Integrated Bioinformatics Data Analysis and Experimental Studies Reveals Prognostic Significance of ALDH Family in Endometria Cancer

Zhou-Tong Dai

Wuhan University of Science and Technology

Yuan Xiang

Wuhan University of Science and Technology

Xing-Hua Liao (✉ xinghualiao@hotmail.com)

Wuhan University of Science and Technology

Research

Keywords: ALDH2, Uterine Corpus Endometrial Carcinoma, miR-135-3p, Overall survival risk score model, bioinformatics.

Posted Date: May 7th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-427053/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Uterine Corpus Endometrial Cancer (UCEC) is one of the three common malignant tumors of the female reproductive tract. According to reports, the cure rate of early UCEC can reach 95%. Therefore, the development of prognostic markers will help UCEC patients to find the disease earlier and develop treatment earlier. The ALDH family was first discovered to be the essential gene of the ethanol metabolism pathway in the body. Recent studies have shown that ALDH can participate in the regulation of cancer.

Methods

We used the gene profile data of 33 cancers in the TCGA database to analyze the expression and survival of the ALDH family. GO, KEGG, PPI multiple functional analysis was used to predict the regulatory role of ALDH family in cancer. In addition, using CCK-8, colony formation, nude mouse tumor formation and other methods, the in vitro function of UCEC cancer cell lines was tested to further confirm the key role of ALDH2 expression in the proliferation of UCEC cell lines. Finally, Lasso and Cox regression methods were used to establish an overall survival prognosis model based on ALDH2 expression.

Result

In our research, we explored the expression of ALDH family in 33 cancers. It was found that ALDH2 was abnormally expressed in UCEC. Besides, in vivo and in vitro experiments were conducted to explore the effect of ALDH2 expression on the proliferation of UCEC cell lines. Meanwhile, the change of its expression is not due to gene mutations, but is regulated by miR-135-3p. At the same time, the impact of ALDH2 changes on the survival of UCEC patients is deeply discussed. Finally, a nomogram for predicting survival was constructed, with a C-index of 0.798 and AUC of 0.764.

Conclusion

This study suggests that ALDH2 may play a crucial role in UCEC progression and has the potential as a prognostic biomarker of UCEC.

Background

Gynecological cancers include ovarian cancer, endometrial cancer, vaginal cancer, cervical cancer and vulvar cancer (1). EC is a common gynecological malignant tumor. According to statistics, in 2018, there were more than 380,000 new cases worldwide (2). In developed countries, endometrial cancer is the most common gynecological cancer. There are more than 60,000 confirmed cases, more than 10,000 deaths, and the death rate is increasing year by year (3). However, in recent decades, studies have found that early screening and intervention can significantly reduce the incidence and mortality of endometrial cancer (4–6). The above

data shows that early screening of patients with a high risk of endometrial cancer is essential, and the discovery of early diagnosis genes will help prevent the occurrence and development of these diseases.

In traditional research, the discovery of new diagnostic markers and prognostic factors through sequencing has the disadvantages of a long time and high cost. The emergence of bioinformatics brings us new research methods for the rapid discovery of new markers and prognostic factors. Recently, more and more studies have shown that bioinformatics can be used as a reliable tool for cancer research [7–9]. In lung cancer, Zhang uses TCGA to conduct a comprehensive analysis of the prognosis, immune function and immune markers based on TNF family markers and developed a risk survival model based on TCGA samples (7). Similarly, some studies have found that not all TP53 mutations (mainly referring to missense and nonsense mutations) with the help of bioinformatics can effectively predict the therapeutic effect of patients with lung adenocarcinoma (LUAD) on ICIs (8). However, Li used multiple sets of bioinformatics in the public database to study the expression characteristics, prognostic value, immune infiltration pattern and biological function of Siglec-15, and verified it in patients (9). In endometrial carcinoma, Wang developed a six-gene prognostic model to predict overall survival (10). These research results show that exploring new cancer markers and developing new prognostic tools through public databases is a reliable research method.

The human acetaldehyde dehydrogenase family (ALDH family) has 19 members, including ALDH1, ALDH3, ALDH2, etc. ALDH2 has become a hot spot in current research due to the importance of its function. They were first discovered to play a critical role in the oxidation of ethanol in humans (11). The alcohol dehydrogenase in the liver cytoplasm first metabolizes alcohol to acetaldehyde, and then the acetaldehyde dehydrogenase in the mitochondria gradually metabolizes acetaldehyde into acetic acid. Finally, it is completely decomposed into water and carbon dioxide in the peripheral circulation. However, more and more recent studies have shown that the ALDH family plays an essential role in the occurrence and development of cancer. In gastric cancer, ALDH1 can be used as a tumor stem cell marker (12, 13). Moreover, the positive expression of ALDH1A1 is associated with low overall survival time and progression-free survival time (14). In breast cancer, citral reduces the growth of breast cancer tumors by inhibiting the breast stem cell marker ALDH1A3 (15). However, in prostate cancer, contrary to tumor tissues, the expression of matrix ALDH1 improves clinical outcome, and is less frequent in PCa metastases (16). However, there is no report on the relationship between ALDH expression and clinical in Uterine Corpus Endometrial Carcinoma.

In this study, we comprehensively explained the prognostic significance of ALDH2 in UCEC patients. At the same time, a new 6-factor prognostic risk scoring model was developed based on the expression of ALDH2. Furthermore, the effect of ALDH2 expression on the proliferation ability of UCEC was preliminarily discussed. And the regulation pathway of ALDH2.. These findings can provide theoretical support for UCEC patients to develop new personalized treatments in the future.

