan.ruppin@nih.gov
Building 15C1, Bethesda, MD 20892; (240)-858-3169

Word count of Main Text, including figure captions: 5007
Keywords: Cancer incidence, autoimmune disease incidence, sex bias, mitochondria

The authors declare no potential conflicts of interest.

Abstract

Cancer occurs more frequently in men while autoimmune diseases (AIDs) occur more frequently in women. To explore whether these sex biases have a common basis, we collected 167 AID incidence studies from many countries for tissues that have both a cancer type and an AID that arise from that tissue. Analyzing a total of 182 country-specific, tissue-matched cancer-AID incidence rate sex bias data pairs, we find that, indeed, the sex biases observed in the incidence of AIDs and cancers that occur in the same tissue are positively correlated across human tissues. The common key factor whose levels across human tissues are most strongly associated with these incidence rate sex biases is the sex bias in the expression of the 37 genes encoded in the mitochondrial genome.

Introduction

Both autoimmune diseases (AIDs) and cancers have notably sex-biased incidence rates. Most AIDs occur more often in women [1, 2], and most cancers occur more often in men [3, 4, 5]. While sex differences in several key biological factors have been implicated in the biased incidence rates observed for both AIDs and cancer, including inflammation and immunity, metabolism and sex hormones, their mechanistic underpinnings remain largely unexplained [2, 6, 7].

Given these observations, we asked whether the sex biases observed in the incidence of AIDs and cancers that occur in the same tissue are correlated *across* human tissues. This question is of fundamental interest, since an affirmative answer may suggest that there are common factors underlying their incidence. Establishing such a link between AIDs and cancers

could further prompt researchers to explore how pertaining findings in AIDs could cross fertilize cancer risk studies and and vice versa, potentially enhancing our ability to prevent and treat these diseases.

To explore whether these sex biases are correlated across human tissues, we collected population-based AID incidence studies for tissues that have both a cancer type and an AID that arise from that tissue. For countries for which we collected AID incidence data, we gathered incidence data for corresponding cancer types from national cancer registries. Analyzing a total of 182 country-specific, tissue-matched cancer-AID incidence rate sex bias data pairs, we find that the incidence rate sex biases observed for AIDs and cancers that occur in the same tissue are positively correlated across human tissues. In addition, we analyzed gene expression data from non-diseased tissue samples to determine if sex biases in gene set expression in these tissues are correlated with AID and cancer incidence rate sex biases in the same tissues. We find that the top positively enriched gene set across human tissues whose expression sex bias is most strongly associated with the incidence rate sex biases for AIDs, cancers, and AIDs and cancers considered jointly, is the set of 37 genes encoded in the mitochondrial genome.

Results

We surveyed 167 published AID studies and the cancer registries for 29 countries to assemble 182 country-specific, tissue-matched cancer-AID incidence rate sex bias data pairs (**Methods**; **Table S2**). For each study, we calculated the *incidence rate sex bias* (*IRSB*) as $IRSB = \log_2(IR_{MALE}/IR_{FEMALE})$, where IR_{MALE} and IR_{FEMALE} are the male and female incidence rates, so that a value of zero indicates no bias, a positive value indicates a higher incidence rate in

males (termed a “male bias”) and a negative value indicates a higher incidence rate in females (similarly termed a “female bias”). Having assembled these data, we computed the mean IRSBs to get a view of tissue-matched cancer and AID incidence rate sex bias across tissues, yielding global IRSB values for 17 AIDs and 17 cancer types across 12 human tissues, comprising a total of 24 cancer-AID data pairs. As expected, most AID incidence rates are female-biased (a negative sex-bias score), while most cancer incidence rates are male-biased (a positive sex-bias score) (**Figure 1A**). **Figure 1B** presents the correlation of the IRSB of these disorders across human tissues, summed up across all countries surveyed. Notably, we find an overall positive correlation (Pearson correlation r=0.48 with two-sided t-test p=0.017, Spearman correlation r=0.43 with two-sided t-test, p=0.034). Repeating this analysis using various levels of cancer type classification shows a consistent and robust correlation (**Figures S1-2, Table S4**). (We used Pearson's product-moment correlation coefficient to measure correlation because it takes effect size into account. We also provide correlation test results based on Spearman's rank correlation coefficient as this assesses correlation differently and may be of interest to the reader. We considered a correlation test result significant when the t-test adjusted $p \leq 0.05$. We used the Benjamini-Hochberg method to adjust p-values for multiple tests.) Second, studying this correlation in a country-specific manner for the four countries with at least 18 AID-cancer data pairs, we find a country-specific significant correlation for Sweden, while the correlations for Denmark, the UK and the USA have q-values (p-values corrected for multiple hypotheses testing) > 0.05 but are quite close to this threshold, showing a consistent trend for each country (**Figure 1C**).

