WOMEN IN SCIENCE: GENETICS

EDITED BY: Rana Dajani, Zodwa Dlamini, Aparna Vasanthakumar, Deepika Polineni, Silvia Calo, Jaira Ferreira de Vasconcellos, Malak Abedalthagafi, Bertha Hidalgo and Carine Le Goff

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-526-7
DOI 10.3389/978-2-83250-526-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

WOMEN IN SCIENCE: GENETICS

Topic Editors:
Rana Dajani, Hashemite University, Jordan
Zodwa Dlamini, Pan African Cancer Research Institute (PACRI), South Africa
Aparna Vasanthakumar, AbbVie, United States
Deepika Polineni, University of Kansas Medical Center, United States
Silvia Calo, Pontificia Universidad Católica Madre y Maestra, Dominican Republic
Jaira Ferreira de Vasconcellos, James Madison University, United States
Malak Abedalthagafi, Emory University, United States
Bertha Hidalgo, University of Alabama at Birmingham, United States
Carine Le Goff, Institut National de la Santé et de la Recherche Médicale (INSERM), France

Citation: Dajani, R., Dlamini, Z., Vasanthakumar, A., Polineni, D., Calo, S., de Vasconcellos, J. F., Abedalthagafi, M., Hidalgo, B., Le Goff, C., eds. (2022). Women in Science: Genetics. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-526-7

Table of Contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 03 October 2022
DOI 10.3389/fgene.2022.1038317

OPEN ACCESS

EDITED AND REVIEWED BY
Dov Greenbaum,
Yale University, United States

*CORRESPONDENCE
Jaira Ferreira de Vasconcellos,
vasconjf@jmu.edu

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

RECEIVED 06 September 2022
ACCEPTED 16 September 2022
PUBLISHED 03 October 2022

CITATION
de Vasconcellos JF, Abedalthagafi M, Calo S, Dajani R, Dlamini Z, Hidalgo B, Le Goff C and Vasanthakumar A (2022), Editorial: Women in science: Genetics. *Front. Genet.* 13:1038317. doi: 10.3389/fgene.2022.1038317

Editorial: Women in science: Genetics

Jaira Ferreira de Vasconcellos[1]*, Malak Abedalthagafi[2], Silvia Calo[3], Rana Dajani[4], Zodwa Dlamini[5], Bertha Hidalgo[6], Carine Le Goff[7] and Aparna Vasanthakumar[8]

[1]Department of Biology, James Madison University, Harrisonburg, VA, United States, [2]Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, United States, [3]Pontificia Universidad Católica Madre y Maestra, Santiago de los Caballeros, Dominican Republic, [4]Department of Biology and Biotechnology, Faculty of Science, The Hashemite University, Zarqa, Jordan, [5]SAMRC Precision Oncology Research Unit (PORU), DSI/NRF SARChI Chair in Precision Oncology and Cancer Prevention (POCP), Pan African Cancer Research Institute (PACRI), University of Pretoria, Hatfield, South Africa, [6]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, United States, [7]Université Paris Cité and Université Sorbonne Paris Nord, INSERM U1148, Laboratory of Vascular Translational Science, Bichat Hospital, Paris, France, [8]AbbVie, North Chicago, IL, United States

KEYWORDS

genetics, genomics, genome-wide association studies, genetic diversity, cancer, development

Editorial on the Research Topic
Women in science: Genetics

The reality of the workforce in Science, Technology, Engineering, and Mathematics (STEM) fields is still that women remain significantly underrepresented. Women obtain more than half of the undergraduate degrees in biology, chemistry, and mathematics in the United States and constitute about half of the American workforce. However, the American Community Survey from the United States Census Bureau has shown that women only held approximately 30% of STEM jobs in 2019 (Census, 2019). Of interest, the percentage of women pursuing STEM education is higher in the Middle East in comparison to the West (Study International, 2019), and according to the Organization for Economic Cooperation and Development women test better and feel more comfortable in mathematics than men in Jordan, Qatar, and the United Arab Emirates (Khazan, 2014). However, worldwide women continue to hold more junior positions in science, earn significantly less and do more unpaid work than their male colleagues, and are overall less likely to be supported during their higher education training (Author Anonymous, 2011; Khazan, 2014; Sommerfeld et al., 2017). Some overarching factors to help explain the larger gender gaps include masculine cultures that leave women with a low sense of belonging, the lack of sufficient early experiences in these fields, and gender gaps in self-efficacy (Cheryan et al., 2017). Moreover, more than 50% of women reported personal experiences with gender-related bias in a 2010 survey from AAAS/L'Oreal, compared with 2% of men who responded to the same survey (AAAS, 2010). And more recently, Chatterjee and Werner reported that original research articles written by women as primary authors had fewer citations than original research articles

written by men as primary authors and senior authors, especially when both primary and senior authors were women (Chatterjee and Werner, 2021), demonstrating the existence of gender-based differences in article citations that can directly impact professional trajectory and success.

To promote interdisciplinarity and worldwide representation from women in genetics research, Frontiers in Genetics launched the Research Topic *Women in Science: Genetics*. This Research Topic accepted 17 manuscripts including 15 original articles, 1 review, and 1 case report from a wide variety of research focuses and to this date has over 29,200 total views, and 5,600 article downloads. The breadth of content published includes cutting-edge research questions and techniques in genetics applied to a wide variety of research focuses. Among them, 4 great contributions were made to cancer research. While significant progress has been made in the research and development of novel therapeutic strategies and precision medicine approaches, a lot of variability in the response to treatments as well as overall survival and disease-free survival still exists. In addition, relapses and secondary malignancy are still largely resistant to the current treatment strategies and a better understanding of the tumors underlying molecular mechanisms is still needed. In this Research Topic, we show that Zhang et al. identified five crucial genes associated with breast cancer progression in both primary and metastatic cancer tissues, which may be novel potential targets for the treatment of breast cancer. Yu et al. stratified cervical cancer in two subtypes based on their 5-Methylcytidine RNA modifications and identified a novel gene expression signature that may be used for clinical risk assessment and/or targeted therapeutical strategies for cervical cancer. Li et al. developed a novel scoring model focusing on the interaction of immune infiltration and ferroptosis to predict the overall survival of ovarian cancer patients. Finally, Sun et al. focused on endometrial cancer and correlated a messenger RNA alternative splicing gene signature to the patient's prognosis and the immune-tumor microenvironment.

Additional genetics analyses were the focus of 6 contributions that used, for example, datasets from genome-wide association studies or bulk and single-cell RNA-seq to investigate different aspects of women's health, minority populations' health, or human development. He et al. used summary datasets from genome-wide association studies in a Mendelian randomization analysis to estimate the causal relationship between birth weight and female-related phenotypes and diseases. They demonstrated that birth weight may play a role in women's body mass index, menarche, decreased levels of adult sex hormone-binding globulin, and increased levels of bioavailable testosterone. Liu et al. used bulk RNA-seq ovary datasets and single-cell RNA-seq follicles datasets to determine molecular mechanisms underlying the ovarian reserve, and identified a gene expression signature that was highly correlated to the ovarian reserve of the primordial follicle pool. This gene signature has the potential to be used as clinical biomarkers for the prediction of women's ovarian reserve and in the development of future fertility targeted interventions. Yin et al. focused on competitive endogenous RNAs, including long non-coding RNAs, circular RNAs, and messenger RNAs, and their potential role in the pathogenesis of endometriosis. High-throughput sequencing demonstrated hundreds of long non-coding RNAs and circular RNAs differentially expressed as well as over one thousand messenger RNAs differentially expressed in the ectopic endometria group compared to normal and eutopic endometria groups. These results elucidate novel aspects of the underlying molecular mechanisms of endometriosis. Xie et al. investigated genetic polymorphisms associated with vitamin D anabolism and women's susceptibility for uterine leiomyomas, a common type of benign gynecological tumor. Remarkably, they found that the DHCR7 rs1044482 C > T polymorphism, a vitamin D anabolism-related gene, was associated as a risk factor for uterine leiomyomas. Moreover, Armstrong et al. performed a meta-analysis to investigate genetic contributors to stroke among African Americans with hypertension and identified 10 statistically significant and 90 suggestive variants associated with a stroke incident in this population. These are critical results that shed light on potential genetic determinants for stroke on hypertensive African Americans. And finally, Lin et al. investigated sex-specific differences in microRNA expression during human fetal lung development by microRNA sequencing and identified over 120 microRNAs that were expressed with a specific male or female pattern during human lung development.

The investigations of novel underlying molecular mechanisms continued to identify novel mutations, map different stages of development as well as investigate the regulatory system of a clinically-relevant microorganism. In a case report contribution, Zhang et al. identified novel mutations in antithrombin and coagulation factor II and their downstream clinically relevant reduced activity (protein deficiency) phenotypes. *In silico* analysis also demonstrated that these mutations may destroy the function and structure of the antithrombin and coagulation factor II proteins. With a focus on endometrium development, Zhang et al. used single-cell RNA sequencing to map the mice uteri from the regenerative endometrium stage to the maturational endometrium stage and investigate novel fundamental molecular mechanisms occurring in the transitional states of the endometrium during the estrus cycle. They demonstrated novel transcription factors associated with the differentiation path and that different stages of the estrus have a distinct cell composition. Furthermore, Cholo et al. investigated the role of the *Mycobacterium tuberculosis* KdpDE regulatory system alone and in association with the Trk K^+-uptake systems transporters in modulating *in vitro* bacterial growth.

The genetic diversity and population structure of plant populations were investigated by two contributions that used

microsatellite markers or genotyping by sequencing to identify single nucleotide polymorphisms, respectively. Brooks et al. investigated wild Aspalathus linearis, also known as rooibos, regarding their genetic diversity and population structure and demonstrated that wild Cederberg populations from the Western Cape are genetically distinct from the wild Northern Cape populations, which shows the critical need for appropriate conservation strategies that protect wild ecotypes. Omire et al. focus was on the investigation of genetic diversity and population structure of *Hyphaene compressa* or doum palm, a plant that grows in the Arid and Semi-Arid lands of Kenya. They demonstrated moderate genetic differences among the populations which may inform future breeding and conservation efforts. In addition, phenological and morphological traits to help support novel target genes for breeding efforts in plants were the focus of 1 contribution. Katoch et al. investigated Macrotyloma uniflorum (Lam.) Verdc, commonly known as horse Gram, for their quantitative trait loci associated with phenological and morphological traits. Their findings demonstrated four and seven quantitative trait loci for phenological traits and for morphological traits across different environments, respectively, a knowledge that may inform future breeding strategies.

Finally, Elbardisy and Abedalthagafi wrote a review article focusing on the history and challenges of women in genetics, where they highlighted the obstacles and contributions made by women in science with a particular focus on non-western women's contribution to the field of genetics.

We sincerely thank all our contributors for sharing their work on the Research Topic *Women in Science: Genetics* and all the reviewers and editorial staff involved in this endeavor for their time, expertise, and insights.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

Author AV was employed by the company AbbVie.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

AAAS (2010). Barriers for women scientists survey report. Sponsored by AAAS office of publishing and member services. Conducted exclusively for L'oréal by cell associates. Available at: https://www.aaas.org/sites/default/files/0928loreal_survey_report.pdf.

Author Anonymous (2011). Women in science. *Nat. Cell Biol.* 13, 489. doi:10.1038/ncb0511-489a

Census (2019). *STEM and STEM-related occupations by sex and median earnings: ACS*. United States Census Bureau. Available at: https://www.census.gov/data/tables/time-series/demo/income-poverty/stem-occ-sex-med-earnings.html.

Chatterjee, P., and Werner, R. M. (2021). Gender disparity in citations in high-impact journal articles. *JAMA Netw. Open* 4 (7), e2114509. doi:10.1001/jamanetworkopen.2021.14509

Cheryan, S., Ziegler, S. A., Montoya, A. K., and Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychol. Bull.* 143 (1), 1–35. doi:10.1037/bul0000052

Khazan, O. (2014). There are only 3 countries where girls feel more comfortable with math than boys. Available at: https://www.theatlantic.com/international/archive/2014/03/there-are-only-3-countries-where-girls-feel-more-comfortable-with-math-than-boys/284272/.

Sommerfeld, J., Manderson, L., Ramirez, B., Guth, J. A., and Reeder, J. C. (2017). Infectious disease research and the gender gap. *Glob. Health Epidemiol. Genom.* 2, e9. doi:10.1017/gheg.2017.2

Study International (2019). The rise of women in STEM in the Arab world. Available at: https://www.studyinternational.com/news/the-rise-of-women-in-stem-in-the-arab-world/.

frontiers in Genetics

ORIGINAL RESEARCH
published: 29 July 2021
doi: 10.3389/fgene.2021.698875

Role of the *kdpDE* Regulatory Operon of *Mycobacterium tuberculosis* in Modulating Bacterial Growth *in vitro*

Moloko C. Cholo†, Maborwa T. Matjokotja, Ayman G. Osman and Ronald Anderson†*

Department of Immunology, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

OPEN ACCESS

Edited by:
Silvia Calo, Pontificia Universidad Católica Madre y Maestra, Dominican Republic

Reviewed by:
Monde Ntwasa, University of South Africa, South Africa
Jose Cansado, University of Murcia, Spain

***Correspondence:**
Moloko C. Cholo
moloko.cholo@up.ac.za

†ORCID:
Moloko C. Cholo
orcid.org/0000-0002-0958-2401
Ronald Anderson
orcid.org/0000-0002-5925-6452

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *22 April 2021*
Accepted: *12 July 2021*
Published: *29 July 2021*

Citation:
Cholo MC, Matjokotja MT, Osman AG and Anderson R (2021) Role of the kdpDE Regulatory Operon of Mycobacterium tuberculosis in Modulating Bacterial Growth in vitro. Front. Genet. 12:698875. doi: 10.3389/fgene.2021.698875

Bacteria use K^+-uptake transporters differentially for adaptation in varying growth conditions. In *Mycobacterium tuberculosis*, two K^+-uptake systems, the Trk comprising the CeoB and CeoC proteins and the Kdp consisting of the two-component system (TCS), KdpDE and KdpFABC, have been characterized, but their selective utilization during bacterial growth has not been completely explored. In the current study, the roles of the *M. tuberculosis* KdpDE regulatory system alone and in association with the Trk transporters in bacterial growth were investigated by evaluating the growth of *M. tuberculosis* KdpDE-deletion and KdpDE/Trk (KT)-double knockout mutant strains in planktonic culture under standard growth conditions. The KT-double knockout mutant strain was first constructed using homologous recombination procedures and was evaluated together with the KdpDE-deletion mutant and the wild-type (WT) strains with respect to their rates of growth, K^+-uptake efficiencies, and K^+-transporter gene expression during planktonic growth. During growth at optimal K^+ concentrations and pH levels, selective deletion of the TCS KdpDE (KdpDE-deletion mutant) led to attenuation of bacterial growth and an increase in bacterial K^+-uptake efficiency, as well as dysregulated expression of the *kdpFABC* and *trk* genes. Deletion of both the KdpDE and the Trk systems (KT-double knockout) also led to severely attenuated bacterial growth, as well as an increase in bacterial K^+-uptake efficiency. These results demonstrate that the KdpDE regulatory system plays a key role during bacterial growth by regulating K^+ uptake *via* modulation of the expression and activities of both the KdpFABC and Trk systems and is important for bacterial growth possibly by preventing cytoplasmic K^+ overload.

Keywords: *Mycobacterium tuberculosis*, two-component system KdpDE, KdpFABC system, Trk system, K^+-uptake systems, K^+ concentration, pH level, gene expression

INTRODUCTION

Growth of *Mycobacterium tuberculosis* is optimal in artificial culture media supplemented with potassium (K^+) concentrations of around 14–15 millimolar (mM) at a pH of about 6.8 (Piddington et al., 2000; Cholo et al., 2015; Baker et al., 2019; Salina et al., 2019). *Mycobacterium tuberculosis* uses K^+ to support cellular metabolic activities, such as cell wall biosynthesis, protein synthesis,

lipid metabolism, and aerobic respiration, which are associated with logarithmic growth (Salina et al., 2014). However, in K^+-limiting (Salina et al., 2014) and low pH environments (pH 5.5; Cholo et al., 2015), this bacterial pathogen alters its metabolic rates, transitioning to slow-growing, persistent-to-non-growing, and dormant phenotypes and allowing for bacterial survival in these unfavorable growth conditions.

Other studies have shown that bacteria adapt to these adverse growth conditions through differential utilization of K^+-uptake transport systems (Epstein, 2003). For example, *Escherichia coli* utilizes the Trk system at elevated K^+ concentrations in near-neutral pH, while the K^+-uptake permease (Kup: TrkD) is used at low pH. However, at K^+-limiting conditions at neutral pH, *E. coli* utilizes the Kdp system (Roe et al., 2000; Epstein, 2003). In this context, *M. tuberculosis* possesses two active K^+-uptake transporters, namely the Trk and Kdp systems (Cole et al., 1998). The Trk consists of two TrkA proteins, CeoB and CeoC, encoded by highly homologous *trk* genes (*ceoB* and *ceoC*) of the *ceoBC* operon. We have previously shown that the Trk system is constitutively expressed in *M. tuberculosis* (Cholo et al., 2015), and has a lower affinity for K^+ than the Kdp (Cholo et al., 2006). It plays a role in slowing bacterial growth in optimal conditions in standard 7H9 broth medium (15 mM K^+, pH 6.8; Cholo et al., 2006). The Trk system has also been implicated in bacterial dormancy, with the CeoB protein being expressed at low K^+ concentrations in biofilm cultures (Kerns et al., 2014; Hegde, 2020), while both *trk* genes are upregulated at low extracellular pH levels in planktonic culture (Cholo et al., 2015).

The *M. tuberculosis* Kdp system, on the other hand, is an inducible two-component system (TCS) comprised of the KdpDE sensor-regulator and the high-affinity K^+-uptake transporter complex, KdpFABC (Cholo et al., 2006). The two Kdp components consist of six proteins encoded by a cluster of six genes, arranged in two operons, *kdpDE* and *kdpFABC*. These operons are divergently transcribed on the bacterial genome; being separated by an intergenic region (~234 bp) located between the *kdpD* and *kdpF* genes (Cole et al., 1998; Cholo et al., 2008). Activation of the Kdp system is mediated by the KdpDE, consisting of the sensor kinase KdpD and response regulator KdpE proteins (Steyn et al., 2003; Freeman et al., 2013). The KdpD and KdpE proteins interact with one another under basal conditions, with KdpD sensing the environmental stimulus and undergoing autophosphorylation at the histidine-642 residue, transferring the phosphoryl moiety to the KdpE subunit at the aspartate-52 residue, resulting in its phosphorylation, and leading to induction of the *kdpFABC* operon (Steyn et al., 2003; Agrawal and Saini, 2014). Low extracellular pH levels (Cholo et al., 2015), as well as K^+-limiting conditions (Salina et al., 2014), are environmental stressors that lead to the induction of both the *kdpDE* and *kdpFABC* operons. These adverse growth conditions also result in acquisition of a dormant phenotype, seemingly implicating the involvement of the Kdp system in bacterial survival.

As with growth in adverse conditions of low extracellular K^+ and pH, information on the role of the Kdp system during mycobacterial growth in optimal K^+ concentrations and pH levels *in vitro* is also limited. In this context, we have previously reported that in the setting of optimal growth conditions, the Kdp system is repressed when the Trk system is functional, being induced and activated as a back-up when the Trk system is inactive (Cholo et al., 2006). Despite the limited information on growth, a few studies have identified that only the *kdpE* among the *kdp* genes, is necessary for optimal growth of *M. tuberculosis* (Sassetti et al., 2003; Griffin et al., 2011), as well as that of *Mycobacterium smegmatis* (Ali et al., 2017).

The issue of the differential utilization of these high- and low-affinity K^+-uptake systems of *M. tuberculosis* during planktonic growth of the pathogen, including the seemingly modulatory role of KdpDE, have been explored in the current study. The research strategy used was based on a comparison of the growth of a selective KdpDE-deletion mutant strain and a recently constructed KdpDE/Trk (CeoBC; KT)-double knockout mutant strain with that of the wild-type (WT) strain of *M. tuberculosis* in standard growth conditions.

MATERIALS AND METHODS

Antimicrobial Agents and Chemicals

Unless indicated, all chemicals were purchased from the Sigma Chemical Co (St. Louis, MO, United States). Kanamycin and hygromycin antibiotics were used at 10 and 50 mg/L, respectively, for the selection of antibiotic-resistant colonies; 5-bromo-4-chloro-3-indolyl-B-D-galactoside (X-Gal) at 0.24 mg/L for blue colonies and sucrose at 2% (20 g/L) as a counter-selectable marker for *sacB*-expressing clones.

Rubidium-86 chloride ($^{86}Rb^+$) was purchased from PerkinElmer Life and Analytical Sciences, Du Pont-NEN Research Products, Boston, MA, United States.

Strains and Growth Media

All plasmids and bacterial strains used in this study are as shown in **Table 1**. The *E. coli* DH5α competent cells were used for cloning procedures for plasmid transformation. The pSOUP42 suicide-delivery vector (SDV) carrying the mutated *M. tuberculosis kdpDE* fragment and the KdpDE-deletion mutant strain of *M. tuberculosis*, were kindly provided by Professor N. Stoker, Royal Veterinary College, United Kingdom (Parish et al., 2003). The *M. tuberculosis* Trk-deletion mutant strain, which we had constructed previously (Cholo et al., 2006) was used for the construction of the KT-double knockout. The *M. tuberculosis* KdpDE-deletion (Parish et al., 2003) and the KT-double knockout (current study) mutant strains were compared with the WT strain for phenotypic and genotypic characteristics.

The Psi (Ψ)- broth medium (5 g Bacto yeast extract, 20 g Bacto tryptone, 5 g $MgSO_4$/L, and pH 7.5) was used for preparation of competent *E. coli* cells and Luria-Bertani (LB) broth for growing plasmid-carrying *E. coli* bacteria, while the *E. coli* colonies were developed on Luria agar (LA) medium. For *M. tuberculosis* cultures, 7H9 broth and 7H10 agar (Difco) media supplemented with 10% oleic acid, dextrose, catalase (OADC), and 2/5% glycerol with/without 0.05% Tween 80

TABLE 1 | Plasmids and bacterial strains used in the study.

Strain	Feature or relevant genotype	Source
Plasmids		
pSOUP42	*Pac*I fragment from pGOAL19, unmarked 1,691-bp *Sph*I *kdpDE* deletion in *kdpD* and *kdpE* genes.	Parish et al., 2003
pSOUP43	*Pac*I fragment from pGOAL17, unmarked 1,691-bp *Sph*I *kdpDE* deletion in *kdpD* and *kdpE* genes.	This study
Bacteria		
DH5α	Wild type *Escherichia coli* strain.	
H37Rv	Wild type laboratory *M. tuberculosis* strain ATCC 25618.	Parish et al., 2003
KdpDE-deletion	Deletion of 1,691-bp *Sph*I *kdpDE* fragment in *kdpD* and *kdpE* genes.	Parish et al., 2003
Trk-deletion	Marked mutant, deletion of 348-bp *ceoB*, insertion of 1746-bp *hyg*-resistant gene cassette at *ceoC* gene.	Cholo et al., 2006
KdpDE/Trk-double knockout mutant	Marked mutant, combined mutations of *kdpDE* and *ceoBC* operons.	This study

respectively, were used for liquid-based growth assays and colony development, respectively.

Construction of a KdpDE/Trk (CeoBC)-Double Knockout Mutant

The KT-deletion mutant, which is characterized by inactivation of both the *kdpDE* and *trk* (*ceoBC*) operons, was constructed using homologous recombination following a two-step strategy (Parish and Stoker, 2000). Prior to electroporation, the *hyg*, P_{Ag85}-*lacZ*, P_{hsp60}-*sacB PacI* cassette in the pSOUP42 SDV was replaced with the P_{Ag85}-*lacZ*, P_{hsp60}-*sacB PacI* cassette from pGOAL17 to form pSOUP43 (**Table 1**). Briefly, approximately 5 μg of UV-pretreated pSOUP43 was electroporated into the *M. tuberculosis* Trk-deletion mutant, carrying the mutation of the *ceoBC* operon (Cholo et al., 2006) and plated onto 7H10 agar medium supplemented with hygromycin, kanamycin, and X-Gal for the isolation of blue single crossover (SCO) clones, followed by isolation of white double-crossovers (DCOs) on non-selection media. The DCO clones were characterized for absence of the plasmid phenotypically using sucrose and kanamycin sensitivity testing procedures. DNA samples from white sucrose-resistant, kanamycin-sensitive *M. tuberculosis* clones were extracted and mutations at the *kdpDE* and *ceoBC* operons were confirmed by a non-radioactive Southern blotting procedure using a digoxigenin (DIG)-labeled PCR-synthesized probe as described in the PCR DIG Probe Synthesis kit (Roche Molecular Biochemicals, Mannheim, Germany).

Bacterial Inoculum Preparation

A bacterial inoculum of each strain was prepared as described, with minor modifications (Cholo et al., 2015; Mothiba et al., 2015). Briefly, a seed culture of *M. tuberculosis* cells was inoculated into 50 ml of 7H9 broth and grown to the mid-log phase at 37°C under stirring conditions. The bacterial cells were harvested by centrifugation at 2851 × *g* at room temperature (RT) for 15 min and the supernatant discarded. The pellet was washed twice and re-suspended in 7H9 broth, followed by adjustment of the optical density (OD) to 1.2 at 540 nm, yielding ca. 10^8–10^9 colony-forming units (cfu)/ml. An inoculum of ca. 10^5 cfu/ml was used in all of the assays.

Preparation of Bacterial Cultures at Logarithmic Phases

For each strain, the bacterial cultures were prepared by inoculating approximately 10^5 cfu/ml cells into 7H9 broth followed by incubation of the culture at 37°C under stirring conditions until the early-, mid-, and late-log phases were reached, corresponding to ODs of 0.1–0.3, 0.4–0.6, and 2.0–2.3 at 540 nm, respectively (Cholo et al., 2015).

Rates of Growth

Cultures for determination of the rates of growth were prepared by inoculating ca. 10^5 cfu/ml of cells of the WT and mutant strains of *M. tuberculosis* into 7H9 broth. The cultures were thoroughly mixed followed by incubation at 37°C for 15 days in the dark with continuous stirring. The cultures were sampled every 3 days beginning at day 0 (D0) to day 15 (D15) and growth was determined spectrophotometrically at a wavelength of 540 nm. The rates of growth of each strain were determined as the time taken by the bacteria to reach the different logarithmic growth phases.

Uptake of Rubidium ($^{86}Rb^+$)

Uptake of K^+ by the WT and mutant strains was determined at the mid- and late-log phases using $^{86}Rb^+$ as a surrogate tracer for K^+. Briefly, the bacteria were harvested from cultures grown to the two logarithmic growth phases and resuspended to ca. 10^6 cfu/ml in K^+-free buffer (KONO) containing 2 mCi/L $^{86}Rb^+$ and uptake of the radioisotope determined as absolute counts per minute (cpm) as previously described (Steel et al., 1999; Cholo et al., 2015).

Extracellular Potassium and pH

The K^+ concentrations and pH levels were determined at the early-, mid-, and late-log phases for each strain. Following culture preparation, the supernatants were harvested by centrifugation (2,851 × *g*, 15 min) followed by decontamination by heat treatment at 95°C for 60 min. The K^+ concentrations and pH levels were measured in the undiluted samples by indirect potentiometry utilizing a K^+-selective electrode in conjunction with a Na^+-reference electrode using the Beckman Coulter Synchron LX 20 System (Beckman Coulter, Ireland Inc., Gateway, Ireland) and the Crison micropH2001 pH meter (Crison Instruments, Barcelona, Spain), respectively. These measurements were determined at the initial (D0), intermediate and final (early-, mid-, and late-log) growth phases, as well as in the processed and unprocessed bacteria-free 7H9 broth medium.

Gene Expression Using the Reverse-Transcriptase-PCR

Gene expression was performed at the early-, mid-, and late-log growth phases in standard 7H9 liquid culture medium as described previously (Cholo et al., 2015). Briefly, RNA was extracted following the Trizol method, and complementary deoxyribonucleic acid (cDNA) was synthesized using the Sigma Enhanced Avian HS reverse transcriptase-PCR (RT-PCR) kit and amplified by quantitative (q)PCR using the LightCycler FastStart DNA Master SYBR Green I kit with the LightCycler 2.0 instrument (Roche Molecular Biochemicals, Mannheim, Germany). The quantities of the individual genes were determined using absolute (AQ) and relative (RQ) quantifications with *sigA* as the reference gene. The relative quantifications were determined based on quantification cycles (Cq) using the $2^{-\Delta\Delta Cq}$ method.

Statistical Analysis

All statistical analyses were performed using the INSTAT program and the unpaired and paired Student *t*-test/Mann–Whitney U-test for analysis of growth rates and gene expression data, respectively. The results are expressed as the means ± SDs. Significance levels were taken at a $p \leq 0.05$.

RESULTS

Construction of the KT-Double Knockout Mutant

In order to investigate the roles of the TCS KdpDE system on bacterial growth, alone and in association with the Trk system, acquisition of the K^+-uptake mutant strains lacking the single KdpDE regulatory system (KdpDE-deletion mutant strain), as well as a combination of both the KdpDE and Trk systems (KT-double knockout mutant strain), was necessary.

The TCS KdpDE, together with the KdpFABC transporter, are the main components of the Kdp system, encoded by separate *kdpDE* and *kdpFABC* operons, transcribed in opposite directions, with the start codon of *kdpD* separated by 234 bp from the start codon of *kdpF*. This genomic arrangement of the *kdp* operons is similar to that of *Mycobacterium bovis*, but is different from those of other bacterial and mycobacterial species, which show both operons having similar transcriptional orientations with the *kdpD* being adjacent to the *kdpC* (Cole et al., 1998; Cholo et al., 2008; Agrawal and Saini, 2014).

The construction of the single KdpDE-deletion mutant of *M. tuberculosis* was as previously reported (Parish et al., 2003). Both the *kdpD* (*Rv1028c*: 2583 bp) and *kdpE* (*Rv1027c*: 681 bp) genes are transcribed in the negative direction with *kdpE* located at position 1148.427–1149.107 (681 bp) with its start codon overlapping the stop codon of *kdpD* found at position 1149.104–1051.686 on the chromosome. Mutation of the *kdpDE* operon was achieved by deletion of a 1,691-bp *Sph*I fragment that spans both the *kdpD* and *kdpE* genes, resulting in inactivation of the sensor kinase/response regulator of the KdpFABC system. However, due to separation of the *kdpDE* and promoter-carrying *kdpFABC* operons on the *M. tuberculosis* genome, the promoter region of the *kdpFABC* operon remained genotypically intact.

In the case of the KT-double knockout strain of *M. tuberculosis*, we used the Trk-deletion mutant strain, which we had constructed previously (Cholo et al., 2006). The *trk* genes comprising the *ceoB* (*Rv2691*: 684 bp) and *ceoC* (*Rv2692*: 663 bp) genes are found on the *ceoBC* operon, with the *ceoB* located at position 3009.344–3010.027, with its stop codon overlapping the start codon of *ceoC* found at position 3010.024–3010.686 on the chromosome. Both *trk* genes are transcribed in the positive direction. The Trk-deletion mutant strain was characterized by inactivation of both the *trk* genes, resulting in a 348-bp deletion at *ceoB* gene and insertion of the 1746-bp *Bam*HI-*Bgl*II *hyg* resistance cassette, derived from the plJ963 vector, at the *Nhe*I site of the *ceoC* gene. The KT-double knockout mutant was constructed by introducing a *kdpDE*-deletion fragment-carrying plasmid, pSOUP43, constructed as described (Parish et al., 2003; **Table 1**) into the Trk-deletion mutant (Cholo et al., 2006; **Figure 1**). Successful mutagenesis of the *kdpDE* operon in the KT-double knockout mutant strain was evident by detection of the 1996-bp *Xho*I-*kdpDE* fragment with the 1,010-bp PCR-synthesized *kdpDE* probe (**Figure 1A; Table 2**; Parish et al., 2003). For the *ceoBC* operon, mutation was revealed by detection of the 2,231-bp *Bcl*I-*ceoBC* fragment using the 714-bp PCR-synthesized *ceoB* probe (**Figure 1B; Table 2**; Cholo et al., 2006).

Rates of Growth

We used the WT and mutant strains to determine the role of the TCS KdpDE alone and in association with the Trk system on bacterial growth. This was achieved by assessing the rates of growth of the WT and the K^+-uptake mutant strains in 7H9 broth medium (15 mM K^+, pH 6.7) under aerobic conditions, sampled every 3 days beginning at D0 to D15 for OD determination. We have previously shown that the OD measurements of 0.1–0.3, 0.4–0.6, and 2.0–2.3 at 540 nm corresponded to the early-, mid-, and late-log phases, respectively (Cholo et al., 2015). The rate of growth of each strain was determined as the time, measured in days, taken by the bacteria to reach these growth phases, and the results are shown in **Figure 2**.

The numbers of bacteria were determined at D0 and were shown to be $8.8 \times 10^4 \pm 7.5 \times 10^4$, $2.1 \times 10^5 \pm 1.8 \times 10^5$, and $9.7 \times 10^4 \pm 1.3 \times 10^5$ cfu/ml for the WT, KdpDE-deletion, and KT-double knockout mutant strains, respectively. The results showed that the rates of growth of the WT and the mutant strains were different, entering the three log phases of growth at varying time points. The growth rates of the mutant strains were attenuated, showing prolonged early-log phases and reaching the early-, mid-, and late-log phases at D9, D12, and D15, respectively, while the WT reached these phases of growth at D6, D9, and D12. The rates of growth were significantly different between the WT and KdpDE-deletion mutant strain at D6, D9, and D12 ($p < 0.05$), while the rates of growth were comparable between the two strains at D15 (ODs: 2.468 ± 0.0064 and 2.360 ± 0.118 for the WT and KdpDE-deletion mutant, respectively, $p = 0.132$). However, the KT-double knockout strain was highly attenuated for growth, with the rates of growth of this mutant being significantly slower than those of the WT at D6, D9, D12, and D15 ($p < 0.05$).

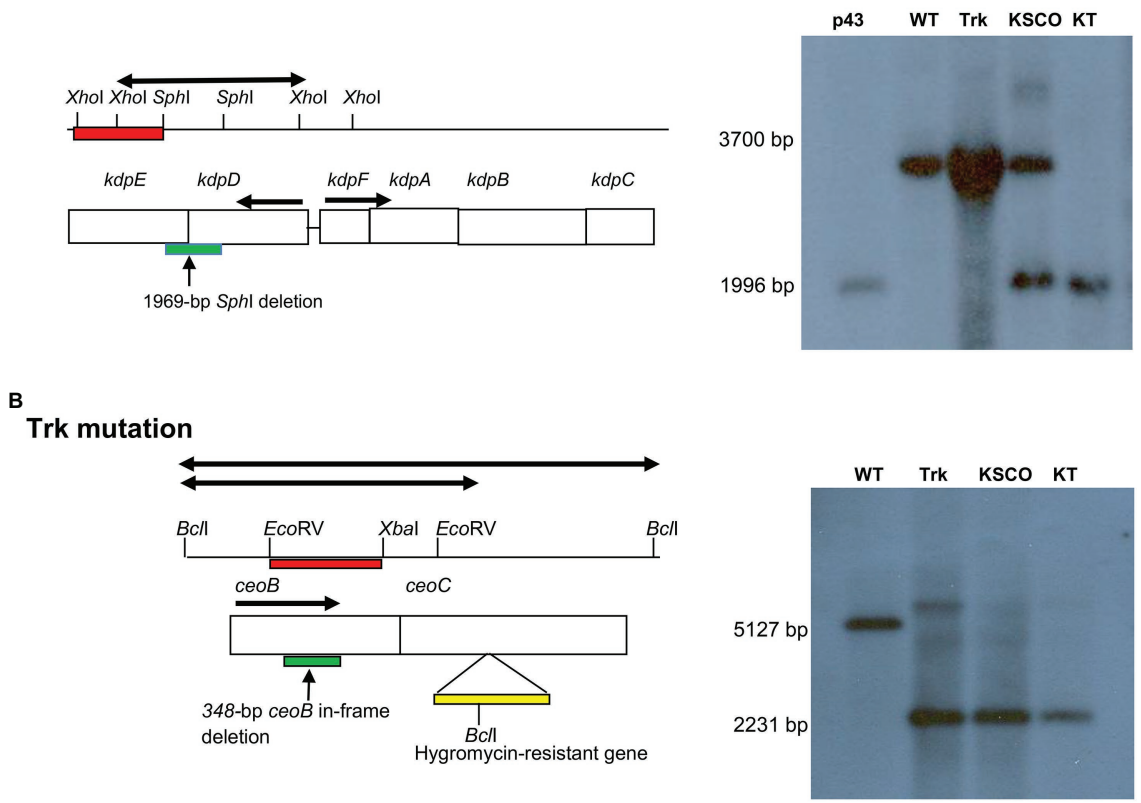

FIGURE 1 | Schematic illustrations of allelic exchange mutagenesis of the *kdpDE* and *ceoBC* mutations in the KT-double knockout strain of *M. tuberculosis* as shown previously for inactivation of the single Trk- and KdpDE-deletion mutant strains (Parish et al., 2003; Cholo et al., 2006). The maps were not drawn to scale. DNA samples were digested with **(A)** *Xho*I and probed with 1,010-bp *kdpDE* PCR-synthesized fragments (red bar) for detection of the 1996-bp *Xho*I-*kdpDE* mutated fragment (thick black arrow) and **(B)** *Bcl*I and probed with 714-bp *ceoB* PCR-synthesized fragment (red bar) for detection of the 2,231-bp *Bcl*I-*ceoBC* mutated fragment (thick black arrow). Deletions in *kdpDE* and *ceoB* genes are shown by green bars, while insertion of the *hyg* gene in *ceoC* gene is shown by yellow bar. p43, pSOUP43; WT, wild type; Trk, Trk-deletion; KSCO, *kdpDE* single crossover; KT, KT-double knockout.

TABLE 2 | Primers used for probe preparation for Southern blotting for KT-double knockout mutant strain construction.

Gene name	Forward primer (bp)	Reverse primer (bp)	Target fragment length (bp)
kdpDE	TCG AGC CCG CAC TGC GCA CCG TGC CGC TGG (30)	CTG GAA ATG CTG GCC CGC AAC CGC GGC AAG (30)	1,010
ceoBC	CCA TCA GGG CGC TGG CAA (18)	CGG CCT GTA GGA CCG TCT (18)	714

Despite the KdpDE-deletion and KT-double knockout mutant strains reaching the early-, mid-, and late-log phases at the same time points (i.e., at D9, D12, and D15, respectively), the growth levels of the two mutant strains determined by comparing their OD measurements at the different time points, were nevertheless different, as shown in **Figure 2**, with the rate of growth of the KT-double knockout mutant strain at D6, D9, and D12 being significantly slower than that of the KdpDE-deletion mutant strain ($p < 0.05$).

As we have previously shown that *M. tuberculosis* utilizes the Trk K^+-uptake transporter exclusively when cultured during standard growth conditions in 7H9 medium, suppressing the activity of the KdpFABC transporter (Cholo et al., 2006, 2015), these observations appear to implicate the KdpDE system in harmonizing the activities of both the Trk and KdpFABC systems to achieve optimum growth.

$^{86}Rb^+$ Uptake

Following determination of growth rates, we then assessed the magnitudes of K^+-uptake by measuring the uptake of $^{86}Rb^+$ by the WT and mutant strains at the mid- and late-log phases of growth as described previously (Steel et al., 1999; Cholo et al., 2015) and the results are shown in **Figure 3**.

Using the $^{86}Rb^+$-uptake model we have, previously shown that the Trk system has a low-affinity for K^+ being responsible for K^+ influx during bacterial growth at the mid- and late-log phases. On the other hand, the Kdp system has been characterized as the high-affinity K^+ transporter, which is suppressed during

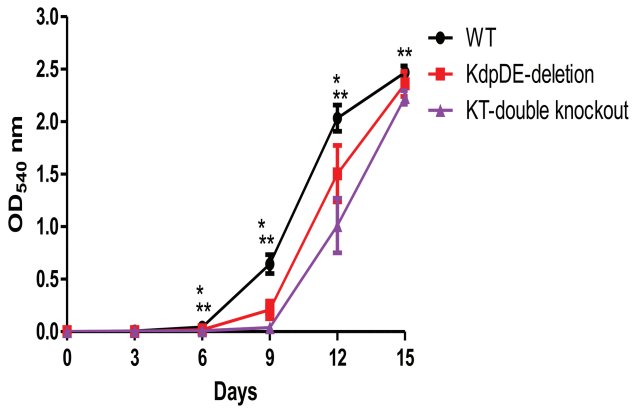

FIGURE 2 | The rates of planktonic growth of the WT and mutant strains of *M. tuberculosis* measured over 15 days. The results are of three different experiments and are expressed as the mean optical density (OD) ± SD at 540 nm. The numbers of bacteria were $8.8 \times 10^4 \pm 7.5 \times 10^4$, $2.1 \times 10^5 \pm 1.8 \times 10^5$, and $9.7 \times 10^4 \pm 1.3 \times 10^5$ cfu/ml for the WT, KdpDE-deletion and KT-double knockout mutant strains at D0, respectively. The OD values for the WT, KdpDE-deletion and KT-double knockout mutants at D6 were: 0.044 ± 0.024, 0.018 ± 0.006, and 0.009 ± 0.006; at D9: 0.644 ± 0.09, 0.208 ± 0.084, and 0.04 ± 0.006; at D12: 2.034 ± 0.126, 1.505 ± 0.269, and 1.011 ± 0.259; and at D15: 2.468 ± 0.0064, 2.360 ± 0.118, and 2.232 ± 0.065, respectively. * and ** represent $p \leq 0.05$ for the KdpDE-deletion and KT-double knockout mutants in relation to the WT, respectively. *Values of p between the WT and KdpDE-deletion mutant at D6, D9, and D12 were 0.026, 0.0022, and 0.0022 respectively; **Values of p between the WT and KT-double knockout at D6, D9, D12, and D15 were 0.0022 at each time point; Values of p between the KdpDE-deletion and KT-double knockout at D6, D9 and D12 were 0.0246, 0.0022, and 0.0260, respectively.

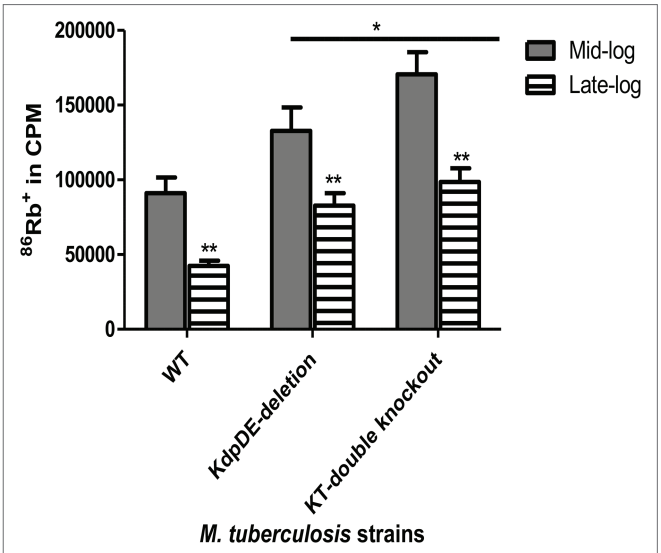

FIGURE 3 | Rubidium ($^{86}Rb^+$) uptake by the WT and the K^+-uptake-deletion mutant strains at the mid- and late-log phases. The results are of five experiments with three replicates for each time point represented as absolute counts per minute (cpm). The gray and striped bars represent uptake of $^{86}Rb^+$ at the mid- and late-log phases, respectively. The cpm values were 91,041 ± 10464.87, 132748.3 ± 15,650, and 170577.3 ± 4877.88 at the mid-log phase and 42431.48 ± 3447.031, 82737.71 ± 8308.834, and 98569.04 ± 9211.451 at the late-log phase for H37Rv (WT), KdpDE-deletion, and KT-double knockout strains, respectively. Statistical differences at $p < 0.05$ are represented by * and **. *Values of p represent comparison of the responses of the WT strain relative to those of the mutant strains at the mid- and late-log phases and were 0.0397 and 0.0006 for the KdpDE-deletion and KT-double knockout mutant strains, respectively at the mid-log phase, while they were 0.0001 for both mutant strains at the late-log phase. The **values of p represent comparison of responses in each strain between mid- and late-log phases.

the logarithmic phases of growth, being induced as a backup when the Trk system is not active (Cholo et al., 2006, 2015).

Somewhat surprisingly, the results revealed that the $^{86}Rb^+$-uptake efficiencies of the mutant strains, were higher than those of the WT strain at both the mid- and late-log phases being highest at the mid-log phase for the KT-double knockout mutant (**Figure 3**). At the late-log phase, the $^{86}Rb^+$-uptake efficiencies of the mutant strains were again higher than that of the WT strain, but not significantly different from one another (82,738 ± 8,309 and 98,569 ± 9,212 cpm for the KdpDE-deletion and KT-double knockout mutant strains, respectively, $p = 0.126$). These findings indicate that deletion of the KdpDE system results in K^+ overload, presumably due to dysregulation of the Trk and KdpFABC K^+ transporters, an event that may disrupt mycobacterial cytoplasmic pH, cellular metabolism, and growth.

For all the strains, the $^{86}Rb^+$-uptake efficiencies were significantly reduced at the late-log phase, being 46, 62, and 57% of the corresponding efficiencies noted at the mid-log phase for the WT, KdpDE-deletion and KT-double knockout mutant strains, respectively, most probably due to decreased extracellular pH (Cholo et al., 2015).

Extracellular Potassium Concentrations and pH Levels

Alterations in the extracellular K^+ concentrations and pH levels that occur during the early-, mid-, and late-log phases of growth of the WT and mutant strains of *M. tuberculosis* under optimal culture conditions in 7H9 broth (15 mM K^+ and pH 6.8) were determined. In the case of the extracellular K^+ concentrations of the culture supernatants, these remained unchanged (14–15.5 mM) with respect to all three strains of *M. tuberculosis* and were comparable during the three log phases (**Supplementary Table S1**). The lack of significant alterations in extracellular K^+ may reflect recycling of K^+ during bacterial growth.

As shown in **Figure 4**, the extracellular pH values of the growth media increased at the early- to mid-log phases and decreased dramatically during the late-log phase, achieving statistical significance in comparison with the corresponding D0 pH value (pH 6.711) for all three strains. The extracellular pH levels of the two mutant strains were comparable at the three log phases of growth, but were higher than those of the WT strain.

Gene Expression

The effects of the KdpDE regulatory system alone and in the presence of the Trk system on the expression of the K^+-uptake genes during bacterial growth were explored by determining the expression levels of all six *kdp* and two *trk* genes in the WT and both mutant strains at the early-, mid-, and late-log

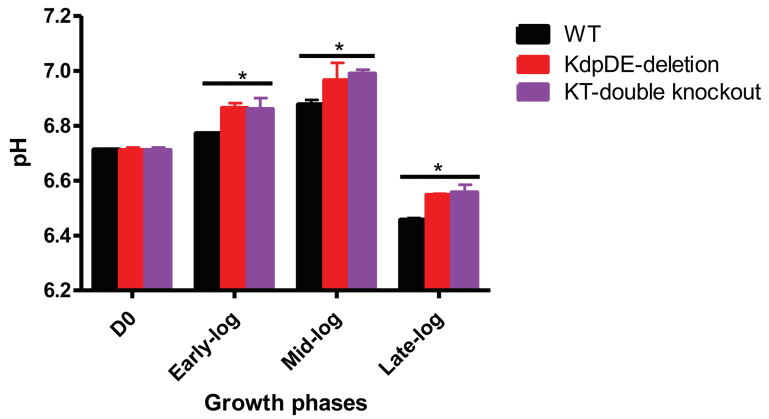

FIGURE 4 | The extracellular pH levels at the three growth phases. The results are of three experiments performed in triplicate. The pH levels were similar at D0 for all the strains 6.711 ± 0.01. The pH levels for the WT, KdpDE-deletion and KT-double knockout mutant strains at early-log were: 6.77 ± 0.011, 6.865 ± 0.017, and 6.86 ± 0.04; at mid-log: 6.876 ± 0.045, 6.965 ± 0.064, and 6.99 ± 0.04; and at late-log: 6.455 ± 0.022, 6.548 ± 0.004, and 6.557 ± 0.028, respectively. Statistically significant differences between D0 vs. the log phases are shown with an (*) representing $p < 0.05$.

growth phases in standard 7H9 liquid culture medium using RT-PCR. We confirmed mutations of the *kdpDE* (*kdpD* and *kdpE*) and *ceoBC* (*ceoB* and *ceoC*) genes by formation of non-specific fragments, with different melting temperatures from those of targeted gene-specific fragments using melting curve analysis data (**Supplementary Table S2**).

The results were analyzed as absolute amounts (μg/ml; AQ) and relative quantifications (RQ) using *sigA* as the reference gene (**Figure 5**; **Supplementary Table S3**). As shown in our previous study under similar conditions (Cholo et al., 2015), the expression levels of the *sigA* gene were comparable between the WT and the KdpDE-deletion and the KT-knockout mutants between the early- and mid-log phases. However, at the late-log phase, expression levels of *sigA* were significantly increased in all the strains, possibly indicative of low pH-induced stress. Similarly, all the K^+-uptake genes in all the strains were elevated during the late-log phase of growth, highlighting responses to alterations in the environmental conditions in the growth medium. Despite this, the levels of increased expression of the *sigA* gene in the WT were much lower than those of the K^+-uptake genes illustrating the constant expression of the *sigA* gene at the logarithmic phase. However, the levels of *sigA* were excessively high in the mutants, increasing by 2–2.5-fold relative to those of the WT at the late-log phase due to low pH stress (Cholo et al., 2015), which are clearly shown by AQ data (**Figure 5A**; **Supplementary Tables S3.1–S3.3**).

For each strain, the expression of all the measured K^+-uptake genes during growth was determined by comparing the expression levels of each gene at the mid- and late-log phases relative to those at the early-log phase.

WT Strain

Absolute Quantification

The AQ data for the WT strain were the same as we have previously reported (Cholo et al., 2015). In summary, during growth, at the early- and mid-log phases, the *kdp* and the *trk* genes were expressed at minimum levels and were upregulated at the late-log phase, with the *ceoB* gene being the most prominently induced gene among all the K^+-uptake genes, followed by *kdpD* and *kdpF* (**Figure 5A**; **Supplementary Table S3.3**). Despite upregulation of the K^+-uptake genes, bacterial growth was slow at the late-log phase, due presumably to low extracellular pH levels (Cholo et al., 2015). Similar findings have been reported in *E. coli*, revealing that expression levels of the *trk* and the *kdp* genes are dependent on the extracellular pH (Epstein, 2003).

Relative Quantification

The RQ values of the genes in the WT strain were determined during growth at the mid- and late-log phases in relation to the early-log phase (**Figure 5B**; **Supplementary Table S3.4**). Expression of the *kdpDE* and *kdpBC* genes was significantly increased at the mid-log phase, while all the *kdp* genes were upregulated at the late-log phase. In the case of the *trk* genes, expression levels of both genes were unchanged at the mid-log phase, while both genes were upregulated at the late-log phase, particularly the *ceoB* gene.

KdpDE-Deletion Mutant Strain

Absolute Quantification

During growth, expression levels of all the *kdpFABC* and *trk* genes were minimal at the early- and mid-log phases. However, the *kdpFABC* genes were upregulated at early-log phase probably as a response to the decrease in extracellular pH (6.7), while they were downregulated at mid-log phase at elevated pH level of 6.9. These genes were significantly increased by ≥1,000-fold at the late-log phase at the lower pH of 6.5 (**Figure 3**; **Supplementary Tables S3.1** and **S3.3**). In relation to the WT (Cholo et al., 2015), the *kdp* genes were increased, while the *trk* genes were decreased at early-log phase. Both *kdpFABC* and *trk* genes were downregulated at mid-log phase. However, with the exception of the *ceoB* gene,

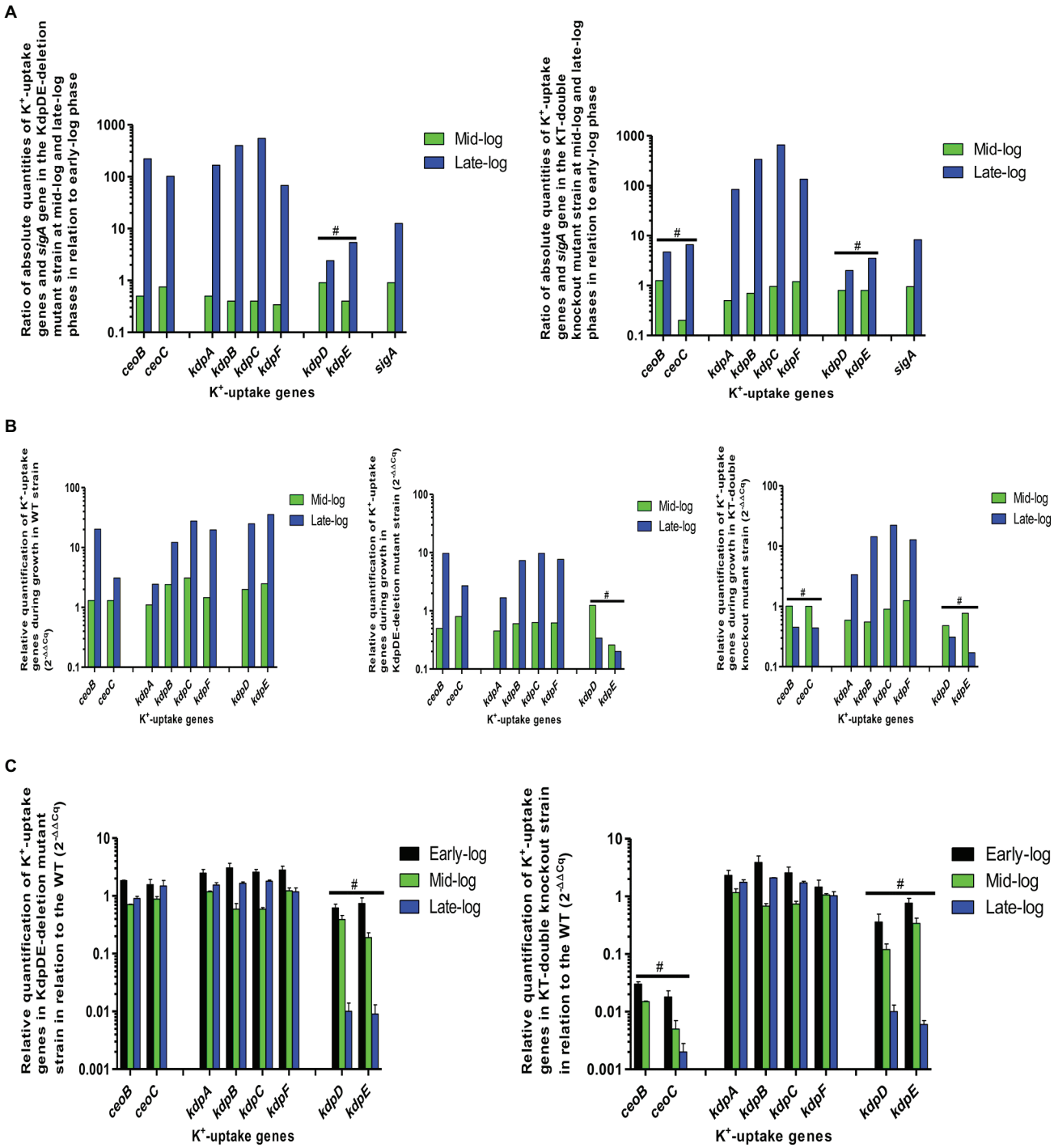

FIGURE 5 | Gene expression measured during the various phases of bacterial growth for the WT and mutant strains. **(A)** Ratio of gene expression of absolute quantities of the K^+-uptake genes and the *sigA* gene in the mutant strains during growth at mid- and late-log phases in relation to early-log phase, **(B)** relative quantification of K^+-uptake genes in the mutant strains during growth relative to early-log phase ($2^{-\Delta\Delta Cq}$), and **(C)** relative quantification of K^+-uptake genes in mutant strains relative to the WT ($2^{-\Delta\Delta Cq}$). #represents mutations of the individual genes.

and similar to those of the WT, the expression levels of the *ceoC* and *kdpFABC* genes were significantly increased by at least 1.5 and up to 7-fold relative to those of the WT at the late-log phase.

Relative Quantification

During growth, all the *kdpFABC* and *trk* genes were downregulated at the mid-log phase and upregulated at the late-log phase (**Figure 5B**; **Supplementary Table S3.4**). In relation

to the WT strain (**Figure 5C**; **Supplementary Table S3.5**), all the *kdpFABC* and *trk* genes were increased at the early-log phase, decreased at mid-log phase, and upregulated at the late-log phase, showing dependency on extracellular pH levels for their expression.

The AQ and RQ results appear to show that inactivation of the KdpDE system leads to dysregulation of both the *kdpFABC* and *trk* genes, resulting in constitutive expression of both operons.

KT-Double Knockout Mutant

Only the *kdpFABC* genes were evaluated as the *kdpDE* and *trk* genes were deleted in this mutant strain as shown by Tm of non-specific fragments due to mutations of the targeted genes (**Supplementary Table S1**).

Absolute Quantification

The expression levels and patterns of the *kdpFABC* genes were similar to those of the KdpDE-deletion mutant at the three growth phases (**Figure 5A**; **Supplementary Tables S3.2** and **S3.3**).

Relative Quantification

During growth, the *kdpA* and *kdpB* genes were downregulated, while expression of the *kdpC* and *kdpF* genes remained unchanged at the mid-log phase (**Figure 5B**; **Supplementary Table S3.4**). All the *kdpFABC* genes were upregulated at the late-log phase. In relation to the WT strain (**Figure 5C**; **Supplementary Table S3.6**) and similar to the KdpDE-deletion mutant, all the *kdpFABC* genes were increased at the early-log phase, decreased at the mid-log phase, and upregulated at the late-log phase.

DISCUSSION

Mycobacterium tuberculosis is able to grow and adapt to varying environmental conditions. We have previously demonstrated that this bacterial pathogen grows exponentially at the mid-log phase at optimum K^+ concentrations and pH levels (14–15 mM K^+, pH 6.8–7.0), but is attenuated for growth, acquiring slow growth-to-dormant status at low pH levels (pH 5.5–6.0), despite maintaining elevated K^+ concentrations at the late-log phase (Cholo et al., 2015).

Most bacteria adapt to varying extracellular K^+ concentrations and pH levels by utilizing different K^+-uptake transporters. Bacterial species, such as *E. coli* and *Salmonella* species, in which these K^+ transporters have been extensively studied, utilize the Trk at elevated K^+ concentrations and neutral pH levels and the Kup at low pH levels (pH 5.5), while they use the Kdp in K^+-limiting conditions at neutral pH (Epstein, 2003; Liu et al., 2013). We have previously shown that *M. tuberculosis*, which encodes the Trk and Kdp as the predominant K^+-uptake transporters (Cole et al., 1998), utilize the Trk system at mid-log (15 mM K^+, pH 6.8; Cholo et al., 2006) and late-log phases (15 mM K^+, pH 6.5; Cholo et al., 2015). However, in the case of the Kdp system, we have observed that the KdpFABC K^+-uptake transporter is suppressed at mid-log while, similar to the Trk transporter, it is upregulated at the late-log phase (Cholo et al., 2015).

However, information on the differential utilization of these K^+-uptake systems during bacterial growth is limited. This was investigated in the current study, in which we determined the role of the TCS KdpDE system in regulating the activities of the two K^+-uptake transporters, the KdpFABC and Trk, during bacterial growth. We achieved this by first constructing the KT-double knockout mutant by transforming the *kdpDE*-deletion fragment constructed previously (Parish et al., 2003; **Table 1**) into the Trk-deletion mutant, also constructed previously (Cholo et al., 2006). Acquisition of the KdpDE-deletion and KT-double knockout mutant strains of *M. tuberculosis* were essential prerequisites to enable us to probe the involvement of the KdpDE system in harmonizing the activities of the KdpFABC and Trk transporters.

The findings of the current study revealed that the TCS KdpDE of *M. tuberculosis* is mechanistically involved in accelerating the rate of bacterial growth by shortening the duration of the early-log phase. This contention is supported by the observation that selective inactivation of the KdpDE system of *M. tuberculosis* caused attenuation of growth that was associated with prolongation of the early-log phase and delayed progression to the mid- and late-log phases, even in the presence of favorable extracellular K^+ concentrations and near-neutral pH, both of which are conducive to exponential growth. Involvement of the KdpDE regulatory system during the exponential phase of bacterial growth was also evident, as demonstrated by the upregulation of both the *kdpD* and *kdpE* genes in the WT strain at the mid-log phase (**Figure 5B**; **Supplementary Table S3.4**). Other studies in *M. tuberculosis* and *M. smegmatis* have emphasized the essentiality of the *kdpE* gene during bacterial growth in standard growth conditions *in vitro* (Sassetti et al., 2003; Griffin et al., 2011; Ali et al., 2017).

As shown previously, *M. tuberculosis* cultured in these conditions utilizes the Trk system as the main K^+-uptake transporter for growth (Cholo et al., 2006). Interestingly, however, in the current study, the growth of *M. tuberculosis* expressing an intact Trk system in the absence of KdpDE system (selective *kdpDE*-gene knockout mutant strain) was significantly attenuated, seemingly, implicating the KdpDE regulatory system in modulating the activity of the Trk system during growth. Not surprisingly, dual inactivation of the KdpDE and the Trk systems resulted in the most severe attenuation of growth relative to both the WT and KdpDE-deletion single knockout mutant. These observations not only underscore the interaction between these systems in promoting optimum bacterial growth, but also the seemingly key involvement of KdpDE in regulating the Trk, as well as the KdpFABC systems.

Using the $^{86}Rb^+$ uptake procedure to determine bacterial K^+-uptake efficiency, we demonstrated that during growth in optimal conditions, *M. tuberculosis* uses the KdpDE system to regulate K^+ uptake by the KdpFABC and Trk transporters. In our previous studies, using the Trk-deletion mutant strain

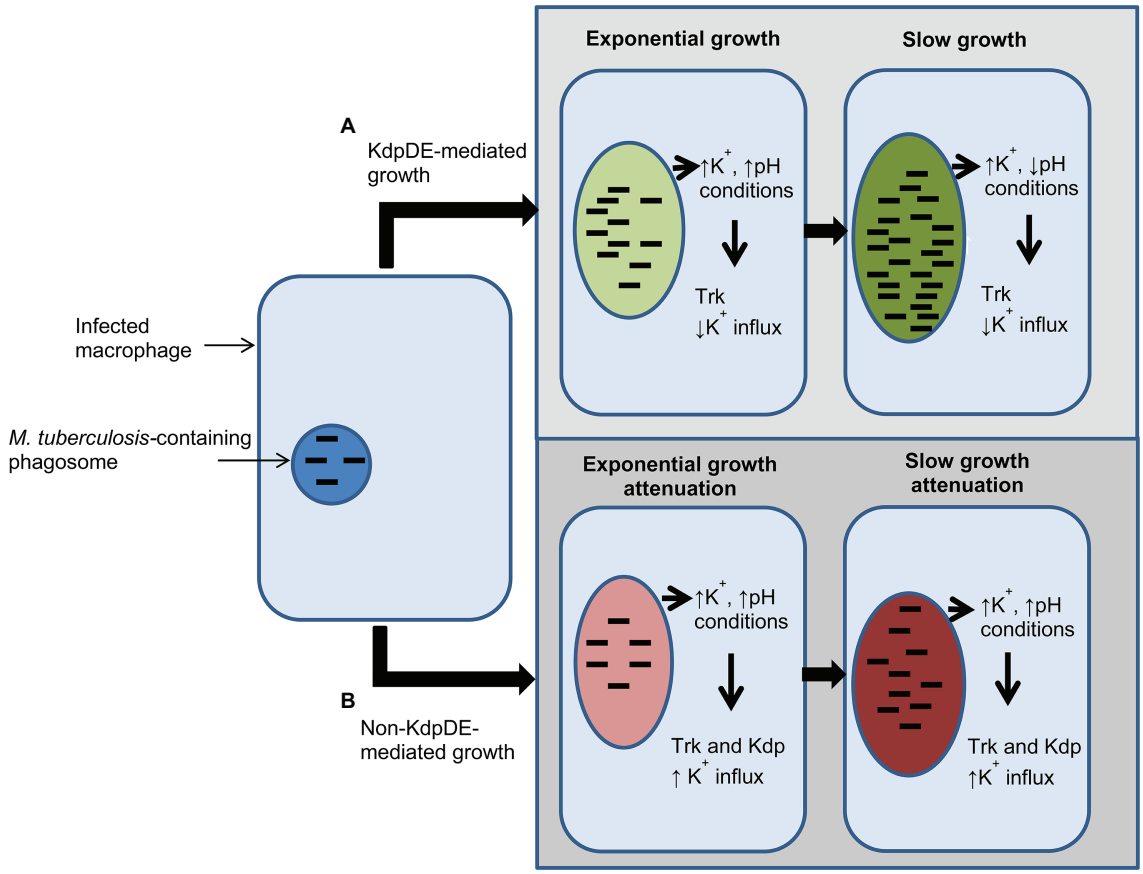

FIGURE 6 | Schematic illustration summarizing events involved in bacterial growth in macrophages in the presence and absence of the two-component system (TCS) KdpDE. **(A)** In the presence of the KdpDE system, bacteria utilize the Trk transporter for K^+ uptake suppressing the KdpFABC transporter, while **(B)** in the absence of the KdpDE both the KdpFABC and the Trk systems are constitutively activated, resulting in excessive influx of K^+, which is detrimental to the bacteria, leading to attenuation of bacterial growth. ↑K^+, phagosomal potassium concentration (19–50 mM); ↑pH, pH 6.7–6.8; ↓pH, pH 6.5.

(Cholo et al., 2006, 2015), we demonstrated that at elevated K^+ concentrations, bacteria utilize the low-affinity Trk system for K^+ uptake, while the high-affinity Kdp system is suppressed, being induced as a back-up in the absence of Trk, conferring high K^+-uptake efficiency on the Trk-deletion mutant (Cholo et al., 2006, 2015). The increase in K^+-uptake efficiency of the KdpDE-deletion mutant strain in the presence of both KdpFABC and Trk K^+-uptake transporters observed in the current study appears to demonstrate failure of the bacteria to differentially regulate the utilization of the two transporters, resulting in both systems being simultaneously operative. Excessive uptake of K^+ is likely to result in dysregulation of bacterial cytoplasmic pH creating an intracellular environment unfavorable for cellular metabolism for growth.

However, at late-log phase (15 mM K^+, pH 6.5), despite bacterial requirements of both the Kdp and Trk systems, the Trk has been shown to be the main K^+-uptake transporter responsible for uptake of the cation. This contention is supported by the $^{86}Rb^+$ uptake data, showing low K^+-uptake efficiency of the WT in relation to the Trk-deletion mutant, together with gene expression data (AQ) in the WT strain showing that the *ceoB* gene is the most highly induced gene among all the K^+-uptake genes at late-log phase (Cholo et al., 2015). In the current study, we have shown that in these conditions, as with optimal growth, the bacteria use the KdpDE regulatory system to regulate the activities and expression of both K^+-uptake transporters. This has been demonstrated by the findings of increased uptake of $^{86}Rb^+$ consistent with dysregulated, simultaneous, excessive functioning of both the Kdp and Trk K^+-uptake transporters, in the absence of the KdpDE regulatory system.

These findings illustrate the constitutive activation of the KdpFABC transporter in the absence of its inducer KdpDE (Epstein, 2003; Steyn et al., 2003; Freeman et al., 2013; Agrawal and Saini, 2014). Similar findings have been demonstrated in previous studies in KdpDE mutant strains of *E. coli* (Asha and Gowrishankar, 1993; Sardesai and Gowrishankar, 2001; Epstein, 2003). These suggest a spontaneous induction of this operon in the absence of its inducer, or alternatively, the presence of an additional mechanism(s) of induction. While these have not been identified in *M. tuberculosis*, such mechanisms that bypass KdpDE have been described in *E. coli* and involve utilization of the histone-like nucleoid-structuring (H-NS) protein, thioredoxin 1, and thioredoxin reductase

(Cole et al., 1998; Sardesai and Gowrishankar, 2001; Epstein, 2016). Although present in *M. tuberculosis*, the involvement of these mechanisms in activation of KdpFABC has not been described (Cole et al., 1998).

We do concede that our findings on the expression of these two mycobacterial K^+-uptake transporters at various stages of growth in the absence of the TCS KdpDE system, are based on quantitation of their mRNA expression, which represents a potential limitation of our study. Nevertheless, we do believe that our findings demonstrate a novel and potentially important dual regulatory role of the KdpDE system in harmonizing the activities of the Trk and Kdp K^+ transporters to ensure stringent control of cytoplasmic pH and growth. We also believe that these findings represent a platform that enables progression to additional confirmatory studies that would include target gene promoter expression using the β-galactosidase assay in this difficult and slow-growing pathogen.

These findings of the current study highlight the critical roles played by the K^+-uptake transporters of *M. tuberculosis* during infection of the host. For example, in macrophages, the primary targets of the pathogen, in high intracellular K^+ concentrations (phagosomal vacuolar K^+ concentration: 19–50 mM; Wagner et al., 2005) and pH levels (pH 6.8; Piddington et al., 2000; Vandal et al., 2009), are probably conducive to bacterial growth. Presumably in this setting, the KdpDE and Trk systems are utilized by *M. tuberculosis* to establish infection (Haydel and Clark-Curtiss, 2004; Rengarajan et al., 2005; MacGilvary et al., 2019), possibly playing a role in bacterial virulence. However, at low pH levels, *M. tuberculosis* bacteria may utilize both the Kdp and Trk systems for survival (**Figure 6**). In this context, absence of the KdpDE system alone, and particularly in combination with the Trk system, is clearly detrimental to bacterial growth, underscoring the potential of these K^+ transporters to serve as potential targets for development of anti-TB drugs.

In conclusion, in *M. tuberculosis*, the KdpDE system plays a key modulatory role in controlling the activities of the KdpFABC and Trk K^+ uptake transporters to regulate growth.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MC and RA contributed to the conception and design of the study and wrote, edited and reviewed the manuscript. MC constructed the mutant strain. MC, MM, and AO performed the phenotypic experiments. MC, MM, AO, and RA contributed to interpretation of the data. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the South African National Research Foundation grant 87649 to MC.

ACKNOWLEDGMENTS

We would like to thank the TB Platform of the South African Medical Research Council (SAMRC) for the provision of the TB laboratory facility and the SAMRC/NHLS/WITS Molecular Mycobacteriology Research Unit, DST-NRF Centre of Excellence for Biomedical Research, School of Pathology, University of the Witwatersrand and National Health Laboratory Service, Johannesburg, South Africa for assistance in generation of the KT-double knockout mutant strain. We would also like to thank N. Stoker for critically reading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.698875/full#supplementary-material

REFERENCES

Agrawal, R., and Saini, D. K. (2014). *Rv1027c-Rv1028c* encode functional KdpDE two-component system in *Mycobacterium tuberculosis*. *Biochem. Biophys. Res. Commun.* 446, 1172–1178. doi: 10.1016/j.bbrc.2014.03.066

Ali, M. K., Li, X., Tang, Q., Liu, X., Chen, F., Xiao, J., et al. (2017). Regulation of inducible potassium transporter KdpFABC by the KdpD/KdpE two-component system in *Mycobacterium smegmatis*. *Front. Microbiol.* 8:570. doi: 10.3389/fmicb.2017.00570

Asha, H., and Gowrishankar, J. (1993). Regulation of *kdp* operon expression in *Escherichia coli*: evidence against turgor as signal for transcriptional control. *J. Bacteriol.* 175, 4528–4537. doi: 10.1128/jb.175.14.4528-4537.1993

Baker, J. J., Dechow, S. J., and Abramovitch, R. B. (2019). Acid fasting: modulation of *Mycobacterium tuberculosis* metabolism at acidic pH. *Trends Microbiol.* 27, 942–953. doi: 10.1016/j.tim.2019.06.005

Cholo, M. C., Boshoff, H. I., Steel, H. C., Cockeran, R., Matlola, N. M., Downing, K. J., et al. (2006). Effects of clofazimine on potassium uptake by a Trk-deletion mutant of *Mycobacterium tuberculosis*. *J. Antimicrob. Chemother.* 57, 79–84. doi: 10.1093/jac/dki409

Cholo, M. C., van Rensburg, E. J., and Anderson, R. (2008). Potassium uptake systems of *Mycobacterium tuberculosis*: genomic and protein organisation and potential roles in microbial pathogenesis and chemotherapy. *South Afr. J. Epidemiol. Infect.* 23, 13–16. doi: 10.1080/10158782.2008.11441327

Cholo, M. C., van Rensburg, E. J., Osman, A. G., and Anderson, R. (2015). Expression of the genes encoding the Trk and Kdp potassium transport systems of *Mycobacterium tuberculosis* during growth *in vitro*. *Biomed. Res. Int.* 2015:608682. doi: 10.1155/2015/608682

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544.