Materials And Methods

Data collection

Expression data and clinical annotation data of 33 types of cancer patients were downloaded from the TCGA data portal (tcga-data.nci.nih.gov/tcga). Perl software (Version 5.3.2) was used in the 33 types of cancer to sort the data to analyze the data for each patient. The expression profile data of normal human endometrial cancer is obtained from the GTEx database (<http://www.gtexportal.org>). All databases data were accessed in September 2020.

Analysis of overall survival

R with survival and survminer packages used the Kaplan-Meier method and Cox regression to analyze survival between different groups. P-value of less than 0.05 is considered statistically significant.

Analysis of different genes

R software with limma package was used to analyze the different genes between the data of each group, and $\log_2$ foldchange > 0.5 and adjust P value < 0.05 was used as the standard.

Analysis of gene and pathway enrichment

In order to determine the biological processes, cell components, molecular functions and biological pathways of differential gene enrichment, R software with clusterProfiler and org.Hs.eg.db packages was used for gene ontology (GO) enrichment and Kyoto Gene and Genome Encyclopedia The whole book (KEGG) approach analysis (17–19). At the same time, Gene Set Enrichment Analysis (GSEA) was used to analyze related functions among different groups. The screening criterion is $P < 0.05$.

Construction of protein-protein interaction (PPI) network

The STRING database was used to construct a protein interaction network of differential genes (20). The screening criterion was a comprehensive score of ≥ 0.9 . Cytoscape software with Cytohubba and Mcode apps was used to visualize and screen highly critical genes and core sub-networks.

Model construction and verification

In order to construct the prognostic risk scoring system for UCEC, univariate Cox regression analysis was used to determine the prognostic gene of DEGs. $P < 0.05$ gene was considered as significant. Then, Lasso penalized Cox regression analysis was used to further select OS-related prognostic genes in UCEC. Finally, UCEC patients in the TCGA database were randomly divided into training sets and validation sets with equal numbers, and univariate Cox analysis was used to regress to establish a risk scoring model gradually. The C-index, AUC value of the model, is calculated by the validation set and the verification of the Nomogram to test the accuracy of the risk score model.

Analysis of intratumoral immune cell composition

CIBERSORT deconvolution algorithm (<https://cibersort.stanford.edu/>) was used to estimate the abundance of 22 immune cell types in UCEC and to evaluate the intratumoral immune cell composition.

Cell culture

Three different endometrial cancer cell lines and 293t were purchased from ATCC (ATCC, USA). The highly differentiated endometrial cancer cell line ISHIKAWA and the moderately differentiated endometrial cancer cell line SPEC-2 were cultured with MEM medium (GIBICO, USA) containing 10% fetal bovine serum (BI, Australia), The poorly differentiated endometrial cancer cell line KLE was cultured in DMEM/F12 (GIBICO, USA) mixed medium containing 10% fetal bovine serum. The 293T cell line and endometrial epithelial cell line hEEC were cultured in DMEM (GIBICO, USA) mixed medium containing 10% fetal bovine serum. All cells are cultured in a 37° incubator containing 5% CO₂.

CCK-8 assay

Inoculate the cell suspension in a 96-well plate and continue to culture for 12h, 24h, 48h, 72h, and then use the Cell Counting Kit 8 (CCK-8) Kit (Dojindo, Japan) to measure cell proliferation ability. The absorbance values at OD 490 nm were measured using a plate reader (Biorad, USA).

Colony formation assay

The cells were seeded in a 6-well plate, and when they were cultured normally to form visible clones, the cells were fixed with 4% formaldehyde, stained with Kimsa's solution (Meilunbio, China) and photographed for analysis.

Xenograft tumor in nude mice

Nude mice were randomly divided into Control group and ALDH2 overexpression group. Nude mice in each group were subcutaneously injected with 0.2 ml of UCEC cell suspension (1×10^7 cells/ml) on the ventral side of the right hind limb. The mice were fed with a normal diet under SPF conditions, and the subcutaneous tumor formation of nude mice was observed. On the 21th day, the nude mice were sacrificed by cervical dislocation, the subcutaneous tumors were completely removed, and weighed with an electronic balance. The Wuhan University of Science and Technology Animal Ethics and Use Committee approved the tumor-forming experiment in nude mice.

Dual-luciferase reporter assay

Luc-3'-UTR of ALDH2 and mutant form were separately subcloned into a pmirGLO (Addgene, USA) vector to establish wt-ALDH2-luc and mut-ALDH2-luc respectively. The miRNA mimic, internal control and Luc plasmid were co-transfected into cells. Then, luciferase activities were detected 48 h after transfection by the Dual-Luciferase Reporter Assay System (Promega, USA).

RNA Pull-Down

RNA pull-down assay was carried out using Magnetic RNA-Protein Pull-Down Kit (Thermo, USA). The RNA-bound beads were added in the cell nuclear lysate. Then, the eluted proteins were detected by western blot analysis.

RNA Immunoprecipitation (RIP)

RNA-protein-antibody complexes were captured using Protein A/G (Thermo, USA). RNA was eluted by adding TRIzol directly to magnetic beads and isolated as per the manufacturer's instructions. cDNA was

synthesized using HiScript® II 1st Strand cDNA Synthesis Kit (Vazyme, China) and analyzed by qRT-PCR.