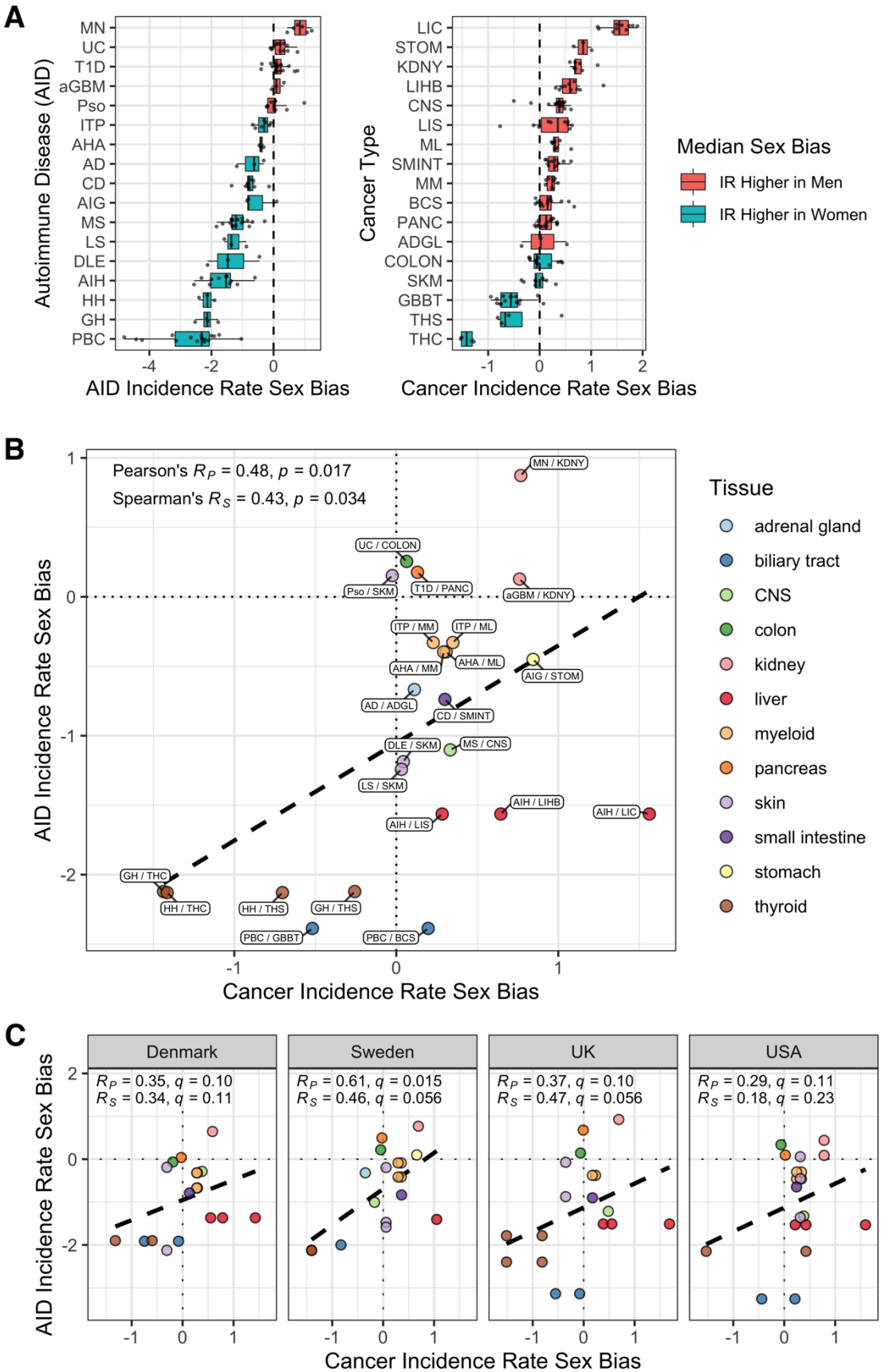
A
Autoimmune Disease (AID)
MN
UC
T1D
aGBM
Pso
ITP
AHA
AD
CD
AIG
MS
LS
DLE
AIH
HH
GH
PBC
-4
-2
0
AID Incidence Rate Sex Bias
Cancer Type
LIC
STOM
KDNY
LIHB
CNS
LIS
ML
SMINT
MM
BCS
PANC
ADGL
COLON
SKM
GBBT
THS
THC
-1
0
1
2
Cancer Incidence Rate Sex Bias
Median Sex Bias
IR Higher in Men
IR Higher in Women
B
Pearson's R_P = 0.48, p = 0.017
Spearman's R_S = 0.43, p = 0.034
AID Incidence Rate Sex Bias
Cancer Incidence Rate Sex Bias
MN / KDNY
UC / COLON
Pso / SKM
T1D / PANC
aGBM / KDNY
ITP / MM
ITP / ML
AHA / MM
AHA / ML
AIG / STOM
AD / ADGL
CD / SMINT
DLE / SKM
MS / CNS
LS / SKM
AIH / LIHB
AIH / LIC
AIH / LIS
GH / THC
HH / THC
HH / THS
GH / THS
PBC / GBBT
PBC / BCS
Tissue
adrenal gland
biliary tract
CNS
colon
kidney
liver
myeloid
pancreas
skin
small intestine
stomach
thyroid
C
Denmark
Sweden
UK
USA
R_P = 0.35, q = 0.10
R_S = 0.34, q = 0.11
R_P = 0.61, q = 0.015
R_S = 0.46, q = 0.056
R_P = 0.37, q = 0.10
R_S = 0.47, q = 0.056
R_P = 0.29, q = 0.11
R_S = 0.18, q = 0.23
AID Incidence Rate Sex Bias
Cancer Incidence Rate Sex Bias

Figure 1. Incidence rate sex biases for cancers and AIDs are positively correlated across tissues of origin. (A) *Distribution across countries of incidence rate sex bias (X-axis) for 17 AIDs and 17 cancer types (Y-axis). All data points are shown. Box shows interquartile range (IQR, first quartile to third quartile), with center bar representing the median (second quartile). Lefthand whisker extends from first quartile (Q1) to Q1-1.5*IQR or to the lowest value point, whichever is greater. Righthand whisker extends from third quartile (Q3) to Q3+1.5*IQR or to the highest value point, whichever is smaller. Positive median sex bias (red) indicates median with higher incidence rate in men; negative median sex bias (blue) indicates median with higher incidence rate in women.* ***(B, C)*** *Tissue-matched incidence rate sex biases for cancers (X-axis) and for autoimmune diseases (Y-axis) are displayed across different tissues of origin (circle color indicates the tissue). Positive values in each of the axes indicate male bias; negative values indicate female bias. The dashed line is the simple linear regression line. Statistics in the top left corner include the Pearson's product-moment correlation r-value (R_p) and t-test p-value; and the Spearman's rank correlation coefficient value (R_s) and a t-test p-value (t-tests were two-sided for the global-level tests and one-sided for the country-level tests). For country-level tests, p-values were corrected for multiple testing using the Benjamini-Hochberg method to produce q-values.* ***(B)*** *Across-population averages, with the cancer-AID pairs labeled.* ***(C)*** *Population-level data for the four countries with the largest numbers of data pairs (at least 18 out of 24 cancer-AID pairs), maintaining the tissue color labels used in the top panel (USA, 20 pairs; Denmark, Sweden, & UK, each 18 pairs).* ***AIDs****: AD, Addison's disease; aGBM, anti-glomerular basement membrane nephritis; AHA, Autoimmune hemolytic anemia; AIG, Autoimmune gastritis; AIH, Autoimmune hepatitis; CD, Celiac disease; DLE, Discoid lupus erythematosus; GH, Graves' hyperthyroidism; HH, Hashimoto's hypothyroidism; ITP, Immune thrombocytopenic purpura;*