Epstein, W. (2003). The roles and regulation of potassium in bacteria. *Prog. Nucleic Acid Res. Mol. Biol.* 75, 293–320. doi: 10.1016/s0079-6603(03)75008-9

Epstein, W. (2016). The KdpD sensor kinase of *Escherichia coli* responds to several distinct signals to turn on expression of the Kdp transport system. *J. Bacteriol.* 198, 212–220. doi: 10.1128/JB.00602-15

Freeman, Z. N., Drus, S., and Waterfield, N. R. (2013). The KdpD/KdpE two-component system: integrating K^+ homeostasis and virulence. *PLoS Pathog.* 9:e1003201. doi: 10.1371/journal.ppat.1003201

Griffin, J. E., Gawronski, J. D., DeJesus, M. A., Ioerger, T. R., Akerley, B. J., and Sassetti, C. M. (2011). High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog.* 7:e1002251. doi: 10.1371/journal.ppat.1002251

Haydel, S. E., and Clark-Curtiss, J. E. (2004). Global expression analysis of two-component system regulator genes during *Mycobacterium tuberculosis* growth in human macrophages. *FEMS Microbiol. Lett.* 236, 341–347. doi: 10.1111/j.1574-6968.2004.tb09667.x

Hegde, S. R. (2020). Computational identification of the proteins associated with quorum sensing and biofilm formation in *Mycobacterium tuberculosis*. *Front. Microbiol.* 10:3011. doi: 10.3389/fmicb.2019.03011

Kerns, P. W., Ackart, D. F., Basaraba, R. J., Leid, J., and Shirtliff, M. E. (2014). *Mycobacterium tuberculosis* pellicles express unique proteins recognized by the host humoral response. *Pathog Dis.* 70, 347–358. doi: 10.1111/2049-632X.12142

Liu, Y., Ho, K. K., Su, J., Gong, H., Chang, A. C., and Lu, S. (2013). Potassium transport of Salmonella is important for type III secretion and pathogenesis. *Microbiology* 159, 1705–1719. doi: 10.1099/mic.0.068700-0

MacGilvary, N. J., Kevorkian, Y. L., and Tan, S. (2019). Potassium response and homeostasis in *Mycobacterium tuberculosis* modulates environmental adaptation and is important for host colonization. *PLoS Pathog.* 15:e1007591. doi: 10.1371/journal.ppat.1007591

Mothiba, M. T., Anderson, R., Fourie, B., Germishuizen, W. A., and Cholo, M. C. (2015). Effects of clofazimine on planktonic and biofilm growth of *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *J. Glob. Antimicrob. Resist.* 3, 13–18. doi: 10.1016/j.jgar.2014.12.001

Parish, T., Smith, D. A., Kendall, S., Casali, N., Bancroft, G. J., and Stoker, N. G. (2003). Deletion of two-component regulatory systems increases the virulence of *Mycobacterium tuberculosis*. *Infect. Immun.* 71, 1134–1140. doi: 10.1128/IAI.71.3.1134-1140.2003

Parish, T., and Stoker, N. G. (2000). Use of a flexible cassette method to generate a double unmarked *Mycobacterium tuberculosis tlyA plcABC* mutant by gene replacement. *Microbiology* 146, 1969–1975. doi: 10.1099/00221287-146-8-1969

Piddington, D. L., Kashkouli, A., and Buchmeier, N. A. (2000). Growth of *Mycobacterium tuberculosis* in a defined medium is very restricted by acid pH and Mg^{2+} levels. *Infect. Immun.* 68, 4518–4522. doi: 10.1128/IAI.68.8.4518-4522.2000

Rengarajan, J., Bloom, B. R., and Rubin, E. J. (2005). Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc. Natl. Acad. Sci. U. S. A.* 102, 8327–8332. doi: 10.1073/pnas.0503272102

Roe, A. J., McLaggan, D., O'Byrne, C. P., and Booth, I. R. (2000). Rapid inactivation of the *Escherichia coli* Kdp K^+ uptake system by high potassium concentrations. *Mol. Microbiol.* 35, 1235–1243. doi: 10.1046/j.1365-2958.2000.01793.x

Salina, E. G., Grigorov, A. S., Bychenko, O. S., Skvortsova, Y. V., Mamedov, I. Z., Azhikina, T. L., et al. (2019). Resuscitation of dormant "non-culturable" *Mycobacterium tuberculosis* is characterized by immediate transcriptional burst. *Front. Cell. Infect. Microbiol.* 9:272. doi: 10.3389/fcimb.2019.00272

Salina, E. G., Waddell, S. J., Hoffmann, N., Rosenkrands, I., Butcher, P. D., and Kaprelyants, A. S. (2014). Potassium availability triggers *Mycobacterium tuberculosis* transition to, and resuscitation from, non-culturable (dormant) states. *Open Biol.* 4:140106. doi: 10.1098/rsob.140106

Sardesai, A. A., and Gowrishankar, J. (2001). Trans-acting mutations in loci other than KdpDE that affect *kdp* operon regulation in *Escherichia coli*: effects of cytoplasmic thiol oxidation status and nucleoid protein H-NS on *kdp* expression. *J. Bacteriol.* 183, 86–93. doi: 10.1128/JB.183.1.86-93.2001

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* 48, 77–84. doi: 10.1046/j.1365-2958.2003.03425.x

Steel, H. C., Matlola, N. M., and Anderson, R. (1999). Inhibition of potassium transport and growth of mycobacteria exposed to clofazimine and B669 is associated with a calcium-independent increase in microbial phospholipase A2 activity. *J. Antimicrob. Chemother.* 44, 209–216. doi: 10.1093/jac/44.2.209

Steyn, A. J., Joseph, J., and Bloom, B. R. (2003). Interaction of the sensor module of *Mycobacterium tuberculosis* H37Rv KdpD with members of the Lpr family. *Mol. Microbiol.* 47, 1075–1089. doi: 10.1046/j.1365-2958.2003.03356.x

Vandal, O. H., Nathan, C. F., and Ehrt, S. (2009). Acid resistance in *Mycobacterium tuberculosis*. *J. Bacteriol.* 191, 4714–4721. doi: 10.1128/JB.00305-09

Wagner, D., Maser, J., Moric, I., Boechat, N., Vogt, S., Gicquel, B., et al. (2005). Changes of the phagosomal elemental concentrations by *Mycobacterium tuberculosis* Mramp. *Microbiology* 151, 323–332. doi: 10.1099/mic.0.27213-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 September 2021
doi: 10.3389/fgene.2021.733715

Development and Validation of a Novel Gene Signature for Predicting the Prognosis by Identifying m5C Modification Subtypes of Cervical Cancer

Jing Yu[1†], Lei-Lei Liang[1†], Jing Liu[1], Ting-Ting Liu[2], Jian Li[1], Lin Xiu[1], Jia Zeng[1], Tian-Tian Wang[1], Di Wang[3], Li-Jun Liang[3], Da-Wei Xie[3], Ding-Xiong Chen[3], Ju-Sheng An[1] and Ling-Ying Wu[1]**

[1]Department of Gynecologic Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2]Department of Blood Grouping, Beijing Red Cross Blood Center, Beijing, China, [3]State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

OPEN ACCESS

Edited by:
Aparna Vasanthakumar, AbbVie, United States

Reviewed by:
Eman Toraih, Tulane University, United States
Shixiong Zhang, Xidian University, China

***Correspondence:**
Ling-Ying Wu
wulingying@csco.org.cn
Ju-Sheng An
anmanman_0@126.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 08 July 2021
Accepted: 07 September 2021
Published: 22 September 2021

Citation:
Yu J, Liang L-L, Liu J, Liu T-T, Li J, Xiu L, Zeng J, Wang T-T, Wang D, Liang L-J, Xie D-W, Chen D-X, An J-S and Wu L-Y (2021) Development and Validation of a Novel Gene Signature for Predicting the Prognosis by Identifying m5C Modification Subtypes of Cervical Cancer. Front. Genet. 12:733715. doi: 10.3389/fgene.2021.733715

Background: 5-Methylcytidine (m5C) is the most common RNA modification and plays an important role in multiple tumors including cervical cancer (CC). We aimed to develop a novel gene signature by identifying m5C modification subtypes of CC to better predict the prognosis of patients.

Methods: We obtained the expression of 13 m5C regulatory factors from The Cancer Genome Atlas (TCGA all set, 257 patients) to determine m5C modification subtypes by the "nonnegative matrix factorization" (NMF). Then the "limma" package was used to identify differentially expressed genes (DEGs) between different subtypes. According to these DEGs, we performed Cox regression and Kaplan-Meier (KM) survival analysis to establish a novel gene signature in TCGA training set (128 patients). We also verified the risk prediction effect of gene signature in TCGA test set (129 patients), TCGA all set (257 patients) and GSE44001 (300 patients). Furthermore, a nomogram including this gene signature and clinicopathological parameters was established to predict the individual survival rate. Finally, the expression and function of these signature genes were explored by qRT-PCR, immunohistochemistry (IHC) and proliferation, colony formation, migration and invasion assays.

Results: Based on consistent clustering of 13 m5C-modified genes, CC was divided into two subtypes (C1 and C2) and the C1 subtype had a worse prognosis. The 4-gene signature comprising FNDC3A, VEGFA, OPN3 and CPE was constructed. In TCGA

Abbreviations: CC, cervical cancer; m5C: 5-methylcytidine; DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; TCGA, The Cancer Genome Atlas; ROC, receiver operating characteristic; HR, hazard ratio; CI, confidence interval; OS, overall survival; PFS, progression free survival; GSVA, Gene set variation analysis; IHC, Immunohistochemistry; NMF, nonnegative matrix factorization; KM, Kaplan-Meier; HPV, Human papillomavirus; GEO, Gene Expression Omnibus; ssGSEA, single sample Gene set enrichment analysis; LASSO, least absolute shrinkage and selection operator; DCA, decision curve analysis; BP, biological processes; CC, cell composition; MF, molecular function.

training set and three validation sets, we found the prognosis of patients in the low-risk group was much better than that in the high-risk group. A nomogram incorporating the gene signature and T stage was constructed, and the calibration plot suggested that it could accurately predict the survival rate. The expression levels of FNDC3A, VEGFA, OPN3 and CPE were all high in cervical cancer tissues. Downregulation of FNDC3A, VEGFA or CPE expression suppressed the proliferation, migration and invasion of SiHa cells.

Conclusions: Two m5C modification subtypes of CC were identified and then a 4-gene signature was established, which provide new feasible methods for clinical risk assessment and targeted therapies for CC.

Keywords: cervical cancer, m5C modification, signature, prognosis, TCGA

INTRODUCTION

It is estimated that 310,000 people die of CC every year worldwide, CC is fourth most common cause of cancer-related death in women and constitutes a major public health problem (Bray et al., 2018; Arbyn et al., 2020). Every 2 min, one woman dies of CC (Knaul et al., 2019). Human papillomavirus (HPV) infection is a major risk factor for CC, with approximately 90% of cases occurring in low-income and middle-income countries lacking organized screening and HPV vaccination programs (Lagheden et al., 2018; Cohen et al., 2019). For underdeveloped countries, the scarcity of resources and infrastructure limits disease prevention and treatment plans, even no prevention and treatment options are available in some areas. Patients with CC often have social difficulties, constipation, diarrhea, severe lymphedema, menopausal symptoms and major financial problems (Cohen et al., 2019). So it is necessary to improve the diagnosis and treatment methods which need to show cost-effective patient-centered improvements compared with the current strategies (Seol et al., 2014). At present, the conventional treatment of CC includes radiotherapy, chemotherapy and surgery. However, patients with advanced-stage disease are prone to resistance to radiotherapy and chemotherapy. Although immunotherapy is becoming an effective adjuvant therapy, most therapeutic vaccines are still in the early experimental stage (Alldredge and Tewari, 2016). Therefore, it is urgent to determine new prognostic indicators and treatment options to improve the survival rate of patients with CC.

In recent years, the epigenetic modification of RNA has become a focus of research; the dynamic regulation and disturbance of these RNA modifications are also significantly related to the occurrence, maintenance and progression of tumors (Han et al., 2020). RNA contains several dynamic modifications, including N6-methyladenosine, 5-methylcytosine and N7-methylguanosine (Schumann et al., 2020; Song et al., 2020; Tang et al., 2021). m5C existing in mRNAs, tRNAs, rRNAs and ncRNAs, is involved in RNA stability and translation efficiency (Squires et al., 2012). Currently, 13 regulatory factors are involved in the process of m5C methylation. The dynamic modification of m5C is regulated by writers (methyltransferase), readers (binding protein), and erasers (demethylase) (Chen et al., 2019). "Writer" complexes, including NOP2, NSUN2, NSUN3, NSUN4, NSUN5, NSUN6, NSUN7, DNMT1, DNMT3A, DNMT3B and TRDMT1, increase methylation at the RNA C5 site (Bohnsack et al., 2019). "Reader" protein ALYREF could recognize and bind to methylated RNA, and "Eraser" protein TET2 could change the modification of m5C by demethylation (Yang et al., 2017). However, the function and molecular mechanism of m5C-related regulators in CC remain unknown.

In this study, we classified CC subtypes according to these 13 currently reported m5C regulatory factors, further explored DEGs in different CC subtypes, and finally identified a 4-gene signature that could predict the prognosis of CC patients.

MATERIALS AND METHODS

Data Download and Preprocessing

Patients with no survival time available and follow-up time of less than 1 month or more than 120 months were excluded, then mRNA data and clinical information of 257 CC patients were downloaded from the TCGA database. The clinical statistical information of the TCGA all set is shown in Additional file 1: **Supplementary Table S1**. Another dataset GSE44001 consisting of 300 CC patient with associated prognostic information was obtained from the Gene Expression Omnibus (GEO) database.

Determination of m5C Modification Subtype

First, we extracted 13 m5C regulatory factors from the TCGA expression matrix. Based on consistent clustering of these 13 genes, 257 CC samples were clustered by the "NMF" method, which was used to select the standard "brunet" option for 50 iterations. The number of clusters k was set at 2 to 10 and the average contour width of the common member matrix was determined by the "NMF" package. The minimum member of each subclass was set to 10. According to cophenetic, rss and silhouette, the optimal number of clusters was determined. KM analysis was used to analyze the difference in prognosis between different subtypes of the patients using the "survival" package and heatmaps were drawn using the "pheatmap" package.

Assessment of Immune Infiltration

In order to identify the immune infiltration differences between different m5C modification subtypes, we used "MCPcounter"

package to evaluate the score of 10 immune cells and the score of 28 immune cells were evaluated by the "single sample Gene set enrichment analysis (ssGSEA)" method in the "gene set variation analysis (GSVA)" package (Charoentong et al., 2017). Besides, we analyzed the mRNA level differences of 13 m5c-related genes between different subtypes.

Identification and GO/KEGG Annotation of m5c Subtype-Related Differentially Expressed Genes

The "limma" package was used to calculate the DEGs between different m5C modification subtypes, and the filter was applied according to the thresholds FDR <0.05 and $|\log_2 FC| > \log_2(1.5)$. Furthermore, 601 upregulated DEGs and 113 downregulated DEGs were analyzed by the "WebGestalt" package for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation.

Construction of a Novel Gene Signature Based on m5c Subtype-Related DEGs

Under the premise there is no preference in the distribution of clinical characteristics of the grouped samples, a total of 257 patients in TCGA all set were randomly divided into training set (n = 128) and test set (n = 129). The TCGA training set and TCGA test set were evaluated by Chi-square test, the sample information is shown in Additional file 1: **Supplementary Table S2**. In addition, we also used the TCGA all set (n = 257) and GSE44001 set (n = 300) as validation set for subsequent verification. In the training set, a univariate Cox regression analysis was conducted by the "survival coxph function" package using the 714 DEGs and survival data, $p < 0.01$ was selected as the threshold for filtering. Next, we used the "glmnet" package for the least absolute shrinkage and selection operator (LASSO) regression to further compress the screened genes to reduce the number of genes, and finally a novel gene signature was established. LASSO retains the advantages of subset shrinkage, and is a biased estimate for processing data with multicollinearity. Based on the LASSO regression results, we developed a prognostic risk score formula, which was calculated as follows:

$$Risk\ score\ (\text{patient}) = \Sigma i\ Coefficient\ (\text{mRNAi}) \times Expression\ (\text{mRNAi})$$

Validation of the Gene Signature

We calculated the risk score of each sample depending on the signature gene and drew the risk score distribution of the sample in the TCGA training set. Furthermore, we used the "timeROC" package to perform receiver operating characteristic (ROC) analysis to explore the prediction accuracy of 1 year, 3 years and 5 years survival rates. Finally, we calculated the risk score and divided the samples with risk score greater than zero into the high-risk group and samples with risk score less than zero into the low-risk group to draw KM curves. To determine the robustness of the signature, we used the same coefficient to perform the same analysis used for the TCGA test set, TCGA all set and external validation data set GSE44001.

Gene Set Variation Analysis

We selected the corresponding gene expression profiles of these samples for ssGSEA via the "GSVA" R package to observe the relationship between the risk score and the KEGG pathway. We calculated the score of each sample in different KEGG pathway and obtained the ssGSEA score of each sample. Next, we calculated the correlation between these pathways and the risk score and selected pathways with a correlation greater than 0.4. Finally, the top 18 KEGG pathways were selected and clustered according to their enrichment score.

Univariate and Multivariate Cox Analysis of the Signature and Construction of a Nomogram

To identify the independence of the gene signature in clinical parameters, we used univariate and multivariate Cox regression to analyze the hazard ratio (HR), 95% confidence interval (CI) of HR and p value in the TCGA all set. We systematically analyzed the clinical information of TCGA patient records, including age, T stage, N stage, FIGO stage, grade, chemotherapy and risk score. According to the results of univariate and multivariate Cox analyses, we constructed a nomogram with the T stage and risk score for predicting survival outcomes (1 year, 3 years and 5 years). Then we performed the calibration curve by the "rms" package to determine the consistency between the actual survival rates and the nomogram-predicted rates. In order to evaluate the reliability of the nomogram, we performed DCA (decision curve analysis) using the "rmda" package. DCA analysis is a method that can assess whether the nomogram improves clinical decision-making. This method can tell us whether it is beneficial to use the model to make clinical decisions, or which of the two models will lead to better decisions.

Tissue Specimens

Fresh adjacent normal tissues and CC tissues were obtained from the Chinese Academy of Medical Sciences and the CAMS & PUMC Medical College. All patients were not treated preoperatively and signed informed consent forms provided by the Cancer Hospital, CAMS & PUMC. The normal surgical margin tissue and the morphology of the primary tumor area were immediately excised from each patient by an experienced pathologist and stored in liquid nitrogen. The study was approved by the Ethics Committee of the Cancer Institute (Hospital), CAMS & PUMC (20/207–2,403).

Cell Culture and Transfection

The human CC cell line SiHa was provided by the Cell Resource Center, IBMS, CAMS/PUMC. The cell lines were cultured in DMEM medium supplemented with 10% fetal bovine serum (Invitrogen, San Diego, CA) at 37°C and 5% CO2 in a humidified incubator. Human specific siRNA sequences are shown in Additional file 1: **Supplementary Table S3**. The

transfection method was described in our previous article (Cao et al., 2018).

Real-Time Quantitative Polymerase Chain Reaction

Total RNA was extracted using RNApure Tissue & Cell Kit (Cwbiotech, Beijing, China). Isolated RNA was used as a template for reverse transcription reactions using HiFiScript cDNA Synthesis Kit (Cwbiotech, Beijing, China). Real-time quantitative PCR analysis was performed using SYBR® Fast qPCR Mix (TaKaRa, Shiga, Japan) and a CFX96 Real-Time System (Bio-Rad, California, United States of America). The primer sequences are shown in Additional file 1: **Supplementary Table S4**. GAPDH served as the internal control.

Western Blotting and Immunohistochemistry Analysis

Western blotting and IHC analysis were performed as described previously (Cao et al., 2018). The antibodies used were as follows: anti-FNDC3A antibody (Abcam, Cambridge, United Kingdom), anti-VEGFA antibody (Proteintech, Wuhan, China), anti-OPN3 antibody (Affinity Biosciences, Cincinnati, United States), and anti-CPE antibody (Proteintech, Wuhan, China). The IHC quantization analysis was calculated by ImageJ software and statistically analyzed in three random fields. Data are shown as mean ± SEM.

Cell Viability Assays

Cells were inoculated into 96-well plates at a concentration of 2000 cells per well. According to the manufacturer's directions, cell viability was determined by the Cell Counting Kit-8 (CCK-8, Dojindo, Japan). The absorbance was measured at 450 nm by an automatic microplate reader (BioTek, Winooski, United States). Measurements were taken every 24 h for seven consecutive days.

Colony Formation Assay

SiHa cells treated with siRNA were plated in 6-well plates at a density of 500 cells per well. After overnight incubation, the cells were cultured for 14 days to form colonies, fixed with methanol and stained with crystal violet. The data represent the mean ± SD of three independent experiments.

Cell Migration and Invasion Assays

700 μL DMEM medium supplemented with 20% serum was added to the lower chamber of the Transwell plates, and 1×10^5 suspended cells were added to the upper compartment. For the invasion experiment, 50 μL Matrigel was added to the membrane of the upper chamber. The Transwell plates were incubated in a carbon dioxide incubator for 16 h in the migration experiment and for 24 h in the invasion experiment. Then, we removed the chamber, washed the cells once with PBS, fixed the cells with solution (methanol: acetone = 1:1) for 30 min, and then stained them with 0.5% crystal violet for 30 min. The chamber was washed with PBS, and then the upper cells were carefully removed, sealed with neutral gum and photographed for counting.

Statistical Analysis

R software 3.5.3 and SPSS 22.0 software (SPSS Inc. Chicago, United States) were used for all statistical analyses. $p < 0.05$ was taken as the probability value to establish statistical significance. Chi-square test was used for statistics of multiple categories, Student's t-test was used to determine the significance of differences between two groups, and ANOVA was used for comparisons among more than two groups.

RESULTS

Determination of m5C Modification Subtype

To clearly illustrate the process of our research, a flow chart is shown in **Figure 1**.

We extracted mRNA levels of 13 m5C regulatory factors from the expression matrix of the TCGA. Then, 257 CC samples were clustered by "NMF" package. The optimal number of clusters was determined according to cophenetic, rss and silhouette analyses, the optimal number of clusters was 2 (**Figures 2A,B**). The expression levels of m5C methylation-related genes in the C1 and C2 subtypes were significantly different (**Figure 2C**). The KM curve revealed that overall survival (OS) rates of the C1 and C2 subtypes were significantly different ($p < 0.05$), and the prognosis of the C1 group was worse than that of the C2 group (**Figure 2D**).

Immune Infiltration Analysis of m5C Modification Subtype

Due to the significant difference in the prognosis of CC patients with two m5C modification subtypes, we next explored the difference in immune cell infiltration between C1 and C2 subtypes. The ssGSEA score suggested that levels of activated CD8 T cells, central memory CD4 T cells, CD56 bright natural killer cells, macrophages, MDSCs and neutrophils were markedly different between two subtypes, and the MCPcounter score indicated that the infiltration levels of CD8 T cells, NK cells, neutrophils, endothelial cells and fibroblasts were significantly different (**Figures 3A,B**). Furthermore, we also analyzed the expression of 13 genes between two subtypes. The expression of eight genes in C1 and C2 subtypes was substantially different; but no difference was found in NOP2, NSUN4, NSUN7, TRDMT1 and DNMT3A expression (**Figure 3C**). The above results indicated that there are significant differences in the immune infiltration of C1 and C2 subtypes, while the expression of most m5C regulatory factors in the two is also different.

Screening DEGs Between m5C Subtypes and Functional Analysis

DEGs between C1 and C2 subtypes consisted of 601 upregulated and 113 downregulated genes. The volcanic map of representative DEGs is represented in **Figure 4A**. Detailed information about

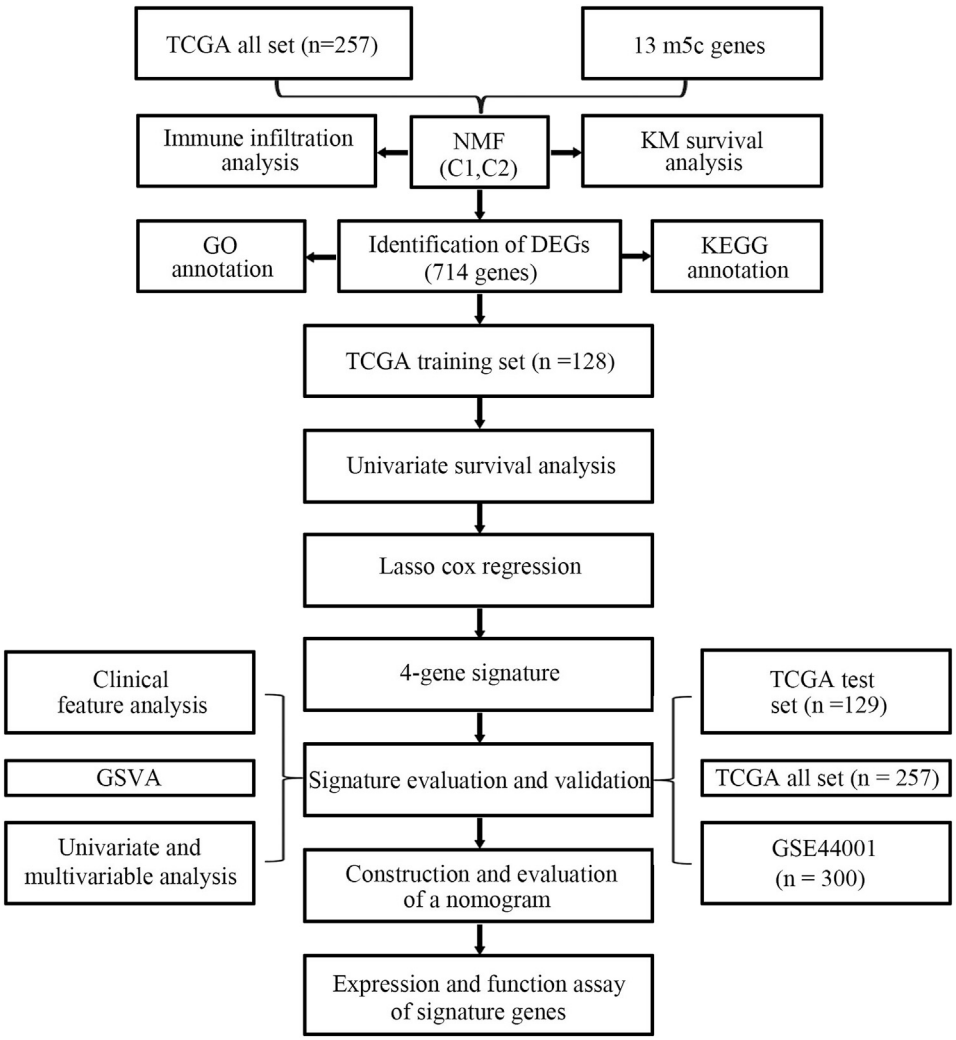

FIGURE 1 | Flow chart for the research.

these DEGs is represented in Additional file 2: **Supplementary Table S5**. We selected 50 genes with the most prominent changes in expression (upregulated and downregulated), as shown in the heat map (**Figure 4B**).

Next, 601 upregulated DEGs were analyzed by GO function and KEGG pathway enrichment annotation. For the GO function analysis of the upregulated DEGs, 493 gene ontologies were annotated to biological process, 88 to cellular component and 81 to molecular function with significant differences ($p < 0.05$). The top 10 annotations are shown in **Figures 4C–E**. For the KEGG pathway enrichment analysis of upregulated DEGs, the top 10 KEGG pathways annotated are shown in **Figure 4F**, including adherens junction, ECM-receptor interaction, amoebiasis, focal adhesion, human papillomavirus infection, PI3K-Akt signaling pathway, and pathways in cancer. More detailed information can be found in Additional file 3: **Supplementary Table S6**. For CC downregulated DEGs, the results of GO function and KEGG pathway enrichment analysis are shown in Additional file 4: **Supplementary Table S7**.

Construction and Evaluation of the Gene Signature in the TCGA Training Set

First, 257 samples in the TCGA-CC dataset were divided into a training set and a test set. The training set consisted of 128 samples, and the test set consisted of 129 samples. The statistical results showed that our groups had no preference, and there was no significant difference between the training set and the test set (Additional file 1: **Supplementary Table S2**).

Using the training set data, a univariate Cox proportional hazards regression analysis was conducted by the "survival coxph function" package for DEGs (714 genes) and the survival data, and $p < 0.01$ was selected as the threshold for filtering. Finally, 27 genes were selected, and the univariate Cox analysis results are shown in **Figure 5A**. Next, we used LASSO regression to further

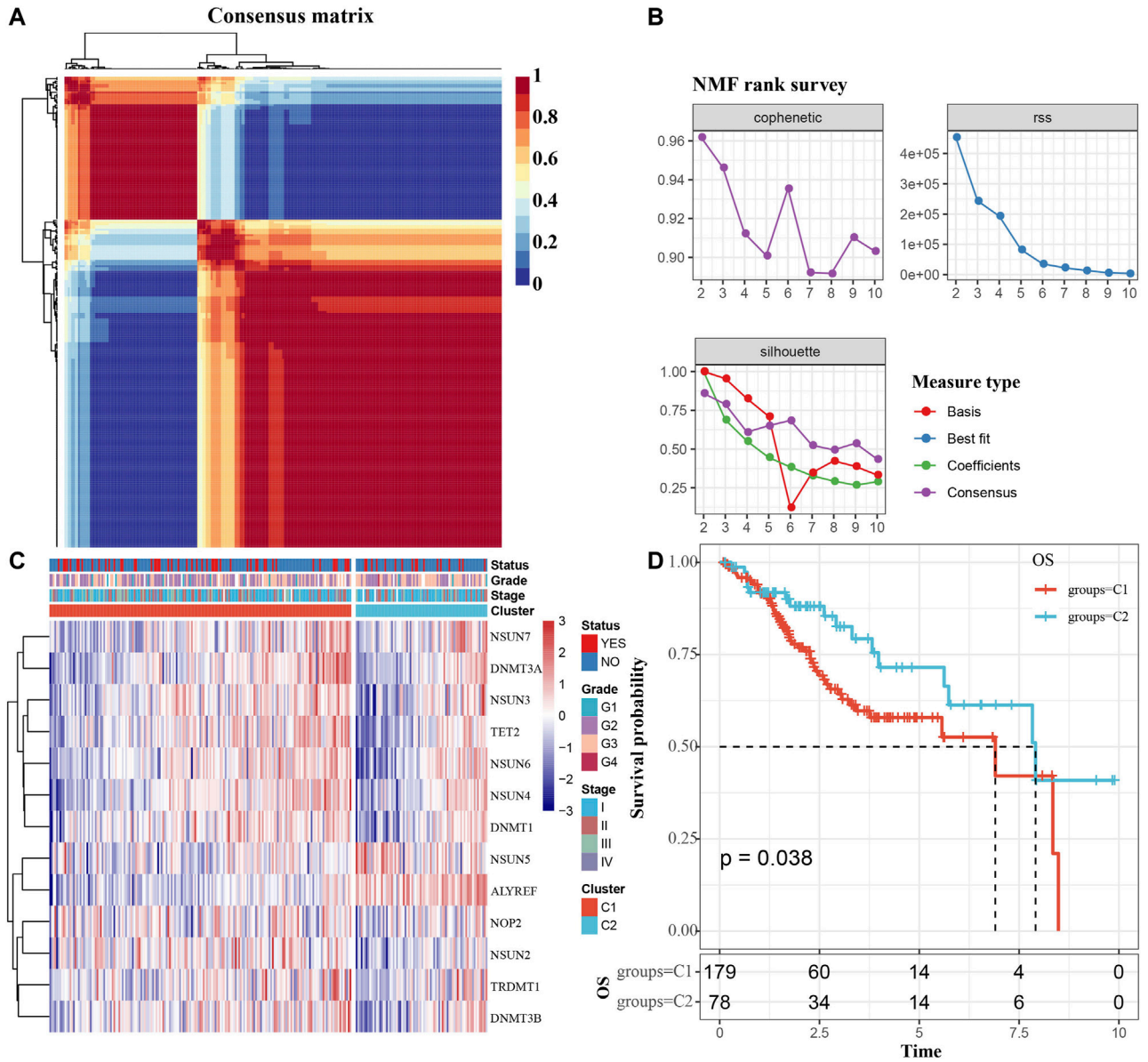

FIGURE 2 | Determination of m5C modification subtype. **(A)** Consensus map for NMF clustering. **(B)** The cophenetic, rss and silhouette distribution when the number of clusters k was set as 2 to 10. The cophenetic correlation is used to reflect the stability of cluster obtained from NMF, we select k value where the magnitude of the cophenetic correlation coefficient begins to fall. The "rss" refers to residual sum of squares, a smaller value represents the effect of model clustering. Silhouette is used to study the distance between clusters of clustering results, When the silhouette value is closer to 1, the clustering is more reasonable. According to cophenetic, rss and silhouette, the optimal number of clusters was two under comprehensive consideration. **(C)** m5c methylation-related gene cluster maps. **(D)** KM survival curves of m5C modification subtypes. Time: years.

compress these 27 genes to reduce the number of genes in the risk signature. We performed 10-fold cross validation to construct the model and analyzed the confidence intervals under each lambda, as shown in **Figure 5B**. We selected the following four final genes with lambda = 0.09,009,939: FNDC3A, VEGFA, OPN3 and CPE.

The KM curves all showed that a higher mRNA level of those four genes indicated worse prognosis in the TCGA training set (Additional file 5: **Supplementary Figure S1**, $p < 0.05$). The final 4-gene signature formula was as follows:

Risk score = 0.3,250,335*FNDC3A (mRNA level) + 0.2,821,988*VEGFA (mRNA level) + 0.3,133,706*OPN3(mRNA level) + 0.1,857,458*CPE (mRNA level).

We calculated the risk score of each sample according to the mRNA level of the signature gene in the training set, and the proportion of deaths in the high-risk group was significantly higher than in the low-risk group. This demonstrated that the risk score is a critical prognostic factor. Consequently, with the increase in risk score, the

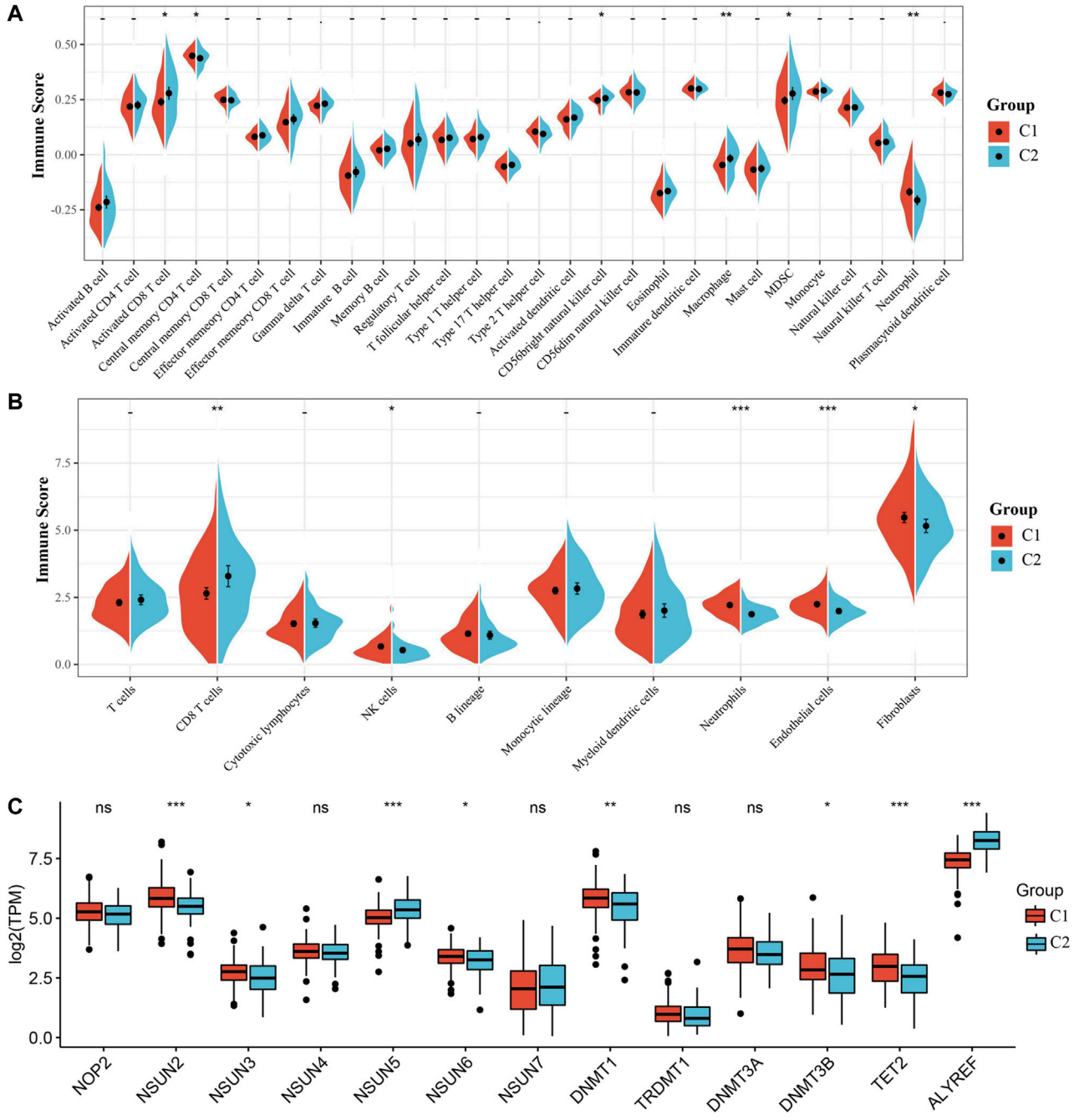

FIGURE 3 | Immune infiltration analysis of the m5C modification subtype. **(A)** Comparison of ssGSEA immune score between C1 and C2 subtypes in the TCGA all set. **(B)** Comparison of MCPcounter immune score between two subtypes in the TCGA all set. **(C)** Expression differences of 13 genes related to m5C modification between two subtypes.

mRNA levels of FNDC3A, VEGFA, OPN3 and CPE were upregulated (**Figure 5C**). To investigate the diagnostic accuracy of the risk signature, the AUC of the time-dependent receiver operating characteristic (ROC) curves was computed. The AUC values of the signature for predicting 1 year, 3 years and 5 years survival rates were 0.74, 0.76 and 0.80 (**Figure 5D**). The KM curve suggested that patients with higher risk score had worse prognosis than those with lower risk score (**Figure 5E**, $p < 0.001$).

Validation of the 4-Gene Signature

To determine the robustness of the model, we used the TCGA test set ($n = 129$) with the same model and coefficient as the training set to calculate the risk score of each sample and drew the risk

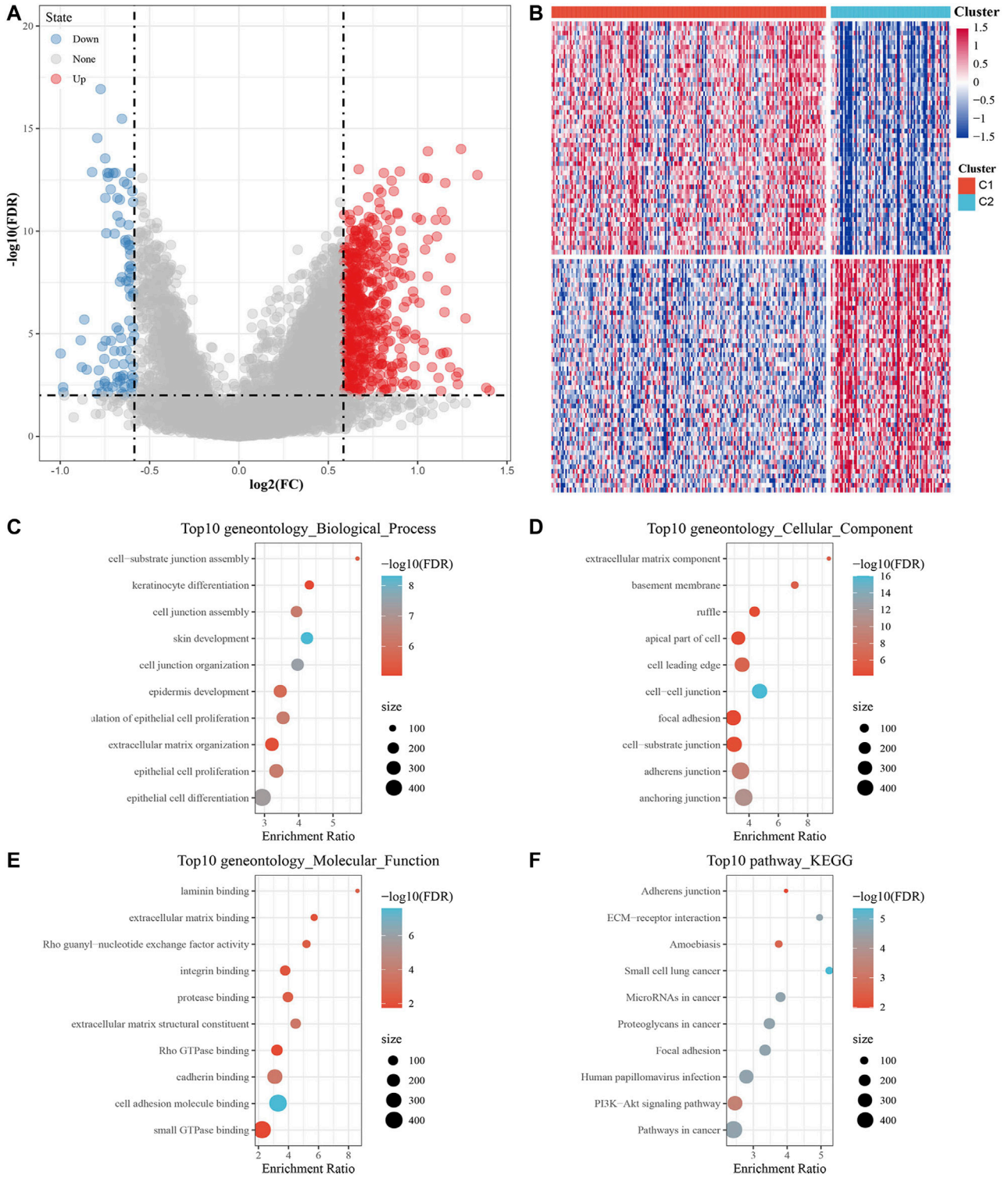

FIGURE 4 | Screening DEGs between m5C subtypes and functional analysis. **(A)** Volcano map of DEGs between the C1 and C2 subtypes. **(B)** Heat map of DEGs between the C1 and C2 subtypes. **(C–E)** The results of GO enrichment analysis of the upregulated DEGs are shown by bubble chart: **(C)** biological processes, **(D)** cell composition and **(E)** molecular function. **(F)** KEGG enrichment analysis of upregulated DEGs.

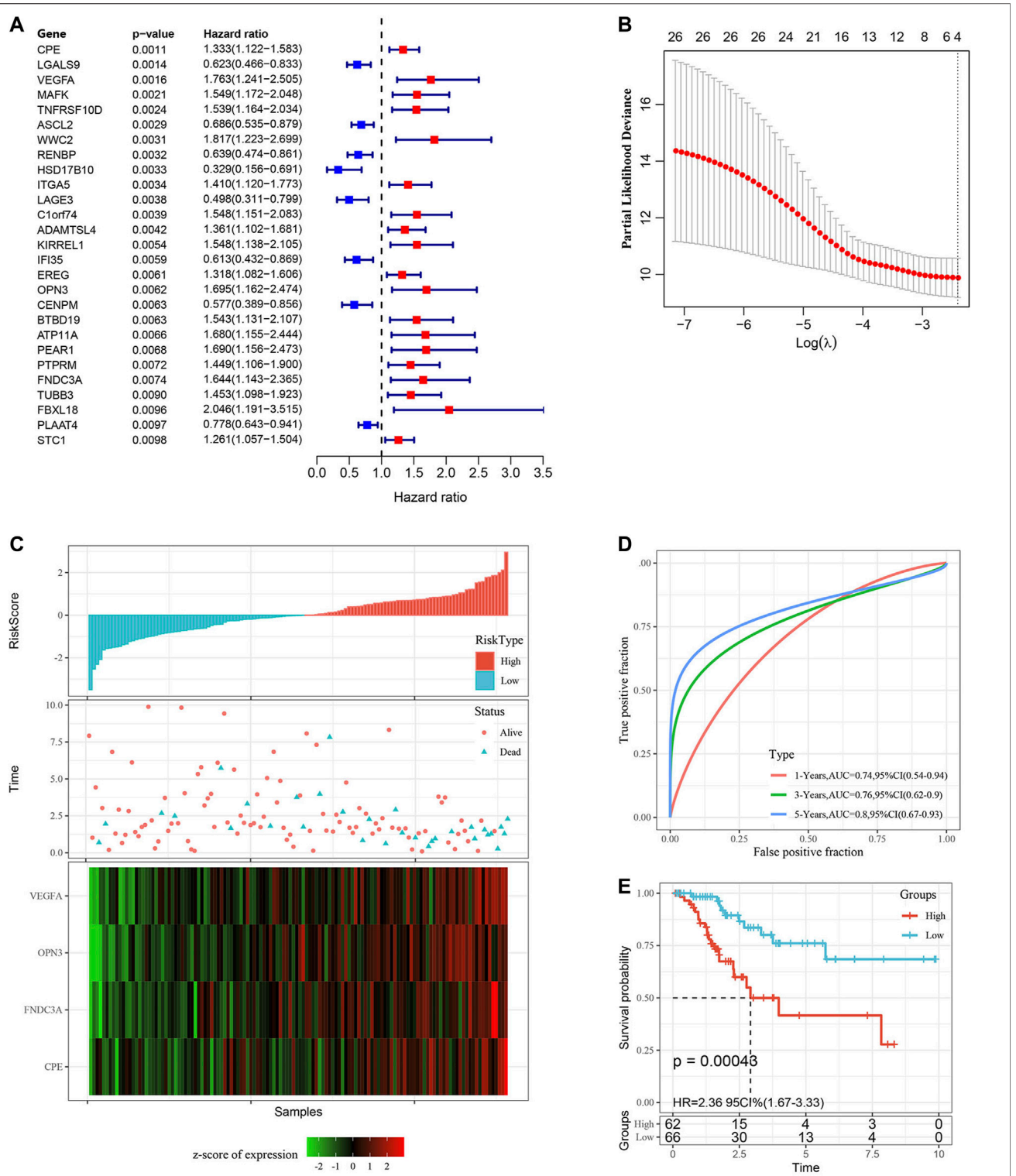

FIGURE 5 | Construction and evaluation of the gene signature in the TCGA training set. **(A)** 27 DEGs were identified by univariate Cox analysis. **(B)** LASSO Cox regression. **(C)** Risk score distribution in the TCGA training set. **(D)** ROC curves were used to assess the efficiency of the risk signature for predicting 1 year, 3 years and 5 years survival rates in the TCGA training set. **(E)** The KM survival curves of the low-risk group and the high-risk group in the TCGA training set. Time: years.

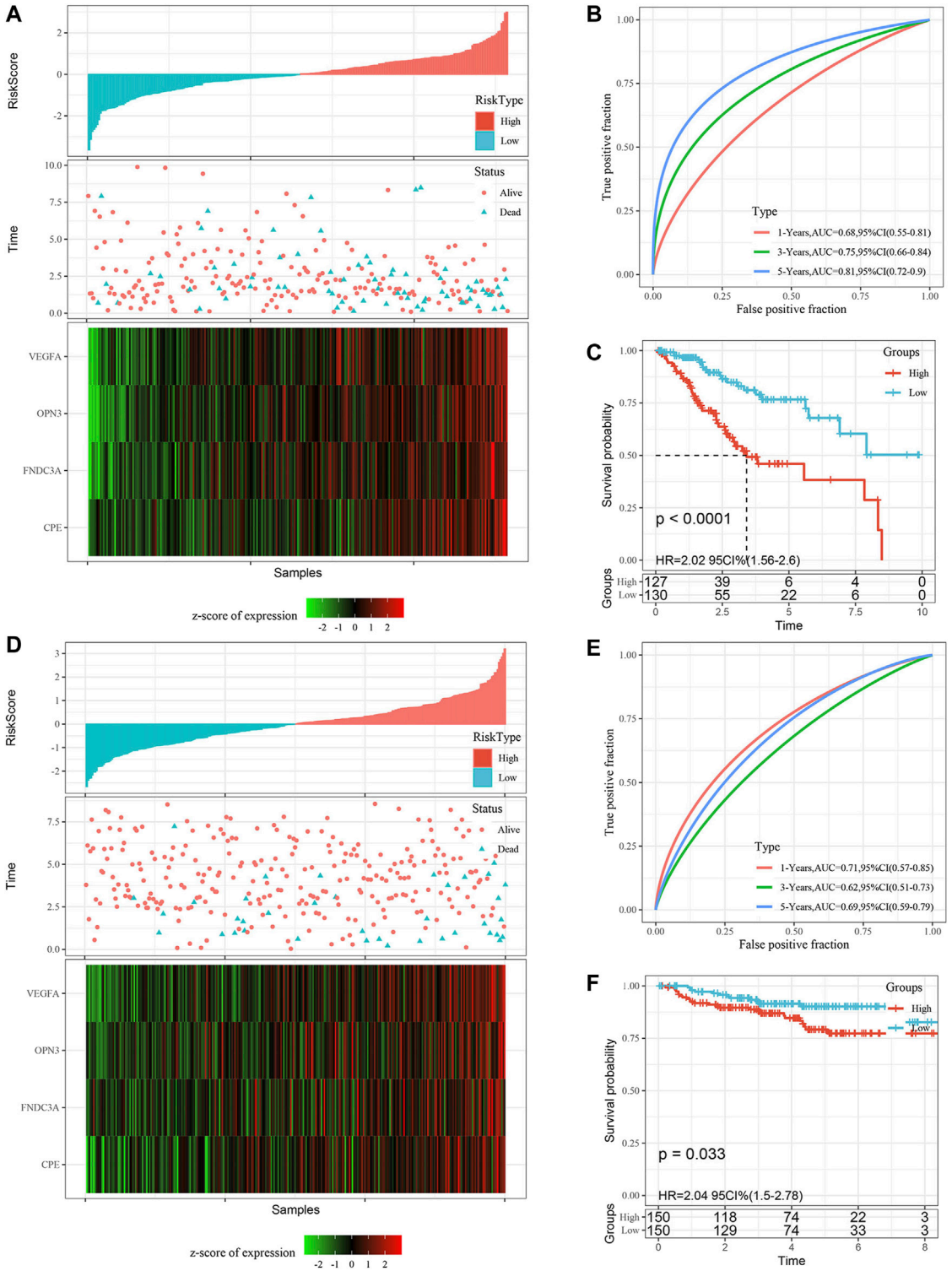

FIGURE 6 | Validation of the gene signature in the TCGA all set and the GSE44001 external validation set. **(A)** Risk score distribution in the TCGA all set. **(B)** ROC curves were used to assess the efficiency of the gene signature for predicting 1 year, 3 years and 5 years survival rates in the TCGA all set. **(C)** The KM survival curves of the low-risk group and the high-risk group in the TCGA all set. **(D)** Risk score distribution in the validation set GSE44001, the survival time of validation set GSE44001 is progression free survival (PFS) time. **(E)** ROC curves were used to assess the efficiency of the gene signature for predicting 1 year, 3 years and 5 years survival rates in the validation set GSE44001. **(F)** The KM survival curves of the low-risk group and the high-risk group in the validation set GSE44001. Time: years.

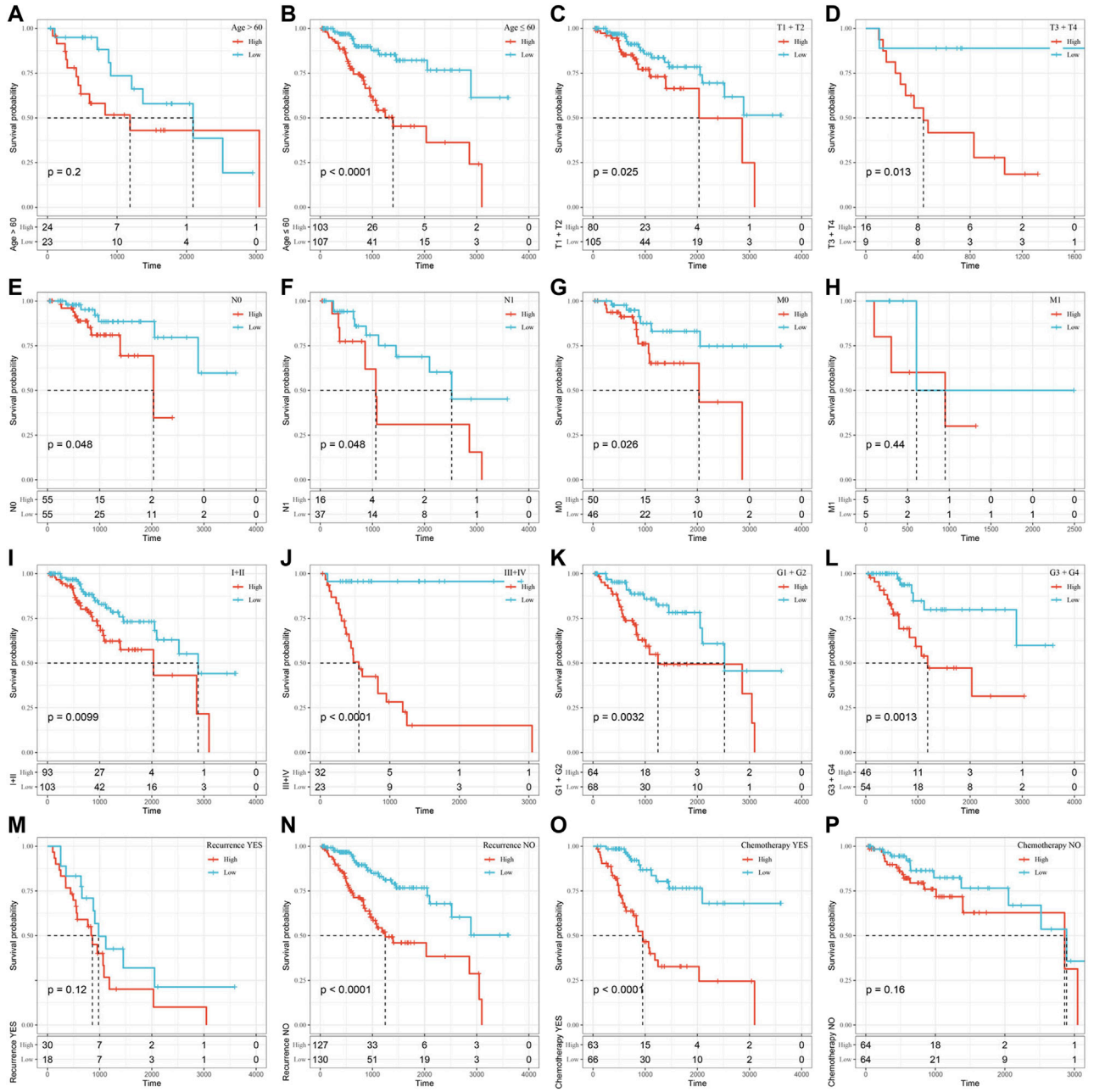

FIGURE 7 | Assessment of the OS rate of the high-risk and low-risk groups based on different clinical subgroups. **(A)** Age>60 subgroup. **(B)** Age≤60 subgroup. **(C)** T1+T2 stage subgroup. **(D)** T3+T4 Stage subgroup. **(E)** N0 Stage subgroup. **(F)** N1 Stage subgroup. **(G)** M0 Stage subgroup. **(H)** M1 Stage subgroup. **(I)**I+II Stage subgroup. **(J)** III+IV Stage subgroup. **(K)** G1+G2 subgroup. **(L)** G3+G4 subgroup. **(M)** Recurrence_ Yes subgroup. **(N)** Recurrence_No subgroup. **(O)** Chemotherapy _Yes subgroup. **(P)** Chemotherapy_No subgroup. Time: days.

score distribution of the patients. Then, we performed ROC analysis on the prognostic classification of the risk score and analyzed the classification efficiency of 1 year, 3 years and 5 years survival rates. Finally, we divided the samples with risk score into high-risk group (n = 65) and low-risk group (n = 64) to perform the KM survival analysis. Importantly, the results of the above analysis were consistent with the performance of the TCGA training set (Additional file 6: **Supplementary Figure S2).**

In addition, we determined the robustness of the signature in the TCGA all set (**Figures 6A–C**) and the external validation dataset GSE44001 (**Figures 6D,E**). The proportion of deaths in the high-risk group was significantly higher than in the low-risk group, which was consistent with the performance of the TCGA training set. The AUC values of the signature showed that the risk score was a good prognostic factor. Finally, the KM curve also

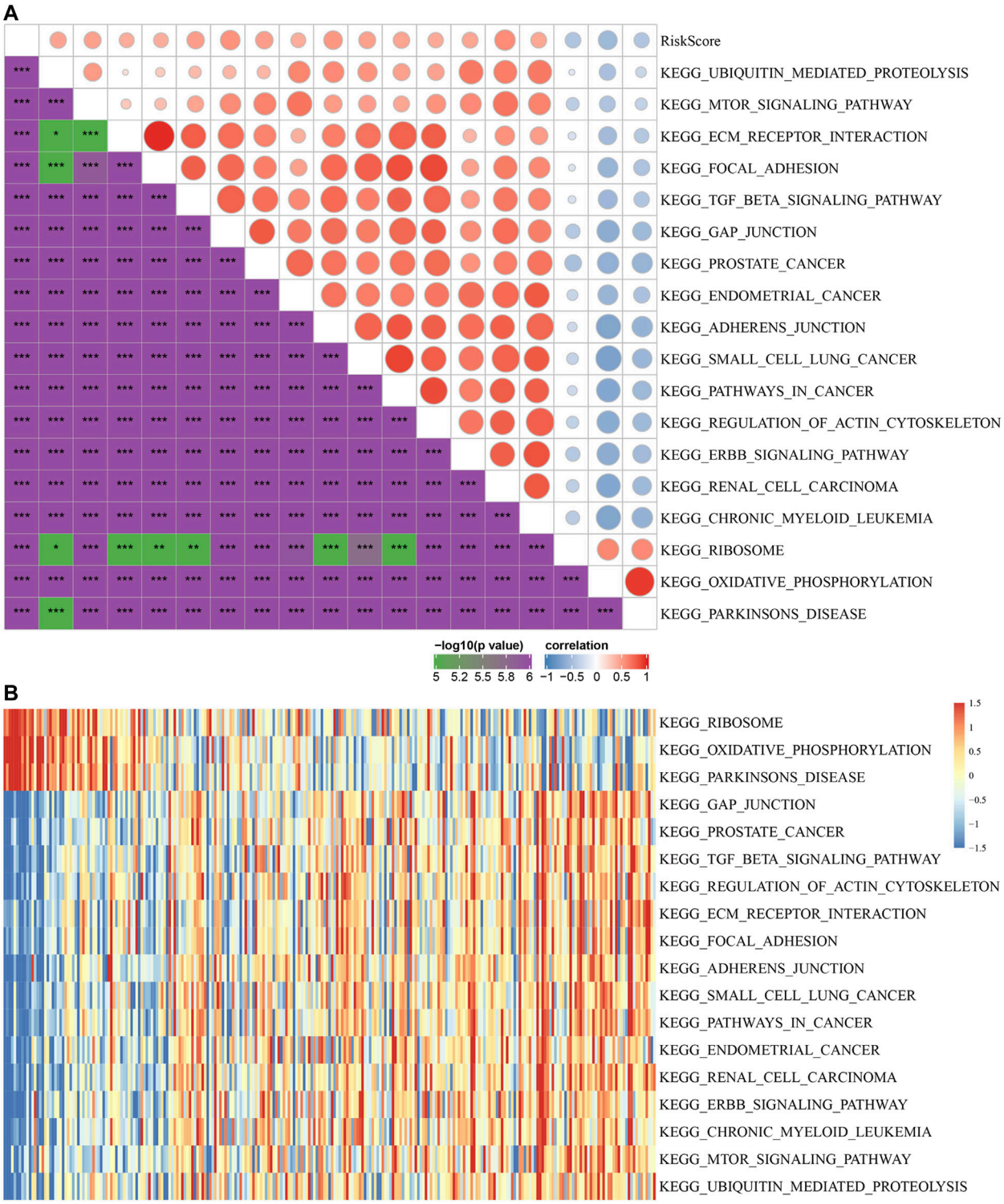

FIGURE 8 | The relationship between risk score and KEGG pathways. **(A)** Correlation coefficient clustering between KEGG pathways and risk score with a risk score correlation greater than 0.4. **(B)** GSVA revealed KEGG pathways associated with the risk score. The horizontal axis represents the sample, and the risk score increases from left to right.

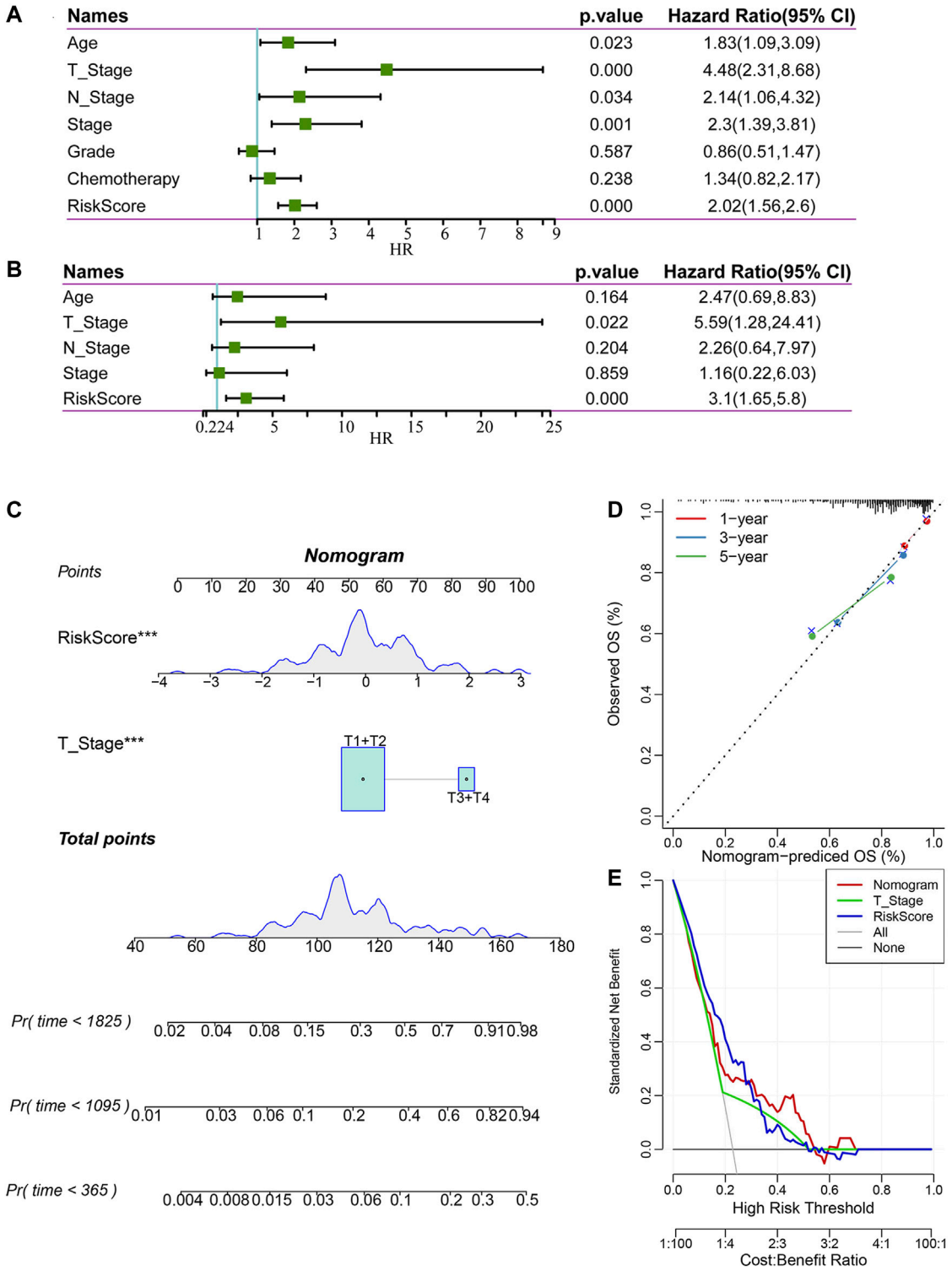

FIGURE 9 | Construction and evaluation of a nomogram. **(A)** Univariate Cox regression analysis of clinical characteristics and risk score. **(B)** Multivariate Cox analysis of clinical characteristics and risk score. **(C)** A nomogram for predicting the 1 year, 3 years and 5 years survival rates of CC patients was established. **(D)** 1 year, 3 years, and 5 years survival rate calibration curves of the line chart. **(E)** The DCA of the nomogram.

showed that there were consistent differences between the high-risk and low-risk groups.

Assessment of the OS Rate of the High-Risk and Low-Risk Groups Based on Different Clinical Subgroups

Furthermore, we performed KM survival analysis according to age, grade, stage, recurrence and chemotherapy treatment in the TCGA all set. For the patients in age < = 60, T1+T2 Stage, T3+T4 Stage, N0 Stage, N1 Stage, M0 Stage,I+II Stage, III+IV Stage, G1+G2, G3+G4, Recurrence _ No and Chemotherapy _ Yes subgroup, the OS interval of patients in the high-risk group was significantly shorter than that of patients in the low-risk group ($p < 0.05$). Only in the age >60, M1, Recurrence _ Yes or Chemotherapy _ No subgroup was the OS interval of patients not different between the high- and low-risk groups ($p > 0.05$). The above findings further showed that the risk signature still has good predictive ability among different clinical subgroups (**Figure 7**).

The Relationship Between Risk Score and KEGG Pathway

To observe the relationship between risk score and the KEGG pathway, we selected the gene expression profiles corresponding to these samples for ssGSEA using the "GSVA" R package. The score of each sample in different KEGG pathway were calculated, and the ssGSEA score of each KEGG pathway corresponding to each sample were obtained. The correlation between these pathways and risk score was further calculated, and the function with correlation greater than 0.4 was selected (**Figure 8A**). Fifteen pathways were positively correlated with the risk score, and three pathways were negatively correlated with the risk score. The top 18 KEGG pathways were selected and clustered according to their enrichment score, as shown in **Figure 8B**. We can find that KEGG_MTOR_SIGNALING_PATHWAY, KEGG_ECM_RECEPTOR_INTERACTION, KEGG_FOCAL_ ADHESION, KEGG _TGF_BETA_SIGNALING_PATHWAY, KEGG_ADHERENS_JUNCTION, KEGG_ PATHWAYS_IN_ CANCER and other tumor-related pathways were activated with increasing risk score.

Construction and Evaluation of a Nomogram

To identify the independence of the 4-gene signature in clinical parameters, we used univariate and multivariate Cox regression to analyze the related HR, 95% CI of HR and *p* value in the clinical information of the TCGA all set. The clinical data of patients were analyzed systematically, including age, T stage, N stage, FIGO stage, grade, chemotherapy and risk score (**Figures 9A,B**). In the TCGA all set, the risk score was an independent prognostic factor. Therefore, the 4-gene signature has good predictive performance and clinical application value.

According to the results of univariate and multivariate Cox regression analyses, we constructed a nomogram with clinical features, T stage and risk score (**Figure 9C**). We found that the risk score had the greatest impact on survival prediction, indicating that the risk score is indispensable in the nomogram. Furthermore, we used the calibration curve to evaluate the prediction accuracy of the signature, as shown in **Figure 9D**. The prediction calibration curves of the three calibration points for 1 year, 3 years and 5 years survival rates were close to the standard curves, indicating that the signature had good prediction performance. In addition, DCA showed that the benefits of the risk score and nomogram were significantly higher than those of the extreme curves. The nomogram curve was higher than that of the risk score, which indicated that the nomogram had good reliability (**Figure 9E**).

Signature Gene Expression Was Upregulated in Cervical Cancer

In order to explore the difference in protein expression of signature genes, we used IHC to detect 6 normal and 21 tumor tissues. The IHC quantization analysis was calculated by ImageJ software and statistically analyzed in three random fields. The results showed that the protein expression levels of the four signature genes in tumor were higher than those in normal (**Figure 10A**). At the same time, we conducted qRT-PCR experiments in 10 normal and 12 tumor tissues to explore the differences in the transcription level of those four genes. It showed that the mRNA levels of the four model genes were significantly increased in tumor tissues, indicating that the expression of model genes in tumor tissues may be abnormally activated (**Figure 10B**).

FNDC3A, VEGFA or CPE Promoted the Proliferation, Invasion and Migration of SiHa Cells

To clarify the functional role of FNDC3A, VEGFA, OPN3 and CPE in CC cells, we applied human-specific siRNA to decrease their protein expression (**Figure 11A**). The CCK-8 assay was applied to detect cell proliferation. The downregulation of FNDC3A, VEGFA or CPE expression significantly suppressed the proliferation (**Figure 11B**) and colony formation capacity of SiHa cells (**Figure 11C**). Transwell assays were applied to detect the invasion and migration ability of SiHa cells *in vitro*, and the number of cells that passed through the polycarbonate membrane was smaller in the FNDC3A, VEGFA or CPE siRNA group than in the negative control group, indicating that FNDC3A, VEGFA or CPE could significantly promote the invasion and migration of SiHa cells (**Figure 11D**). Downregulation of OPN3 expression had no significant effect on the proliferation, colony formation ability, invasion and migration of SiHa cells.

DISCUSSION

m5C is common methylation modification in eukaryotic RNA, which can promote the regulation of nuclear mRNA through the methyltransferase NSUN2 and the binding protein ALYREF and participates in the splicing and protein translation process of

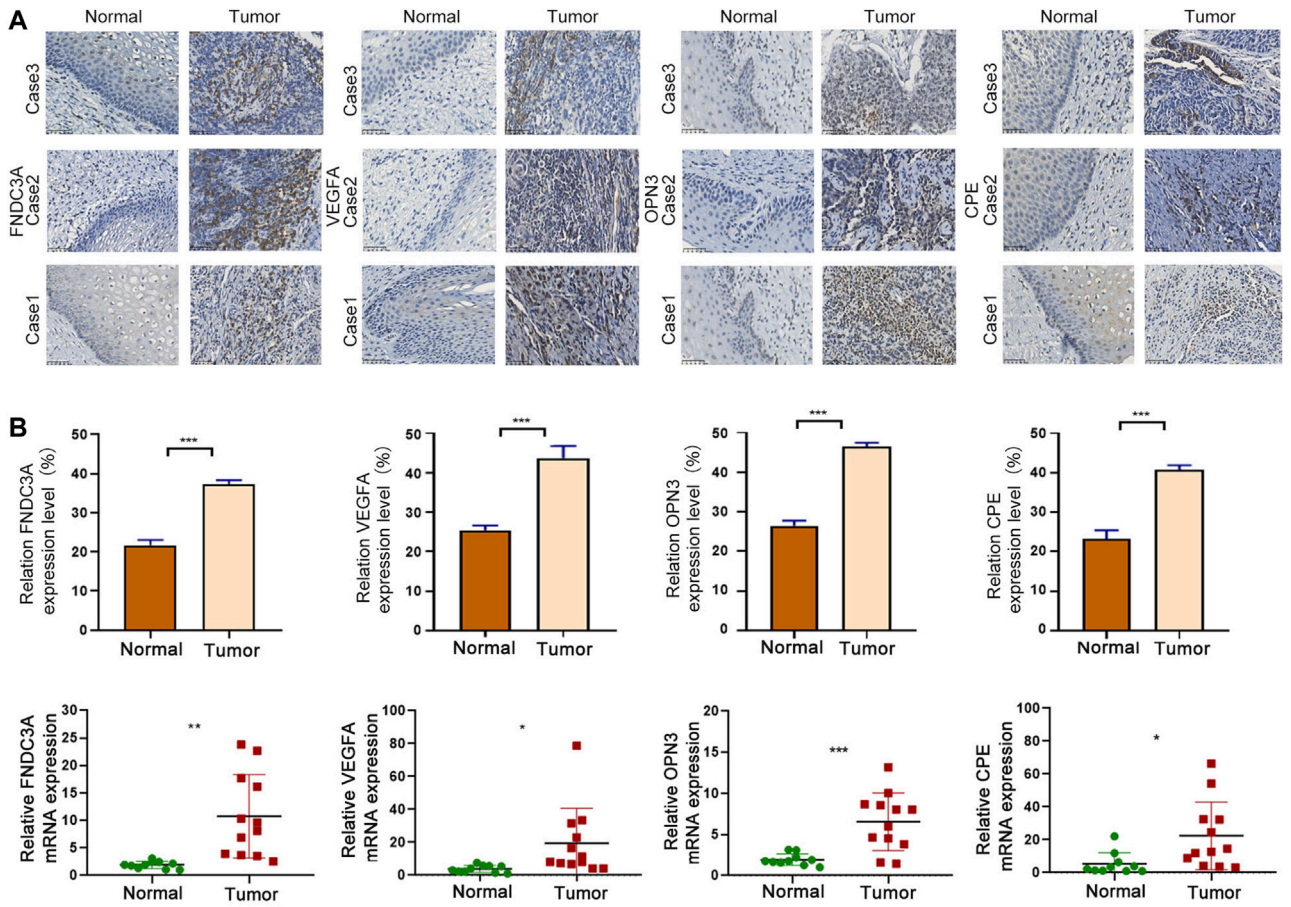

FIGURE 10 | Signature gene expression was upregulated in CC. **(A)** IHC analysis of FNDC3A, VEGFA, OPN3 and CPE protein levels in tumor and normal tissues. Scale bar = 50 μm ***$p < 0.001$. **(B)** qRT-PCR analysis of FNDC3A, VEGFA, OPN3 and CPE mRNA levels in tumor and normal tissues. *, $p < 0.05$. **, $p < 0.01$. ***, $p < 0.001$.

several mRNAs (Yang et al., 2017). It was reported that m5C methylation of the 3′-UTR contributes to an increase in mRNA stability (Zhang et al., 2012); tRNA occurs most frequently on cytosine of the variable arm and C38 of the anticodon ring (Agris, 2008; Trixl and Lusser, 2019), which can maintain the thermal stability of the tRNA secondary structure and improve the recognition ability of codons. Besides, the rRNA of all organisms is modified by m5C, methylation sites of human and yeast 28S rRNA, M5C2870 and M5C2278, are critical for protein translation (Sharma and Lafontaine, 2015). In addition, modification of m5C can also be detected in noncoding RNAs, such as lncRNAs, erRNAs, and vtRNAs (Amort et al., 2013; David et al., 2017).

m5C modification plays an important role in the development of tumors, such as that of NSUN family proteins (RNA m5C methyltransferase). Compared with normal human tissues and cells, the expression of NSUN2 is increased in a variety of tumor tissues, and NSUN2 is considered to be an effective prognostic marker for some cancers, such as squamous cell carcinomas and colon carcinomas (Frye and Watt, 2006). In breast cancer cells, NSUN6 can form a complex with the proteins LLGL2 and lncRNA Maya, which inactivates the kinase Hippo/MST1 through methylation of Hippo/MST1, resulting in promotion of tumor metastasis (Li et al., 2017). However, the role and mechanism of m5C RNA modification in the prognosis of CC have not been studied.

We hypothesized that m5C RNA modification-related genes have broad prospects in the prognostic evaluation of CC. First, we extracted the mRNA levels of 13 m5C regulatory factors from the TCGA expression matrix for clustering and obtained two subtypes of CC, C1 and C2. Next, we identified the differences in the immune infiltration levels between the two molecular subtypes. We screened DEGs between the C1 and C2 subtypes and obtained 601 upregulated genes and 113 downregulated genes. Next, we divided the TCGA all set (257 samples) into a training set and a test set. In the TCGA training set, we used univariate Cox regression and LASSO regression analysis to establish a 4-gene signature comprising FNDC3A, VEGFA, OPN3 and CPE. Then, we conducted risk distribution analysis, ROC, curve analysis and survival analysis in the TCGA training set, the TCGA test set, the TCGA all set and the GSE44001 data set to verify our signature. Furthermore, we

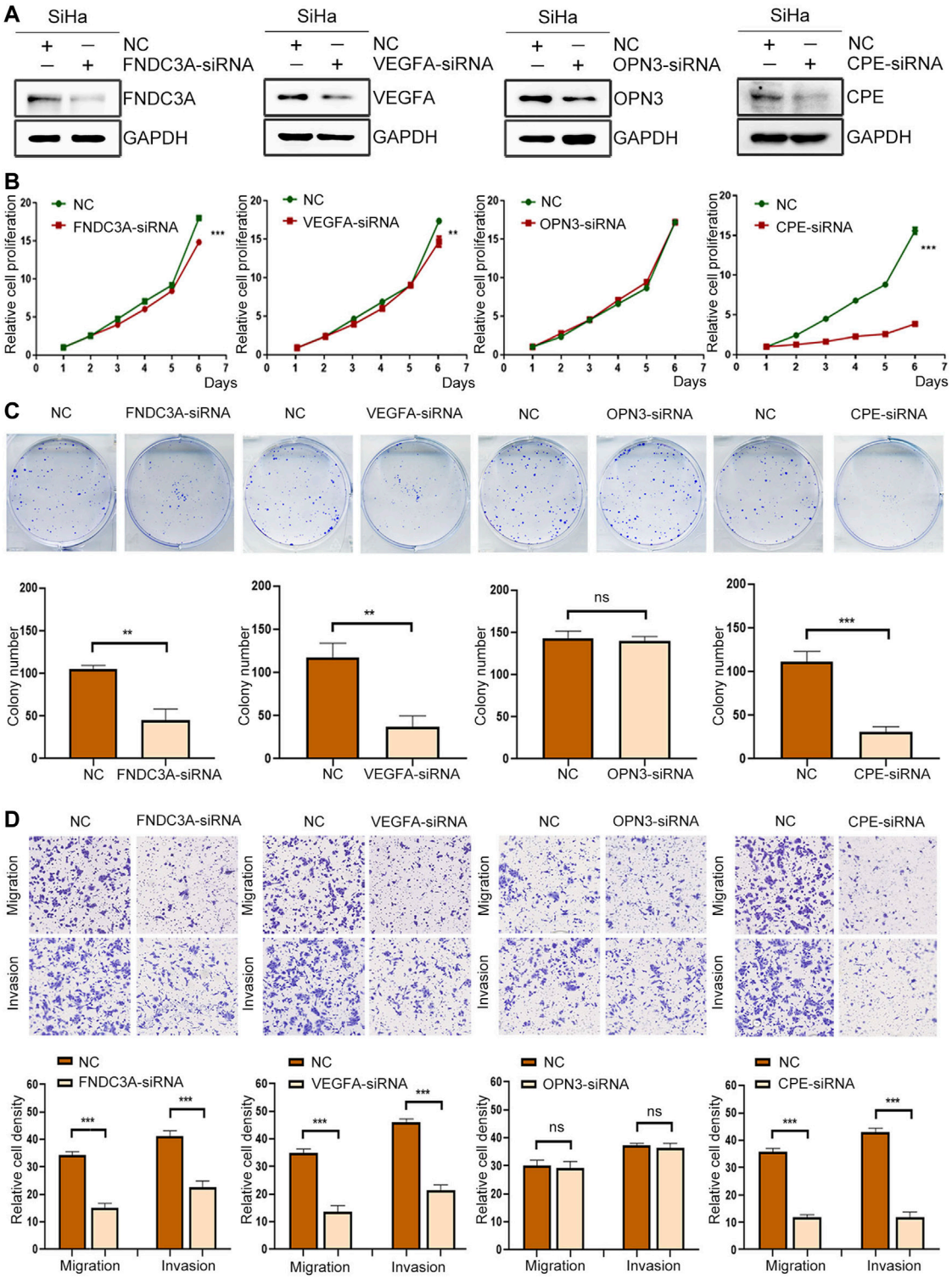

FIGURE 11 | FNDC3A, VEGFA or CPE promoted the proliferation, invasion and migration of SiHa cells. **(A)** Western blotting analysis of FNDC3A, VEGFA, OPN3 or CPE expression in SiHa cells transfected with their siRNA. **(B)** Cell proliferation abilities were detected by CCK-8. **(C)** Colony number were detected. **(D)** Cell migration and invasion abilities were evaluated by Transwell assay. NC: Negative Control. **, $p < 0.01$; ***, $p < 0.001$.

found that in most clinical feature subgroups, the OS interval of patients in the high-risk group was significantly shorter than that of patients in the low-risk group ($p < 0.05$). In addition, we proved that the risk score was an independent risk factor and constructed an effective nomogram to predict the 1 year, 3 years and 5 years survival rates of CC patients.