RNA isolation and qRT-PCR

RNeasy Plus Universal Mini Kit (QIAGEN, USA) was used to isolate total RNA from cell lines, and HiScript® II 1st Strand cDNA Synthesis Kit (Vazyme, China) was used to reverse transcribe into cRNA. At the same time, the miRNeasy Micro Kit (QIAGEN, USA) was used to isolate miRNA from cell lines, and the miRNA 1st Strand cDNA Synthesis Kit (Vazyme, China) was used to reverse transcription of miRNA. For the extracted RNA and miRNA, qRT-PCR was performed on the Bio-Rad CFX-96 (Biorad, USA) system using the SYBR Green (Yisen, USA) method to determine the relative expression level. β -actin and U6 were used As an endogenous control. The primer sequences of ALDH2 are as follows: ALDH2-F: 5-ATGGCAAGCCCTATGTCATCT-3, ALDH2-R: 5-CCGTGGTACTTATCAGCCCA-3. The synthesis of primers, plasmid sequencing, miRNA reverse transcription sequences, primers and probes are all designed by Ribobio (Ribobio, China).

Western blot

Western blot assay was performed according to the standard protocol. The ALDH2 antibody was purchased from CST (CST, USA), and the antibody was diluted according to the recommended ratio in the instructions

Statistical Analyses

Comparisons between groups and normality were performed using R 3.6.3 software. Comparisons were completed using a one-way analysis of variance (ANOVA), two-tailed Student's t-test, non-parametric tests. Kaplan–Meier analyses with log-rank tests and the Cox proportional hazard model were used to analysis for survival. All data are presented as the mean $\pm$ SD. Statistical significance was set at $p < 0.05$.

Results

1. The expression of ALDH family in all cancers of TCGA.

Through the TCGA database, the expression data of all ALDH families in all cancers are obtained. As shown in Figure 1, the expression genes of the ALDH family genes in cancer are very different (Figure 1.A), and the expressions of the ALDH family genes are highly correlated (Figure 1.B). In detail, ALDH9H1, ALDH18A1, ALDH3A2, ALDH1A1, ALDH1B1 and ALDH2 are expressed higher than other ALDH family genes in all cancers. The expression of ALDH2 and ALDH8A1 is highly positively correlated. Further analysis of the first six ALDH family genes that are highly expressed in cancer. According to the classification of normal tissues and tumor tissues, it was found that the expression of ALDH3B1, ALDH18A1, etc., increased in a variety of cancers, but ALDH9A1 and ALDH2 showed the opposite trend (Figure 1.C) in UCEC. Except for ALDH3B1, which was statistically significant, the expression of the other 5 ALDH family members decreased in tumor tissues. These results show that the expression of ALDH family genes changes in a variety of cancers, and the changes in their expression may be involved in the regulation of tumors.

2. The prognostic significance of ALDH family.

To further confirm the effect of ALDH family genes that are highly expressed in cancer on UCEC patients. We retrieved the survival information of each patient in the TCGA database, combined with the expression analysis of ALDH family genes. Changes in other ALDH family genes do not affect the overall survival of UCEC patients. Only changes in the expression of ALDH2 gene will significantly affect the overall survival of UCEC patients (Figure 2.A). The results showed that the overall survival of the low expression group of ALDH2 and ALDH18A1 was low. At the same time, the survival of DSS, DFI, and PFI of the low expression group of ALDH2 was the same as the overall survival result, all showing a worse survival rate (Figure 2.B). For DSS in the ALDH18A1 low expression group, DFI and PFI were not different from the high expression group (Figure 2.C). These results indicate that patients in the ALDH2 low expression group are associated with worse prognostic survival in different survival groups. In addition to survival time and survival status, TCGA data also contains complete clinical information of the patient. The results of the clinical analysis found that the expression of ALDH2 was not significantly correlated with the age and race of the patients (Figure 2.D).

3. Analysis of GO, KEGG, GSEA function enrichment.

In this study, we downloaded the expression data of all UCEC samples in the TCGA database. The deletion was divided into 2 groups according to the median value and the expression of ALDH2, and the different genes between the groups were analyzed. A total of 477 differential genes were identified. Among them, 322 genes were up-regulated, and 155 genes were down-regulated (Figure 3.A, B). In order to understand the possible regulatory functions of these differential genes. The six pathways with the highest correlation found in GO enrichment are (Figure 3.C), humoral immune response mediated by circulating immunoglobulin, complement activation, classical pathway, complement activation, immunoglobulin mediated immune response, B cell mediated immunity, and protein activation cascade. The 6 pathways with the highest correlation of KEGG enrichment are (Figure 3.D), Cell adhesion molecules, Type I diabetes mellitus, Intestinal immune network for IgA production, Hematopoietic cell lineage, Th1 and Th2 cell differentiation, and Phagosome. GSEA enrichment found that the GO processes of the main ALDH2 up-regulated group included AZUROPHIL GRANULE LUMEN, RESPONSE TO INTERFERON GAMMA, MYELOID LEUKOCYTE MIGRATION, IMMUNE RECEPTOR ACTIVITY, TETRAPYRROLE METABOLIC PROCESS (Figure 3.E). The GO processes of the down-regulation group include MITOCHONDRIAL GENOME MAINTENANCE, REGULATION OF DNA RECOMBINATION, CELL CYCLE CHECKPOINT, DNA REPLICATION CHECKPOINT, SULFUR AMINO ACID BIOSYNTHETIC PROCESS (Figure 3.F). The above enrichment results indicate that the alteration of ALDH2 expression may be related to cellular immunity. Therefore, the immune infiltration score of each UCEC patient was calculated by R software with cibersort package. The results showed that patients in the ALDH2 low expression group had lower scores for CD8⁺ T cells and plasma cells (Figure 3.G).