LS, Localized scleroderma; MN, primary autoimmune membranous nephritis; MS, Multiple sclerosis; PBC, Primary biliary cholangitis; Pso, Psoriasis; T1D, Type 1 diabetes; UC, Ulcerative colitis. ***Cancers****: ADGL, adrenal gland cancer; BCS, liver (biliary) cholangiosarcoma; COLON, colon cancer; CNS, central nervous system cancer; GBBT, gallbladder & biliary tract cancer; KDNY, kidney cancer; LIC, liver carcinoma; LIHB, liver hepatoblastoma; LIS, liver sarcoma; ML, myeloid leukemia (acute and chronic); MM, multiple myeloma; PANC, pancreatic cancer; SKM, skin melanoma; SMINT, small intestine cancer; STOM, stomach cancer; THC, thyroid carcinoma; THS, thyroid sarcoma.*

Observing this fundamental correlation, we next asked if we could identify factors that might jointly modulate both the incidence rate sex bias observed in cancer and in AID across human tissues. We conducted both an unbiased general investigation and a hypothesis-driven one. We specifically examined four major factors that have been previously associated in the literature with the incidence rates of cancers and AIDs and/or their incidence rate sex biases. Those include (1) inflammatory or immune activity in the tissue [8, 9]; (2) expression of immune checkpoint genes [10, 11]; (3) the extent of X-chromosome inactivation [6, 12]; and finally, (4) mitochondrial activity [13, 14] and mitochondrial DNA copy number [15, 16].

Having these literature-driven specific hypotheses in mind, we still have chosen to begin by systematically charting the landscape of gene sets whose sex-biased enrichment in normal tissues is associated with IRSB in cancers and AIDs in an unbiased manner (see subsection 'Gene expression analysis of human tissues' in **Methods**). We analyzed gene expression data from non-diseased tissue samples from GTEx v8 [17], for tissues in which both cancer and AID arise;

GTEx data were available for 10 of the 12 tissues we studied above (**Table S3**). First, (1) for each gene in each tissue we calculated the *expression sex bias (ESB)* as $ESB = \log_2(GE_{MALE}/GE_{FEMALE})$, where GE_{MALE} or GE_{FEMALE} denote the average gene expression in TPM (transcripts-per-million) for male or female samples of the tissue. (2) Second, we computed the correlation of the expression sex bias of each gene with AID or cancer IRSBs (we abbreviate these correlations as $\text{corr}_{\text{ESB/IRSB}}$). We also computed aggregated or "joint" $corr_{ESB/IRSB}$ values as the average for each gene of its $corr_{ESB/IRSB}$ value for AID IRSBs and its $corr_{ESB/IRSB}$ value for cancer IRSBs. (3) Finally, for each of these three phenotypes, we ranked all the genes from top to bottom by the $\text{corr}_{\text{ESB/IRSB}}$ values and performed a gene set enrichment analysis (GSEA) [18, 19] to identify gene sets and pathways that were either significantly positively or negatively associated with IRSB. In total, this analysis covered 7763 gene sets, including gene ontology biological process sets and chromosome-location based sets from MSigDB [20], three X-chromosome gene sets (fully escape X-inactivation, variably escape X-inactivation, and pseudoautosomal region) [21], and finally, the two separate sets of nuclear-encoded genes whose protein products localize to the mitochondria and the 37 mitochondrial-genome-encoded genes [22].

Figure 2 shows the top positively and negatively $\text{corr}_{\text{ESB/IRSB}}$ enriched sets with $p \leq 10^{-3}$ after multiple hypotheses test correction for AID incidence (positive, Panel A; negative Panel B), for cancer incidence (positive, Panel C; negative Panel D), and their joint aggregate enrichment for both AID and cancer incidence (positive, Panel E, negative, Panel F). Strikingly, the top enriched gene set (highest normalized enrichment score (NES)) in *all* three phenotypes is the set of 37 genes encoded on the mitochondrial genome, including many genes with high $\text{corr}_{\text{ESB/IRSB}}$

values. In contrast, while the (much larger) set of all genes encoding proteins that localize to the mitochondria is significantly enriched for cancer IRSB, it is not significantly enriched for AID IRSB, where it is only ranked 3842 out of 6420 (negatively) enriched gene sets. Several immune-related gene sets also show high and significant $corr_{ESB/IRSB}$ positive enrichments in accordance with one of our initial hypotheses (**Figure 2**). However, the three different X chromosome gene sets studied in light of another one of our original hypotheses are not significantly enriched in $corr_{ESB/IRSB}$ values. Finally, several mRNA processing gene sets show strong negative significant correlations and high negative NES scores with AID and cancer incidence.