To explore the function of signature genes and their prognostic correlation in CC patients, we carried out qRT-PCR and IHC experiments and found that the mRNA levels and protein expression of those genes were all higher in cervical cancer tissues than in normal tissues. In addition, downregulation of FNDC3A, VEGFA or CPE expression suppressed the proliferation, migration and invasion of SiHa cells *in vitro*. KM survival analysis showed that high expression of the four hub genes was a risk factor for CC patients. FNDC3A can be used as a prognostic marker for colorectal cancer and is highly expressed in colon cancer tissues, and high expression of FNDC3A increases the mortality rate of colon cancer patients (Meyer et al., 2012; Wuensch et al., 2019). In multiple myeloma, high expression of FNDC3A can lead to ROS accumulation, ATP deficiency and cell death in multiple myeloma cells. VEGFA can be used as a prognostic biomarker for head and neck squamous cell carcinoma, esophageal squamous cell carcinoma, glioblastoma and papillary thyroid carcinoma (He et al., 2020; Stuchi et al., 2020; Yang et al., 2020; Zheng and Tao, 2020). Downregulation of VEGFA expression can inhibit the proliferation, angiogenesis and metastasis of osteosarcoma cells, ovarian cancer and lung squamous cell carcinoma (Chen et al., 2020a; Chen et al., 2020b; Li et al., 2020; Qin et al., 2020). Upregulation of VEGFA expression can promote the proliferation, angiogenesis and metastasis of gastric cancer cells and breast cancer cells (Chen et al., 2020a; Wang et al., 2020). OPN3 can be used as a prognostic biomarker for lung adenocarcinoma. With the increase in OPN3 expression, the mortality rate of lung adenocarcinoma patients increases, and the survival time decreases (Wang et al., 2019). It has been reported that the OPN3 gene enhances the metastasis of lung adenocarcinoma, and its overexpression promotes epithelial-mesenchymal transition (Xu et al., 2020). In lung carcinoids, patients with high OPN3 expression are more likely to experience relapse and metastasis (Miyanaga et al., 2020). In addition, OPN3 can also sensitize liver cancer cells to 5-fluorouracil treatment by regulating the apoptosis pathway (Jiao et al., 2012). In a recent study, CPE was used to predict recurrence of early lung adenocarcinoma (Jones et al., 2021). In addition, CPE expression was upregulated in patients with extranasal nodal natural killer cell/T cell lymphoma (NKTCL) after cytarabine chemotherapy and could be used as a chemotherapy index for NKTCL patients (Gong et al., 2018).

In summary, we developed a novel 4-gene signature based on m5c modification, which had good AUC in the training set and three validation sets. Based on the signature, we constructed an effective nomogram to predict the 1 year, 3 years and 5 years survival rates of CC patients. We suggested using this classifier as a molecular diagnostic test to evaluate the prognostic risk of CC patients. Furthermore, we found that three of the signature genes (FNDC3A, VEGFA or CPE) function as oncogenes to promote the proliferation, invasion and migration of cervical cancer cells and could be potential therapeutic targets for CC. The advantage of this study is that we identified a prognostic 4-gene signature with a relatively high AUC in the training and three validation datasets, which can accurately predict survival rates. Then we explored the expression and function of the signature genes to explore the potential of these genes as therapeutic targets. The limitation of this study is that we should further carry out animal experiments to verify the function of model genes, in addition, the molecular mechanisms of model genes regulating the progression of CC still need to be further explored.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The study was approved by the Ethics Committee of the Cancer Institute (Hospital), CAMS & PUMC. Samples were obtained with written informed consent from all patients, and informed consent was obtained in accordance with the Declaration of Helsinki; Consent for publication; Written informed consent for publication was obtained from all participants.

AUTHOR CONTRIBUTIONS

JY, L-LL, L-YW, and J-SA conceived and designed the study. JY, L-LL, and JL performed the experiments. JY, L-LL, and T-TL analyzed the data and drafted the manuscript. L-YW and J-SA secured financing of the study. JL, LX, JZ, T-TW, DW, L-JL, D-WX, and D-XC contributed to the review and editing. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81972450), the CAMS Innovation Fund for Medical Sciences (CIFMS) (No. 2016-I2M -1- 001), and the National Natural Science Foundation for Young Scholars of China (81802747).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.733715/full#supplementary-material

REFERENCES

Agris, P. F. (2008). Bringing Order to Translation: the Contributions of Transfer RNA Anticodon-domain Modifications. *EMBO Rep.* 9 (7), 629–635. doi:10.1038/embor.2008.104

Alldredge, J. K., and Tewari, K. S. (2016). Clinical Trials of Antiangiogenesis Therapy in Recurrent/Persistent and Metastatic Cervical Cancer. *The Oncologist* 21 (5), 576–585. doi:10.1634/theoncologist.2015-0393

Amort, T., Soulière, M. F., Wille, A., Jia, X. Y., Fiegl, H., Wörle, H., et al. (2013). Long Non-coding RNAs as Targets for Cytosine Methylation. *RNA Biol.* 10 (6), 1003–1008. doi:10.4161/rna.24454

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., et al. (2020). Estimates of Incidence and Mortality of Cervical Cancer in 2018: a Worldwide Analysis. *Lancet Glob. Health* 8 (2), e191–e203. doi:10.1016/s2214-109x(19)30482-6

Bohnsack, K. E., Höbartner, C., and Bohnsack, M. T. (2019). Eukaryotic 5-methylcytosine (m^5C) RNA Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease. *Genes (Basel)* 10 (2), 102. doi:10.3390/genes10020102

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Cao, Y.-Y., Yu, J., Liu, T.-T., Yang, K.-X., Yang, L.-Y., Chen, Q., et al. (2018). Plumbagin Inhibits the Proliferation and Survival of Esophageal Cancer Cells by Blocking STAT3-PLK1-AKT Signaling. *Cell Death Dis* 9 (2), 17. doi:10.1038/s41419-017-0068-6

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cel Rep.* 18 (1), 248–262. doi:10.1016/j.celrep.2016.12.019

Chen, J., Liu, A., Wang, Z., Wang, B., Chai, X., Lu, W., et al. (2020). LINC00173.v1 Promotes Angiogenesis and Progression of Lung Squamous Cell Carcinoma by Sponging miR-511-5p to Regulate VEGFA Expression. *Mol. Cancer* 19 (1), 98. doi:10.1186/s12943-020-01217-2

Chen, J., Li, X., Yang, L., Li, M., Zhang, Y., and Zhang, J. (2020). CircASH2L Promotes Ovarian Cancer Tumorigenesis, Angiogenesis, and Lymphangiogenesis by Regulating the miR-665/VEGFA Axis as a Competing Endogenous RNA. *Front. Cel Dev. Biol.* 8, 595585. doi:10.3389/fcell.2020.595585

Chen, L., Wang, P., Bahal, R., Manautou, J. E., and Zhong, X.-b. (2019). Ontogenic mRNA Expression of RNA Modification Writers, Erasers, and Readers in Mouse Liver. *PLoS One* 14 (12), e0227102. doi:10.1371/journal.pone.0227102

Cohen, P. A., Jhingran, A., Oaknin, A., and Denny, L. (2019). Cervical Cancer. *The Lancet* 393 (10167), 169–182. doi:10.1016/s0140-6736(18)32470-x

David, R., Burgess, A., Parker, B., Li, J., Pulsford, K., Sibbritt, T., et al. (2017). Transcriptome-Wide Mapping of RNA 5-Methylcytosine in Arabidopsis mRNAs and Noncoding RNAs. *Plant Cell* 29 (3), 445–460. doi:10.1105/tpc.16.00751

Frye, M., and Watt, F. M. (2006). The RNA Methyltransferase Misu (NSun2) Mediates Myc-Induced Proliferation and Is Upregulated in Tumors. *Curr. Biol.* 16 (10), 971–981. doi:10.1016/j.cub.2006.04.027

Gong, Y., Pu, W., Jin, H., Yang, P., Zeng, H., Wang, Y., et al. (2018). Quantitative Proteomics of CSF Reveals Potential Predicted Biomarkers for Extranodal NK-/T-cell Lymphoma of Nasal-type with Ethmoidal Sinus Metastasis. *Life Sci.* 198, 94–98. doi:10.1016/j.lfs.2018.02.035

Han, X., Wang, M., Zhao, Y. L., Yang, Y., and Yang, Y. G. (2020). RNA Methylations in Human Cancers. *Semin. Cancer Biol.* S1044-579X (20), 30241–30248. doi:10.1016/j.semcancer.2020.11.007

He, W., Leng, X., Yang, Y., Peng, L., Shao, Y., Li, X., et al. (2020). Genetic Heterogeneity of Esophageal Squamous Cell Carcinoma with Inherited Family History. *Ott* 13, 8795–8802. doi:10.2147/ott.s262512

Jiao, J., Hong, S., Zhang, J., Ma, L., Sun, Y., Zhang, D., et al. (2012). Opsin3 Sensitizes Hepatocellular Carcinoma Cells to 5-fluorouracil Treatment by Regulating the Apoptotic Pathway. *Cancer Lett.* 320 (1), 96–103. doi:10.1016/j.canlet.2012.01.035

Jones, G. D., Brandt, W. S., Shen, R., Sanchez-Vega, F., Tan, K. S., Martin, A., et al. (2021). A Genomic-Pathologic Annotated Risk Model to Predict Recurrence in Early-Stage Lung Adenocarcinoma. *JAMA Surg.* 156 (2), e205601. doi:10.1001/jamasurg.2020.5601

Knaul, F. M., Rodriguez, N. M., Arreola-Ornelas, H., and Olson, J. R. (2019). Cervical Cancer: Lessons Learned from Neglected Tropical Diseases. *Lancet Glob. Health* 7 (3), e299–e300. doi:10.1016/s2214-109x(18)30533-3

Lagheden, C., Eklund, C., Lamin, H., Kleppe, S. N., Lei, J., Elfström, K. M., et al. (2018). Nationwide Comprehensive Human Papillomavirus (HPV) Genotyping of Invasive Cervical Cancer. *Br. J. Cancer* 118 (10), 1377–1381. doi:10.1038/s41416-018-0053-6

Li, C., Wang, S., Xing, Z., Lin, A., Liang, K., Song, J., et al. (2017). A ROR1-HER3-lncRNA Signalling axis Modulates the Hippo-YAP Pathway to Regulate Bone Metastasis. *Nat. Cel Biol* 19 (2), 106–119. doi:10.1038/ncb3464

Li, S., Liu, M., Do, M. H., Chou, C., Stamatiades, E. G., Nixon, B. G., et al. (2020). Cancer Immunotherapy via Targeted TGF-β Signalling Blockade in TH Cells. *Nature* 587 (7832), 121–125. doi:10.1038/s41586-020-2850-3

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and Near Stop Codons. *Cell* 149 (7), 1635–1646. doi:10.1016/j.cell.2012.05.003

Miyanaga, A., Masuda, M., Motoi, N., Tsuta, K., Nakamura, Y., Nishijima, N., et al. (2020). Whole-exome and RNA Sequencing of Pulmonary Carcinoid Reveals Chromosomal Rearrangements Associated with Recurrence. *Lung Cancer* 145, 85–94. doi:10.1016/j.lungcan.2020.03.027

Qin, Y., Zhang, B., and Ge, B. J. (2020). MicroRNA-150-5p Inhibits Proliferation and Invasion of Osteosarcoma Cells by Down-Regulating VEGFA. *Eur. Rev. Med. Pharmacol. Sci.* 24 (18), 9265–9273. doi:10.26355/eurrev_202009_23008

Schumann, U., Zhang, H.-N., Sibbritt, T., Pan, A., Horvath, A., Gross, S., et al. (2020). Multiple Links between 5-methylcytosine Content of mRNA and Translation. *BMC Biol.* 18 (1), 40. doi:10.1186/s12915-020-00769-5

Seol, H.-J., Ulak, R., Ki, K.-D., and Lee, J.-M. (2014). Cytotoxic and Targeted Systemic Therapy in Advanced and Recurrent Cervical Cancer: Experience from Clinical Trials. *Tohoku J. Exp. Med.* 232 (4), 269–276. doi:10.1620/tjem.232.269

Sharma, S., and Lafontaine, D. L. J. (2015). 'View from A Bridge': A New Perspective on Eukaryotic rRNA Base Modification. *Trends Biochem. Sci.* 40 (10), 560–575. doi:10.1016/j.tibs.2015.07.008

Song, B., Tang, Y., Chen, K., Wei, Z., Rong, R., Lu, Z., et al. (2020). m7GHub: Deciphering the Location, Regulation and Pathogenesis of Internal mRNA N7-Methylguanosine (m7G) Sites in Human. *Bioinformatics* 36 (11), 3528–3536. doi:10.1093/bioinformatics/btaa178

Squires, J. E., Patel, H. R., Nousch, M., Sibbritt, T., Humphreys, D. T., Parker, B. J., et al. (2012). Widespread Occurrence of 5-methylcytosine in Human Coding and Non-coding RNA. *Nucleic Acids Res.* 40 (11), 5023–5033. doi:10.1093/nar/gks144

Stuchi, L. P., Castanhole-Nunes, M. M. U., Maniezzo-Stuchi, N., Biselli-Chicote, P. M., Henrique, T., Padovani Neto, J. A., et al. (2020). VEGFA and NFE2L2 Gene Expression and Regulation by MicroRNAs in Thyroid Papillary Cancer and Colloid Goiter. *Genes (Basel)* 11 (9), 954. doi:10.3390/genes11090954

Tang, Y., Chen, K., Song, B., Ma, J., Wu, X., Xu, Q., et al. (2021). m6A-Atlas: a Comprehensive Knowledgebase for Unraveling the N6-Methyladenosine (m6A) Epitranscriptome. *Nucleic Acids Res.* 49 (D1), D134–D143. doi:10.1093/nar/gkaa692

Trixl, L., and Lusser, A. (2019). The Dynamic RNA Modification 5-methylcytosine and its Emerging Role as an Epitranscriptomic Mark. *Wiley Interdiscip. Rev. RNA* 10 (1), e1510. doi:10.1002/wrna.1510

Wang, H., Lu, D., Liu, X., Jiang, J., Feng, S., Dong, X., et al. (2019). Survival-related Risk Score of Lung Adenocarcinoma Identified by Weight Gene Co-expression Network Analysis. *Oncol. Lett.* 18 (5), 4441–4448. doi:10.3892/ol.2019.10795

Wang, X., Che, X., Yu, Y., Cheng, Y., Bai, M., Yang, Z., et al. (2020). Hypoxia-autophagy axis Induces VEGFA by Peritoneal Mesothelial Cells to Promote Gastric Cancer Peritoneal Metastasis through an Integrin α5-fibronectin Pathway. *J. Exp. Clin. Cancer Res.* 39 (1), 221. doi:10.1186/s13046-020-01703-x

Wuensch, T., Wizenty, J., Quint, J., Spitz, W., Bosma, M., Becker, O., et al. (2019). Expression Analysis of Fibronectin Type III Domain-Containing (FNDC) Genes in Inflammatory Bowel Disease and Colorectal Cancer. *Gastroenterol. Res. Pract.* 2019, 3784172. doi:10.1155/2019/3784172

Xu, C., Wang, R., Yang, Y., Xu, T., Li, Y., Xu, J., et al. (2020). Expression of OPN3 in Lung Adenocarcinoma Promotes Epithelial-mesenchymal Transition and Tumor Metastasis. *Thorac. Cancer* 11 (2), 286–294. doi:10.1111/1759-7714.13254

Yang, J., Jiang, Q., Liu, L., Peng, H., Wang, Y., Li, S., et al. (2020). Identification of Prognostic Aging-Related Genes Associated with Immunosuppression and Inflammation in Head and Neck Squamous Cell Carcinoma. *Aging* 12 (24), 25778–25804. doi:10.18632/aging.104199

Yang, X., Yang, Y., Sun, B.-F., Chen, Y.-S., Xu, J.-W., Lai, W.-Y., et al. (2017). 5-methylcytosine Promotes mRNA export - NSUN2 as the Methyltransferase and ALYREF as an m^5C Reader. *Cell Res* 27 (5), 606–625. doi:10.1038/cr.2017.55

Zhang, X., Liu, Z., Yi, J., Tang, H., Xing, J., Yu, M., et al. (2012). The tRNA Methyltransferase NSun2 Stabilizes p16INK4 mRNA by Methylating the 3′-untranslated Region of P16. *Nat. Commun.* 3, 712. doi:10.1038/ncomms1692

Zheng, S., and Tao, W. (2020). Identification of Novel Transcriptome Signature as a Potential Prognostic Biomarker for Anti-angiogenic Therapy in Glioblastoma Multiforme. *Cancers (Basel)* 12 (9), 2368. doi:10.3390/cancers12092368

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 November 2021
doi: 10.3389/fgene.2021.774400

Comprehensive Analysis of the Tumor Microenvironment and Ferroptosis-Related Genes Predict Prognosis with Ovarian Cancer

*Xiao-xue Li, Li Xiong, Yu Wen and Zi-jian Zhang**

Department of General Surgery, The Second Xiangya Hospital of Central South University, Changsha, China

The early diagnosis of ovarian cancer (OC) is critical to improve the prognosis and prevent recurrence of patients. Nevertheless, there is still a lack of factors which can accurately predict it. In this study, we focused on the interaction of immune infiltration and ferroptosis and selected the ESTIMATE algorithm and 15 ferroptosis-related genes (FRGs) to construct a novel E-FRG scoring model for predicting overall survival of OC patients. The gene expression and corresponding clinical characteristics were obtained from the TCGA dataset (n = 375), GSE18520 (n = 53), and GSE32062 (n = 260). A total of 15 FRGs derived from FerrDb with the immune score and stromal score were identified in the prognostic model by using least absolute shrinkage and selection operator (LASSO)–penalized COX regression analysis. The Kaplan–Meier survival analysis and time-dependent ROC curves performed a powerful prognostic ability of the E-FRG model via multi-validation. Gene Set Enrichment Analysis and Gene Set Variation Analysis elucidate multiple potential pathways between the high and low E-FRG score group. Finally, the proteins of different genes in the model were verified in drug-resistant and non–drug-resistant tumor tissues. The results of this research provide new prospects in the role of immune infiltration and ferroptosis as a helpful tool to predict the outcome of OC patients.

Keywords: ovarian cancer, tumor infiltrating immune cells, ferroptosis, prognostic, the cancer genome atlas

OPEN ACCESS

Edited by:
Zodwa Dlamini,
University of Pretoria, South Africa

Reviewed by:
Binbin Wang,
National Cancer Institute, National Institutes of Health (NIH), United States
Ke Han,
Harbin University of Commerce, China

****Correspondence:***
Zi-jian Zhang
178211082@csu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *11 September 2021*
Accepted: *18 October 2021*
Published: *15 November 2021*

Citation:
Li X-x Xiong L, Wen Y and Zhang Z-j (2021) Comprehensive Analysis of the Tumor Microenvironment and Ferroptosis-Related Genes Predict Prognosis with Ovarian Cancer. Front. Genet. 12:774400. doi: 10.3389/fgene.2021.774400

INTRODUCTION

Ovarian cancer (OC) is one of the most serious gynecological diseases and the second most common gynecological disease that causes female deaths worldwide, seriously endangering women's health and safety (BRAY et al., 2018). Surgery combined with chemotherapy and targeted immunotherapy have greatly improved the survival rate of OC patients in recent times (CORRADO et al., 2019; LI et al., 2019). However, the survival rate of OC has not changed even in developed countries, such as the United States (GIAMPAOLINO et al., 2019). This is mainly because about 70% of OC patients are already at an advanced stage once diagnosed and have lost the opportunity for radical surgery, so the 5-year survival rate is only 30% (STEWART et al., 2019). Therefore, from a long-term perspective, the prognosis of patients always depends on early diagnosis and prevention of recurrence in OC. Although the diagnosis of OC has been developed in recent decades, the prediction of diagnosis and prognosis in OC patients is still unsatisfactory. The recognized risk factors for OC include genetic risk, obesity, age, and the use of perineal talcum powder

(PENNINKILAMPI and Eslick, 2018; LHEUREUX et al., 2019). But these factors are not yet considered a good source of help in predicting the prognosis of patients. Therefore, it is urgent to establish new biomarker model for OC diagnosis and prognosis prediction.

Tumor infiltrating immune cells (TIICs) are composed of a variety of cells in the tumor microenvironment, such as T cells, macrophages, neutrophils, and stromal cells (CHEW et al., 2012). Immunotherapy has a certain effect on OC, and the antitumor effect of infiltrating T lymphocytes in OC has already been observed (ZAMARIN, 2019). Recent studies have confirmed that the infiltration of immune-related cells and the abnormal expression of certain genes in these cells may be related to the occurrence and development of tumors, so these factors can also be used to predict the prognosis of OC (BACI et al., 2020; JIANG et al., 2020; NOWAK and KLINK, 2020). Similarly, stromal cells in tumors are also involved in tumor growth and drug resistance regulation, such as tumor-associated fibroblasts (YEUNG et al., 2016; De NOLA et al., 2019). The analysis and evaluation of immune cells and stromal cells can help us get a deeper understanding of the relationship between the tumor microenvironment (Corn et al., 2020) and prognosis of OC patients and help develop a reliable prognostic and predictive model.

Ferroptosis was proposed by DIXON in 2012 (DIXON et al., 2012). It is a new kind of programmed cell death that occurs after ferrous ions catalyze the formation of lipid peroxides. More and more evidences show that ferroptosis plays an important regulatory role in the occurrence and development of liver cancer (ZHANG et al., 2019), gastric cancer (LEE et al., 2020), and OC (CARBONE and MELINO, 2019; LIN and CHI, 2020; WANG et al., 2021). It can not only instruct the research of antitumor drugs but can also be used for OC biomarker screening. More importantly, ferroptosis-mediated iron ions, amino acid, reactive oxygen species, and lipid metabolism are closely related to the tumor immune microenvironment (Friedmann Angeli et al., 2019). At present, there is no report about the prognostic evaluation of the immune-stromal score combined with ferroptosis in OC patients, which has potential research value. In this article, we downloaded samples of OC patients from the public data sets TCGA and GEO. After standardizing the data, we constructed a prediction model of the immune-stromal score combined with ferroptosis-related genes through the TCGA training set and verified it in the TCGA training set and two validation sets of GEO. In addition, we analyzed the immune matrix infiltration of OC, ferroptosis gene co-expression network, and potential pathways. Compared with previous studies, this study has better credibility and more comprehensive data.

MATERIALS AND METHODS

Data Source and Pre-processing

The gene expression data of the OC patients were downloaded from the TCGA and Gene Expression Omnibus (GEO) databases. The GEO datasets should fulfill the criteria that expression profiles were detected by array or high-throughput sequencing and contained corresponding clinical information, including age, histologic type, and overall survival (OS), at least. We obtained 375 samples in total with gene expression profiles from the TCGA OV dataset. In GEO datasets we collected GSE18520 which contained 53 OC samples and GSE32062 which contained 260 samples. OS refers to the time interval from the date of the patient's first diagnosis to death. Among these datasets, we applied the TCGA OC dataset as the training set and the GEO datasets (GSE18520 and GSE32062) as validation sets. Affymetrix Human Genome U133 Plus 2.0 Array GPL570 served as the microarray platform for the GEO datasets, and the coding genes of patients with missing values were excluded. According to the annotation file provided by the platform, we annotated the microarray probe set to the gene name one by one. Then, log2 (Affy RMA) was used to unify the gene expression values. For genes containing several probes, the average value represented the expression value of the gene. For quality control and standardization during the data analysis, we used R package limma.

The Analysis of Tumor-Infiltrating Immune Cells in Ovarian Cancer

First, the immune infiltration of TCGA patients was evaluated. R package "CIBERSORT" was applied to acquire the standardized abundance index of immune cells. Then, we used the LM22 gene signature and CIBERSORT algorithm to define 22 immune cells sensitively and specifically. The immune cells included the B cell family (naïve B cells, memory B cells, and plasma cells), T cell family ($CD8^+$ T cells and naïve $CD4^+$ T cells), M0 macrophages, M1 macrophages, resting NK cells, activated NK cells, resting dendritic cells, activated dendritic cells, resting mast cells, and monocytes). CIBERSORT is a deconvolution algorithm based on support vector regression and non-negative matrix factorization. We downloaded the reference gene expression value standard file LM22 corresponding to various immune cell types and called the corresponding package R script to calculate the proportion of different immune cells. Among them, LM22 is an expression matrix containing 547 genes, which serves as a standard control for distinguishing the proportion of immune cells. The main components of normal cells in tumor tissues are stromal cells and immune cells. These cells interfere with tumor signals in mechanism research and help promote tumor immune escape. By using expression data and R package "ESTIMATE", the ESTIMATE algorithm could estimate the abundance of stromal cells and immune cells in tumors and speculate the ESTIMATE scores.

Ferroptosis-Related Gene Acquisition

We collected 275 FRGs from the ferroptosis-related dataset FerrDb while removing the excess genes (http://www.zhounan.org/ferrdb). Regarding the 275 FRGs, we extracted a total of 244 genes which coexisted in TCGA-OV, GSE18520, and GSE32062 datasets. Then, the 244 genes were applied for the construction of a prognostic model.

The Construction of the E-FRG Score Model

Through the "glmnet" package in R, we applied the least absolute shrinkage and selection operator (LASSO) regression analysis to determine the best weighting coefficient of FRGs in predicting OC prognosis. LASSO is an improvement of the least squares analysis. The core is to use penalty terms and regularization methods for statistical modeling and suppress overfitting. The best value of the penalty coefficient λ with the smallest partiality deviation was determined by running the lambda-min test, which gives the smallest cross-validation error. Therefore, the following formula was applied to speculate the E-FRGs score of each sample: E-FRGs score = ∑expgenei*βi. Expgenei represents the immune score, stromal score, and expression of the genes and βi represents the optimal coefficient for each factor included.

Validation of the Estimate and Ferroptosis-Related Genes-Score Model

We applied the TCGA OC cohort as a training set for evaluating ESTIMATE and ferroptosis-related gene (E-FRG) score models, which contained 375 samples with integral gene expression profiles and necessary patient information. Then, we applied the GSE18520 and GSE32062 datasets to verify the predictive effect of the E-FRG score of OC. According to the critical value in TCGA, OC patients were divided into higher or lower E-FRG score groups. We choose the median as the critical value and use the median to verify the robustness of the model in the GSE18520 and GSE32062 datasets. Then we used Kaplan–Meier analysis to predict OS of OC patients while employing univariate and multivariate Cox regression analyses to determine independent prognostic factors between genes and other scores in E-FRGs in the training set. The receiver operating characteristic (ROC) curve was used to verify the preciseness and predictive ability of E-FRGs. The area under the curve (AUC) values of each risk model were calculated to determine the optimal risk signature. When the maximum AUC value was reached, the calculation procedure was terminated. The predictive ability of the risk signature for 1-/3-/5-year OS was assessed using the "survivalROC" R package. All patients were then categorized into the high-risk and low-risk groups based on the cutoff value identified in the training set. Kaplan–Meier (K-M) survival curves along with the log-rank test were used to identify differences in OS between the two groups using the R packages "survival" and "survminer".

The Analysis of Coexpression Gene Network

The "WGCNA" software package in R was used and weighted coexpression gene network analysis on FRGs in TCGA OC was performed. The main functions of WGCNA include clustering analysis of genes and calculating the association between FRGs and phenotypes. First, the correlation coefficients were calculated between the genes and determined gene modules. Then, a coexpression gene network was constructed, and the association between clinical features and gene modules was determined. Then, we used the integrated capability in WGCNA software to set the soft threshold power β to 10. FRGs were classified into four modules which illustrated the analogous expression modes on the basis of the hybrid dynamic cutting tree. We applied the cluster dendrograms to show the consequence of gene merging and classification. Finally, through Pearson's correlation, the correlation between different modular genes and clinical characteristics has been evaluated.

Clinical Tissues and Western Blotting Assays

We randomly collected a dozen human OC fresh tissues from patients undergoing surgery after obtaining their consent at The Second Xiangya Hospital of Central South University. Then, we analyzed the condition of these patients after receiving cisplatin treatment to select five specimens from cisplatin-sensitive patients (S1-S5) and five specimens from cisplatin-resistant patients (R1-R5). The experiment was approved by the Human Ethics Committee of The Second Xiangya Hospital of Central South University. Then, cancer cell lines in 6-cm dishes and tissues were lysed using RIPA buffer to obtain protein samples. Then, the samples were centrifuged at 12,000 g for 5 min at 4°C in a 1.5-ml tube, and the cell extract was transferred to a new tube. Then, 5*sodium dodecyl sulfate (SDS)–loading buffer was added and heated for 5 min at 95°C. The electrophoresis conditions were 120 V and 50 min. After electrophoresis, a polyvinylidene difluoride (PVDF) membrane was used for protein transfer at conditions of 400 mA at 45 min. The PVDF membrane was blocked using 5% skim milk diluted with Tris-buffered saline Tween 20 (TBST). The membrane was incubated with the following primary antibodies SLC7A5 (1: 1,000, #13752-1-AP, Proteintech, United States), ACSL4 (1: 2000, # 22401-1-AP, Proteintech, United States), XCT (1:2000, #26864-1-AP, Proteintech, United States), GPX4 (1:2 000, #14432-1-AP, Proteintech, United States), ALOX5 (1:2000, # 10021-1-Ig, Proteintech, United States), STEAP3 (1:4,000, # 17186-1-AP, Proteintech, United States), ZFP36(1:1,000, # 12737-1-AP, Proteintech, United States), GABARAPL1 (1: 2000, #11010-1-AP, Proteintech, United States), NRAS (1: 3,000, #10724-1-AP, Proteintech, United States) and beta actin (1:1,000, #20536-1-AP, Proteintech). After overnight incubation, the membrane was labeled using horseradish peroxide (HRP)–conjugated secondary antibody. Then, we washed the PVDF membrane and applied the electrochemiluminescence (ECL) luminescent solution to develop the pictures.

Cell Culture and Cell Viability

OC cell lines, namely, OVCAR3 and SKOV3 were obtained from Sciencell (California, United States) and maintained in RPMI 1640 medium supplemented with 10% fetal bovine serum. To establish CDDP-resistant cells, namely, OVCAR3/CDDP and SKOV3/CDDP, the cells were treated with cisplatin (#HY-17394, MedChemExpress, China) in a stepwise manner from 0.2 to 2 μg/ml. Then, the cells were transferred to a cisplatin-free medium for three days before the beginning of the experiments to reduce the influence of cisplatin. The drugs and their concentrations involved in cell viability, GSH, and MDA

assays included cisplatin (2.2 or 6.9 μg/ml), erastin (10 μM, #S7242, Selleck, United States), RSL3 (0.1 μM, #S8155, Selleck, United States) and acetylcysteine (N-acetyl-l-cysteine, NAC) (1 mM, #S1623, Selleck, United States), while their treatment time was 24 h. The cells were seeded on 96-well plates at a density of 1.5×10^4cells/ml and cultured 24 h before drug treatment. Then, cell viability was detected by using the CellTiter Blue® reagent (Promega, G8082, United States). A water bath maintained at 37°C was employed to remove the reagent, and 20 μl/well of the CellTiter Blue® Reagent was added. Then, the plates were incubated in standard cell culture conditions for 2 h, and the fluorescence intensity was measured at 560/590 nm.

Glutathione Assay and Lipid Peroxidation Assay

The cells were treated with 5% 5-sulfosalicylic acid (SSA) solution. Then, the GSH level was determined by using the reduced GSH Assay kit (#K464-100, BioVision) according to the manufacturer's protocol. The results were normalized to total protein concentration for each sample. Lipid peroxidation could be detected by the combination of malondialdehyde (MDA) with thiobarbituric acid (TBA) through the MDA assay kit (#MAK085, Sigma). The cells were treated in MDA lysis buffer and then centrifugated at 13,000 g for 3 min. The absorbance wavelength of products formed by MDA and TBA was measured at 532 nm.

Comparison of Enriched Oncogenic Pathways

We used Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis to figure out potential signaling pathways of FRGs while using gene ontology (GO) enrichment analysis to explain the functions of FRGs. And the "clusterProfiler" software package was used for visualization. We conducted GO analyses at a standard of p-value < 0.01 and FDR<0.05. In addition, we performed Gene Set Enrichment Analysis (GSEA) on FRGs at a standard of p-value < 0.05. In order to further explain the potential relationship between the pathways involved in FRGs in OC and immune cells, the Gene Set Variation Analysis (GEVA) method was applied to obtain the pathway-score matrix. Pearson's correlation analysis was applied between the GSVA and CIBERSORT matrices.

Statistical Analysis

All statistical analyses were conducted using R and Rstudio software (R version 4.0.3). The Wilcoxon test was used to judge the difference of two groups in single gene expression. Pearson's correlation test was used to determine the correlation between immune cell ratios and pathway scores. According to the risk score, the sample was divided into two groups, and the Kaplan–Meier survival curve was established using the log-rank test. An ROC curve was constructed to evaluate the sensitivity and specificity of the risk score of the E-FRG survival prediction model. All statistical tests were two-sided tests, and a p-value <0.05 was considered statistically significant.

RESULTS

The Relationship Between the Tumor-Infiltrating Immune Cells and Clinical Characteristics of Ovarian Cancer Patients

The flow diagram of the procedures of our study is illustrated in **Figure 1**. In this study, three datasets were enrolled which contained a total of 688 OC samples. The tumor microenvironment cells modulate the antitumor response and play important roles in predicting clinical prognosis and treatment effect. Using the CIBERSORT algorithm, we summarized the 22 subpopulations of TIICs of the 375 OC samples in the training set, including the B cell family (naïve B cells, memory B cells, and plasma cells), T cell family ($CD8^+$ T cells and naïve $CD4^+$ T cells), M0 macrophages, M1 macrophages, resting NK cells, activated NK cells, resting dendritic cells, activated dendritic cells, resting mast cells, and monocytes. The distribution of immune cells was significantly different among individuals, which depicted a comprehensive individual immune characteristic in OC (**Figure 2A**). Then, the relationship between each subtype of TIICs and clinical characteristics in the training set was analyzed, including stage, grade, RFS, and OS. We found that M0 macrophages were identified to be associated with grade (p=.069), which was elevated while the TNM stage increased (**Figure 2B**). Then, the correlation analysis showed a comprehensive landscape of TIIC interactions in OC. The results showed that some tumor immune cells had a high connection with others, including CD8 T cells and M0 macrophages, CD8 T cells and resting dendritic cells, naïve B cells and memory B cells, follicular helper T cells and M0 macrophages, activated dendritic cells and M0 macrophages, gamma delta T cells and resting mast cells, and M1 macrophages and monocytes. (**Figure 2C**)

To construct a prognostic model, the univariable Cox regression of all 22 TIICs in the training set was constructed to find independent prognostic factors for OS. We defined plasma cells, M1 macrophages, monocytes, follicular helper T cells, M2 macrophages, and activated mast cells as risky prognostic factors, whose p-values are 0.013, 0.030, 0.031, 0.001, 0.009, and 0.003 respectively. Then, we applied multivariable Cox proportional hazard regression analysis with the two subtypes of TIICs. As a result, the included two cells showed a poor predictive effect on the prognosis of OC patients. So, we determined to introduce other factors to establish a prognostic model for OC (**Supplementary Table S1**).

Correlation With the Estimation of Tumor-Infiltrating Cells and Prognosis

To figure out the degree of tumor immune cell infiltration in samples, the ESTIMATE algorithm was applied using the gene expression profiles in OC. The TCGA, GSE18520, and GSE32062 datasets were applied. The ESTIMATE algorithm could generate three scores: immune score, stromal score, and ESTIMATE score. These scores were tested by Kaplan–Meier survival analysis and

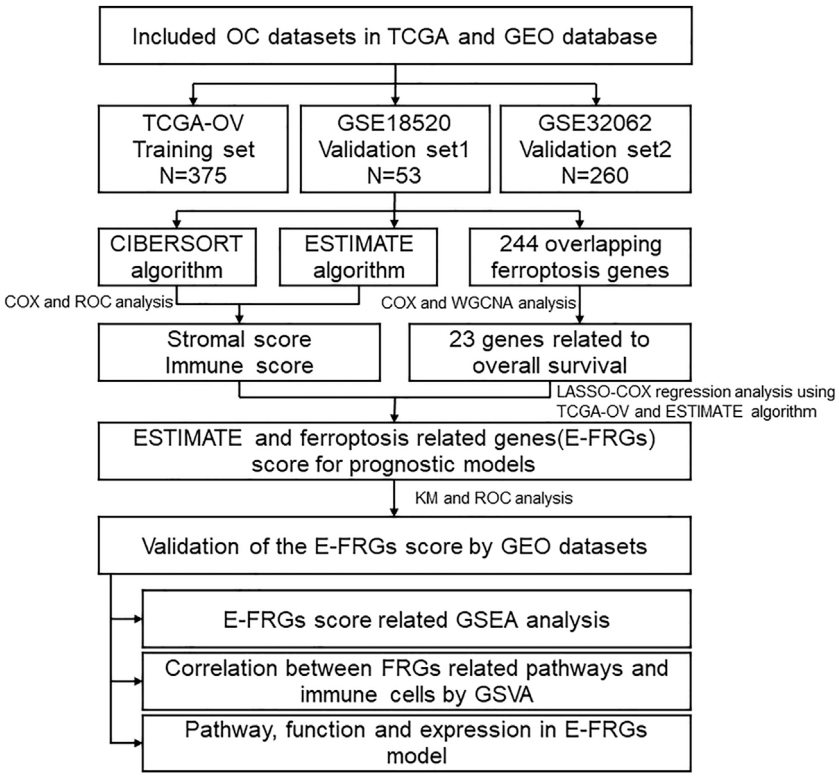

FIGURE 1 | Overall flowchart. First, use ESTIMATE and CIBERSORT were used to analyze the relationship with the prognosis of OC patients. Furthermore, the ESTIMATE scores and ferroptosis-related gene expression were combined to construct the E-FRG model. Finally, the possible regulatory pathway was analyzed based on the E-FRGS model.

ROC curves to assess their prediction capability. The immune score was related to the survival in TCGA(p = 0.06) (**Figure 3A**), GSE18520 (p = 0.11) (**Figure 3B**) and GSE32062 (p=.0059) datasets (**Figure 3C**). The stromal score was significantly related to the survival in TCGA(p=.0058) (**Figure 3D**) and GSE18520 (p = 0.0031) (**Figure 3E**), but not related to GSE32062 (p=.09) (**Figure 3F**). The ESTIMATE score was significantly related to the survival in all three datasets, whose p-values were 0.031, 0.023, and 0.011 respectively. Because the ESTIMATE score is the sum of the aforementioned two scores, we display it in **Supplementary Figure S1**. The stromal score achieved the highest area under the curve of ROC (AUC) in 1-, 3-, and 5-year ROC analysis in GSE18520 (**Figures 3G–I**), while the ESTIMATE score and immune score acquired the highest AUC in 1-year ROC analysis in it. These results indicate that the prognostic prediction capability of the immune score was not effective enough, so we needed more factors to establish a predictive model.

The Analyses of Survival-Related Ferroptosis-Related Genes in Ovarian Cancer

Though there have been some studies about ferroptosis in OC in recent years, the relationship between ferroptosis, cisplatin resistance, and prognosis of OC has not been clear (LIN and CHI, 2020). We collected three clinical tissues from platinum-sensitive patients (S1-S3) and three tissues from platinum-resistant patients (R1-R3) to analyze the protein level of ferroptosis pivot proteins. Then, we found that proteins which inhibited ferroptosis (SLC7A5, XCT, and GPX4) were elevated in platinum-resistant tissues, while ACSL4 which inhibited ferroptosis was elevated in platinum-sensitive tissues (**Figure 4A**). The protein expression had the same tendency in OC cell lines and corresponding platinum-resistant cell lines (OVCAR3/OVCAR3-CDDP and SKOV3/SKOV3-CDDP) (**Figure 4B**). In order to further verify the relationship between ferroptosis and platinum resistance in OC cells, we tested the antitumor effect of the ferroptosis agonist (erastin or RSL3) combined with cisplatin. Eastin and RSL3 can inhibit the activity of XCT and GPX4, respectively, to induce ferroptosis, while acetylcysteine (N-acetyl-l-cysteine, NAC) could inhibit ferroptosis through a mitochondrial-dependent pathway. The IC50 values of cisplatin were 2.2 and 6.9 μg/ml for SKOV3 and SKOV3/CDDP cells, respectively, which was applied in the following experiments (**Figure 4C**). The cell viability assays demonstrated that ferroptosis agonists, namely, erastin and RSL3 could both reinforce the cytoxic effect of cisplatin in SKOV3 and SKOV3-CDDP, while NAC could rescue these lethal combinations (**Figure 4D**). We then examined the accumulation

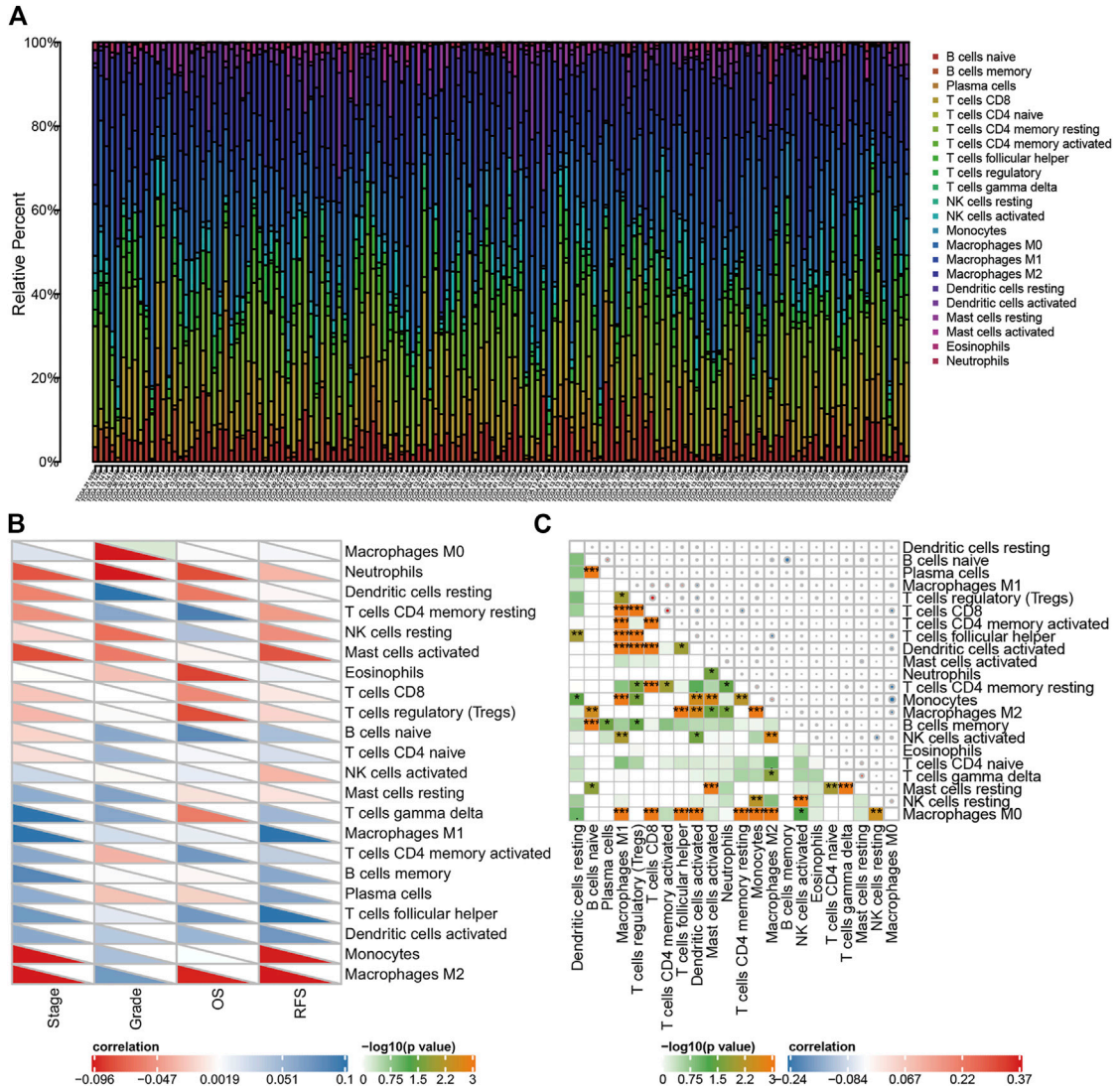

FIGURE 2 | Distribution of tumor immune cells in OC patients, and the relationship between immune infiltration cells and clinical features. **(A)** Proportion of 22 kinds of TIICs in TCGA OC samples was shown in the bar plot. Horizontal axis represents different patient samples and the vertical axis represents the percentage of TIIC. **(B)** Correlation of 22 TIICs with clinicopathologic-grade characteristics was calculated by Pearson's correlation test. **(C)** Correlation among 22 kinds of TIICs. Color in the lower left corner represents the size of the *p*-value, and the size and color of the circle in the upper right corner represent Pearson's correlation coefficient. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

of the end product of lipid oxidation, namely, malondialdehyde (MDA) in the treated cells, which was a marker for oxidative stress and ferroptosis. The content of GSH was also quantified. It showed that erastin and RSL3 could induce ferroptosis by GSH depletion and lipid peroxidation in SKOV3 and SKOV3-CDDP–treated with cisplatin (**Figures 4E,F**). Western blotting assays demonstrated the different protein expression in GPX4 and XCT after the treatment of ferroptosis agonists, namely, NAC and cisplatin, which proved the aforementioned points (**Figure 4G**). These data together suggested that ferroptosis agonists could induce ferroptosis to strengthen the lethal effect of the IC50 concentration of cisplatin in OC cells, which reminded us that ferroptosis may participate in recovering platinum resistance and had potential efficacy in cooperating with cisplatin in OC cells. The gene symbols of the three datasets were integrated with 244 ferroptosis-related genes (FRGs) obtained from the FerrDb database (http://www.zhounan.org/ferrdb) for further analysis (**Figure 5A**). We used WGCNA to analyze gene expression data of overlapping FRGs in the TCGA dataset to establish a co-expression gene network. The soft-thresholding power was defined on the basis of scale-free R2 (R2 = 0.95). Through the power and average linkage hierarchical clustering, we figured out four modules (**Figures 5B,C**). The co-expression gene network has been illustrated in eigengenes. Subsequently, the correlation analysis of each eigengene with clinical characteristics, including OS, grade, stage, and RFS, was

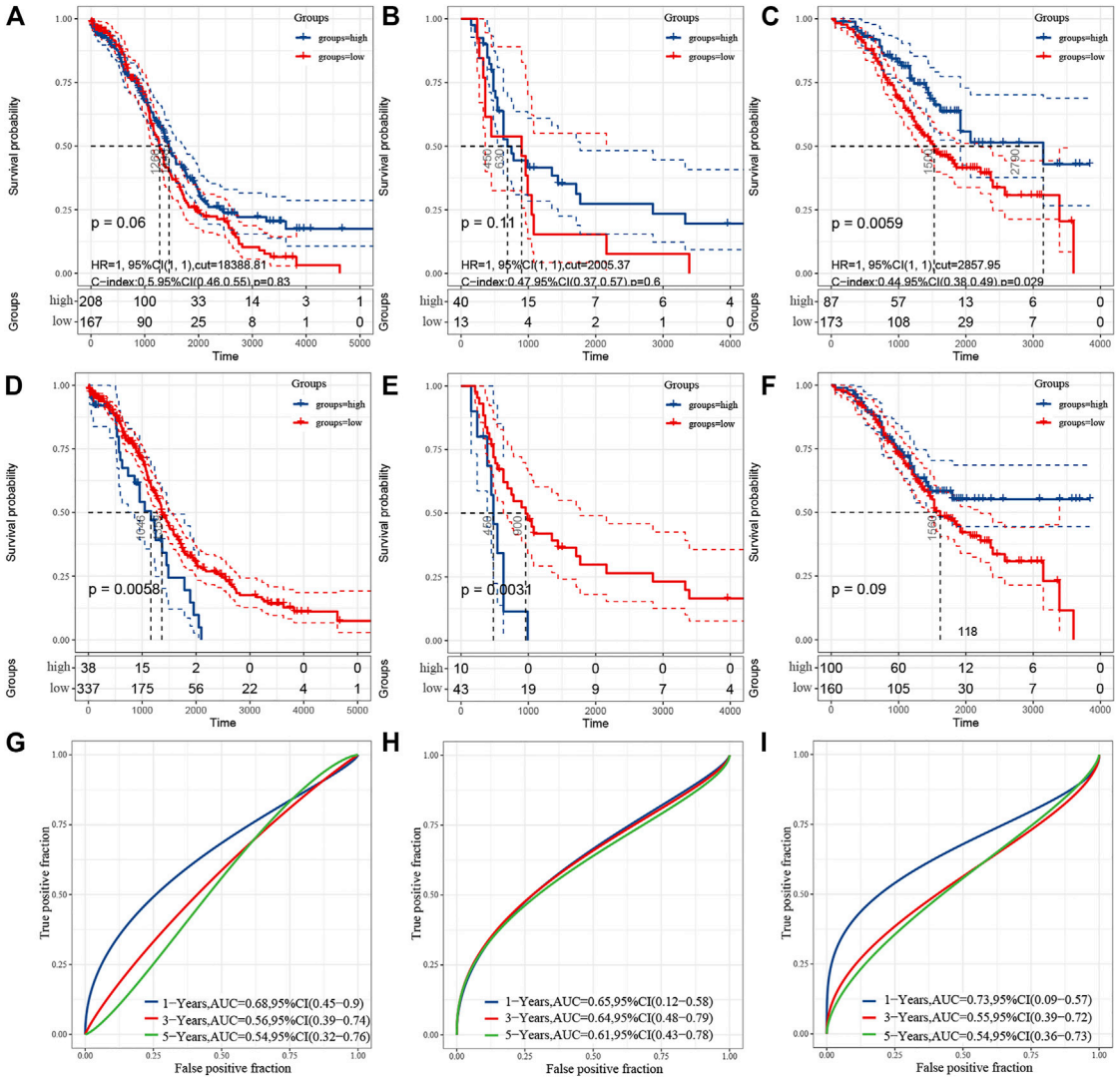

FIGURE 3 | Correlation with the estimation of tumor-infiltrating cells and prognosis. **(A–C)** Kaplan–Meier survival analysis showing association between immune score and OS. **(D–F)** Kaplan–Meier survival analysis showing association between stromal score and OS. **A** and **D** represent TCGA, **B** and **E** represent GSE18520, C and F represent GSE32062. G–I, ROC curve for measuring the predictive value of immune score or stromal score for OS in GSE18520. **(G)** represents the immune score. **(H)** represents the stromal score. **(I)** represents the ESTIMATEscore.

conducted. As presented in **Supplementary Figure S2**, the gray module was positively correlated with RFS (p = 0.01) in OC patients. The blue module was negatively correlated with OS (p = 0.02), and the turquoise module was negatively correlated with OS (p = 0.006), stage (p = 1e-04) and RFS (p = 0.01). The gray module contained 44 FRGs and the turquoise contained 115 FRGs. It is reported that the ferroptosis-related genes are involved in tumor stromal invasion and immune process (STOCKWELL and JIANG, 2019), but it is not clear how the correlation works in OC. We first selected 23 genes related to survival from the data of the aforementioned 244 gene sets and performed Pearson's correlation analysis with the results calculated by the ESTIMATE algorithm. The results showed that 15 genes each were significantly related to immune scores and stromal scores in the ESTIMATE algorithm. This suggests that the ferroptosis-related genes (FRGs) are closely related but not completely collinear with ESTIMATE (**Figure 5D**). The combination of these two indexes may obtain a more accurate prognostic model.

Construction and Validation of the Estimate and Ferroptosis-Related Gene Model by Integrating Estimate and Ferroptosis-Related Genes

Since the ESTIMATE algorithm cannot precisely distinguish patients with OC with high-risk well, we decided to introduce FRGs into the prognostic risk prediction model for OC. Recently reported FRGs are associated with platinum resistance and poor prognosis in OC, while ferroptosis could affect the activation and function of immune cells (Friedmann Angeli et al., 2019; CHAN

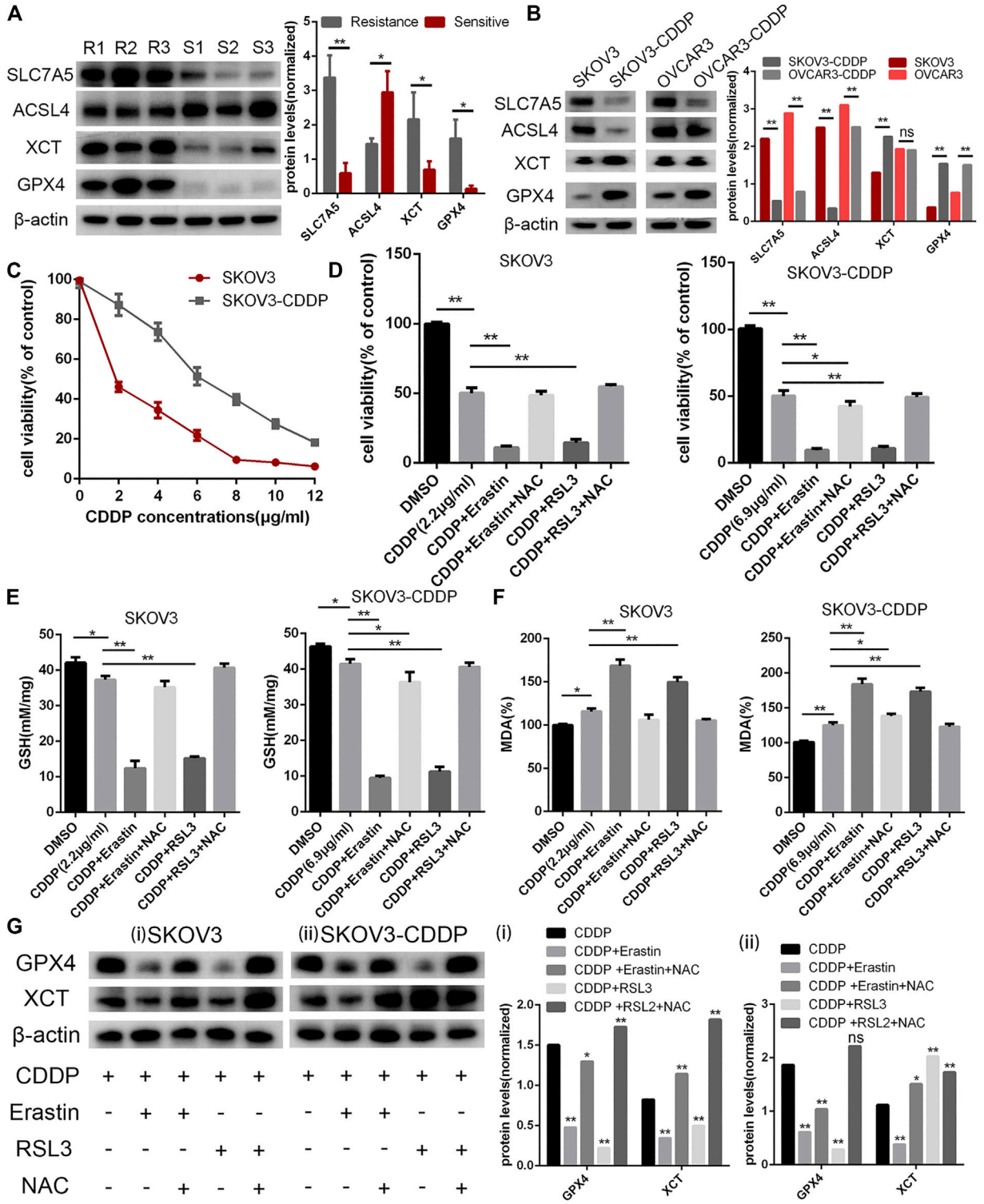

FIGURE 4 | Analyses of survival-related FRGs in OC. **(A)** Ferroptosis pivot genes, namely, SLC7A5, ACSL4, XCT, GPX4, and β-actin protein expression in platinum-resistant tissues (R1-R3) and platinum-sensitive tissues (S1-S3). **(B)** SLC7A5, ACSL4, XCT, GPX4, and β-actin protein expression in ovarian cell lines (OVCAR3 and SKOV3) and platinum-resistant OC cell lines (OVCAR3-CDDP and SKOV3-CDDP). **(C)** Cell viability assays under cisplatin treatment; SKOV3/CDDP showed higher cisplatin tolerance, representing the resistance phenotype. Cells were treated with cisplatin (2.2 μg/ml for SKOV3 and 6.9 μg/ml for SKOV3-CDDP) in the absence or presence of erastin (10 μM), RSL3 (0.1 μM), and NAC (1 mM) for 24 h and cell viability. **(D)** Glutathione (GSH) **(E)** and lipid peroxidation (MDA) assays **(F)** were performed. **(G)** XCT, GPX4, and β-actin protein expression in SKOV3 **(i)** and SKOV3-CDDP **(ii)** under the abovementioned treatment. Error bars represent the standard deviation (s.d.) of triplicate measurements. $^{*}p < 0.05$ and $^{**}p < 0.01$.

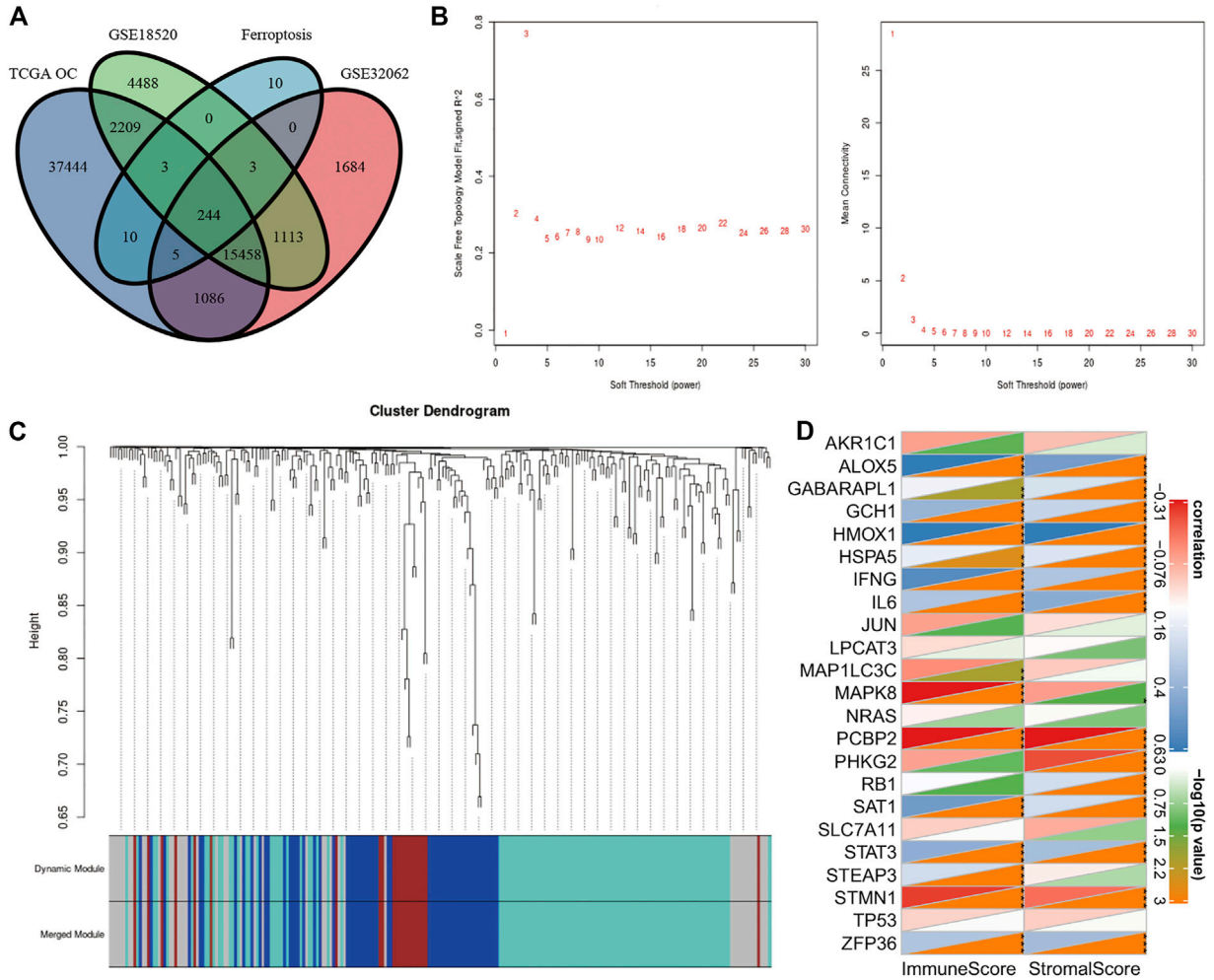

FIGURE 5 | (A) Venn diagrams are used to show the overlapping FRGs in FerrDb, TC GA OV, GSE18520, and GSE32062. **(B)** Scale-free fit index for soft-thresholding powers. Ordinate on the left represents the square of the correlation coefficient between log (k) and log [p (k)] in the FRG network. Ordinate on the right represents the mean value of the adjacent function of all genes in the corresponding gene module. **(C)** Cluster dendrogram of ferroptosis-related genes. Each branch represents a gene, and the colors below represent specific co-expression modules. **(D)** Pearson's correlation analysis between 23 ferroptosis-related genes and the results calculated by ESTIMATE. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

et al., 2020; WANG et al., 2021; MA et al., 2021). It could be inferred that the integration of the immune and FRGs may be able to assess the condition of OC patients and form an effective prognostic predictive risk model. Univariable Cox proportional hazard regression analysis was applied with initial screening of 244 FRGs, and 23 FRGs were finally identified based on the log-rank test. Next, the LASSO–Cox regression model was conducted integrating the ESTIMATE algorithm and these FRGs (E-FRGs) (**Figures 6A,B**). Finally, the immune score, stromal score, and 15 FRGs were determined in the E-FRG model, which included solute carrier family seven member 11 (SLC7A11), RB transcriptional corepressor 1 (RB1), GTP cyclohydrolase 1 (GCH1), lysophosphatidylcholine acyltransferase 3 (LPCAT3), poly (rC)-binding protein 2 (PCBP2), zinc finger RNA-binding protein (ZFP36), STEAP3 metalloreductase (STEAP3), mitogen-activated protein kinase 8 (MAPK8), GABA type A receptor–associated protein like 1 (GABARAPL1), interferon gamma (IFNG), phosphorylase kinase catalytic subunit gamma 2 (PHKG2), NRAS, hot shock protein A5 (HSPA5), microtubule-associated protein one light chain three gamma (MAP1LC3C), polyunsaturated fatty acid 5-lipoxygenase (ALOX5), ESTIMATE score, and stromal score (**Figure 6C**). In the E-FRG model, the risk score (E-FRG score) was generated using the following formula:

E-FRG score=(0.019×ALOX5) + (0.108×GABARAPL1) + (-0.122×GCH1) + (-0.054×HSPA5) + (-0.326×IFNG) + (0.105×LPCAT3) + (-0.061×MAP1LC3C) + (-0.181×MAPK8) + (-0.145×NRAS) + (-0.240×PCBP2) + (-0.104×PHKG2) + (0.142×RB1) + (-0.281×SLC7A11) + (0.011×STEAP3) + (0.041×ZFP36) + (-1.19E-06×stromal score) + (-9.91E-06-06×immune score).

Then the E-FRG score was calculated for the patients enrolled in the TCGA dataset with the corresponding score. The optimal cutoff value was determined by the median value of -2.26, -8.75,

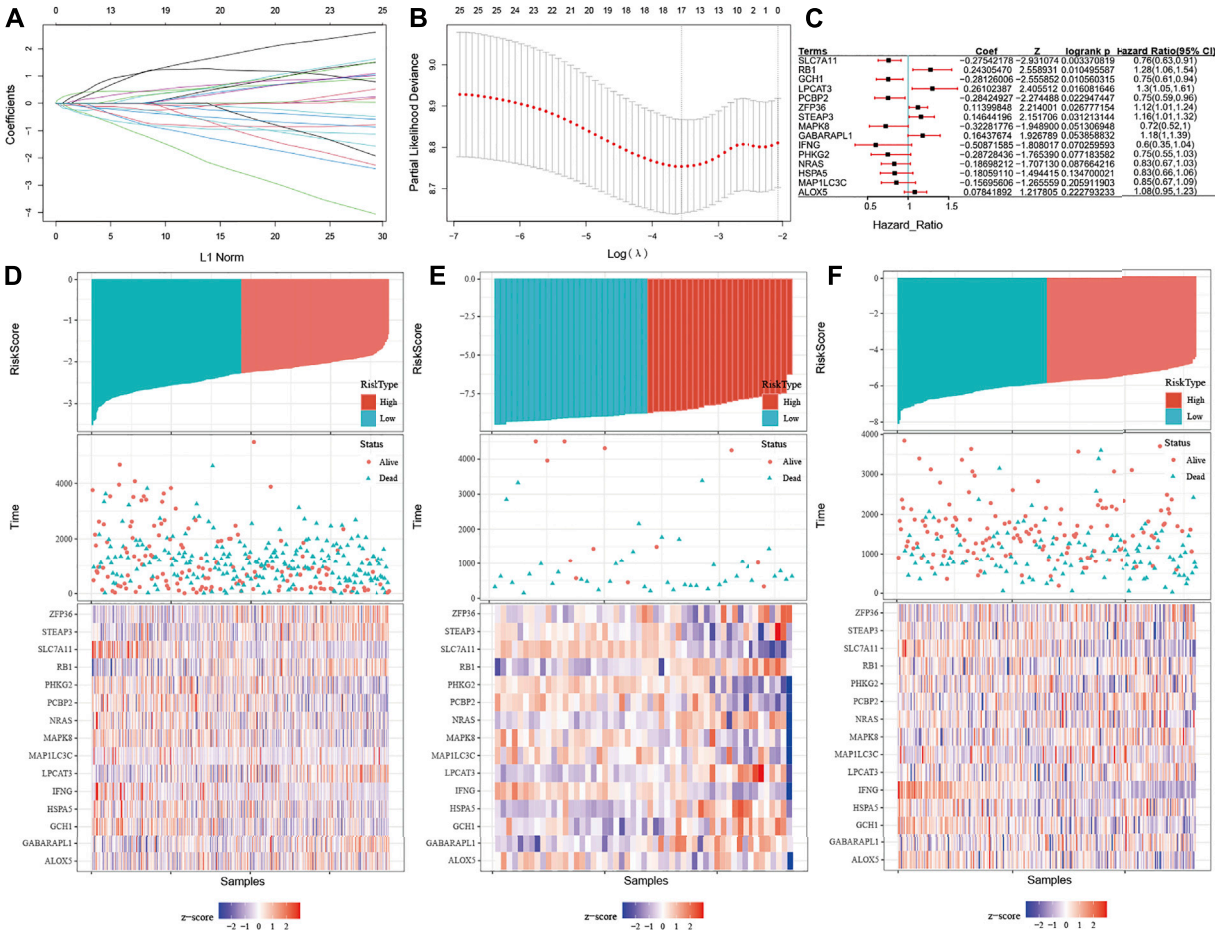

FIGURE 6 | Construction of the E-FRG model by integrating ESTIMATE and FRGs. **(A)** LASSO coefficient profiles of the 23 FRGs, immune score, and stromal score. Coefficient profile plot was produced versus the log (λ). **(B)** Tuning parameter (λ) selection in the LASSO-Cox regression model was performed using 10-fold cross-validation via the 1 standard error for the minimum criteria. Black vertical lines were plotted at the optimal λ based on the minimum criteria and 1 standard error for the minimum criteria. **(C)** Forest plot showing *p*-values of FRGs in the model indicated by the color scale presented on the side. Horizontal bars represent 95% confidence intervals. **(D–F)** Distribution between the high E-FRG score group and low E-FRG score group in the TCGA dataset **(D)**, GSE18520 **(E)**, and GSE32062 **(F)**, respectively.

and -5.81 in TCGA, GSE18520, and GSE32062, respectively (**Figures 6D–F**). Subsequently, the included OC patients could be classified into high-risk and low-risk groups on the basis of the median value. The Kaplan–Meier curve analysis showed that the E-FRG model could distinguish OC patients with good or bad prognosis. The high–E-FRG group manifested a shorter OS than the OS of low–E-FRGs group in the TCGA dataset ($p < 0.0001$, **Figure 7D**). In GSE18520, patients with high risk showed shorter OS with a marginally significant *p*-value of 0.15 (**Figure 7E**). In GSE32062, the OS was significantly shorter in the patients with high risk ($p = 0.0071$, **Figure 7F**).

Time-dependent ROC analysis showed that AUC of E-FRG scores for the prediction of 1-, 3-, and 5-year OS in the training set were 0.60, 0.67, and 0.70, respectively (**Figure 7A**). In GSE18520, the AUC value of the E-FRG scores for predicting 1-, 3-, and 5-year OS were 0.46,0.60, and 0.70, respectively (**Figure 7B**), while those in GSE32062 were 0.65,0.60, and 0.64, respectively (**Figure 7C**). It is worth noting that the E-FRG score showed better predictive effect than other potential prognostic markers on the basis of time-dependent ROC analysis (**Supplementary Figure S3**).

The Development of a Nomogram for Improving the Estimate and Ferroptosis-Related Gene Model

As a regression model illustrated in images, the nomogram is widely applied in the diagnosis of tumors and the prognostic analysis. The nomogram and calibration curve were applied in our study in order to illustrate the E-FRG model more vividly and improve the practicality of this model (**Figures 7G,H**). The nomogram included more than 17 traits, including the 15 FRGs, immune score, and stromal score. The score of each characteristic was determined by the scale on the top. The sum of the scores of the 17 traits was defined as the final score. We could estimate the prognosis of 1-, 3-, and 5-year

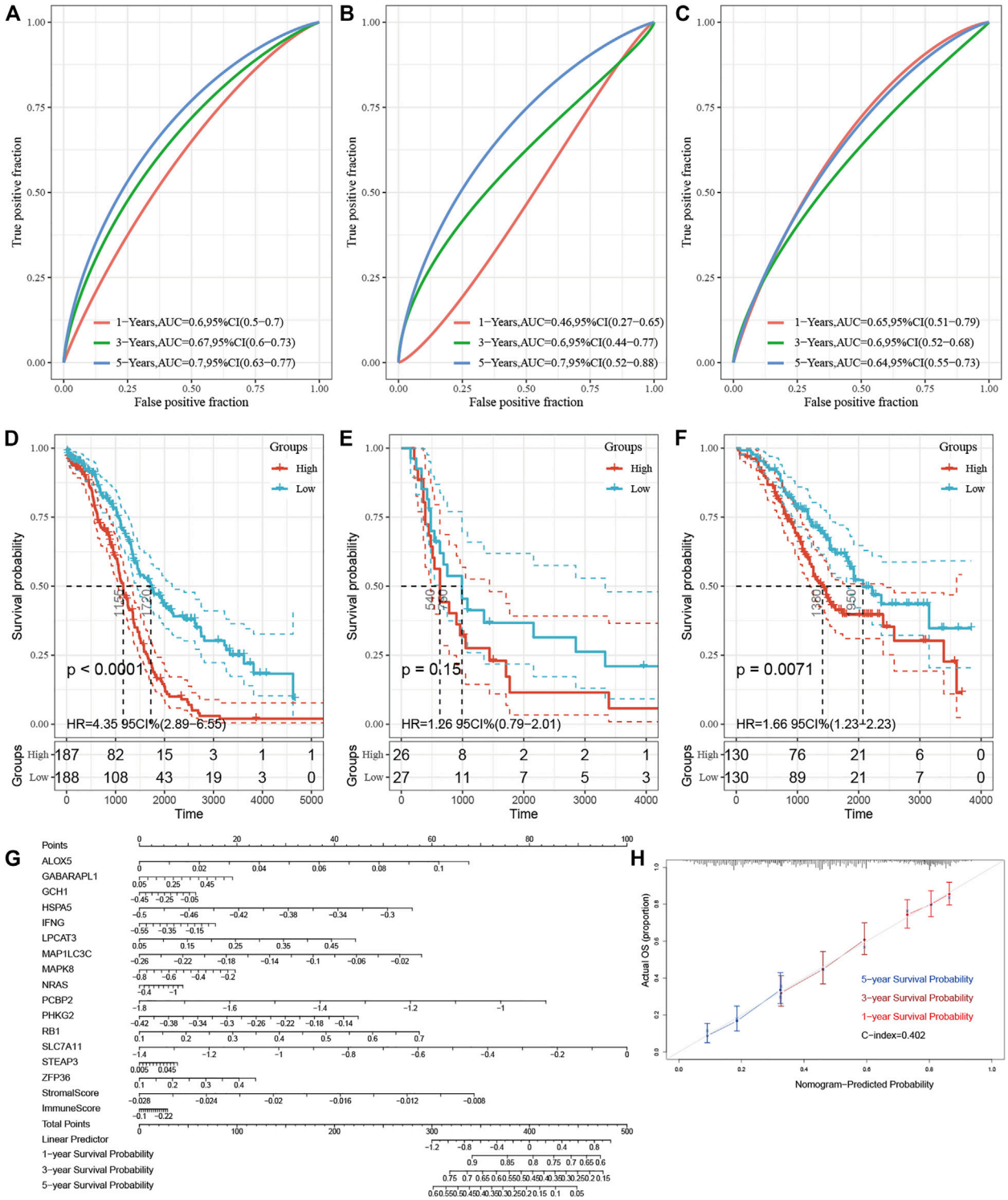

FIGURE 7 | Validation of the E-FRG model and development of a nomogram for improving the model. **(A–C)** Area under the curve of the ROC curve is used to visually indicate the predictive power of E-FRG scores for 1-, 3-, and 5-year OS of OC patients in the TCGA dataset **(A)**, GSE18520 **(B)**, and GSE32062 **(C)**, respectively. **(D–F)** Kaplan–Meier curves for 1-, 3-, and 5-year OS with low and high E-FRG scores in the TCGA dataset **(D)**, GSE18520 **(E)**, and GSE32062 **(F)**, respectively. **(G)** Nomogram for predicting the 1-, 3-, and 5-year OS probabilities of OC patients. The total score is composed of 17 scoring features. Each score of the nomogram corresponds to the genes or ESTIMATE scores in the E-FRG model. **(H)** Calibration of the nomogram for predicting the probability of survival at 1-, 3-, and 5-year in TCGA (C-dex = 0.402).

