



Quantitative estimation of intracellular oxidative stress in human tissues

Jun Bai [†], Renbo Tan[†], Zheng An  and Ying Xu

Corresponding author: Ying Xu, Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA. Tel.: 0431-89876558;

Fax: 0431-89876558; E-mail: xyn@uga.edu

[†]Jun Bai and Renbo Tan contributed equally to this work.

Abstract

Oxidative stress is known to be involved in and possibly a key driver of the development of numerous chronic diseases, including cancer. It is highly desired to have a capability to reliably estimate the level of intracellular oxidative stress as it can help to identify functional changes and disease phenotypes associated with such a stress, but the problem proves to be very challenging. We present a novel computational model for quantitatively estimating the level of oxidative stress in tissues and cells based on their transcriptomic data. The model consists of (i) three sets of marker genes found to be associated with the production of oxidizing molecules, the activated antioxidation programs and the intracellular stress attributed to oxidation, respectively; (ii) three polynomial functions defined over the expression levels of the three gene sets are developed aimed to capture the total oxidizing power, the activated antioxidation capacity and the oxidative stress level, respectively, with their detailed parameters estimated by solving an optimization problem and (iii) the optimization problem is so formulated to capture the relevant known insights such as the oxidative stress level generally goes up from normal to chronic diseases and then to cancer tissues. Systematic assessments on independent datasets indicate that the trained predictor is highly reliable and numerous insights are made based on its application results to samples in the TCGA, GTEx and GEO databases.

Keywords: oxidative stress, computational prediction, chronic disease, cancer, disease driver

Introduction

Oxidative stress is an essential component and possibly a driving force of many chronic diseases, including cancer, Alzheimer disease and diabetes [1]. In cancer, oxidative stress is believed to be involved in multiple phases of the disease development, such as cancerous transformation from normal cells, angiogenesis and metastasis [2]. Generally, the development of any chronic disease may involve multiple types of stressors like oxidative stress, hypoxia and pH imbalance. Each of these stressors may contribute to the development of a disease in distinct ways. Hence, sorting out the stress types and their levels present in a disease tissue is highly desired; and having such a capability could prove to be essential to the full elucidation of key drivers of many diseases, including cancer.

The level of oxidative stress refers to the gap between the total oxidizing power and the antioxidation capacity available in a cell. Under stressful conditions of any kind, oxidizing molecules such as reactive oxygen [3, 4] or/and nitrogen [5, 6] species will be produced intracellularly and/or by local immune cells. Throughout evolution,

most, if not all cells have developed capacities to cope with the excessive oxidizing molecules by designated processes or via moonlighting by certain molecular species to protect the essential cellular components from being oxidatively damaged. For example, the glutathione (GSH) system is the designated antioxidation capability in human cells. Under severe stressful conditions, the oxidizing molecules produced may overpower the designated reducing capacity, leading to oxidative stress. Our goal here is to develop an algorithm and software for predicting the level of oxidative stress in a human tissue or cell based on its gene-expression data.

Several researchers have attempted to tackle this prediction problem, but the issue has proven to be very challenging [7, 8]. Published studies generally focus on the prediction of specific biomarkers for oxidative stress instead of the oxidative stress level, such as carbonylated proteins [9], oxidized low-density lipoproteins, oxidized products of lipids like 4-hydroxynonenal and malondialdehyde [10, 11] and protein thiols [12]. These biomarkers have two general limitations: (i) they reflect the oxidation level due to specific molecules; and (ii) more

Jun Bai is a researcher in the Cancer Systems Biology Center, China-Japan Union Hospital of Jilin University and is a postgraduate student in the School of Artificial Intelligence, Jilin University.

Renbo Tan is a researcher in the Cancer Systems Biology Center, China-Japan Union Hospital of Jilin University and College of Computer Science and Technology, Jilin University.

Zheng An is a researcher and PhD student in Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, the University of Georgia.

Ying Xu is an endowed professor of both the University of Georgia and Jilin University.

Received: March 3, 2022. **Revised:** April 28, 2022. **Accepted:** May 4, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

importantly, they are not suited for large-scale data analyses to study cellular changes associated with oxidative stress [13].

One key reason for the challenging nature is that some functional but non-essential molecules could be used to neutralize the persistently produced oxidizing molecules, such as lipids, enzymes and ribonucleic acid (RNA) when the designated antioxidation capacity reaches its limit. The oxidized levels of such molecules are rather difficult to estimate based on omic data. Another challenge is that no omic data with matching experimentally measured oxidative stress data are publicly available, to the best of our knowledge, making computational model development and validation difficult. Those are probably the reasons that there are no publicly available computer servers for making such predictions.

To overcome these challenges, we have taken a broad-brush systems-level approach. We consider three classes of marker genes and their expressions to estimate the levels of the total oxidizing power (O), the activated antioxidation capacity (R) and the intracellular stress level attributed to oxidation (OS), respectively. Under the assumption of $OS \approx O - R$, we have estimated the parameters for integrating the expressions of the selected marker genes to determine each of the three quantities through solving an optimization problem. To reliably assess the quantity of O , we have collected as many proteins as possible, whose functions are known to produce oxidizing molecules. To estimate R , we have considered multiple molecular species known to be oxidized in cancer tissues, such as lipids, enzymes and RNA, in addition to the GSH level. For OS , we have assessed several pathways whose levels reflect the oxidative stress level. A non-linear model is developed, which is most consistent with the expressions of the marker genes in normal tissue samples, non-cancerous disease and cancer tissues as well as our knowledge about the general stress levels across these tissue types.

Systematic analyses and validation of the trained predictor on transcriptomic data of normal, human disease tissues and cancer tissues of different types indicate that the predictor represents a reliable tool for accurate prediction of the oxidative stress level in human tissues, diseased or normal, which should be useful for studies of a wide range of biological problems associated with oxidative stress.

Results

Model construction

We use the following to model the (intracellular) level of oxidative stress:

$$OS = O - R, \quad (1)$$

where O is the average production rate of all the major oxidizing molecules in a tissue, R is the activated antioxidation capacity in the same tissue and OS denotes the oxidative stress level. Our first goal is to identify three sets of marker genes: MG-O, MG-R and MG-S whose

integrated expression levels well reflect the above three quantities, O , R and OS , respectively, so that either side of Equation (1) can be regarded as an estimate of the overall oxidative stress level.

We have examined and collected all genes whose proteins are known to produce oxidizing molecules, including immune cell genes whose proteins produce superoxide [14] and H_2O_2 that can diffuse into cells via chloride channels and aquaporins [15]; electron transport chain complexes I and III [16] that can diffuse into cytosol via VDACs [17]; intracellular NADPH oxidases [2]; nitric oxide synthases, which catalyze nitric oxide-generating reactions [18]; genes relevant to lipid peroxidation, a process that produces reactive intermediates when electrons are taken away from lipids [19]; cytosolic Fenton reaction, known to be a major electron sink that drives continuous Fenton reaction [20] and the other oxidase genes such as the cytochrome P450, monoamine oxidase, D-amino acid oxidase, myeloperoxidase, protein-methionine sulfoxide oxidase and lysyl oxidase [2].