4. PPI network construction.

PPI of DEGs was constructed by using the String network tool. After hiding all the individual nodes, it is found that there are 392 nodes and 295 edges (Figure 4.A). Statistics found that the first six genes with the number of interaction relationships are ORM2, C3, CTSH, HLA-DPA1, HLA-DPB1, HLA-DQA1 (Figure 4.B). At the same time, import the results into Cytospace software. Use MCODE and Cytohubba apps to calculate the core subnet (Figure 4.C) and core gene (Figure 4.D) in the PPI network. The result is shown in the figure.

5. The risk score model of UCEC construction.

Our data show that the high expression ALDH2 group has a better OS, DSS, and PFI (Figure 2). Therefore, a prognostic model based on the overall survival score of ALDH2 expression was constructed. The expression profile of UCEC in the TCGA database was randomly divided into an equal number of training groups and test groups. COX univariate results showed that a total of 62 genes were significantly related to the overall survival of UCEC patients. Further use lasso regression to screen out 15 genes to prevent the model from overfitting (Figure 5.A, B). Finally, a 6-factor overall prognostic risk score model was established using COX stepwise regression. The coef of risk score model of overall survival was:

$CDKN2A \times 0.18793 + WNT10A \times 0.23536 + AQP5 \times -0.21242 + STX18 \times -0.30617 + GZMA \times -0.44271 + LCN2 \times -0.15148$. COX multifactorial analysis found that they were all highly significantly related to prognosis (Figure 5.C).

Next, use the test group to test the model. The risk scores of 268 UCEC patients in the verification group are shown in the figure (Figure 5.D). The statechart shows that the number of deaths in the high-risk group is more (Figure 5.E). Moreover, the survival prognosis of the high-risk group is worse (Figure 5.F). Besides, the C-index of the model is calculated to be 0.798. The AUC value of 1 and 3 years is 0.738, and the AUC value of 5 years is 0.764 (Figure 5.G). These results show that the model shows that the model has a better ability to predict. Therefore, we draw a nomogram (Figure 5.H) according to the risk score model, and the verification result is shown in the Figure 5.I.

6. Effect Of ALDH2 Overexpression On Tumor Progression In Vitro And In Vivo.

To determine whether ALDH2 expression is reduced in endometrial cancer cell lines. It was detected by Western Blot and qRT-PCR. Compared with normal human endometrial epithelial cells hEEC, the expression of ALDH2 is reduced in ISHIKAWA, SPEC-2, and KLE (Figure 6.A, B). At the same time, we constructed a lentiviral plasmid pLKO.1-ALDH2 that overexpressed ALDH2. The virus was collected after transfection of 293T cells. After being concentrated by PEG8000, it infects ISHIKAWA and KLE cell lines. The overexpression efficiency was verified by Western Blot and qRT-PCR (Figure 6.C, D). Next, the relationship between the expression of ALDH2 and the proliferation ability of endometrial cancer was verified by CCK-8 and colony formation analysis (Figure 6.E, F). The results showed that restoring the overexpression of ALDH2 reduces the proliferation ability of endometrial cancer cell lines. In addition, the tumor xenograft model is used to determine whether overexpression of ALDH2 has the same function. The results showed that the tumor

mass and volume of mice in the overexpression group ALDH2 group were significantly smaller than those of the control group (Figure 6.G).

7. Hsa-miR-135b-3p binds to the 3UTR of.

The change of ALDH2 plays a crucial role in the survival and immune infiltration of UCEC patients. Therefore, there is an urgent need to find a way to regulate ALDH2. Analysis of gene mutation frequency found that in most patients, it is not the mutation that caused the effect of ALDH2 to disappear (Figure 7.A, B).

Numerous studies have found that miRNA can silence gene expression by targeting gene 3UTR. We speculate that ALDH2 may also be regulated by this mechanism. Screen the difference miRNA between UCEC patients and the normal group. The volcano plot showed that a total of 122 up-regulated miRNAs were screened (Figure 7. C). At the same time, visit the bioinformatics prediction website TargetScan. As a result, 312 miRNAs were predicted to bind to ALDH2's 3UTR. A total of 10 miRNAs were selected from the intersection, and the correlation between their expression and ALDH2 expression is shown in the figure (Figure 7D, E). The top three miRNAs with the highest negative correlation are hsa-miR-301b-5p, hsa-miR-3187-3p and hsa-miR-135b-3p.

To verify which miRNA regulates the expression of ALDH2, we constructed the Luciferase plasmid of ALDH2 3'UTR. The dual fluorescein report experiment results showed that only the hsa-miR-135b-3p transfection group had a decrease in relative fluorescence intensity. However, there was no significant difference between hsa-miR-301b-5p and hsa-miR-3187-3p after transfection (Figure 7F). The mutation group showed that the mutation could eliminate the inhibitory effect of hsa-miR-135b-3p on ALDH2 (Figure 7G). At the same time, RIP and RNA pull-down found that hsa-miR-135b-3p can enrich the expression of ALDH2, but the control group did not have this phenomenon (Figure 7H, I). qRT-PCR showed that after transfection of hsa-miR-135-3p mimic, the expression of hsa-miR-135b-3p increased significantly (Figure 7J), while the results of WB and qRT-PCR showed that the expression of ALDH2 decreased (Figure 7K, L). Besides, TCGA survival data showed that the high expression group of hsa-miR-135b-3p showed a lower prognosis (Figure 7M). These results all indicate that hsa-miR-135b-3p can down-regulate the expression of ALDH2.