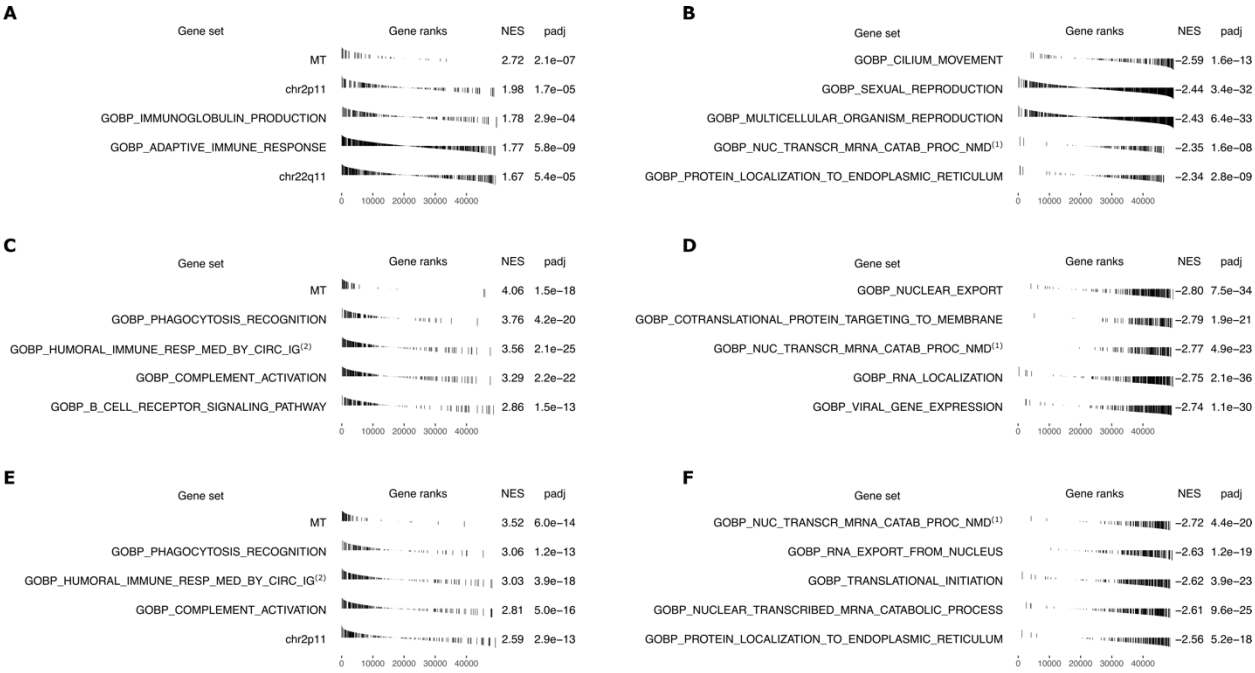

Figure 2. GSEA results for correlations of gene expression sex bias with IRSB. *Top 5 positively and negatively enriched gene sets, with adjusted $p \leq 10^{-3}$, for AID incidence (positive, **A**; negative, **B**), cancer incidence (positive, **C**; negative, **D**), and AIDs and cancers jointly (positive, **E**; negative, **F**). For each gene set the plot shows: (**Gene set**) name;(**Gene ranks**) bar*

plot of $corr_{ESB/IRSB}$ *values ordered from highest correlation at the left to lowest at the right (bars for genes in the gene set are black); (**NES**) normalized enrichment score; and (**padj**) Benjamini-Hochberg corrected p-value.* ***Abbreviated gene set names:*** *(1) Gene Ontology Biological Process (GOBP) Nuclear transcribed mRNA catabolic process nonsense-mediated decay; (2) GOBP Humoral immune response mediated by circulating immunoglobulin; (3) GOBP Nuclear transcribed mRNA catabolic process.*

To obtain a clearer visualization of the key positively enriched gene sets described above, we summarized the expression of the genes composing a given gene set in a normal GTEx tissue by computing their geometric mean, giving us a single activity summary value (see subsection 'Gene expression analysis of human tissues' in **Methods**). We then computed the correlation across tissues between these summary values of the gene sets in each normal tissue and the IRSBs of cancers or AIDs (**Figure 3**). In concordance with the results of the unbiased analysis presented above, we do not observe a significant correlation between cancer or AID incidence rate sex bias and the expression of key immune checkpoint genes (CTLA-4, PD-1, or PD-L1, **Figure S3**), or the extent of X-chromosome inactivation (quantified by the expression of XIST lncRNA [23], **Figure S4**). We also do not find such significant consistent correlations for the top immune gene sets found via the unbiased analysis (previously shown in **Figure 2**). However, we do find strong correlations between these summary values for the mitochondrial gene set, which was ranked highest in **Figure 2** (gene set "MT"): Remarkably, we find that the sex bias of mtRNA expression in GTEx tissues is positively correlated both with AID incidence rate sex bias (Pearson r=0.56, one-sided t-test p=0.018) and with cancer incidence rate sex bias (Pearson r=0.67, one-sided t-test p=0.0058) (**Figures 3A** and **3B**; the correlations between mtRNA

expression and cancer and AID incidence rates for each of the sexes individually are provided in **Figure S5**). The significance of these two associations is further supported by observing that the basic correlation between cancer and AID IRSBs becomes insignificant when we compute the partial correlation between these two variables while controlling for the mtRNA expression bias (Pearson r=0.21, two-sided t-test p=0.42). Overall, these findings are in line with previous reports linking mitochondrial activity [13, 14] and mtDNA copy number [15, 16] with higher AID and cancer risk.

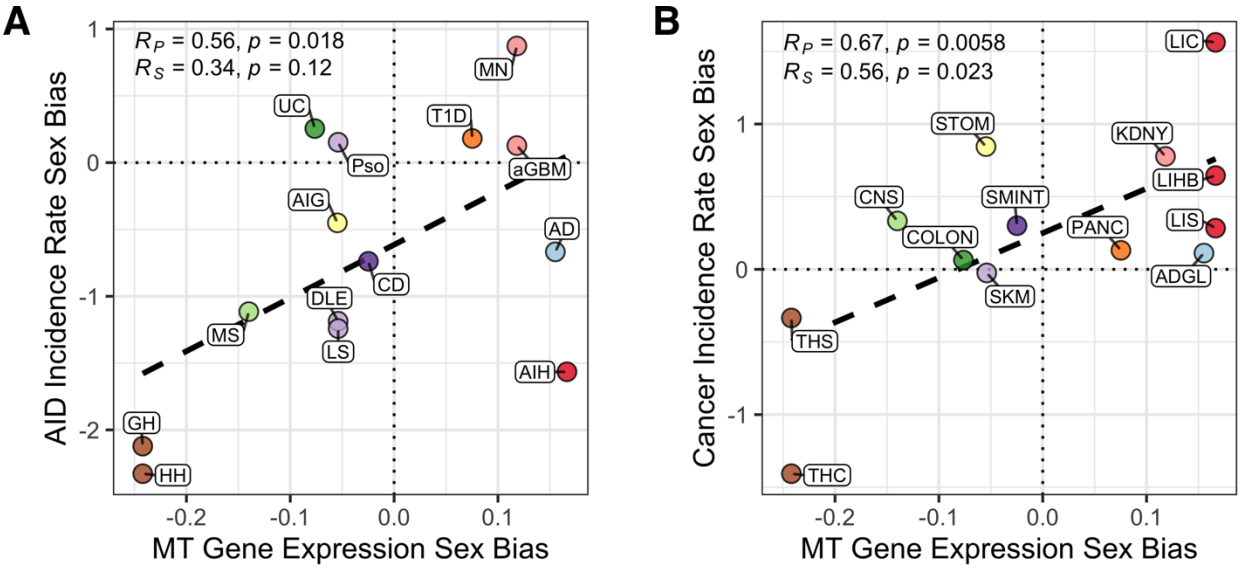

Figure 3. Mitochondrial gene expression is a strong correlate of sex biases in incidence rate of autoimmune diseases and cancer types across tissues. *The correlation between expression ratio of mitochondrial gene expression in male vs female tissues (X-axis) with the incidence rate sex biases of* ***(A)*** *autoimmune diseases (Y-axis) and* ***(B)*** *cancer types (Y-axis) across human tissues (circle color indicates the tissue).* ***AIDs****: AD, Addison's disease; aGBM, anti-glomerular basement membrane nephritis; AIG, Autoimmune gastritis; AIH, Autoimmune hepatitis; CD, Celiac disease; DLE, Discoid lupus erythematosus; GH, Graves' hyperthyroidism; HH,*