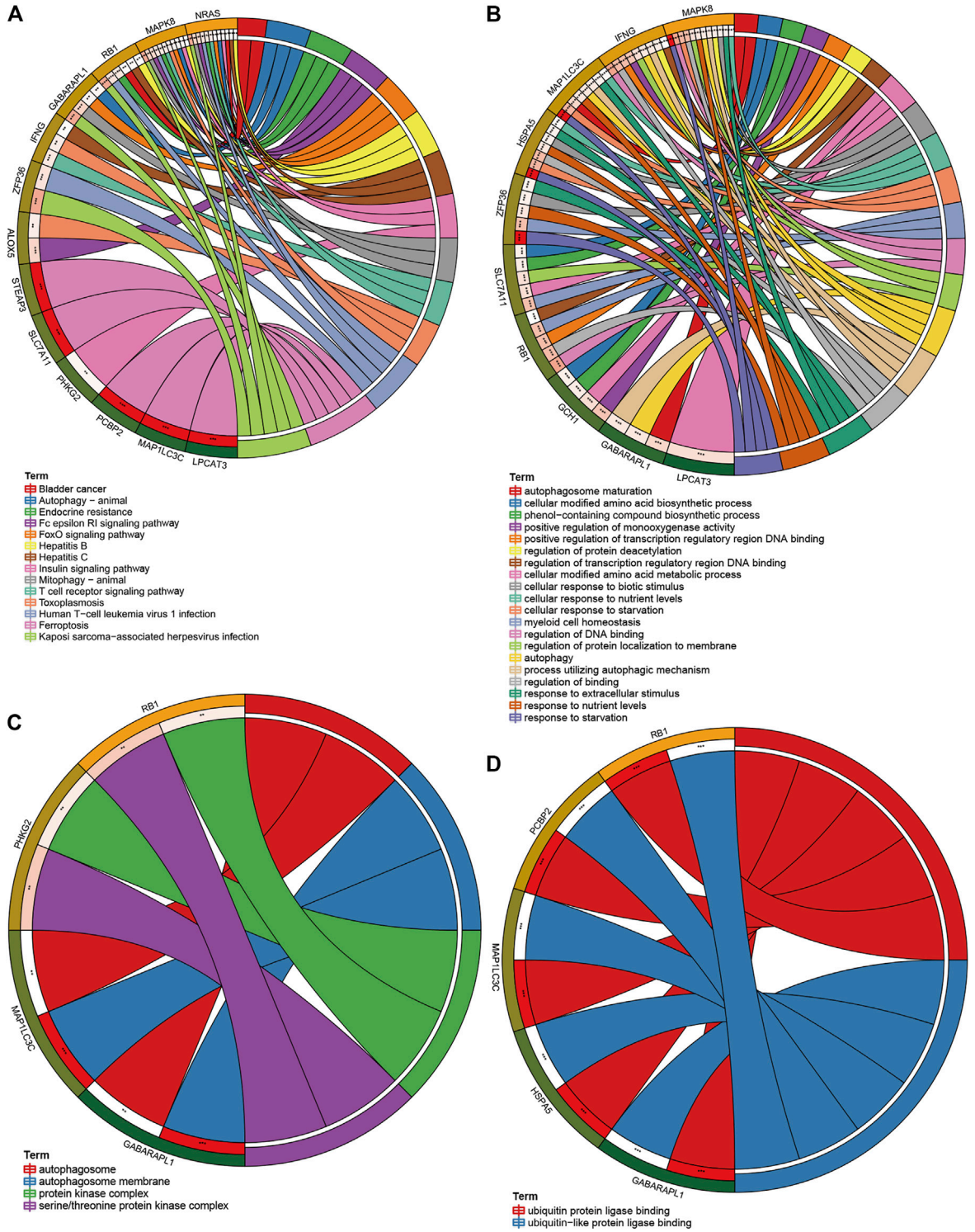

FIGURE 8 | Defining FRG-related pathways and immune cells by GSEA and GSVA analysis. **(A,B)** GSEA analysis between the low and high E-FRG score groups. **(C)** GSVA analysis on FRGs in the TCGA training set which provided 22 FRG-related pathways. **(D)** Correlation of 22 TIICs with 22 FRG-related pathways. Pearson's coefficient was used for significance test. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

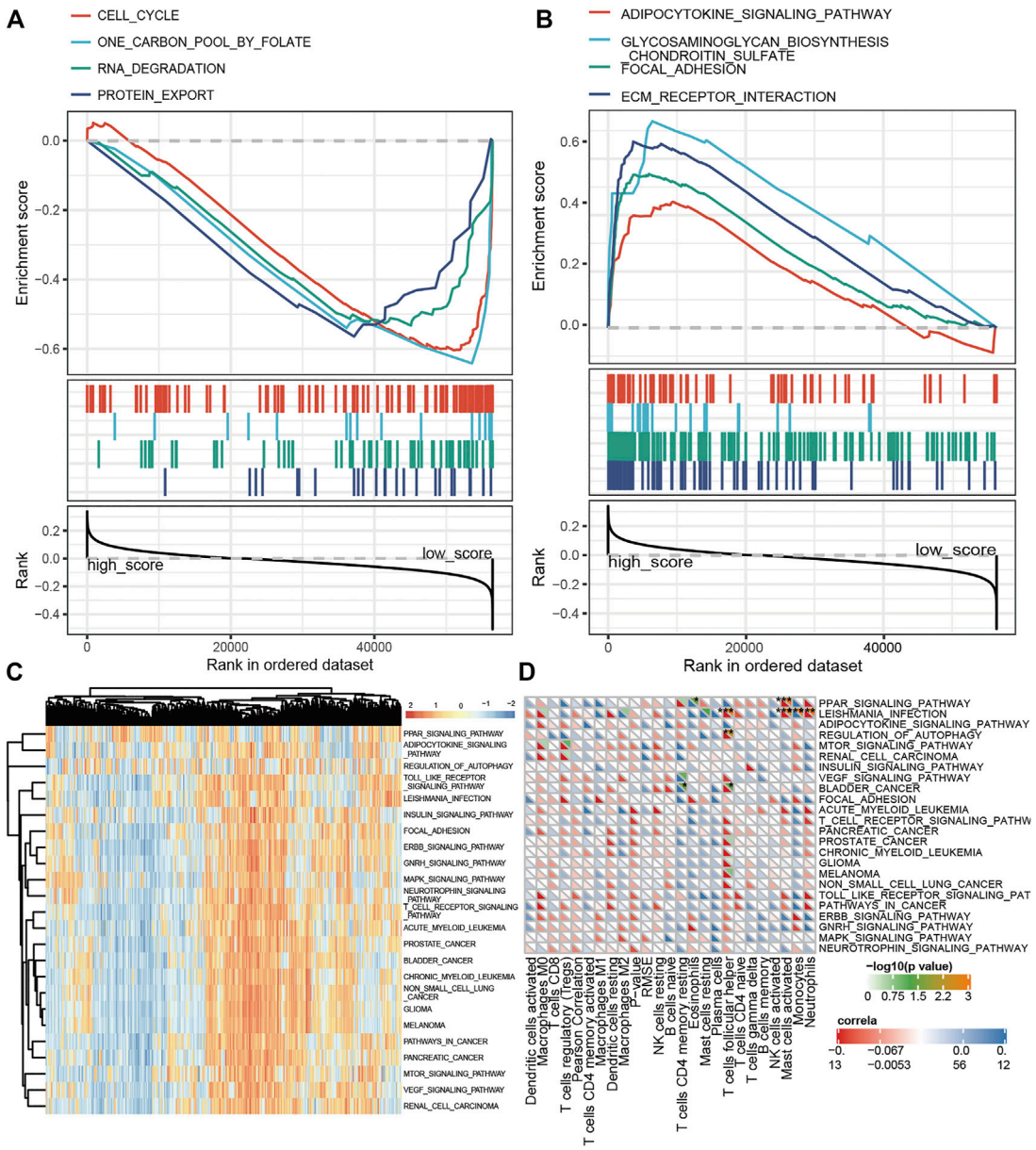

FIGURE 9 | KEGG, GO analysis of FRGs in the E-FRG model. **(A)** KEGG analysis showing main pathways related to ferroptosis. (B–D) GO analysis showing significant GO terms related to FRGs. **(B)** FRGs related to biological processes. **(C)** FRGs related to cell components. **(D)** FRGs related to molecular function.

OS for OC patients by the perpendicular line from the total point axis to the two-outcome axis.

Defining Ferroptosis-Related Gene-Related Pathways and Immune Cells by Gene Set Enrichment Analysis and GSVA Analysis

The enrichment analysis of ferroptosis is helpful to explain why the E-FRG score can better predict the prognosis of OC patients. We first used E-FRG scores to perform KEGG–GSEA analysis and selected eight pathways related to E-FRG scores. Among them, adipocytokine, glycosaminoglycan biosynthesis chondroitin sulfate, focal adhesion, and ECM receptor interaction signaling pathway are positively correlated with E-FRG scores (**Figure 8A**) and cell cycle, one carbon pool by folate, RNA degradation, and protein export are negatively correlated (**Figure 8B**).

In addition, ferroptosis is closely related to the regulation and activation of immune cells (HASCHKA et al., 2020). Exploring the immune-related FRG pathway is conducive to the discovery of drugs that have the ability to activate ferroptosis and immunity at the same time. We first performed GSVA analysis on FRGs (**Figure 8C**) and calculated the correlation between the GSVA and CIBERSORT scores (**Figure 8D**). The results showed that

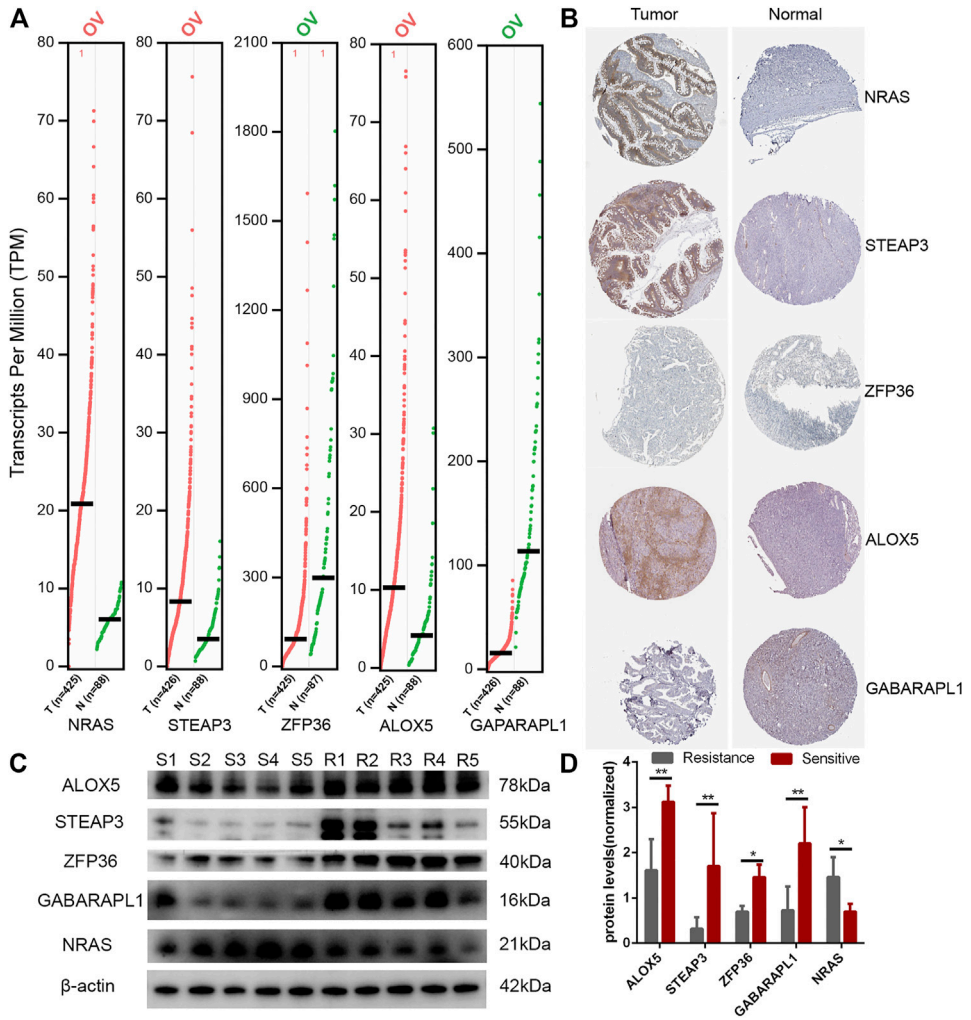

FIGURE 10 | Expression of FRGs in the E-FRG model. **(A)** GEPIA database showing the expression of TOP5 different genes in OC tissues compared with normal tissues. **(B)** HPA database showing the expression of TOP5 different genes in OC tissues compared with normal tissues. **(C)** Protein expression of TOP5 genes in the E-FRG model (ALOX5, STEAP3, ZFP36, GABARAPL1, and NRAS) and β-actin in platinum-resistant tissues (R1-R5) and platinum-sensitive tissues (S1-S5). **(D)** Error bars represent the standard deviation (s.d.) of quintuplicate measurements. *$p < 0.05$ and **$p < 0.01$.

autophagy, leishmania infection, and the PPAR signaling pathway were positively correlated with follicular helper T cells, mast cells, and neutrophils and negatively correlated with monocytes and eosinophils.

Encyclopedia of Genes and Genomes, Gene Ontology Analysis, and Expression of Genes in the Estimate and Ferroptosis-Related Gene Model

After the overall enrichment analysis of FRG, we performed KEGG and GO analysis on 15 genes in the E-FRG model and also collected their RNA and protein expression. In the results of KEGG analysis, autophagy, endocrine resistance, and immune pathways are the main pathways related to ferroptosis (**Figure 9A**). Ten genes are enriched in GO biological processes, and these processes are mainly involved in autophagy, amino acid metabolism, various oxidative stress reactions, and subsequent DNA damage (**Figure 9B**). Cell component analysis suggests that four genes are involved in the autophagosome and protein kinase complex (**Figure 9C**). The results of GO molecular function suggest that five genes, such as RB1 and HSPA5, are enriched in the ubiquitination process (**Figure 9D**).

Finally, we showed TOP5 different expression genes in the E-FRG model. Compared with normal tissues, NRAS, STEAP3, and ALOX5 are highly expressed in OC and ZFP36 and GABARAPL1 are lowly expressed (**Figure 10A**). In the HPA database, NRAS, STEAP3, and ALOX5 proteins are significantly expressed in tumor tissues, but the difference in the expression of ZFP36 and GABARAPL1 proteins is not obvious (**Figure 10B**). Then, we analyzed the expression of TOP5 genes in five clinical tissues from platinum-sensitive patients (S1-S5) and five tissues from platinum-resistant patients (R1-R5). It illustrated that the

genes negatively correlated with E-FRG model (NRAS) were elevated in platinum-sensitive tissues, while genes positively correlated with the E-FRG model (ALOX5, STEAP3, ZFP36, and GABARAPL1) were elevated in platinum-resistant tissues (**Figures 10 C,D**).

DISCUSSION

OC is the most difficult malignant tumor to treat among common gynecological tumors (BRAY et al., 2018). Immunotherapy and targeted therapy have improved the prognosis of patients to a certain extent. However, the current prognosis of patients based on immunotherapy is still unpredictable, so it is necessary to explore models that can accurately predict the prognosis of OC patients. More and more studies have shown that immune cells or stromal cell infiltration are related to the survival of OC patients (LI et al., 2021). Ferroptosis is a new type of programmed cell death, and many ferroptosis-related genes are involved in immune regulation. These genes have been found to have expression changes after targeted therapies in tumors, so they may be used to predict the prognosis of OC patients (YE et al., 2020). Research studies on ferroptosis or immunity have also been reported. In 2021, Yang et al. reported a prognostic model consisting of nine ferroptosis-related genes by multi COX regression analysis, but only 60 ferroptosis-related genes were included and the immune–stromal score was not involved in the study (YANG et al., 2021). Here, we first evaluated the immune cell infiltration of OC and the evaluation effect of the ESTIMATE score on the prognosis of OC. Based on collecting ferroptosis-related genes as much as possible, the E-FRG model combining ESTIMATE scores and FRGs was developed and verified by the LASSO-COX method. The E-FRG model used to identify high-risk patients has good discrimination in both training and validation datasets.

In this study, the constructed prognostic model consisted of two ESTIMATE scores and 15 genes related to ferroptosis. These genes regulate ferroptosis through a variety of metabolic pathways. STEAP3 is an endosomal ferrireductase which functions as an iron and copper transporter in erythroid cells. STEAP3 maintains the homeostasis of iron ions in cells by reducing Fe3+ to Fe2+, thus participating in the regulation of ferroptosis in bone marrow injury (WILKINSON et al., 2019). SLC7A11, also called XCT, is a signature protein of ferroptosis and participates in intracellular and extracellular cysteine transport to regulate the production of glutathione (GSH) (KOPPULA et al., 2020). PCBP2 is the main poly (rC)-binding protein in the cell. It was previously thought to act as an adaptor between the mitochondrial antiviral signaling protein (MAVS) and E3 ubiquitin ligase, triggering the ubiquitination and degradation of MAVS. It was also thought to participate in the transduction of mitochondrial antiviral signals (YANATORI et al., 2016). NRAS regulates the activity of GTPase and is one of the most common targets of targeted drugs, as well as the upstream target of drugs, such as sorafenib. Many studies believed that RAS gene activation is an important prerequisite for tumor cell ferroptosis (SU et al., 2020). RB1 is a key regulator of cell division. The active form of RB1 binds to E2F transcription factor 1 and inhibits the transcriptional activity of it, leading to cell cycle arrest. Louandre found in sorafenib-treated liver cancer cells in which the downregulation of RB1 promoted the occurrence of ferroptosis (LOUANDRE et al., 2015). MAPK8 is an important member of the serine/threonine protein kinase family, responsible for regulating the proliferation and apoptosis of glioblastoma cells (XU et al., 2018). In ferroptosis, the MAPK pathway is often studied as a downstream of RAS (NGUYEN et al., 2020).

MAP1LC3C, also known as LC3C, is a key protein in the autophagy pathway and the ubiquitin-like modification of heterophagy (CHO et al., 2020). This model also includes GABARAPL1, which also belongs to the LC3 family of autophagy and is responsible for the formation of autophagic vesicles (PARK et al., 2016). Ferritinophagy is a specific autophagosome formed by ferritin and autophagy. This study suggests that MAP1LC3C may be involved in this process, but it has not been further confirmed by studies. ZFP36 is a C2H2 zinc finger protein, which destabilizes the transcription containing the AU-rich element (ARE) by removing or deadenylating the poly (A) tail of RNA, thus attenuating protein synthesis. Since ZFP36 is a key factor regulating ROS homeostasis, it has been confirmed that ZFP36 prevents ferroptosis in hepatic stellate cells through the autophagy signaling pathway (ZHANG et al., 2020). PHKG2 is the catalytic subunit of PHK which participates in glycogen decomposition and hormone regulation by activating glycogen phosphorylase. PHKG2 regulates the effectiveness of iron ions on lipoxygenase and then drives ferroptosis through the peroxidation of polyunsaturated fatty acids (PUFA) at the diallyl position (YANG et al., 2016). LPCAT3 is a phosphatidylcholinease that participates in and maintains the homeostasis of phospholipid metabolism and regulates tumorigenesis. Consistent with the study of Yang et al., the model we constructed also believes that LPCAT3 is an important iron death gene for predicting the prognosis of OC (YANG et al., 2021).

ALOX5 belongs to the lipoxygenase family, which is rich in iron and is responsible for lipid oxidation. ALOX5 catalyzes the peroxidation of PUFA, which is the limit of the leukotriene biosynthesis and ferroptosis process. Ferroptosis can further lead to positive feedback amplification of itself by reducing ALOX5-mediated inflammation (LIU et al., 2015). IFNG is an important macrophage activator which has a regulatory effect on the phenotypic transformation of tumor-associated macrophages and can resist the growth of pancreatic tumors (HUANG et al., 2020). Studies have confirmed that IFNG released by cytotoxic T cells downregulates the expression of glutamate transport systems (SLC3A2 and XCT), thereby promoting lipid peroxidation and iron drop in cancer cells, which explains the close relationship between immunity and ferroptosis (WANG et al., 2019). Endoplasmic reticulum stress is also believed to interact with ferroptosis. HSPA5 is a kind of heat shock protein. As a downstream of UPR, it may inhibit ferroptosis (ZHU et al., 2017). GCH1 induces lipid remodeling by regulating the synthesis of tetrahydrobiopterin/dihydrobiopterin, thereby selectively preventing phospholipid consumption and inhibiting ferroptosis (KRAFT et al., 2020).

In summary, we constructed a model that can effectively predict the prognosis of OC patients based on ESTIMATE scores and ferroptosis-related genes. Our E-FRG model provides new perspectives for the improvement of individualized management of OC patients. In addition, this study found that the level of the E-FRG score is related to the degree of infiltration of immune cells and stromal cells. These ferroptosis-related genes also interact with signal pathways, such as autophagy and immunity, suggesting that these genes may be key nodes in the crosstalk of pathways.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Cancer Genome Atlas (TCGA) database for TCGA-OV at https://www.cancer.gov/t-cga, the Gene Expression Omnibus (GEO) database for GSE32062 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32062, for GSE73293 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73293.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Human Ethics Committee of The Second Xiangya Hospital of Central South University (2017S109). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

X-L (XXL) conducted the research and drafted the manuscript. LX (LX) and YW (YW) helped with implementation. Z-Z (ZJZ) analyzed bioinformation data. LX and YW reviewed the manuscript. All authors read and approved the submitted manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China, No. 81970569, Fundamental Research Funds for the Central Universities of Central South University, No. 2021zzts0367 and Hunan Provincial Innovation Foundation For Postgraduate, No. CX20210369.

ACKNOWLEDGMENTS

The Medical Experiment Center of the Second Xiangya Hospital of Central South University provided assistance in equipment and experimental design for this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.774400/full#supplementary-material

REFERENCES

Baci, D., Bosi, A., Gallazzi, M., Rizzi, M., Noonan, D. M., Poggi, A., et al. (2020). The Ovarian Cancer Tumor Immune Microenvironment (TIME) as Target for Therapy: A Focus on Innate Immunity Cells as Therapeutic Effectors. *INT. J. MOL. SCI.* 21, 21. doi:10.3390/ijms21093125

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68, 394–424. doi:10.3322/caac.21492

Carbone, M., and Melino, G. (2019). Stearoyl CoA Desaturase Regulates Ferroptosis in Ovarian Cancer Offering New Therapeutic Perspectives. *Cancer Res.* 79, 5149–5150. doi:10.1158/0008-5472.can-19-2453

Chan, D. W., Yung, M. M., Chan, Y.-S., Xuan, Y., Yang, H., Xu, D., et al. (2020). MAP30 Protein from Momordica Charantia Is Therapeutic and Has Synergic Activity with Cisplatin against Ovarian Cancer *In Vivo* by Altering Metabolism and Inducing Ferroptosis. *Pharmacol. Res.* 161, 105157. doi:10.1016/j.phrs.2020.105157

Chew, V., Toh, H. C., and Abastado, J. P. (2012). Immune Microenvironment in Tumor Progression: Characteristics and Challenges for Therapy. *J. Oncol.* 2012, 608406. doi:10.1155/2012/608406

Cho, H. S., Park, S. Y., Kim, S. M., Kim, W. J., and Jung, J. Y. (2020). Autophagy-Related Protein MAP1LC3C Plays a Crucial Role in Odontogenic Differentiation of Human Dental Pulp Cells. *Tissue eng. Regen. Med.* 18 (2), 265–277. doi:10.1007/s13770-020-00310-3

Corn, K. C., Windham, M., Kenzie, A., and Rafat, M. (2020). Lipids in the Tumor Microenvironment: From Cancer Progression to Treatment. *PROG. LIPID RES.* 80, 101055. doi:10.1016/j.plipres.2020.101055

Corrado, G., Palluzzi, E., Bottoni, C., Pietragalla, A., Salutari, V., Ghizzoni, V., et al. (2019). New Medical Approaches in Advanced Ovarian Cancer. *MINERVA MED.* 110, 367–384. doi:10.23736/S0026-4806.19.06139-1

De Nola, R., Menga, A., Castegna, A., Loizzi, V., Ranieri, G., Cicinelli, E., et al. (2019). The Crowded Crosstalk between Cancer Cells and Stromal Microenvironment in Gynecological Malignancies: Biological Pathways and Therapeutic Implication. *INT. J. MOL. SCI.* 20, 20. doi:10.3390/ijms20102401

Dixon, S. J., Lemberg, K. M., Lamprecht, M. R., Skouta, R., Zaitsev, E. M., Gleason, C. E., et al. (2012). Ferroptosis: an Iron-dependent Form of Nonapoptotic Cell Death. *Cell* 149, 1060–1072. doi:10.1016/j.cell.2012.03.042

Friedmann Angeli, J. P., Krysko, D. V., and Conrad, M. (2019). Ferroptosis at the Crossroads of Cancer-Acquired Drug Resistance and Immune Evasion. *Nat. Rev. Cancer* 19, 405–414. doi:10.1038/s41568-019-0149-1

Giampaolino, P., Della Corte, L., Foreste, V., Vitale, S. G., Chiofalo, B., Cianci, S., et al. (2019). Unraveling a Difficult Diagnosis: the Tricks for Early Recognition of Ovarian Cancer. *Minerva med.* 110, 279–291. doi:10.23736/S0026-4806.19.06086-5

Haschka, D., Hoffmann, A., and Weiss, G. (2020). Iron in Immune Cell Function and Host Defense. *Semin. Cell dev. Biol.* 115, 27–36.

Huang, J., Chen, P., Liu, K., Liu, J., Zhou, B., Wu, R., et al. (2020). CDK1/2/5 Inhibition Overcomes IFNG-Mediated Adaptive Immune Resistance in Pancreatic Cancer. *Gut* 70 (5), 890–899. doi:10.1136/gutjnl-2019-320441

Jiang, Y., Wang, C., and Zhou, S. (2020). Targeting Tumor Microenvironment in Ovarian Cancer: Premise and Promise. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1873, 188361. doi:10.1016/j.bbcan.2020.188361

Koppula, P., Zhuang, L., and Gan, B. (2020). Cystine Transporter SLC7A11/xCT in Cancer: Ferroptosis, Nutrient Dependency, and Cancer Therapy. *Protein Cell* 12 (8), 599–620. doi:10.1007/s13238-020-00789-5

Kraft, V. A. N., Bezjian, C. T., Pfeiffer, S., Ringelstetter, L., Müller, C., Zandkarimi, F., et al. (2020). GTP Cyclohydrolase 1/Tetrahydrobiopterin Counteract Ferroptosis through Lipid Remodeling. *ACS Cent. Sci.* 6, 41–53. doi:10.1021/acscentsci.9b01063

Lee, J.-Y., Nam, M., Son, H. Y., Hyun, K., Jang, S. Y., Kim, J. W., et al. (2020). Polyunsaturated Fatty Acid Biosynthesis Pathway Determines Ferroptosis

Sensitivity in Gastric Cancer. *Proc. Natl. Acad. Sci. USA* 117, 32433–32442. doi:10.1073/pnas.2006828117

Lheureux, S., Braunstein, M., and Oza, A. M. (2019). Epithelial Ovarian Cancer: Evolution of Management in the Era of Precision Medicine. *CA Cancer J. Clin.* 69, 280–304. doi:10.3322/caac.21559

Li, N., Li, B., and Zhan, X. (2021). Comprehensive Analysis of Tumor Microenvironment Identified Prognostic Immune-Related Gene Signature in Ovarian Cancer. *Front. Genet.* 12, 616073. doi:10.3389/fgene.2021.616073

Li, X., Zhang, Y., Chai, X., Zhou, S., Zhang, H., He, J., et al. (2019). Overexpression of MEF2D Contributes to Oncogenic Malignancy and Chemotherapeutic Resistance in Ovarian Carcinoma. *Am. J. Cancer res.* 9, 887–905.

Lin, C.-C., and Chi, J.-T. (2020). Ferroptosis of Epithelial Ovarian Cancer: Genetic Determinants and Therapeutic Potential. *Oncotarget* 11, 3562–3570. doi:10.18632/oncotarget.27749

Liu, Y., Wang, W., Li, Y., Xiao, Y., Cheng, J., and Jia, J. (2015). The 5-Lipoxygenase Inhibitor Zileuton Confers Neuroprotection against Glutamate Oxidative Damage by Inhibiting Ferroptosis. *Biol. Pharm. Bull.* 38, 1234–1239. doi:10.1248/bpb.b15-00048

Louandre, C., Marcq, I., Bouhlal, H., Lachaier, E., Godin, C., Saidak, Z., et al. (2015). The Retinoblastoma (Rb) Protein Regulates Ferroptosis Induced by Sorafenib in Human Hepatocellular Carcinoma Cells. *Cancer Lett.* 356, 971–977. doi:10.1016/j.canlet.2014.11.014

Ma, L.-L., Liang, L., Zhou, D., and Wang, S.-W. (2021). Tumor Suppressor miR-424-5p Abrogates Ferroptosis in Ovarian Cancer through Targeting ACSL4. *neo* 68, 165–173. doi:10.4149/neo_2020_200707n705

Nguyen, T. H. P., Mahalakshmi, B., and Velmurugan, B. K. (2020). Functional Role of Ferroptosis on Cancers, Activation and Deactivation by Various Therapeutic Candidates-An Update. *Chemico-Biological Interactions* 317, 108930. doi:10.1016/j.cbi.2019.108930

Nowak, M., and Klink, M. (2020). The Role of Tumor-Associated Macrophages in the Progression and Chemoresistance of Ovarian Cancer. *Cells* 9, 9. doi:10.3390/cells9051299

Park, S., Choi, J., Biering, S. B., Dominici, E., Williams, L. E., and Hwang, S. (2016). Targeting by AutophaGy Proteins (TAG): Targeting of IFNG-Inducible GTPases to Membranes by the LC3 Conjugation System of Autophagy. *Autophagy* 12, 1153–1167. doi:10.1080/15548627.2016.1178447

Penninkilampi, R., and Eslick, G. D. (2018). Perineal Talc Use and Ovarian Cancer. *EPIDEMIOLOGY* 29, 41–49. doi:10.1097/ede.0000000000000745

Stewart, C., Ralyea, C., and Lockwood, S. (2019). Ovarian Cancer: An Integrated Review. *Semin. Oncol. Nurs.* 35, 151–156. doi:10.1016/j.soncn.2019.02.001

Stockwell, B. R., and Jiang, X. (2019). A Physiological Function for Ferroptosis in Tumor Suppression by the Immune System. *Cel Metab.* 30, 14–15. doi:10.1016/j.cmet.2019.06.012

Su, Y., Zhao, B., Zhou, L., Zhang, Z., Shen, Y., Lv, H., et al. (2020). Ferroptosis, a Novel Pharmacological Mechanism of Anti-cancer Drugs. *Cancer Lett.* 483, 127–136. doi:10.1016/j.canlet.2020.02.015

Wang, W., Green, M., Choi, J. E., Gijón, M., Kennedy, P. D., Johnson, J. K., et al. (2019). CD8+ T Cells Regulate Tumour Ferroptosis during Cancer Immunotherapy. *NATURE* 569, 270–274. doi:10.1038/s41586-019-1170-y

Wang, Y., Zhao, G., Condello, S., Huang, H., Cardenas, H., Tanner, E. J., et al. (2021). Frizzled-7 Identifies Platinum-Tolerant Ovarian Cancer Cells Susceptible to Ferroptosis. *CANCER RES.* 81, 384–399. doi:10.1158/0008-5472.can-20-1488

Wilkinson, H. N., Upson, S. E., Banyard, K. L., Knight, R., Mace, K. A., and Hardman, M. J. (2019). Reduced Iron in Diabetic Wounds: An Oxidative Stress-dependent Role for STEAP3 in Extracellular Matrix Deposition and Remodeling. *J. Invest. Dermatol.* 139, 2368–2377. doi:10.1016/j.jid.2019.05.014

Xu, P., Zhang, G., Hou, S., and Sha, L.-g. (2018). MAPK8 Mediates Resistance to Temozolomide and Apoptosis of Glioblastoma Cells through MAPK Signaling Pathway. *Biomed. Pharmacother.* 106, 1419–1427. doi:10.1016/j.biopha.2018.06.084

Yanatori, I., Richardson, D. R., Imada, K., and Kishi, F. (2016). Iron Export through the Transporter Ferroportin 1 Is Modulated by the Iron Chaperone PCBP2. *J. Biol. Chem.* 291, 17303–17318. doi:10.1074/jbc.m116.721936

Yang, L., Tian, S., Chen, Y., Miao, C., Zhao, Y., Wang, R., et al. (2021). Ferroptosis-Related Gene Model to Predict Overall Survival of Ovarian Carcinoma. *J. ONCOL.* 2021, 6687391. doi:10.1155/2021/6687391

Yang, W. S., Kim, K. J., Gaschler, M. M., Patel, M., Shchepinov, M. S., and Stockwell, B. R. (2016). Peroxidation of Polyunsaturated Fatty Acids by Lipoxygenases Drives Ferroptosis. *Proc. Natl. Acad. Sci. USA* 113, E4966–E4975. doi:10.1073/pnas.1603244113

Ye, Z., Liu, W., Zhuo, Q., Hu, Q., Liu, M., Sun, Q., et al. (2020). Ferroptosis: Final Destination for Cancer. *Cell Prolif* 53, e12761. doi:10.1111/cpr.12761

Yeung, T.-L., Leung, C., Li, F., Wong, S., and Mok, S. (2016). Targeting Stromal-Cancer Cell Crosstalk Networks in Ovarian Cancer Treatment. *Biomolecules* 6, 3. doi:10.3390/biom6010003

Zamarin, D. (2019). Novel Therapeutics: Response and Resistance in Ovarian Cancer. *Int. J. Gynecol. Cancer* 29, s16–s21. doi:10.1136/ijgc-2019-000456

Zhang, X., Du, L., Qiao, Y., Zhang, X., Zheng, W., Wu, Q., et al. (2019). Ferroptosis Is Governed by Differential Regulation of Transcription in Liver Cancer. *Redox Biol.* 24, 101211. doi:10.1016/j.redox.2019.101211

Zhang, Z., Guo, M., Li, Y., Shen, M., Kong, D., Shao, J., et al. (2020). RNA-binding Protein ZFP36/TTP Protects against Ferroptosis by Regulating Autophagy Signaling Pathway in Hepatic Stellate Cells. *Autophagy* 16, 1482–1505. doi:10.1080/15548627.2019.1687985

Zhu, S., Zhang, Q., Sun, X., Zeh, H. J., Lotze, M. T., Kang, R., et al. (2017). HSPA5 Regulates Ferroptotic Cell Death in Cancer Cells. *Cancer res.* 77, 2064–2077. doi:10.1158/0008-5472.can-16-1979

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

REVIEW
published: 25 November 2021
doi: 10.3389/fgene.2021.759662

The History and Challenges of Women in Genetics: A Focus on Non-Western Women

*Hadeel Elbardisy and Malak Abedalthagafi**

Genomics Research Department, Saudi Human Genome Project, King Fahad Medical City, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

"Women in much of the world lack support for fundamental functions of a human life." This truthful portrait was pointed out by Martha Nussbaum in her book "*Introduction: Feminism & International Development."* Throughout history, gender inequality has been persistent in many aspects of life, including health and empowerment. Unfortunately, this inequality has not been excluded from the field of science. Perpetual assumption that women's absence or restriction to secondary roles in various disciplines is an acceptable law of nature misrepresents women's contribution to science and maintains hurdles for participation in the future. According to a recent UNESCO's report, women make up only 30% of researchers worldwide. But despite all the obstacles, women made major contributions with discoveries that shaped the progress in many scientific fields. In the field of genetics, Rosalind Franklin is an example of unwittingly compromised women's scientific achievements. Franklin was an expert in X-ray crystallography; her data, especially the "photo 51," was critical to James Watson and Francis Crick along with their own data to publish the discovery of the double helix DNA structure in 1953. Her contribution was acknowledged posthumously in Watson's memoir in 1968. Barbara McClintock was a 20th century American cytogeneticist who remains up to date the only woman receiving an unshared Nobel prize in Physiology or Medicine. McClintock dedicated her work to cytogenetics and discovered the phenomenon of mobile genes. Her research was initially subjected to skepticism in the 1950s. It was not until the late 1960s that the community realized the significance of McClintock's discovery. The history of science is occupied with a myriad of similar tales of such inspiring women that, after tremendous struggles, thrived and achieved breakthroughs in their respective fields. It is prominent our limited knowledge of women's experience and struggle in science in non-western world. Addressing the stories of this outstanding minority is critical to expand the understanding of the gender disparity factors embedded in diverse cultures. In this article, we attempt to put the spotlight on some fascinating non-western women and their significant contributions to the field of genetics.

Keywords: women in science, non-western, genetics, gender, career

OPEN ACCESS

Edited by:
Thomas Liehr,
Friedrich Schiller University Jena, Germany

Reviewed by:
Aparna Banerjee,
Catholic University of the Maule, Chile
Iris Paola Guzmán-Guzmán,
Autonomous University of Guerrero, Mexico
Heike Petermann,
University of Münster, Germany

****Correspondence:***
Malak Abedalthagafi
malthagafi@kacst.edu.sa

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *16 August 2021*
Accepted: *26 October 2021*
Published: *25 November 2021*

Citation:
Elbardisy H and Abedalthagafi M (2021) The History and Challenges of Women in Genetics: A Focus on Non-Western Women. Front. Genet. 12:759662. doi: 10.3389/fgene.2021.759662

WOMEN IN GENETICS

Today women are able to succeed in a still male-dominated science community and prove their pivotal roles, although there are still many significant obstacles present, including social norms, political systems, and religious backgrounds resulting in gender disparities. Policymakers' lack of awareness reform these obstacles sterner with less priority given to address the gender gap (Andres, 2011). Women under representation in the science field could also be rooted to a diversified restrain faced in higher education, career path, working environment, role stereotypes and the family work balance (Handelsmann et al., 2005). In a recent study, gender inequality in scientific careers was analyzed through a large scale longitudinal bibliometric analysis. The analysis showed that women starting their publication career have proportionally increased over the course of recent years from 30% (2000) up to almost 40%. Yet, male researchers often publish an average of 15–20% more than female researchers (Boekhout et al., 2021). Also, the tendency that women discontinue publications was slightly higher than men–an indication of their dropout and it revealed that about 25% of men are more often the last authors than women of similar career years–an indication of more senior positions (Sanderson, 2021; Boekhout et al., 2021). This gender imbalance is also presented in high prestigious research awards as Nobel prize, etc. Between 2001 and 2020, a total of 2011 men were awarded compared to 262 women only (Watson, 2021). Worth mentioning, female scientists annual share of awards has raised from 6% up to 19% between 2016 and 2020 (Meho, 2021). But the gender disparity is still consistent if we consider the average numbers of full-time female researchers particularly in biological, life sciences, computer science and mathematics (Meho, 2021). In 2020, Emmanuelle Charpentier and Jennifer Doudna were announced winners of Nobel prize in chemistry for the development of the CRISPR/Cas9 genome editing methodology, a revolutionary novel tool for gene editing (Nobel Prize, 2020). The CRISPR/Cas9 system is present naturally in archaea and bacteria, acting as a defensive shield against pathogens and thus providing immunity against these viruses and plasmids (Terns and Terns, 2011). The system works as precise genetic scissors, allowing double strand cleavage at specific regions in the DNA that are complementary to a subset of mature CRISPR RNAs (crRNAs). The crRNAs base pairs form a dual RNA with trans-activating crRNAs (TracrRNA). These dual RNAs are able to target specific sites and sequences within the genomic DNA permitting precise genome editing (Jinek et al., 2012). This CRISPR technology could be applied for genome-wide screening, editing of gene coding sequences, epigenome editing and transcriptional regulation. Thus, the CRISPR technology represents a promising therapeutics tool for genetic disorders and targeted cancer therapies (Barrangou and Doudna, 2016). Moreover, CRISPR genome editing represents an exciting new tool for wide range of applications in the agricultural, food, and industrial biotechnology sciences (Barrangou and Doudna, 2016). Throughout the awards history, this is the first time for two women to share a Nobel prize in chemistry (Rincon, 2020). In an interview with C and EN, Doudna said "It certainly makes me happy that it could be the case that because two women were involved in the early days of CRISPR that we could have established a culture that is welcoming to other women in the field. That's kind of cool." This was her answer to if she believes that CRISPER could be the unique tool that welcomes and pave the road for female scientists (Satyanarayana, 2020).

By the end of 2019, the world witnessed the emergence of a new coronavirus leading to the ongoing COVID-19 pandemic. Unexpectedly people's daily life came to a pause with lockdowns imposed all over the world (Chams et al., 2020). History portrays human survival against several outbreaks and pandemics affecting thousands to millions lives. Although our advancements in technology and medicine are of critical importance, we came to realize our vulnerability towards this novel pathogen. The escalating rates in mortality and morbidity alerted the science community to act rapidly to develop strategies against this pandemic (Hu et al., 2021). From understanding the virus, it's pathogenesis, methods of diagnosis and description of clinical manifestations to genome sequencing and using molecular and genetic data to seek a therapeutic route or effective vaccination, scientists were in a race against the pandemic spread of COVID-19. Several women were at the frontline of developing COVID-19 vaccines (Bora, 2021). Professor Sarah Gilbert, the women behind the Oxford/AstraZeneca vaccine, led a whole team in a race to develop a vaccine and push for the preclinical and clinical testing (BBC, 2020; Lane, 2020). Her previous work on malaria vaccine research, which focused on developing vaccines that preferably trigger a T-cell response over triggering B-cell antibody responses alone paved the route for creating vaccines that contain specific antigens within the viral host, a technology known as recombinant viral vector vaccines which outweighed the risk concerns associated with traditional live attenuated vaccines. Gilbert's team focused on this technology in their work on the COVID-19 vaccine (Lane, 2020). Over a year, the Oxford/AstraZeneca vaccine was approved and now it is being administrated worldwide to impede the virus spread (WHO, 2020). A work of excellence credited to the oxford team which composed of two third female researchers led by Dr Gilbert (PA Media, 2021). At a point of her life, Dr Gilbert did consider leaving science for good especially during her doctorate study. Back then, she was fueled with energy to gain experience from diverse disciplines but was faced with "one subject focus" ideology. Luckily for us, she did not leave science, although her career got more challenging after giving birth to triplets (BBC, 2020). She is a true example of a committed scientist and a devoted mother who excelled at both. Diverse heroines also had outstanding roles in the development of novel vaccination strategies to combat COVID-19 (Bora, 2021; Romero et al., 2020). Dr Kizzmkia Corbertt, an African American scientist, had an eloquent role in the development of Moderna's COVID-19 vaccine. She also took a deepened commitment to outreach communities of color to alleviate the vaccine skepticism (Subbaraman, 2021). Dr Özlem Türeci, a notorious scientist, physician and entrepreneur who was instrumental in developing the first FDA approved mRNA vaccine (BioNTech-Pfizer) within just a year (Bryer, 2021). Dr Türeci, the descent of

TABLE 1 | The average percentage of females involved in R&D positions (full time and Part time jobs) in different regions globally, according UIS fact sheet (UNESCO Institue for Statistics 2020).

Region	Average percentage (%)
Central Asia	48.5
Latin America and the Caribbean	45.8
Arab States	40.9
Central and Eastern Europe	39.0
North America and Western Europe	32.9
Sub-Saharan Africa	31.1
East Asia and the Pacific	25.0
South and West Asia	23.1

Turkish immigrants and the cofounder of BioNTech company, confronts the gender inequality by adopting a balanced workforce in her company with 54% being females. She credited the rapid vaccine release to this equitable workforce (PA Media, 2021). These scientists and many more show the tangible impact of women's inextricably role during the global crisis. In this context, the power of role models for young girls is critical. Mattel company released six Barbie dolls to honor women in science and their contributions in the fight against COVID-19 in last August (Joly and Shea, 2021)- a splendid step to reform the image of an unrealistic plastic doll into a doll depicting real and successful women scientists as role model for our future generations of female scientists.

NON-WESTERN WOMEN IN GENETICS

A study discussing the gender inequality paradox in STEM fields pointed out that women living in non-western countries with a greater gender equality gap are more likely to be engaged in STEM as a result of the society pressure and pursuing an improved overall quality of life (Stoet and Geary, 2018). The UNESCO Institute for Statistics fact sheet released in June 2020 showed that women in R&D represent 25% for East Asia and the pacific, 23.1% for South and West Asia (~50%) and 40.9% for Arab States, while only 32.9% of researchers were female in North America and Western Europe (**Table 1**) (UNESCO Institue for Statistics, 2020). Here, we present selected recounts of nonwestern women from these regions, our selection is based on several reasons. To begin with, regions like East Asia and the Pacific progress to narrow gender disparities in diverse fields is tardy over the years with 2.5% improvement from 2006 to 2019, a concerning alarm for a region of one of highest women percentages globally (1.13 billion women) (World Economic Forum, 2020). On contrary, Southeast Asia region attains a grappling progress towards narrowing the gender gap (Bekhouche, 2013). Important factors mastered this improvement. For example, Singapore ranks number one as one of the safest location for women in Asia Pacific (Evlanova, 2019). With a powerful law protecting women's rights and a labor force comprised of 60% women, resources to fair treatment for women outstand clear in Singapore (Setianto, 2020). To bring visibility to today's women's position in these regions, we selected three stories particularly in the field of genetics representing South Korea, Singapore, and Thailand.

Narry Kim, a South Korean molecular cell biologist, has made critical contributions to our understanding of RNA biology (The Royal society, 2021). Her journey began with receiving her BSc from Seoul National University in 1992 and her PhD from Oxford University in 1998, where her research focused on retroviral proteins and their role in constructing gene transfer vectors. Her post-doctoral research took place in the laboratory of Gideon Dreyfuss at the University of Pennsylvania in Philadelphia where she worked on studying mRNA surveillance pathways (Cell Symposia, 2019). Organisms have these pathways to ensure the fidelity of mRNA during its biogenesis (Hoof and Wagner, 2011). In 2001, She returned to Seoul National University (SNU) as a faculty member, and by 2013 she became a SNU distinguished Professor (Seoul National University, 2021). Her research team is focused on the control of gene regulation by RNA. RNA molecules have a crucial role in post-translational gene regulation and are key players in some diseases (Mattick, 2011). Understanding their pathways and mechanisms of action promises novel insights and the development of improved therapeutic solutions, for example in cancer therapies and stem cell engineering (Narry Kin Lab, 2021). Three main research directions [microRNA (miRNA), RNA tails, and RNA binding proteins (RBP)] are her laboratory's main topics of interest (Seoul National University, 2021). Dr. Kim's group was able to elucidate the mechanisms of miRNA biogenesis through two sequential steps: pre-miRNAs generation in the nucleus and processing of these pre-miRNA into mature miRNAs in the cytoplasm (Lee et al., 2002). They identified several key factors within the pathway such as DROSHA, DGCR8, RNA polymerase II and a terminal nucloetidyl transferases known as uridyltransferases (The Royal society, 2021). Her research group was able to identify other key factors, including the RNA binding proteins Lin28a and Lin28b important in stem cell programming (Heo et al., 2008). They also developed a novel experimental tool called TAIL-seq for genome-wide screening for mRNAs tails [poly(A) and 3′ end modification]. This tool has helped to reveal the roles of these mRNA tails in diseases (Chang et al., 2014). In an approach to understand the complexity of RNA binding proteins (RBP), Kim's lab has founded a proteomic facility to develop novel techniques to scrutinize RBP networks (Seoul National University, 2021). Narry Kim was recognized nationally and internationally by the scientific community for her significant contributions. In 2007, she was named Woman Scientist of the Year by the Ministry of Science and Technology of South Korea. She received the L'Oreal-UNESCO Women in Science Award in 2008. By that time, Dr Kim was one of the very few female scientists being recognized in the Asian pacific region. In an interview upon her winning, Dr Kim stated the limited independent positions and almost no leadership positions for female researchers must not endure. She already witnessed the loss of talented females in the field and called for more efforts to be done by the government to push women in science (Kim, 2008). She was named a National Honor Scientist by the Ministry

of Education, Science and Technology in 2010. In 2013, Dr. Kim was awarded the S-Oil Leading Scientist of the Year Award, the Korea S&T Award, and the Gwanak Grand Prize Honor Sector by Seoul National University. In 2017, she received the Chen Award. In 2019, she received the Asian Award in Medicine. She was elected as a Foreign Associate of the prestigious European Molecular Biology Organization (EMBO) in 2013, a Foreign Associate of the US National Academy of Science (NAS) in 2014, and Member of Korean Academy of Science and Technology (KAST) in 2014 (Narry Kin Lab, 2021). Currently, she is the director of the RNA research at the Institute for Basic Science (IBS) leading the research on high resolution mapping of the new coronavirus RNA that will help in finding more accurate strategies against COVID-19 (Kim 2020).

Chanchao Lorthongpanich is a young leading principal investigator and developmental stem cell biologist at the Siriraj Center of Excellence for Stem Cell Research in Bangkok, Thailand (SiSCR, 2021). Early on she understood the urgency of resolving the problem of limited blood supply in hospitals which can be life-threatening for patients. Dr. Lorthongpanich and her research team developed an alternative intervention for blood supply shortage in hospital patients by establishing an *in vitro* production system (UTAR, 2021). They were able to revert differentiated adult cells from patients to a stem cell state and therefore generate patient-specific induced pluripotent stem cells (iPSCs). The adult cells were retrieved from patients suffering from paroxysmal nocturnal hemoglobinuria (PNH), a disease developing from genetic mutations specifically in hematopoietic stem cells, resulting in severe hemolytic anemia, thrombosis and peripheral blood cytopenia (Brodsky, 2014). By utilizing iPSCs to generate autologous hematopoietic stem cells (HSCs), they created HSC without the disease-causing mutation which thus can then be used as a conventional treatment available for PNH. This method of retrieving cells from a patient and directing them to differentiate into HSCs avoids the complications resulting from allogenic HSCs transfusions, lack of matching HLA donors and post-transplant complications (Phondeechareon et al., 2016). In an interview with the journal *Nature*, she referred to her current focus on developing human platelets inside the laboratory, a lifesaving approach to overcome the continuous shortage of platelet donors, especially in Thailand (Nogrady, 2019). Enhancing the large scale production of platelets from hematopoietic stems cells holds the potential of providing abundant donor-independent platelets that would be lifesaving for many patients with critical conditions (Bangkok UNESCO, 2018). She advocates for Thailand to be an ideal destination to establish biotechnology factories due to the decreased expenses of living (Nogrady, 2019). Lorthongpanich addresses the burden of inadequate funding during her career but a deepened commitment to science and her country steered her will. For that, Chanchao Lorthongpanich was awarded the L'Oreal-UNESCO for Women in Science Awards in 2018 for her outstanding contributions to stem cell research.

Yue Wan is a senior research scientist at the Genome Institute of Singapore, a junior principal investigator at the Agency for Science, Technology and Research (A*STAR), and an adjunct assistant professor in the Department of Biochemistry at the National University of Singapore (Wan, 2021). Her research interest is focused on understanding RNA functional structures and their roles in the regulation of cellular processes. Her interests include the impact of RNA structures on cellular states, RNA interactions networks in different organisms, and the genome of RNA viruses to better understand infectious diseases and their impact on human health (Wan, 2021). Dr. Wan has developed several ingenious novel laboratory tools for the study of RNA. She and her team were able to reveal the intramolecular and intermolecular RNA-RNA interactions in eukaryotes utilizing a high-throughput approach called Sequencing of Psoralen cross-linked, Ligated, And Selected Hybrids (SPLASH) (Aw et al., 2016). They showed that SPLASH could be an informative tool in understanding the complexity of eukaryotes transcriptomes, providing a deeper understanding of how RNAs interact with each other and the surrounding cellular molecules (Aw et al., 2016). Wan and her team also developed a new technology called Parallel Analysis of RNA Structures (PARS) to determine thousands of RNA structure simultaneously, a method that enables better insight of RNA structures to fully comprehend their function in cellular states (Wan et al., 2011; Kertesz and Yue, 2010). Recently, they published their work on developing a nanopore sequencing method called PORE-cupine to determine distinct RNA isoforms of the same gene and their ability to adapt different structural conformations (Aw et al., 2021). Wan was recognized for her research in Singapore and abroad. In 2015, she received the Genome Web Young Investigator Award from the Singapore National Academy of Science, and the Young Scientist Award and EmTech MIT TR35 Asia Honoree, an award given to promising innovators under the age of 35 organized by MIT Technology Reviews (The Branco Weiss Fellowship, 2021). She was the first Singaporean scientist to win the Branco Weiss Fellowship granted by a Swiss based philanthropic organization which every year awards 10 outstanding scientists (Asian Scitentist Newsroom, 2014). In 2016, she was awarded the Ten Outstanding Young Person Award Finalist in Singapore, an A*STAR Investigator position, and the 2016 L'Oréal Singapore For Women in Science National Fellowship (Asian Scientist Newsroom, 2016). As a young mother of two, she strongly supports a family friendly policy to be offered to female researchers that would spin their productivity and wellness (A*Star Talent Times, 2021).

Accounts of an older generation of women scientists shows how influential women have been in the field of science. With a false myth of botany being imposed as an "amusement" for women with restricted role as home gardeners, plant gatherers and housewives' herbalist. Studying the science of plants was viewed as a field being exclusively owned by men (Howard, 2001). A lackluster belief that encouraged us to choose and acknowledge Dr. Archana Sharma, the godmother of Botany. Dr. Sharma, being born to an academic family in India, completed her B.Sc. from Bikaner University and pursued her master's and Ph.D. degrees in the Department of Botany at the University of Calcutta (Sopory, 2009). In 1960, she became the second women receiving her doctorate in botany (D.Sc.) from the University of Calcutta, one of the oldest universities in India (Pathiki et al., 2015). Her research was not only focused on botany, but it also expanded to

cytogenetics, human genetics, and environmental mutagenesis. After her education, Dr. Sharma became a faculty member in the department of Botany at the University of Calcutta. By 1972, along with her husband professor AK Sharma, she was appointed a professor of genetics in the Center of Advanced Studies in Cell and Chromosome Research after the foundation of a school of cytogenetics (Shah, 2018). In 1981, she was promoted to Head of the Department of Botany at the University of Calcutta. Dr. Sharma inventing revolutionary novel methods for visualizing chromosome structures that soon became gold standard techniques for plant cell and chromosome research worldwide (Sharma and Archana, 1956). Later, Sharma and her husband published a book summarizing their research and findings on chromosomes, aptly entitled: "Chromosome techniques–theory and practice" (Sharma and Sharma, 1980). This represented an informative repository of molecular and histological techniques, including pre-treatment and hypotonic techniques, fixation, staining and processing of cells, understanding chromosome structures and analysis after culture of cells, and various techniques to visualize the banding patterns of chromosomes (Sharma and Sharma, 1980). This textbook remains very popular as one of the standard curriculums in the field of chromosome research and botany. Prof. Sharma's laboratory produced multiple novel discoveries: her findings of the speciation in the vegetative reproduction of plants raised a new concept of how new species evolve after plants propagate asexually through regular inconsistent chromosome mosaicism (Pathiki et al., 2015). Furthermore, she studied methods of inducing cell divisions in mature nuclei and the causes of polyteny in plants. In human genetics, she focused on comparative studies in genetic polymorphism between normal populations versus those with pathological conditions and investigated the sex abnormalities that are prominent among the Indian population (Nigli et al., 1980). Elevated levels of environmental pollution due to industrial discharges and increased use of pesticides alerted scientists to investigate the effects of this pollution on the Indian population. Her group investigated the impact of pesticides and heavy metals on various biological systems, and the clastogenic impact and hazardous effects on chromosomal abnormalities, mitosis inhibition and cell division (Giri et al., 1984; Agarwal et al., 1990; Mukherjee and Sharma, 1987). She established *The Nucleus*, an international journal for cytology and allied topics of cell and chromosome research (Sopory, 2009). Her research and contributions were recognized nationally and internationally. She was a fellow of the Indian National Science Academy in 1977 and was elected President of the Indian Botanical Society in 1989. Between 1986 and 1987 she was the general president of the Indian Science Congress Association. In 1990, she was a member of the International Academy of Science in Germany (Pathiki et al., 2015). She was awarded the Shanti Swarup Bhatnagar Prize for Science and Technology in 1976. She was the recipient of the Padma Bhushan Award in 1984 and received the Women Scientist Award and the Ashutosh Mukherji Medal by the Indian Science Congress Association in 1999 (Pathiki et al., 2015). Moreover, she was a policy maker in the Indian government as she was a member of the Science and Engineering Research Council and the Environmental Research Council. In addition, she was also a member of the panel for the cooperation with the UNESCO (Sopory, 2009).

The position of African women in STEM remains a prominent concern as poverty, health and education inequality are more potent barriers in Africa. An embedded culture of women's role as an exclusive family caregiver plays a pivotal role of young girls drop out after primary education (Andres, 2011). The racial discrimination reinforces the issue especially when it comes to fair opportunities in funded scholarships, health equity and workforce infrastructure (Ighodaro et al., 2021). In 2015, the African Union announced the Year of Women's Empowerment and Development Towards Africa Agenda 2063. This Agenda had the goal to further develop the Science, Technology, and Innovation Strategy for Africa 2024 into an initiative to advocate for women's inclusion in these areas (Muthumbi and Sommerfeld, 2015) and to promote women in Africa to participate and become key members in the field of science. We selected our upcoming scientist from Nigeria, as it is one of the highest population country in the region with accentuated under-representation of women in research and a paucity of them in senior positions as well (WHO, 2020). Alongside with poverty, religious and cultural obstacles mentioned above (Andres, 2011). We share the inspiring story of Adeyinka Falusi, a scientist that devoted her career to human genetics, bioethics and inherited hematological diseases in her home country. Her early inspiration came from a childhood friend, Grace Olaniyan, and her interest in science. Out of curiosity, she read Grace's science textbooks and became deeply interested in the topics discussed. Adeyinka Falusi earned her M.Phil. and PhD at the University of Ibadan in 1986, with her research dedicated to elucidating the various types of anemia and sickle cell diseases (SCD) as inherited genetic disorders in the Nigerian population. Along with her team, she was able to screen for and discover exclusive genetic markers for sickle cell anemia (kulozik et al., 1986), a remarkable paradigmatic shift in the prevalence of published sickle cell variants. Her research extended to the molecular epidemiology of the compounding impact of several hematological diseases including Malaria and Thalassemia (Higgs et al., 1986; Fey et al., 1990). She also studied the genetics of breast cancer among Nigerian women (Yonglan et al., 2018), which was one of the first and foremost genetic studies conducted in the Nigerian population. Her work in the field of SCD granted her a L'OREAL UNESCO Outstanding Woman of Science Award in 2001, and she received the Rare Gem Award in the Category of Science and Technology in 2003. She was awarded the National Productivity Order of Organization for Women in Science for Developing Worlds (Adeyinka, 2021) and received a Vocational Excellence Award for Impact in Science in 2014. Recognizing the lack of community awareness for SCD, she founded the Sickle Cell Association of Nigeria (SCAN) and later the Sickle Cell Hope Alive Foundation (Adeyinka, 2021; The Network of African Science Academies, 2017). In an interview, she recounted that 1 day her director encouraged her to engage with the community and share her knowledge with the people of Nigeria. That day she went to the local Yemetu and Adeoyo hospitals and started spreading awareness, support, and information, which later

morphed into the establishment of SCHAF where she donated her retirement money (Gesinde, 2020). Under her leadership, the foundation promoted community awareness with free parent's handbooks for sickle cell patients (The Network of African Science Academies, 2017). One of her other notable contributions is her eminent role in research bioethics. She extensively worked on the development of institutional ethics review board at her university that grew to an ethical model review board nationally. From 2001 to 2005 she served as the chairperson of the University of Ibadan/University Collage Hospital Institutional Review Committee, where she was keen to restructure the review committee and establish the first guidelines for ethical reviews in the University. She was the country coordinator for Nigeria in the Networking for Ethics of Biomedical Research in Africa (NEBRA), promoting research ethics in central and western Africa from 2005 to 2006. Her work in bioethics was awarded with the EDCTP award for her outstanding role in establishing research ethics review boards in Africa. Dr. Falusi also established the Nigerian Bioethics Initiative (NIBIN) (African Success, 2009), and serves as a truly inspiring example of an African women scientist. She was privileged with support from her husband and family, an advantage not common in Nigeria. Although, she sacrificed many years prior joining academics to take care of her five children yet remained persist to achieve her dream and utilize her knowledge to help Nigerian people (Gesinde, 2020).

Although women in the Middle East have made progress in STEM fields and R&D, with the percentage of women exceeding 40% (UNESCO Institue for Statistics, 2020), there are still challenges faced by women in this cultural environment. Social norms, gender inequality, religious background, early marriage, and childbirth are glimpses of the many obstacles' women are confronted with in the Middle East. These societal obstacles, together-a lack of equal opportunities, salaries and support, cause woman to still struggle to balance a healthy work–family life. The dominating ideology that men should be the primary source of income is one of the root causes of this inequality and salary gaps. It is estimated that three out of four women do not work even after completing their education (Gatti et al., 2013). In some countries like Egypt, Jordan and Tunisia women with higher education degrees suffer lower employment opportunities than women with lower educational levels. On the contrary, Gulf countries are promoting employment opportunities for educated women (International labour office, 2016). A statement that influenced our selection from this region to represent fruitful stories from both proportions. On the bright side, there is a rising awareness for the need to support employment of women, with governments pushing women's enrollment in education, labor force and managerial positions (Patel, 2019).

We here present some Arab women scientists that made remarkable contributions in the field of genetics in the Middle East. Nadia Sakati is a distinguished pediatrician who revolutionized genetics in Saudi Arabia. She had a dream of being a doctor wearing a white gown since she was in grade 8 (Takreem, 2018). By 1965, she had completed her medical degree from the Medical School of Damascus University. Dr. Sakati then worked as a pediatric resident at the American University of Beirut and the Jackson Memorial Hospital in Miami in 1966. By 1969, she became a fellow in the Genetics and Metabolism Department at the University of California. She joined King Faisal Specialist Hospital and Research Centre (KFSHRC) in Riyadh in 1978, where she established one of the first genetic departments in Saudi Arabia (Cadogan, 2020). With a high rate of consanguinity in the population along with prominent genetic disorders among children in Saudi Arabia, the establishment of this Department of Genetics was crucial to unearth the role of genetics in such cases. During her stay in KFSHRC, she held positions as chairman of Pediatrics from 1987 to 1989, and as director of the Genetics/Endocrinology and Metabolic Disease Fellowship Program from 1989 to 1995. Dr. Sakati was the head of the Department of Medical Genetics from 1995 to 2001. Her major breakthrough was the discovery of three rare genetic syndromes that were named after her. Sakati-Nyhan-Tisdale syndrome, a disease she reported in 1971 along with her colleagues William Leo Nyhan and William Tisdale, was described in an 8 year old boy with bone malformations and congenital heart disease (Sakati et al., 1971). Sanjad-Sakati syndrome, another rare genetic condition that was first recorded in Saudi Arabia by Sakati along with Sami A. Sanjad in 1991, is characterized by congenital hypoparathyroidism associated with severe failure to grow along with dysmorphism (Sanjad et al., 1991). This discovery helped with the diagnosis of children showing similar phenotypes and brought a step closer the mapping of rare genetic disorders across Saudi Arabia (Sanjad et al., 1991). Woodhouse-Sakati syndrome, the third rare genetic disorder discovered by Sakati in collaboration with Nichols Woodhouse in 1983, was discovered by investigating seven Saudi patients suffering from hypogonadism, alopecia, diabetes mellitus, mental retardations, and deafness, along with ECG abnormalities (Sakati and Woodhouse, 1983). Her discoveries and work on these rare diseases are summarized in two books she coauthored with William L. Nyhan entitled "Genetic and Malformation Syndrome in Clinical Medicine" (1976) and "Diagnostic Recognition of Genetic Disease" (1987) (Nyhan and Sakati, 1976; Nyhan and Sakati, 1987). Due to Dr. Sakati's outstanding accomplishments and commitment to genetic research in Saudi Arabia, His Royal Highness, King Fahad bin Abdulaziz granted her the Saudi Nationality in 1993 (Takreem, 2018). In 2001, she was recognized as a distinguished senior consultant in KFSHRC, a position she still holds, and in 2018 she was awarded The Special Distinction Award by Takreem, an organization that aims to honor Arab laureates and recognize their great achievements (Takreem, 2018). During her 40-years tenure at King Faisal Hospital, she trained and educated many Saudi residents and fellows, and her discoveries changed the life of many families in Saudi Arabia by providing hope of having healthy children through chorionic villus sampling and segregation analysis, and by being able to allow families to comprehend the diagnosis of their children with a controlled management and treatment plan.

From Egypt, we want to honor Samia Aly Temtamy, the founder of human genetics in Egypt. Her inspiration to be a physician was seeded when she was only 10 years old. She fell ill

with typhoid fever and was treated by Dr. Ibrahim Nagui, whom she subsequently considered a role model and an inspiration to become a doctor. Her passion for medicine grew when she started reading her older brother's medical textbooks (Temtamy, 2019). She was one of the first female graduates in her class at the Faculty of Medicine at Cairo University in 1957, where then-president of Egypt Gamal Abd Elnasser bestowed her the graduate certificate. She spent 1 year as an intern at Cairo University Hospital, followed by 2 years as a pediatric resident at Cairo University Children Hospital. In 1960, She completed her Diploma in Child Health (D.C.H) from Faculty of Medicine, Cairo University. During her residency, children suffering from various congenital malformations were seen by her-at a time when very limited knowledge and research was available in this area. While accompanying her husband who completed his fellowship training in cancer research at John Hopkins University (Temtamy, 2019), she was accepted into a new PhD program in Human Genetics at Johns Hopkins University. Temtamy earned her PhD degree in 1966 as one of the first Arab females to earn a doctorate in this field (National Research Centre in Egypt, 2021). During her PhD research, she was able to propose for the first time a classification of hand malformations based on anatomical positions of these anomalies and their morphology. She later published a book with her mentor Victor McKusick summarizing her extensive work on hand malformations entitled "The Genetics of Hand Malformation" (Temtamy and McKusick, 1978). In addition to this brilliant nosology, she discovered over 30 syndromes in collaboration with coworkers through patients she investigated (Temtamy, 2019). Several of these syndromes were named after her, including Temtamy syndrome (Temtamy et al., 1996) and Temtamy Preaxial Brachydactyly syndrome (Li et al., 2010). After completing her PhD, she was determined to return to Egypt to establish the field of human genetics in her country and deliver the knowledge she gained during her stay in the US. In 1977, she founded the first human genetics department in the National Research Center. Later, it expanded until it became the Center of Excellence of Human Genetics in 2014. The Center is now comprised of eight departments including Clinical Genetics, Orodental Genetics, Cytogenetics, Prenatal Diagnosis and Fetal Medicine, Biochemical Genetics, Immunogenetics, Medical Molecular Genetics, and Molecular Genetics and Enzymology, with more than 200 researchers enrolled. Dr. Temtamy initiated a national program for neonatal screening and served as its principal investigator. This program screened over 15,000 newborns and identified a high frequency rate of Hypothyroidism and PKU among the Egyptian population (Temtamy, 1998). In 1985, she was appointed the head of the Division of Genetic Engineering and Biotechnology (National Research Centre in Egypt, 2021). In 1995, she became professor Emeritus of Human Genetics at NRC. She was a vital member in the African Society of Human Genetics and the National Society of Human Genetics. In 2017, Dr. Temtamy was invited by the Human Genome Organization (HUGO) and awarded the HUGO African prize for lifetime contributions to human genetics. In Egypt, she was awarded the Nile Prize by the Academy of Scientific Research and Technology in 2011, and the State Prize of Merit in Medical Sciences by the Egyptian Academy of Scientific Research and Technology in 2000 (The women and memory forum, 2018). Unfortunately, in June 2021 Samia Temtamy passed away, leaving behind a treasury of knowledge in genetic diseases. She will be remembered as a role model of an ambitious, successful, and determined woman.

From Jordon hails to Rana Dajani, a scientist and feminist that committed herself to speak up for Arab women and their perspective challenges faced in science. Her impressive scientific journey started at the University of Jordon, where she completed her bachelor's degree and master's degree with first honor awards, and then pursued her PhD in molecular cell biology at the University of Iowa in 2005 (HU University, 2021). Being a young mother by that time, her spouse resigned his job to relocate to the States providing the family support during her career progression (Abedalthagafi 2018). Her research focused on cell signaling and inter- and intra-regulatory networks within a cell through interdisciplinary approaches (HU University, 2021). Her research interests also extended to genome-wide associations studies (GWAS) in the fields of diabetes, cancer, and stem cell research. Her research on stem cells initiated the necessity to establish the foundations of stem cell research ethics laws in Jordan (Hauser, 2017). She pleaded for the theory of biological evolution in Islam (Hauser, 2017). She is an international expert on the genetics of Circassia and Chechen populations in Jordan. Currently, Dr. Dajani is a tenured professor of biology and biotechnology at the Hashemite University in Jordan. She has been awarded a 2019–2021 Zuzana Simoniova Cmelikova Visiting Scholar Award at the Jepson School of Leadership Studies at the University of Richmond, the first institution dedicated to leadership education. Prior to this, she has been a fellow at the Radcliffe Institute for Advanced Studies at Harvard University. During her stay at Radcliffe, she dedicated some of her time to publish her book "Five Scarves: Doing the Impossible-If we can reverse cell fate, why can't we redefine success?" (Dijana, 2018). In her book, she documents her personal journey from growing up and starting a family to being a genetic expert and advocating for biological evolution within an Islam perspective. The title "Five scarves" serves as a reference to her own roles as mother, teacher, scientist, social entrepreneur, and feminist. The book is a brilliant approach to document the hurdles she faced as an Arab Muslim woman in academia, and reveals the different views and women's experiences in various cultures and religions. Moreover, she wrote several articles in *Nature* about women's education and struggles in science as part of her enduring efforts for women's liberation (Dajani, 2012). She also is the president of the Jordan chapter for the Organization for Women in Science for the Developing World, an organization that provides training opportunities for women scientists. She is an advisor for the UN Women Advisory Jordan Council and founded a mentoring program named *The Three Circles* that provides support for Arab females in science (Dajani, 2021a). She received several other honors as well, including the Eisenhower and two Fulbright fellowships, and was invited as visiting professor to Yale University and as a visiting scholar to Cambridge University (Dajani, 2021b). In 2015, She was chosen by *Arabian Business* to be among the 100 most powerful Arab women in science and

healthcare (Arabian Business, 2015). Beyond being a trailblazer for women, Dajani established a non-governmental organization known as *We Love Reading*–an idea born upon her return to Jordan from the United States where she realized the scarcity of community libraries in Jordan. She started by recounting weekly story sessions for children in her community mosque that later evolved to establish an organization to nourish the love of reading in children from a very young age. Her essential dogma is that reading is a powerful tool that every child needs to conquer under any circumstance. This program has trained around 700 women to be storytellers, built more than 300 community libraries across Jordan and refugee camps, and distributed more than 250,000 books worldwide reaching as far as Mexico, Turkey, Thailand, Azerbaijan, and Uganda (Arabian Business, 2015). Dr. Dajani earned numerous awards, including the UNHCR Nansen Refugee Award for the Middle East, the Synergos Award for Arab World Social Innovators and a WISE award. Nationally, she was honored with the Order of Al Hussein for Distinguished Contributions of the Second Class in 2014, earned by those who made distinguished contributions to the Jordan society. She was acknowledged by his Majesty King Abdullah II of Jordan as a women leader in 2015 (Dajani, 2021a).

SUMMARY

Throughout history, powerful women changed the course of events, reformed policy decisions, asserted their stance and defied the status quo. Queens Cleopatra of Egypt and Victoria of the United Kingdom played central historic roles (Froelich, 2020). At the same time, however, women that have not been fortunate enough to have a royal or upper-class heritage have very often been subjected to gender inequality, cultural and societal hurdles, and experienced countless difficulties imposed on their education and careers (Andres, 2011). This review, aimed to specifically recognize and celebrate non-western women in the field of genetics, attempts to shed light on the life stories and outstanding accomplishments of female scientist that did not accept the stereotypical roles that society had tried to force on them. The selected examples presented here show what non-western women scientists have been able to accomplish. Our selected heroines with diverse backgrounds, each was vulnerable to at least a form of barrier. Narry Kim sets an example of a rigorous woman that currently holds a managerial appointment confronting the fact of women's under-representation in high ranked positions in East Asia (World Economic Forum, 2020). Meanwhile, Thailand is one of the South-East countries that approached the finish line to gender equity in several fields like heath sector (Bekhouche, 2013), the limited fund availability for biotechnology research imposes a hurdle to young female researchers (Nogrady, 2019). Chanchao Lorthongpanich, devoted her research to find cost effective alternatives to resolve blood donation issue in Thailand since thriving in an equally compelling environment never fails to foster women that could make an enormous difference. The impact of adopting potent laws to protect women's rights, governmental imperatives to support women's inclusion in labor workforce is implemented in Singapore. We could relate the rising star Yue Wan, as an example. Prior approaching her mid-thirties, she led a team as a junior PI pioneering in innovative sequencing technology to decipher RNA functional role. On Contrary, we witnessed Adeyinka Falusi's tenacious efforts to pursue her career in a lower women's supportive environment. Her primitive support was based solely on family and spouse. We could categorize older generation examples, Archana Sharma, Nadia Sakati, and Samia Temtamy in a category of conquering the ideology of male dominating society by introducing a new field of science in their respective country. A pressing stance for women's contribution to science when provided the right resources for knowledge and education. Archana had a vital role to India's national development post the independence and setting a blueprint for botany education globally (Khan 2020). Nadia's knowledge and staunch research permitted the discovery of the Saudi genetic profile distinctiveness and establishment of medical genetics in Saudi Arabia (Takreem, 2018) while Samia is the main contributor of founding the center of excellence in human genetics in Egypt (Temtamy, 2019).Our last selected scientist, Rana Dajani, sets a triumphant image of modern muslim women in science. Besides, advocating to abolish the hijab stereotype as an indication of oppression in western communities (Dajani, 2012). She conveyed modesty for evolution and stem cell research in Islam. An eccentric challenge that would only be rectified by empowered education. Resembling the husband's support Adeyina Falusi acknowledged, Dajani's spouse support during her graduate studies was a key for her career nourishment. Being descendant of Palestinian and Syrian parents, she explicitly promoted for educational and mentorship programs in Arab world and in particular refugees' campuses. We believe our selection, however limited, could provide an improved understanding of the diverse hurdles experienced in non-western countries for women. Confronting the society determinants, struggling a family work balance, acceptance of evolution in a religious setting, introducing an ideology of a female favorable environment and many more. We can point out three influential factors that is common among most of our selected scientists: 1) the fundamental impact of travelling abroad for education. 2) The motivated resilience experienced upon family/spouse support. 3) Achieving a balanced work-family or female positive setting. We summarized our listed scientists and the influential factors affecting their career progression weather it is a challenge or a support, a limited attempt to comprehend women's position in these regions (**Table 2**).

Today the world's advancement is driven by progress in science and technology. Many of our most pressing problems, including global warming, viral pandemics, strained energy resources and environmental pollution can effectively only be solved by advances in science and technology (UNESCO Report Science, 2021). For this, we must combine and harness intellectual power of all humans, and hence it is imperative

TABLE 2 | Summarizing our selected scientists and some of the influential factors that affected their career progress.