As for the antioxidation capacity, we consider both the designated antioxidative enzymes, such as SOD, CAT, GPX and PRDX, enzymes crucial in the syntheses of GSH and TXN1/2, the two designated antioxidation systems, and other molecular species known to serve as moonlighting scavengers for oxidizing molecules such as lipids [21, 22], proteins [23–25] and RNA [26, 27], which have different levels of reductive propensities and hence been used when the oxidative stress reaches beyond certain levels, resulting in their damages. Considering that there are no direct marker genes for reflecting the levels of damages of these molecular species, we have used the expression levels of the degradation and/or repair genes for each class of such biomolecules to approximate their damage levels and hence the level for scavenging of oxidizing molecules.

To assess the oxidative stress level, we have collected all the genes annotated to be stress-related, including genes related to the ER stress [28, 29], unfolded protein response [30], apoptosis [30, 31] and necrosis [32, 33]. The expressions of these genes may reflect stresses due to other reasons such as hypoxia or pH imbalance. To deal with this issue, we have applied a set of selection criteria (Materials and Methods) to choose the gene list MG-S. To check the validity of this selection, we note that the first principal component of the selected genes correlates with 96% of MG-S genes having Pearson Correlation Coefficient (PCC) > 0.5 , providing strong evidence that the selected gene set is highly coherent and hence possibly due to the same cause. The detailed gene lists MG-O, MG-R and MG-S, along with the biological function of each gene, are summarized in [Supplementary Tables S1 and S2](#) available online at <https://academic.oup.com/bib>.

Now, Equation (1) for each sample can be written as

$$F_1(x_1, \dots, x_r) = F_2(y_1, \dots, y_m) - F_3(z_1, \dots, z_n) \quad (2)$$

where $\{x_i\}$, $\{y_j\}$ and $\{z_k\}$ are the expressions of the marker genes in MG-S, MG-O and MG-R, with r , m and n being

the numbers of genes in the three gene sets, respectively; and $F_1()$, $F_2()$ and $F_3()$ are (to be determined) functions for integrating the expression levels of the selected genes for oxidative stress, oxidizing molecule production and antioxidation molecules, respectively.

Model parameterization

Our goal here is to train a model for predicting the oxidative stress level in each given sample based on the genes selected above and their expressions in the training dataset. Let $\{x_i\}$, $\{y_k\}$ and $\{z_p\}$ be the expressions of the marker genes in MG-S, MG-O and MG-R, respectively, and X be the set of normal tissues, Y be the cancer-adjacent control tissues and Z be the cancerous tissues in our training data. Assume that $F_1()$, $F_2()$ and $F_3()$ can each be reliably approximated using a quadratic function defined over MG-S, MG-O and MG-R, respectively:

$$\begin{aligned} F_1(\{a_i\}, \{b_{ij}\}) &= \sum_{0 < i \leq r} a_i x_i + \sum_{0 < i, j \leq r, i \neq j} b_{ij} x_i x_j \\ F_2(\{c_k\}, \{d_{k,l}\}) &= \sum_{0 < k \leq m} c_k y_k + \sum_{0 < k, l \leq m, k \neq l} d_{k,l} y_k y_l \\ F_3(\{e_p\}, \{f_{p,q}\}) &= \sum_{0 < p \leq n} e_p z_p + \sum_{0 < p, q \leq n, p \neq q} f_{p,q} z_p z_q \end{aligned} \quad (3)$$

with $\{a_i\}$, $\{b_{ij}\}$, $\{c_k\}$, $\{d_{k,l}\}$, $\{e_p\}$ and $\{f_{p,q}\}$ being unknown parameters to be determined through solving the following optimization problem over the training data. Now the goal is to select values for these parameters that minimize the following function:

$$\min_{\alpha, \beta, \{a_i\}, \{b_{ij}\}, \{c_k\}, \{d_{k,l}\}, \{e_p\}, \{f_{p,q}\}} \sum_{s \in X \cup Y \cup Z} (F_1(s) - (\alpha F_2(s) - \beta F_3(s)))^2$$

subject to:

1. $0 < F_1(s) \leq C_1$ for $s \in X$,
2. $C_1 < F_1(s) \leq C_2$ for $s \in Y$,
3. $C_2 < F_1(s) \leq C_3$ for $s \in Z$,
4. $0 < \alpha, \beta$,
5. $0 \leq \frac{\partial F_1(s)}{\partial x_i}$ for each x_i and $s \in X \cup Y \cup Z$,
6. $0 \leq \frac{\partial F_2(s)}{\partial y_j}$ for each y_j and $s \in X \cup Y \cup Z$,
7. $0 \leq \frac{\partial F_3(s)}{\partial z_k}$ for each z_k and $s \in X \cup Y \cup Z$,

where α and β are two to-be-determined scaling factors. Constraints 1–3 enforce that the to-be-derived $F_1()$ should have higher values for cancer samples than those for cancer-adjacent disease samples, which should have higher values than normal samples. Constraints 5–7 require that $F_i()$ should be a monotonic function with respect to each variable for $i = 1, 2, 3$.

In the current study, $F_1()$ has 36 variables, $F_2()$ has 49 and $F_3()$ has 58, giving rise to a total of 143 constraints under Constraints 5–7. The three functions are trained using gene-expression data of a total of 16 718 samples, consisting of 5620 normal tissues from 18 organs in the GTEx database, 742 cancer-adjacent control samples and 10 356 cancer samples, both from TCGA. This optimization problem is solved as a quadratic programming problem using Python software package SciPy optimize [34].

The trained parameters α , β , $\{a_i\}$, $\{b_{ij}\}$, $\{c_k\}$, $\{d_{k,l}\}$, $\{e_p\}$ and $\{f_{p,q}\}$ that give rise to the minimum solution to the above objective function are given in [Supplementary Table S3](#) available online at <https://academic.oup.com/bib>. To demonstrate the stability of our solved parameters, we have randomly partitioned the training dataset into two halves and solved the minimization problem over each half of the dataset; and repeated this 50 times. [Figure 1A](#) shows the 100 minimum values for the 100 different datasets, indicating that our trained predictor is highly stable. [Figure 1B–D](#) and [Supplementary Figure S1A–H](#), available online at <https://academic.oup.com/bib>, show the predicted oxidative stress levels in 3658 normal tissues, 667 cancer-adjacent tissues and 6762 cancer tissues from GTEx and TCGA, respectively.

Model validation

We have applied the trained predictor to several tissue/cell-based gene-expression datasets, not involved in our model training, whose (relative) levels of oxidative stress are known.

Cell samples treated with H_2O_2 versus controls

H_2O_2 is an oxidizing molecule commonly associated with oxidative stress. We have tested our predictor on gene-expression data in dataset GSE143155 retrieved from the GEO database, which consists of 12 neuron samples treated with 800 μM H_2O_2 for 17 h and 18 untreated samples. Our predictor accurately predicted that the 12 treated samples have considerably higher levels of oxidative stress than the untreated samples, as shown in [Figure 2A](#). We also carried out similar validations on other two datasets (GSE6607 and GSE10896) to show that our predictor performs equally well on datasets independently collected under different levels of oxidative stress, as shown in [Figure S5](#).