Hashimoto's hypothyroidism; LS, Localized scleroderma; MN, primary autoimmune membranous nephritis; MS, Multiple sclerosis; Pso, Psoriasis; T1D, Type 1 diabetes; UC, Ulcerative colitis. ***Cancers****: ADGL, adrenal gland cancer; CNS, central nervous system cancer; COLON, colon cancer; KDNY, kidney cancer; LIC, liver carcinoma; LIHB, liver hepatoblastoma; LIS, liver sarcoma; PANC, pancreatic cancer; SMINT, small intestine cancer; SKM, skin melanoma; STOM, stomach cancer; THC, thyroid carcinoma; THS, thyroid sarcoma.*

Discussion

The correlative findings between the expression of mitochondrially encoded genes and cancer and AID IRSBs across human tissues are quite surprising, giving rise to two further fundamental questions. First, what biological mechanisms may be associated with sex differences in overall mitochondrial functioning? One potential candidate may be estrogen signaling, which has been shown to regulate at least four mitochondrial functions relevant to health and disease [24], including, (1) biogenesis of mitochondria, whose levels differ across sexes and tissues [25], (2) T-cell metabolism (including mitochondrial activity measured by Seahorse assays) and T-cell survival (estimated by retention of inner membrane potential) [26], (3) unfolded protein response [27] (mediated partly via mitochondrial superoxide dismutase) [28], and (4) generation of reactive oxygen species (ROS) [29]. Second, how might sex differences in mitochondria functioning modulate the sex-biased incidence observed in cancers and AIDs? One possible mechanism is through differences in ROS production, which notably involves quite a few mitochondrially encoded genes: Increased mitochondrial ROS generation has been associated with both the initiation and intensification of autoimmunity in several organ-specific AIDs [13] and with cancer initiation and progression [14]. More generally, alterations in mtDNA copy

number have been associated with increased risk of lymphoma and breast cancer, [15] and somatic mtDNA mutations producing mutated peptides may trigger autoimmunity [16].

Our analyses have a few limitations and we list three main ones. First, the majority of our AID-cancer data pairs are from European countries (113 of 182 [62%]), which might introduce geographic, ethnic, or social biases. Second, factors beyond biological drivers, such as sex differences in the propensity to seek medical care or reporting of specific diseases, are not characterized in the datasets studied. However, putative disease-specific effects may be somewhat mitigated given the opposite tendency of sex biases for AIDs and cancers in a study of tissue-specific correlations like ours. Third, although much of the incidence rate data is age-standardized, we could not take additional steps to account for age-related incidence rate differences as the sample sizes available are too small to enable doing such an analysis in a robust manner.

As in humans, sex differences have been reported in animal studies of diseases, which has prompted us to search the literature and survey previous studies of sex bias in disease incidence in rodent models of cancers and AIDs. We focused on studies of sex difference in spontaneous and/or autochthonous carcinogenesis by either carcinogen treatment or genetic engineering, excluding transplantation of syngeneic animals because these animals do not model disease development (representative examples are listed in **Tables S5-6** for cancer and AIDs respectively.) **Table S5** lists our cancer incidence findings, where the sex bias was male skewed in colon, liver, kidney, pancreas, and stomach, and higher in females in the thyroid, consistent with the human reports. Interestingly, for colon, liver, kidney, pancreas, and thyroid, the sex bias

disappeared or was reduced when the animals were subjected to castration/ovariectomy or hormone treatment, supporting the notion that the differences in these organs are likely to be driven by sex hormones. **Table S6** lists AID rodent models that allow for direct comparisons to the human data. The AID sex bias reported is however generally higher in males than in females, in difference from the human findings, but the higher male bias observed in kidney, colon, pancreas, and skin compared to the thyroid is maintained.

In summary, we find a surprising overall positive correlation between cancer and AID incidence rate sex biases across many different human tissues. Among key factors that have been previously associated with sex bias in either AID or cancer incidence, we find that the sex bias in the expression of mitochondrially encoded genes (and possibly in the expression of a few immune pathways) stands out as a key factor whose aggregate level across human tissues is quite strongly associated with these incidence rate sex biases. Our findings thus call for further mechanistic studies on the role of mitochondrial gene expression in determining cancer and AID incidence and their incidence rate sex biases.

Methods

Overview

Our analysis is divided into two main parts: curation and analysis of disease incidence rate data; and investigation of associations between incidence rates and gene expression in corresponding non-diseased human tissue samples. First we studied the association of incidence rates for AIDs and cancers occurring in the same tissue. We collected incidence data for AIDs from published studies, and for each country for which we found incidence data for a given AID, we collected

incidence data from that country's national cancer registry for cancers occurring in the same tissue as the AID. We matched AID and cancer incidence data by tissue and by country to produce country-specific tissue-matched AID-cancer incidence rate data pairs. We then computed across-tissue correlations between AIDs and cancers for male incidence rates, female incidence rates, overall incidence rates, and incidence rate sex biases, at both the individual country level and the across-country global level.

Next we used non-diseased human tissue transcriptomic data from GTEx [17] to investigate possible factors across human tissues that might be associated with incidence rate sex biases. We computed correlations between incidence rate sex biases and either expression of individual genes or enrichment of human functional gene sets across tissues.

Autoimmune disease incidence data curation

We first performed an extensive literature search for sex-specific incidence data for AIDs. For each AID, we searched for original studies mentioning the disease and epidemiology, prevalence, incidence, incidence rate, or sex bias using Google Scholar. We considered only population-based studies that use clinical inclusion criteria and have at least 25 cases for a given disease. We evaluated whether or not a study was population-based using either (a) the characteristics of the existing data source used in the study (e.g., a mandatory country-wide reporting registry) or (b) estimates showing that the data collected in the study were likely representative of the overall population. We evaluated whether or not a study used clinical diagnostic criteria by looking for use of a disease-specific blood test, a histological assay, or other evidence used to confirm diagnosis and rule out similar non-autoimmune conditions. Additionally, we considered only AIDs

with a focal primary tissue (e.g., we included ulcerative colitis but excluded Crohn's disease), for which we could find incidence data for at least three countries. We excluded sex-specific tissues.