Female scientists	Region	Selected factors
Narry Kim	East Asia	Under representation of women in managerial and high ranked positions in East Asia
Chanchao Lorthongpanich	Southeast Asia	Gender balanced environment paving the road to address and find solutions for national issues
Yue Wan	Southeast Asia	Gender balanced environment in workforce/laws protecting women's right
Archana Sharma	South Asia	Thriving in a male dominated society
Adeyinka Falusi	West Africa	Under representation of women in research/Poverty/Religious/Cultural factor/Family support
Nadia Sakati	Middle East	Thriving in a male dominated society/Promoting employment opportunity
Samia Temtamy	Middle East	Thriving in a male dominated society/Challenging employment opportunity
Rana Dajani	Middle East	Muslim women Stereotyping/Modesty in Science/Family support

that women must be part of this process. Specifically, the female population in the non-western world represents a vast reservoir of intellectual potential that has not been allowed to effectively contribute to scientific and technological advances. To change this dilemma, women must be empowered and enabled to have the same education paths and career choices available to them as their male counterparts. National and international initiatives need to promote women for managerial and leadership positions, and not stop at ensuring only basic education for young girls. This change will necessitate an ongoing reformation of social and societal attitudes to enable women to thrive in education and career. While much more needs to be done, governmental entities and science institutes in non-western countries has been taking important steps to enable women's access to equal rights, education, and career opportunities.

With the immense knowledge gained by the sequencing of the human genome, the deciphering of cellular signaling pathways in normal and disease states, and the development of novel therapeutic approaches, genetics and related fields are at the forefront of a medical revolution of an unprecedented magnitude. Ethical and legal concerns associated with this progress will need to be addressed on an ongoing basis. Steering and controlling this scientific and medical revolution and associated regulatory issues will be more difficult without considering women a pillar of our science community and the establishment of a supportive environment encouraging more women to be involve in STEM.

AUTHOR CONTRIBUTIONS

HE wrote the manuscript; MA designed the study and edited the manuscript. Both authors finalized and approved the content.

ACKNOWLEDGMENTS

The authors thanks King Abdulaziz City for Science and Technology and the Saudi Human Genome Project for the support.

REFERENCES

Abedalthagafi, M. (2018). A Jordanian Biologist Redefines Success for Women in Science. Available at: https://www.nature.com/articles/d41586-018-05891-7 (Accessed October 24, 2021).

Adeyinka, G. (2021). FALUSI. Schaf. Available at: http://schafng.org/volunteer/prof-adeyinka-g-falusi/ (Accessed August 2, 2021).

African Success (2009). Biography of Adeyinka Falusi. African Success. Available at: https://web.archive.org/web/20140316200256/http://www.africansuccess.org/visuFiche.php?id=755&lang=en (Accessed August 02, 2021).

Agarwal, K., Sharma, A., and Talukder, G. (1990). Clastogenic Effects of Copper Sulphate on the Bone Marrow Chromosomes of Mice *In Vivo*. *Mutat. Res. Lett.* 243 (1), 1–6. doi:10.1016/0165-7992(90)90115-z

Andres, J. T. (2011). *Overcoming Gender Barriers in Science: Facts and Figures*. London, UK: Sci Dev Net.

Arabian Business. 2015. The 100 Most Powerful Arab Women 2015 in Science and Healthcare. Arabian Business. March 02. Accessed August 1, 2021

Asian Scientist Newsroom (2016). *Asia's Rising Scientists: Wan Yue*. Singapore: Wildtype Media Group Pte Ltd. Available at: https://www.asianscientist.com/2016/12/features/asias-rising-scientists-wan-yue/ (Accessed July 15, 2021).

Asian Scitentist Newsroom (2014). *Asian Scientist*. Singapore: Wildtype Media Group Pte Ltd. Available at: https://www.asianscientist.com/2014/07/topnews/wan-yue-singaporean-win-branco-weiss-fellowship-2014/ (Accessed July 19, 2021).

AStar Talent Times (2021). Uncovering Hidden Pattern. Available at: https://www.a-star.edu.sg/docs/librariesprovider1/default-document-library/news-events/talent-times/a-star-talenttimes-vol5-uncovering-hidden-potential (Accessed Sep 26, 2021).

Aw, J. G. A., Lim, S. W., Wang, J. X., Lambert, F. R. P., Tan, W. T., Shen, Y., et al. (2021). Determination of Isoform-specific RNA Structure with Nanopore Long Reads. *Nat. Biotechnol.* 39 (3), 336–346. doi:10.1038/s41587-020-0712-z

Aw, G. J., Shen, Y., Wilm, A., Sun, M., Lim, X. N., Boon, K.-L., et al. (2016). *In Vivo* Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol. Cel* 62 (4), 603–617. doi:10.1016/j.molcel.2016.04.028

Bangkok UNESCO (2018). Five Women Set the STEM Example: Fellowships to Motivate Next Generation of Researchers. Available at: https://bangkok.unesco.org/content/five-women-set-stem-example-fellowships-motivate-next-generation-researchers (Accessed July 18, 2021).

Barrangou, R., and Doudna, J. A. (2016). Applications of CRISPR Technologies in Research and beyond. *Nat. Biotechnol.* 34 (9), 933–941. doi:10.1038/nbt.3659

BBC (2020). *Prof Sarah Gilbert: The Woman Who Designed the Oxford Vaccine*. London, UKBBC COMPANY. Available at: https://www.bbc.com/news/uk-55043551 (Accessed July 18, 2021).

Bekhouche, Y. (2013). Top Five Countries for Gender equality in South-East Asia. Available at: https://www.weforum.org/agenda/2013/06/top-five-countries-for-gender-equality-in-south-east-asia/ (Accessed Oct 24, 2021).

Boekhout, H., van der Weijden, I., and Waltman, L. (2021). Gender Differences in Scientific Careers: A Large Scale Bibliometric Analysis. Digital library arXiv, Available at: https://arxiv.org/abs/2106.12624.

Bora, S. (2021). She the People: The Women's Channel. Available at: https://www.shethepeople.tv/home-top-video/meet-10-female-scientists-instrumental-in-developing-covid-19-vaccines-around-the-world/ (Accessed july 13, 2021).

Brodsky, R. A. (2014). Paroxysmal Nocturnal Hemoglobinuria. *Blood* 124 (18), 2804–2811. doi:10.1182/blood-2014-02-522128

Bryer, T. (2021). Covid Will Become Manageable: BioNTech Co-founder Says Teh Virus Will Be with Us for Years. Available at: https://www.cnbc.com/2021/09/30/biontech-co-founder-ozlem-tureci-says-covid-will-be-with-us-for-years.html (Accessed Oct 17, 2021).

Cadogan, A. A. (2020). Nadia Sakati. Life in the Fast Lane. Available at: https://litfl.com/nadia-sakati/ (Accessed July 28, 2021).

Cell Symposia (2019). *Cell Symposia*. Amsterdam: Elsevier Inc. Available at: http://www.cell-symposia.com/rnas-2019/bio-kim.asp (Accessed July 10, 2021).

Chams, N., Chams, S., Badran, R., Shams, A., Araji, A., Raad, M., et al. (2020). COVID-19: A Multidisciplinary Review. *Front. Public Health* 8 (383). doi:10.3389/fpubh.2020.00383

Chang, H., Lim, J., Ha, M., and Kim, V. N. (2014). TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3′ End Modifications. *Mol. Cel* 53 (6), 1044–1052. doi:10.1016/j.molcel.2014.02.007

Dajani, R. (2012). How Women Scientists Fare in the Arab World. *Nature* 491 (9), 9. doi:10.1038/491009a

Dajani, R. (2021a). One of the World's Leading Muslim Female Scientists. About Her. Available at: https://www.abouther.com/node/40556/people/leading-ladies/dr-rana-dajani-one-world%E2%80%99s-leading-muslim-female-scientists (Accessed Augst 1.

Dajani, R. (2021b). *PhD.* Harvard University. Available at: https://scholar.harvard.edu/rdajani/biography (Accessed August 1, 2021).

Dijana, R. (2018). Five Scarves: Doing the Impossible - if We Can Reverse Cell Fate, Why Can T We Redefine Success? in *Social Issues, Justice and Status*. New York, USA: Nova Science Publishers.

Evlanova, A. (2019). Top 5 Safest Countries in Asia Pacific for Women. Available at: https://www.valuechampion.sg/top-5-safest-countries-asia-pacific-women (Accessed Oct 24, 2021).

Fey, M. F., Wainscoat, J. S., Mukwala, E. C., Falusi, A. G., Vulliamy, T. J., and Luzzatto, L. (1990). A PvuII Restriction Fragment Length Polymorphism of the Glucose-6-Phosphate Dehydrogenase Gene Is an African-specific Marker. *Hum. Genet.* 84 (5), 471–472. doi:10.1007/BF00195822

Froelich, P. (2020). The 10 Most Powerful Queens in History, from Catherine the Great to Queen Victoria. Available at: https://nypost.com/2020/05/30/the-10-most-powerful-queens-from-catherine-the-great-to-queen-victoria/(Accessed Sep 27, 2021).

Gatti, R., Morgandi, M., Grun, R., Bordmann, S., and Mata Lorenzo, E. (2013). *Jobs for Shared Prosperity:Time for Action in the Middle East and North Africa*. Washington: The world Bank.

Gesinde, T. (2020). *I Almost Blew up the Laboratory in Secondary School —Prof Adeyinka Falusi.* Available at: https://tribuneonlineng.com/i-almost-blew-up-the-laboratory-in-secondary-school-prof-adeyinka-falusi/ (Accessed July 19, 2021).

Giri, A. K., Singh, O. P., Sanyal, R., Sharma, A., Talukder, G., and Talukder, G. (1984). Comparative Effects of Chronic Treatment with Certain Metals on Cell Division. *Cytologia* 49 (3), 659–665. doi:10.1508/cytologia.49.659

Handelsman, J., Cantor, N., Carnes, M., Denton, D., Fine, E., Grosz, B., et al. (2005). Careers in Science. More women in Science. *Science* 309 (5738), 1190–1191. doi:10.1126/science.1113252

Hauser, R. (2017). Harvard Radcliffe Institute *Fellowships/Fellow.* President and Fellows of Harvard College. Available at: https://www.radcliffe.harvard.edu/people/rana-dajani (Accessed August 1, 2021).

Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V. N. (2008). Lin28 Mediates the Terminal Uridylation of Let-7 Precursor MicroRNA. *Mol. Cel* 32 (2), 276–284. doi:10.1016/j.molcel.2008.09.014

Higgs, D. R., Wainscoat, J. S., Flint, J., Hill, A. V., Thein, S. L., Nicholls, R. D., et al. (1986). Analysis of the Human Alpha-Globin Gene Cluster Reveals a Highly Informative Genetic Locus. *Proc. Natl. Acad. Sci.* 83 (14), 5165–5169. doi:10.1073/pnas.83.14.5165

Howard, P. L. (2001). *Women in the Plan World: The Significance of Women and Gener Bias for Botany and for Biodiversity Conservation*. Wageningen: Wageningen University.

Hu, B., Guo, H., Zhou, P., Li-Shi, Z., et al. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* 19, 141–154. doi:10.1038/s41579-020-00459-7

HU University (2021). *Faculty Staff Website.* The Hashemite University ICET. Available at: https://staff.hu.edu.jo/CV_e.aspx?id=0gnQeXkZsCc= (Accessed August 1, 2021).

Ighodaro, E. T., Littlejohn, E. L., Akhetuamhen, A. I., and Benson, R. (2021). Giving Voice to Black Women in Science and Medicine. *Nat. Med.* 27, 1316–1317. doi:10.1038/s41591-021-01438-y

International labour office (2016). *Promoting Women's Empowerment*. Geneva: IFAD.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *science* 337 (6096), 816–821. doi:10.1126/science.1225829

Joly, J., and Shea, L. (2021). *Life In Plastic, It's Fantastic! COVID Vaccine Scientist Gets Barbie*. Euronews. Augst 4. Available at: https://www.euronews.com/2021/08/04/life-in-plastic-covid-19-vaccine-scientist-honoured-with-barbie-doll (Accessed August 5, 2021).

Kertesz, M., Yue, M. E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide Measurement of RNA Secondary Structure in Yeast. *Nature* 467 (7311), 103–107. doi:10.1038/nature09322

Khan, S. J. (2020). Archana Sharma: Sci-illlustrate Stories. Available at: https://medium.com/sci-illustrate-stories/archana-sharma-2f6be6b8d819 (Accessed October 24, 2021).

Kim, Joon. Ha. (2020). The Kaist Herald. May 26. Available at: http://herald.kaist.ac.kr/news/articleView.html?idxno=10256 (Accessed Sep 26, 2021).

Kim, N. (2008). Nature Interviewed V. Narry Kim As One Of the Prize Winners Of the Year (Dec 29). Available at: https://en.snu.ac.kr/snunow/snu_media/news?md=v&bbsidx=71751 (Accessed sep 26, 2021).

Kulozik, A. E., Wainscoat, J. S., Serjeant, G. R., Kar, B. C., Al-Awamy, B., Essan, G. J., et al. (1986). Geographical Survey of Beta S-Globin Gene Haplotypes: Evidence for an Independent Asian Origin of the Sickle-Cell Mutation. *Am. J. Hum. Genet.* 39 (2), 239–244.

Lane, R. (2020). Sarah Gilbert: Carving a Path towards a COVID-19 Vaccine. *The lancet* 395 (10232), 1247. doi:10.1016/S0140-6736(20)30796-0

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., and narry Kim, V. (2002). MicroRNA Maturation: Stepwise Processing and Subcellular Localization. *EMBO J.* 21, 4663–4670. doi:10.1093/emboj/cdf476

Li, Y., Laue, K., Temtamy, S., Aglan, M., Kotan, L. D., Yigit, G., et al. (2010). Temtamy Preaxial Brachydactyly Syndrome Is Caused by Loss-Of-Function Mutations in Chondroitin Synthase 1, a Potential Target of BMP Signaling. *Am. J. Hum. Genet.* 87 (6), 757–767. doi:10.1016/j.ajhg.2010.10.003

Mattick, J. S. (2011). The central Role of RNA in Human Development and Cognition. *FEBS Lett.* 585 (11), 1600–1616. doi:10.1016/j.febslet.2011.05.001

Meho, L. I. (2021). The Gender gap in Highly Prestigious International Research Awards, 2001-2020. *Quantitative Sci. Stud.*, 1–14. doi:10.1162/qss_a_00148

Mukherjee, A., and Sharma, A. (1987). Effects of Cadmium and Zinc on Cell Division and CHromosomal Aberrations in Allium Sativum. *Curr. Sci.* 56 (21), 1097–1100.

Muthumbi, J., and Sommerfeld, J. (2015). *Africa's Women in Science*. Geneva: World Health Organization. Available at: https://www.who.int/tdr/research/gender/Women_overview_piece.pdf (Accessed August 2, 2021).

Narry Kin Lab (2021). RNA Biology Institute for Basic Science and Seoul National University. Institute for Basic Science and Seoul National University. Available at: https://narrykim.org/en/about/(Accessed July 18, 2021).

National Research Centre in Egypt (2021). NRC.SCI.EG. NRC. Available at: https://www.nrc.sci.eg/system/BasicData/basicdata_english_view.php?editid1=2412 (Accessed July 29, 2021).

Nigli, M., Talukder, G., and Sharma, A. (1980). Sex Chromosomal Abnormalities in India. *Trop. Geogr. Med.* 32 (3), 206–215.

Nobel Prize (2020). Press Release: The Nobel Prize in Chemistry 2020. Available at: https://www.nobelprize.org/prizes/chemistry/2020/press-release/ (Accessed July 13, 2021).

Nogrady, B. (2019). The Biotechnologists Making Their Mark on the International Stage. *Nature* 569, S37-S39. 10.1038/d41586-019-01693-7.

Nyhan, W. L., and Sakati, N. (1987). *Diagnostic Recognition of Genetic Disease*. Philadelphia: Lea & Febiger.

Nyhan, W. L., and Sakati, N. (1976). *Genetic & Malformation Syndromes in Clinical Medicine*. First Edition. Missouri: Year Book Medical Publishers.

PA Media (2021). BioNTech Co-founder Says Gender equality Made Vaccine Possible. Available at: https://www.theguardian.com/world/2021/mar/08/biontech-co-founder-says-gender-equality-made-vaccine-possible (Accessed Oct 17, 2021).

Patel, D. (2019). The Indigenous Challenges Facing Arab Women in the Middle East and North Africa Economies. LSE. Available at: https://blogs.lse.ac.uk/

internationaldevelopment/2019/09/06/the-indigenous-challenges-facing-arab-women-in-the-middle-east-and-north-africa-economies/(Accessed July 28, 2021).
Pathiki, S. N., Thammineni, P., and Dharani, V. (2015). *Indian Botanists.* Indian Botanists. March 7. Available at: http://www.indianbotanists.com/2015/03/archana-sharma-indian-woman-botanist.html (Accessed july 26, 2021).
Phondeechareon, T., Wattanapanitch, M., U-Pratya, Y., Damkham, C., Klincumhom, N., Lorthongpanich, C., et al. (2016). Generation of Induced Pluripotent Stem Cells as a Potential Source of Hematopoietic Stem Cells for Transplant in PNH Patients. *Ann. Hematol.* 95 (10), 1617–1625. doi:10.1007/s00277-016-2756-1
Rincon, P. (2020). Two Women Share Chemistry Nobel in Historic Win for 'genetic Scissors. Available at: https://www.bbc.com/news/science-environment-54432589 (Accessed Sep 23, 2021).
Romero, L., Salzman, S., and Folmer, K. (2020). *Kizzmekia Corbett, an African American Woman, Is Praised as Key Scientist behind COVID-19 Vaccine.* Decemeber 13. Available at: https://abcnews.go.com/Health/kizzmekia-corbett-african-american-woman-praised-key-scientist/story?id=74679965 (Accessed July 15, 2021).
Sakati, N., Nyhan, W. L., and Tisdale, W. K. (1971). A New Syndrome with Acrocephalopolysyndactyly, Cardiac Disease, and Distinctive Defects of the Ear, Skin, and Lower Limbs. *J. Pediatr.* 79 (1), 104–109. doi:10.1016/s0022-3476(71)80066-5
Sanderson, K. (2021). More Women Than Ever Are Starting Careers in Science. Available at: https://www.nature.com/articles/d41586-021-02147-9 (Accessed Sep 27, 2021).
Sanjad, S. A., Sakati, N. A., Abu-Osba, Y. K., Kaddoura, R., and Milner, R. D. (1991). A New Syndrome of Congenital Hypoparathyroidism, Severe Growth Failure, and Dysmorphic Features. *Arch. Dis. Child.* 66 (2), 193–196. doi:10.1136/adc.66.2.193
Satyanarayana, M. (2020). Crispr Technology: Where Female Entrepreneurs Thrive. March 8. Available at: https://cen.acs.org/biological-chemistry/gene-editing/CRISPR-technology-Where-female-entrepreneurs-thrive/98/i9 (Accessed September 23, 2021).
Seoul National University (2021). n.d. *Biological Sciences.* BIOSCI. Available at: https://biosci.snu.ac.kr/en/people/faculty?mode=view&profidx=5 (Accessed July 13, 2021).
Setianto, N. (2020). Advancing Gender Equality in Southeast Asia: Case Studies from the Philippines and Singapore. Available at: https://www.internationalaffairs.org.au/australianoutlook/advancing-gender-equality-in-southeast-asia-case-studies-from-the-philippines-and-singapore/(Accessed Oct 24, 2021).
Shah, A. (2018). Intersectional Feminism. FII media Private Limited. Available at: https://feminisminindia.com/2018/07/30/dr-archana-sharma-pioneering-botanist/(Accessed July 26, 2021).
Sharma, A. K., and Sharma, A. (1956). Fixity in Chromosome Number of Plants. *Nature* 177, 335–336. doi:10.1038/177335a0
Sharma, A. K., and Sharma, A. (1980). *Chromosome Techniques Theory and Practice*. london: Butterworth-Heinemann. doi:10.1016/C2013-0-01036-5
SiSCR. (2021). Developmental Stem Cell Biologist. Available at: https://siscr.org/developmental-stem-cell-biologists-dr-chanchao-lorthongpanich-has-recently-joined-siscr/(Accessed July 15, 2021).
Sopory, S. (2009). Archana Sharma Biographical Memoir. Available at: https://www.insaindia.res.in/BM/BM35_0913.pdf (Accessed July 20, 2021).
Stoet, G., and Geary, D. C. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychol. Sci.* 29 (4), 581–593. doi:10.1177/0956797617741719
Subbaraman, Nidhi.. 2021. This COVID-Vaccine Designer Is Tackling Vaccine Hesitancy — in Churches and on Twitter. Feb 11. Available at: nature.com/articles/d41586-021-00338-y.(Accessed Oct 17, 2021).
Takreem (2018). *Nadia Sakati - KINGDOM of SAUDI ARABIA.* WEBNEOO. Available at: http://takreem.org/profile-details-265 (Accessed july 28, 2021).
Temtamy, S. A., and McKusick, V. A. (1978). *The Genetics of Hand Malformations*. New York: Liss.
Temtamy, S. A. (1998). Prevention of Genetic Diseases and Malformations in Newborns. *Sci. J. Minist. Health Popul.* 2, 22–27.
Temtamy, S. A., Salam, M. A., Aboul-Ezz, E. H. A., Hussein, H. A., H Helmy, S. A., and Shalash, B. A. (1996). New Autosomal Recessive Multiple Congenital Abnormalities/mental Retardation Syndrome with Craniofacial Dysmorphism Absent Corpus Callosum, Iris Colobomas and Connective Tissue Dysplasia. *Clin. Dysmorphol.* 5 (3), 231–240. doi:10.1097/00019605-199607000-00007
Temtamy, S. A. (2019). The Development of Human Genetics at the National Research Centre, Cairo, Egypt: A Story of 50 Years. *Annu. Rev. Genom. Hum. Genet.* 20, 1–19. doi:10.1146/annurev-genom-083118-015201
Terns, M. P., and Terns, R. M. (2011). CRISPR-based Adaptive Immune Systems. *Curr. Opin. Microbiol.* 14 (321), 321–327. doi:10.1016/j.mib.2011.03.005
The Branco Weiss Fellowship (2021). Taking Research beyond the Mainstream. Available at: https://brancoweissfellowship.org/fellow/wan.html (Accessed July 19, 2021).
The Network of African Science Academies (2017). *Women in Science - Inspiring Stories from Africa*. Nairobi: The Network of African Science Academies.
The Royal society (2021). *V. Narry Kim.* London, UK: The Royal society. Available at: https://royalsociety.org/people/v-narry-kim-narrykim-9776/(Accessed July 15, 2021).
The women and memory forum (2018). The Women and Memory Forum: Samia Temtamy. The Women and Memory Forum. Available at: http://whoisshe.wmf.org.eg/expert-profile/samia-temtamy (Accessed July 29, 2021).
UNESCO Institue for Statistics (2020). Women in Science, Fact Sheet no.60." June. Available at: http://uis.unesco.org/sites/default/files/documents/fs60-women-in-science-2020-en.pdf (Accessed July 19, 2021).
UNESCO Report Science (2021). *UNESCO SCIENCE REPORT: The Race against Time for Smarter Development*. Paris: United Nations Educational,Scientific and Cultural Organization.
UTAR (2021). *Producing Blood from Stem Cell for Blood Transfusion*. Perak: Universiti Tunku Abdul Rahman. Available at: https://news.utar.edu.my/news/2021/April/27/02/02.html (Accessed July 15, 2021).
van Hoof, A., and Wagner, E. J. (2011). A Brief Survey of mRNA Surveillance. *Trends Biochem. Sci.* 36 (11), 585–592. doi:10.1016/j.tibs.2011.07.005
Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011). Understanding the Transcriptome through RNA Structure. *Nat. Rev. Genet.* 12, 641–655. doi:10.1038/nrg3049
Wan, Y. (2021). National university of Singapore. Available at: https://medicine.nus.edu.sg/bch/faculty/yue-wan/(Accessed July 18, 2021).
Watson, C. (2021). Women Less Likely to Win Major Research Awards. Available at: https://www.nature.com/articles/d41586-021-02497-4 (Accessed Sep 27, 2021).doi:10.1038/d41586-021-02497-4
WHO (2020). Vaccine, WHO Recommendation AstraZeneca/EU Approved Sites COVID-19. Available at: https://extranet.who.int/pqweb/vaccines/covid-19-vaccine-chadox1-s-recombinant-0 (Accessed July 18, 2021).
Woodhouse, N. J., and Sakati, N. A. (1983). A Syndrome of Hypogonadism, Alopecia, Diabetes Mellitus, Mental Retardation, Deafness, and ECG Abnormalities. *J. Med. Genet.* 20 (3), 216–219. doi:10.1136/jmg.20.3.216
World Economic Forum (2020). Global Gender Gap Index Report 2020. Available at: https://reports.weforum.org/global-gender-gap-report-2020/the-global-gender-gap-index-2020/performance-by-region-and-country/(Accessed Oct 24, 2021).
Zheng, Y., Walsh, T., Gulsuner, S., Casadei, S., Lee, M. K., Ogundiran, T. O., et al. (2018). Inherited Breast Cancer in Nigerian Women. *J. Clin. Oncol.* 36 (28), 2820–2825. doi:10.1200/JCO.2018.78.3977

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fgene.2021.761988

Resprouters Versus Reseeders: Are Wild Rooibos Ecotypes Genetically Distinct?

J. Brooks[1], *N. P. Makunga*[1]*, *K. L. Hull*[2], *M. Brink-Hull*[2], *R. Malgas*[3] *and R. Roodt-Wilding*[2]*

[1]Department of Botany and Zoology, Stellenbosch University, Matieland, South Africa, [2]Department of Genetics, Stellenbosch University, Matieland, South Africa, [3]Department of Conservation Ecology and Entomology, Stellenbosch University, Matieland, South Africa

Aspalathus linearis (Burm. F.) R. Dahlgren (Fabaceae) or rooibos, is a strict endemic species, limited to areas of the Cederberg (Western Cape) and the southern Bokkeveld plateau (Northern Cape) in the greater Cape Floristic Region (CFR) of South Africa. Wild rooibos, unlike the cultivated type, is variable in morphology, biochemistry, ecology and genetics, and these ecotypes are broadly distinguished into two main groups, namely, reseeders and resprouters, based on their fire-survival strategy. No previous assessment of genetic diversity or population structure using microsatellite markers has been conducted in *A. linearis*. This study aimed to test the hypothesis that wild rooibos ecotypes are distinct in genetic variability and that the ecotypes found in the Northern Cape are differentiated from those in the Cederberg that may be linked to a fire-survival strategy as well as distinct morphological and phytochemical differences. A phylogeographical and population genetic analyses of both chloroplast (*trn*LF intergenic region) and newly developed species-specific nuclear markers (microsatellites) was performed on six geographically representative wild rooibos populations. From the diversity indices, it was evident that the wild rooibos populations have low-to-moderate genetic diversity (He: 0.618–0.723; Ho: 0.528–0.704). The Jamaka population (Cederberg, Western Cape) had the lowest haplotype diversity ($H = 0.286$), and the lowest nucleotide diversity ($\pi = 0.006$) even though the data revealed large variations in haplotype diversity ($h = 0.286$–0.900) and nucleotide diversity ($\pi = 0.006$–0.025) between populations and amongst regions where wild rooibos populations are found. Our data suggests that populations of rooibos become less diverse from the Melkkraal population (Suid Bokkeveld, Northern Cape) down towards the Cederberg (Western Cape) populations, possibly indicative of clinal variation. The largest genetic differentiation was between Heuningvlei (Cederberg, Western Cape) and Jamaka ($F_{ST} = 0.101$) localities within the Cederberg mountainous region, and, Blomfontein (Northern Cape) and Jamaka (Cederberg) ($F_{ST} = 0.101$). There was also a significant isolation by distance ($R^2 = 0.296$, $p = 0.044$). The presence of three main clusters is also clearly reflected in the discriminant analysis of principal components (DAPC) based on the microsatellite marker analyses. The correct and appropriate management of wild genetic resources of the species is urgently needed, considering that the wild Cederberg populations are genetically distinct from the wild Northern Cape plants and are delineated in

OPEN ACCESS

Edited by:
Zodwa Dlamini, SAMRC Precision Oncology Research Unit (PORU), South Africa

Reviewed by:
Susana Caballero, University of Los Andes, Colombia
Mateusz Labudda, Warsaw University of Life Sciences-SGGW, Poland

***Correspondence:**
R. Roodt-Wilding
rouvayroodtwilding@gmail.com
N. P. Makunga
makunga@sun.ac.za

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *20 August 2021*
Accepted: *27 October 2021*
Published: *20 December 2021*

Citation:
Brooks J, Makunga NP, Hull KL, Brink-Hull M, Malgas R and Roodt-Wilding R (2021) Resprouters Versus Reseeders: Are Wild Rooibos Ecotypes Genetically Distinct? Front. Genet. 12:761988. doi: 10.3389/fgene.2021.761988

accordance with ecological functional traits of reseeding or resprouting, respectively. The haplotype divergence of the ecotypes has also provided insights into the genetic history of these populations and highlighted the need for the establishment of appropriate conservation strategies for the protection of wild ecotypes.

Keywords: genetic diversity, medicinal plants, microsatellites, phylogeography, population genetic structure, rooibos, wild populations

INTRODUCTION

Aspalathus linearis (Burm. F.) R. Dahlgren (Fabaceae), is a commercially important South African legume, and a strict endemic of the Cape Floristic Region (CFR). It is more popularly known for its production of rooibos tea, an herbal beverage traditionally harvested in the wild, and now commercially produced for the global export market (Hawkins et al., 2011; Joubert and de Beer, 2011; Van Wyk and Gorelik, 2017). It occurs naturally in the Cederberg region of the Western Cape and in a few areas of the south-western parts of the Northern Cape (e.g., the Suid Bokkeveld and the Noord Bokkeveld Plateau near the rural town of Nieuwoudtville). There are populations within the distribution range that consist of different wild ecological types (ecotypes) (Van der Bank et al., 1999). Variable colour morphs are displayed by wild rooibos populations as the needle-like leaves between populations may range from a light grey-green to a bright green. These ecotypes vary in size/height of the plant, branching structure, leaf size, leaf colour, and leaf and stem thickness (Malgas et al., 2011). *Aspalathus linearis* is particularly important for its role in nitrogen-fixing in N- and P-limited fynbos environments. The fynbos region is a unique biome featuring over 7,000 species that are found in the Western and some parts of the Eastern provinces of South Africa. There are three main plant families that show a high level of species radiation and richness within the fynbos region namely, Restionaceae, Proteaceae and Ericaceae and nutrient poor soils of that are found in the fynbos biome are thought to have led to the high diversity of the Fabaceae plants in this region (Rebelo et al., 2006). It is also a pioneer species in a fire-prone vegetation type, relying on either resprouting from the underground lignotuber of burnt parent plants (resprouters), or as fire-triggered germination of new individuals (reseeders). Congeneric fire-survival strategies are common in several Fynbos taxa, e.g., Proteaceae; Ericaceae and Fabaceae (Marais et al., 2014; Pausas and Keeley, 2014). Rooibos exhibits a plethora of health benefits and is widely used for commercial products such as tea, food products and cosmetics as it has powerful antioxidant properties due to the abundance of flavonoids and other phenolic compounds found throughout the plant (Van Heerden et al., 2003; Smith and Swart, 2018; Bond and Derbyshire, 2020). The commercial importance of rooibos and the value rooibos provides to the livelihood to the local farmers, thus provides further impetus in understanding phylogeographic patterns that are linked to both its metabolites and population genetic structure and evolutionary history (Feliner, 2014). Combined phylogeographic and population genetic level research may also provide useful information for conservation studies by highlighting spatial conservation priorities, and broadening the scope of genetic diversity amongst wild ecotypes, protecting species diversity, similar to studies of rare and endangered species (Pollock et al., 2015; Médail and Baumel, 2018).

Over the past decade, some population genetic studies have been conducted within the Fabaceae family of plants, focusing on *Astralagus bibullatus* (Barneby and E. L. Bridges), *Anthonotha macrophylla* (P. Bauv), and commercial *Cyclopia* species (Baskauf and Burke, 2009; Demenou and Hardy, 2017; Potts, 2017; Niemandt et al., 2018; Galuszynski and Potts, 2020). Van der Bank et al. (1995) studied the genetic variation of wild *A. linearis* and the relationship of four geographically isolated populations, to determine levels of genetic variation and genetic differentiation using isozyme analysis. The study of Van der Bank et al. (1999), also based on isozyme analyses, concluded that resprouters likely evolved from reseeder plants and that this life history strategy was set at the population level. The ecological research of Hawkins et al. (2011) which included more extensive population surveys in the Cederberg showed no overlap between reseeding and resprouting populations in this particular region. However, the influence of the two fire survival strategies of reseeding or resprouting on the genetic diversity of rooibos still remains largely unknown.

Apart from its value in the commercialisation of rooibos, research has also contributed significantly to the understanding of rooibos ecotype diversity, genetic variation and the evolutionary history of wild rooibos (Joubert and de Beer, 2011). Edwards et al. (2008) investigated the barcoding potential of three DNA regions for the genus *Aspalathus*. These included nuclear ribosomal *ITS*, plastid *psb*A—*trn*H and *trn*T—*trn*L intergenic regions. Overall, the *trn*T*trn*L region was the most discriminatory between the *Aspalathus* species. Very few studies have investigated the complexity and variability between wild rooibos populations on a molecular level and this may be due to morphology, fire-survival strategy, reproductive strategy and biochemical variability (Dahlgren, 1968). The comparative study by Malgas et al. (2010) assessed haplotype variation and morphological variation among wild rooibos populations, using chloroplast $trnL^{UAA}F$—$trnF^{GAA}$, $trnT^{GGU}$—$trnD^{GUC}F$, $trnS^{GCU}$—$trnG^{UCC}$, $trnT^{UGU}F$—$trnL^{UAA}R$ intergenic regions and a nuclear marker, *PIII-PIV*, and observed a correlation between morphology and haplotypic variation. It was speculated that a genetic basis for the observed differences in morphology was important in the inherent morphotypes that are known to occur in the wild, and that are popularly reflected in the local ecological knowledge amongst resource-users (Malgas et al., 2011). The authors postulated that genetic differences between resprouter

FIGURE 1 | Collection sites of wild-growing *Aspalathus linearis* in the Cederberg area of the Western Cape and Northern Cape of South Africa.

and reseeder types may play a significant role in the diversification of *A. linearis* as a whole.

There is currently no study that has taken a complementary phylogeographic and population genetics approach in evaluating genetic variability between wild rooibos populations. This study was successful in investigating the phylogeography and population genetics of six wild rooibos ecotypes using both chloroplast sequencing and microsatellite marker analyses. This combined study approach provides foundational genetics research that is novel and may be added to the body of knowledge on rooibos. There are few genetics studied previously performed on rooibos yet wild rooibos ecotypes are well characterised in terms of metabolomic profiles. These ecotypes demonstrate metabolite variability that are linked to geographic localities (Lötter and le Maitre, 2014; Stander et al., 2017; Brooks, 2021). Additionally, the combination of genetic analyses with metabolomics may provide novel insights into understanding the rooibos species. This diversity of ecotypes may be an opportunity for novel products within the rooibos industry; moreover, ecotype diversity may be considered an advantage in the face of climate change (Lötter and le Maître, 2014). By maintaining competitive diversity between ecotypes, wild species are at lower risk of needing protection as microclimates continue to change and populations decline. It is important to emphasise that genetic diversity, even in strict endemic species such as *A. linearis,* is important for long-term conservation planning and for ensuring future sustainability of wild populations. This is because inherent genetic diversity facilitates better adaptation to changing environments, allowing for better population fitness (Potts, 2017). Populations that are continuously declining, often result in reduction in genetic variation, and may lead to inbreeding and/or genetic drift which ultimately reduces the natural fitness and potential adaptability of plants (Baskauf and Burke, 2009).

For these reasons, phylogeographic and population level analyses were conducted in this study using both chloroplast DNA (*trn*LF) sequencing analysis and a panel of 11 nuclear microsatellite marker loci to investigate wild populations of *Aspalathus linearis.* This dual-marker approach was chosen as it would allow for a historical and a contemporary assessment of species diversity and genetic differentiation (Wang et al., 2019). This study aimed to test the hypothesis that wild rooibos ecotypes are variable and distinct in genetic variability at the intra- and inter-population levels, and to discriminate wild ecotypes from various geographical regions. This was achieved through the collection of wild accessions from Cederberg in the Western Cape and Nieuwoudtville in the Northern Cape (**Figure 1**) before investigating the genetic diversity within and between the collected ecotypes using a dual-marker approach.

METHODS AND MATERIALS

Plant Material

Collections of rooibos wild plants were gathered from four localities in Nieuwoudtville and in the Suid Bokkeveld in the Northern Cape with permission from the Heiveld Co-operative and land owners (**Table 1**; **Figure 1**). Field harvests were conducted in mid-February 2018. Field guides from local

TABLE 1 | Wild rooibos (*Aspalathus linearis*) sampling sites and details of collections from the Cederberg region of the Western Cape and the Suid Bokkeveld of the Northern Cape.

Sample site/region	GPS coordinates	Number of individuals	Elevation (m)	Voucher	Fire-survival strategy (resprouter/ reseeder)	Distance to nearest town
Heuningvlei, Cederberg	32°12′ S 19°05′ E	11	868	*A.lin*_H2018	Reseeder	67.3 km to Clanwilliam, Cederberg
Jamaka, Cederberg	32°21′ S 19°02′ E	15	405	*A.lin*_J2018	Reseeder	23.8 km to Clanwilliam, Cederberg
Blomfontein, Nieuwoudtville	31°73′ S 19°13′ E	15	740	*A.lin*_B2018	Resprouter	47.4 km to Nieuwoudtville, Northern Cape
Dobbelaarskop, Nieuwoudtville	31°47′ S 19°11′ E	15	718	*A.lin*_D2018	Resprouter	54 km to Nieuwoudtville, Northern Cape
Matarakopje, Nieuwoudtville	31°94′ S 19°11′ E	15	480	*A.lin*_Ma2018	Resprouter	63.5 km to Nieuwoudtville, Northern Cape
Melkkraal, Nieuwoudtville	31°37′ S 19°21′ E	15	780	*A.lin*_M2018	Resprouter	11.5 km to Nieuwoudtville, Northern Cape

communities assisted with the identification of ecotypes of these plants and these were verified by a botanist (Nokwanda Pearl Makunga). Accessions from the Western Cape were also collected in the Cederberg mountainous region with a flora collection permit issued by CapeNature (Permit number: CN35-28-268) at two locations (**Table 1**). Branches with leaves near the top of the plant were collected and used for genetic analysis. The individuals that were collected per population ranged from 11 to 15 and were never mixed with other individuals. All samples were placed in individually labelled plastic Ziploc® bags with silica gel granules. These samples were stored in the dark at room temperature until further analysis. These populations can also be distinguished by their fire survival strategies, namely resprouters and reseeders. The Cederberg (Western Cape) populations are typically of the reseeder type, while the Northern Cape populations are commonly of the resprouter type (Malgas et al., 2010). In total, the collected populations cover a distance of 100 km. Representative voucher specimens were deposited in the herbarium of the Department of Botany and Zoology, at Stellenbosch University after confirmation of their taxonomic identity (**Table 1**).

DNA Extraction

Total genomic DNA was extracted from collected leaf material preserved in silica gel, according to a CTAB protocol described by Borse et al. (2011) with specific modifications in order to optimise genomic DNA (gDNA) quality from wild rooibos. This was important as the phenolics of rooibos could potentially influence downstream applications; the modifications are described below. The extraction buffer [10 ml; 2% (w/v) Cetyl Trimethyl Ammonium Bromide (CTAB); 100 mM Tris; 20 mM Ethylenediaminetetraacetic Acid (EDTA); 1.4 M NaCl, with added 2% (w/v) Polyvinylpyrrolidone (PVP)] was preheated in a water bath at 65°C (Sigma Aldrich®). Plant tissue (100 mg) was ground into a fine powder in liquid nitrogen using a mortar and pestle and 1 ml pre-warmed extraction buffer was immediately added to the ground plant material (2 ml Eppendorf tube per reaction). After 20 min, 2 µL of β-mercaptoethanol was added and further incubated at 65°C for 1 h and 30 min. The mixture was then thoroughly vortexed and placed at 65°C for another 20 min heating period before the tubes were cooled down to room temperature and centrifuged at 11,000 rpm (Microlitre centrifuge, Mikro 120, Hettich Zentrifugen) for 10 min. The supernatant was transferred into a new tube, whereas the pellet with the cellular debris was discarded. Chloroform isoamyl alcohol (24:1; v/v) was added to the supernatant. The tubes were again centrifuged at 11,000 rpm for 10 min at 4°C and thereafter, the upper aqueous phase was transferred to new tubes and twice the volume of molecular grade absolute ethanol (BioUltra, Sigma -Aldrich®) was added to precipitate the DNA at −20°C overnight. To collect the DNA, all samples were centrifuged at 11,000 rpm for 10 min at 4°C before the supernatant was discarded and left to dry at room temperature. Once dried, the pellet was dissolved in 60 µL of Tris-EDTA (TE) buffer (10 mM, 1 mM, pH 8.0). The extracted DNA was quantified using a Nanodrop spectrophotometer (NanoDrop™ Lite, Thermo Fischer Scientific). Samples were diluted to a concentration of 50 ng/µL using sterile dH_20 and stored at −20°C until further use. The DNA was visualised on a 1% (m/v) agarose gel (6 µL EtBr) to confirm the presence of high-quality DNA.

Sequencing of Chloroplast *trn*L^{UAA}F—*trn*F^{GAA} Gene Region

Extracted genomic DNA was subjected to polymerase chain reaction (PCR) amplification (BioRad T100—Applied Biosystems™) using the *trn*L^{UAA}F and *trn*F^{GAA} primers: 5′-CGAAATCGGTAGACGCTACG-3′ and 5′-ATTTGAACT GGTGACACGAG-3′ (Integrated DNA Technologies, United States), respectively, designed based on the work of Taberlet et al. (1991). The PCRs were performed in volumes of 25 µL containing 1 µL of template DNA, 12.5 µL of Qiagen Multiplex PCR Master Mix (Whitehead Scientific, South Africa), 1.5 µL of each primer (10 µM) and 8.5 µL of sterile Milli-Q H_20 (Ultrapure water purification system Barnstead™ MicroPure™, Thermo Fischer Scientific). The PCR reaction was performed with an initial 2-min denaturation step at 95°C. This was followed by 35 cycles, consisting of denaturation at 95°C for 1 min, annealing (adjusted from Malgas et al., 2010) at 48.6 °C for 30 s, and

TABLE 2 | Multiplex assay for 13 *Aspalathus linearis* species-specific nuclear microsatellite markers, where the repeat motif, dye, size range (bp), and annealing temperature (T_A) are indicated.

Multiplex	Marker name	Repeat motif	Dye	Dye colour	Size range (bp)	T_A (°C)
1	ROI82	(AG)30	VIC	Green	180 (160–200)	58
	ROI65	(CT)17	NED	Yellow	225 (200–250)	
	ROI66	(CT)18	6-FAM	Blue	155 (130–180)	
2	ROI72B	(CT)27	TET	Green	183 (160–210)	60
	ROI70	(GA)5	PET	Red	127 (110–150)	
	ROI70B	(AG)27	NED	Yellow	200 (180–220)	
	ROI71B	(AG)36	6-FAM	Blue	250 (220–270)	
3	ROI64	(GT)14	PET	Red	150 (130–170)	58
	ROI69	(CT)5	VIC	Green	108 (90–130)	
	ROI73	(AG)12	VIC	Green	211 (190–230)	
4	ROI67	(AG)31	TET	Green	158 (140–190)	60
	ROI83	(AG)10	PET	Red	120 (100–140)	
	ROI85	(AG)8	6-FAM	Turquoise	228 (210–260)	

extension at 72°C for 40 s. The reaction was concluded with a final extension step at 72°C for 5 min (Malgas et al., 2010). Following PCR amplification, amplicons were visualised by means of agarose gel (1% w/v) electrophoresis at 110 V.

The PCR products were purified using a Sephadex A® G-50 column (Sigma Aldrich®), according to the manufacturer's specifications, and, bidirectional sequencing reactions were performed using a BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific), without any changes to the manufacturer's instructions. Reactions were performed in volumes of 10 µL. Cycling conditions included an initial denaturation period of 1 min at 96°C, followed by 25 cycles of 10 s at 96°C, 5 s at 50°C, and 4 min at 60°C, as per the manufacturer's instructions. Following this step, the sequencing reactions were sent for visualisation at the Central Analytical Facility (CAF) (DNA Sequencing Unit) at Stellenbosch University. The sequencing files were then manually trimmed on either end to a final length of 501 bp and edited using BioEdit v7.2.6.1 (Hall, 1999). The sequences of each population were then aligned in MEGA v7.0 (Kumar et al., 2016) using the ClustalW alignment algorithm with default parameters.

Genetic Data Analyses Based on Chloroplast Gene Sequences

Diversity indices were calculated in DNASP v5.0 (Librado and Rozas, 2009) for all of the wild rooibos populations. These included the total number of haplotypes (H), haplotype diversity (h), nucleotide diversity (π), and the average number of nucleotide substitutions (k). A hierarchical analysis of molecular variance (AMOVA) was performed using ARLEQUIN v3.5.2 (Excoffier and Lischer, 2010) ($p < 0.05$) to investigate potential population differentiation based on the chloroplast sequences. The AMOVA tested the hypothesis of panmixia, whereby there are no restrictions between populations (global population) (F_{ST}). The AMOVA also tested for genetic differentiation between the Cederberg (Western Cape) and Northern Cape regions (among regions, F_{ST}), among populations within the two regions (F_{SC}) as well as the genetic differentiation within populations (F_{CT}). A Median-Joining haplotype network (Bandelt et al., 1999). was constructed using NETWORK v5.0.1.1 (http://www.fluxus-engineering.com), to investigate the evolutionary relationships among haplotypes.

Nuclear SSR Amplification and Genotyping

Intact genomic DNA samples were sent to Genetic Marker Services (GMS) (United Kingdom) for the development of 18 dinucleotide microsatellite markers (Short Sequence Repeats—SSRs) specific to *Aspalathus linearis*. Sequence information was obtained from GMS for 13 developed markers (**Table 2**). The remaining 5 markers were not polymorphic and where therefore excluded from further analyses. The forward primers for polymorphic markers were fluorescently labelled (PET, NED, 6-FAM, TET, and VIC) by ThermoFisher (**Table 2**). The markers were then optimised into four multiplex groups (3–4 markers per multiplex) and amplified across a total of 86 individuals.

Polymerase chain reaction amplifications were performed in order to test for successful amplification of the markers and to optimise the PCR conditions. Each reaction consisted of a total volume of 25 µL using 1 µL of 50 ng template gDNA and 1x Qiagen Multiplex PCR Master Mix (Whitehead Scientific, South Africa). The PCR reaction was run at 95°C for 2 min as the initial denaturation step followed by 35 cycles of denaturation at 95°C for 30 s, annealing at the appropriate annealing temperature for each multiplex (**Table 2**) for 30 s, and an extension step at 72°C for 40 s. The reaction was completed with a final extension step at 72°C for 5 min. The amplicons were diluted 10 x with ddH_2O and sequenced at CAF (Stellenbosch University) using the 500 LIZ® size standard. Electropherograms were analysed using the software program GeneMapper v5.0 (Applied Biosystems) for the detection of peaks, bin calling and genotyping.

Genetic Diversity Using Nuclear Microsatellite Markers

Microsatellite genotypes were evaluated for allele stuttering, allelic dropout and the presence of null alleles while the frequency of null alleles per locus per population was calculated using MICROCHECKER v2.2.3 (Van Oosterhout et al., 2004). The software, GENEPOP ON THE WEB v4.2

TABLE 3 | Polymorphic nucleotide positions of the 4 haplotypes determined across 6 wild rooibos populations.

	Nucleotide positions				
	210	**218**	**278**	**355**	**463**
Haplotypes	G	G	A	A	C
H1	.	.	.	.	.
H2	.	.	.	T	.
H3	.	.	.	.	T
H4	A	T	T	.	.

TABLE 4 | Genetic diversity indices of collected wild rooibos populations of the chloroplast gene region, *trn*L^{UAA}F—*trn*F^{GAA}. *n*—sample size; H—total number of haplotypes; *h*—haplotype diversity; π—nucleotide diversity; *k*—average number of nucleotide substitutions.

Sampling region	***n***	**H**	***h***	**π**	***k***
Cederberg, Western Cape	**13**	**2**	**0.538**	**0.001320**	**0.538**
Heuningvlei	6	1	0.800	0.002024	1
Jamaka	7	2	0.286	0.000685	0.285
Suid Bokkeveld, Northern Cape	**26**	**3**	**0.748**	**0.004700**	**1.717**
Blomfontein	6	2	0.733	0.002554	1.266
Dobbelaarskop	8	1	0.571	0.001516	0.571
Matarakopje	5	2	0.900	0.010569	5.2
Melkkraal	7	1	0.571	0.001176	0.571
Overall	**39**	**4**	**0.428**	**0.001930**	**0.566**

Boldtype face indicates the two regions, namely the Cederberg region of the Western Cape and the Suid Bokkeveld, Northern Cape region as well as the diversity indices across all of the populations.

(Rousset, 2008) was used to test for loci deviating from Hardy-Weinberg Equilibrium (HWE) expectations (10,000 dememorisations, 100 batches, and 10,000 iterations per batch) and for between-loci linkage disequilibrium (LD) within and across sampling populations. The inbreeding coefficient (F_{IS}) for each sampling population and region (two populations from Cederberg, Western Cape, and four populations from Suid Bokkeveld, Northern Cape) were also estimated in GENEPOP. Markers under selection were then tested for in ARLEQUIN v3.5.2 (Excoffier and Lischer, 2010) ($p < 0.05$).

Genetic diversity indices were calculated for two datasets: 1) sampling populations treated separately (global dataset) and 2) sampling populations grouped into broad geographic regions (regional dataset—Cederberg vs Suid Bokkeveld). This included the average number of alleles per locus (An), the effective number of alleles per locus (Ae), allelic richness, scaled to each population size of 15 individuals (A_R), observed and expected heterozygosity (Ho and He), Shannon's index (I), and fixation index (F) calculated in GENALEX v6.501 (Peakall and Smouse, 2006). Polymorphism information content (PIC) of each marker was determined in MSATTOOLS v3.1.1 (Park, 2001).

Genetic Differentiation Using Nuclear Microsatellite Markers

Principal Coordinate Analysis (PCoA) was performed in GENALEX v6.501 (Peakall and Smouse, 2006) to determine the clustering patterns across all populations. Pairwise F_{ST} estimates were calculated (999 permutations, $p < 0.05$) in order to determine the degree of genetic differentiation. A hierarchical AMOVA was performed in ARLEQUIN v3.5.2 ($p < 0.05$). The AMOVA was used to interrogate panmixia and degree of genetic differentiation as previously described (refer to *Genetic Data Analyses Based on Chloroplast Gene Sequences*). Multivariate discriminant analysis of principal components (DAPC) was performed in R Studio (R v3.5.3) using the *K*-means clustering method in the *adegenet* package to determine the genetic structure. This was achieved by the estimation of the alpha score, determining the optimal number of principal components to retain. The clustering method was run at $k = 20$. The Bayesian Information Criterion (BIC, Schwarz, 1978) was used to determine the optimal *K* value. A Bayesian clustering analysis was implemented in Structure v2.3.4 (Pritchard et al., 2000), assuming an admixture ancestry model with correlated allelic frequencies. Ten replicates were run for each *K* tested ($k = 1$–6), using a burn-in of 50,000 followed by 500,000 steps where data points were retained. The optimal *K* values were determined based on Delta *K* (Evanno et al., 2005) and the four tests of Puechmaille (2016), namely *MedMedK*, *MedMeaK*, *MaxMedK*, and *MaxMeaK*, which were determined by StructureSelector (Li and Liu 2018). Assignment plots were generated and visualised using the web service Clumpak (Kopelman et al., 2015). An assessment of relatedness (*r*) within the wild rooibos populations was performed in GENALEX v6.501, using the Queller and Goodnight (1989) estimator of relatedness (RQG). Lastly, isolation by distance (IBD) was tested using a Mantel test in GENALEX v6.501 to determine the relationship between genetic distance and geographical distance between sampling populations.

Landscape Genetics Data Analyses Using Nuclear Microsatellite Markers

To investigate genetic stratification of wild rooibos populations as a result of landscape features, the R package *Geneland* was used (Guillot et al., 2011). First, geographic positioning system (GPS) coordinates for each of the six sampling populations from Cederberg region (Heuningvlei and Jamaka) and Northern Cape region (Blomfontein, Dobbelaarskop, Matarakopje, Melkkraal) were converted to Universal Transverse Mercator (UTM) coordinates for each sample per sampling location using the R packages *mapproj* and *PBSmapping*.

Genotypic information and UTM coordinates were used as input files for the Geneland pipeline. The population cluster range specified was $k = 1$–6, which tested the hypotheses of complete panmixia ($k = 1$) to complete isolation ($k = 6$). Allele frequencies were assumed to be correlated between populations, and a spatial model was stipulated to explain spatial patterns due to gene flow between locations. This was performed for 1,00,000 MCMC runs across 10 independent iterations.

RESULTS

Genetic Diversity: Chloroplast *trn*LF Region

A total of 39 individuals from six ecologically distinct wild populations were successfully sequenced. Analysis of DNA

TABLE 5 | Haplotype distribution of collected wild rooibos populations. *n*—number of individuals per population; H1—haplotype 1; H2—haplotype 2; H3—haplotype 3; H4—haplotype 4.

Sampling population	*n*	H1	H2	H3	H4
Blomfontein	6	3	3	0	0
Dobbelaarskop	8	8	0	0	0
Heuningvlei	6	6	0	0	0
Jamaka	7	1	0	6	0
Matarakopje	5	4	0	0	1
Melkkraal	7	7	0	0	0

polymorphisms revealed 5 polymorphic sites consisting of two transitions and three transversions, three of which were singletons and two were parsimony informative sites (**Table 3**). A total of four distinct haplotypes were observed across all populations (**Tables 4**, **5**). The overall haplotype diversity was 0.428 and the nucleotide diversity was 0.001930. The Jamaka population (Cederberg, Western Cape) had the lowest haplotype diversity (0.286), and the lowest nucleotide diversity (0.0006) even though haplotype diversity generally varied considerably across the wild populations (h = 0.286–0.900). Nucleotide diversity ranged from 0.0006 to 0.010 across all the populations but the Matarakopje plants displayed the highest nucleotide diversity (**Table 4**). Among the populations studied, Matarakopje (Suid Bokkeveld, Northern Cape) had the highest genetic diversity.

The haplotype network for the four distinct haplotypes observed in this study, appears to consist of an ancestral haplotype (H1) represented by 29 individuals across all populations whereas the remaining 3 haplotypes (H2, H3, H4) are private haplotypes representing the Blomfontein, Jamaka and Matarakopje populations, respectively (**Figure 2**). H2 is represented by three individuals, while H3 is represented by six individuals. It is important to note that H4 is only represented by one individual.

Genetic Diversity: Nuclear Microsatellite Markers

In total, 86 individuals were successfully genotyped for 13 species-specific markers, with the average number of alleles ranging from 1 to 10 per marker (**Supplementary Table S1**). The frequencies of null alleles reached a maximum of 0.366 for locus ROI72B (**Supplementary Table S1**). Across all wild rooibos populations, several loci deviated from HWE, namely ROI82, ROI72B, and ROI73. The loci that were in LD include ROI65 and ROI70, ROI72B and ROI70, and lastly, ROI66 and ROI70B. Locus ROI72B showed evidence for null alleles, deviations from HWE, LD as well as being under selection ($p < 0.05$).

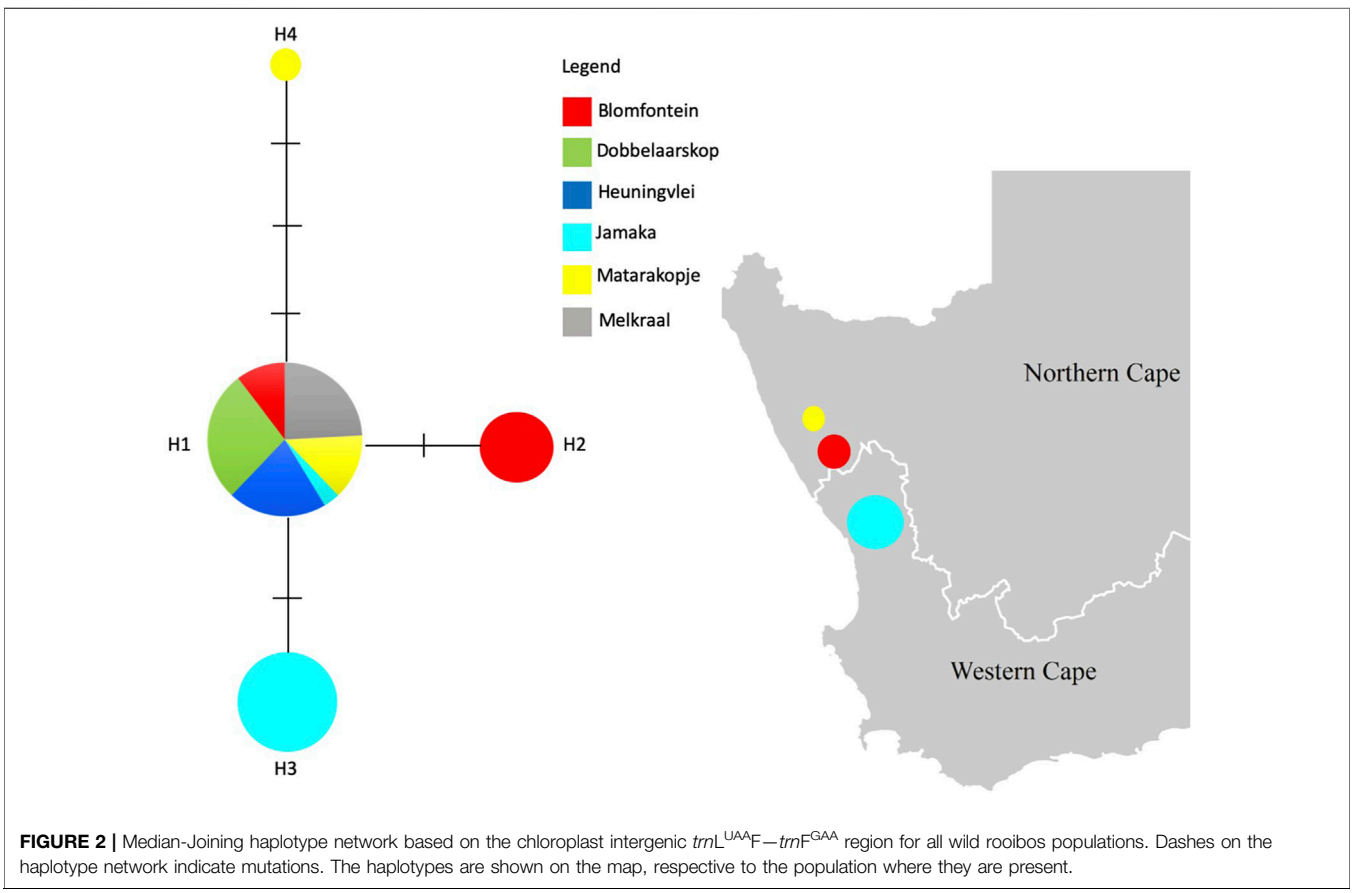

FIGURE 2 | Median-Joining haplotype network based on the chloroplast intergenic *trn*L^{UAA}F—*trn*F^{GAA} region for all wild rooibos populations. Dashes on the haplotype network indicate mutations. The haplotypes are shown on the map, respective to the population where they are present.

TABLE 6 | Genetic diversity indices for six wild *Aspalathus linearis* populations based on 11 microsatellite loci. *n*—sample size; PIC—Polymorphic Information Content; An—mean number of alleles per locus; Ae—mean number of effective alleles; A_R—allelic richness; I—Shannon's index; Ho—observed heterozygosity; He—expected heterozygosity; uHe—unbiased expected heterozygosity; F—fixation index; F_{IS}—inbreeding coefficient.

	n	PIC	An	Ae	A_R	I	Ho	He	uHe	F	F_{IS}
Blomfontein	15	0.652	6.364	3.959	4.580	1.365	**0.528**	**0.618**	0.646	0.121	0.147
Dobbelaarskop	15	0.682	7.455	4.533	5.250	1.605	0.700	0.714	0.747	–0.018	0.019
Heuningvlei	11	0.594	5.182	**3.411**	4.250	1.287	**0.704**	0.640	0.681	–0.114	–0.100
Jamaka	15	0.606	6.909	4.217	4.640	1.418	0.565	0.636	0.661	0.077	0.113
Matarakopje	15	0.689	7.273	**4.966**	5.170	1.602	0.631	**0.723**	0.759	0.091	0.128
Melkkraal	15	0.647	7.545	4.213	5.010	1.529	0.610	0.677	0.708	0.065	0.099
Average	14.333	**0.645**	**6.788**	**4.217**	4.816	**1.468**	0.623	0.668	0.700	0.037	**0.038**

Boldtype face indicates those values discussed in text.

TABLE 7 | Hierarchical analysis of molecular variance (AMOVA) based on *trn*L^{UAA}F—*trn*F^{GAA} sequences for different structuring hypotheses of wild *Aspalathus linearis* based on six wild populations in two different sampling regions (Cederberg in the Western Cape and Suid Bokkeveld in the Northern Cape).

Hypothesis tested	Source of variation	Variation (%)	Fixation index
Panmixia	Among populations	53.15	Φ_{ST} = 0.531[a]
	Within populations	46.85	
Inter-region	Among regions	16.89	Φ_{ST} = 0.568[a]
	Among populations within regions	39.93	Φ_{SC} = 0.480[a]
	Within populations	43.18	Φ_{CT} = 0.168

[a]*Indicates statistical significance at* $p < 0.05$.

For these reasons, this marker was excluded from further analysis. Locus ROI85 yielded little to no genotyping information, and was therefore also disregarded. Locus ROI73 presented with deviations from HWE across all populations and showed evidence of null alleles but was retained for further analysis as it did not display LD with any other markers and was also not found to be under selection. A total of 11 markers were therefore retained for further analyses. Overall, the mean number of observed alleles (An) was 6.788 while the mean number of effective alleles (Ae) was 4.217, ranging from 3.411 (Heuningvlei) to 4.966 (Matarakopje) (**Table 6**; **Supplementary Table S1**). Shannon's information index reported an average of 1.468 and F-statistics revealed moderate genetic differentiation (F_{ST} = 0.101, $p < 0.05$). The observed heterozygosity varied from 0.528 to 0.704 while the expected heterozygosity was recorded at values of 0.618–0.723. The inbreeding coefficient (F_{IS}) averaged at 0.038, indicating little to no inbreeding. Polymorphic Information Content (PIC) of the SSRs showed an average of 0.645 (**Table 6**). It should be noted that only 15 individuals were available per population.

Population Differentiation and Genetic Structure: Chloroplast *trn*LF Region

The hierarchical AMOVA (**Table 7**) highlighted genetic differentiation based on the *trn*L^{UAA}F—*trn*F^{GAA} chloroplast region. Significant differentiation was observed among populations when testing for panmixia (Φ_{ST} = 0.531, $p < 0.05$), as well as when assessing differentiation among the regions (Φ_{ST} = 0.568, $p < 0.05$) and among populations within regions (Φ_{SC} = 0.480, $p < 0.05$). However, there was no significant differentiation reported within populations (Φ_{CT} = 0.168, $p > 0.05$). The AMOVA results support a divergence between regions (Cederberg and Suid Bokkeveld) as well as among populations within regions. There was no significant divergence within populations, which correlates to the presence of haplotype 1 in the majority of the total individuals that were sampled.

Population Differentiation and Genetic Structure: Nuclear Microsatellite Markers

Pairwise F_{ST} estimates extended from 0.005 to 0.101 ($p < 0.05$). This indicates moderate genetic differentiation between the wild rooibos populations (**Table 8**; **Supplementary Table S2**). The largest genetic differentiation was between Heuningvlei and Jamaka (F_{ST} = 0.101), and Blomfontein and Jamaka (F_{ST} = 0.101). The hierarchical AMOVA supported this genetic differentiation across all but one level (**Table 8**). Significant differentiation was reported among the regions (F_{ST} = 0.064, $p < 0.05$) and among populations within regions (F_{SC} = 0.053, $p < 0.05$). However, there was no significant differentiation reported within populations (F_{CT} = 0.011, $p > 0.05$).

The principal coordinate analysis (PCoA) revealed no patterns of genetic structure (**Figure 3A**), although clustering was presented by the Discriminant Analysis of Principal Components (DAPC) (**Figure 3B**). The two-dimensional distribution pattern observed from the PCoA (**Figure 3A**) totalled 13.55% variance, accumulated on the first two components (5.87 and 5.44%, respectively). The DAPC analysis revealed separation of the resprouter populations (Northern Cape populations, occurring on the positive side of PC1) versus the reseeder populations (Cederberg populations, occurring on the negative side of PC1). This was supported by the

TABLE 8 | Hierarchical analysis of molecular variance (AMOVA) based on 11 microsatellite markers for different structuring hypotheses of wild *Aspalathus linearis* based on six wild populations in two different sampling regions (Cederberg in the Western Cape and Suid Bokkeveld in the Northern Cape).

Hypothesis tested	Source of variation	Variation (%)	Fixation index
Panmixia	Among populations	5.897	F_{ST} = 0.058[a]
	Within populations	94.102	
Inter-region	Among regions	1.190	F_{ST} = 0.064[a]
	Among populations within regions	5.272	F_{SC} = 0.053[a]
	Within populations	93.536	F_{CT} = 0.011

[a]*Indicates statistical significance at* p < *0.05.*

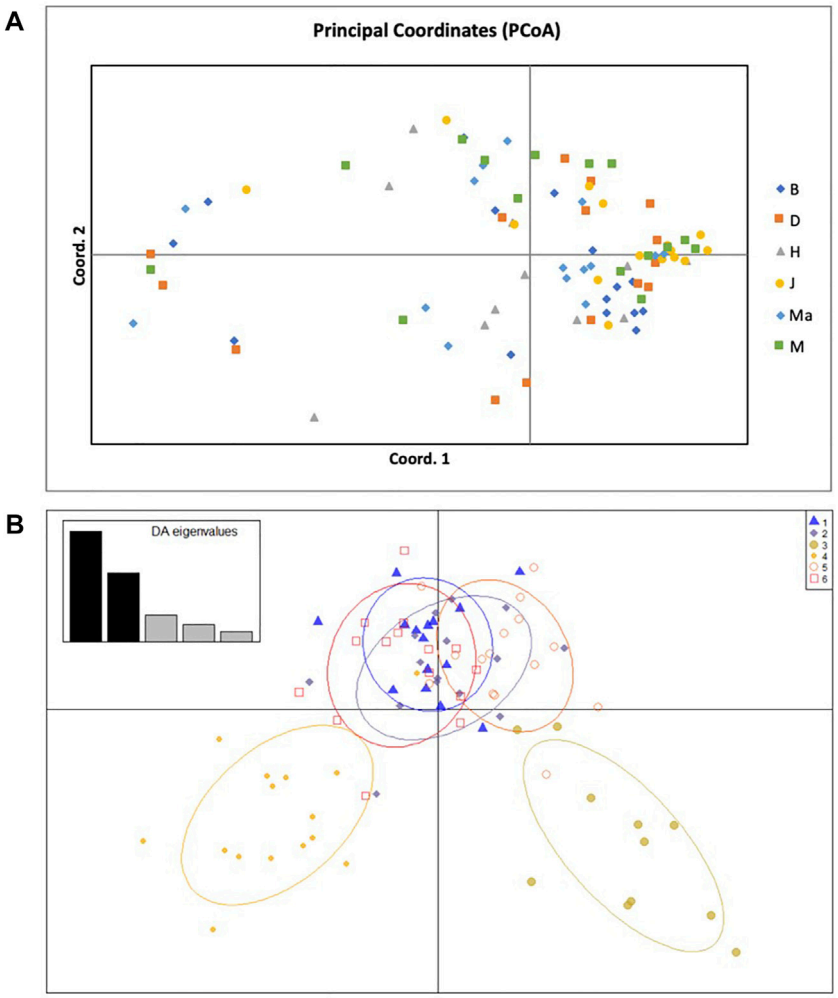

FIGURE 3 | (A) Principal Coordinate Analysis (PCoA) of 86 individuals across six wild rooibos populations based on 11 SSR loci. B—Blomfontein, D—Dobbelaarskop, H—Heuningvlei, J—Jamaka, Ma—Matarakopje, M—Melkkraal. **(B)** Discriminant Analysis of Principal Components (DAPC) analysis of 86 individuals across six wild rooibos populations based on 11 SSR loci. B—Blomfontein, D—Dobbelaarskop, H—Heuningvlei, J—Jamaka, Ma—Matarakopje, M—Melkkraal. Each grouping represents a genetic cluster.

Bayesian Structure results, which showed a similar distinction between Northern Cape and Cederberg populations at k = 3 (**Figure 4**). An assessment of relatedness within the wild rooibos populations showed the highest relatedness to be within the Heuningvlei and Jamaka populations (**Figure 5A**).

The Isolation by Distance Mantel test revealed that there is a significant correlation (R^2 = 0.296, p = 0.044) between genetic distance and geographical distance (**Figure 5B**). Additionally, based on the landscape analyses in the *Geneland* pipeline, a total of four clusters were identified (**Figure 6**), with Heuningvlei and

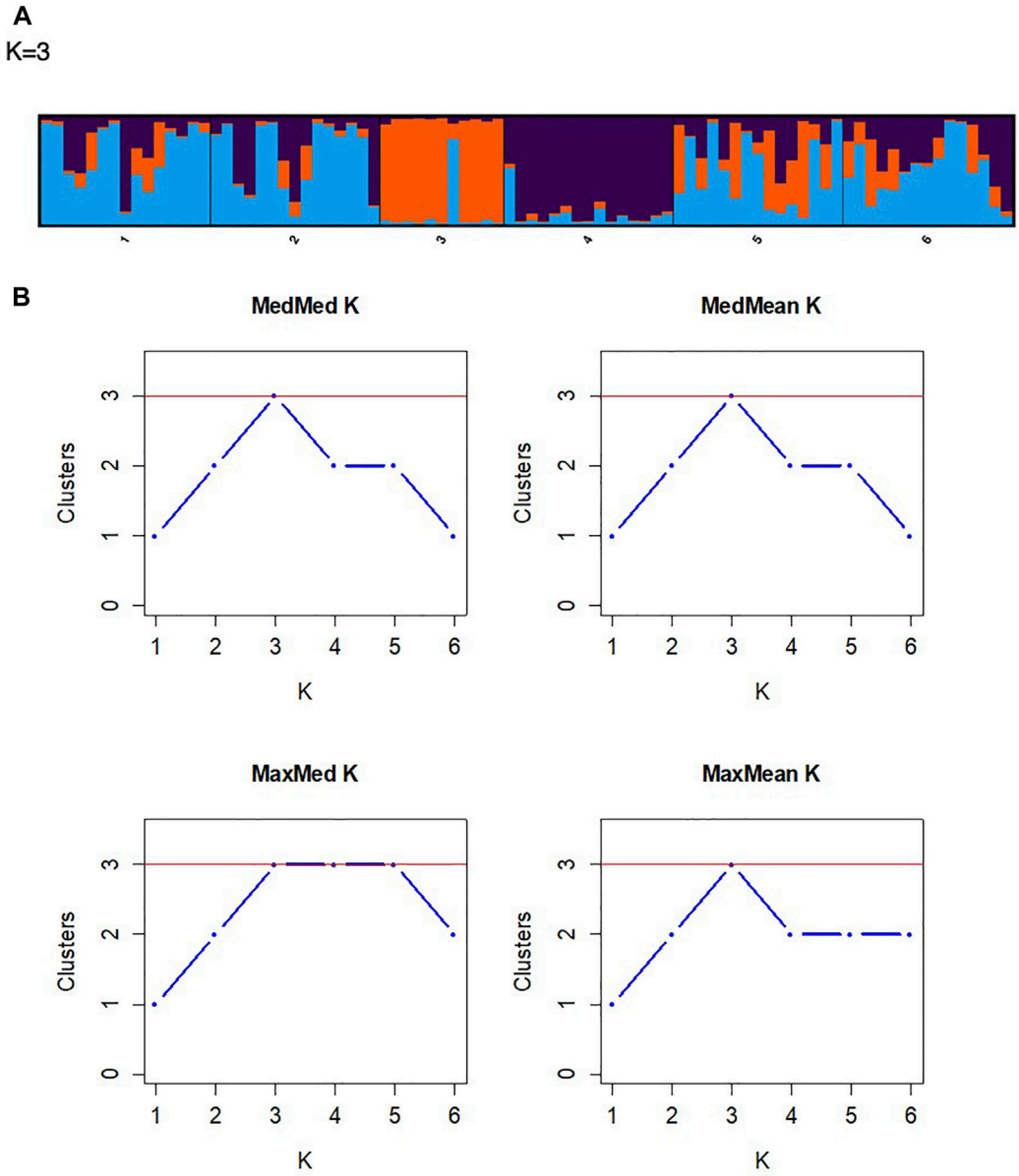

FIGURE 4 | (A) A STRUCTURE bar plot illustrating the distribution of the collected wild rooibos populations (k = 3). **(B)** Assignment plots of the optimal K values using the four tests of Puechmaille, namely *MedMedK*, *MedMeaK*, *MaxMedK*, and *MaxMeaK*.

Jamaka clustering independently from each other, and from the other sampling locations (**Table 9**). This is in congruence with the multivariate DAPC analysis, which showed a distinction between the reseeders from the Western Cape (Cederberg region) and the resprouters from the Northern Cape (Nieuwoudtville region).

Additionally, a map was constructed based on the genetic and UTM coordinate information, which unfortunately did not show clear distinctions between the previously defined clusters (**Figure 7**), but rather areas or zones where these clusters occur. Notably, all clusters share a degree of overlap, pointing to a level of gene flow between the groups. This was further supported by the low pairwise F_{ST} estimates between the clusters, with the Heuningvlei population in cluster four being the most distinct of the group, although none of the values were statistically significantly different between the clusters (**Table 10**).