Tissue samples from smokers versus non-smokers

It is known that long-time cigarette smoking leads to oxidative stress in lung airways [35, 36]. We have tested our predictor on gene-expression data in GSE10006, also from GEO, which consists of 20 large-airway and 18 small-airway tissues of smokers and nine large-airway and 13 small-airway tissues of non-smokers. Our model predicts that the smoker samples have significantly higher levels of oxidative stress than those of the non-smokers, as shown in [Figure 2B](#). We also tested the predictor on gene-expression data in GSE37768 from GEO, consisting of peripheral lung tissues from 11 smokers and 9 non-smokers. Again, the smoker samples have a significantly higher level of oxidative stress than those of the non-smokers, as shown in [Figure 2C](#).

Early versus advanced chronic obstructive pulmonary disease tissues

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease and oxidative stress is a known contributing factor [37, 38]. We have applied our predictor to

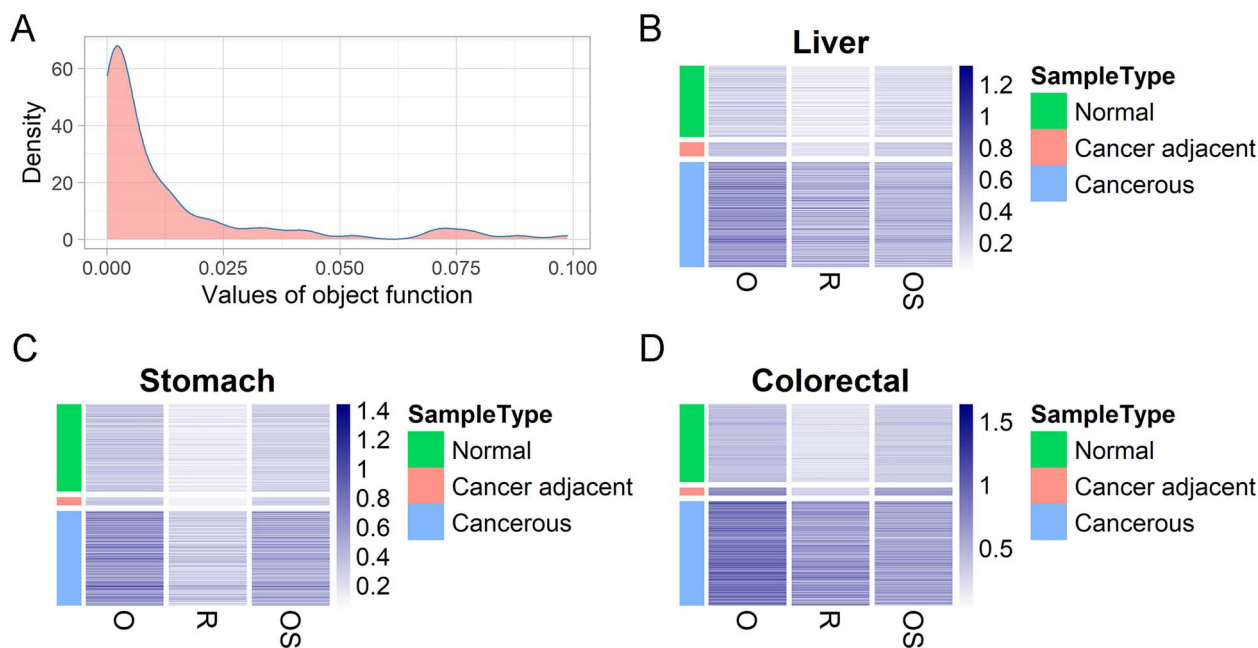


Figure 1. Model validation. (A) The density distribution of the 100 minimum values of the objective function for 100 simulations. The x-axis is for the minimized objective value achieved over randomly selected 50% of the training data and the y-axis is for the density of each achieved objective value over 100 runs (Note: the pink color has no special meaning). Heatmaps for predicted production rates of oxidizing molecules (O), activated capacity for antioxidation molecules (R) and the oxidative stress level (OS) across normal, cancer-adjacent, and cancer tissue samples of different organs: (B) Liver tissues, (C) stomach tissues and (D) colorectal tissues. Results for eight other organs are shown in [Supplementary Figure S1](#) available online at <https://academic.oup.com/bib>.