Discussion

Endometrial cancer is one of the three most common malignant tumors of the female reproductive tract, and it is the sixth most common cancer in women worldwide (2). With the increase in the average life expectancy of the population and the change in living habits, the incidence of EC has continued to rise and become younger in the past two decades (21). Although the cause of endometrial cancer is not very clear so far, it may be due to genetics, obesity, or the use of drugs (22). However, there are clear results showing that the cure rate of early UCEC can reach 95% (23)

Therefore, the development of prognostic markers will help UCEC patients to find the disease earlier and develop treatment earlier. However, due to traditional experimental methods, the time period is long, and the

research and development costs are high. The emergence of bioinformatics has brought us the possibility of studying new prognostic markers. A number of research results have shown that bioinformatics is a reliable research tool. In gastric cancer, GDF-15 can be used as a biomarker for gastric cancer (24). In liver cancer, YTHDF1 expression is elevated in liver cancer patients (25). In UCEC, Li found that Mammaglobin B is a prognostic marker of UCEC (26). Similarly, we downloaded the expression data of 33 cancers in TCGA. Extracting ALDH family genes found that the expression of ALDH family has changed significantly in most cancers. The ALDH family was first discovered to be mainly involved in the process of alcohol metabolism (27). However, recent studies have found that ALDH family genes can participate in the regulation of cancer. In detail, ALDH1 has been identified as a tumor stem cell marker involved in the development of cancer (28). And ALDH2, Li found that restoring the expression of ALDH2 in lung adenocarcinoma can inhibit the migration of lung cancer cells (29). This is consistent with our findings. We found that ALDH2 expression in most cancers not only decreases in LUAD but also in UCEC. Also expressed decline. The characteristic information found that the change of ALDH2 was not significantly correlated with age group and race. Combined with the clinical information of UCEC patients in the TCGA database, it was found that in UCEC, the OS, DSS, DFI and PFI of the low expression group of ALDH2 had worse survival conditions. Bioinformatics enrichment function enrichment found that the changes of ALDH2 participate in cancer mainly through the regulation of immune function. The previous discussion showed that ALDH1 had been identified as a tumor cell marker (28). Christophe found that ALDH1 cell subsets have higher cancer stem cell characteristics (30). Mohamed identified that the expression of ALDH1 is highly correlated with the expression of Colorectal Carcinoma's tumor stem cell markers Notch1 and CD44 (31). And we calculated the characteristic value of immune infiltration of UCEC patients through CIBERSORT and found that the change of ALDH2 is highly correlated with CD8 + T cell. However, there are no reports about the changes of ALDH2 and the research of cancer stem cells and immune-related research. Our findings provide new ideas for further research on the regulation of ALDH2 on cancer. In addition, we initially explored the relationship between the expression of ALDH2 and the progress of UCEC in vivo and in vitro. It was found that the expression of ALDH2 was reduced in the three UCEC cell lines ISHIKAWA, SPEC-2, and KLE. At the same time, restoring the expression level of ALDH2 in the cell line can reduce the proliferation ability of tumor cells. And in vivo experiments have also verified our findings.

Gene mutation is one of the main factors of gene inactivation. Through the Cbioportal database, it was found that the changes in ALDH2 were not due to genetic mutations. In recent years, a new way of regulating genes non-coding RNA regulating genes has been widely verified. Non-coding RNA is a type of RNA that cannot encode the protein. It has been considered a "junk" transcription product for a long time. However, research in recent years has changed people's understanding of ncRNA, and more and more studies have found that ncRNA. It is a class of functional regulatory molecules that can regulate a series of cellular processes, including chromosome remodeling, transcription, post-transcriptional modification, and signal transduction, and plays a crucial role in developmental and disease processes (32). NcRNA mainly includes micro miRNA, lncRNA, and circRNA, which coordinate their parts and play an essential role in the complex function regulation network of cells. MiRNA is a kind of endogenous non-coding RNA, which is generally composed of 22 nucleotides. It regulates gene expression by binding to the 3'UTR of target mRNA (33). In 2002, Calin et al. confirmed the presence of miRNA in chronic leukemia cells. The expression of -15

and miR-16 is reduced or even missing. This is the earliest direct evidence that the abnormal expression of miRNA is related to tumorigenesis (34). Subsequently, more and more tumor miRNA chips have confirmed that miRNA can indeed regulate proto-oncogenes. Or the expression of tumor suppressor genes, which play an important role in the occurrence and metastasis of a variety of malignant tumors (35–37). By taking the intersection of the UCEC data, it was found that the expression of miR-135-3p in UCEC increased, which was negatively correlated with the expression of ALDH2. Therefore, through dual-luciferase reporter gene experiments, RIP, RNA pull-down, it was found that miR-135-3p can target ALDH2. At the same time, the expression of miR-135 was up-regulated in NSCLC cells. Silencing miR-135 can inhibit cell viability, migration, and invasion (38). In lung cancer, overexpression of ALDH2 inhibits the malignant characteristics of lung adenocarcinoma cells, such as proliferation, stemness, and migration, while knocking down ALDH2 increases these characteristics (29). These results verify this regulatory relationship from the side.

We found earlier that the low expression group of ALDH2 had a worse survival status. Therefore, we developed a risk score model based on ALDH2 expression to predict the survival of UCEC patients. In order to construct a risk scoring model, we screened 478 differentially expressed mRNAs from 538 originals in the UCEC dataset in the TCGA database according to the expression of ALDH2. According to COX regression and Lasso Cox regression model analysis, screening and constructing a prognostic risk scoring model. The risk scoring model can divide UCEC patients into high-risk and low-risk groups to predict their prognosis. In addition, whether it is through the time-dependent ROC curve or the C-index of the calculation model, the risk scoring model we build has high predictive sensitivity and performance.