We collected 188 AID-country incidence rate datasets from 167 studies. For each dataset, we calculated the *incidence rate sex bias* (*IRSB*) as

$$IRSB = \log_2(IR_{MALE}/IR_{FEMALE}) \qquad (A)$$

so that a value of zero indicates no bias, a positive value indicates a higher incidence rate in males (termed a "male bias") and a negative value indicates a higher incidence rate in females (similarly termed a "female bias"). A majority of the studies provided sex-specific (123 of 188 datasets, 65%) and total (143 of 188, 76%) incidence rates (IR):

$$IR_{POP} = cases_{POP}/population_{POP} \qquad (B)$$

(where: "POP" stands for either the "MALE", "FEMALE", or "TOTAL" population; $cases_{TOTAL} = cases_{MALE} + cases_{FEMALE}$; and $population_{TOTAL} = population_{MALE} + population_{FEMALE}$). Most studies reported IR as cases per year per 10^5 persons; those using a different scale were converted to this scale. We used "crude" incidence rates (as defined above) when available; some studies provided only age-adjusted incidence rates.

Estimating incidence rates

For each of the four incidence rate measures we consider (IRSB, IR_F, IR_M, and R_{TOTAL}) the majority of studies provided a value, while other studies gave values for other measures (i.e. different incidence rates or case counts) that can be used to estimate the value of that measure. For a given measure we can divide our AID-country datasets into four groups (**Table 1**): (1) those with the measure's value but not the values of other measures we can use to estimate that

value; (2) those with the measure's value and the values of other measures we can use to estimate that value; (3) those without the measure's value but with the values of other measures we can use to estimate that value; and (4) those with neither the measure's value nor the values of measures we can use to estimate that value. For each measure, we assessed the accuracy of our estimator by comparing the actual and estimated values for datasets in group (2), and then used that same estimator to estimate values for datasets in group (3).

Table 1. Measures to estimate, measures needed for estimators, and numbers of datasets with values for these measures. *Numbers indicate dataset count and percentage (out of 188 total datasets) for each group of datasets (1-4) described in the text.*

(a) Measure to estimate	(b) Measures needed for estimator	(1) Datasets with (a)	(2) Datasets with (a) & (b)	(3) Datasets with (b) but not (a)	(4) Datasets with neither (a) nor (a)
IRSB	$cases_M/cases_F$	125 (66%)	105 (56%)	63 (34%)	0 (0%)
IR_M, IR_F	IR_{TOTAL}, $cases_M/cases_F$	123 (65%)	84 (45%)	41 (22%)	24 (13%)
IR_{TOTAL}	IR_M, IR_F	143 (76%)	101 (54%)	22 (12%)	23 (12%)

The estimators for all four measures require a value for the population's sex ratio

$$Sex_Ratio = \frac{population_{FEMALE}}{population_{MALE}} \qquad (C)$$

As only one study provided the background population sex ratio, we estimated measures using either a sex ratio of 1:1 or the sex ratio for the corresponding population (matching the specific

country during the years the study was conducted) according to United Nations estimates [30]. Based on (A), (B), and (C), we estimated IRSB, IR_F, IR_M, and R_{TOTAL} as:

$$IRSB = \log_2\left(\frac{cases_{MALE}}{cases_{FEMALE}} \times Sex_Ratio\right)$$

$$IR_{FEMALE} = IR_{TOTAL} \times \frac{cases_{FEMALE}}{cases_{TOTAL}} \times (1 + 1/Sex_Ratio)$$

$$IR_{MALE} = IR_{TOTAL} \times \frac{cases_{MALE}}{cases_{TOTAL}} \times (1 + Sex_Ratio)$$

$$IR_{TOTAL} = IR_M \times \left(\frac{1}{1 + Sex_Ratio}\right) + IR_F \times \left(\frac{Sex_Ratio}{1 + Sex_Ratio}\right)$$

To assess the accuracy of our estimators we compared the actual and estimated values for datasets in group (2) in two ways (**Table 2**). First, we computed the Pearson's correlation coefficient *r* between the two values. All estimators were accurate: for each the correlation coefficient was close to 1 and the one-sided t-test was significant. Second, we computed a simple linear model of the form $x_{actual} = \beta \times x_{estimate} + \alpha$. All estimators were accurate: for each the coefficient β was close to 1 and the r^2 close to 1 (where *r* is the Pearson's correlation coefficient). For all four measures the estimators performed well, but for each measure the estimator using a sex ratio of 1:1 performed as good as or slightly better than the estimator using the sex ratio based on the United Nations estimates. Accordingly, for our analyses we used estimators with a sex ratio of 1:1. For all of our analyses, results computed using only given values, and not estimates, were consistent with results computed using both given and estimated values (the code for this paper includes scripts to reproduce all tests and figures using data that either includes or excludes estimated values.)

Table 2. Pearson's correlation coefficient and simple linear model results for estimators. *For each measure, each of two estimators (with sex ratio as 1:1 or from the United Nations estimates) is shown with its simple linear model coefficient β, intercept α, and r^2, and its Pearson's correlation coefficient r and one-sided t-test p-value.*

Measure	Estimator	β	α	r^2	r	p
$IRSB$	$IRSB_{1:1}$	0.922	-0.0254	0.966	0.983	6.04E-78
	$IRSB_{UN}$	0.923	-0.0585	0.963	0.981	5.54E-76
IR_{FEMALE}	$IR_{FEMALE,1:1}$	1.020	-0.303	0.997	0.998	1.15E-107
	$IR_{FEMALE,UN}$	1.035	-0.307	0.997	0.999	5.04E-105
IR_{MALE}	$IR_{MALE,1:1}$	0.975	0.228	0.997	0.999	6.10E-108
	$IR_{MALE,UN}$	0.959	0.236	0.997	0.999	2.26E-105
IR_{TOTAL}	$IR_{TOTAL;1:1}$	1.013	-0.123	1.000	1.000	1.09E-182
	$IR_{TOTAL;UN}$	1.013	-0.136	1.000	1.000	1.39E-182

Combining data for each country

When multiple studies were available for an AID in a country, we used the across-study arithmetic mean of each incidence rate measure as the measure value for that AID-country pair (for IR_{MALE}, IR_{FEMALE}, IR_{TOTAL}, or $IRSB$ measures). Overall, surveying 167 published studies (**Supplementary References**), we calculated 133 country-specific AID incidence rate sex bias

data points for 17 AIDs in 33 countries (**Table S1**, e.g., the mean incidence rate sex bias for Type 1 diabetes in Spain is one such data point).