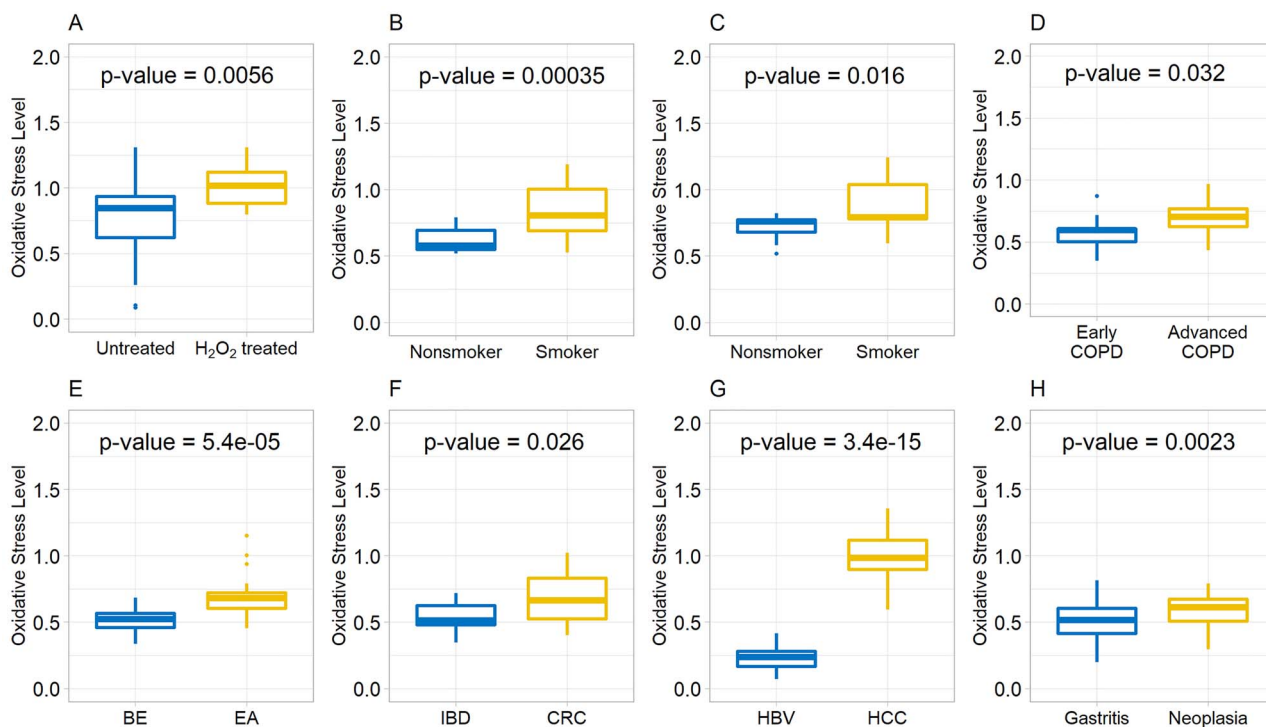


Figure 2. (A) Boxplots of predicted oxidative stress levels on samples treated with H₂O₂ versus untreated. (B) Boxplots of the predicted oxidative stress levels on airway samples of smokers versus non-smokers. (C) Boxplots of the predicted oxidative stress levels in lung tissues of smokers versus non-smokers. (D) Boxplots of the predicted oxidative stress levels in early versus advanced COPD samples. (E) Boxplots of the predicted oxidative stress levels in Barrett's esophagus (BE) versus esophageal adenocarcinoma (EA) tissues. (F) Boxplots of predicted oxidative stress levels in IBD versus CRC samples. (G) Boxplots of predicted oxidative stress levels in HBV versus HCC samples. (H) Boxplots of predicted oxidative stress levels in gastritis versus gastric neoplasia samples.

Table 1. PCCs between the predicted oxidative stress levels and Cys-p and Cys-t across each of the four types of tissue samples

	Normal	Cancer-adjacent	Primary cancer	Metastatic cancer
PCC between Cys-t and OS level	0.5	0.4	0.77	0.79
PCC between Cys-p and OS level	0.52	0.35	0.43	0.31

Table 2. PCCs between cysteine statistics and estimated oxidative stress levels across different organs

Organ	Cys-p	Organ	Cys-t
Brain	0.937	Pancreas	0.763
Adrenal gland	0.908	Adrenal gland	0.692
Kidney	0.905	Stomach	0.688
Thyroid	0.743	Liver	0.686
Breast	0.715	Brain	0.686
Bladder	0.701	Prostate	0.674
Esophagus	0.697	Esophagus	0.663
Bone marrow	0.692	Colorectal	0.644
Ovary	0.685	Lung	0.607
Cervix	0.677	Breast	0.538
Uterus	0.586	Lymph nodes	0.509
Stomach	0.57	Ovary	0.483
Testis	0.539	Bladder	0.441
Liver	0.501	Kidney	0.404
Prostate	0.487	Cervix	0.25
Colorectal	0.232	Bone marrow	0.195
Lung	0.191	Testis	0.149
Pancreas	-0.724	Thyroid	0.119

the gene-expression data in GSE10006 in GEO, consisting of 13 small-airway tissues of early COPD patients and 14 samples of advanced COPD patients. [Figure 2D](#) displays our prediction results, showing that the oxidative stress levels in early disease samples are consistently lower than the advanced ones as expected.

Oxidative stress versus cysteine metabolism

Amino acid cysteine plays key antioxidation roles since its sulfur atom can donate up to eight electrons [39]. We have assessed the relationships between the predicted oxidative stress levels and the intracellular levels of cysteine over the 16 718 tissues discussed earlier, with 5620 normal, 742 cancer-adjacent, 9962 primary cancer tissues and 394 metastatic cancer tissues ([Supplementary Tables S4B](#) and [C](#) available online at <https://academic.oup.com/bib>). We have used the following two quantities to reflect the intracellular cysteine level: Cys-p for the total number of cysteine residues in all the expressed proteins and Cys-t for the total expression level of cysteine importers (Materials and Methods). [Tables 1](#) and [2](#) list the PCCs between the predicted oxidative stress levels and Cys-p and Cys-t across each of the four types of tissue samples.

We note that (i) the predicted oxidative stress levels strongly correlate with both Cys-p and Cys-t; (ii) when the oxidative stress level increases, different organs utilize different ways to use cysteine as a response. Some organs, like thyroid and testis, produce more cysteine-containing proteins, while some organs, such as pancreas

and lung, import more cysteines into the cells, which could be used for GSH syntheses. Most organs, including brain and adrenal gland, do both.

Overall, these validation results provide strong supporting evidence that our predictor is a reliable tool for estimating the oxidative stress level in both cell and tissue samples.

Model application

We have applied the validated predictor to numerous disease tissues of different types and made a number of discoveries.

Oxidative stress in chronic diseases versus corresponding cancers

We have studied the oxidative stress levels in a few chronic diseases and cancers in the same organ types. The following summarizes our findings.

Barrett's esophagus versus esophageal adenocarcinoma

Barrett's esophagus is known to be precancerous. We have applied our predictor to the gene-expression data in GSE26886 of GEO, which consists of 20 Barrett's esophagus tissues and 21 esophageal adenocarcinoma tissues. The esophageal adenocarcinoma tissues have significantly higher levels of oxidative stress than Barrett's esophagus tissues, as predicted by our predictor and detailed in [Figure 2E](#). This result is consistent with published studies [40–42].

Inflammatory bowel disease (IBD) versus colorectal cancer (CRC)

A similar study was conducted between IBD and CRC tissues. We have applied our predictor to the gene-expression data in GSE4183 of GEO, consisting of 15 IBD samples and 15 CRC samples. As expected, CRC tissues have higher oxidative stress levels than the IBD samples, as shown in [Figure 2F](#).

Chronic hepatitis B virus (HBV) infection versus hepatocellular carcinoma (HCC)

We have compared the oxidative stress levels between 21 pairs of HCC and non-neoplastic HBV-infected liver tissues using gene-expression data in GSE94660 from GEO. As shown in [Figure 2G](#), the oxidative stress levels of HCC samples are significantly higher than the HBV-infected non-cancerous samples, as predicted by our predictor.

Gastritis versus gastric neoplasia

We have predicted and compared the oxidative stress levels between 31 pairs of precancerous gastric lesions and

non-neoplastic gastritis samples using gene-expression data in GSE130823 from GEO. As shown in Figure 2H, the oxidative stress levels in precancerous samples are significantly higher than in the non-cancerous samples.

Oxidative stress levels across different cancers

We have estimated the levels of oxidative stress over 10 336 tissue samples of 33 cancer types and 742 cancer-adjacent control samples from TCGA and 5620 normal samples of 18 organs from GTEx and derived the following information.

Oxidative stress level versus survival rate

We have considered all samples of 16 of all the 33 cancer types in TCGA, each having at least 100 samples with survival data to study the relationship between the oxidative stress level and the survival rate (Supplementary Table S5 available online at <https://academic.oup.com/bib>). For each cancer type, we divide all its samples into two groups: samples whose predicted oxidative stress levels are among the top 50% of all cancer samples and the remaining 50%. Our finding is that for 11 of the 16 cancer types, the more oxidatively stressed cancer tissues have significantly lower survival rates compared to the less oxidatively stressed tissues (P -value < 0.1), as shown in Figure 3A and B and Supplementary Figure S2A–N available online at <https://academic.oup.com/bib>.