The TCGA database provides complete data related to cancer, including gene editing, mutation information, clinical survival information, etc. In this study, by using the information provided by the TCGA database, we found that most of the ALDH family genes were abnormally expressed in 33 cancers. Among them, in UCEC, the expression of ALDH2 decreased, and the low expression group had a worse prognosis. In addition, we preliminarily discussed how the changes in ALDH2 are involved in the regulation of cancer and preliminarily verified that miR-301a-5p targets the 3UTR of ALDH2. Finally, a risk score model is constructed. The risk scoring model has high performance in terms of specificity and accuracy. It provides new ideas and directions for further exploring the mechanism and treatment strategies of UCEC. However, current research still has some limitations.

Conclusion

This study analyzed the expression of ALDH2 and its family genes in 33 cancers. And in UCEC, the expression of ALDH2 and the role of survival were deeply explored. It was found that the expression of ALDH2 was reduced in UCEC patients, and the survival status of the low expression group was worse. At the same time, bioinformatics and experimental methods are used to verify the function of ALDH2 in UCEC patients. The results showed that restoring the expression of ALDH2 can inhibit tumor progression, and the expression of ALDH2 is regulated by miR-135-3p. In addition, we propose a survival model based on the expression of ALDH2, whose C-index is 0.798 and AUC is 0.764.

Abbreviations

MKL-1: Megakaryoblastic leukemia 1

SRF: serum response factor

FITC: fluorescein isothiocyanate

DAPI: 4',6-diamidino-2-phenylindole

qRT-PCR: real-time PCR

WB: western blot

NC: negative control

Declarations

Ethics approval and consent to participate

All mouse experimental procedures and procedures were evaluated and authorized in strict accordance with the guiding principles of the Animal Protection and Use Committee of Wuhan University of Science and Technology and in accordance with the "Hubei Province Experimental Animal Management Regulations."

TCGA data is publicly available.

Consent for publication

All authors have read this manuscript and approved for submission.

Availability of data and material

The data generated during this study are included in this article and its supplementary information files are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was financially supported by the National Natural Science Foundation of China (No. 31501149, 31770815, 31570764) and the Hubei Natural Science Foundation (2017CFB537, 2019CFB529).

Author contributions

XHL designed research; XHL, YX, and ZTD performed research., ZTD analyzed data; and XHL wrote the paper. ZTD participated in the revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA: a cancer journal for clinicians* (2015) 65(2):87-108. doi: 10.3322/caac.21262. PubMed PMID: 25651787.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* (2018) 68(6):394-424. doi: 10.3322/caac.21492. PubMed PMID: 30207593.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA: a cancer journal for clinicians* (2015) 65(1):5-29. doi: 10.3322/caac.21254. PubMed PMID: 25559415.
4. Willoughby BJ, Faulkner K, Stamp EC, Whitaker CJ. A descriptive study of the decline in cervical screening coverage rates in the North East and Yorkshire and the Humber regions of the UK from 1995 to 2005. *Journal of public health* (2006) 28(4):355-60. doi: 10.1093/pubmed/fdl062. PubMed PMID: 17062818.
5. Quinn M, Babb P, Jones J, Allen E. Effect of screening on incidence of and mortality from cancer of cervix in England: evaluation based on routinely collected statistics. *Bmj* (1999) 318(7188):904-8. doi: 10.1136/bmj.318.7188.904. PubMed PMID: 10102852; PubMed Central PMCID: PMC27810.
6. Barnholtz-Sloan JS, Schwartz AG, Qureshi F, Jacques S, Malone J, Munkarah AR. Ovarian cancer: changes in patterns at diagnosis and relative survival over the last three decades. *American journal of obstetrics and gynecology* (2003) 189(4):1120-7. doi: 10.1067/s0002-9378(03)00579-9. PubMed PMID: 14586365.
7. Zhang C, Zhang G, Sun N, Zhang Z, Zhang Z, Luo Y, et al. Comprehensive molecular analyses of a TNF family-based signature with regard to prognosis, immune features, and biomarkers for immunotherapy in lung adenocarcinoma. *EBioMedicine* (2020) 59:102959. doi: 10.1016/j.ebiom.2020.102959. PubMed PMID: 32853987; PubMed Central PMCID: PMC7452643.
8. Sun H, Liu SY, Zhou JY, Xu JT, Zhang HK, Yan HH, et al. Specific TP53 subtype as biomarker for immune checkpoint inhibitors in lung adenocarcinoma. *EBioMedicine* (2020) 60:102990. doi: 10.1016/j.ebiom.2020.102990. PubMed PMID: 32927274; PubMed Central PMCID: PMC7494676.
9. Li B, Zhang B, Wang X, Zeng Z, Huang Z, Zhang L, et al. Expression signature, prognosis value, and immune characteristics of Siglec-15 identified by pan-cancer analysis. *Oncoimmunology* (2020) 9(1):1807291. doi: 10.1080/2162402X.2020.1807291. PubMed PMID: 32939323; PubMed Central PMCID: PMC7480813.
10. Wang Y, Ren F, Chen P, Liu S, Song Z, Ma X. Identification of a six-gene signature with prognostic value for patients with endometrial carcinoma. *Cancer medicine* (2018) 7(11):5632-42. doi: 10.1002/cam4.1806. PubMed PMID: 30306731; PubMed Central PMCID: PMC6247034.
11. Chi YC, Lee SL, Lee YP, Lai CL, Yin SJ. Modeling of Human Hepatic and Gastrointestinal Ethanol Metabolism with Kinetic-Mechanism-Based Full-Rate Equations of the Component Alcohol