DISCUSSION

Historical Influences on Genetic Diversity and Genetic Differentiation

Aspalathus linearis is an extremely complex and variable species (Van der Bank et al., 1999; Van Heerden et al., 2003) and this

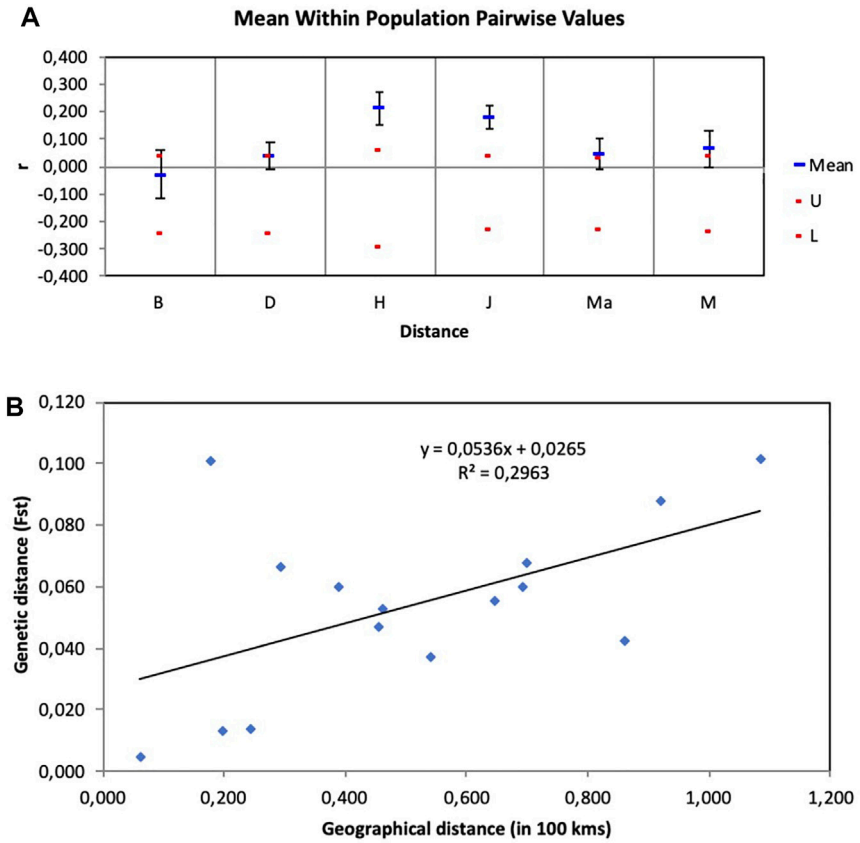

FIGURE 5 | (A) Assessment of relatedness within wild rooibos populations. B—Blomfontein, D—Dobbelaarskop, H—Heuningvlei, J—Jamaka, Ma—Matarakopje, M—Melkkraal. U—upper confidence limit, L—lower confidence limit. **(B)** Isolation by distance (IBD) graph using genetic distance measured in F_{ST} estimates and geographical distance measured in 100 km.

variability is evident in the morphology, phytochemistry, ecology, and genetics with respect to wild rooibos populations (Van der Bank et al., 1999; Hawkins et al., 2011; Stander et al., 2017). The development of the fire-survival mechanisms that occur exclusively on a population level allow them to maintain their unique ecology as resprouters or reseeders. Biogeographic spatial distribution of rooibos is also associated with different chemical signatures which have been hypothesised to be associated with population-based genetic variation. In this study, rooibos was confirmed to exhibit three unique haplotypes connected to an ancestral haplotype, representing all ecotypes (**Figure 2**) and these data also lends support to the previous study of Malgas et al. (2010).

Overall, the haplotype and nucleotide diversity observed across both regions was low (h = 0.428, π = 0.002). Malgas et al. (2010) conducted a molecular study on wild rooibos populations using the same *trn*L^{UAA}F—*trn*F^{GAA} region, but no haplotype or nucleotide diversity was presented for comparison. There is, notably, large variation in haplotype diversity and nucleotide diversity between populations and between regions (**Table 4**). This corresponds with the conclusions by Hawkins et al. (2011) that wild rooibos ecotypes are indeed ecologically distinct.

Rooibos may well be included in predictions of species range shifts due to climate change widely forecast for the Fynbos biome (Midgley et al., 2003; Lötter and le Maître, 2014). Recent droughts in the Western Cape of South Africa, together with soil nutrient depletion, has lowered yields from commercial rooibos production (Smith et al., 2018). Apart from this, climate change also influences wild harvesting by locals and drives patterns of wild rooibos collections especially when cultivated rooibos yields may be low and subsistence farmers become more reliant on wild populations to bulk up their yield of rooibos for trade with commercial producers (Lötter and le Maître, 2014). Across all biomes, drought and historical climate changes over time result in declining population sizes (Lötter and le Maître, 2014). Historically, during the Plio-Pleistocene (5–2.5 mya) glacial cycles, the climate became cooler, more arid and then humid and warmer between the glacial phases. This had a significant impact on the flora of Africa, particularly their ability to adapt to and survive harsh climates (Tolley et al., 2014). The chloroplast data revealed only five polymorphic sites, showing little differences between the populations, suggesting recent variability. Populations that are more genetically similar would have a more recent common ancestor as genetic variability accumulates over time (Schaal

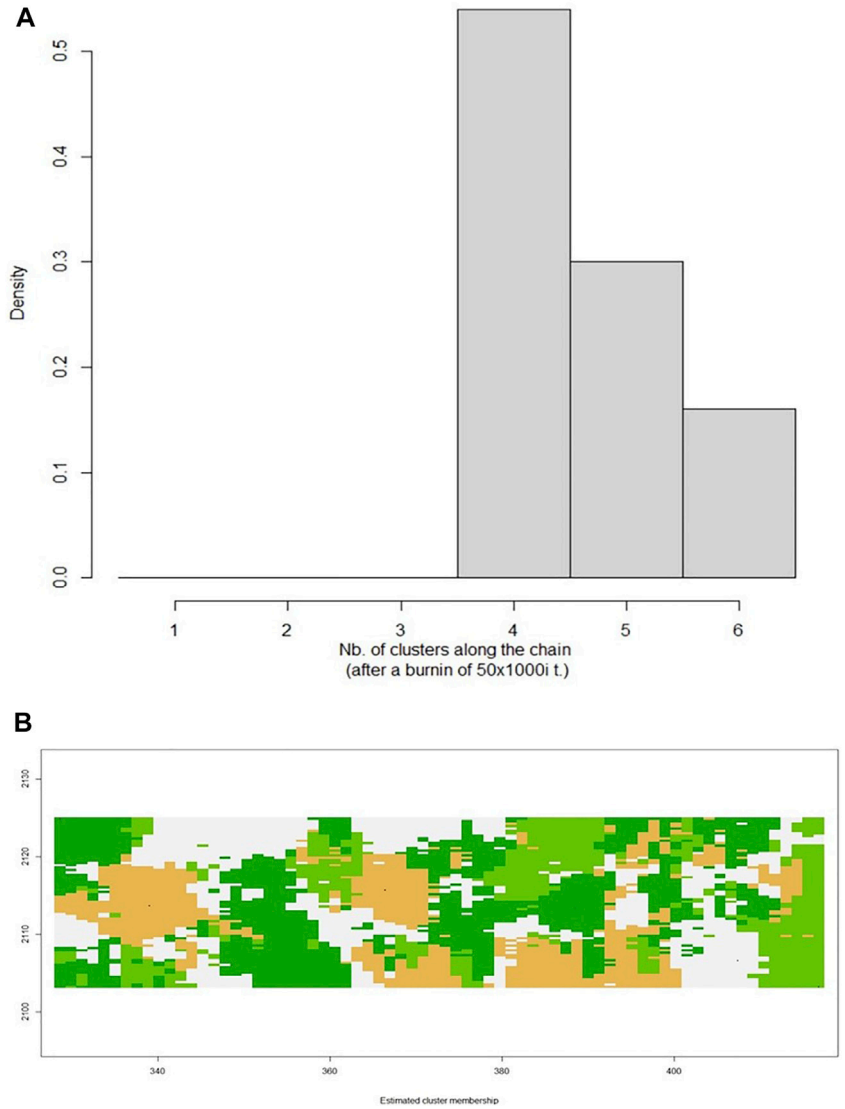

FIGURE 6 | (A) Graphical representation of the number of clusters defined in the data set with Geneland, with k = 4 being the optimal number of clusters. **(B)** Posterior mode of population membership. Different colours denote different clusters.

TABLE 9 | Assignment probabilities of individuals from the six geographical locations to the identified clusters.

Sampling locations	Posterior probability of population membership			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Heuningvlei	0.111	0.148	0.074	**0.667**
Jamaka	0.185	**0.630**	0.185	-
Blomfontein	0.037	0.111	**0.741**	0.111
Dobbelaarskop	0.037	0.111	**0.741**	0.111
Matarakopje	**0.667**	0.111	-	0.222
Melkkraal	**0.667**	0.111	-	0.222

Bolded values indicate the greatest assignment probability of the four options.

et al., 1998). It can be said that the most recent diversification event within the Cape Floristic Region, among populations within species, can be dated to the Pleistocene (Tolley et al., 2014). The diversification of wild rooibos, resulting in these four haplotypes could be hypothesised to have taken place during this time. There is however no evidence or discussion of this in previous studies.

The hierarchical AMOVA indicated significant genetic differentiation across all but one level. It revealed a high level of differentiation between regions (Cederberg and Northern Cape populations, Φ_{ST} = 0.568, $p < 0.05$) as well as among the populations within the two regions (Φ_{SC} = 0.480, $p < 0.05$). There was, however, no significant differentiation within populations (**Table 7**). This could be evidence of isolation between the two regions and between certain populations within the regions as many of these populations are separated by physical barriers such as the Cederberg Mountains and the chasm in the eastern part of the escarpment near Nieuwoudtville. The Cederberg Mountains span roughly 100 km, separating the

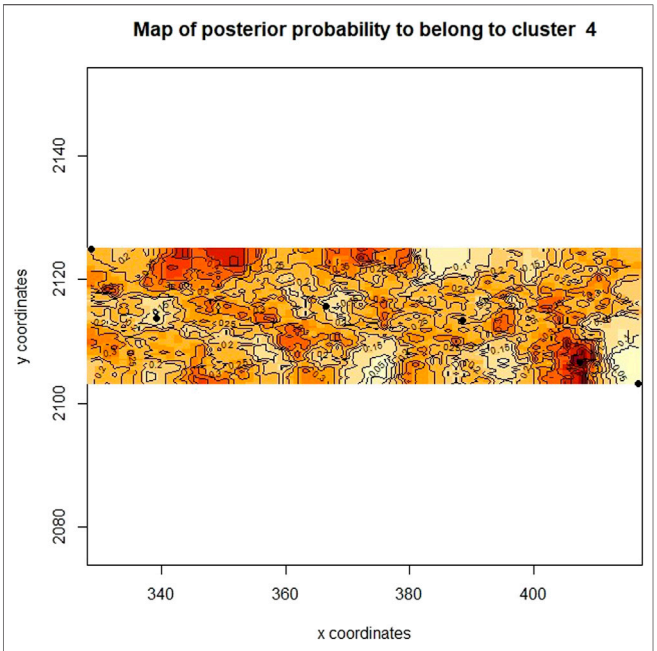

FIGURE 7 | Bayesian clustering analysis displayed as a map of posterior probabilities to belong to cluster 4 (k = 4) as determined in Geneland.

Cederberg populations from the Bokkeveld plateau of the Northern Cape. The Cederberg region forms part of the Cape Fold Mountain range and thus acts as a physical barrier with a channelling effect, restricting and/or limiting dispersal routes. There are also several rivers, namely the Olifants River in the Cederberg area and the Doring River spanning from the Cederberg to the Suid Bokkeveld area (Malgas et al., 2010). These rivers are also important to consider as genetic barriers (Davis et al., 2018). More significantly, the chasm that separates the Dobbelaarskop population from the other Northern Cape populations in the Suid Bokkeveld should be considered as another genetic barrier.

Within this Cape Floristic Region, ants are often responsible for seed dispersal, as with rooibos. This has a significant effect on gene flow. Ants have a relatively low dispersal capacity, only having a range of a few meters (Lötter and le Maître, 2014). This possibly contributes to limited gene flow, which in turn creates genetic structuring between populations that are separated by large physical distances. Habitat specificity also drives local colonisation and this may lead to geographically distinct clades (Wang et al., 2019). *Aspalathus linearis* is known to be a strict endemic and major historical losses of habitat as a result of anthropogenic activities may have led to population contractions, leading to highly fragmented wild populations of rooibos that are low in plant numbers, also contributing to shaping genetic variation and spatial genetic patterns (Hawkins et al., 2011).

The threat to genetic diversity is exacerbated by environmental change and the conversion of wild rooibos habitats for agricultural use. It can be argued that the local mainstream rooibos industry does not necessarily place much value on the wild rooibos (Lötter and le Maitre, 2014; Wynberg, 2017) and as a result, commercial land-users are under economic pressure to cultivate rooibos as their main source of income. For small-scale resource-poor farmers, economic pressures are worse, but over-exploitation is circumvented in two ways. First, wild rooibos is highly valued in niche overseas markets, fetching premium prices that help to offset the profit that would have been made from land conversion and that would have come at the cost of wild rooibos habitats (Malgas et al., 2010). Secondly, the Heiveld Co-operative holds its members strictly accountable for conservation and husbandry of wild rooibos, cultivating an ethics of care for populations in the wild that include co-occurring species and biodiversity in general. In these organisations, the genetic diversity of wild rooibos is also valued. Farmers know from their own local ecological knowledge that wild rooibos is more resilient to climate change, pests and disease (Louw, 2006). Malgas et al. (2010) suggested that some of these distinct population morphotypes and/or chemotypes that were prominent in some areas in the past may no longer exist in the present. This loss of diversity impacts genetic variability and has the potential to lead to the fixation of alleles, reduced adaptability to environmental stressors and heightens the chances of inbreeding.

Malgas et al. (2010) discussed the correspondence between molecular analyses and morphology-based analyses on geographical locations when investigating haplotypic variation. Because *A. linearis* is highly specialised in its spatial distribution, which contributes to its phenology, molecular physiology, the microenvironment and other ecosystem-driven attributes, it may thus have played a greater contribution to adaptive evolutionary genetic traits. Often, phylogeographical patterns are dynamic in nature and are continually being influenced by adaptive potential, ecological interactions, and climate change (Tolley et al., 2014).

TABLE 10 | Pairwise F_{ST} estimates between clusters.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1				
Cluster 2	0.049			
Cluster 3	0.005	0.049		
Cluster 4	0.060	0.119	0.069	

Contemporary Influences on Rooibos Genetic Diversity and Genetic Differentiation

To the best of our knowledge, no previous assessment of genetic diversity or population structure using microsatellite markers has been conducted in *Aspalathus linearis*. Previous genetic studies on this species utilised isozymes (Van der Bank et al., 1995; Van der Bank et al., 1999), as well as chloroplast sequencing and a single nuclear region (Malgas et al., 2010), focusing on the evolution of resprouters versus reseeders, haplotypic variation, and phylogenetic relationships. This current study found a higher ratio of observed number of alleles (An) to effective number of alleles (Aee). This could indicate that the alleles represented

across all of the sampled populations are quite variable between populations and not all alleles are present in every population. A low-to-moderate genetic diversity was found based on the diversity indices of the wild rooibos populations (**Table 7**; Wang et al., 2019). Lower genetic diversity could be explained by species with small population sizes and this concurs with the observation that some of the collected wild rooibos populations were, noticeably, growing as small and patchy populations. This could be due to seasonal changes, which could lead to overexploitation when populations are thriving. Such practices are likely to continue unabated in the future and this particular study also serves to highlight the urgent need for the conservation of unique genetic populations of rooibos. Biological characteristics, reproductive ecology, and geography are key factors that influence genetic diversity (Ellegren and Galtier, 2016; Selseleh et al., 2019) and those plants that are found to occur as small populations with a sporadic distribution may thus show limited genetic variation as individual plants in isolation are thus likely to reproduce with each other. It is important to note that low genetic diversity could also be owing to the number of individuals that were collected. Additionally, low genetic diversity may be associated with high levels of relatedness between certain populations (**Table 8**; **Figure 6A**). For this reason, inbreeding (F_{IS}) was investigated but little to no inbreeding was detected in the populations focused on in this study (**Table 5**; $F_{IS} = 0.038$).

Gene flow is directly influenced through seed dispersal and pollinators (Schaal et al., 1998) and spatial connectedness of populations may thus be important for pollination where insects are the main pollinators. Pollination, in particular, plays a fundamental role in species diversity and cross-pollinated plants have more genetic variation than those plants that are self-pollinated. Rooibos is dependent on flying pollinators and ants for seed dispersal (Herbst, 2011; Lötter and le Maître, 2014; Melin et al., 2014). Although flying pollinators have the potential to reach further distances, seeds that are dispersed by ants may be limited in their dispersal mechanisms (Lötter and le Maître, 2014). Intrinsic genetic variation enables plants to respond to changing environmental conditions and large seasonal variation and as a result, fluctuation in pollination patterns. Micro-climates are known to contribute to localised adaptations that display epigenetic changes and these traits become heritable over time (Malgas et al., 2010; Kronholm and Collins, 2016).

The Heuningvlei and Jamaka populations showed the highest relatedness within populations (**Figure 6A**) and were also the most differentiated according to the F_{ST} estimates (**Table 6**). These two populations both occur within the Cederberg Mountain range, are located only 18 km apart, yet spatially separated by these mountains, with the Heuningvlei population situated in a valley. It may thus be expected that higher levels of gene flow are likely to occur. As these populations occur in closer proximity, greater genetic exchange between these two populations would thus be expected and the Cederberg Mountains represent a geographic boundary, isolating these populations from the Northern Cape group into a distinct lineage. The landscape genetic analysis performed in this study (Geneland cluster analysis and IBD Mantel test) supports the restriction of gene exchange resultant from these vast regions with geographic boundaries. The overlapping areas in posterior probabilities of population membership allude to the possibility of barriers to gene flow, and these coincide with the Cederberg Mountainous region between the two sampling regions, Cederberg and Nieuwoudtville (**Figure 7**). Furthermore, the microclimates between Heuningvlei and Jamaka are more similar in comparison to those plants found in the Northern Cape where the average temperature is lower and more rainfall occurs than in the Cederberg. Other factors that may be considered as being important to drive greater genetic relatedness within the Heuningvlei and Jamaka groups may be the influence of reproductive ecology, particularly reproductive barriers and the status of pollination and seed dispersal. In fact, reproductive barriers are known to limit gene flow, greatly influencing population structure and support genetic differentiation (Schaal et al., 1998). These two Cederberg populations are genetically differentiated ($F_{SC} = 0.053$, $p < 0.05$), most likely as a result of physical isolation, despite some obvious similarities in terms of their morphological appearance and chemical profiles (Stander et al., 2017). This results in distinct populations which are often overexploited, reducing the number of individuals within the population itself. The limited gene flow can influence genetic variation and over time, a high level of relatedness within these populations could result, as seen in this present study.

Available scientific information regarding the reproductive ecology associated with rooibos is tenuous. As far as we are aware, the exact plant-pollinator networks for *Aspalathus* remain ill-defined but wasps and bees are thought to be the important animals for pollinating rooibos although no study has focused on this directly (Gess, 2000; Herbst, 2011). For this reason, it thus becomes more difficult to explain likely effects linked to genetic structure based on a reproductive ecology context. Wild rooibos populations do not display both mechanisms of fire survival strategy; these mechanisms are mutually exclusive on a population level (Van der Bank et al., 1999). Because rooibos populations display either reseeding or resprouting mechanisms for the vegetative establishment of new plants, various adaptations such as this, may also then influence intra- and inter-population dynamics at local and regional scales.

Our data suggests that genetic diversity of wild rooibos populations decline along a gradient, from Melkkraal in the north to Jamaka in the south (**Figure 1**). This may contribute to the understanding that the Cederberg is the center of endemism of rooibos and that the populations in the Suid Bokkeveld have radiated out of the Cederberg over time explaining the increased diversity in the Northern Cape. This data illustrates the potential of clinal variation through gradual variation of a trait being inherited over time across a geographic gradient though there is not sufficient evidence to substantiate this claim. This geographical gradient could be altitude, climate or other environmental influences. Additionally, it is important to consider that these two regions have different biomes and the transition zones between them could result in increased species richness and thus could explain the increased diversity in the Northern Cape populations. Clinal variation could imply restricted gene flow and results in phenotypic diversity

(Takahashi et al., 2019) and in the case of rooibos, interpopulation metabolomic differences (Stander et al., 2017). Similar results were observed by Van der Bank et al. (1999), where it was inferred that speciation would be more likely to occur in reseeder populations and that resprouters have a higher possibility for clinal variation. In this current study, all the populations investigated in the Suid Bokkeveld are resprouters and the Cederberg populations are reseeders even though resprouters do occur in the Cederberg.

There was a significant correlation between geographical distance and genetic distance (**Figure 5B**). It is important to consider spatial differences as well as geographical barriers and how that might influence seed dispersal and pollination and its subsequent contribution to genetic population structure. Seasonal variation also largely determines the distribution of flying pollinators. Rooibos is typically pollinated from September to November with a few still flowering in January (Malgas et al., 2011). It is possible that these wild populations flower at different times and flowering may not always be synchronised amongst diverse population groups, and that is related to the microclimates or specific locality where these unique populations are found (Joubert et al., 2008). Differences in flowering strategies are likely to influence the likelihood of reproduction between these ecotypes. The intensification of agriculture has caused significant changes to the landscape of both the Western Cape and Northern Cape over time, leading to major biodiversity losses for both plants and animals (Vlok and Raimondo, 2011; Schutte-Vlok and Raimondo, 2020). The wild plants and the ecosystems of the Cape floral region are vastly different from what was observed in the past, and natural plant stands that act as refugia for insect pollinators are patchy and fragmented with continuous habitats that once were in existence, no longer available (Tolley et al., 2014). This has an impact on foraging distances for pollinators and unfortunately alters dispersal mechanisms as traveling distances for nesting, and nutrient resources become further apart. The connection of ecosystems to each other becomes less possible and so geospatial distance can in such a way influence the genetic makeup of plants of the same species.

Pairwise F_{ST} analyses revealed moderate genetic differentiation whereas DAPC analysis determined genetic structure across all sampled populations. These results show a high level of congruence between the genetic data sets, confirming the genetic patterns resolved using the chloroplast intergenic region (**Table 5**). The DAPC is a powerful multivariate approach to resolve the number of genetic clusters that are present between the populations without prior knowledge of their genetic relationship and thus reduces population bias. The DAPC plot reveals genetic structure without the assumption that the populations are panmictic (Jombart et al., 2010). In this study, the DAPC indicated that the reseeder populations (Jamaka and Heuningvlei ecotypes) are genetically distinct from the Northern Cape resprouter populations (**Figure 3B**). These fire survival strategies may limit genetic hybridisation; leading to population-isolated types that express particular phenotypes. The DAPC also supports the moderate genetic differentiation that is evident from the pairwise F_{ST} analysis. Additionally, the Bayesian structure analysis revealed the separation of populations into three clusters, the Northern Cape populations clustering together, and the two Western Cape populations clustered individually as distinct populations (k = 3; **Figure 4**). This consolidates the DAPC, strengthening the evidence of reseeders and resprouters being genetically distinct. This finding is not necessarily new as similar results have been reported by Van der Bank et al. (1999) but it serves to corroborate that particular study which was based on isozymes. Van der Bank et al. (1999) proposed that resprouters are derived from reseeders. Unfortunately, that study did not mention the evolutionary time period whereby this diversification might have occurred. It is interesting that this proposed separation of reseeders and resprouters has been maintained in present times. Historical events such as population bottlenecks are strong indicators of genetic structure (Schaal et al., 1998; Davis et al., 2018). The separation of resprouters versus reseeders highlights the potential effects of genetic drift, isolation by distance, fire survival strategies, and environmental differences between the region sampled as well as the combination of all of these factors and how these could be responsible for the populations decreasing in numbers and may need to be investigated further (Clarke et al., 2013). Furthermore, the DAPC showed separation of the two Cederberg populations (Western Cape), corroborating pairwise F_{ST} values (**Table 6**), and likely indicating that they are genetically distinct. This is interesting because although these populations are in close proximity to each other, the biogeographic landscape may create a channelling effect, limiting extensive genetic exchange. This could be due to the geographical barrier between these two populations as the Cederberg Mountains create a terrain of valleys and peaks with the Heuningvlei population found in a deep valley. Moreover, these two populations are also morphologically different. The Jamaka population was particularly unique as it was smaller in size compared to most plants and had blue-green coloured leaves as opposed to bright green. The presence of three main clusters is clearly reflected in the DAPC, based on the current suite of microsatellite markers. The inclusion of additional microsatellites could potentially lead to higher resolution in terms of population structure at the intra-regional level.

CONCLUSION

Haplotype divergence of the ecotypes from the Cederberg and Suid Bokkeveld provided insights into the genetic history of these populations and there was a clear separation between resprouters and reseeders corroborating the original hypothesis. Through using both nuclear markers and chloroplast sequencing, a comprehensive and complementary portrait of the genetic structure of wild rooibos was evident and there was an ancestral haplotype consisting of both reseeders and resprouters. This data may be indicative of clinal genetic variation that suggests decreased diversity from the Suid Bokkeveld populations down into the Cederberg region. Overall, low intra-specific population diversity was strongly evident in wild collected rooibos, particularly in the reseeder populations of Jamaka and Heuningvlei. Wild rooibos

populations occur as sparse collections with few individuals and limited options for allowing frequent gene flow, making them more susceptible to biodiversity loss. The results presented herein thus highlight the importance of assessing genetic variability and the need for implementing strategies for conservation priorities for rooibos that is a highly restricted endemic growing in biogeographic regions that face both habitat degradation and future climate changes. In this particular context, the correct and appropriate management of wild genetic *A. linearis* resources is thus strongly encouraged as distinct gene pools have been confirmed in this study. Many land-users have commented on the population decline over the last few decades and these knowledge-holders have emphasised that more recently it is becoming more difficult every year to find a suitable harvest of wild populations due to populations declining. Conservation initiatives may prove to be of value for both *in situ* and *ex situ* strategies. Additionally, it is important to prioritise conservation efforts at every step of the supply chain, particularly for such a uniquely endemic species such as rooibos. Conservation strategies could include thorough monitoring and record-keeping of wild harvesting that supports livelihoods as well as deposits of wild populations to a gene bank for the conservation of distinct populations.

DATA AVAILABILITY STATEMENT

The data presented in this study are deposited in NCBI, accession numbers OK771546-OK771592 (trnLF chloroplast sequences) and OL438920-OL438931 (nuclear microsatellite sequences).

AUTHOR CONTRIBUTIONS

This paper is based on the data generated by JB in partial fulfilment of degree requirements (http://scholar.sun.ac.za). JB, RM, and RR-W assisted in building on the idea that was initially proposed and conceptualised by NM. JB conducted the laboratory experiments prior to KH and MB-H assisting with raw data analyses and interpretation the *trn*LF and microsatellite data. The first draft of this paper was generated by JB and edited by NM, RM, and RR-W. All authors declare that they read, commented on, and approved the final article.

FUNDING

The National Research Foundation (NRF) of South Africa provided funding for this study (Grant number: 95442 and 76555; awarded to NM) and the DAAD-NRF scholarship was awarded to JB in 2018.

ACKNOWLEDGMENTS

The opinions, findings, and conclusion or recommendations expressed are those of the author(s) alone, and the NRF accepts no liability whatsoever in this regard. Plant collections were conducted under a CapeNature (Permit number: CN35-28-268) and the Heiveld Cooperation is thanked for their assistance and permission to collect the wild rooibos samples. We are grateful to the Heuningvlei community, particularly Dalene van der Westhuizen. The DNA sequencing was conducted at the Central Analytical Facility of Stellenbosch University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.761988/full#supplementary-material

REFERENCES

Bandelt, H. J., Forster, P., and Rohl, A. (1999). Median-joining Networks for Inferring Intraspecific Phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi:10.1093/oxfordjournals.molbev.a026036

Baskauf, C. J., and Burke, J. M. (2009). Population Genetics of *Astragalus bibullatus* (Fabaceae) Using AFLPs. *J. Hered.* 100, 424–431. doi:10.1093/jhered/esp033

Bond, T. J., and Derbyshire, E. J. (2020). Rooibos tea and Health: A Systematic Review of the Evidence from the Last Two Decades. *Nutr. Food Techn.* 6, 1–11. doi:10.16966/2470-6086.166

Borse, T., Joshi, P., and Chaphalkar, S. (2011). Biochemical Role of Ascorbic Acid during the Extraction of Nucleic Acids in Polyphenol Rich Medicinal Plant Tissues. *J. Plant Mol. Biol. Biotechnol.* 2, 1–7.

Brooks, J. (2021). *Assessing the Phylogeography and Metabolomic Signatures of Wild Rooibos (Aspalathus linearis) Ecological Populations*. Stellenbosch, (South Africa): The Botany Department, Stellenbosch University. Masters of Science dissertation.

Clarke, P. J., Lawes, M. J., Midgley, J. J., Lamont, B. B., Ojeda, F., Burrows, G. E., et al. (2013). Resprouting as a Key Functional Trait: How Buds, protection and Resources Drive Persistence after Fire. *New Phytol.* 197, 19–35. doi:10.1111/nph.12001

Collevatti, R. G., Grattapaglia, D., and Hay, J. D. (2003). Evidences for Multiple Maternal Lineages of *Caryocar brasiliense* Populations in the Brazilian Cerrado Based on the Analysis of Chloroplast DNA Sequences and Microsatellite Haplotype Variation. *Mol. Ecol.* 12, 105–115. doi:10.1046/j.1365-294x.2003.01701.x

Dahlgren, R. (1968). Revision of the Genus *Aspalathus*. II. The Species with Ericoid and Pinoid Leaflets. 7. Subgenus Nortieria, with Remarks on Rooibos tea Cultivation. *Botaniska Notiser* 121, 165–208.

Davis, C. D., Epps, C. W., Flitcroft, R. L., and Banks, M. A. (2018). Refining and Defining Riverscape Genetics: How Rivers Influence Population Genetic Structure. *WIREs Water* 5, 1269–1295. doi:10.1002/wat2.1269

Demenou, B. B., and Hardy, O. J. (2017). Development, Characterization, and Cross-Amplification of Microsatellite Markers in the Understudied African Genus *Anthonotha* (Fabaceae). *Appl. Plant Sci.* 5, 160–167. doi:10.3732/apps.1600120

Edwards, D., Horn, A., Taylor, D., Savolainen, V., and Hawkins, J. A. (2008). DNA Barcoding of a Large Genus, *Aspalathus* L. (Fabaceae). *Taxon* 57, 1317–4E. doi:10.1002/tax.574021

Ellegren, H., and Galtier, N. (2016). Determinants of Genetic Diversity. *Nat. Rev. Genet.* 17, 422–433. doi:10.1038/nrg.2016.58

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the Number of Clusters of Individuals Using the Software Structure: a Simulation Study. *Mol. Ecol.* 14, 2611–2620. doi:10.1111/j.1365-294x.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: a New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Galuszynski, N. C., and Potts, A. J. (2020). Applied Phylogeography Of Cyclopia Intermedia (Fabaceae) Highlights The Need For "Duty Of Care" When Cultivating Honeybush. *PeerJ.* 8, 9818. doi:10.7717/peerj.9818

Gess, S. (2000). Rooibos: Refreshment for Humans, Bees and Wasps. *Veld & Flora* 86, 19–21.

Guillot, G., Santos, F., and Estoup, A. (2011). *Populations Genetics Analysis Using R and the Geneland Program*. Lyngby, Denmark: Technical University of Denmark.

Hall, T. A. (1999). BioEdit: A User-friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Hawkins, H.-J., Malgas, R., and Biénabe, E. (2011). Ecotypes of Wild Rooibos (*Aspalathus linearis* (Burm. F) Dahlg., Fabaceae) Are Ecologically Distinct. *South Afr. J. Bot.* 77, 360–370. doi:10.1016/j.sajb.2010.09.014

Herbst, M. (2011). *Ecosystem Functioning, Ecosystem Services and Rooibos Production as Affected by Connectivity to Natural Vegetation and Agrochemical Use in Rooibos tea (Aspalathus linearis) Farming*. Town: Unpublished MSc dissertation Botany Department, University of Cape.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant Analysis of Principal Components: a New Method for the Analysis of Genetically Structured Populations. *BMC Genet.* 11, 94–109. doi:10.1186/1471-2156-11-94

Joubert, E., and de Beer, D. (2011). Rooibos (*Aspalathus linearis*) beyond the Farm Gate: from Herbal tea to Potential Phytopharmaceutical. *South Afr. J. Bot.* 77, 869–886. doi:10.1016/j.sajb.2011.07.004

Joubert, E., Gelderblom, W. C. A., Louw, A., and de Beer, D. (2008). South African Herbal Teas: *Aspalathus linearis*, *Cyclopia* Spp. And *Athrixia phylicoides*-A Review. *J. Ethnopharmacology* 119, 376–412. doi:10.1016/j.jep.2008.06.014

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak : a Program for Identifying Clustering Modes and Packaging Population Structure Inferences across K. *Mol. Ecol. Resour.* 15, 1179–1191. doi:10.1111/1755-0998.12387

Kronholm, I., and Collins, S. (2016). Epigenetic Mutations Can Both Help and Hinder Adaptive Evolution. *Mol. Ecol.* 25, 1856–1868. doi:10.1111/mec.13296

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054

Li, Y.-L., and Liu, J.-X. (2018). StructureSelector : A Web-Based Software to Select and Visualize the Optimal Number of Clusters Using Multiple Methods. *Mol. Ecol. Resour.* 18, 176–177. doi:10.1111/1755-0998.12719

Librado, P., and Rozas, J. (2009). DnaSP V5: a Software for Comprehensive Analysis of DNA Polymorphism Data. *Bioinformatics* 25, 1451–1452. doi:10.1093/bioinformatics/btp187

Lötter, D., and Maitre, D. (2014). Modelling the Distribution of *Aspalathus linearis* (Rooibos tea): Implications of Climate Change for Livelihoods Dependent on Both Cultivation and Harvesting from the Wild. *Ecol. Evol.* 4, 1209–1221. doi:10.1002/ece3.985

Louw, R. (2006). *Sustainable Harvesting of Wild Rooibos (Aspalathus linearis) in the Suid Bokkeveld, Northern Cape*. Northern Cape: Unpublished MSc dissertation. Botany department, University of Cape Town.

Malgas, R., Oettle, N., and Koelle, B. (2011). The Heiveld Co-operative:. *Case Stud. Emerging Farmers Agribusinesses South Africa*, 172–191. doi:10.2307/j.ctv1v7zcch.16

Malgas, R. R., Potts, A. J., Oettlé, N. M., Koelle, B., Todd, S. W., Verboom, G. A., et al. (2010). Distribution, Quantitative Morphological Variation and Preliminary Molecular Analysis of Different Growth Forms of Wild Rooibos (*Aspalathus linearis*) in the Northern Cederberg and on the Bokkeveld Plateau. *South Afr. J. Bot.* 76, 72–81. doi:10.1016/j.sajb.2009.07.004

Marais, K. E., Pratt, R. B., Jacobs, S. M., Jacobsen, A. L., and Esler, K. J. (2014). Postfire Regeneration of Resprouting Mountain Fynbos Shrubs: Differentiating Obligate Resprouters and Facultative Seeders. *Plant Ecol.* 215, 195–208. doi:10.1007/s11258-013-0289-4

Médail, F., and Baumel, A. (2018). Using Phylogeography to Define Conservation Priorities: The Case of Narrow Endemic Plants in the Mediterranean Basin Hotspot. *Biol. Conservation* 224, 258–266. doi:10.1016/j.biocon.2018.05.028

Melin, A., Rouget, M., Midgley, J. J., and Donaldson, J. S. (2014). Pollination Ecosystem Services in South African Agricultural Systems. *Sajs* 110, 1–9. doi:10.1590/sajs.2014/20140078

Midgley, G. F., Hannah, L., Millar, D., Thuiller, W., and Booth, A. (2003). Developing Regional and Species-Level Assessments of Climate Change Impacts on Biodiversity in the Cape Floristic Region. *SAJS* 112, 87–97. doi:10.1016/s0006-3207(02)00414-7

Moyo, M., Bairu, M. W., Amoo, S. O., and van Staden, J. (2011). Plant Biotechnology in South Africa: Micropropagation Research Endeavours, Prospects and Challenges. *South Afr. J. Bot.* 77, 996–1011. doi:10.1016/j.sajb.2011.06.003

Niemandt, M., Roodt-Wilding, R., Tobutt, K. R., and Bester, C. (2018). Microsatellite Marker Applications in *Cyclopia* (Fabaceae) Species. *South Afr. J. Bot.* 116, 52–60. doi:10.1016/j.sajb.2018.02.408

Nieto Feliner, G. (2014). Patterns and Processes in Plant Phylogeography in the Mediterranean Basin. A Review. *Perspect. Plant Ecol. Evol. Syst.* 16, 265–278. doi:10.1016/j.ppees.2014.07.002

Park, S. (2001). *The Excel Microsatellite Toolkit*. Dublin, Ireland: Animal Genomics Laboratory University College.

Pausas, J. G., and Keeley, J. E. (2014). Evolutionary Ecology of Resprouting and Seeding in Fire-prone Ecosystems. *New Phytol.* 204, 55–65. doi:10.1111/nph.12921

Peakall, R., and Smouse, P. E. (2006). GenAlEx 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research. *Mol. Ecol. Notes* 6, 288–295. doi:10.1111/j.1471-8286.2005.01155.x

Pollock, L. J., Rosauer, D. F., Thornhill, A. H., Kujala, H., Crisp, M. D., Miller, J. T., et al. (2015). Phylogenetic Diversity Meets Conservation Policy: Small Areas Are Key to Preserving Eucalypt Lineages. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370, 20140007–20140010. doi:10.1098/rstb.2014.0007

Potts, A. J. (2017). Genetic Risk and the Transition to Cultivation in Cape Endemic Crops-The Example of Honeybush (*Cyclopia*). *South Afr. J. Bot.* 110, 52–56. doi:10.1016/j.sajb.2016.09.004

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–959. doi:10.1093/genetics/155.2.945

Puechmaille, S. J. (2016). The Program Structure does not reliably recover the Correct Population Structure when Sampling Is Uneven: Subsampling and New Estimators Alleviate the Problem. *Mol. Ecol. Resour.* 16, 608–627. doi:10.1111/1755-0998.12512

Queller, D. C., and Goodnight, K. F. (1989). Estimating Relatedness Using Genetic Markers. *Evolution* 43, 258–275. doi:10.1111/j.1558-5646.1989.tb04226.x

Rebelo, A. G., Boucher, C., Helme, N., Mucina, L., and Rutherford, M. C. (2006). "Fynbos Biome," in *The Vegetation of South Africa, Lesotho and Swaziland*. Editors L. Mucina and M. C. Rutherford (Pretoria: South African National Biodiversity Institue) Strelitzia 19, 53–219.

Rousset, F. (2008). genepop'007: a Complete Re-implementation of the Genepop Software for Windows and Linux. *Mol. Ecol. Resour.* 8, 103–106. doi:10.1111/j.1471-8286.2007.01931.x

Schaal, B. A., Hayworth, D. A., Olsen, K. M., Rauscher, J. T., and Smith, W. A. (1998). Phylogeographic Studies in Plants: Problems and Prospects. *Mol. Ecol.* 7, 465–474. doi:10.1046/j.1365-294x.1998.00318.x

Schutte-Vlok, A. L., and Raimondo, D. (2020). Threatened Species Programme | SANBI Red List of South African Plants. National Assessment: Red List of South African Plants 2020. Available at: http://redlist.sanbi.org/species.php?species=446-41 (accessed 22 6, 21).

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Stat.* 6, 461–464. doi:10.1214/aos/1176344136

Selseleh, M., Hadian, J., Nejad Ebrahimi, S., Sonboli, A., Georgiev, M. I., and Mirjalili, M. H. (2019). Metabolic Diversity and Genetic Association between Wild Populations of *Verbascum songaricum* (Scrophulariaceae). *Ind. Crops Prod.* 137, 112–125. doi:10.1016/j.indcrop.2019.03.069

Smith, C., and Swart, A. (2018). *Aspalathus linearis* (Rooibos) - a Functional Food Targeting Cardiovascular Disease. *Food Funct.* 9, 5041–5058. doi:10.1039/c8fo01010b

Smith, J. F. N., Botha, A., and Hardie, A. G. (2018). Role of Soil Quality in Declining Rooibos (*Aspalathus linearis*) tea Yields in the Clanwilliam Area, South Africa. *Soil Res.* 56, 252–263. doi:10.1071/sr17029

Stander, M. A., Joubert, E., and de Beer, D. (2019). Revisiting the Caffeine-free Status of Rooibos and Honeybush Herbal Teas Using Specific MRM and High Resolution LC-MS Methods. *J. Food Compost. Anal.* 76, 39–43. doi:10.1016/j.jfca.2018.12.002

Stander, M. A., van Wyk, B.-E., Taylor, M. J. C., and Long, H. S. (2017). Analysis of Phenolic Compounds in Rooibos tea (*Aspalathus Linearis*) with a Comparison of Flavonoid-Based Compounds in Natural Populations of Plants from Different Regions. *J. Agric. Food Chem.* 65, 10270–10281. doi:10.1021/acs.jafc.7b03942

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., et al. (2007). Power and Limitations of the Chloroplast *trn*L (UAA) Intron for Plant DNA Barcoding. *Nucleic Acids Res.* 35, e14–8. doi:10.1093/nar/gkl938

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. (1991). Universal Primers for Amplification of Three Non-coding Regions of Chloroplast DNA. *Plant Mol. Biol.* 17, 1105–1109. doi:10.1007/bf00037152

Takahashi, D., Teramine, T., Sakaguchi, S., and Setoguchi, H. (2019). Genetic Data Reveals a Complex History of Multiple Admixture Events in Presently Allopatric Wild Gingers (*Asarum* spp.) Showing Intertaxonomic Clinal Variation in Calyx Lobe Length. *Mol. Phylogenet. Evol.* 137, 146–155. doi:10.1016/j.ympev.2019.05.003

Tolley, K. A., Bowie, R. C. K., John Measey, G., Price, B. W., and Forest, F. (2014). *The Shifting Landscape of Genes since the Pliocene: Terrestrial Phylogeography in the Greater Cape Floristic Region*. Oxford, England: Fynbos: ecology, evolution and conservation of a megadiverse region. Oxford University Press, 142–163. doi:10.1093/acprof:oso/9780199679584.003.0007

Van der Bank, M., van der Bank, F. H., and van Wyk, B.-E. (1999). Evolution of Sprouting versus Seeding in *Aspalathus linearis*. *Pl Syst. Evol.* 219, 27–38. doi:10.1007/bf01090297

Van der Bank, M., van Wyk, B.-E., and van der Bank, H. (1995). Biochemical Genetic Variation in Four Wild Populations of *Aspalathus linearis* (Rooibos tea). *Biochem. Syst. Ecol.* 23, 257–262. doi:10.1016/0305-1978(95)00016-n

Van Heerden, F. R., van Wyk, B.-E., Viljoen, A. M., and Steenkamp, P. A. (2003). Phenolic Variation in Wild Populations of *Aspalathus Linearis* (Rooibos tea). *Biochem. Syst. Ecol.* 31, 885–895. doi:10.1016/s0305-1978(03)00084-x

Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. M., and Shipley, P. (2004). MICRO-CHECKER: Software for Identifying and Correcting Genotyping Errors in Microsatellite Data. *Mol. Ecol. Notes* 4, 535–538. doi:10.1111/j.1471-8286.2004.00684.x

Van Wyk, B.-E., and Gorelik, B. (2017). The History and Ethnobotany of Cape herbal teas. *South Afr. J. Bot.* 110, 18–38. doi:10.1016/j.sajb.2016.11.011

Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. d. F. (2016). Microsatellite Markers: What They Mean and Why They Are So Useful. *Genet. Mol. Biol.* 39, 312–328. doi:10.1590/1678-4685-gmb-2016-0027

Vlok, J. H., and Raimondo, D. (2011). *Cyclopia genus, National Assessment: Red List of South African Plants version 2015.1*. Pretoria: South African National Biodiversity Institute.

Wang, H. Y., Yin, X., Yin, D. X., Li, L., and Xiao, H. X. (2019). Population Genetic Structures of Two Ecologically Distinct Species *Betula platyphylla* and *B. ermanii* Inferred Based on Nuclear and Chloroplast DNA Markers. *Ecol. Evol.* 9, 11406–11419. doi:10.1002/ece3.5643

Wang, Y., Zhou, T., Li, D., Zhang, X., Yu, W., Cai, J., et al. (2019). The Genetic Diversity and Population Structure of *Sophora alopecuroides* (Faboideae) as Determined by Microsatellite Markers Developed from Transcriptome. *PLOS ONE* 14, e0226100–17. doi:10.1371/journal.pone.0226100

Wynberg, R. (2017). Making Sense of Access and Benefit Sharing in the Rooibos Industry: Towards a Holistic, Just and Sustainable Framing. *South Afr. J. Bot.* 110, 39–51. doi:10.1016/j.sajb.2016.09.015

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 December 2021
doi: 10.3389/fgene.2021.781451

Genetic Contributors of Incident Stroke in 10,700 African Americans With Hypertension: A Meta-Analysis From the Genetics of Hypertension Associated Treatments and Reasons for Geographic and Racial Differences in Stroke Studies

Nicole D. Armstrong[1], Vinodh Srinivasasainagendra[2], Amit Patki[2], Rikki M. Tanner[1], Bertha A. Hidalgo[1], Hemant K. Tiwari[2], Nita A. Limdi[3], Ethan M. Lange[4], Leslie A. Lange[4], Donna K. Arnett[5] and Marguerite R. Irvin[1]*

[1]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, United States, [2]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, United States, [3]Department of Neurology, University of Alabama at Birmingham, Birmingham, AL, United States, [4]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [5]College of Public Health, University of Kentucky, Lexington, KY, United States

OPEN ACCESS

Edited by:
Alpo Juhani Vuorio, University of Helsinki, Finland

Reviewed by:
Jiang Li, Geisinger Medical Center, United States
Chikashi Terao, riken, Japan

***Correspondence:**
Nicole D. Armstrong
nmda@uab.edu

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *22 September 2021*
Accepted: *23 November 2021*
Published: *21 December 2021*

Citation:
Armstrong ND, Srinivasasainagendra V, Patki A, Tanner RM, Hidalgo BA, Tiwari HK, Limdi NA, Lange EM, Lange LA, Arnett DK and Irvin MR (2021) Genetic Contributors of Incident Stroke in 10,700 African Americans With Hypertension: A Meta-Analysis From the Genetics of Hypertension Associated Treatments and Reasons for Geographic and Racial Differences in Stroke Studies. Front. Genet. 12:781451. doi: 10.3389/fgene.2021.781451

Background: African Americans (AAs) suffer a higher stroke burden due to hypertension. Identifying genetic contributors to stroke among AAs with hypertension is critical to understanding the genetic basis of the disease, as well as detecting at-risk individuals.

Methods: In a population comprising over 10,700 AAs treated for hypertension from the Genetics of Hypertension Associated Treatments (GenHAT) and Reasons for Geographic and Racial Differences in Stroke (REGARDS) studies, we performed an inverse variance-weighted meta-analysis of incident stroke. Additionally, we tested the predictive accuracy of a polygenic risk score (PRS) derived from a European ancestral population in both GenHAT and REGARDS AAs aiming to evaluate cross-ethnic performance.

Results: We identified 10 statistically significant (p < 5.00E-08) and 90 additional suggestive (p < 1.00E-06) variants associated with incident stroke in the meta-analysis. Six of the top 10 variants were located in an intergenic region on chromosome 18 (*LINC01443-LOC644669*). Additional variants of interest were located in or near the *COL12A1, SNTG1*, *PCDH7*, *TMTC1*, and *NTM* genes. Replication was conducted in the Warfarin Pharmacogenomics Cohort (WPC), and while none of the variants were directly validated, seven intronic variants of *NTM* proximal to our target variants, had a p-value <5.00E-04 in the WPC. The inclusion of the PRS did not improve the prediction accuracy compared to a reference model adjusting for age, sex, and genetic ancestry in either study and had lower predictive accuracy compared to models accounting for established stroke risk factors. These results demonstrate the necessity for PRS derivation in AAs, particularly for diseases that affect AAs disproportionately.

Conclusion: This study highlights biologically plausible genetic determinants for incident stroke in hypertensive AAs. Ultimately, a better understanding of genetic risk factors for stroke in AAs may give new insight into stroke burden and potential clinical tools for those among the highest at risk.

Keywords: incident stroke, hypertension, antihypertensives, disparities, polygenic risk score, genome wide association studies

INTRODUCTION

As the second-leading global cause of death and a leading cause of disability-adjusted life-years lost (Katan and Luft, 2018), stroke is a major public health burden especially among African Americans (AAs) who have a 50% higher risk of stroke (Howard et al., 2011). The role of genetics on stroke risk has been evidenced through twin studies, where monozygotic twins are more likely to be concordant than dizygotic twins (odds ratio for concordance (OR): 1.65, 95% CI: 1.2–2.3) (Flossmann et al., 2004). Previously reported heritability estimates of stroke based on data from genome-wide association studies (GWAS) are comparable for AAs and individuals of European Ancestry (EAs); 38% for EAs (Bevan et al., 2012) and 35% for AAs (Traylor et al., 2017), yet AAs contribute far less data to a large body of literature on stroke genetic risk factors (Bentley et al., 2020).

GWAS of stroke in European populations have identified susceptibility loci located on chromosomes 4q25 and 9p21 (Gretarsdottir et al., 2008), 7p21.1 (International Stroke Genetics, 2012), 6p21 (Holliday et al., 2012), 12p13 (Ikram et al., 2009), 12q24 (Kilarski et al., 2014), and 16q22 (Traylor et al., 2012). Overall, there has been a lack of similar primary GWAS discovery efforts in AAs, and results from EAs have not generally replicated in AAs (Carty et al., 2015; Peprah et al., 2015). For example, of 520,000 participants in the large trans-ethnic GWAS meta-analysis from the MEGASTROKE consortium, only ~4% of participants were AAs (Malik et al., 2018). The only other study with a large number of AAs, the Consortium of Minority Population Genome-Wide Association Studies of Stroke (COMPASS), confirmed the need for race-specific stroke discovery, finding weak or no validation across ethnic groups (Carty et al., 2015; Keene et al., 2020). Given the relative lack of data on this race group in the published literature, additional stroke variant discovery in this population remains needed.

GWAS data capitalizes on the polygenic nature of common diseases and has been collected on a large scale to provide useful health information over traditional clinical risk factors. So far GWAS data at the single variant level, even from large consortia studies such as those described above, has been difficult to translate to the clinic. More recently, cumulative single variant effects from GWAS, at thousands to millions of variants, are being used to estimate polygenic risk scores (PRS), which may prove useful for stroke risk prediction and prevention. Unfortunately, few stroke PRS have been developed in populations including AAs and, to the best of our knowledge, only one PRS was developed exclusively in individuals of African descent, consisting of only 29 variants (Traylor et al., 2017).

Abraham et al. (2019) comprised a composite meta-risk PRS (metaPRS) for ischemic stroke leveraging publicly available GWAS data for stroke and 19 stroke-related phenotypes generated from the British European dataset of the United Kingdom Biobank (UKB, n = 407,388). Under a split sample design, participants with 'any stroke' event were oversampled for the derivation set (n = 11,995 total; n = 2,065 with "any stroke" events) followed by validation in the remaining ~395,000 participants. The authors did not report on score validation in the small sample of available individuals of African descent from the UKB (N~8,000). Given difficulties in cross-ethnic validation of stroke variants in GWAS, we suspected that the publicly available Abraham composite metaPRS will not perform adequately in AAs.

The goals of the current study, which we set among AAs at elevated risk for stroke due to hypertension (HTN), were two-pronged. First, we aimed to increase stroke GWAS data available in this race group using data from the Genetics of Hypertension Associated Treatments (GenHAT) study and the Reasons for Geographic and Racial Differences in Stroke (REGARDS) studies. Second, we aimed to assess the predictive ability of the UKB derived stroke PRS (Abraham et al., 2019) in these two populations given the lack of previous cross-ethnic validation.

MATERIALS AND METHODS

Study Participants

GenHAT was an ancillary pharmacogenetic study to the Antihypertensive and Lipid-Lowering Treatment To Prevent Heart Attack Trial (ALLHAT) and REGARDS is an ongoing cohort study of stroke risk in the continental US. The GenHAT and REGARDS studies contributed genetic and phenotypic data on 10,717 AAs with HTN at baseline (GenHAT n = 6,908; REGARDS n = 3,809), further detailed in the **Supplementary Material**. Participants in GenHAT were randomized to either chlorthalidone, a thiazide diuretic (TD), or lisinopril, an angiotensin-converting enzyme (ACE) inhibitor, while REGARDS participants were taking a TD, ACE inhibitor, or a combination of both for inclusion in the analysis (**Supplementary Table S1**). All studies were approved by local institutional review boards and/or ethics committees. All participants provided written informed consent.

Definition of Incident Stroke

In GenHAT, incident stroke was defined as the rapid onset of persistent neurologic deficit attributable to an obstruction or rupture of the arterial system, including stroke occurring

TABLE 1 | Demographic characteristics of GenHAT and REGARDS participants stratified by incident stroke case-control status.

	GenHAT			REGARDS		
	Cases (*n* = 366)	**Controls (*n* = 6,542)**	***p***	**Cases (*n* = 280)**	**Controls (*n* = 3,529)**	***p***
Age, years	68.70 ± 8.19	66.02 ± 7.70	<0.001	67.41 ± 9.01	64.42 ± 8.81	<0.001
Sex						
Male	191 (52.2%)	2,893 (44.2%)	0.003	102 (36.4%)	1,310 (37.1%)	0.868
Female	175 (47.8%)	3,649 (55.8%)		178 (63.6%)	2,219 (62.9%)	
Antihypertensive class						
Thiazide diuretic	190 (51.9%)	4,107 (62.8%)	<0.001	88 (31.4%)	1,418 (40.2%)	0.013
ACE Inhibitor	176 (48.1%)	2,435 (37.2%)		123 (43.9%)	1,311 (37.1%)	
Combination therapy[a]	NA	NA		69 (24.6%)	800 (22.7%)	
Cigarette smoking	89 (31.7%)	1,506 (27.1%)	0.104	39 (14.0%)	591 (16.8%)	0.255
DM	188 (51.4%)	2,588 (39.6%)	<0.001	136 (49.5%)	1,325 (38.2%)	<0.001
BMI	29.65 ± 6.34	30.52 ± 6.60	0.014	31.20 ± 5.81	31.83 ± 6.76	0.126
SBP, mmHg	148.37 ± 15.98	146.13 ± 15.75	0.008	135.65 ± 16.90	132.03 ± 16.78	0.001
DBP, mmHg	84.64 ± 10.90	84.88 ± 10.12	0.667	78.42 ± 10.41	78.74 ± 10.04	0.603
eGFR, mL/min/1.73 m^2	78.62 ± 22.88	82.95 ± 21.68	<0.001	83.44 ± 25.52	86.47 ± 27.59	0.079

[a]*GenHAT participants were taking either thiazide diuretic or ACE inhibitor at baseline.*

Categorical variables are described as N (%), while continuous variables are described as mean ± standard deviation. Abbreviations: TD, thiazide diuretic class; ACE, angiotensin-converting enzyme; DM, diabetes mellitus; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate.

during surgery, not known to be secondary to brain trauma, tumor, infection, or other non-ischemic cause and must last more than 24 h unless death supervenes or there is a demonstrable lesion compatible with acute stroke on CT or MRI scan (ALLHAT Protocol, 2000). In REGARDS, any suspected stroke events were identified every 6 months *via* telephone interview. The medical records associated with these events were retrieved and adjudicated by a physician, using the World Health Organization definition of stroke as focal neurologic symptoms lasting more than 24 h or those with neuroimaging data consistent with stroke (WHO Task Force on Stroke and other Cerebrovascular Disorders, 1989; Howard et al., 2011). For REGARDS, all stroke events occurring before or on September 30, 2019 were included in this analysis. REGARDS participants with a history of previous stroke were excluded.

Genotyping, Quality Control, and Imputation

Genome-wide genotyping was performed within each study independently using Illumina Infinium Multi-Ethnic AMR/AFR Extended BeadChip arrays (MEGA chip; Illumina, Inc., San Diego, CA). **Supplementary Table S2** includes detailed information on the variant and sample quality control (QC). Filtered genotype calls were imputed using the NHLBI Trans-omics for Precision Medicine (TOPMed) release 2 (Freeze 8) reference panel, which leverages data on ~186,000 samples (~30% AA) (Fuchsberger et al., 2015; Das et al., 2016; Taliun et al., 2021). Post-imputation QC excluded variants with imputation quality scores (Rsq) < 0.3 and a Minor Allele Count (MAC) < 20 in each cohort, as previously described using TOPMed data (Sarnowski et al., 2021). The variants not represented in both GenHAT and REGARDS were excluded from subsequent analyses.

Statistical Analysis

Baseline descriptive statistics for cases and controls are presented as counts (percentages) for categorical variables or mean ± standard deviation (SD) for continuous variables, and were compared using χ^2 tests and *t*-tests, respectively (**Table 1**).

Firth logistic regression models implemented in PLINK2 (Ma et al., 2013; Chang et al., 2015) were used for genome-wide association analysis of incident stroke status. Models of the imputed effect allele dosage were adjusted for age, sex, and the top 10 principal components (PCs) for ancestry derived in EIGENSTRATv6.1.4 (Price et al., 2006). An inverse variance-weighted, fixed effects meta-analysis was performed on the summary statistics from GenHAT and REGARDS, using METAL software (Willer et al., 2010). Statistical heterogeneity was evaluated using Cochran's chi-square test in METAL. Regional plots were created using LocusZoom v0.12 (Pruim et al., 2010; Boughton et al., 2021). Gene annotation was completed using ANNOVAR (Wang et al., 2010). Genome-wide significance was set at $p < 5.00E\text{-}08$ after a Bonferroni correction for multiple testing. A randomization drug-adjusted sensitivity model was performed in the GenHAT data to account for any effects of the antihypertensive agent, and the results were similar to the discovery model (data not shown). To address potential issues associated with case-control imbalance, we ran sensitivity analyses in the GenHAT and REGARDS data using the saddle point approximation (SPA) in the Scalable and Accurate Implementation of Generalized (SAIGE) program and meta-analyzed the results (Zhou et al., 2018). This approach provides good control of type 1 error for binary traits, however the SPA approach in SAIGE is very conservative and has been described to result in inflated effect-size estimates (Mbatchou et al., 2021). To determine if there are any shared genetic risk factors for stroke with EAs, we conducted a fixed effects, inverse-variance weighted meta-analysis incorporating top ($p < 1.00E\text{-}07$) variants ($n = 356$) from the publicly available MEGASTROKE European analysis summary statistics (Malik et al., 2018).

To analyze the putative biological mechanisms of a subset of significant and suggestive variants identified in the meta-analysis with $p < 1.00E-06$, we utilized the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) version 1.3.6a online platform and the GENE2FUNC process (Watanabe et al., 2017). Within FUMA, over 22,000 genes underwent zero-mean normalization and log2 transformation of the expression values [zero mean of log2 (reads per kilobase per million +1)]. Differentially expressed gene (DEG) sets for each of the 53 specific tissue types from the Genotype-Tissue Expression (GTEx) project version eight RNA sequencing data (GTEx Consortium, 2017), were determined from two-sided Student's *t*-test performed per gene per tissue type against all other tissue types. Genes with $p \leq 0.05$ after Bonferroni correction and absolute log fold change of ≥0.58, were included in the DEG set for a given tissue (Watanabe et al., 2017). Furthermore, FUMA distinguished between genes that were upregulated and downregulated in a specific tissue type compared to other tissues *via* the sign of the t-statistic (Watanabe et al., 2017). Genes identified from the meta-analysis were uploaded into FUMA and mapped to Ensembl identifiers. Our prioritized genes were tested against biologically relevant tissue (brain n = 12, artery n = 3, heart n = 2, and kidney n = 2) DEG sets using hypergeometric tests to evaluate if the targeted genes were overrepresented in FUMA DEG sets in specific tissue types. Multiple testing correction was performed using a Benjamini-Hochberg adjustment. Statistical significance was calculated using a p-threshold of $p < 0.05$.

We utilized the publicly available risk score weights for the Abraham metaPRS to assess the predictive accuracy of this score in both GenHAT and REGARDS AAs (Abraham et al., 2019). Specifically, the PRS were generated separately for each study population using the allelic scoring option in PLINK2, resulting in the inclusion of 466,657 and 466,614 variants in GenHAT and REGARDS, respectively. We then employed a series of nine logistic regression models. Model 1 (reference model) regressed the 'any stroke' outcome on age, sex, and PCs 1–10. The metaPRS and clinical risk factors (SBP, DBP, DM, baseline cigarette smoking, or BMI) were individually added to Model 1. A clinical model added each clinical risk factor (SBP + DBP + DM + baseline cigarette smoking + BMI) to model 1. Lastly, we generated a composite model that consisted of Model one plus clinical factors and the metaPRS (SBP + DBP + DM + baseline cigarette smoking + BMI + metaPRS). Using the predicted values for stroke, the performance of each model was determined by the area under the receiver operator characteristic (ROC) curve (AUC). Delong's test for two correlated ROC curves compared the pairwise performance of the nested models (i.e., model one to each of the subsequent models). The fit of each model was determined by Nagelkerke's pseudo-R^2. All predictive regression modeling and model fit calculations were performed in R version 3.6.2.

Replication for Meta-analysis

We sought replication in AAs from the University of Alabama at Birmingham Warfarin Pharmacogenomics Cohort (WPC) (Shendre et al., 2016; Yanik et al., 2017). The WPC replication population included 790 AAs, of which 145 had an incident stroke. In the WPC, incident stroke case-control status was regressed onto genotypes imputed to the TOPMed release 2 (Freeze 8) reference panel, adjusting for age, sex, the top 4 PCs, and genotyping platform using PLINK2. Further details are described in the **Supplementary Material**. We also expanded our replication lookups to the region around our index variants from the meta-analysis. Based on prior research, we considered a window within 100 KB of the target variant (Genomes Project et al., 2015).

Further replication was obtained using the publicly available summary statistics from the recent COMPASS meta-analysis study (Keene et al., 2020). In the COMPASS meta-analysis, 22,051 AAs were included, of which 3,734 had a physician-adjudicated stroke. Due to the imputation reference panel differences (TOPMed vs. 1,000 Genomes), we did not have sufficient overlap to replicate on the variant level. Our replication efforts examined all variants located in the gene region [5′ untranslated region (UTR) through the 3′ UTR] of the target variant, or in the case of intergenic variants, the entire region between the two flanking genes. Of note, there is an overlap of 864 participants (66 cases) from our discovery REGARDS population and those included in the published COMPASS meta-analysis of IS, which could inflate the results on a single-variant level.

RESULTS

The baseline characteristics for GenHAT and REGARDS participants are presented in **Table 1**. In the 6,908 GenHAT participants, approximately 5% (n = 366) had an incident stroke. Males comprised approximately 52% of stroke cases and 44% of controls. The average age of cases was nearly 69 years of age, while the controls were younger at 66 years ($p < 0.001$). In GenHAT, stroke cases were more likely current cigarette smokers (32% v. 27%; $p = 0.104$) and diabetic (51% v. 40%; $p < 0.001$) compared to controls. Likewise, stroke cases had worse kidney function as estimated by the mean glomerular filtration rate (eGFR) (78.62 v. 82.95 ml/min/1.73 m^2; $p < 0.001$), higher mean SBP (148.37 v. 146.13 mmHg; $p = 0.008$) and a negligible difference in mean DBP (84.64 v. 84.88 mmHg; $p = 0.667$) compared to controls.

In the 3,809 REGARDS participants, approximately 7% (n = 280) had an incident stroke. Males comprised approximately 36% of incident stroke cases versus 37% of controls. The average age of stroke cases was 67 vs. 64 years for controls ($p < 0.001$). Of stroke cases, 14% were current cigarette smokers compared to 17% of controls ($p = 0.255$), and almost 50% had DM versus 38% of controls ($p < 0.001$). The stroke cases had slightly higher baseline SBP (135.65 mmHg v. 132.03 mmHg; $p = 0.001$), lower baseline DBP (78.42 mmHg v. 78.74 mmHg; $p = 0.603$), and worse kidney function (83.44 ml/min/1.73 m^2 v. 86.47 ml/min/1.73 m^2; $p = 0.079$) compared to controls (**Table 1**).

In **Table 2**, we present 21 top variants ($p < 1.00E-07$) across nine unique genes, of which 10 variants were statistically significant ($p < 5.00E-08$) from the inverse variance-weighted meta-analysis (**Figure 1**). An additional 79 variants were

TABLE 2 | Top variants (p < 1.00E-07) associated with incident stroke from inverse variance-weighted meta-analysis.

rsID	CHR	BP (hg38)	A1/A2	EAF	OR	95% CI	[a]p	[b]Direction	Location	Gene(s)
rs117880209	18	14,996,132	C/T	0.012	2.98	2.06, 4.31	6.45E-09	−−	intergenic	*LINC01443; LOC644669*
rs144162260	6	75,188,146	G/T	0.005	4.32	2.61, 7.15	1.25E-08	−−	intronic	*COL12A1*
rs117791256	18	15,006,517	C/G	0.011	2.95	2.03, 4.28	1.31E-08	++	intergenic	*LINC01443; LOC644669*
rs536017124	12	30,539,571	A/C	0.004	4.79	2.79, 8.22	1.32E-08	++	intergenic	*TMTC1; IPO8*
rs142422295	18	15,015,740	C/T	0.011	2.93	2.02, 4.25	1.51E-08	−−	intergenic	*LINC01443; LOC644669*
rs145341988	18	15,015,569	T/C	0.011	2.93	2.02, 4.25	1.51E-08	++	intergenic	*LINC01443; LOC644669*
rs140550089	18	15,003,860	C/T	0.011	2.93	2.02, 4.25	1.51E-08	−−	intergenic	*LINC01443; LOC644669*
rs192840029	12	30,513,818	A/G	0.004	4.67	2.72, 8.00	2.17E-08	++	intergenic	*TMTC1; IPO8*
rs568505299	12	70,890,293	T/C	0.005	3.95	2.44, 6.41	2.64E-08	++	intronic	*PTPRR*
rs58633304	18	14,988,000	C/A	0.013	2.76	1.93, 3.96	2.79E-08	−−	intergenic	*LINC01443; LOC644669*
rs186234470	15	48,470,292	T/C	0.011	2.77	1.92, 4.00	5.20E-08	++	intronic	*FBN1*
rs117962542	8	50,357,386	A/G	0.003	5.59	3.00, 10.41	5.91E-08	++	intronic	*SNTG1*
rs116671900	5	1,833,726	T/C	0.002	7.60	3.64, 15.87	6.59E-08	++	intergenic	*NDUFS6; LINC02116*
rs145817478	6	75,173,029	G/A	0.004	4.27	2.52, 7.23	6.70E-08	−−	intronic	*COL12A1*
rs74599173	4	30,537,733	A/G	0.004	4.48	2.59, 7.74	8.09E-08	++	intergenic	*LINC02472; PCDH7*
rs77853510	4	30,553,234	A/G	0.004	4.48	2.59, 7.74	8.09E-08	++	intergenic	*LINC02472; PCDH7*
rs118141576	8	50,326,925	A/T	0.003	5.09	2.81, 9.22	8.18E-08	++	intronic	*SNTG1*
rs541454296	12	70,806,203	C/T	0.003	5.52	2.95, 10.31	8.72E-08	−−	intronic	*PTPRR*
rs143089250	6	28,694,541	C/T	0.003	5.79	3.04, 11.04	9.02E-08	−−	intergenic	*LINC00533; LINC01623*
rs117306,905	18	15,004,526	T/C	0.013	2.67	1.86, 3.84	9.50E-08	++	intergenic	*LINC01443; LOC644669*
rs75989184	4	30,534,014	G/C	0.004	4.42	2.56, 7.63	9.93E-08	−−	intergenic	*LINC02472; PCDH7*

[a]*Genome-wide statistical significance after multiple testing correction,* p < *5.00e-08.*
[b]*Direction order: GenHAT, REGARDS.*
Abbreviations: rsID, reference SNP cluster ID; CHR, chromosome number; BP, base position from hg38; A1, effect allele; A2, allele 2; EAF, effect allele frequency; OR, odds ratio; CI, confidence interval.

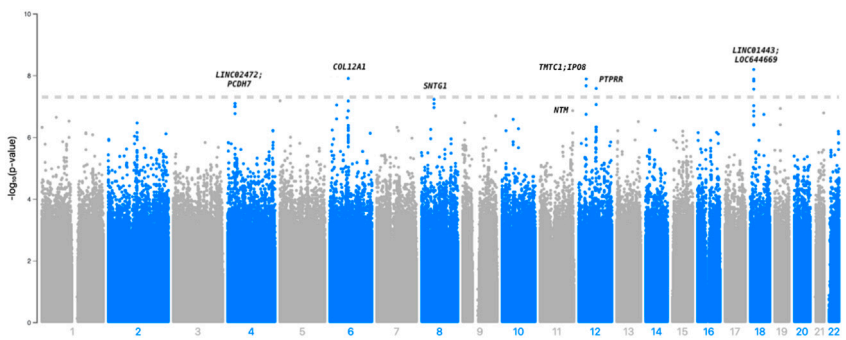

FIGURE 1 | Manhattan plot depicting the top associations with incident stroke in GenHAT-REGARDS meta-analysis. The dotted line is representative of genome-wide significance (p < 5.00E-08).

marginally significant (p < 1.00E-06) and are described in **Supplementary Table S3**. In the sensitivity analysis using SAIGE, we observed consistent results in our top finding, although we saw marginal reduction in the significance (**Supplementary Table S4**). The most significant variant, rs117880209 (p = 6.45E-09), was located in the intergenic region between *LINC01443* and *LOC644669* on chromosome 18. At this variant the odds ratio for incident stroke was 2.98 (95% CI 2.06-4.31) for the C (v. T) allele and the direction of the effect was consistent across both GenHAT and REGARDS. An additional 10 variants in this intergenic region met or exceeded p < 1.00E-06.

Furthermore, we observed 13 intronic variants in *COL12A1* on chromosome 6 with p < 1.00E-6. One intronic variant in this gene, rs144162260, reached statistical significance (p = 1.25E-08; OR = 4.32; 95% CI = 2.61–7.15). Three additional variants exceeded genome-wide significance, including two intergenic variants between *TMTC1* and *IP O 8* (top variant: rs536017124; p = 1.32E-08; OR = 4.79; 95% CI = 2.79–8.22), and one intronic variant of *PTPRR* (rs568505299; p = 2.64E-08, OR = 3.95, 95% CI = 2.44–6.41) (**Table 2**).

Other variants of biological interest include 11 intergenic variants between *LINC02472* and *PCDH7* (top variant: rs74599173, p = 8.09E-08), three intronic variants of *SNTG1* (top variant: rs117962542, p = 5.91E-08), and two intronic variants of *NTM* (top variant: rs185159493, p = 1.37E-07) (**Table 2**; **Supplementary Table S3**). The genomic inflation factor (λ) from the individual cohorts (GenHAT λ = 0.952;

REGARDS λ = 1.002) and the meta-analysis (λ = 0.992) showed no evidence of systematic inflation (**Supplementary Table S5**; **Supplementary Figure S1**). LocusZoom plots of the top variants in *COL12A1, PTPRR, NTM,* and the intergenic regions of *LINC01443- LOC644669*, *TMTC1- IP O 8,* and *LINC02472-PCDH7*, show that the suggestive variants within these regions are in moderate-to-strong linkage disequilibrium (LD) with the sentinel variant (**Supplementary Figures S2–S7**). Variant-specific LD estimates generated from the GenHAT/REGARDS data are shown in **Supplementary Table S6**.

None of our single variant results were replicated in individuals from the WPC at a threshold of p < 5.00E-04 (p = 0.05/100 suggestive variants from **Table 2** and **Supplementary Table S3**). Of the 10 statistically significant variants, rs144162260 in *COL12A1*, rs536017124 and rs192840029 in the *TMTC1-IPO8* intergenic region, and rs58633304 in the *LINC01443-LOC644669* intergenic region had the same direction of effect but non-significant p-values (**Supplementary Table S3**). Our extended look-up of variants ±100 kb of the target variant, identified seven intronic variants within the *NTM* gene that had p-values <5E-04. These variants were all located within 60 kb from either rs185159493 or rs184866696 (**Supplementary Table S7**).

Our lookup efforts in the COMPASS summary statistics provided marginal support for replication at the gene level, as none of the variants from our discovery were identified in COMPASS, most likely due to differences in imputation reference panels. Genes of interest based on the meta-analysis results included *COL12A1, NTM, PTPRR,* and *SNTG1*, as well as the *LINC01443-LOC64466, LINC02472-PCDH7,* and *TMTC1-IPO8* intergenic regions. While we could not replicate our meta-analysis findings on the single-variant level, rs192315401, an upstream variant of the *LINC02472-PCDH7* reached nominal significance in COMPASS (**Supplementary Table S8**, p = 3.20E-04). Results from the meta-analysis of MEGASTROKE top findings (N = 356 variants with p < 1.00E-07) with our AA data are presented in **Supplementary Table S9**. The top findings from MEGASTROKE were not significant in our data and the race-combined meta-analysis did not notably change any MEGASTROKE findings. We specifically focused on variants located within two well-characterized stroke loci: *PITX2* (n = 118) and *HDAC9* (n = 7). For the 118 *PITX2* variants, 102 had a consistent direction of effect across all three cohorts, while the remaining 16 were consistent across MEGASTROKE and REGARDS. In the *HDAC9* gene locus, all seven variants had the same direction of effect across all three cohorts (**Supplementary Table S9**).

We utilized FUMA to identify any tissue specificity of genes represented in our top findings. Of 64 unique gene names identified among variants with p < 1.00E-06, 57 were mapped to Ensembl identifiers by FUMA. Tissue analysis on an *a priori* selected 19 specific tissue types from GTEx revealed statistically significant, differential upregulation in brain tissues, specifically the hippocampus (p_{adj} = 9.67E-05), hypothalamus (p_{adj} = 2.05E-03), amygdala (p_{adj} = 5.28E-03), frontal cortex BA9 (p_{adj} = 6.36E-03), putamen basal ganglia (p_{adj} = 1.56E-02), anterior cingulate cortex BA24 (p_{adj} = 1.62E-02), and the caudate basal ganglia (p_{adj} = 3.17E-02) (**Supplementary Table S10**).

Accuracy in predicting incident stroke was assessed in GenHAT and REGARDS separately, using variant weights for over 400,000 overlapping variants in the UKB metaPRS (Abraham et al., 2019) (**Supplementary Figure S8**). In GenHAT, model 1 (AUC 62.68%; 95% CI 59.89–65.48%) was not statistically different than the model adding the metaPRS (Model 1 + metaPRS; AUC 62.68%; 95% CI 59.89–65.48%; p = 0.981). However, the models adding DM (Model 1 + DM; AUC 64.64%; 95% CI 61.68–67.43%; p = 0.025), all the clinical factors (Model 1 + All Clinical Factors; AUC 66.44%; 95% CI 63.29–69.60%; p = 0.003), and all the clinical factors plus the metaPRS (Model 1 + All Clinical Factors + metaPRS; AUC 66.46%; 95% CI 63.31–69.62; p = 0.003) were significantly different from the reference (i.e., Model one; **Supplementary Table S11**). Less than 0.01% of the variance was attributed to the metaPRS in GenHAT based on the difference of Nagelkerke pseudo-R^2 values between the reference and metaPRS models (**Figure 2**; **Supplementary Table S11**). Similar results were observed in the REGARDS population. The reference model 1 (AUC 60.88%; 95% CI 57.45–64.30%) was not statistically different to the metaPRS model (AUC 60.87%; 95% CI 57.45–64.30%; p = 0.856). The models adding DM (AUC 62.94%; 95% CI 59.57–66.32%; p = 0.035), all the clinical factors (AUC 64.15%; 95% CI 60.76–67.54; p = 0.006), and all the clinical factors plus the metaPRS (AUC 64.16%; 95% CI 60.77–67.54; p = 0.006) were statistically significant, while the model adding SBP (AUC 62.37%; 95% CI 58.94–65.80%; p = 0.071) was marginally significant. Likewise, less than 0.001% of the variance was explained by adding the risk score for REGARDS based on the pseudo-R^2 values between the reference and metaPRS model (**Figure 2**; **SupplementaryTable S12**).

DISCUSSION

While numerous stroke outcome GWAS have been published in the past several years, few studies have been performed exclusively in AAs. Using data from over 10,700 individuals from the GenHAT and REGARDS studies, we identified 10 statistically significant and an additional 90 suggestive genetic variants associated with incident stroke in individuals with HTN. While none of our findings were directly replicated, their gene level associations with stroke warrant future replication efforts, particularly variants located in or near the *NTM* or *PCDH7* genes.