We have predicted the oxidative stress levels for all cancer tissues of all cancer types in TCGA with matching organs in GTEx, resulting in 22 cancer types (Supplementary Table S6 available online at <https://academic.oup.com/bib>). For each of these cancer types, we have used the median of the oxidative stress levels predicted for the GTEx samples of each organ as the baseline oxidative-stress level, B_c of the organ. We note that the ratio OS/B_c provides a strong indicator for 5-year survival as shown in Figure 3C and D, where OS is the average oxidative stress level across all cancer tissues of each cancer type under consideration.

For the considered 22 cancer types, we have conducted regression analyses of the average 5-year survival rate against OS and OS/B_c , respectively. As shown in Figure 3C and D, the normalized oxidative stress level (with respect to B_c of the organ) is a better predictor for survival compared to the absolute oxidative stress level. The linear regression analysis against OS/B_c achieves $R^2 = 0.33$ with a P -value of 0.005, while the linear regression analysis against OS is not as good with P -value being 0.33.

For each of the 22 cancer types, we have conducted a regression analysis of its average 5-year survival rate y against $v_1 = OS/B_c$ of the cancer type and $v_2 = B_c$ of the organ, as below:

$$y = 2.49 - 0.30v_1 - 6.56v_2 + 9.63v_2^2 - 4.45v_2^3$$

which achieves $R^2 = 0.46$ with a P -value of 0.03.

Oxidative stress level versus cancer stage

We have examined how the oxidative stress level changes as a cancer progresses. Specifically, we have considered all cancer types in TCGA having at least 10 tissue samples for each of the four stages, resulting in a total of 10 cancer types (Supplementary Table S7 available online at <https://academic.oup.com/bib>). Our results revealed that in 8 of the 10 cancer types except for KIRC and SKCM, the oxidative stress level averaged over all samples generally increases with stage, with at most one stage out of order, as shown in Figure 4A–C and Supplementary Figure S3A–G available online at <https://academic.oup.com/bib>.

Oxidative stress level versus cancer grade

A similar analysis is conducted between oxidative stress levels versus cancer grades. Out of all cancer types in TCGA, 11 types have grade information, totaling 2976 samples (Supplementary Table S8 available online at <https://academic.oup.com/bib>). For each cancer type, we divide all samples into two groups: low-grade (G1 and G2) and high-grade (G3 and G4). We note that the level of oxidative stress goes up from the low-grade to the high-grade group in 8 of the 11 cancer types with exception of ESCA, KIRC and STAD, as shown in Figure 4D–F and Supplementary Figure S4A–H available online at <https://academic.oup.com/bib>. The result suggests that the oxidative-stress level may play an important role in dictating the grade of a cancer, a novel insight into the best of our knowledge.

Oxidative stress versus cancer metastasis

It has been suggested that metastasis represents a survival strategy for cancer cells to escape from the excessive reactive oxygen species in their primary sites [43]. We have calculated the first principal component of the expression data of all epithelial-mesenchymal transition (EMT)-related genes (Supplementary Table S9 available online at <https://academic.oup.com/bib>) across all 10 336 cancer samples in TCGA and noted that the PCC between the first principal component of the EMT genes and the predicted oxidative stress levels is 0.80 with a P -value of $< 10^{-5}$.

We have then studied three cancer types in TCGA, each having at least five non-metastatic primary cancer samples and at least five samples known to have metastasized to a distant organ (Supplementary Table S10 available online at <https://academic.oup.com/bib>). We note that the predicted oxidative stress levels for samples known to have metastasized are considerably higher than the samples not metastasized yet in all three cancer types (Figure 4G–I). This and the above result strongly suggest that oxidative stress is highly involved in cancer metastasis.

To further support this postulation, we have applied our predictor to another independent dataset GSE7553 in GEO, with 40 melanoma samples known to have metastasized and 14 that have not metastasized. Again, the samples that have metastasized have a considerably