- Dehydrogenase Isozymes and Allozymes. *Chemical research in toxicology* (2018) 31(7):556-69. doi: 10.1021/acs.chemrestox.8b00003. PubMed PMID: 29847918.
12. Lim RT, Jr., Gentry RT, Ito D, Yokoyama H, Baraona E, Lieber CS. First-pass metabolism of ethanol is predominantly gastric. *Alcoholism, clinical and experimental research* (1993) 17(6):1337-44. doi: 10.1111/j.1530-0277.1993.tb05250.x. PubMed PMID: 8116851.
 13. Druesne-Pecollo N, Tehard B, Mallet Y, Gerber M, Norat T, Hercberg S, et al. Alcohol and genetic polymorphisms: effect on risk of alcohol-related cancer. *The Lancet Oncology* (2009) 10(2):173-80. doi: 10.1016/S1470-2045(09)70019-1. PubMed PMID: 19185835.
 14. Orywal K, Szmitkowski M. Alcohol dehydrogenase and aldehyde dehydrogenase in malignant neoplasms. *Clinical and experimental medicine* (2017) 17(2):131-9. doi: 10.1007/s10238-016-0408-3. PubMed PMID: 26886278; PubMed Central PMCID: PMC5403859.
 15. Thomas ML, de Antueno R, Coyle KM, Sultan M, Cruickshank BM, Giacomantonio MA, et al. Citral reduces breast tumor growth by inhibiting the cancer stem cell marker ALDH1A3. *Molecular oncology* (2016) 10(9):1485-96. doi: 10.1016/j.molonc.2016.08.004. PubMed PMID: 27592281; PubMed Central PMCID: PMC5423215.
 16. Nastaly P, Filipiska M, Morrissey C, Eltze E, Semjonow A, Brandt B, et al. ALDH1-positive intratumoral stromal cells indicate differentiated epithelial-like phenotype and good prognosis in prostate cancer. *Translational research : the journal of laboratory and clinical medicine* (2019) 203:49-56. doi: 10.1016/j.trsl.2018.08.007. PubMed PMID: 30287243.
 17. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* (2012) 16(5):284-7. doi: 10.1089/omi.2011.0118. PubMed PMID: 22455463; PubMed Central PMCID: PMC3339379.
 18. Yu G, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* (2015) 31(4):608-9. doi: 10.1093/bioinformatics/btu684. PubMed PMID: 25677125.
 19. Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular bioSystems* (2016) 12(2):477-9. doi: 10.1039/c5mb00663e. PubMed PMID: 26661513.
 20. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* (2015) 43(Database issue):D447-52. doi: 10.1093/nar/gku1003. PubMed PMID: 25352553; PubMed Central PMCID: PMC4383874.
 21. Sheikh MA, Althouse AD, Freese KE, Soisson S, Edwards RP, Welburn S, et al. USA endometrial cancer projections to 2030: should we be concerned? *Future oncology* (2014) 10(16):2561-8. doi: 10.2217/fon.14.192. PubMed PMID: 25531045.
 22. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E. Endometrial cancer. *Lancet* (2016) 387(10023):1094-108. doi: 10.1016/S0140-6736(15)00130-0. PubMed PMID: 26354523.
 23. Luckett R, Pena N, Vitonis A, Bernstein MR, Feldman S. Effect of patient navigator program on no-show rates at an academic referral colposcopy clinic. *Journal of women's health* (2015) 24(7):608-15. doi:

- 10.1089/jwh.2014.5111. PubMed PMID: 26173000.
24. Liu JY, Dong XX, Lu JN, Zhang Y, Liu KF, Liu LF, et al. Utility of GDF-15 as a diagnostic biomarker in gastric cancer: an investigation combining GEO, TCGA and meta-analysis. *FEBS open bio* (2019) 9(1):35-42. doi: 10.1002/2211-5463.12537. PubMed PMID: 30652072; PubMed Central PMCID: PMC6325603.
25. Zhao X, Chen Y, Mao Q, Jiang X, Jiang W, Chen J, et al. Overexpression of YTHDF1 is associated with poor prognosis in patients with hepatocellular carcinoma. *Cancer biomarkers : section A of Disease markers* (2018) 21(4):859-68. doi: 10.3233/CBM-170791. PubMed PMID: 29439311.
26. Li J, Xu W, Zhu Y. Mammaglobin B may be a prognostic biomarker of uterine corpus endometrial cancer. *Oncology letters* (2020) 20(5):255. doi: 10.3892/ol.2020.12118. PubMed PMID: 32994818; PubMed Central PMCID: PMC7509766.
27. Gross ER, Zambelli VO, Small BA, Ferreira JC, Chen CH, Mochly-Rosen D. A personalized medicine approach for Asian Americans with the aldehyde dehydrogenase 2*2 variant. *Annual review of pharmacology and toxicology* (2015) 55:107-27. doi: 10.1146/annurev-pharmtox-010814-124915. PubMed PMID: 25292432; PubMed Central PMCID: PMC4435945.
28. Li W, Ma H, Zhang J, Zhu L, Wang C, Yang Y. Unraveling the roles of CD44/CD24 and ALDH1 as cancer stem cell markers in tumorigenesis and metastasis. *Scientific reports* (2017) 7(1):13856. doi: 10.1038/s41598-017-14364-2. PubMed PMID: 29062075; PubMed Central PMCID: PMC5653849.
29. Li K, Guo W, Li Z, Wang Y, Sun B, Xu D, et al. ALDH2 Repression Promotes Lung Tumor Progression via Accumulated Acetaldehyde and DNA Damage. *Neoplasia* (2019) 21(6):602-14. doi: 10.1016/j.neo.2019.03.008. PubMed PMID: 31071657; PubMed Central PMCID: PMC6506700.
30. Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, Dutcher J, Brown M, et al. ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell stem cell* (2007) 1(5):555-67. doi: 10.1016/j.stem.2007.08.014. PubMed PMID: 18371393; PubMed Central PMCID: PMC2423808.
31. Mohamed SY, Kaf RM, Ahmed MM, Elwan A, Ashour HR, Ibrahim A. The Prognostic Value of Cancer Stem Cell Markers (Notch1, ALDH1, and CD44) in Primary Colorectal Carcinoma. *Journal of gastrointestinal cancer* (2019) 50(4):824-37. doi: 10.1007/s12029-018-0156-6. PubMed PMID: 30136202.
32. Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. *Nature reviews Cancer* (2018) 18(1):5-18. doi: 10.1038/nrc.2017.99. PubMed PMID: 29170536; PubMed Central PMCID: PMC6337726.
33. Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. *Cell* (2012) 148(6):1172-87. doi: 10.1016/j.cell.2012.02.005. PubMed PMID: 22424228; PubMed Central PMCID: PMC3308137.
34. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, et al. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America* (2002) 99(24):15524-9. doi: 10.1073/pnas.242606799. PubMed PMID: 12434020; PubMed Central PMCID: PMC137750.