Cancer incidence data curation

Cancer incidence rates were calculated from GLOBOCAN [31] data for all but three countries. For each country for which we had AID data, we computed each cancer type's incidence rate measure for each year and then averaged the yearly measure values to produce a single measure value for each country-cancer pair (for IR_{MALE}, IR_{FEMALE}, IR_{TOTAL}, or $IRSB$ measures). Cancer data for Finland [32], Sweden [33], and Taiwan [34] were collected from country-specific databases. For Finland and Sweden we calculated each incidence rate measure as the across-year average yearly measure for each cancer type for the most recent 20 years (1999-2019) for each country. For Taiwan we calculated each measure as the average of the measure for the two available time periods (1998-2002, 2003-2007). Overall, we calculated 165 country-specific cancer incidence rate sex bias data points for 17 cancer subtypes in 29 countries (**Table S2**; for an additional four countries we were unable to find population-level cancer incidence data).

Pairing AID and cancer incidence data

Across 12 human tissues we paired 17 AIDs with 17 cancer types for a total of 24 cancer-AID data pairs. To compute the correlation between AIDs and cancer incidence rate sex biases across tissues, we grouped AIDs with matched cancers occurring in the same tissue in the same country (**Table S3**). For example, for the UK, we paired thyroid AID data points for Hashimoto's hypothyroidism and Graves' hyperthyroidism with cancer data points for thyroid carcinoma and thyroid sarcoma, resulting in 4 possible thyroid cancer-AID pairs. The 133 country-specific AID incidence rate sex bias data points were matched to the 165 country-specific cancer incidence

rate sex bias data points, yielding a total of 182 country-specific, tissue-matched cancer-AID incidence rate sex bias data pairs that are jointly present in both the AID and cancer datasets (**Table S2**).

Gene expression analysis of human tissues

Gene expression was calculated from GTEx v8 data [17] provided in transcripts-per-million (TPM). For gene *i* and tissue *k* with *m* samples we calculated the within-tissue gene expression (GE) as the arithmetic mean TPM across samples as (where "POP" stands for either the "MALE", "FEMALE", or "TOTAL" population):

$$GE_{POP,i,k} = \frac{1}{m}\sum_{j=1}^{m} TPM_{POP,i,j,k}$$

and the gene expression sex bias (ESB) as:

$$ESB_{i,k} = \log_2\left(GE_{MALE,i,k}/GE_{FEMALE,i,k}\right)$$

where both $GE_{MALE,i,k}$ and $GE_{FEMALE,i,k}$ are positive.

For a set of *n* genes *N* and tissue *k* with *m* samples, we calculated the within-tissue gene set activity (GSA) as the geometric mean of gene expression across genes:

$$GSA_{POP,N,k} = \left(\prod_{i=1}^{n} GE_{POP,i,k}\right)^{1/n}$$

and the gene set activity sex bias (ASB) as:

$$ASB_{N,k} = \log_2\left(GSA_{MALE,N,k}/GSA_{FEMALE,N,k}\right)$$

where both $GSA_{MALE,N,k}$ and $GSA_{FEMALE,N,k}$ are positive.

Gene set enrichment analysis across human functional pathways

We performed gene set enrichment analysis (GSEA) in three steps. First, for each gene in each tissue we calculated the gene expression sex bias (ESB). Second, we computed the across-tissue

Spearman correlation of the ESB of each gene with AID or cancer IRSBs (we abbreviate these correlations as $corr_{ESB/IRSB}$). We also computed aggregated or "joint" $corr_{ESB/IRSB}$ values as the average for each gene of its $corr_{ESB/IRSB}$ value for AID IRSBs and its $corr_{ESB/IRSB}$ value for cancer IRSBs. Finally, for each of these three phenotypes, we ordered all the genes from greatest to least by the $corr_{ESB/IRSB}$ values and performed a GSEA [19] to identify gene sets and pathways that were either significantly positively or negatively associated with IRSB (for gene sets used see Results). We considered GSEA results significant if the adjusted $p \leq 10^{-3}$ (we used the Benjamini-Hochberg method to adjust p-values for multiple tests) and ranked the results by normalized enrichment score (NES) [19].

Code and data availability

Code and data used for analysis are available on Zenodo at [link]. Statistical analyses and figure preparation were performed on a Macintosh computer (OS 12.5.1; 32GB memory; 8-core 2.3GHz processor) in *RStudio* (v2021.09.0+351 "Ghost Orchid" release) [35] running the *R* language (v4.1.2) [36]. Processed data files are provided with the scripts. All data used was publicly available. Plots produced in *R* were aligned and lettered using *Inkscape* (v1.0) (inkscape.org) to produce multi-plot figures.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, NCI. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Competing Interests

The authors declare no competing interests.

Author contributions

DRC and ER conceived of and designed project. DRC collected published human study data, performed all data analyses. DRC, AAS and ER wrote the manuscript. CPD collected and interpreted published rodent study data and wrote the relevant discussion. SS reviewed the data analysis code. All authors (including NUN, BMR, JSBS, SMM, AE, KA, PEC, PSR, AAS) participated in discussions of the project and contributed valuable ideas and feedback and helped in refining the manuscript writeup.

Data & code availability

All data & code used for analysis and preparation of figures is available on Zenodo for reviewers.

References

[1] S. T. Ngo, F. J. Steyn and P. A. McCombe, "Gender differences in autoimmune disease," *Frontiers in Neuroendocrinology,* vol. 35, pp. 347-369, 2014.

[2] L. Moroni, I. Bianchi and A. Lleo, "Geoepidemiology, gender and autoimmune disease," *Autoimmunity Reviews,* vol. 11, pp. A386-392, 2012.

[3] A. Clocchiatti, E. Cora, Y. Zhang and G. P. Dotto, "Sexual dimorphism in cancer," *Nature Reviews Cancer,* vol. 16, pp. 330-339, 2016.

[4] A. R. Costa, M. L. de Oliveira, Cruz, I., C. R. Santos, J. F. Cascalheira and C. R. A. Santos, "The sex bias of cancer," *Trends in Endocrinology & Metabolism,* vol. 31, pp. 785-799, 2020.

[5] S. Haupt, F. Caramia, S. L. Klein, J. B. Rubin and Y. Haupt, "Sex disparities matter in cancer development and therapy," *Nature Reviews Cancer,* vol. 21, pp. 393-407, 2021.

[6] S. C. Credendino, C. Neumayer and I. Cantone, "Genetics and Epigenetics of Sex Bias: Insights from Human Cancer and Autoimmunity," *Trends in Genetics,* vol. 36, pp. 650-663, 2020.

[7] G. Edgren, L. Liang, H.-O. Adami and E. T. Chang, "Enigmatic sex disparities in cancer incidence," *European Journal of Epidemiology,* vol. 27, pp. 187-196, 2012.