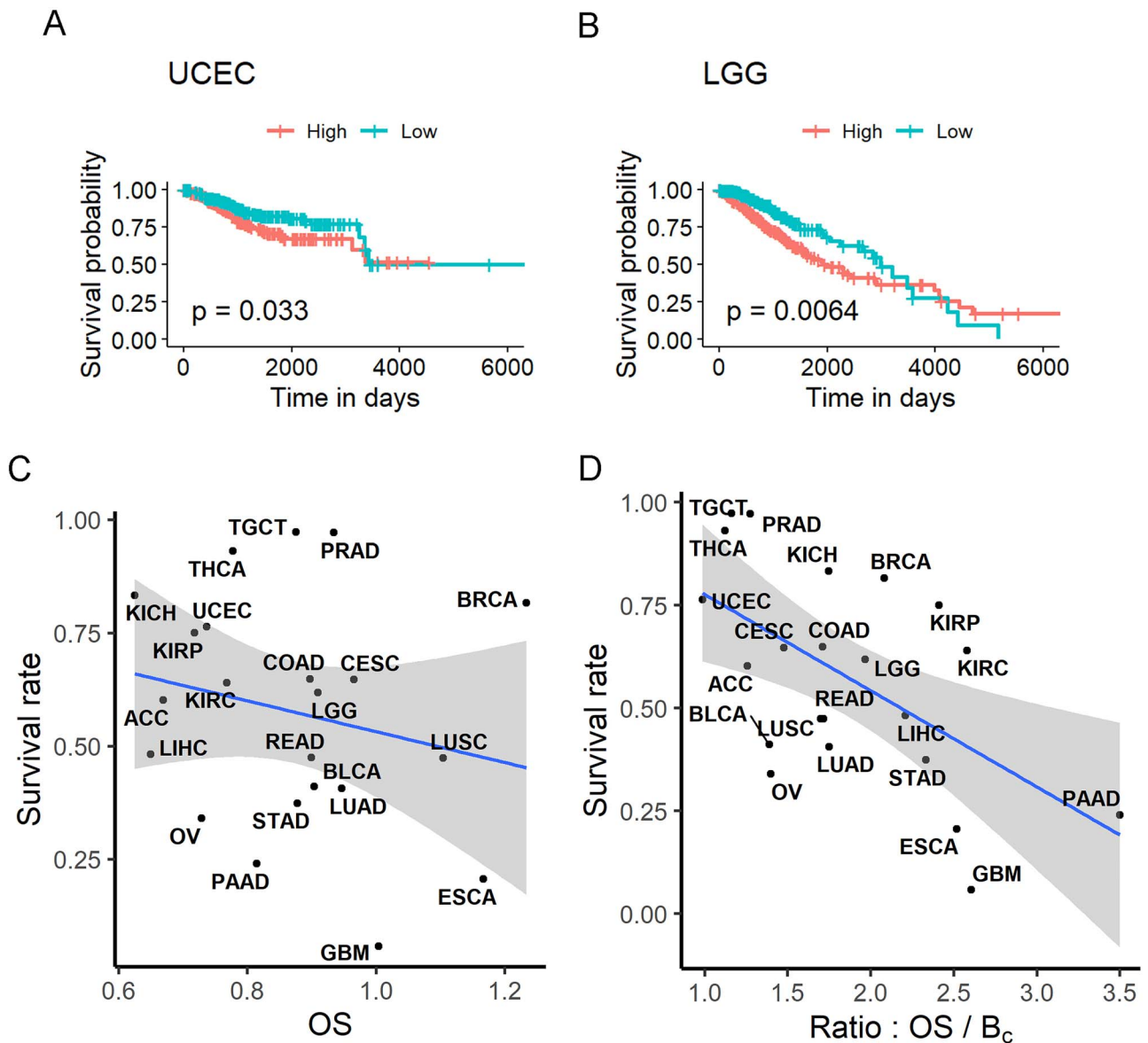


Figure 3. Oxidative stress level versus survival rate. Survival curves for patients in the low oxidative stress group (blue) and in the high oxidative stress group (pink) for (A) UCEC and (B) LGG. Results for 14 other cancer types are shown in [Supplementary Figure S2](#) available online at <https://academic.oup.com/bib>. (C) Relationship between the absolute oxidative stress and five-year survival rate across 22 cancer types. The x-axis is for the median values of the predicted oxidative stress (OS) for 22 cancer types, and the y-axis is for the 5-year survival rate. (D) Relationship between the normalized oxidative stress and the five-year survival rate across 22 cancer types. The x-axis is for the ratio between the median values of the predicted oxidative stress for 22 cancer types and the baseline oxidative stress in each organ: OS/ B_c , and the y-axis is for the 5-year survival rate.

higher level of oxidative stress than those that have not metastasized (Figure 4).

Discussion

A novel and general framework for estimating the intracellular oxidative stress level in human tissue or cell samples is presented, the first of its kind based on the best of our knowledge. The novel idea lies in estimating three main quantities: the total oxidizing power, the activated antioxidation capacity and the intracellular oxidative stress through the selection of three sets of marker genes and integrating them via solving an optimization problem.

To accurately estimate each of the three quantities, we have methodically gone through all the human protein-encoding genes and assessed their relevance to any of the three quantities for inclusion. Under the assumption that each of the three quantities can be reliably approximated as a quadratic function of its contributing genes, we have formulated the problem of estimating the oxidative stress level as a quadratic programming problem and solved it rigorously. The validity of the predictor is assessed over a large number of validation and application problems as the prediction results are highly consistent with numerous published experimental datasets and our general understanding about the relevant biology.

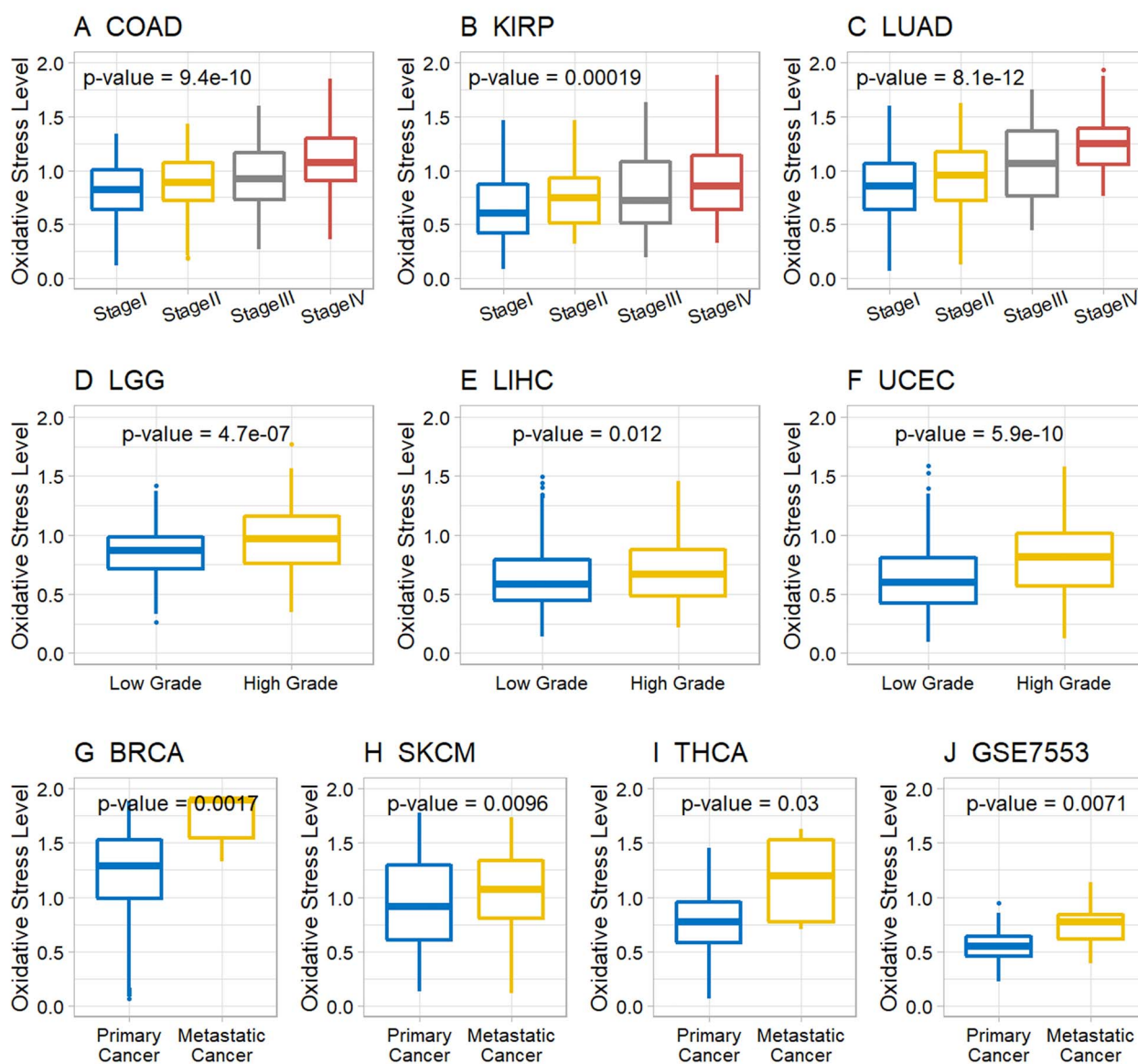


Figure 4. Boxplots of the predicted oxidative stress levels for different cancer stages, grades and primary versus metastatic cancers. Boxplots of predicted oxidative stress levels for different cancer stages in (A) COAD, (B) KIRP and (C) LUAD. Results for seven other cancer types are shown in [Supplementary Figure S3](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>. Boxplots of predicted oxidative stress levels for different cancer grades in (D) LGG, (E) LIHC and (F) UCEC. Results for eight other cancer types are shown in [Supplementary Figure S4](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>. Boxplots of predicted oxidative stress levels in primary and corresponding metastatic cancers in (G) BRCA, (H) SKCM, (I) THCA and (J) GSE7553.

A key reason that we have modeled the three quantities as quadratic functions instead of higher-degree polynomial functions is largely limited by our computing power as the number of constraints under Constraints 5–7 will be considerably increased if cubic or higher-degree polynomial functions are used, making direct applications of the quadratic programming solver infeasible. We anticipate that this is not an unsolvable issue as two approaches are currently being undertaken, with one through simplification of the current model via analytic methods to reduce the number of constraints and another through seeking collaboration with labs with more powerful computing power.