- # Figures

Page 16/23

expression levels of ALDH1A1, ALDH1B1, ALDH2, ALDH9H1, ALDH3B1, and ALDH18A1 in the TCGA database compared with normal tissues. (The number of normal tissue samples less than 3 was hidden). * $p < 0.05$, *** $p < 0.001$.

Figure 2

The prognostic value of the ALDH family. A. Kaplan–Meier survival curves by OS. B. Kaplan–Meier survival curves by PFI, DSS, and DFI of ALDH2. C. Kaplan–Meier survival curves by PFI, DSS, and DFI of ALDH18A1. D. The relationship between clinicopathological characteristics and the expression of ALDH2.

Figure 3

Analysis of GO, KEGG, GSEA function enrichment. A. Heatmap showing differential gene expression (FDR < 0.05) between low expression group and high expression group of ALDH2. B. Volcano plot of all differential gene expression analysis. C. GO enrichment analysis. D. KEGG enrichment analysis. E. GSEA enrichment analysis of high expression group of ALDH2. F. GSEA enrichment analysis of low expression group of ALDH2. G. Enrichment scores for 22 immune cell subpopulations based on deconvolution by CIBERSORT between low expression group and high expression group of ALDH2. * $p < 0.05$, *** $p < 0.001$.

Figure 4

PPI network construction. A. PPI network of all DEGs. The individual nodes were hidden. The interaction relationship prediction threshold is ≥ 0.900 . B. Top 30 genes with the highest number of nodes. C. The core subnet of the PPI network by using the MCODE app. D. The core gene of the PPI network by using the Cytohubba app.

Figure 5

The risk score model of UCEC construction. A. DEGs were identified by the LASSO logistic regression model with non-zero coefficients. B. The graph of the relationship between likelihood deviation and log (Lambda). The vertical dashed lines indicate the λ value and the maximum λ value with the smallest error. C. Forest plot of the hazard ratio for OS of parameters. D. Predicted risk of overall survival by the risk score model. E. Scatter plot of UCEC death predicted by risk score model. F. The survival curve of the high expression group and the low expression progenitor in the risk score model. G. ROC curve of risk score model. H. Predicts the OS of patients with the Nomogram. I. Calibration curve of the nomogram in the training dataset.

Figure 6

Effect Of ALDH2 Overexpression On Tumor Progression In Vitro And In Vivo. A. Expression of ALDH2 in the UCEC cells by Western Blot. B. Expression of ALDH2 in the UCEC cells by qRT-PCR. C. Overexpression efficiency of ALDH2 by Western Blot. D. Overexpression efficiency of ALDH2 by qRT-PCR. E. CCK-8 assays OD 450nm. F. Colony formation ability was determined using colony formation assays. G. Tumor xenografts from nude mice subsequent assays. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 7

Hsa-miR-135b-3p binds to the 3'UTR of ALDH2. A. Gene mutation distribution of UCEC patients detected in TCGA. B. KM survival curve of the mutant group and non-mutation group. C. The volcano plot of miRNAs in UCEC patients with TCGA: Red dots and green dots indicate differentially expressed miRNAs based on the fold change $p < 0.05$. D. Venn diagram of up-regulated miRNA and predicted binding miRNA. E. Correlation analysis between the expression of predicted bound miRNA and the expression of ALDH2. F. Relative Luciferase activity was measured with a dual-luciferase reporter assay. G. Dual-luciferase assay of the

mutation group. H. RNA pull-down analysis with ALDH2 antibody. I. RIP assay was further verified for a direct association between hsa-miR-135-3p and ALDH2. J. Overexpression efficiency of hsa-miR-135-3p mimic. K. WB analysis the expression of ALDH2 with overexpression of hsa-miR-135-3p. L. qRT-PCR analysis relative expression of ALDH2 with overexpression of hsa-miR-135-3p. M. Kaplan-Meier survival curve for patients with UCEC (According to the median value of aldh2 expression, it is divided into two groups of high and low expression). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.