In our meta-analysis, we identified 10 statistically significant variants, including six intergenic variants between a long-intergenic non-protein coding RNA (*LINC01443*) and a pseudogene (*LOC644669*). Four of these six variants were identified in a 2020 intracranial aneurysm (IA) GWAS in an East Asian population (Bakker et al., 2020). Additional significant findings include those variants located in *COL12A1.* This gene encodes the collagen alpha chain of type XII collagen, which interacts with type 1 collagen-containing fibrils (UniProt, 2021), and is a predicted target of the human microRNA-21, which is induced by Angiotensin II (Wang et al., 2017). Angiotensin II is the principal effector hormone of the renin-angiotensin system and increases blood pressure through vasoconstriction and

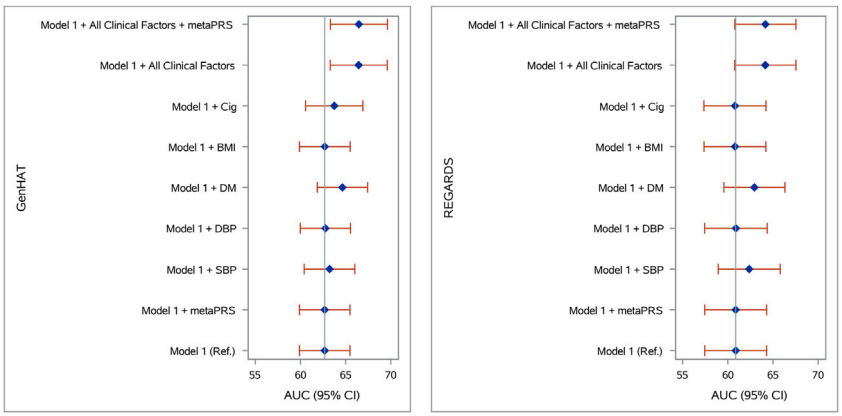

FIGURE 2 | Stroke prediction model comparison in **(A)** GenHAT and **(B)** REGARDS populations. Shown are the area under the receiver operator characteristic curve (AUC) for eight logistic regression models (with and without the Abraham et al. metaPRS and clinical risk factors) used to predict stroke risk in comparison to a basic model 1 (age + sex +10 ancestry PCs). Error bars represent the 95% confidence intervals of the AUC. Abbreviations: PRS, polygenic risk score; Ref, reference model; SBP, systolic blood pressure; DBP, diastolic blood pressure; DM, diabetes mellitus; BMI, body mass index, kg/m^2; Cig, cigarette smoking at baseline.

increased sodium and water retention (de Leeuw, 1999), which is clinically relevant to our hypertensive study base.

Of additional interest were the variants downstream of the *TMTC1* (transmembrane O-Mannosyltransferase Targeting Cadherins 1) gene. *TMTC1* encodes a protein that transfers mannosyl residues to the hydroxyl group of serine or threonine residues and is primarily dedicated to the cadherin superfamily (UniProt, 2021). In a prior meta-analysis, variants in *TMTC1* were associated with heart failure (HF) in African ancestry populations (Smith et al., 2010). Other *TMTC1* variants have been associated with lipid metabolism (Talmud et al., 2009; Della-Morte et al., 2011). A 2011 study using data from the Northern Manhattan Stroke Study identified an interaction of *TMTC1* with abdominal obesity contributing to phenotypic variation in left ventricular mass (LVM). Increased LVM is a known risk factor for HF, stroke, and CVD (Della-Morte et al., 2011).

The only variants with gene-based replication were located in the first intron of *NTM* and upstream of *PCDH7*. *NTM* encodes a member of the IgLON glycosylphosphatidylinositol-anchored cell adhesion molecular family (Gene, 2004) and is primarily expressed in the heart and lungs (GTEx Consortium, 2015). A 2012 study found a balanced translocation break in *NTM* in a family with intracranial and thoracic aortic aneurysms (Luukkonen et al., 2012). Furthermore, intronic *NTM* variants have been previously implicated in intracerebral hemorrhage and small vessel ischemic stroke in Europeans (Chung et al., 2019). Additional studies have reported associations between *NTM* and CVD risk factors such as triglyceride levels (Li et al., 2015) and elevated protein levels in plasma serum (Cao et al., 2015). A recent study from the International Consortium for Antihypertensive Pharmacogenomics Studies concluded that variation in the *NTM* gene is associated with an increased risk of adverse cardiovascular outcomes in patients treated with beta-blockers, as well as an increase in blood pressure after beta-blocker treatment (McDonough et al., 2021). Although not statistically significant, we identified 11 variants located upstream of the *PCDH7* (protocadherin 7) gene. In previous studies, *PCDH7* was differentially expressed in aneurysm wall tissue compared to superficial temporal artery tissue (Shi et al., 2009), as well as being associated with white matter hyperintensities in EAs with ischemic stroke (Traylor et al., 2016).

Of the three marginally associated intronic variants within the *SNTG1* (syntrophin gamma-1) gene, rs117962542 has been previously implicated with stroke risk in a sample of ~70,000 individuals of European descent from MEGASTROKE (Malik et al., 2018). *SNTG1* encodes a protein that mediates gamma-enolase trafficking to the plasma membrane and enhances its neurotrophic activity (UniProt, 2021). An epistasis analysis performed in 2,800 EAs found an association between an *SNTG1* variant and a history of arterial HTN (Zhou et al., 2020). Tissue specificity of our top identified genes, specifically *SNTG1* and *NTM*, showed differential expression in specific brain tissues, compared to other available tissues from GTEx RNA sequence data.

Complex diseases, such as stroke, have shown additive genetic architecture in previous association studies, making PRS a widely lucrative approach. PRS have been utilized to estimate an individual's lifetime genetic risk of disease. While the current discriminative ability is low in the overall population, PRS may be useful in populations where there is a higher probability of disease to assist in prevention or diagnosis, or to inform treatment choices (Lewis and Vassos, 2020). Currently, there are several pitfalls of PRS implementation (Arnold and Koenig, 2021). One of the most impactful is the shortage of data describing PRS performance in African populations, especially since differences in allele frequencies and/or LD may limit cross-ethnic utility (Martin et al., 2019). Our data show that the application of the genome-wide Abraham metaPRS (derived and validated in >400K EAs) to >10,000 AAs does not aid in stroke prediction beyond age, sex, and genetic ancestry (reference model). We chose to test the Abraham metaPRS compared to other previously

published PRS due to the metaPRS being similarly predictive to several stroke risk factors, including family history, BP, BMI, and smoking, as well as reports that the metaPRS (HR 1.26; 95% CI 1.22–1.31 per SD) doubled the predictive accuracy of ischemic stroke compared to the 90-variant score (HR 1.13, 95% CI 1.10–1.17) (Rutten-Jacobs et al., 2018; Abraham et al., 2019) in EAs. In both GenHAT and REGARDS, the models accounting for clinical risk factors (SBP, DBP, DM, BMI, and cigarette smoking) had the highest predictive accuracy and a negligible improvement was observed when adding the metaPRS. This suggests that Abraham score utility is not transferable to AA adults. Therefore, there remains a strong need for the generation and validation of stroke scores in minority populations, specifically AAs.

This study has several strengths. Both GenHAT and REGARDS collected adjudicated stroke data on a large sample of AAs with HTN who are disproportionally at-risk for stroke. We also utilized contemporary genotyping and imputation methods designed to be more inclusive for research in minority populations, allowing for more accurate genetic interrogation of our population (e.g., linkage patterns). We also must note some weaknesses. We were unable to focus on stroke type or sub-type due to a lack of that data in ALLHAT/GenHAT. Additionally, because of our inclusion/exclusion criteria focused on HTN, our findings may not be generalizable to younger, healthier populations, or individuals taking other antihypertensives (i.e., not a thiazide diuretic or ACE inhibitor). With our study base we may not find HTN genes related to stroke. However, this is an issue in older AA population studies of stroke in general as the prevalence of HTN is high (e.g., ~70% in the parent REGARDS study). This could also be reflected in the FUMA results, where there was limited differential expression in vascular, non-brain tissues (e.g., artery, heart). Additionally, the GTEx data used in the FUMA analysis is comprises only 12.9% AA samples. Also, the lack of single-variant replication in the WPC was limited by the lower allele frequency variants from the discovery being underrepresented in the smaller sample size of the WPC. Finally, while the relatively recent and more racially inclusive TOPMed imputation panel allowed for interrogation of millions of novel variants, many of which were of lower allele frequency, it limited our replication efforts in published data that was imputed into earlier reference datasets, which highlights the necessity for additional stroke datasets in AAs.

In conclusion, we identified 10 statistically significant variants associated with incident stroke in AAs with HTN. The individual variants were not independently validated in the WPC or previously reported. However, many gene regions were biologically plausible and we found gene-based validation in previously published data. When accounting for case-control imbalance using SAIGE, our top findings remained significant and the same direction of the effect size was observed. This highlights the need for additional validation in large, stroke studies with AA populations and contemporary methods (e.g., use of whole genome sequence data and/or TOPMed imputed data) to capture genetic variation in non-European populations. The majority of the results identified genes (*NTM, PCDH7, COL12A1*) related to stroke or other CVD related diseases in Asian and European ancestral populations, with the exception of one region near *TMTC1*. This suggests both ancestry-common and ancestry-specific stroke risk genes are present in AAs. As hypothesized, this will necessitate discovery efforts to be more inclusive as genetic diagnostics trained on GWAS data are considered for use in the clinic. To that point, the published Abraham composite metaPRS trained in a large sample of EAs was not validated in the REGARDS or GenHAT study AAs. This research highlights the need to collect additional data through large biobanks and consortia efforts that can alleviate the potential for genetic discovery disparities.

DATA AVAILABILITY STATEMENT

The raw GenHAT genotypic and phenotypic data used in this study are deposited in the National Center for Biotechnology Information (NCBI) Database for Genotypes and Phenotypes (dbGaP), accession number phs002716.v1.p1. The raw REGARDS genotype and phenotype data used in this study can be found in dbGaP, accession number phs002719.v1.p1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of Alabama at Birmingham Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NA, BH, HT, DA, and MI contributed to the conception and design of the study. NA, RT, BH, HT, NL, EL, LL, DA, and MI contributed to the interpretation of results. MI, LL, and NL provided funding for the study. NA, VS, and AP performed the statistical analysis. NA wrote the first draft of the manuscript. All authors contributed to the manuscript revisions, read, and approved the final, submitted version.

FUNDING

The study was supported by the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute (NHLBI) grants R01HL123782 (MI) and R01HL136666 (MI, LL). NA was supported by an NIH NHLBI T32 Fellowship (T32HL007457). The Warfarin Pharmacogenomics Cohort was supported by NHLBI grant R01HL092173 (NL). The REGARDS study is supported by a cooperative agreement U01 NS041588 from the National Institute of Neurological Disorders and Stroke, National Institutes of Health, U.S. Department of Health and Human Services. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the

National Institutes of Health. Representatives of the funding agency have been involved in the review of the manuscript but not directly involved in the collection, management, analysis, or interpretation of the data.

ACKNOWLEDGMENTS

The authors thank the other investigators, the staff, and the participants of the REGARDS study for their valuable contributions. A full list of participating REGARDS investigators and institutions can be found at http://www.regardsstudy.org. The MEGASTROKE project received funding from sources specified at http://www.megastroke.org/acknowledgments.html.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.781451/full#supplementary-material

REFERENCES

Abraham, G., Malik, R., Yonova-Doing, E., Salim, A., Wang, T., and Danesh, J. (2019). Genomic Risk Score Offers Predictive Performance Comparable to Clinical Risk Factors for Ischaemic Stroke. *Nat. Commun.* 10(1), 5819. doi:10.1038/s41467-019-13848-1

ALLHAT Protocol (2000). Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial. (ALLHAT) Protocol [Online]. Available: https://ccct.sph.uth.tmc.edu/ALLHAT/Documents/Protocol.pdf.

Arnold, N., and Koenig, W. (2021). Polygenic Risk Score: Clinically Useful Tool for Prediction of Cardiovascular Disease and Benefit from Lipid-Lowering Therapy? *Cardiovasc. Drugs Ther.* 35 (3), 627–635. doi:10.1007/s10557-020-07105-7

Bakker, M. K., van der Spek, R. A. A., van Rheenen, W., Morel, S., Bourcier, R., Hostettler, I. C., et al. (2020). Genome-wide Association Study of Intracranial Aneurysms Identifies 17 Risk Loci and Genetic Overlap with Clinical Risk Factors. *Nat. Genet.* 52 (12), 1303–1313. doi:10.1038/s41588-020-00725-7

Bentley, A. R., Callier, S. L., and Rotimi, C. N. (2020). Evaluating the Promise of Inclusion of African Ancestry Populations in Genomics. *NPJ Genom Med.* 5, 5. doi:10.1038/s41525-019-0111-x

Bevan, S., Traylor, M., Adib-Samii, P., Malik, R., Paul, N. L., Jackson, C., et al. (2012). Genetic Heritability of Ischemic Stroke and the Contribution of Previously Reported Candidate Gene and Genomewide Associations. *Stroke* 43 (12), 3161–3167. doi:10.1161/STROKEAHA.112.665760

Boughton, A. P., Welch, R. P., Flickinger, M., VandeHaar, P., Taliun, D., Abecasis, G. R., et al. (2021). LocusZoom.js: Interactive and Embeddable Visualization of Genetic Association Study Results. *Bioinformatics*. doi:10.1093/bioinformatics/btab186

Cao, T. H., Quinn, P. A., Sandhu, J. K., Voors, A. A., Lang, C. C., Parry, H. M., et al. (2015). Identification of Novel Biomarkers in Plasma for Prediction of Treatment Response in Patients with Heart Failure. *Lancet* 385 (Suppl. 1), S26. doi:10.1016/S0140-6736(15)60341-5

Carty, C. L., Keene, K. L., Cheng, Y. C., Meschia, J. F., Chen, W. M., Nalls, M., et al. (2015). Meta-Analysis of Genome-wide Association Studies Identifies Genetic Risk Factors for Stroke in African Americans. *Stroke* 46 (8), 2063–2068. doi:10.1161/STROKEAHA.115.009044

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of Larger and Richer Datasets. *Gigascience* 4, 7. doi:10.1186/s13742-015-0047-8

Chung, J., Marini, S., Pera, J., Norrving, B., Jimenez-Conde, J., Roquer, J., et al. (2019). Genome-wide Association Study of Cerebral Small Vessel Disease Reveals Established and Novel Loci. *Brain* 142 (10), 3176–3189. doi:10.1093/brain/awz233

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation Genotype Imputation Service and Methods. *Nat. Genet.* 48 (10), 1284–1287. doi:10.1038/ng.3656

de Leeuw, P. W. (1999). How Do Angiotensin II Receptor Antagonists Affect Blood Pressure? *Am. J. Cardiol.* 84 (2A), 5K–6K. doi:10.1016/s0002-9149(99)00399-9

Della-Morte, D., Beecham, A., Rundek, T., Wang, L., McClendon, M. S., Slifer, S., et al. (2011). A Follow-Up Study for Left Ventricular Mass on Chromosome 12p11 Identifies Potential Candidate Genes. *BMC Med. Genet.* 12, 100. doi:10.1186/1471-2350-12-100

Flossmann, E., Schulz, U. G., and Rothwell, P. M. (2004). Systematic Review of Methods and Results of Studies of the Genetic Epidemiology of Ischemic Stroke. *StrokeAA* 35 (1), 212–227. doi:10.1161/01.STR.0000107187.8439010.1161/01.STR.0000107187.84390.AA

Fuchsberger, C., Abecasis, G. R., and Hinds, D. A. (2015). minimac2: Faster Genotype Imputation. *Bioinformatics* 31 (5), 782–784. doi:10.1093/bioinformatics/btu704

Gene (2004). National Library of Medicine, National Center for Biotechnology Information. Available at: https://www.ncbi.nlm.nih.gov/gene/.

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A Global Reference for Human Genetic Variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Gretarsdottir, S., Thorleifsson, G., Manolescu, A., Styrkarsdottir, U., Helgadottir, A., Gschwendtner, A., et al. (2008). Risk Variants for Atrial Fibrillation on Chromosome 4q25 Associate with Ischemic Stroke. *Ann. Neurol.* 64 (4), 402–409. doi:10.1002/ana.21480

GTEx Consortium (2017). Genetic Effects on Gene Expression across Human Tissues, Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, Enhancing, G.g., Fund, N.I.H.C. *Nature* 550 (7675), 204–213. doi:10.1038/nature24277

GTEx Consortium (2015). Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science* 348 (6235), 648–660. doi:10.1126/science.1262110

Holliday, E. G., Maguire, J. M., Evans, T. J., Koblar, S. A., Jannes, J., Sturm, J. W., et al. (2012). Common Variants at 6p21.1 Are Associated with Large Artery Atherosclerotic Stroke. *Nat. Genet.* 44 (10), 1147–1151. doi:10.1038/ng.2397

Howard, V. J., Kleindorfer, D. O., Judd, S. E., McClure, L. A., Safford, M. M., Rhodes, J. D., et al. (2011). Disparities in Stroke Incidence Contributing to Disparities in Stroke Mortality. *Ann. Neurol.* 69 (4), 619–627. doi:10.1002/ana.22385

Ikram, M. A., Seshadri, S., Bis, J. C., Fornage, M., DeStefano, A. L., Aulchenko, Y. S., et al. (2009). Genomewide Association Studies of Stroke. *N. Engl. J. Med.* 360 (17), 1718–1728. doi:10.1056/NEJMoa0900094

International Stroke Genetics (2012). Genome-wide Association Study Identifies a Variant in HDAC9 Associated with Large Vessel Ischemic Stroke, Wellcome Trust Case Control, Bellenguez, C., Bevan, S., Gschwendtner, A., Spencer, C.C. *Nat. Genet.* 44 (3), 328–333. doi:10.1038/ng.1081

Katan, M., and Luft, A. (2018). Global Burden of Stroke. *Semin. Neurol.* 38 (2), 208–211. doi:10.1055/s-0038-1649503

Keene, K. L., Hyacinth, H. I., Bis, J. C., Kittner, S. J., Mitchell, B. D., Cheng, Y. C., et al. (2020). Genome-Wide Association Study Meta-Analysis of Stroke in 22 000 Individuals of African Descent Identifies Novel Associations with Stroke. *Stroke* 51 (8), 2454–2463. doi:10.1161/STROKEAHA.120.029123

Kilarski, L. L., Achterberg, S., Devan, W. J., Traylor, M., Malik, R., Lindgren, A., et al. (2014). Meta-analysis in More Than 17,900 Cases of Ischemic Stroke Reveals a Novel Association at 12q24.12. *Neurology* 83 (8), 678–685. doi:10.1212/WNL.0000000000000707

Lewis, C. M., and Vassos, E. (2020). Polygenic Risk Scores: from Research Tools to Clinical Instruments. *Genome Med.* 12 (1), 44. doi:10.1186/s13073-020-00742-5

Li, C., Bazzano, L. A., Rao, D. C., Hixson, J. E., He, J., Gu, D., et al. (2015). Genome-wide Linkage and Positional Association Analyses Identify Associations of Novel AFF3 and NTM Genes with Triglycerides: the GenSalt Study. *J. Genet. Genomics* 42 (3), 107–117. doi:10.1016/j.jgg.2015.02.003

Luukkonen, T. M., Poyhonen, M., Palotie, A., Ellonen, P., Lagstrom, S., Lee, J. H., et al. (2012). A Balanced Translocation Truncates Neurotrimin in a Family with Intracranial and Thoracic Aortic Aneurysm. *J. Med. Genet.* 49 (10), 621–629. doi:10.1136/jmedgenet-2012-100977

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., and Go, T. D. i. (2013). Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. *Genet. Epidemiol.* 37 (6), 539–550. doi:10.1002/gepi.21742

Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., et al. (2018). Multiancestry Genome-wide Association Study of 520,000 Subjects Identifies 32 Loci Associated with Stroke and Stroke Subtypes. *Nat. Genet.* 50 (4), 524–537. doi:10.1038/s41588-018-0058-3

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., et al. (2021). Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits. *Nat. Genet.* 53 (7), 1097–1103. doi:10.1038/s41588-021-00870-7

McDonough, C. W. W., Jack, J. R., Motsinger-Reif, A. A., Armstrong, N. D., Bis, J. C., House, J. S., et al. (2021). Adverse Cardiovascular Outcomes and Antihypertensive Treatment: A Genome-wide Interaction Meta-Analysis in the International Consortium for Antihypertensive Pharmacogenomics Studies (ICAPS). *Clin. Pharmacol. Ther.* 110 (3), 723–732. doi:10.1002/cpt.2355

Peprah, E., Xu, H., Tekola-Ayele, F., and Royal, C. D. (2015). Genome-wide Association Studies in Africans and African Americans: Expanding the Framework of the Genomics of Human Traits and Disease. *Public Health Genomics* 18 (1), 40–51. doi:10.1159/000367962

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies. *Nat. Genet.* 38 (8), 904–909. doi:10.1038/ng1847

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. (2010). LocusZoom: Regional Visualization of Genome-wide Association Scan Results. *Bioinformatics* 26 (18), 2336–2337. doi:10.1093/bioinformatics/btq419

Rutten-Jacobs, L. C., Larsson, S. C., Malik, R., and Rannikmae, K.; MEGASTROKE consortium, International Stroke Genetics Consortium (2018). International Stroke GeneticsGenetic Risk, Incident Stroke, and the Benefits of Adhering to a Healthy Lifestyle: Cohort Study of 306 473 UK Biobank Participants. *BMJ* 363, k4168. doi:10.1136/bmj.k4168

Sarnowski, C., Chen, H., Biggs, M. L., Wassertheil-Smoller, S., Bressler, J., Irvin, M. R., et al. (2021). Identification of Novel and Rare Variants Associated with Handgrip Strength Using Whole Genome Sequence Data from the NHLBI Trans-omics in Precision Medicine (TOPMed) Program. *PLoS One* 16 (7), e0253611. doi:10.1371/journal.pone.0253611

Shendre, A., Brown, T. M., Liu, N., Hill, C. E., Beasley, T. M., Nickerson, D. A., et al. (2016). Race-Specific Influence of CYP4F2 on Dose and Risk of Hemorrhage Among Warfarin Users. *Pharmacotherapy* 36 (3), 263–272. doi:10.1002/phar.1717

Shi, C., Awad, I. A., Jafari, N., Lin, S., Du, P., Hage, Z. A., et al. (2009). Genomics of Human Intracranial Aneurysm wall. *Stroke* 40 (4), 1252–1261. doi:10.1161/STROKEAHA.108.532036

Smith, N. L., Felix, J. F., Morrison, A. C., Demissie, S., Glazer, N. L., Loehr, L. R., et al. (2010). Association of Genome-wide Variation with the Risk of Incident Heart Failure in Adults of European and African Ancestry: a Prospective Meta-Analysis from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. *Circ. Cardiovasc. Genet.* 3 (3), 256–266. doi:10.1161/CIRCGENETICS.109.895763

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program. *Nature* 590 (7845), 290–299. doi:10.1038/s41586-021-03205-y

Talmud, P. J., Drenos, F., Shah, S., Shah, T., Palmen, J., Verzilli, C., et al. (2009). Gene-centric Association Signals for Lipids and Apolipoproteins Identified via the HumanCVD BeadChip. *Am. J. Hum. Genet.* 85 (5), 628–642. doi:10.1016/j.ajhg.2009.10.014

Traylor, M., Farrall, M., Holliday, E. G., Sudlow, C., Hopewell, J. C., Cheng, Y. C., et al. (2012). Genetic Risk Factors for Ischaemic Stroke and its Subtypes (The METASTROKE Collaboration): a Meta-Analysis of Genome-wide Association Studies. *Lancet Neurol.* 11 (11), 951–962. doi:10.1016/S1474-4422(12)70234-X

Traylor, M., Rutten-Jacobs, L., Curtis, C., Patel, H., Breen, G., Newhouse, S., et al. (2017). Genetics of Stroke in a UK African Ancestry Case-Control Study: South London Ethnicity and Stroke Study. *Neurol. Genet.* 3 (2), e142. doi:10.1212/NXG.0000000000000142

Traylor, M., Zhang, C. R., Adib-Samii, P., Devan, W. J., Parsons, O. E., Lanfranconi, S., et al. (2016). Genome-wide Meta-Analysis of Cerebral white Matter Hyperintensities in Patients with Stroke. *Neurology* 86 (2), 146–153. doi:10.1212/WNL.0000000000002263

UniProt, C. (2021). UniProt: the Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi:10.1093/nar/gkaa1100

Wang, G., Wu, L., Chen, Z., and Sun, J. (2017). Identification of Crucial miRNAs and the Targets in Renal Cortex of Hypertensive Patients by Expression Profiles. *Ren. Fail.* 39 (1), 92–99. doi:10.1080/0886022X.2016.1244083

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* 38 (16), e164. doi:10.1093/nar/gkq603

Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional Mapping and Annotation of Genetic Associations with FUMA. *Nat. Commun.* 8 (1), 1826. doi:10.1038/s41467-017-01261-5

WHO Task Force on Stroke and other Cerebrovascular Disorders (1989). Stroke--1989. Recommendations on Stroke Prevention, Diagnosis, and Therapy. Report from the WHO Task Force on Stroke and Other Cerebrovascular Disorders. *Stroke* 20 (10), 1407–1431. doi:10.1161/01.str.20.10.1407

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans. *Bioinformatics* 26 (17), 2190–2191. doi:10.1093/bioinformatics/btq340

Yanik, M. V., Irvin, M. R., Beasley, T. M., Jacobson, P. A., Julian, B. A., and Limdi, N. A. (2017). Influence of Kidney Transplant Status on Warfarin Dose, Anticoagulation Control, and Risk of Hemorrhage. *Pharmacotherapy* 37 (11), 1366–1373. doi:10.1002/phar.2032

Zhou, J., Passero, K., Palmiero, N. E., Muller-Myhsok, B., Kleber, M. E., Maerz, W., et al. (2020). Investigation of Gene-Gene Interactions in Cardiac Traits and Serum Fatty Acid Levels in the LURIC Health Study. *PLoS One* 15 (9), e0238304. doi:10.1371/journal.pone.0238304

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently Controlling for Case-Control Imbalance and Sample Relatedness in Large-Scale Genetic Association Studies. *Nat. Genet.* 50 (9), 1335–1341. doi:10.1038/s41588-018-0184-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fgene.2021.723477

Identification of Five Cytotoxicity-Related Genes Involved in the Progression of Triple-Negative Breast Cancer

Yan Zhang[1,2,3], Gui-hui Tong[4], Xu-Xuan Wei[2], Hai-yang Chen[2], Tian Liang[2], Hong-Ping Tang[5], Chuan-An Wu[6], Guo-Ming Wen[6], Wei-Kang Yang[6], Li Liang[1,7]* and Hong Shen[1]**

[1]Department of Pathology, School of Basic Medical Sciences, Southern Medical University/Nanfang Hospital, Southern Medical University, Guangzhou, China, [2]Department of Pathology, The First Affiliated Hospital of Guangdong University Of Pharmacy, Guangzhou, China, [3]Department of Pathology, Shenzhen Longhua District Maternity & Child Healthcare Hospital, Shenzhen, China, [4]Department of Pathology,The first Affiliated Hospital,Guangzhou Medical University, Guangzhou, China, [5]Department of Pathology, Shenzhen Maternity & Child Healthcare Hospital, Shenzhen, China, [6]Department of Prevention and Health Care, Shenzhen Longhua District Maternity & Child Healthcare Hospital, Shenzhen, China, [7]Guangdong Province Key Laboratory of Molecular Tumor Pathology, Guangzhou, China

OPEN ACCESS

Edited by:
Zodwa Dlamini,
SAMRC Precision Oncology Research Unit (PORU), South Africa

Reviewed by:
Ryan Spengler,
University of Wisconsin-Madison, United States
Shuna Cui,
Yangzhou University, China

***Correspondence:**
Wei-Kang Yang
9043948@qq.com
Li Liang
lli@smu.edu.cn
Hong Shen
shenhong2010168@163.com

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: 11 June 2021
Accepted: 29 October 2021
Published: 03 January 2022

Citation:
Zhang Y, Tong G-h, Wei X-X, Chen H-y, Liang T, Tang H-P, Wu C-A, Wen G-M, Yang W-K, Liang L and Shen H (2022) Identification of Five Cytotoxicity-Related Genes Involved in the Progression of Triple-Negative Breast Cancer. Front. Genet. 12:723477. doi: 10.3389/fgene.2021.723477

Background: Breast cancer is one of the deadly tumors in women, and its incidence continues to increase. This study aimed to identify novel therapeutic molecules using RNA sequencing (RNA-seq) data of breast cancer from our hospital.

Methods: 30 pairs of human breast cancer tissue and matched normal tissue were collected and RNA sequenced in our hospital. Differentially expressed genes (DEGs) were calculated with raw data by the R package "edgeR", and functionally annotated using R package "clusterProfiler". Tumor-infiltrating immune cells (TIICs) were estimated using a website tool TIMER 2.0. Effects of key genes on therapeutic efficacy were analyzed using RNA-seq data and drug sensitivity data from two databases: the Cancer Cell Line Encyclopedia (CCLE) and the Cancer Therapeutics Response Portal (CTRP).

Results: There were 2,953 DEGs between cancerous and matched normal tissue, as well as 975 DEGs between primary breast cancer and metastatic breast cancer. These genes were primarily enriched in PI3K-Akt signaling pathway, calcium signaling pathway, cAMP signaling pathway, and cell cycle. Notably, $CD8^+$ T cell, M0 macrophage, M1 macrophage, regulatory T cell and follicular helper T cell were significantly elevated in cancerous tissue as compared with matched normal tissue. Eventually, we found five genes (*GALNTL5, MLIP, HMCN2, LRRN4CL,* and *DUOX2*) were markedly corelated with $CD8^+$ T cell infiltration and cytotoxicity, and associated with therapeutic response.

Conclusion: We found five key genes associated with tumor progression, $CD8^+$ T cell and therapeutic efficacy. The findings would provide potential molecular targets for the treatment of breast cancer.

Keywords: DUOX2, $CD8^+$ T cell, GALNTL5, breast cancer, therapeutic efficacy

INTRODUCTION

Breast cancer is a deadliest type of female carcinoma (Xu et al., 2021), the incidence of breast cancer is increasing in the past few years (Hu et al., 2021; Su et al., 2021; Tagliamento et al., 2021; Xu et al., 2021). Although the treatment of breast cancer has achieved great progress (Li et al., 2021; Mouabbi et al., 2021), 5-years survival rate for advanced breast cancer is still poor (Kang et al., 2018; Wang et al., 2021a). Accordingly, the molecular mechanisms of breast cancer occurrence and progression are still largely unclear.

Triple-negative breast cancer (TNBC) is an aggressive and heterogeneous subtype of breast cancer. The characteristics of TNBC on immunohistochemical examination are estrogen receptor negative (ER-), progesterone receptor negative (PR-) and human epidermal growth factor receptor 2 negative (HER2-) (Dent et al., 2007; Aysola et al., 2013; Sporikova et al., 2018; Garrido-Castro et al., 2019; Yin et al., 2020). The mortality and recurrence rate of TNBC is higher than other types of breast cancer; especially in the first 5 years after diagnosis, the mortality and recurrence rate are significantly higher than other types of breast cancer (Dent et al., 2007). Primary or secondary resistance to the treatment restrained the present therapeutic strategy. Facing this grim situation, it is very urgent to identify new molecular targets to treat resistant TNBC.

Meanwhile, there is also a growing recognition of the impact of the tumor microenvironment on the fate of tumors. Tumor infiltrating immune cells (TIICs) not only affect the growth of the tumor, but also affect the effect of treatment (Ren et al., 2021; Wan et al., 2021; Zhang et al., 2021). Comprehending the interactions between cancer cells and TIICs is critical for identifying key pathogenic molecules, improving drug sensitivity, and developing new therapeutic strategies.

With the rapid development of sequencing technology, human awareness of disease has entered the genetic molecular level. Analysis of RNA sequencing (RNA-seq) data has revealed potential pathogenic genes and key molecules in various types of disease including cancer (Park, 2021; Sohrabi et al., 2021). The vigorous development of high-throughput sequencing technology and bioinformatics has provided a powerful tool for revealing the underlying molecular mechanism of breast cancer.

Here, we performed RNA sequencing of 30 pairs of tumorous tissue and matched normal tissue of 30 TNBC patients from our hospital, and conducted a combined analysis of RNA-seq data using bioinformatics methods. This study is expected to identify the potential key genes associated with tumor progression, tumor immunity and therapeutic efficacy. The findings would provide potential molecular targets for the treatment of TNBC.

MATERIALS AND METHODS

Sample Acquisition and Pathological Diagnosis

30 pairs of cancerous tissue and matched normal tissue from 30 breast cancer patients, including 15 breast cancer with lymph node metastasis and 15 breast cancer without lymph node metastasis, were collected from our hospital. All participating patients received a standard mastectomy. During the operation, once the breast tissue was removed by the surgeon, the research technicians waiting on the side would sample part of the removed breast tissue, including the tumor and surrounding normal breast tissue. The sampled tissue would be sent for pathological examination by two professional pathologists to determine whether they were cancer or normal breast tissue. Then these diagnosed samples would be used for subsequent RNA sequencing. Written informed consent was obtained from all enrolled patients. This study was approved by the Ethics Committee of our hospital.

RNA-seq data and clinical data from the TCGA cohort of patients with breast cancer were used to investigate the survival value of key genes in the development of breast cancer.

RNA Extraction

Total RNA was isolated and purified using TRIzol (Life, cat.265709, CA, United States) following the manufacturer's procedure. After the quality inspection of Agilent 2,100 Bioanalyzer (Agilent, cat. G2939AA, CA, United States) and NanoPhotometer® (Implen, cat. N60, Munich, Germany), mRNA with poly(A) is purified from 1 μg total RNA using VAHTS® mRNA Capture Beads with Oligo (dT) (Vazyme, cat. N401-01, Nanjing, China) through two rounds of purification.

Library Generation and RNA Sequencing

Subsequently, mRNA fragment was interrupted using VAHTS® Universal V6 RNA-seq Library Prep Kit (Vazyme, cat. NR604, Nanjing, China) under 94°C 8 min and reversed transcription into cDNA which would use to synthesise U-labeled second-stranded DNAs. An A-base was added to the blunt ends of each strand to ligase the indexed adapters which contains a T-base at the tail end. After UDG enzyme treatment of the U-labeled double-strand DNA, size selection was performed with VAHTS® DNA Clean Beads (Vazyme, cat. N411, Nanjing, China).

Then the ligated products are amplified with PCR by the following conditions: initial denaturation at 98°C for 5 min; 12–17 cycles of denaturation at 98°C for 10 s, annealing at 60°C for 30 s, and extension at 72°C for 30 s; final extension at 72°C for 5 min. The average insert size of cDNA library was 280 ± 80 bp. After purification by VAHTS® DNA Clean Beads (Vazyme, Quality control and normalization of sequencing data cat. N411-02, Nanjing, China), quality control of concentration and fragment size is performed by Agilent 2,100 Bioanalyzer (Agilent, cat. G2939AA, CA, United States) and Qubit assay tubes (Life, cat. 1604220, CA, United States).

At last, we performed the 2 × 150 bp paired-end sequencing (PE150) on an Illumina Novaseq™ 6,000 system (Illumina Corporation, San Diego, United States) following the vendor's recommended protocol by Guangzhou Huayin Health Medical Group CO.,Ltd. (Guangzhou, China).

Data Quality Control and Genome Alignment

Raw reads were trimmed adapters using Cutadapt (https://cutadapt.readthedocs.io/en/stable/, v1.16) and a self-made program for removing contamination and low-quality reads which bases with a quality score lower than Q20 exceeds 50%, respectively. rRNA contamination was filtered by Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, version: bowtie2, v2.3.3.1) (Langmead and Salzberg, 2012).

Clean reads were mapped to the genome GRCh37 using TopHat (http://ccb.jhu.edu/software/tophat/index.shtml,v2.1.1) (Trapnell et al., 2009). The parameters used were all default.

Quantitation

The expression level of mRNA were calculated using RSEM (RNA-Seq by Expectation Maximization) (v1.3.1) by normalized to FPKM (Fragments Per kilobase Per Million reads) (Li and Dewey, 2011). FPKM data was further converted into TPM data for estimation of the abundance of tumor-infiltrating immune cells and for correlation analysis between interested genes and interested immune cells.

Differential Expression Analysis

Differentially expressed genes (DEGs) were calculated with original read counts using R package "edgeR" (McCarthy et al., 2012). Note that "edgeR" is designed to work with actual read counts. Normalized data, including FPKM, RPKM and TMP, is not recommended to be used in place of actual counts in edgeR.

DEGs between 30 cancerous tissue and 30 matched normal tissue were calculated using a paired design, which can be achieved by formula of "design = model. matrix (~patient + group)" in R. DEGs between 15 TNBC tissue with metastasis and 15 TNBC tissue without metastasis were computed using the regular analysis process.

R package "edgeR" implements novel statistical methods based on the negative binomial distribution as a model for count variability, including empirical Bayes methods, exact tests, and generalized linear models, and quasi-likelihood tests to calculate DEGs.

Prior to further analysis, "edgeR" provides a procedure to filter out low expressed genes and normalize the raw data into TMM data which is subsequently used to calculate DEGs. The R function of filterByExpr is used to filter out low expressed genes. The filterByExpr function keeps rows that have worthwhile counts in a minimum number of samples.

Selection criteria for DEGs were as follows: $|\log FC| > 1$ and FDR <0.05. Benjamini-Hochberg method was used to adjust *p*-values.

Functional Annotation

Functional annotation of DEGs was conducted using R package "clusterProfiler" (version: 3.18.1) (Yu et al., 2012), which provides a comprehensive set of functional annotation tools for researchers to comprehend the biological meaning behind specific gene sets. The clusterProfiler package depends on the Bioconductor annotation data GO. db and KEGG. db to obtain the maps of the entire GO and KEGG corpus. Bioconductor annotation packages org. Hs.eg.db, org. Mm.eg.db, and org. Sc.sgd.db were imported for genome-wide annotation of mapping Entrez gene identifiers or ORF identifiers for humans, mice, and yeast, respectively. Functional annotation consists of gene oncology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, which allows one can investigate what biological functions and signaling pathways a given gene set is involved in. We also conducted gene set enrichment analysis (GSEA) based on a ranked gene set using "clusterProfiler". GSEA could reveal some enriched signaling pathways missed in GO analysis. Key parameters were as follows: pAdjustMethod = "BH", pvalueCutoff = 0.05, qvalueCutoff = 0.2, nPerm = 1,000, minGSSize = 10, maxGSSize = 500.

Estimation of Tumor Infiltrating Immune Cells

Tumor-infiltrating immune cells (TIICs) were estimated using a website tool TIMER 2.0 (Newman et al., 2015; Li et al., 2020). TIMER 2.0 provides multiple computational methods based on deconvolution to characterize immune cell composition of complex tissues from their gene expression profiles. TIMER 2.0 should enable large-scale analysis of bulk RNA-seq data for cellular biomarkers and therapeutic targets. The accuracy of TIMER has been demonstrated by immunohistochemistry and flow cytometry.

Quantification of Immune Cell Cytotoxicity

Immune cell cytotoxicity of each sample was quantified based on expression levels of *CD8A, CD8B, GZMA, GZMB* and *PRF1*, using single-sample GSEA (ssGSEA). ssGSEA is an extension of GSEA and can be used to calculate separate enrichment scores for each pairing of a sample and gene set. Each ssGSEA enrichment score represents the degree to which the genes in a particular gene set are coordinately up- or down-regulated within a sample. Key parameters were as follows: kcdf = "Gaussian", min. sz = 1, max. sz = Inf, tau = 0.25, abs. ranking = TRUE.

Correlation Between Hub Genes and Drug Sensitivity

Data on breast cancer cell lines were obtained from two large-scale cancer profiling studies: the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012; Ghandi et al., 2019), which profiles gene expression in cancer cells, and the Cancer Therapeutics Response Portal (CTRP) (Seashore-Ludlow et al., 2015), which characterizes the response of cancer cell lines to a collection of drugs. We categorized breast cancer cell lines into low-and high-expression groups based on the median of each hug gene expression levels, and compared sensitivity to therapeutic drugs in the high-versus low-group using *t*-test. The IC50 value of each drug was used as a measure of drug response, which was available in CTRP database.

TABLE 1 | Clinical characteristics of 30 triple-negative breast cancer patients.

Patient ID	Gender	Age at diagnosis	Tumor size (cm)	Pathological grade	Lymph node	Relapse
102548	F	NA	2*1	II	yes	no
104338	F	NA	NA	II	yes	no
105094	F	NA	1*3	NA	yes	no
109745	F	NA	2.5*2.5*2	III	no	no
1906415	F	49	1.3*1.7	II	no	yes
1912627	F	65	3.7*2.7*2.3	II	yes	no
1924346	F	46	2*1.3	III	no	no
1926760	F	37	4.8*2.1	III	no	no
1927842	F	36	3.6*1.3*1.5	III	yes	yes
1933414	F	40	2.9*1.5*1.8	III	no	no
1940640	F	66	3.1*1.2	II	yes	no
2004407	F	64	3*2.5*1.5	III	no	no
2005288	F	46	2.5*1.8*2	III	no	no
2006047	F	60	2.8*1.9	III	yes	no
2008260	F	59	2*1.9	III	no	no
2009329	F	37	2.2*1.8	III	yes	NA
2009381	F	47	2	NA	no	no
2009850	F	49	2.6*2.4	III	yes	no
2017611	F	57	1.7*1.2	II	yes	no
2039179	F	42	2.3*1.7	III	yes	no
2040686	F	40	1.7*1*1	II	no	no
2045012	F	40	1.9*1.1	III	no	no
2046297	F	37	5.5*5*1.2	II	yes	no
348981	F	56	8*6*2	II-III	yes	no
354300	F	43	2.5*2.5*2	NA	no	no
359448	F	30	1.5	NA	no	no
94377	F	NA	2*2	II	yes	yes
98389	F	NA	2	II	no	no
98475	F	NA	7.5*2*2	NA	no	no
99145	F	NA	0.7*0.9	II	yes	no

NA, means no data is available.

Statistics

All statistical analyses were completed using R software (Version 4.0.1). Based on the data homogeneity of variance and normal distribution, either the independent sample *t* test or Wilcoxon signed rank test was chosen. The log-rank test was used to evaluate survival significance. Spearman's correlation coefficient was used to assess the correlation between two continuous variables. $p < 0.05$ was considered statistically significant.

RESULTS

Patient Characteristics

A total of 30 patients with diagnosed triple-negative breast cancer were included in this study, including 15 breast cancer with lymph node metastasis and 15 breast cancer without lymph node metastasis from our hospital. 30 pairs of cancerous tissue and matched normal tissue were obtained from these 30 breast cancer patients. Cancer tissue and matched normal tissue were all pathologically diagnosed by two professional pathologists to determine whether they were cancer or normal tissue. Main clinical characteristics, including age at diagnosis, tumor size, pathological grade, metastasis status and recurrence, were showed in **Table 1**.

Key Genes Throughout the Oncogenesis and Progression of Breast Cancer

To identify key genes throughout the oncogenesis and progression of breast cancer, we first performed differential expression analysis using RNA-seq data of 30 pairs of breast tissues (cancerous tissue and matched normal tissue) from 30 TNBC patients enrolled in this study, and then performed differential expression analysis using RNA-seq data from 15 TNBC with lymph node metastasis and 15 TNBC without lymph node metastasis. The results showed that there were 2,953 DEGs between cancerous and matched normal tissue (**Figure 1A**), as well as 975 DEGs between TNBC with lymph node metastasis and TNBC without lymph node metastasis (**Figure 1B**). There were 117 overlapping DEGs between the above two gene sets (**Figure 1C**), which were potential key genes involved in the oncogenesis and progression of TNBC. Meanwhile, heatmap analysis showed a distinct gene expression pattern between cancerous tissue and matched normal tissue (**Figure 1D**), and between TNBC with lymph

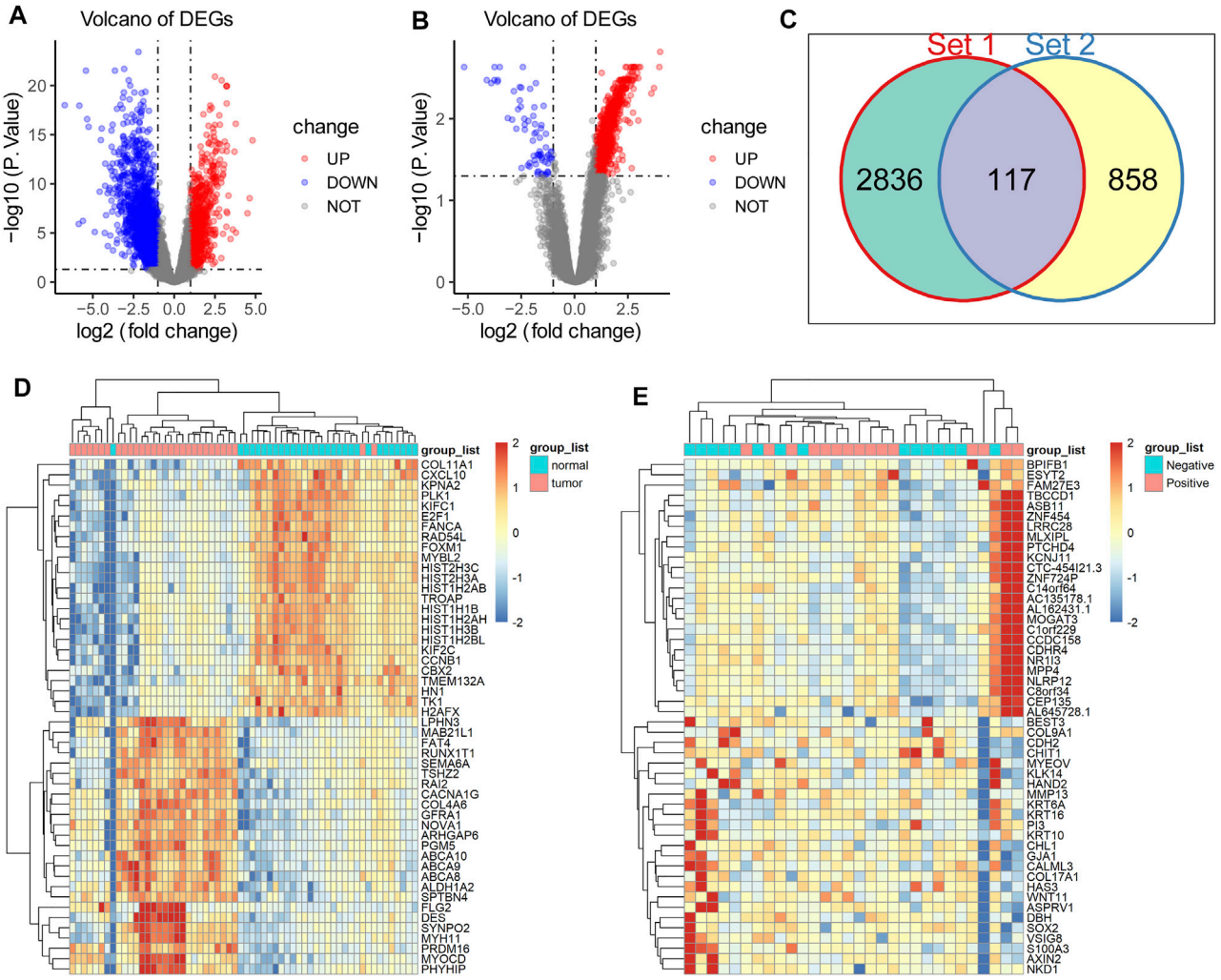

FIGURE 1 | Acquisition of DEGs in TNBC. **(A)** There were 2,953 DEGs between normal and cancerous tissue, including 981 upregulated DEGs (red dots) and 1972 downregulated genes (blue dots). **(B)** There were 975 DEGs between TNBC with lymph node metastasis and TNBC without lymph node metastasis, including upregulated 905 DEGs (red dots) and 70 downregulated genes (blue dots). **(C)** There were 117 overlapping DEGs between the above two gene sets. **(D)** Heatmap anaysis showed a distinct gene expression pattern between cancerous tissue and matched normal tissue. **(E)** Heatmap analysis showed a distinct gene expression pattern between TNBC with lymph node metastasis and TNBC without lymph node metastasis.

node metastasis and TNBC without lymph node metastasis (**Figure 1E**).

Functional Annotation of DEGs

Since we identified 117 key DEGs throughout the development and progression of TNBC, and revealed a distinct expression present between different tissue, we next wondered to know the underlying biological function and signaling pathways. To explore the potential affected biological function and signaling pathways, we performed functional enrichment analysis for DEGs using R package “clusterProfiler”. The findings showed that extracellular matrix organization, extracellular structure organization and regulation of trans-synaptic signaling were the most enriched biological process in gene oncology (**Figure 2A**). PI3K-Akt signaling pathway, calcium signaling pathway, cAMP signaling pathway, and cell cycle were the most enriched KEGG pathways (**Figure 2C**). GSEA findings showed that cell cycle and p53 signaling pathways were the most enriched KEGG pathways (**Figure 2 B, D**). These signaling pathways were involved in the development of breast cancer, and potential targeted pathways in the research and treatment of breast cancer.

Exploration of Tumor Immune Microenvironment in Breast Cancer

Understanding the interactions between cancer and the host immune system is critical for identifying key pathogenic molecules, improving drug sensitivity, and developing new therapeutic strategies. To investigate the effects of tumor immune microenvironment on breast cancer, we estimated the abundance of immune cells in normal tissue and cancerous tissue

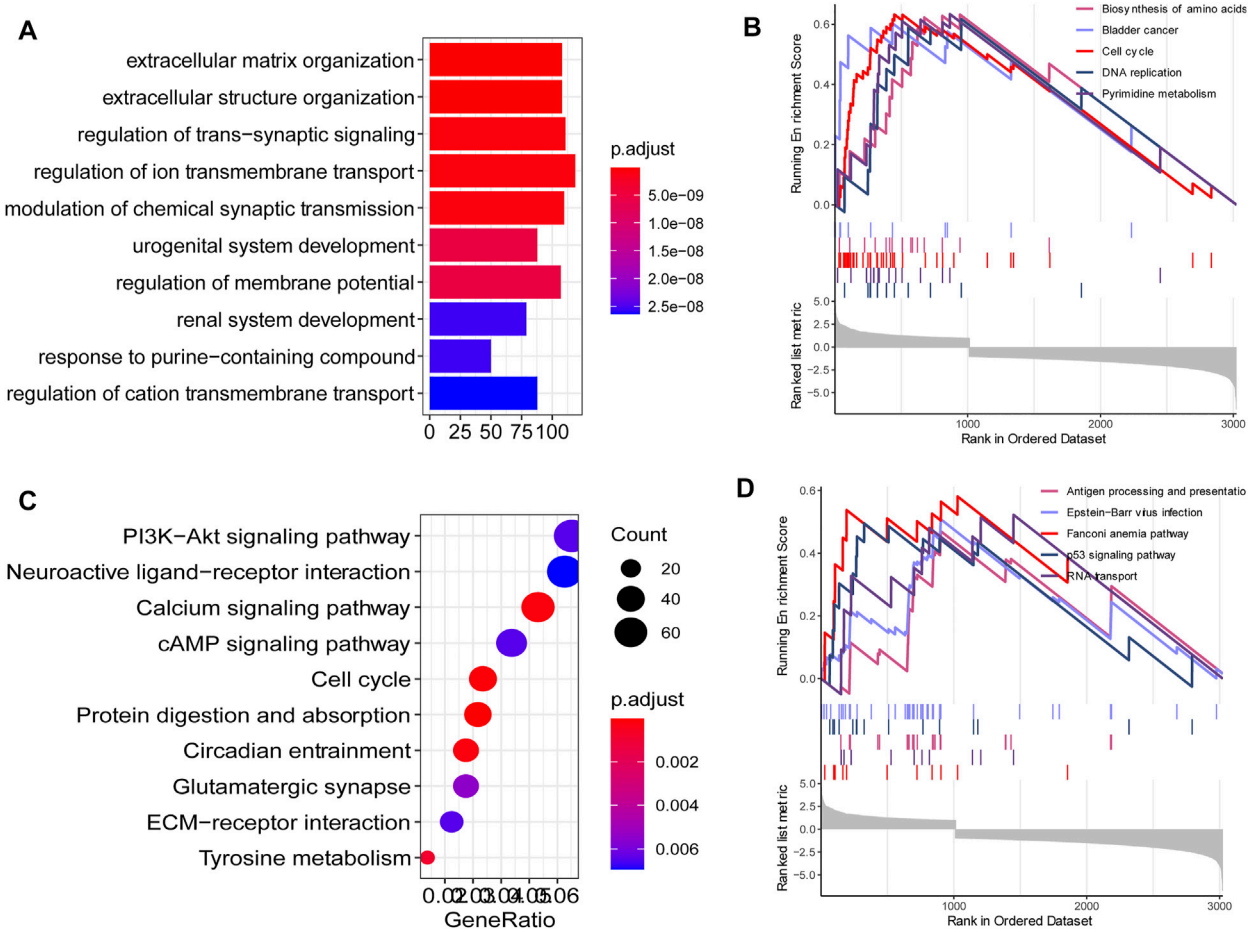

FIGURE 2 | Functional annotation. **(A)** The top 10 enriched biological process in gene oncology. **(C)** The top 10 KEGG pathways in gene oncology. **(B, D)** The top 10 KEGG pathways in GSEA findings.

based on RNA-seq data using TIMER 2.0. The results showed that M2 macrophage, B cell plasma and $CD8^+$ T cell were the top three immune cell types in the normal tissues (**Figure 3A**), while $CD8^+$ T cell, M2 macrophage and B cell plasma were the top three immune cell types in the cancerous tissues (**Figure 3B**). Notably, $CD8^+$ T cell, M0 macrophage, M1 macrophage, regulatory T cell and follicular helper T cell were significantly elevated in cancerous tissue as compared with normal tissue, suggesting an elevated immune response in the tumor (**Figure 3C**).

Identification of Cytotoxicity-Associated Key Genes in TNBC

Considering $CD8^+$ T cell infiltration was significantly elevated in cancerous tissue, we wondered to identify key genes associated with $CD8^+$ T cell infiltration. We performed a correlation analysis between the expression levels of 117 key DEGs and $CD8^+$ T cell infiltration. We found there were 22 genes significantly associated with $CD8^+$ T cell infiltration.

To investigate whether these 22 pivotal genes were also implicated in immune cell cytotoxicity, we analyzed the association of each pivotal gene expression and immune cell cytotoxicity using Pearson correlation analysis. Immune cell cytotoxicity of each sample was quantified based on expression levels of *CD8A, CD8B, GZMA, GZMB* and *PRF1*, using ssGSEA. The expression of cytotoxicity genes *CD8A, CD8B, GZMA, GZMB* and *PRF1* can represent immune cell cytotoxicity (Huson et al., 2016; Mitchell et al., 2020) (Balint et al., 2020; Bassez et al., 2021). The results showed that five genes (*GALNTL5, MLIP, HMCN2, LRRN4CL, DUOX2*) were markedly corelated with cytotoxicity ($p < 0.05$, $R < -0.3$; **Figure 4A**). These five genes were inversely associated with $CD8^+$ T cell infiltration (**Figures 4B–F**), and also negatively correlated with cytotoxicity (**Figures 4G–K**), suggesting their important role in tumor immunity.

Investigation on the Role of the Five Cytotoxicity-Associated Key Genes in TNBC

Since the above analysis revealed five key genes which were negatively correlated with $CD8^+$ T cell infiltration and

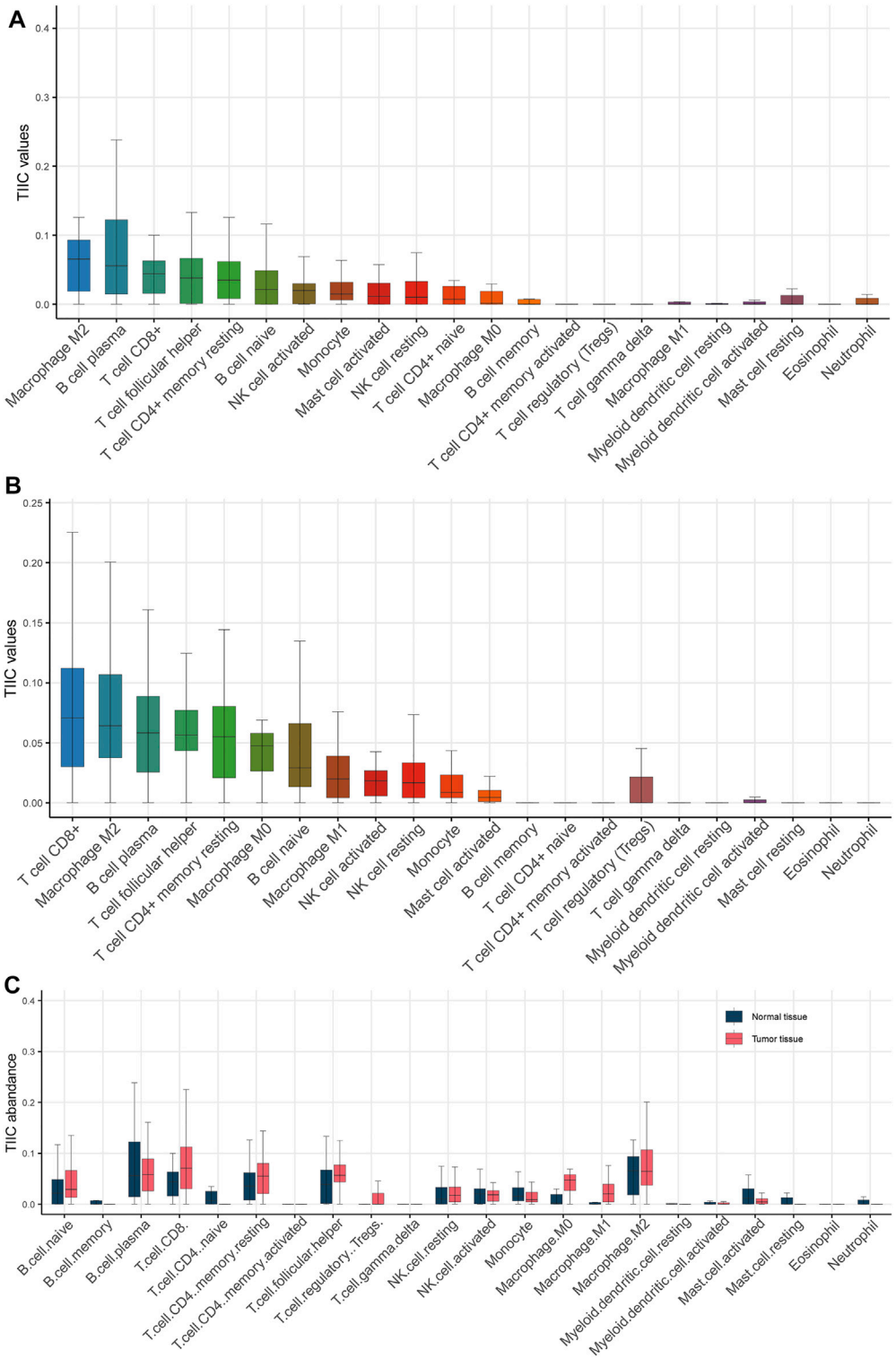

FIGURE 3 | Estimation of the abundance of tumor infiltrating immune cells. **(A)** Various types of immune cells in normal tissue. **(B)** Various types of immune cells in cancerous tissue. **(C)** Boxplot showed that $CD8^+$ T cell, M0 macrophage, M1 macrophage, regulatory T cell and follicular helper T cell were significantly elevated in cancerous tissue as compared with normal tissue.

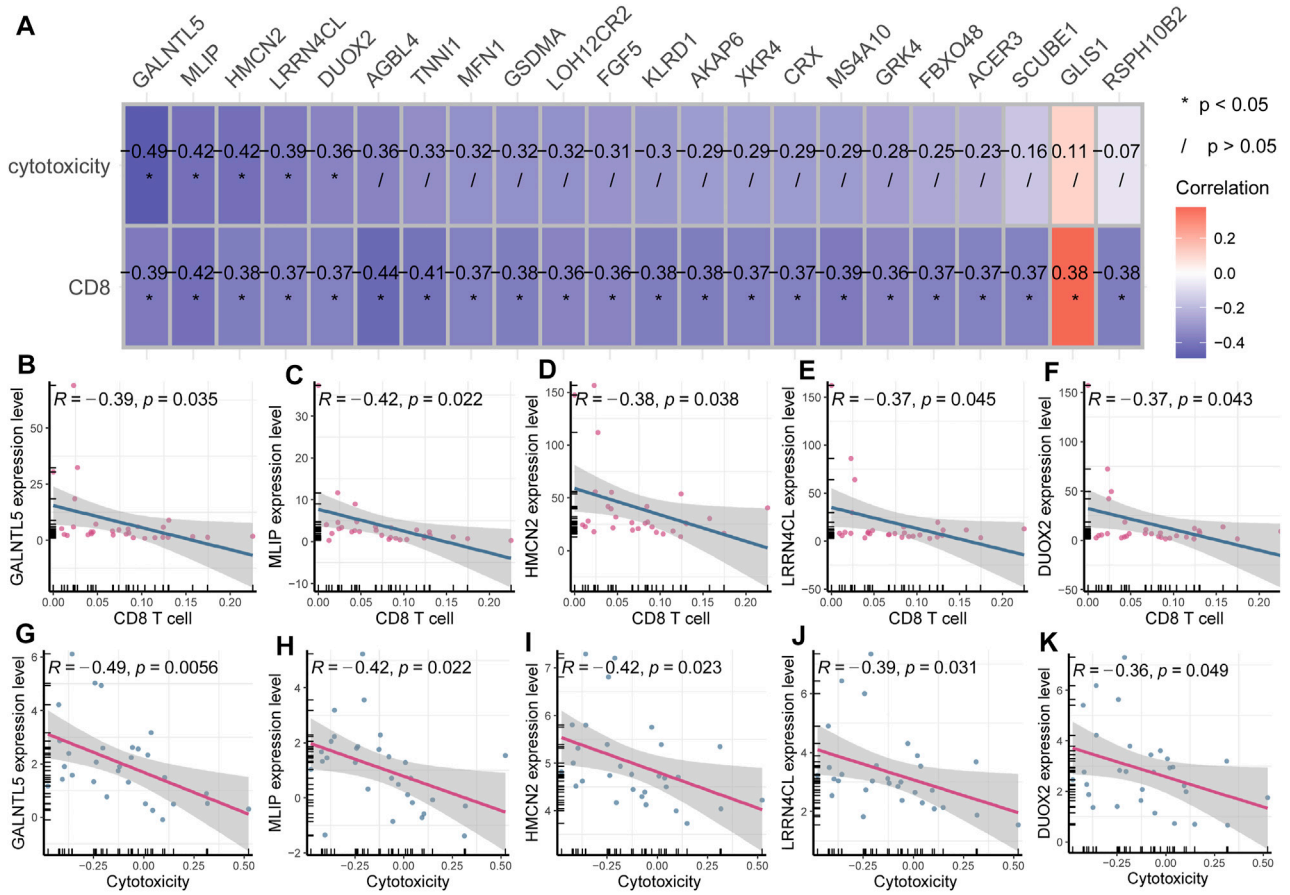

FIGURE 4 | Identification of cytotoxicity-associated key genes in breast cancer **(A)** Heatmap plot displayed key genes associated with $CD8^+$ T cell infiltration and cytotoxicity. **(B–F)** *GALNTL5, MLIP, HMCN2, LRRN4CL* and *DUOX2* were inversely associated with $CD8^+$ T cell infiltration. **(G–K)** *GALNTL5, MLIP, HMCN2, LRRN4CL* and *DUOX2* were also negatively correlated with immune cell cytotoxicity.

cytotoxicity, we wondered whether they played a role in the progression of TNBC and had a survival value. To validate their expression in the breast cancer tissue, we analyzed the expression of these five genes using RNA-seq data of breast cancer patients from our hospital. Consistent with the above findings, these four key genes (*MLIP, HMCN2, LRRN4CL*, and *DUOX2*) were critically upregulated in the cancerous tissue than in the normal tissue (t-test; $p < 0.05$; **Figure 5A–E**), further highlighting their protumor effects in breast cancer and their potential as a therapeutic target in cancer treatment.

Next, we analyzed the survival value of these five genes using RNAs-seq data and clinical data from the TCGA cohort of breast cancer patients, and found that *MILP*, *LRRN4CL*, and *DUOX2* had a significant survival relevance (log-rank test; $p < 0.05$), while *GALNTL5* and *HMCN2* had no survival value (log-rank test; $p > 0.05$; **Figure5 F–J**).

(F-J) Survival analysis was performed for *GALNTL5, MLIP, HMCN2, LRRN4CL* and *DUOX2* using log-rank test for RNA-seq data and the corresponding survival data from the TCGA cohort of patients with breast cancer.

Association of Cytotoxicity-Associated Genes With Therapeutic Response

As *GALNTL5, MLIP, HMCN2, LRRN4CL* and *DUOX2* were identified as cytotoxicity-associated genes, we next investigated their effects on therapeutic response by analyzing RNA-seq data and drug sensitivity of multiple breast cancer cell lines using Pearson coefficient analysis. As expected, the results showed the five cytotoxicity-associated genes were reflective of drug sensitivity of multiple agents in breast cancer cell lines (**Figure 6**). The targeted signaling pathways of each agent were shown in **Supplementary Table S1**.

DISCUSSION

We performed RNA sequencing for 30 pairs of TNBC tissue and matched normal tissue from our hospital, and identified five pivotal genes (*GALNTL5, MLIP, HMCN2, LRRN4CL,* and *DUOX2*), which were correlated with associated with $CD8^+$ T cell infiltration, tumor progression and therapeutic efficacy.

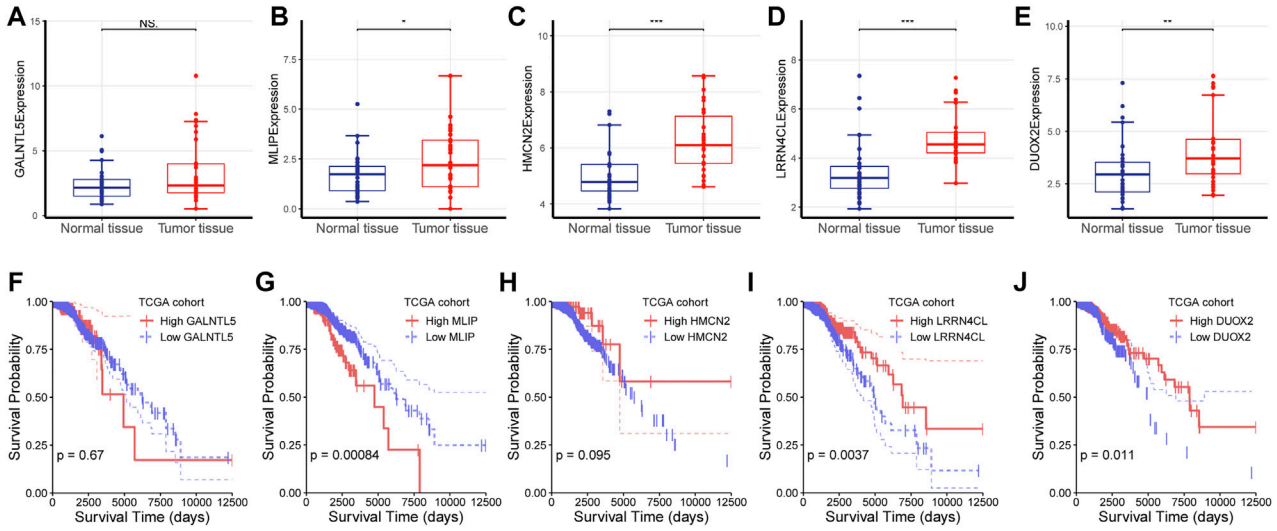

FIGURE 5 | Investigation on the role of the five cytotoxicity-associated key genes **(A–E)** The expression of each gene was characterized between normal tissue and tumor tissue. NS: No significance, ∗: $p < 0.05$, ∗∗: $p < 0.05$, ∗∗∗: $p < 0.001$.

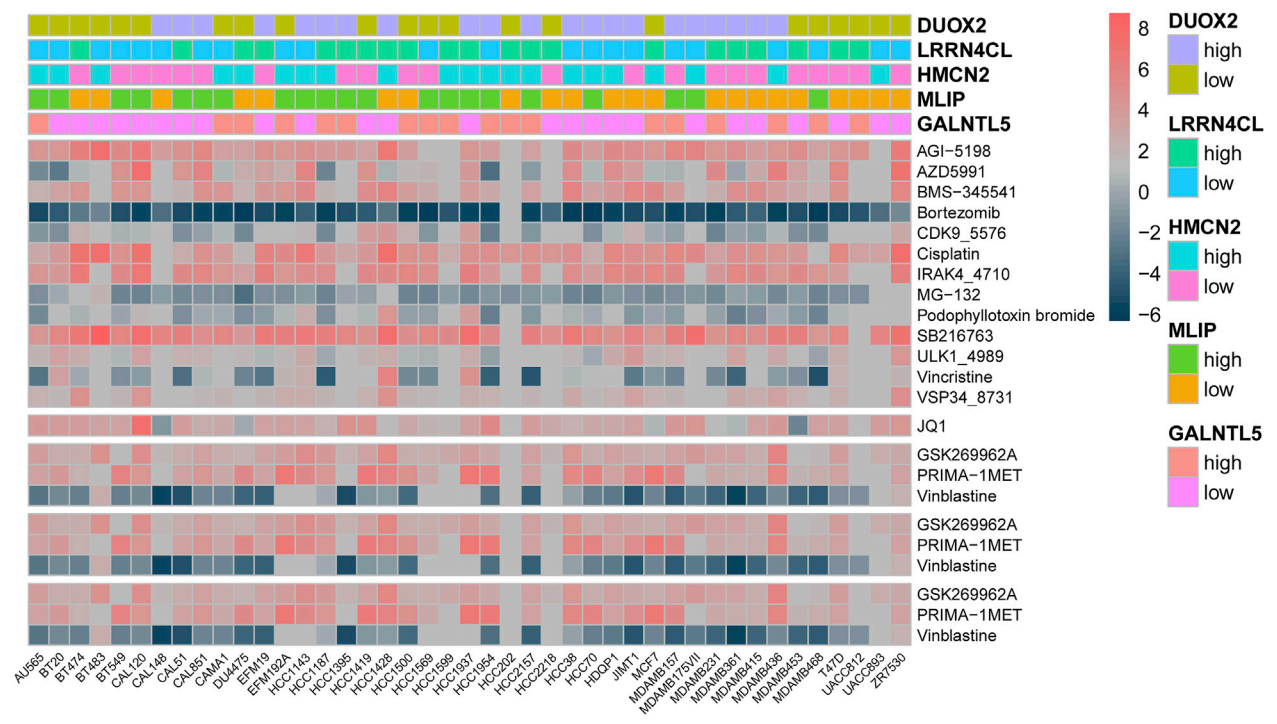

FIGURE 6 | Heatmap analysis showed that *GALNTL5, MLIP, HMCN2, LRRN4CL* and *DUOX2* were associated with drug sensitivity of multiple agents in breast cancer cell lines.

These findings will facilitate the understanding of the mechanism underlying the progression of breast cancer and the function of tumor immune microenvironment.

We observed that $CD8^+$ T cell, M0 macrophage, M1 macrophage, regulatory T cell and follicular helper T cell were significantly elevated in cancerous tissue compared with normal tissue. Tumor microenvironment is a dynamic and complex system that consists of various immunocytes, including regulatory tissue-resident $CD8^+$ T cells, macrophages, regulatory T cells, tumor-associated macrophages (TAMs), and

so on (Chew et al., 2017). The interaction between cancerous cells and the surrounding immune cells affect the expansion, migration and invasion of tumor cells (Hernandez-Gea et al., 2013). In previous studies, $CD8^+$ T cells are the predominant T cell subset in the tumor microenvironment, and correlated with improved survival outcomes in various cancers, including colorectal cancer (Mansuri et al., 2021), esophageal cancer (Hao et al., 2020), and gastric cancer (Lee et al., 2018). Macrophages play a core role in tumor immune evasion, and are expected to be the next Frontier in the immunotherapy for cancer (Qiu et al., 2021). Generally, macrophages can be divided into two categories: classically activated macrophages (M1) and alternatively activated macrophages (M2), respectively. M1 macrophages are characterized by CD68, CD86, and CD80, and secrete cytokines and chemokines like TNF-α, IL-1β, IL-12, CXCL9, CXCL10, to promote the pro-inflammatory Th1 response. M2 macrophages are featured by CD163, CD204, and CD206, and exert immunomodulatory effects, inhibiting endogenous antitumor immunity. The interaction of these tumor-infiltrating immune cells is complicated and requires to further profiling. Our findings will help to clarify the tumor microenvironment in breast cancer and the design of immunocyte-based immunotherapies.

This study still found five genes (*GALNTL5, MLIP, HMCN2, LRRN4CL, DUOX2*) associated with $CD8^+$ T cell infiltration and cytotoxicity. The polypeptide N-acetylgalactosaminyltransferase-like protein 5 (GALNTL5) is involved in male fertility; however, its involvement in the development of breast cancer remains unclear (Yao et al., 2017). Muscle-enriched A-type lamin-interacting protein (MlIP) is a recently discovered Amniota gene that encodes proteins of unknown biological function (Ahmady et al., 2021). And roles of *MLIP, HMCN2* in breast cancer are unknown. *DUOX1* and *DUOX2* is an H2O2-generating enzyme related to a wide range of biological features, such as hormone synthesis, host defense, cellular proliferation, and fertilization (Fortunato et al., 2018). *DUOX1* has been involved in breast cancer, whereas the role of *DUOX2* on breast cancer is still unreported. These five newly identified genes are potential therapeutic targets in breast cancer therapy.

This study has several contributions to breast cancer research. First of all, the previously identified key biomarkers of breast cancer are either related to the pathogenesis or related to progression (Wang et al., 2021b; Dameri et al., 2021; Gonzalez-Ericsson et al., 2021; Xing et al., 2021), and few biomarkers have been identified that are related to both occurrence and progression of TNBC. Here, we identified five key genes that played a role in the pathogenesis and progression of TNBC, suggesting their potential to be candidate therapeutic targets that benefit more patients. Second, we used five key molecules to connect tumor development, tumor immunity, and drug therapy in tandem, emphasizing the central role of tumor immunity in tumor development and clinical treatment. Finally, this study included a total of 30 TNBC patients, including 30 pairs of paired RNA-seq data and corresponding detailed clinical information, which can provide research resources for others' research.

This work has several limitations to further address. First, the study was mainly based on RNA-seq data, thereby needing further experiments *in vitro* and *in vivo*. Secondly, although we revealed the function of five key genes on $CD8^+$ T cells in the tumor microenvironment, we did not explore its role in the dendritic cells and related chemokines that are involved in the function of $CD8^+$ T cell.

In conclusion, we performed RNA sequencing of 30 pairs of normal and tumorous tissues from our hospital, and found several key gens associated with tumor progression and therapeutic efficacy. The findings would provide potential molecular targets for the treatment of breast cancer.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the GEO repository, accession number GSE183947.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee in the Nanfang Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YZ, GT, HC, TL, CW, GW, and HS were responsible for the literature review and writing Introduction and Discussion of the manuscript. XW, HT, WY, and LL analyzed the bioinformatics data and wrote Material and Methods and Results sections of the manuscript.