A few discoveries are made, enabled by this new predictor. Human cells seem to have a great capacity to cope

with oxidative stress for survival through employing non-designated molecules to neutralize the increasing oxidizing power when the designated antioxidation systems are saturated as in the case of utilizing increasingly more cysteine-containing proteins when coping with increasing oxidative stress levels. There have been numerous studies on metabolic reprogramming (MR) in cancer [44] and other diseases but there have not been any systematic studies on MRs driven by oxidative stress. We anticipate that the new predictor will enable such studies.

A highly unexpected result is the observation that oxidative stress may play a driving role in cancer metastasis, a hypothesis postulated by a previous study [45]. This is also consistent with a well-known observation that more hypoxic cancer tends to have poorer prognosis

[46] as hypoxia is essentially the other side of the same coin of oxidative stress since the main reason for hypoxia in cancer is the large consumption of O₂ for producing oxidizing molecules. We anticipate that systematic studies to identify all MRs associated with the oxidative stress could lead to a novel understanding about the cellular processes leading to metastasis.

We speculate, based on data here and previous studies, that improved understanding could be enabled by this new predictor: we could use this model to study the association between oxidative stress and other issues, such as cancer occurrence, drug resistance and so on, by exploring whether oxidative stress is involved and what roles it plays in these issues [47, 48]. It would facilitate a better understanding of many biological processes in a redox perspective.

Compared to related studies, our model represents the only predictor for predicting the level of intracellular oxidative stress based on given omic data except for one predictor our group published two years ago, which is basically an infant version of the current predictor [13]. To the best of our knowledge, other predictors are all for predicting biomarkers for specific types of oxidative stressors such as the carbonylated proteins [9], oxidized low-density lipoproteins and oxidized lipids [10, 11] rather than predicting the levels of oxidative stress.

It is noteworthy that our predicted stress level is a numerical value without a physical 'unit'. This problem could be potentially resolved through (i) collecting samples with experimentally measured stress levels and matching transcriptomic data; and (ii) calibrating our predicted stress levels against such measured stress levels. Unfortunately, there is no publicly available transcriptome data with matching experimentally measured oxidative stress, to the best of our knowledge. Also, the current predictor treats the intracellular stress as a whole and does not distinguish among stresses in individual cellular compartments as some compartments such as mitochondria may have higher levels of oxidative stresses compared to other compartments in cancer. This will be one of the areas that our future work will focus on to improve.

Materials and methods

Data

RNA-seq data of 11 098 tissue samples of 33 cancer types were retrieved from the TCGA database, along with those of 5620 samples from 18 organs from GTEx and 366 samples from eight additional datasets in GEO, which are detailed in [Supplementary Table S4](#) available online at <https://academic.oup.com/bib>.

The recount3 program in the R package was used for normalization of the RNA-seq data from TCGA and GTEx [49], so cross-platform gene-expressions can be compared directly. In addition, GEOquery in R was used for downloading gene-expression data from GEO.

Molecular Signatures Database and published studies [1, 2, 20] were used for selecting oxidative stress-related genes.

Selection of oxidative stress-related genes

We have selected a subset of human genes known to associate with oxidative stress using the following criteria: (i) they are involved in oxidative stress-induced processes as collected above and (ii) their expressions correlate with at least 50% genes involved in the production of oxidizing molecules across all the TCGA and GTEx samples.

Cysteine statistics

We have estimated the intracellular level of cysteine using two indicators: Cys-p for the total number of cysteines present in the expressed proteins and Cys-t for the total expression level of cysteine importers. For each expressed protein p , let $C(p)$ be the number of cysteines in the protein and $E(p)$ be the gene-expression level of p . Then, Cys-p is defined as follows across all expressed proteins in each sample:

$$\text{Cys-p} = \sum_p E(p)C(p)$$

Cys-t is defined as the total expression level of three major cysteine importers, namely, *SLC7A11*, *SLC1A4* and *SLC1A5*.

Significance test

t-test is used to assess if the oxidative stress levels between two groups of samples are significantly different. The R function 't.test' is used to calculate the P-value. The ANOVA test is used to assess if the oxidative stress levels among more than two different groups are significantly different. The R function 'aov' is used to calculate the P-value.

Kaplan–Meier survival analysis

The Kaplan–Meier survival analysis is conducted to evaluate the difference in survival rates between the group with high oxidative-stress levels and the group with low oxidative-stress level using R function 'survfit'.

Key Points

- The paper presents a novel model for quantitatively estimating the level of oxidative stress in human tissues and cells based on their transcriptomic data.
- The novel idea lies in estimating three main quantities: the total oxidizing power, the activated antioxidation capacity and the intracellular oxidative stress through the selection of three sets of marker genes and integrating them via solving an optimization problem. Systematic assessments indicate that the predictor is highly reliable.
- Several discoveries have been made, enabled by the new predictor, such as that oxidative stress may play a driving role in cancer metastasis.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Data availability

The data used in this study are openly available in TCGA (<https://portal.gdc.cancer.gov/>), GTEx (<https://GTExportal.org/home/>) and GEO (<https://www.ncbi.nlm.nih.gov/geo/>). The code to conduct the simulations and reproduce the analyses is available at <https://github.com/zaihebian/OXIS-oxidative-stress-predictor>.

Authors' contributions

Y.X. conceived the project and conducted the study design; J.B., R.T. and Z.A. collected the data and did literature review; J.B. performed the computational analyses; and J.B., R.T. and Y.X. wrote the paper. All authors read and approved the final manuscript.

Acknowledgement

All authors thank colleagues in the Computational Systems Biology Lab in the Department of Biochemistry and Molecular Biology, the University of Georgia for helpful discussions regarding this project.