[8] S. I. Grivennikov, F. R. Greten and M. Karin, "Immunity, inflammation, and cancer," *Cell,* vol. 140, pp. 883-899, 2010.

[9] S. L. Klein and K. L. Flanagan, "Sex differences in immune responses," *Nature Reviews Immunology,* vol. 16, pp. 626-638, 2016.

[10] C. Huang, H.-X. Zhu, Y. Yao, Z.-H. Bian, Y. Zheng, L. Li, H. M. Moutsopoulos, M. E. Gershwin and Z.-X. Lian, "Immune checkpoint molecules. Possible future therapeutic implications in autoimmune diseases.," *Journal of Autoimmunity,* vol. 104, p. 102333, 2019.

[11] M. Wagner, M. Jasek and L. Karabon, "Immune checkpoint molecules--inherited variations as markers for cancer risk," *Frontiers in Immunology,* vol. 11, p. 606721, 2021.

[12] A. Dunford, D. M. Weinstock, V. Savova, S. E. Schumacher, J. P. Cleary, A. Yoda, T. J. Sullivan, J. M. Hess, A. A. Gimelbrant, R. Beroukhim, M. S. Lawrence, G. Getz and A. A. Lane, "Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias," *Nature Genetics,* vol. 49, pp. 10-16, 2017.

[13] G. Di Dalmazi, J. Hirschberg, D. Lyle, J. B. Freij and P. Caturegli, "Reactive oxygen species in organ-specific autoimmunity," *Autoimmunity Highlights,* vol. 7, p. 11, 2016.

[14] S. S. Sabharwal and P. T. Schumacker, "Mitochondrial ROS in cancer: initiators, amplifiers, or Achilles' heel?," *Nature Reviews Cancer,* vol. 14, pp. 709-721, 2014.

[15] L. Hu, X. Yao and Y. Shen, "Altered mitochondrial DNA copy number contributes to human cancer risk: evidence from an updated meta-analysis," *Scientific Reports,* vol. 6, p. 35859, 2016.

[16] L. Chen, B. Duvvuri, J. Grigull, R. Jamnik, J. E. Wither and G. E. Wu, "Experimental evidence that mutated self-peptides derived from mitochondrial DNA somatic mutations have the potential to trigger autoimmunity," *Human Immunology,* vol. 75, pp. 873-879, 2014.

[17] GTEx Consortium, "The GTEx Consortium atlas of genetic regulatory effects across human tissues," *Science,* vol. 369, pp. 1318-1330, 2020.

[18] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *PNAS,* vol. 102, pp. 15545-15550, 2005.

[19] G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov and A. Sergushichev, "Fast gene set enrichment analysis," *bioRxiv,* p. https://doi.org/10.1101/060012, 01 02 2021.

[20] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics,* vol. 27, pp. 1739-1740, 2011.

[21] T. Tukiainen, A.-C. Villani, A. Yen, M. A. Rivas, J. L. Marshall, R. Satija, M. Aguirre, L. Gauthier, M. Fleharty, A. Kirby, B. B. Cummings, S. E. Castel, K. J. Karczewski, F. Aguet, A. Byrnes, GTEx_Consortium, T. Lappalainen, A. Regev and A, "Landscape of X chromosome inactivation across human tissues," *Nature,* vol. 550, pp. 244-248, 2017.

[22] P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. A. Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson and S. Lee, "A subcellular map of the human proteome," *Science,* vol. 356, p. 820, 2017.

[23] D. B. Pontier and J. Gribnau, "Xist regulation and function eXplored," *Human Genetics,* vol. 130, pp. 223-236, 2011.

[24] D. N. Di Florio, J. Sin, M. J. Coronado, P. S. Atwal and D. Fairweather, "Sex differences in inflammation, redox biology, mitochondria and autoimmunity," *Redox Biology,* vol. 31, p. 101482, 2020.

[25] R. Ventura-Clapier, M. Moulin, J. Piquereau, C. Lemaire, M. Mericskay, V. Veksler and A. Garnier, "Mitochondria: a central target for sex differences in pathologies," *Clinical Science,* vol. 131, pp. 803-822, 2017.

[26] I. Mohammad, I. Starskaia, T. Nagy, J. Guo, E. Yatkin, K. Väänänen, W. T. Watford and Z. Chen, "Estrogen receptor alpha contributes to T cell-mediated autoimmune inflammation by promoting T cell activation and proliferation," *Science Signaling,* vol. 11, p. eaap9415, 2018.

[27] L. Papa and D. Germain, "Estrogen receptor mediates a distinct mitochondrial unfolded protein response," *Journal of Cell Science,* vol. 124, pp. 1396-1402, 2011.

[28] T. C. Kenny, A. J. Craig, A. Villaneuva and D. Germain, "Mitohormesis Primes Tumor Invasion and Metastasis," *Cell Reports,* vol. 27, pp. 2292-2303, 2019.

[29] T.-L. Liao, Y.-C. Lee, C.-R. Tzeng, Y.-P. Wang, H.-Y. Chang, Y.-F. Lin and S.-H. Kao, "Mitochondrial translocation of estrogen receptor beta affords resistance to oxidative insult-induced apoptosis and contributes to the pathogenesis of endometriosis," *Free Radical Biology and Medicine,* vol. 134, pp. 359-373, 2019.

[30] United Nations, Department of Economic and Social Affairs, Population Division, "World Population Prospects," 2022. [Online]. Available: https://population.un.org/wpp/Download/Standard/Population/. [Accessed 23 June 2022].

[31] F. Bray, M. Colombet, L. Mery, M. Piñeros, A. Znaor, R. Zanetti and J. Ferlay, "Cancer Incidence in Five Continents," Lyon, IARC, 2017. [Online]. Available: http://ci5.iarc.fr. [Accessed 18 07 2021].

[32] Finnish Cancer Registry, 27 04 2021. [Online]. Available: https://cancerregistry.fi/statistics/cancer-statistics/. [Accessed 21 07 2021].

[33] The Swedish Cancer Register, 06 12 2020. [Online]. Available: https://www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/swedish-cancer-register/. [Accessed 21 07 2021].

[34] Taiwan Cancer Registry, 04 05 2012. [Online]. Available: http://tcr.cph.ntu.edu.tw/main.php?Page=N2. [Accessed 21 07 2021].

[35] RStudio Team, "RStudio: Integrated Development for R.," RStudio, Inc., Boston, MA, 2018.

[36] R Core Team, "R: A language and environment for statistical computing.," R Foundation for Statistical Computing, Vienna, Austria, 2021.