References

- Forman HJ, Zhang H. Targeting oxidative stress in disease: promise and limitations of antioxidant therapy. *Nat Rev Drug Discov* 2021;**20**:689–709.
- Hayes JD, Dinkova-Kostova AT, Tew KD. Oxidative stress in cancer. *Cancer Cell* 2020;**38**:167–97.
- He L, He T, Farrar S, et al. Antioxidants maintain cellular redox homeostasis by elimination of reactive oxygen species. *Cell Physiol Biochem* 2017;**44**:532–53.
- Reczek CR, Chandel NS. The two faces of reactive oxygen species in cancer. *Annu Rev Cancer Biol* 2017;**1**:79–98.
- Walker LM, York JL, Imam SZ, et al. Oxidative stress and reactive nitrogen species generation during renal ischemia. *Toxicol Sci* 2001;**63**:143–8.
- Weidinger A, Kozlov AV. Biological activities of reactive oxygen and nitrogen species: oxidative stress versus signal transduction. *Biomolecules* 2015;**5**:472–84.
- Selvaraj N, Bobby Z, Das AK, et al. An evaluation of level of oxidative stress and protein glycation in nondiabetic undialyzed chronic renal failure patients. *Clin Chim Acta* 2002;**324**:45–50.
- Farah IO. Assessment of cellular responses to oxidative stress using MCF-7 breast cancer cells, black seed (*N. Sativa* L.) extracts and H₂O₂. *Int J Environ Res Public Health* 2005;**2**:411–9.
- Dalle-Donne I, Rossi R, Giustarini D, et al. Protein carbonyl groups as biomarkers of oxidative stress. *Clin Chim Acta* 2003;**329**:23–38.
- Ho E, Karimi Galoughi K, Liu C-C, et al. Biological markers of oxidative stress: applications to cardiovascular research and practice. *Redox Biol* 2013;**1**:483–91.
- Dalleau S, Baradat M, Guéraud F, et al. Cell death and diseases related to oxidative stress: 4-hydroxynonenal (HNE) in the balance. *Cell Death Differ* 2013;**20**:1615–30.
- Zinellu A, Fois AG, Sotgia S, et al. Plasma protein thiols: an early marker of oxidative stress in asthma and chronic obstructive pulmonary disease. *Eur J Clin Invest* 2016;**46**:181–8.
- Liu L, Cui H, Xu Y. Quantitative estimation of oxidative stress in cancer tissue cells through gene expression data analyses. *Front Genet* 2020;**11**:494.
- Thomas DC. The phagocyte respiratory burst: historical perspectives and recent advances. *Immunol Lett* 2017;**192**:88–96.
- Fisher AB. Redox signaling across cell membranes. *Antioxid Redox Signal* 2009;**11**:1349–56.
- Chio IIC, Tuveson DA. ROS in cancer: the burning question. *Trends Mol Med* 2017;**23**:411–29.
- Csordás G, Hajnóczky G. SR/ER-mitochondrial local communication: calcium and ROS. *Biochim Biophys Acta* 2009;**1787**:1352–62.
- Grishko VI, Druzhyna N, LeDoux SP, et al. Nitric oxide-induced damage to mtDNA and its subsequent repair. *Nucleic Acids Res* 1999;**27**:4510–6.
- Wauchope OR, Mitchener MM, Beavers WN, et al. Oxidative stress increases M1dG, a major peroxidation-derived DNA adduct, in mitochondrial DNA. *Nucleic Acids Res* 2018;**46**:3458–67.
- Sun H, Zhang C, Cao S, et al. Fenton reactions drive nucleotide and ATP syntheses in cancer. *J Mol Cell Biol* 2018;**10**:448–59.
- Auten RL, Davis JM. Oxygen toxicity and reactive oxygen species: the devil is in the details. *Pediatr Res* 2009;**66**:121–7.
- Perillo B, di Donato M, Pezone A, et al. ROS in cancer therapy: the bright side of the moon. *Exp Mol Med* 2020;**52**:192–203.
- Friguet B. Oxidized protein degradation and repair in ageing and oxidative stress. *FEBS Lett* 2006;**580**:2910–6.
- Stadtman ER. Protein oxidation in aging and age-related diseases. *Ann NY Acad Sci* 2001;**928**:22–38.
- Dasuri K, Zhang L, Keller JN. Oxidative stress, neurodegeneration, and the balance of protein degradation and protein synthesis. *Free Radic Biol Med* 2013;**62**:170–85.
- Kong Q, Lin C-LG. Oxidative damage to RNA: mechanisms, consequences, and diseases. *Cell Mol Life Sci* 2010;**67**:1817–29.
- Yan LL, Zaher HS. How do cells cope with RNA damage and its consequences? *J Biol Chem* 2019;**294**:15158–71.
- Bhattacharai KR, Riaz TA, Kim HR, et al. The aftermath of the interplay between the endoplasmic reticulum stress response and redox signaling. *Exp Mol Med* 2021;**53**:151–67.
- Liu MQ, Chen Z, Chen LX. Endoplasmic reticulum stress: a novel mechanism and therapeutic target for cardiovascular diseases. *Acta Pharmacol Sin* 2016;**37**:425–43.
- Kupsco A, Schlenk D. Oxidative stress, unfolded protein response, and apoptosis in developmental toxicity. *Int Rev Cell Mol Biol* 2015;**317**:1–66.
- Kannan K, Jain SK. Oxidative stress and apoptosis. *Pathophysiology* 2000;**7**:153–63.
- Hanus J, Zhang H, Wang Z, et al. Induction of necrotic cell death by oxidative stress in retinal pigment epithelial cells. *Cell Death Dis* 2013;**4**:e965.
- Fiers W, Beyaert R, Declercq W, et al. More than one way to die: apoptosis, necrosis and reactive oxygen damage. *Oncogene* 1999;**18**:7719–30.
- Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;**17**:261–72.
- Liguori I, Russo G, Curcio F, et al. Oxidative stress, aging, and diseases. *Clin Interv Aging* 2018;**13**:757–72.
- Isik B, Ceylan A, Isik R. Oxidative stress in smokers and non-smokers. *Inhal Toxicol* 2007;**19**:767–9.

37. Rahman I. The role of oxidative stress in the pathogenesis of COPD: implications for therapy. *Treat Respir Med* 2005;**4**: 175–200.
38. Kirkham PA, Barnes PJ. Oxidative stress in COPD. *Chest* 2013;**144**: 266–73.
39. Araki K, Kusano H, Sasaki N, et al. Redox sensitivities of global cellular cysteine residues under reductive and oxidative stress. *J Proteome Res* 2016;**15**:2548–59.
40. O'Farrell NJ, Phelan JJ, Feighery R, et al. Differential expression profiles of oxidative stress levels, 8-oxo-dG and 4-HNE, in Barrett's esophagus compared to esophageal adenocarcinoma. *Int J Mol Sci* 2019;**20**:4449.
41. Song JH, Han YM, Kim WH, et al. Oxidative stress from reflux esophagitis to esophageal cancer: the alleviation with antioxidants. *Free Radic Res* 2016;**50**:1071–9.
42. Hardikar S, Onstad L, Song X, et al. Inflammation and oxidative stress markers and esophageal adenocarcinoma incidence in a Barrett's esophagus cohort. *Cancer Epidemiol Biomark Prev* 2014;**23**:2393–403.
43. Pani G, Galeotti T, Chiarugi P. Metastasis: cancer cell's escape from oxidative stress. *Cancer Metastasis Rev* 2010;**29**:351–78.
44. Sun H, Zhou Y, Skaro MF, et al. Metabolic reprogramming in cancer is induced to increase proton production. *Cancer Res* 2020;**80**:1143–55.
45. Liao Z, Chua D, Tan NS. Reactive oxygen species: a volatile driver of field cancerization and metastasis. *Mol Cancer* 2019;**18**:65.
46. Walsh JC, Lebedev A, Aten E, et al. The clinical importance of assessing tumor hypoxia: relationship of tumor hypoxia to prognosis and therapeutic opportunities. *Antioxid Redox Signal* 2014;**21**:1516–54.
47. Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Prz Gastroenterol* 2019;**14**:89–103.
48. Boursi B, Arber N. Small bowel malignancies: Why are they so rare? *Harefuah* 2004;**143**:727–732, 765.
49. Wilks C, Zheng SC, Chen FY, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol* 2021;**22**:323.