FUNDING

This work was supported by the National Key R and D program of China (2017YFC1309002) and The Science Foundation of Shenzhen Science and Technology Innovation Committee (2018) 16598 (NO. 20180214150032959).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.723477/full#supplementary-material

REFERENCES

Ahmady, E., Blais, A., and Burgon, P. G. (2021). Muscle Enriched Lamin Interacting Protein (Mlip) Binds Chromatin and Is Required for Myoblast Differentiation. *Cells.* 10, 615. doi:10.3390/cells10030615

Aysola, K., Desai, A., Welch, C., Xu, J., Qin, Y., Reddy, V., et al. (2013). *Triple Negative Breast Cancer - An Overview*. Hereditary Genet. Suppl 2, 001. doi:10.4172/2161-1041.S2-001

Balint, Š., Müller, S., Fischer, R., Kessler, B. M., Harkiolaki, M., Valitutti, S., et al. (2020). Multiprotein Particles From T Cells Deliver Cytotoxic Cargo to Targets. *Cancer Discov.* 10, 899. doi:10.1158/2159-8290.CD-RW2020-072

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature.* 483, 603–607. doi:10.1038/nature11003

Bassez, A., Vos, H., Van Dyck, L., Floris, G., Arijs, I., Desmedt, C., et al. (2021). A Single-Cell Map of Intratumoral Changes During Anti-PD1 Treatment of Patients With Breast Cancer. *Nat. Med.* 27, 820–832. doi:10.1038/s41591-021-01323-8

Chew, V., Lai, L., Pan, L., Lim, C. J., Li, J., Ong, R., et al. (2017). Delineation of an Immunosuppressive Gradient in Hepatocellular Carcinoma Using High-Dimensional Proteomic and Transcriptomic Analyses. *Proc. Natl. Acad. Sci. USA.* 114, E5900–e5909. doi:10.1073/pnas.1706559114

Dameri, M., Ferrando, L., Cirmena, G., Vernieri, C., Pruneri, G., Ballestrero, A., et al. (2021). Multi-Gene Testing Overview With a Clinical Perspective in Metastatic Triple-Negative Breast Cancer. *Int. J. Mol. Sci.* 22, 7154. doi:10.3390/ijms22137154

Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., et al. (2007). Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. *Clin. Cancer Res.* 13, 4429–4434. doi:10.1158/1078-0432.ccr-06-3045

Fortunato, R. S., Gomes, L. R., Munford, V., Pessoa, C. F., Quinet, A., Hecht, F., et al. (2018). DUOX1 Silencing in Mammary Cell Alters the Response to Genotoxic Stress. *Oxid Med. Cel Longev.* 2018, 3570526. doi:10.1155/2018/3570526

Garrido-Castro, A. C., Lin, N. U., and Polyak, K. (2019). Insights Into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment. *Cancer Discov.* 9, 176–198. doi:10.1158/2159-8290.cd-18-1177

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., 3rd, et al. (2019). Next-Generation Characterization of the Cancer Cell Line Encyclopedia. *Nature.* 569, 503–508. doi:10.1038/s41586-019-1186-3

Gonzalez-Ericsson, P. I., Wulfkhule, J. D., Gallagher, R. I., Sun, X., Axelrod, M. L., Sheng, Q., et al. (2021). Tumor-Specific Major Histocompatibility-II Expression Predicts Benefit to Anti-PD-1/l1 Therapy in Patients With HER2-Negative Primary Breast Cancer. *Clin. Cancer Res.* 27, 5299–5306. doi:10.1158/1078-0432.ccr-21-0607

Hao, J., Li, M., Zhang, T., Yu, H., Liu, Y., Xue, Y., et al. (2020). Prognostic Value of Tumor-Infiltrating Lymphocytes Differs Depending on Lymphocyte Subsets in Esophageal Squamous Cell Carcinoma: An Updated Meta-Analysis. *Front. Oncol.* 10, 614. doi:10.3389/fonc.2020.00614

Hernandez-Gea, V., Toffanin, S., Friedman, S. L., and Llovet, J. M. (2013). Role of the Microenvironment in the Pathogenesis and Treatment of Hepatocellular Carcinoma. *Gastroenterology.* 144, 512–527. doi:10.1053/j.gastro.2013.01.002

Hu, Y., Wu, D., Feng, X., and Shi, Z. (2021). Research on the Effect of Interfering With miRNA-155 on Triple-Negative Breast Cancer Cells. *Genes Genomics.* 28 doi:10.1007/s13258-021-01106-y

Huson, M. A. M., Scicluna, B. P., van Vught, L. A., Wiewel, M. A., Hoogendijk, A. J., Cremer, O. L., et al. (2016). The Impact of HIV Co-Infection on the Genomic Response to Sepsis. *PLoS One.* 11, e0148955. doi:10.1371/journal.pone.0148955

Kang, S. Y., Lee, S. B., Kim, Y. S., Kim, Z., Kim, H. Y., Kim, H. J., et al. (2018). Breast Cancer Statistics in Korea. *J. Breast Cancer.* 24 (2021), 123–137. doi:10.4048/jbc.2021.24.e22

Langmead, B., and Salzberg, S. L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods.* 9, 357–359. doi:10.1038/nmeth.1923

Lee, J. S., Won, H. S., Sun, D. S., Hong, J. H., and Ko, Y. H. (2018). Prognostic Role of Tumor-Infiltrating Lymphocytes in Gastric Cancer. *Medicine (Baltimore).* 97, e11769. doi:10.1097/md.0000000000011769

Li, B., and Dewey, C. N. (2011). RSEM: Accurate Transcript Quantification from RNA-Seq Data With or Without a Reference Genome. *BMC Bioinformatics.* 12, 323. doi:10.1186/1471-2105-12-323

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48, W509–w514. doi:10.1093/nar/gkaa407

Li, X., Sun, H., Liu, Q., Liu, Y., Hou, Y., and Jin, W. (2021). A Pharmacophore-Based Classification Better Predicts the Outcomes of HER2-Negative Breast Cancer Patients Receiving the Anthracycline- And/or Taxane-Based Neoadjuvant Chemotherapy. *Cancer Med.* 10, 4658–4674. doi:10.1002/cam4.4022

Mansuri, N., Birkman, E.-M., Heuser, V. D., Lintunen, M., Ålgars, A., Sundström, J., et al. (2021). Association of Tumor-Infiltrating T Lymphocytes With Intestinal-Type Gastric Cancer Molecular Subtypes and Outcome. *Virchows Arch.* 478, 707–717. doi:10.1007/s00428-020-02932-3

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential Expression Analysis of Multifactor RNA-Seq Experiments With Respect to Biological Variation. *Nucleic Acids Res.* 40, 4288–4297. doi:10.1093/nar/gks042

Mitchell, K. G., Diao, L., Karpinets, T., Negrao, M. V., Tran, H. T., Parra, E. R., et al. (2020). Neutrophil Expansion Defines an Immunoinhibitory Peripheral and Intratumoral Inflammatory Milieu in Resected Non-Small Cell Lung Cancer: a Descriptive Analysis of a Prospectively Immunoprofiled Cohort. *J. Immunother. Cancer.* 8, e000405. doi:10.1136/jitc-2019-000405

Mouabbi, J. A., Chand, M., Asghar, I. A., Sakhi, R., Ockner, D., Dul, C. L., et al. (2021). Lumpectomy Followed by Radiation Improves Survival in HER2 Positive and Triple-Negative Breast Cancer With High Tumor-Infiltrating Lymphocytes Compared to Mastectomy Alone. *Cancer Med.* 10, 4790–4795. doi:10.1002/cam4.4050

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets From Tissue Expression Profiles. *Nat. Methods.* 12, 453–457. doi:10.1038/nmeth.3337

Park, Y. Y. (2021). Genomic Analysis of Nuclear Receptors and miRNAs Identifies a Role for the NR3C1/miR-200 Axis in Colon Cancer. *Genes Genomics.* 43, 913–920. doi:10.1007/s13258-021-01112-0

Qiu, Y., Chen, T., Hu, R., Zhu, R., Li, C., Ruan, Y., et al. (2021). Next Frontier in Tumor Immunotherapy: Macrophage-Mediated Immune Evasion. *Biomark Res.* 9, 72. doi:10.1186/s40364-021-00327-3

Ren, S., Wang, W., Zhang, C., Sun, Y., Sun, M., Wang, Y., et al. (2021). The Low Expression of NUP62CL Indicates Good Prognosis and High Level of Immune Infiltration in Lung Adenocarcinoma. *Cancer Med.* 10, 3403–3412. doi:10.1002/cam4.3877

Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., et al. (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* 5, 1210–1223. doi:10.1158/2159-8290.cd-15-0235

Sohrabi, E., Moslemi, M., Rezaie, E., Nafissi, N., Khaledi, M., Afkhami, H., et al. (2021). The Tissue Expression of MCT3, MCT8, and MCT9 Genes in Women With Breast Cancer. *Genes Genomics.* 43, 1065–1077. doi:10.1007/s13258-021-01116-w

Sporikova, Z., Koudelakova, V., Trojanec, R., and Hajduch, M. (2018). Genetic Markers in Triple-Negative Breast Cancer. *Clin. Breast Cancer.* 18, e841–e850. doi:10.1016/j.clbc.2018.07.023

Su, N., Liu, L., He, S., and Zeng, L. (2021). Circ_0001666 Affects miR-620/WNK2 Axis to Inhibit Breast Cancer Progression. *Genes Genomics.* 43, 947–959. doi:10.1007/s13258-021-01114-y

Tagliamento, M., Agostinetto, E., Bruzzone, M., Ceppi, M., Saini, K. S., de Azambuja, E., et al. (2021). Mortality in Adult Patients with Solid or Hematological Malignancies and SARS-CoV-2 Infection With a Specific Focus on Lung and Breast Cancers: A Systematic Review and Meta-Analysis. *Crit. Rev. Oncology/Hematology.* 163, 103365. doi:10.1016/j.critrevonc.2021.103365

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering Splice Junctions With RNA-Seq. *Bioinformatics.* 25, 1105–1111. doi:10.1093/bioinformatics/btp120

Wan, Y., Wang, X., Liu, T., Fan, T., Zhang, Z., Wang, B., et al. (2021). Prognostic Value of CCR2 as an Immune Indicator in Lung Adenocarcinoma: A Study Based on Tumor-Infiltrating Immune Cell Analysis. *Cancer Med.* 10, 4150–4163. doi:10.1002/cam4.3931

Wang, N., Gu, Y., Chi, J., Liu, X., Xiong, Y., Zhong, C., et al. (2021a). Screening of DNA Damage Repair Genes Involved in the Prognosis of Triple-Negative Breast Cancer Patients Based on Bioinformatics. *Front. Genet.* 12, 721873. doi:10.3389/fgene.2021.721873

Wang, X., Shao, X., Huang, J., Lei, L., Huang, Y., Zheng, Y., et al. (2021b). Exploring the Concepts and Practices of Advanced Breast Cancer Treatment: a Narrative Review. *Ann. Transl Med.* 9–721. doi:10.21037/atm-21-1458

Xing, Z., Wang, R., Wang, X., Liu, J., Zhang, M., Feng, K., et al. (2021). CircRNA Circ-PDCD11 Promotes Triple-Negative Breast Cancer Progression via Enhancing Aerobic Glycolysis. *Cell Death Discov.* 7, 218. doi:10.1038/s41420-021-00604-y

Xu, S., Liu, Y., Zhang, T., Zheng, J., Lin, W., Cai, J., et al. (2021). The Global, Regional, and National Burden and Trends of Breast Cancer From 1990 to 2019: Results From the Global Burden of Disease Study 2019. *Front. Oncol.* 11, 689562. doi:10.3389/fonc.2021.689562

Yao, X., Ei-Samahy, M. A., Feng, X., Zhang, T., Li, F., Zhang, G., et al. (2017). Expression and Localization of Polypeptide N-Acetylgalactosaminyltransferase-Like Protein 5 in the Reproductive Organs and Sperm of Hu Sheep. *Anim. Reprod. Sci.* 187, 159–166. doi:10.1016/j.anireprosci.2017.10.020

Yin, L., Duan, J.-J., Bian, X.-W., and Yu, S.-c. (2020). Triple-negative Breast Cancer Molecular Subtyping and Treatment Progress. *Breast Cancer Res.* 22, 61. doi:10.1186/s13058-020-01296-5

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zhang, G., Xu, Q., Zhang, X., Yang, M., Wang, Y., He, M., et al. (2021). Spatial Cytotoxic and Memory T Cells in Tumor Predict superior Survival Outcomes in Patients With High-Grade Serous Ovarian Cancer. *Cancer Med.* 10, 3905–3918. doi:10.1002/cam4.3942

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 January 2022
doi: 10.3389/fgene.2021.762604

Identification of QTLs Linked to Phenological and Morphological Traits in RILs Population of Horsegram (*Macrotyloma uniflorum*)

Megha Katoch, Rushikesh Sanjay Mane and Rakesh Kumar Chahota*

Department of Agricultural Biotechnology, College of Agriculture, CSK HP Krishi Vishvavidyalaya, Himachal Pradesh, India

Horsegram [*Macrotyloma uniflorum* (Lam.) Verdc.] is an important legume but understudied in terms of its genetic improvement. Genetic information on various phenological and morphological traits may help in the utilization of new genes for breeding in horsegram and thus affect agronomic practices and crop yield. A total of 162 recombinant inbred lines derived from intraspecific crosses between HPKM249 × HPK4 was used to construct a genetic linkage map and to identify quantitative trait loci (QTLs) associated with phenological and morphological traits. Of the total 2011 molecular markers, which were screened on parental lines for polymorphism survey, 493 markers were found to be polymorphic and used for genotyping of recombinant inbred line population. Out of 493 polymorphic markers, 295 were mapped on ten linkage groups at LOD 3.5 spanning a total distance of 1,541.7 cM with an average distance between markers of 5.20 cM. Phenotypic data of two years at two different locations were used to identify QTLs by composite interval mapping A total of four QTLs (LOD ≥2.5) for phenological traits (days to 50% flowering, reproductive period and days to maturity) and seven QTLs (LOD ≥2.5) for morphological traits (plant height, primary branches and secondary branches) were detected across different environments. The phenotypic variation explained by QTLs ranged from 6.36 to 47.53%. The present study will help to augment scanty genomic information in this orphan crop that would provide genomics tools to breeders for its genetic enhancement through molecular-assisted selection.

Keywords: *Macrotyloma uniflorum*, genetic linkage map, QTL map, phenotypic variation, phenological traits, morphological traits

OPEN ACCESS

Edited by:
Rana Dajani,
Hashemite University, Jordan

Reviewed by:
Alireza Pour-Aboughadareh,
Seed and Plant Improvement Institute, Iran
Cengiz Toker,
Akdeniz University, Turkey

****Correspondence:***
Megha Katoch
meghakatoch24@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *22 August 2021*
Accepted: *09 December 2021*
Published: *25 January 2022*

Citation:
*Katoch M, Mane RS and Chahota RK (2022) Identification of QTLs Linked to Phenological and Morphological Traits in RILs Population of Horsegram (*Macrotyloma uniflorum*). Front. Genet. 12:762604. doi: 10.3389/fgene.2021.762604*

INTRODUCTION

Macrotyloma uniflorum (Lam.) Verdc (commonly called Horsegram) is an important legume and fodder crop of Asia and Africa, where it is grown as a staple food crop from prehistoric times. It has diploid chromosome number 2n = 20 (Cook et al., 2005) with a genome size of 398 Mbps (Shirasawa et al., 2021). The genus *Macrotyloma* comprises 32 wild species distributed in African, Australian, and Indian subcontinents. Amongst them, *Macrotyloma uniflorum* var. *uniflorum* is regarded as the only cultivated species grown in the Indian subcontinent andconsidered to be originated in Southern India (Vavilov, 1951; Zohary, 1970).

Horsegram is cultivated as food legume in India, Sri Lanka, Mauritius, Nepal, Malaysia, and Myanmar by the poor people of the marginal areas, whereas in Africa and Australia it is mainly

cultivated for animal feed (Asha et al., 2006). Owing to its valuable medicinal properties mentioned in Ayurveda, it is cultivated on a larger areas in India as compared to other countries. In India, Andhra Pradesh, Karnataka, and Tamil Nadu are the major horsegram producing states and cover approximately an area of 0.31 million ha and the collective production is 0.13 million tonnes with a yield of 430 kg/ha (DES 2016–2017).

The ever-increasing global population accompanied by degradation in cultivated land has put precocious pressure on breeders to enhance food grain production of non conventional crops so that food can be provided for everyone. Of the several hundred plant species known, only 120 species are cultivated for human food. However, of these, only nine crops supply approximately 75% of global plant-derived energy, wheat, rice, and maize being the top three crops (FAOSTAT 2021). Therefore, there is a dire need to explore other plant species/crops that bears the capacity to meet the increasing food supply-demand with simultaneous better nutritional values. Horsegram is one of the underprivileged crop and exhibit immense potential to cope with the increasing demand of food. Itpossesses inherent capability to grow under drought-like situations (Reddy et al., 1990), grow in varied temperature conditions (Ramya et al., 2013), tolerant to heavy metal stress (Sudhakar et al., 1992), and having a high percentage of protein, antioxidants, fiber and several important vitamins like Vitamin A, Vitamin B1, Vitamin B2, Vitamin B3 and vitamin C (Sodani et al., 2004; Reddy et al., 2005). Additionally, it has nitrogen fixation capacity which aids in improving the fertility of the soil. Also, the description of horsegram in Ayurveda is known for centuries and is widely used in the treatment of urinary stones and urinary diseases, regulates the abnormal menstrual cycle in women and is used to treat high fever, throat infection, cough, hiccups, and worms (Yadava and Vyas, 1994; Chunekar and Pandey, 1998; Neelam, 2007; Ravishankar and Vishnupriya, 2012). Its valuable nutritional and medicinal properties make it a crop of interest and potential food source of the future (National Academy of Sciences 1978).

The major bottleneck in horsegram productivity is the low genetic potential of most of the released varieties, which are generally developed from a narrow genetic base. Attempts should have been made to genetically improve such crops by combining favorable QTLs for various target traits in a single plant genotype (Wu, 1998). Lack of genetic variation and genomic information on important plant traits is a major obstruction to initiate a systematic breeding program in horsegram. The scarcity of genomic resources challenged/ prompted us to undertake Simple Sequence Repeat (SSR) marker developed in the related well-characterized legume species. We initiated the horsegram marker development program in the year 2012 and currently, we have a repertoire of more than 10,000 SSRs identified from well-characterized legumes including genomic and genic SSRs developed from horsegram transcriptome and genomic sequences. These genomic resources can be assessed in horsegram database (www.hillagric.ac.in:1005) and now being employed to improve grain yield and other agronomic traits. Phenological and morphological traits are important traits and knowledge of their genetics may help in the utilization of new genes for breeding and thus affect agronomic practices and crop yield. However, these traits have a complex phenotype, polygenic in nature, and are quantitatively inherited. Hence, mapping quantitative trait loci (QTLs) associated with genomic regions harboring genes for these traits represent a promising strategy for undertaking marker-aided breeding and trait improvement. Till now no such study has been reported in horsegram and thus needs to be determined. Therefore, for the genetic improvement of horsegram construction of fine linkage map and identifying QTLs linked to various important agronomic traits is essential which will increase the genomic resources and knowledge of genetics of these traits.

In the present study, we report the development of a linkage map generated with 295 molecular markers and quantitative trait loci (QTL) mapping of phenological and morphological traits in a recombinant inbred line (RIL) population derived from the intraspecific cross between HPKM249 × HPK4. The information presented in the study will help to dissect morphological and phonological traits with the help of molecular markers and provide breeders with genomics tools to select desirable plant type.

MATERIALS AND METHODS

Mapping Population

Mapping population consisting of 162 RILs derived from an intraspecific cross of HPKM249 × HPK4 was evaluated for morphological and phonological traits. The mapping population was developed through Single Seed Descent (SSD) method from F_2 to F_8 generations. The parental lines (HPKM249 & HPK4) show contrasting characteristics to each other for the traits under study (**Figure 1** and **Table 1**).

Phenotyping

During the crop season, F_8 progenies were planted in the experimental areas of department of Agriculture Biotechnology, CSK HPKV, Palampur, H.P. (latitude, 32.11 and longitude, 76.53) and at Hill Agricultural Research & Extension Centre, Bajaura, H.P. (latitude, 31.85 and longitude, 77.16) using Augmented Block Design (ABD) with two replications and four checks, VLG-1, HPKM249, HPK4 and HPK317 repeated after 20 rows. Each 1-m line consisted of 10 plants spaced 30 cms apart. Standard agronomic practices were followed for raising healthy crop. Phenotyping of various phenological and morphological traits were carried out at these locations (Palampur and Bajaura) over a period of two seasons (2016 and 2017). Data on five randomly taken plants in between the first and last plants of the same line was recorded. Three phenological traits viz. days to 50% flowering (FL), days to maturity (MT) and reproductive period (RP) and three morphological traits viz. plant height (PH, cm), number of primary branches (PB), and number of secondary branches (SB) were recorded. Plant height, number of primary

FIGURE 1 | Morphology of the two contrasting parents.

TABLE 1 | Variations in parents for different traits.

Trait	HPK4	HPKM249
Days to flowering	60–65	30
Days to maturity	120–124	80–82
Plant height (cm)	100.0	35.0–40.0
Growth type	Indeterminate	Determinate
Growth habit	Twining	Bush type
Maturity type	Asynchronous	Synchronous

branches, and number of secondary branches were measured just before the physiological maturity of the plant by taking readings on five plants in each line and averaging before analysis. Days from the date of sowing to the date when 50% of the plants in a line showed the first fully open flower and days from sowing to physiological maturity when 90% of the plants had turned brown were recorded for calculating days to 50% flowering (FL) and days to maturity (MT) respectively. The reproductive growth period (RP) was calculated as the days between the start of flowering and physiological maturity.

Statistical Analysis

Traits distribution was studied using skewness and kurtosis statistics by Past 3.25 software. Pearson correlation coefficients and frequency distribution among different traits were calculated using the same software. The parental phenotypic variance was analyzed using ANOVA.

Genotyping

Young leaf tissues (0.5–1 g) of 162 RILs individuals along with parents (HPK4 and HPKM249) were used for isolation of genomic DNA using the modified cTAB method (Murray and Thompson, 1980). Concentration (ng/μl) and purity of isolated DNA was checked on agarose gel electrophoresis and quantified on a microvolume spectrophotometer (Biospec-nano, Shimadzu Biotech, United States) using Tris EDTA as blank. PCR-based markers from different sources were used to screen

polymorphism between parent HPK4 and HPKM249. A total of 2011 PCR primers consisting of 63 Expressed Sequence Tag Simple Sequence Repeats (EST SSRs), 403 genic Simple Sequence Repeats (gSSRs), 387 genomic Simple Sequence Repeats (geSSRs), 24 drought specific Simple Sequence Repeats (dsSSRs), 300 Simple Sequence Repeats (SSR) from other legumes, 450 Random Amplified Polymorphic DNA (RAPD) markers, and 384 Conserved Ortholog Set (COS) markers of Cook's Lab UCD, United States were employed in the present study. The polymorphic primer pairs were then used for genotyping of the mapping population.

For amplification using SSR markers, a total reaction mixture of 10.0 µl volume was prepared to contain 4.80 µl of sterilized distilled water, 2.0 µl of template DNA (13 ng/µl), 0.5 µl each of forward and reverse primer (5 µM), 0.5 µl of MgCl2 (25 mM), 1.0 µl of 10X PCR buffer (10 mM Tris-HCl, 50 mM KCl, pH 8.3), 0.5 µl of dNTP mix (0.2 mM each of dATP, dGTP, dCTP and dTTP) and 0.2 µl of *Taq* polymerase (5U/µl). The reaction mixture for RAPD markers consisted of 1.5 of µl DNA, 2.5 µl of 10X PCR buffer (10 mM Tris, pH 9.0, 50 mM KCl, 0.01% Gelatin, 1.5 mM MgCl2), 0.2 µl of dNTPs (25 mM), 1.0 µl of MgCl2 (25 mM), 1.0 µl of primer (2 µM/µl), 0.1 µl of Taq DNA polymerase (5 U/µl) and made final volume of 25 µl by adding deionised water. The PCR amplification of genomic DNA was performed on Veriti 384® Thermal Cycler (Applied Biosystems, CA, United States). The PCR program for amplification was an initial denaturation cycle at 94°C for 5 min, followed by 45 cycles of denaturation at 94°C for 1 min, annealing at 50°C–65°C (for SSRs), and 37°C (for RAPDs) for 1 min and extension at 72°C for 2 min and final extension at 72°C for 7 min. The amplified products were resolved on either 6% Polyacrylamide Gel Electrophoresis in 1X TBE or 3% metaphor agarose gel (Lonza) in 1X TAE buffer depending on the size difference between amplified DNA along with the size markers (100 bp DNA ladder). The fragments were visualized using Gel-Documentation Unit (ENDUROTM GDS Gel Documentation System, United States) or silver-staining procedure.

The input file for linkage map construction was prepared manually by the scoring of amplified bands. HPKM249 type banding pattern was scored as "A", HPK4 type banding pattern was scored as "B" and heterozygous loci was scored as "H", whereas for RAPD markers the absence (0) and presence 1) of bands were recorded. Sizes of amplified fragments were noted by using a 100-bp DNA ladder (Fermentas, Lithuania).

Genetic Linkage Map and QTL Analysis

The genetic linkage map construction was performed with scored genotypic data file using JOINMAP® 4.1 program (van Ooijen, 2006). The LOD threshold >3.0 and <8.0 with a step-up of 0.5 was considered significant for identifying different linkage groups and clustering of markers on them. The linkage groups which show the highest number of markers with maximum linkage among them at different LODs values were selected.

For quantitative trait analysis, QTLs were identified using the composite interval mapping (CIM) method (Zeng, 1993; Zeng, 1994) implemented in Windows QTL Cartographer V2.5 software (Wang et al., 2005). The walking speed selected for QTLs was 2 cM with a window size of 10 cM using the Zmapqtl standard model 6. The forward regression algorithm was used to obtain cofactors. To calculate a genome-wide threshold for LOD score a 1000-permutation test at a significance level of $p = 0.05$ for shuffling genotypes with the phenotype means was performed (Doerge and Churchill, 1996). A LOD threshold score of ≥2.5 was selected for identification of the QTLs on the horsegram LGs. The location of QTLs with 95% confidence intervals was estimated by one LOD interval around the QTL peak (Mangin et al., 1994). The estimated additive effect and the percentage of phenotypic variation explained (based on the R2 value) by each putative QTL was calculated by the Zmapqtl procedure using the software with the CIM model. The QTL map was prepared with MapChart 2.32 software (Voorrips, 2002).

RESULTS

Trait Variations in Parents and RIL Population

Horsegram being an orphan crop, very little has been understood about the genetic structure of its traits. We analyzed six phenological and morphological traits as described in **Table 2** which represents the descriptive statistics for all the traits. Parental lines, HPKM249 and HPK4 which were used for developing the mapping population found to have significant differences for all the traits studied. Three phenological traits, namely days to 50% flowering (FL), reproductive period (RP), and days to maturity (MT) are important indicators of maturity and were used for phenotyping of the RILs population. Phenotyping of FL showed significant genetic variability for RILs in different years and locations. HPKM249 flowered in 36 days as compared to 54 days of HPK4 during 2016 at Palampur and similar results were observed at Palampur during 2017, whereas at Bajaura during 2017, HPKM249 flowered in 32 days as compared to 57 days of HPK4. The range for days to flowering among RILs varied from 30–58 days in 2016 at Palampur, 32–52 days in 2017 at Palampur, and 31–57 days in 2017 at the Bajaura location. Further, no significant difference was found among RILs in different years and locations. A similar trend was observed for RP and MT with no significant difference among RILs in different years and locations.

RIL population was phenotyped for various morphological traits like plant height (PH), number of primary branches (PB), and number of secondary branches (SB). Plant height (PH) varied from 34–98 cm and 48–106 cm at Palmapur for the year 2016 & 2017, respectively. A significant difference was observed for PH in 2017 at Bajaura (60–145 cm) as compared to the Palampur location. Similarly, PB varied from one to six branches at Palampur in both years (2016 and 2017), however a significant difference was observed for PB in 2017 at Bajaura, which varied from 3–10. Further, significant differences for PH and PB among RILs were observed in both seasons and locations. A similar result was observed for SB with significant differences among RILs in both years and locations.

TABLE 2 | Mean performance of parents and RILs for phenological and morphological traits.

Traits	Year	Location	HPKM249	HPK4	Range (RIL)	Mean	Sd
Morphological							
Plant height (PH)	2016	PLP	38.0	101.0	34.0–98.0	68.3	14.0
	2017	PLP	41.0	99.0	48.0–106.0	72.9	12.0
	2017	BJR	39.0	99.0	60.0–145.0	91.7	18.8
	Combined	—	39.8	99.3	52.1–116.8	78.4	13.3
Primary branches (PB)	2016	PLP	6.7	2.6	1.0–6.0	2.5	0.9
	2017	PLP	6.4	2.5	1.6–5.3	3.0	0.7
	2017	BJR	10.0	3.0	2.5–9.5	6.0	1.4
	Combined	—	7.7	2.7	1.9–5.7	3.9	0.7
Secondary branches (SB)	2016	PLP	8.0	5.2	1.8–14.0	5.0	1.5
	2017	PLP	12.2	3.6	3.7–10.7	6.5	1.4
	2017	BJR	14.0	5.0	5.0–18.0	10.8	2.4
	Combined	—	12.1	4.3	5.0–11.0	7.7	1.2
Phenological							
Days to 50% flowering (FL)	2016	PLP	36.0	54.0	30.0–58.0	41.2	5.9
	2017	PLP	36.0	54.3	32.7–52.7	41.6	4.7
	2017	BJR	32.0	57.5	31.0–57.0	40.9	4.9
	Combined	—	34.7	55.3	33.5–52.2	41.3	4.6
Reproductive period (RP)	2016	PLP	39.0	64.0	19.0–77.0	50.4	10.6
	2017	PLP	46.7	62.7	32.3–74.3	52.8	8.7
	2017	BJR	48.0	56.5	37.0–73.5	55.4	7.8
	Combined	—	45.8	60.8	34.7–73.7	53.3	7.8
Days to maturity (MT)	2016	PLP	75.0	118.0	71.0–115.0	91.6	10.4
	2017	PLP	82.7	117.0	71.6–114.0	94.4	9.7
	2017	BJR	80.0	114.0	78.5–112.5	96.3	9.2
	Combined	—	80.5	116.1	74.1–111.0	94.6	9.0

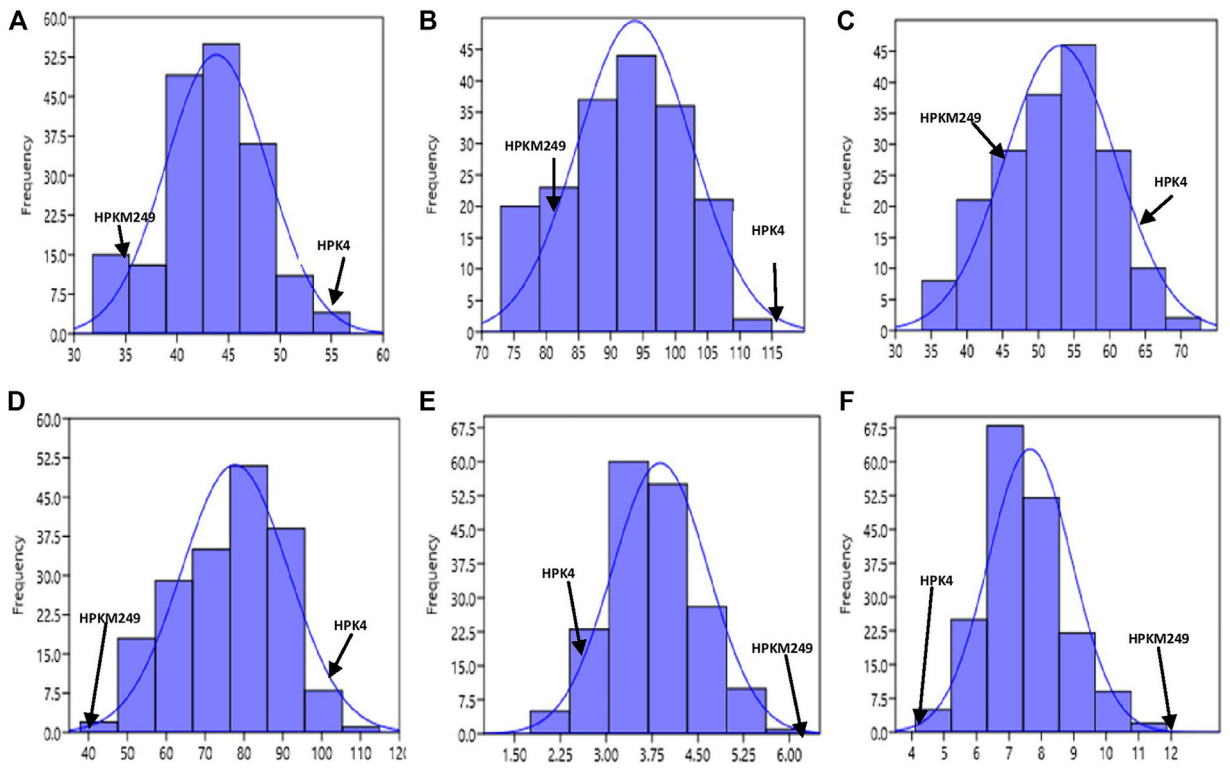

FIGURE 2 | Frequency distribution curve of **(A)** Days to 50% flowering **(B)** Days to maturity **(C)** Reproductive period **(D)** Plant height **(E)** No. of primary branches **(F)** No. of secondary branches.

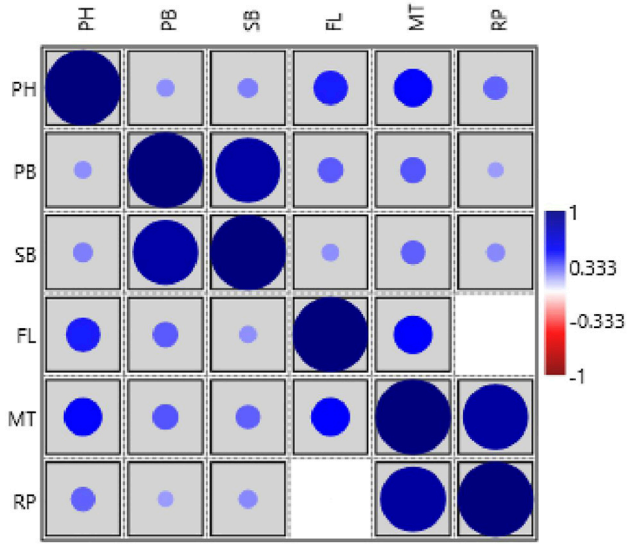

FIGURE 3 | Pearson's correlation matrix among different traits analyzed in the HPM249 HPK4 RILs (PH-Plant height, PB-No. of primary branches, SB-No. of secondary branches, FL-Days to 50% flowering, MT- Days to maturity, RP- Reproductive period).

The phenotypic values showed a normal frequency distribution, which is a typical characteristic of quantitative traits (**Figure 2**). Correlations values among traits in RIL populations showed that the traits were positively correlated with each other and were statistically significant ($p < 0.05$). The maximum correlation was found between MT and RP (r = 0.86) followed by PB and SB (r = 0.85) (**Figure 3** and **Table 3**). All traits showed positive correlations among themselves except for FL and RP which showed negative correlation with each other in all environments. The ANOVA of 162 RILs for different locations and environments showed significant variation for all the traits.

Parental Polymorphism and Genotyping of Mapping Population

The polymorphism survey was done on parental lines using 1177 SSR primer pairs [63 (Horsegram EST SSRs) + 403 (Horsegram genic SSRs) + 387 (Horsegram genomic SSRs) + 24 (drought specific SSRs) + 300 (SSRs from other legumes viz. red clover and *Medicago*)] of which 430 were found to be polymorphic (**Table 4**). Along with this, 450 RAPD primers were also screened out of which 55 were found to be polymorphic and in addition 384 COS markers of *Medicago truncatula* were screened out of which 8 were found to be polymorphic (**Table 4**). A total of 2011 primers were screened, of which 493 were found to be polymorphic and were further used for genotyping of 162 individuals (**Table 4**). The size of amplified bands produced by 493 polymorphic primers ranged between 100 and 250 bp. The genotyping data thus obtained was scored manually and used as the input file for the construction of the horsegram linkage map.

TABLE 3 | Pearson correlation coefficients among traits evaluated in the RIL population.

	PH	PB	SB	FL	MT	RP
Plant Height	1.00	0.22	0.25	0.44	0.49	0.31
No. of Primary branches	0.22	1.00	0.85	0.32	0.33	0.19
No. of Secondary branches	0.25	0.85	1.00	0.22	0.31	0.23
Days to 50% flowering	0.44	0.32	0.22	1.00	0.50	–0.01
Days to maturity	0.49	0.33	0.31	0.50	1.00	0.86
Reproductive period	0.31	0.19	0.23	–0.01	0.86	1.00

Construction of Genetic Linkage Map

Using JoinMap software, version 4.0, of the total 493 polymorphic markers, 295 (59.84%) were mapped on 10 LGs at LOD 3.5. The linkage map spanned 1,541.7 cM (Kosambi cM) length and an average marker interval size of 5.20 cM (**Table 5**). These 295 mapped markers include 15 EST SSRs, 87 genomic SSRs, 87 genic SSRs, 22 RAPDs, 73 SSRs from other species, four drought-specific markers, and seven COS markers. Each of the ten linkage groups differed from one another in terms of their length and the total number of markers mapped. Of the total 295 mapped markers, LG1 harbored 89 markers followed by LG2 which contained 58 markers. 35 markers were mapped on LG3, 29 markers were mapped on LG4, 19 were mapped on LG5 and LG7 whereas 18 markers were mapped on LG6. LG eight and LG nine contained the least number of markers with seven and 6 markers, respectively, and linkage group 10 harbored 15 markers (**Figure 4** and **Table 5**). Though LG7 is having a maximum length of 238.5 cM due to very less number of markers 19) present on it with an average marker density of 12.5 however is of less importance while mapping of different QTLs in comparison to linkage groups LG1 and LG2 having marker density of 2.0 and 2.7 respectively even having smaller linkage size (182.9 and 159 cM). Segregation distortion for all the 493 polymorphic markers was determined and of these 295 (59.83%) followed the expected segregation ratio, whereas 198 markers (40.16%) were found to show deviation. These distorted markers were excluded from the final analysis.

The maximum distance recorded between two markers was 61.2 cM on LG7 and the minimum distance was 0.003 cM on LG2. The number of markers present in different linkage groups was unequal. Four large groups having 12–19 markers within a length of 10 cM and five groups having 28–31 markers within a length of 30 cM were found. The length of the linkage groups did not reflect the number of markers linked on it as the distance between markers varied across different linkage groups. For example, LG1 carrying 89 markers having a length of 182.9 cM and an average marker distance of 2.0 cM, whereas LG4 having a length of 188.0 cM carrying 29 markers and an average marker distance of 6.5 cM, and LG5 having a length of 192.5 cM covered by 19 markers and average spacing of 10.0 cM.

QTL Mapping

The QTL analysis identified four QTLs (LOD ≥2.5) for phenological and seven QTLs (LOD ≥2.5) for morphological traits (**Figure 4** and **Table 6**). One QTL for days to 50% flowering (qFT01) was detected on LG2 flanked by markers MUGR644-MUMST80 at LOD score of 2.8 and explaining 6.62% of the

TABLE 4 | List of molecular markers used for construction of linkage map of horsegram.

S. No	Marker	Markers screened	Polymorphic markers	Percent polymorphism (%)	Markers mapped	Source
1	HUGMS	63	36	57.14	15	EST SSRs (Sharma et al., 2015)
2	MUMS	200	55	27.50	45	Genic SSRs (Sharma et al., 2015)
3	MUMST	100	37	37.0	22	Genic trirepeats (Sharma et al., 2015)
4	MUMSD	103	44	42.72	20	Genic Direpeats (Sharma et al., 2015)
5	MUGSSR	99	42	42.42	31	Genomic SSRs (Chahota et al., 2017)
6	MUSSR	50	24	48.0	16	Genomic SSRs (Chahota et al., 2017)
7	MUGR	94	30	31.91	20	Genomic SSRs (Chahota et al., 2017)
8	MUD	96	28	29.17	13	Genomic SSRs (Kaldate et al., 2017)
9	MUGSR	48	8	16.67	7	Genomic SSRs (Chahota et al., 2017)
10	RAPD	450	55	12.22	22	Operon Tech, United States of America and Fred Muehlbauer, United States of America
11	Drought specific primers	24	5	20.83	4	Charu and Manoj 2011
12	RcSSRs	196	88	44.90	56	Sato et al., 2005
13	MtSSRs	104	33	31.73	17	Eujayl et al., 2004
14	COS	384	8	2.08	7	Douglas R. Cook, UC, Davis, United States of America
—	Total	2011	—	493	24.52	295

TABLE 5 | Distribution of 295 markers on ten linkage groups of an intra-specific linkage map of horsegram.

LGs	Markers mapped	Map length (cM)	Average marker density (cM)
LG1	89	182.9	2.0
LG2	58	159.0	2.7
LG3	35	129.0	3.7
LG4	29	188.0	6.5
LG5	19	192.5	10.1
LG6	18	165.6	9.2
LG7	19	238.5	12.5
LG8	7	71.6	10.2
LG9	6	46.4	7.7
LG10	15	168.2	11.2
Total	295	1,541.7	5.2

phenotypic variation with allelic contribution by HPKM249 resulted in reduced flowering time by about >2 days. One QTL for the reproductive period (*qRP01*) was detected on LG5 flanked by MUGSSR10-RCS6448 at a LOD score of 2.7 and explaining 6.36% of the phenotypic variation (**Table 6**). Two QTLs namely *qMT01* and *qMT02* were detected for days to maturity on LG7 flanked by markers MUGSSR241-HUGMS39 at LOD 2.6 explaining 7.25% of the phenotypic variation and on LG9 flanked by HUGMS3-MUGR607 marker interval at LOD 2.9 explaining 47.53% of the phenotypic variation, respectively and in combination explained 54.78% of total phenotypic variation. Additive effect is the difference in the average performance of the RIL carrying early maturity allele of first parent with respect to those carrying the allele of second parent at the particular locus (QTL). A positive value (+) of the additive effect indicates the allele originating from HPKM249 and a negative value (–) of the additive effect indicates the allele originating from HPK4.

Additive effect demonstrated that HPKM249 contributed alleles for a reproductive period and days to maturity with QTL named *qMT01* resulted in reduced days to maturity by > 3days and *qMT02* resulted in reduced days to maturity by > 8 days (**Table 6**).

One QTL was detected for plant height namely *qPH01* on LG1 flanked by RCS6168-RCS6169 and explaining 6.6% of the phenotypic variation at a LOD value of 2.7. This QTL had an additive effect of 3.96 cm and was contributed by the allele from HPKM249. Three QTLs were detected for primary branches namely qPB01, qPB02, and qPB03 with two on LG6 (qPB01 and qPB02) both flanked by OPI66-MUMST29 and explaining 22.0 and 17.0% of the phenotypic variation, respectively, and one QTL on LG9 (qPB03) flanked by HUGMS3-MUGR607 explaining 32.4%of the phenotypic variation (**Table 6**). These QTLs in combination explained 71.4% of total phenotypic variation for primary branches. Additive effect demonstrated allelic contribution from both the parents. Three QTLs were detected for secondary branches all on LG7 namely *qSB01, qSB02,* and *qSB03*. Both *qSB01* and *qSB03* were flanked by MUD77-HUGMS13 marker interval explaining 23.6 and 15.5% phenotypic variation respectively. *qSB02* was flanked by MUMS13-MUMS95 marker interval explaining 7.5% of phenotypic variation. These QTLs in combination explained 46.6% of total phenotypic variation for secondary branches. Additive effect demonstrated allelic contribution from both the parents. The position of QTLs for phenological and morphological traits at LOD ≥2.5 on 10 linkage groups on horsegram has also been shown in **Figure 5**.

DISCUSSION

The main concern of the breeders and farmers is to enhance the yield of the important crops which can provide sufficient nutrition to the human population. Genetic improvement of underutilized crops and including them in commercial agriculture for more production to address this issue.

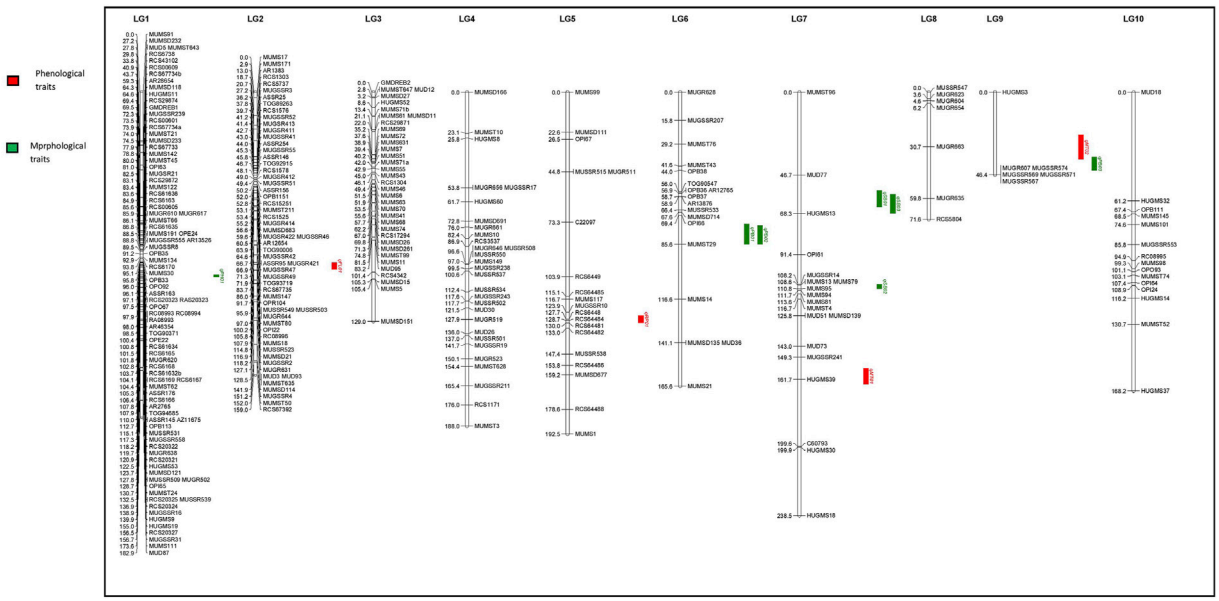

FIGURE 4 | Quantitative trait loci (QTLs) linked to phenological and morphological traits on 10 linkage groups of horsegram (Red color indicates QTLs for Phenological traits; Green color indicates QTLs for Morphological traits.

TABLE 6 | QTLs for various drought related traits identified using QTL Cartographer.

Trait	RIL (HPKM249 × HPK4)		LG	QTL name	Marker interval	LOD score	Additive effect[c]	Pve (R^2%)[d]
	Year	Loc						
Morphological								
Plant height	2016-2017	COMBINED	1	*qPHT01*	RCS6168-RCS6169	2.7	3.96	6.6
Primary branches	2017	PLP	6	*qPB01*	OPI66-MUMST29	4.2	0.37	22.0
	2017	PLP	9	*qPB03*	HUGMS3-MUGR607	5.4	–0.63	32.4
	2016-2017	COMBINED	6	*qPB02*	OPI66-MUMST29	3.8	0.34	17.0
Secondary branches	2017	BJR	7	*qSB01*	MUD77-HUGMS13	4.9	1.21	23.6
	2017	BJR	7	*qSB02*	MUMS13-MUMS95	3.3	–0.75	7.5
	2016-2017	COMBINED	7	*qSB03*	MUD77-HUGMS13	3.7	0.50	15.5
Phenological								
Days to 50% flowering	2016-2017	COMBINED	2	*qFL01*	MUGR644-MUMST80	2.8	1.19	6.62
Reproductive Period	2017	BJR	5	*qRP01*	MUGSSR10-RCS6448	2.7	3.87	6.36
Days to Maturity	2016	PLP	7	*qMT01*	MUGSSR241-HUGMS39	2.6	2.86	7.25
	2017	PLP	9	*qMT02*	HUGMS3-MUGR607	2.9	7.82	47.53

Horsegram is an important legume with great potential due to its medicinal, nutraceutical and capable in growing harsh environmental advantages but is still underutilized and understudied. It serves as an important source of protein for the poorest of poor society. It is grown as food by local people in some areas of developing countries and as feed for animals in dry areas (Chahota et al., 2013; Fuller and Murphy, 2018). However, its genetic improvement is ignored by both at scientists as well as at institutional levels. For genetic improvement of any crop, genetic information of its important traits is required. Phenological and morphological traits play an important role in enhancing productivity, adaptation and yield stability of any crop. With the current scenario of climate change and increase in feeding population the need to develop high yielding, early maturing and climate resilient varieties has increased. Different horsegram varieties exhibit variation in their flowering and maturity time therefore genetic knowledge of their phenology aids in the development of varieties with desired and useful characteristics such as early maturing and with high yield. Days to flowering has a direct implementation on other phenological traits like time to podding and maturity (Gaur et al., 2015) and to enhance the yield, balance between flowering time and maturity is essential (Kong et al., 2018). These traits are quantitatively inherited and are influenced by the environment (Sari et al., 2021). Thus knowing genetics and environmental interactions can help to explicate the intrinsic process of flowering time and maturity. Similarly improving morphological and plant architectural traits is essential for enhancing crop yield. Modulating important plant architectural traits can aid plant breeders in optimizing crop

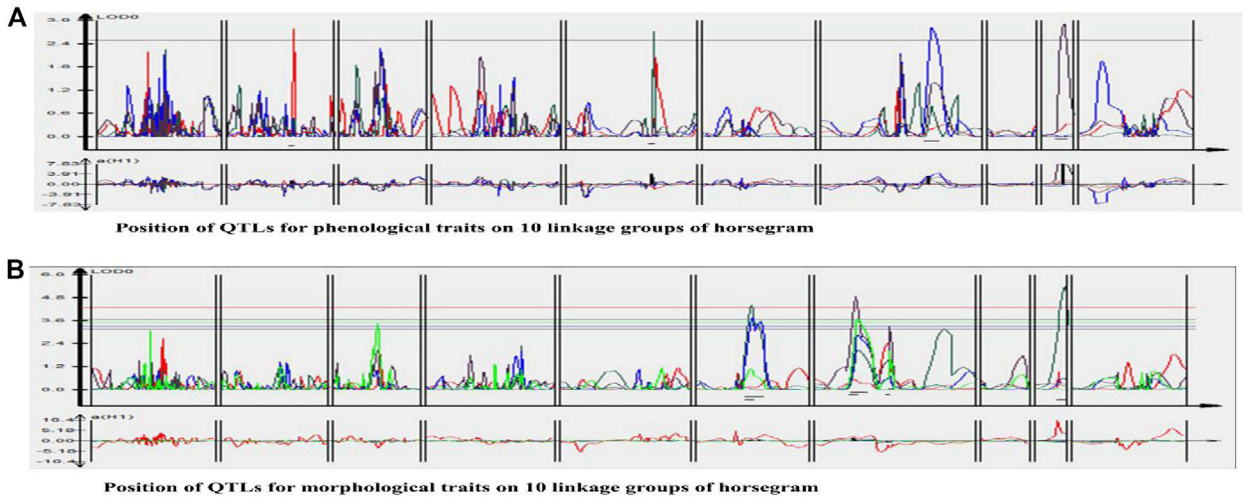

FIGURE 5 | Position of QTLs for **(A)** Phenological traits **(B)** Morphological traits on 10 linkage groups of horsegram.

performance and yield (Horton, 2000). Various studies have been reported in legumes explaining the positive association between various morphological traits like plant height, more primary and secondary branches, erect and determinate growth type with improved yield (Jain, 1975; Bahl and Jain, 1977; Sedgley et al., 1990).

Linkage maps and genetic information of a trait are prerequisites for starting any breeding program of desirable crop varieties. Construction of saturated linkage maps using molecular markers will further aid in the localization and mapping of genes/QTLs of different important traits. Since QTLs were first identified in tomato using RFLP-based linkage maps (Paterson et al., 1988), different QTLs for various important traits have been identified in many crops. However, the genetics of important traits in horsegram has not yet been determined and similarly, no QTLs for important traits have been mapped. This may be due to limited variation among cultivars and less available genetic resources in horsegram. In this study, we employed genic and genomic SSRs of horsegram, SSRs from well-characterized legume species and RAPDs to develop a genetic linkage map of horsegram and also identified QTLs linked to phenological and morphological traits. Identification of genomic regions controlling these important traits is the first step towards implementing genomic-assisted breeding in this orphan legume species. Development of genomic resources in horsegram lagged much behind compared to other major pulses. However, in recent years efforts have been made to develop molecular markers through the mining of transcriptome and genomic sequencing data (Sharma et al., 2015; Chahota et al., 2017; Kaldate et al., 2017) which is utilized in the study. Both parents HPK4 and HPKM249 crossed for development of RILs population have a good amount of variation for various phenological and morphological traits (**Figure 1** and **Table 1**). The level of polymorphism in parental line using molecular markers is 24.52% which is comparable to polymorphism observed in other legumes such as 22.1% in chickpea (Radhika et al., 2007), 23.6% in peanut (Hong et al., 2010), 26.8% in adzuki bean (Chaitieng et al., 2006) and 27.02% in soybean (Hwang et al., 2009). The present map is a well saturated intraspecific molecular linkage map of horsegram based on DNA markers containing 10 linkage groups and covering a map distance of 1,541.7 cM (**Table 4**). Large mapping population size along with better pairing and crossing over of chromosomes of two varieties belonging to the same species between them is the main reason for the development of a saturated linkage map. The length of linkage groups ranged from 46.4 cM in LG9 to 238.5 cM in LG7. The marker density ranged from 2.0 to 12.5 cM, with an average marker density of 5.2 cM showing variation in degrees of saturation of all linkage groups. The variation in saturation of markers on different linkage groups showed that the distribution of markers on each LGs was random. The maximum numbers of markers were mapped on LG1, which harbored 89 markers with the average marker density of 2.0 cM and minimum on LG9 which embraced 6 markers with an average marker density of 7.7 cM (**Figure 4** and **Table 5**). Such discrepancies could probably be eliminated by further saturating the map with more SSRs and SNPs markers (Grisi et al., 2007).

The main bottleneck in the development of high-yielding varieties particularly in grain legumes is the dearth of genomic and genetic information of many important traits. Phenological and morphological traits are considered complex traits exhibiting quantitative inheritance because their phenotypic expression relies on many factors (Kover et al., 2009). The genetic information of the important traits is revealed by dissecting the potential genomic regions harboring QTLs controlling these traits. Also, the breeding improvement of the crops is accelerated through marker-assisted selection (MAS) which is aided by the identification of linked markers to important QTLs. Thus identification of major QTLs is a very essential step for the improvement of grain legumes (Bocianowski, 2013) and in identifying closely linked markers to the specific trait for

marker-assisted selection and positional cloning. The present study allows the identification of important genomic regions linked to phenological and morphological traits in horsegram. Based on this genetic map and marker-trait associations, a total of 11 QTLs for these traits were identified. Four QTLs (LOD ≥2.5) for phenological traits (days to 50% flowering, reproductive period and days to maturity) and seven QTLs (LOD ≥2.5) for morphological traits (plant height, primary branches and secondary branches) were detected at different LGs across different environments (**Figure 4** and **Table 6**). The phenotypic variation explained by QTLs ranged from 6.36 to 47.53%. In this study, we reported QTLs for all traits recorded at different environments for two years at two different locations as well as based on the pooled data of all the environments. To the best of our knowledge, this study is the first comprehensive study that reports QTLs for phenological and morphological traits like days to flowering, maturity, reproductive period, number of branches in horsegram. A precise understanding of horsegram architecture and phenology through validation and utilization of the markers linked to the trait of interest in marker-assisted breeding will help in designing better breeding strategies.

Since horsegram lagged behind other grain legumes in the development of genomic resources, recent efforts for the construction of linkage map and identification of important QTLs aids in genetic upliftment of horsegram by enhancing the availability of molecular markers. To further strengthen the application of these QTLs in horsegram genomic-assisted breeding, a saturation of linkage map with more molecular markers and locating more tightly linked markers to important genomic regions should be pursued. Also, the construction of a second-generation high-density linkage map with the inclusion of SNP markers would increase the resolution of QTLs and provide a better picture of the occurrence of these QTLs for future genetic and genomic studies.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

MK performed the experiment, recorded phenotypic data, performed molecular data analysis and written the manuscript, RK recorded phenotypic data and RC planned the study and finalized manuscript.

REFERENCES

Asha, K. I., ltha, M., Abraham, Z., Jayan, P. K., Nair, M. C., and Mishra, S. K. (2006). "Genetic Resources," in *Horsegram in India*. Editor D. Kumar (Jodhpur: Scientific Publisher), 11–28.

Bahl, P. N., and Jain, H. K. (1977). Association Among Agronomic Characters and Plant Ideotype in Chickpea (Cicer Arietinum L.). *J. Plant Breed.* 79, 154–159.

Bocianowski, J. (2013). Epistasis Interaction of QTL Effects as a Genetic Parameter Influencing Estimation of the Genetic Additive Effect. *Genet. Mol. Biol.* 36, 093–100. doi:10.1590/s1415-47572013000100013

Chahota, R. K., Sharma, T. R., Sharma, S. K., Kumar, N., and Rana, J. C. (2013). "Horsegram," in *Genetic and Genomic Resources of Grain Legume Improvement*. Editors M. Singh, H. D. Upadhyaya, and I. S. Bisht (Elsevier insight), 293–305. doi:10.1016/b978-0-12-397935-3.00012-8

Chahota, R. K., Shikha, D., Rana, M., Sharma, V., Nag, A., Sharma, T. R., et al. (2017). Development and Characterization of SSR Markers to Study Genetic Diversity and Population Structure of Horsegram Germplasm (*Macrotyloma Uniflorum*). *Plant Mol Biol Rep* 35 (5), 550–561. doi:10.1007/s11105-017-1045-z

Chaitieng, B., Kaga, A., Tomooka, N., Isemura, T., Kuroda, Y., and Vaughan, D. A. (2006). Development of a Black Gram [*Vigna mungo* (L.) Hepper] Linkage Map and its Comparison with an Azuki Bean [*Vigna angularis* (Willd.) Ohwi and Ohashi] Linkage Map. *Theor Appl Genet* 113 (7), 1261–1269. doi:10.1007/s00122-006-0380-5

Chunekar, K. C., and Pandey, G. S. (1998). *Bhavaprakash Nighantu (Indian Materia Medica) of Sri Bhavamisra (C. 1500–1600 AD)*. Varanasi: Chaukhamba Bharati Academy, 984.

Cook, B. G., Pengelly, B. C., Brown, S. D., Donnelly, J. L., Eagles, D. A., Franco, M. A., et al. (2005). *Tropical Forages: An Interactive Selection Tool. Web Tool*. Brisbane, Australia: CSIRO, DPI&F (Qld), CIAT and ILRI.

Doerge, R. W., and Churchill, G. A. (1996). Permutation Tests for Multiple Loci Affecting a Quantitative Character. *Genetics* 142 (1), 285–294. doi:10.1093/genetics/142.1.285

Fuller, D. Q., and Murphy, C. (2018). The Origins and Early Dispersal of Horsegram (*Macrotyloma Uniflorum*), a Major Crop of Ancient India. *Genet Resour Crop Evol* 65 (1), 285–305. doi:10.1007/s10722-017-0532-2

Gaur, P. M., Samineni, S., Tripathi, S., Varshney, R. K., and Gowda, C. L. L. (2015). Allelic Relationships of Flowering Time Genes in Chickpea. *Euphytica* 203 (2), 295–308. doi:10.1007/s10681-014-1261-7

Grisi, M. C., Blair, M. W., Gepts, P., Brondani, C., Pereira, P. A., and Brondani, R. P. (2007). Genetic Mapping of a New Set of Microsatellite Markers in a Reference Common Bean (*Phaseolus vulgaris*) Population BAT93 X Jalo EEP558. *Genet Mol Res* 6 (3), 691–706.

Hong, Y., Chen, X., Liang, X., Liu, H., Zhou, G., Li, S., et al. (2010). A SSR-Based Composite Genetic Linkage Map for the Cultivated Peanut (*Arachis hypogaea* L.) Genome. *BMC Plant Biol* 10 (1), 1–13. doi:10.1186/1471-2229-10-17

Horton, P. (2000). Prospects for Crop Improvement through the Genetic Manipulation of Photosynthesis: Morphological and Biochemical Aspects of Light Capture. *J. Exp. Bot.* 51, 475–485. doi:10.1093/jexbot/51.suppl_1.475

Hwang, T.-Y., Sayama, T., Takahashi, M., Takada, Y., Nakamoto, Y., Funatsuki, H., et al. (2009). High-density Integrated Linkage Map Based on SSR Markers in Soybean. *DNA Res.* 16 (4), 213–225. doi:10.1093/dnares/dsp010

Jain, H. K. (1975). Breeding for Yield and Other Attributes in Grain Legumes. *Indian J Genet Plant Breed.* 35 (2), 169–187.

Kaldate, R., Rana, M., Sharma, V., Hirakawa, H., Kumar, R., Singh, G., et al. (2017). Development of Genome-wide SSR Markers in Horsegram and Their Use for Genetic Diversity and Cross-Transferability Analysis. *Mol Breed.* 37 (8), 1–10. doi:10.1007/s11032-017-0701-1

Kong, L., Lu, S., Wang, Y., Fang, C., Wang, F., Nan, H., et al. (2018). Quantitative Trait Locus Mapping of Flowering Time and Maturity in Soybean Using Next-Generation Sequencing-Based Analysis. *Front. Plant Sci.* 9, 995. doi:10.3389/fpls.2018.00995

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., et al. (2009). A Multiparent Advanced Generation Inter-cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. *Plos Genet* 5 (7), e1000551–15. doi:10.1371/journal.pgen.1000551

Mangin, B., Goffinet, B., and Rebaï, A. (1994). Constructing Confidence Intervals for QTL Location. *Genetics* 138 (4), 1301–1308. pmid: 7896108. doi:10.1093/genetics/138.4.1301

Murray, M. G., and Thompson, W. F. (1980). Rapid Isolation of High Molecular Weight Plant DNA. *Nucl Acids Res* 8 (19), 4321–4326. doi:10.1093/nar/8.19.4321

National Academy of Sciences (1978). *Moth Bean in Tropical Legumes: Resources for the Future*. Washington, DC: National Academy of Sciences.

Neelam, D. A. (2007). *Identification and Quantification of Nutraceuticals from Bengal Gram and Horse Gram Seed Coat*. [dissertation/Bachelor of Technology thesis]. India: [Department of Biotechnology] Sathyabama University Chennai.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. (1988). Resolution of Quantitative Traits into Mendelian Factors by Using a Complete Linkage Map of Restriction Fragment Length Polymorphisms. *Nature* 335 (6192), 721–726. doi:10.1038/335721a0

Radhika, P., Gowda, S. J. M., Kadoo, N. Y., Mhase, L. B., Jamadagni, B. M., Sainani, M. N., et al. (2007). Development of an Integrated Intraspecific Map of Chickpea (*Cicer arietinum* L.) Using Two Recombinant Inbred Line Populations. *Theor. Appl. Genet.* 115 (2), 209–216. doi:10.1007/s00122-007-0556-7

Ramya, M., Reddy, K. E., Sivakumar, M., Pandurangaiah, M., Nareshkumar, A., Sudhakarbabu, O., et al. (2013). Molecular Cloning, Characterization and Expression Analysis of Stress Responsive Dehydrin Genes from Drought Tolerant Horsegram [*Macrotyloma Uniflorum* (Lam.) Verdc.]. *International J. Biotechnology and Biochemistry* 9 (3), 293–312.

Ravishankar, K., and Vishnu, P. P. S. (2012). *In Vitro* antioxidant Activity of Ethanolic Seed Extracts of *Macrotyloma Uniflorum* and *Cucumis melo* for Therapeutic Potential. *International Journal on Res. Methodologies in Physics and Chemistry* 2 (2), 442–445.

Reddy, A. M., Kumar, S. G., Jyothsnakumari, G., Thimmanaik, S., and Sudhakar, C. (2005). Lead Induced Changes in Antioxidant Metabolism of Horsegram (*Macrotyloma Uniflorum* (Lam.) Verdc.) and Bengalgram (*Cicer arietinum* L.). *Chemosphere* 60 (1), 97–104. doi:10.1016/j.chemosphere.2004.11.092

Reddy, P. S., Sudhakar, C., and Veeranjaneyulu, K. (1990). Water Stress Induced Changes in Enzymes of Nitrogen Metabolism in Horsegram, *Macrotyloma Uniflorum* (Lam), Seedlings. *Indian J Exp Biol* 28 (3), 273–276.

Sari, H., Sari, D., Eker, T., and Toker, C. (2021). De Novo super-early progeny in interspecific crosses Pisum sativum L. × P. fulvum Sibth. et Sm. *Sci Rep* 11 (1), 19706–19714. doi:10.1038/s41598-021-99284-y

Sedgley, R. H., Siddique, K. H. M., and Walton, G. H. (1990). Chickpea Ideotypes for Mediterranean Environments, in Chickpea in the Nineties, Proceedings of the Second International Workshop on Chickpea Improvement. Patancheru, India, 4–8.12.1989: ICRISAT, 87–91.

Sharma, V., Rana, M., Katoch, M., Sharma, P. K., Ghani, M., Rana, J. C., et al. (2015). Development of SSR and ILP Markers in Horsegram (*Macrotyloma Uniflorum*), Their Characterization, Cross-Transferability and Relevance for Mapping. *Mol Breed.* 35 (4), 1–22. doi:10.1007/s11032-015-0297-2

Shirasawa, K., Chahota, R., Hirakawa, H., Nagano, S., Nagasaki, H., Sharma, T., et al. (2021). A Chromosome-Scale Draft Genome Sequence of Horsegram (Macrotyloma Uniflorum), *Gigabyte*, 1. 1–23. doi:10.46471/gigabyte.30

Sodani, S. N., Paliwal, R. V., and Jain, L. (2004). *Phenotypic Stability for Seed Yield in Rainfed Horsegram (*Macrotyloma uniflorum *[Lam.] Verdc)*. Paper presented in National Symposium on Arid Legumes for Sustainable Agriculture and Trade. Jodhpur, 5–7.11.2004 (Central Arid Zone Research Institute).

Sudhakar, C., Syamalabai, L., and Veeranjaneyulu, K. (1992). Lead Tolerance of Certain Legume Species Grown on lead Ore Tailings. *Agric. Ecosysy. Environ.* 41 (3-4), 253–261. doi:10.1016/0167-8809(92)90114-Q

Van Ooijen, J. W. (2006). JoinMap® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations. *Kyazma BV, Wageningen* 33 (10), 1371.

Vavilov, N. I. (1951). "The Origin, Variation, Immunity and Breeding of Cultivated Plants," in Phytogeographic Basis of Plant Breeding. *Chronica Botanica* 13, 1–366.

Voorrips, R. E. (2002). MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *J Hered.* 93, 77–78. doi:10.1093/jhered/93.1.77

Wang, S., Basten, C. J., and Zeng, Z. B. (2005). *Windows QTL Cartographer 2.5 Department of Statistics*. Raleigh, NC: North Carolina State University.

Wu, R. L. (1998). Genetic Mapping of QTLs Affecting Tree Growth and Architecture in Populus: Implication for Ideotype Breeding. *Theor Appl Genet* 96 (3-4), 447–457. doi:10.1007/s001220050761

Yadava, N. D., and Vyas, N. L. (1994). *Arid Legumes*. India: Agro Botanical Publ.

Zeng, Z. B. (1994). Precision Mapping of Quantitative Trait Loci. *Genetics* 136 (4), 1457–1468. doi:10.1093/genetics/136.4.1457

Zeng, Z. B. (1993). Theoretical Basis for Separation of Multiple Linked Gene Effects in Mapping Quantitative Trait Loci. *Pnas* 90 (23), 10972–10976. doi:10.1073/pnas.90.23.10972

Zohary, D. (1970). "Centres of Diversity and Centres of Origin," in *Genetic Resources in Plants- Their Exploration and Conservation*. Editors O. H. Frankel and E. Bennett (Blackwell: Oxford), 33–42.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2022
doi: 10.3389/fgene.2022.762202

Genetic Diversity and Population Structure of Doum Palm (*Hyphaene compressa*) Using Genotyping by Sequencing

Agnes Omire[1], *Johnstone Neondo*[2], *Nancy L. M. Budambula*[3], *Laura Wangai*[4], *Stephen Ogada*[2] *and Cecilia Mweu*[2]*

[1]Department of Botany, School of Biological Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya, [2]Institute for Biotechnology Research (IBR), Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya, [3]Department of Biological Sciences, School of Pure and Applied Sciences, University of Embu, Embu, Kenya, [4]Department of Biomedical Sciences, School of Health Sciences, Kirinyaga University, Kerugoya, Kenya

OPEN ACCESS

Edited by:
Rana Dajani,
Hashemite University, Jordan

Reviewed by:
Lizandra Jaqueline Robe,
Federal University of Santa Maria, Brazil
Mukesh Choudhary,
ICAR-Indian Institute of Maize Research, India
Amol N. Nankar,
Center of Plant Systems Biology and Biotechnology, Bulgaria
Aleksandra Dimitrijevic,
Institute of Field and Vegetable Crops, Serbia

***Correspondence:**
Cecilia Mweu
cmweu@jkuat.ac.ke

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *21 August 2021*
Accepted: *03 January 2022*
Published: *04 February 2022*

Citation:
*Omire A, Neondo J, Budambula NLM, Wangai L, Ogada S and Mweu C (2022) Genetic Diversity and Population Structure of Doum Palm (*Hyphaene compressa*) Using Genotyping by Sequencing. Front. Genet. 13:762202. doi: 10.3389/fgene.2022.762202*

Doum palm (*Hyphaene compressa*) is a perennial economic plant primarily growing in Kenya's Arid and Semi-Arid Lands (ASALs). It is heavily relied upon for food, animal feed, construction materials and medicine, making it an ideal plant for resource sustainability. However, the limited information on its genetic resources has hindered its breeding and conservation studies. This study used the genotyping by sequencing approach to identify Single Nucleotide Polymorphisms. These SNPs were further used to assess the genetic diversity and population structure of 96 *H. compressa* accessions from Coastal, Northern and Eastern ASAL regions of Kenya using two approaches; reference-based and *de novo*-based assemblies. STRUCTURE analysis grouped the sampled accessions into two genetic clusters (Cluster 1 and Cluster 2). Cluster 1 included accessions from the Northern region, whereas Cluster 2 included all accessions from Eastern and Coastal regions. Accessions from Kwale (Coastal) had mixed ancestry from both Cluster 1 and Cluster 2. These STRUCTURE findings were further supported by principal components analysis, discriminant analysis of principal components and phylogenetic analysis. Analysis of molecular variance indicated greater genetic variation within populations (92.7%) than among populations (7.3%). An overall F_{ST} of 0.074 was observed, signifying moderate genetic differentiation among populations. The results of this study will provide information useful in breeding, marker-assisted selection and conservation management of *H. compressa.*

Keywords: genetic diversity, GBS, single nucleotide polymorphisms, population structure, *Hyphaene compressa*, doum palm

INTRODUCTION

Hyphaene compressa H. Wendl., also known as the East African doum palm, is a member of the Arecaceae family. It is integral in the agroforestry system of coastal and riverine parts of Africa (Amwatta, 2004; Uhl and Moore, 2019). The doum palm also grows in the arid and semi-arid lands (ASALs) of Kenya (Maundu and Tengnas, 2005). It is a valuable source of food, animal feed, medicine for headaches and worms, as well as non-wood products for construction and weaving.

Thus, it is a substantial income-generating plant, particularly for women in ASALs who derive their livelihoods from the sale of woven leaf products (Amwatta, 2004; Maundu and Tengnas, 2005; Omire et al., 2020a). The ability to withstand waterlogging, drought and salinity makes the doum palm a reliable economic plant with ability to avert natural calamities including drought in such areas (Amwatta, 2004; Omire et al., 2020a).

In Kenyan ASALs, non-timber products are restricted to a few plant species such as *H. compressa*, subsequently threatening its existence. Thus, *H. compressa* is classified as a threatened and a national priority species in the ASALs of Kenya with a high potential for genetic erosion due to overexploitation by the rural communities (Kigomo, 2001). However, the International Union for Conservation of Nature (IUCN) considers it a species of least concern with an unknown population trend due to its wide geographical distribution throughout East Africa (Cosiaux et al., 2017). Whereas species might exist as large populations they could be regionally threatened. *Hyphaene. compressa* resources are known to be strained and overexploited in the Eastern part of Kenya (Omire et al., 2020a). Despite this knowledge on the status, there are no known interventions to reverse the current trend (Kigomo, 2001). This could exacerbate the risk of extinction of such species (Cosiaux et al., 2018).

Threats to *H. compressa* include human interference as well as biotic and abiotic stress (Omire et al., 2020a). Overgrazing by pastoralist communities, particularly along the riverine areas, is a significant threat to this palm since livestock graze and browse on *H. compressa* (Kigomo, 2001). The strain on *H. compressa* resources has been aggravated by the sedentarization of the nomadic pastoralists (Amwatta, 2004). Sedentarization leads to the assemblage of pastoralists around limited resources and ultimately to land degradation (Johnson, 1993). Another source of pressure on *H. compressa* is overharvesting and harvesting of immature sword leaves. These leaf pressures have been shown to cause dwarfing in a sister palm, *Hyphaene thebaica* (Kahn and Luxereau, 2008). Other selection pressures on *H. compressa* include logging, burning and wine tapping from the apical meristem. These pressures collectively lead to genetic drift through the loss of specific genotypes, which might eventually affect the *H. compressa* gene pool (Kigomo, 2001).

There is scanty information on the genetic diversity of doum palm which limits access to its important traits and thus hinders its improvement. Previous diversity studies on *H. compressa* focussed on accessing the variability of its morphological traits (Omire et al., 2020b). The study identified five morphotypes with accessions from the Kenyan Coastal area being the most heterogeneous. However, this cannot be used to adequately delineate the doum palm since morphological markers may also be affected by the environment, are limited in number, unstable, slow and some appear late in plant development making them difficult to score (Andersen and Lubberstedt, 2003; Mokhtar et al., 2016). Furthermore, using a single marker like morphology is not adequate to assess diversity (Khan et al., 2012). Overall, genetic markers are superior to morphological markers (Ganie et al., 2015) and may or may not agree with phenotypic expression of a genomic trait.

For non-model plants like doum palm with no reference genome, sequencing the whole genome to mine the SNPs would be ideal. There are other methods like Genotyping by Sequencing (GBS) that are able to acquire in depth data on parts of the genome and are as effective but less costly compared to whole genome sequencing (Wallace and Mitchell, 2017). GBS is a method that provides reduced libraries for Illumina next generation sequencing (NGS) by targeting the subsets with restriction enzymes followed by ligation of DNA barcoded adapters (Elshire et al., 2011). PCR amplification and high throughput NGS of the genomic subsets on a single lane of flow cells is then done (Elshire et al., 2011; He et al., 2014; Burghardt, et al., 2017). GBS is simple, rapid and highly reproducible (Davey et al., 2011; Burghardt et al., 2017). These features make GBS highly attractive for several genetic applications, including genetic diversity, phylogeny, genome-wide association studies, association maps, genomic selection, physical and linkage maps (Burghardt et al., 2017). GBS is an ideal tool for genetic diversity studies with the advantage of being able to identify SNPs, insertions, deletions and microsatellites (Elshire et al., 2011) even in non-model organisms with no prior genome information (Taranto et al., 2016). Diversity studies can be combined with phylogenetic studies to provide more information on the origin and domestication of the germplasm for conservation purposes (Burghardt et al., 2017). Earlier studies have alluded to the fact that the evolution of a population is guided by its local interactions in the environment (Klimova et al., 2018). Gene flow has a tendency to homogenize populations and reduce genetic variability (Brunet et al., 2012). However, there needs to be enough gene pool on which selection can take place for effective speciation.

Thus far, the genetic diversity of *H. compressa* remains unknown despite its economic and subsistence role in Africa's ASALs. There were no *H. compressa* or other palms in the genus *Hyphaene* with assembled genomes at the time of this study. Due to the scanty genetic information coupled with the pressure already demonstrated on this palm, there is a need to decipher its genetics. This study assumes that the different accessions of *H. compressa* growing in Kenya are diverse. The present study aimed to identify genome-wide SNPs, assess the level of genetic diversity, determine the population structure and estimate gene flow between *H. compressa* accessions collected from four ASAL regions of Kenya using GBS approach. The data obtained from this study will be an important genomic resource that will be used to inform the conservation, management, breeding and propagation of *H. compressa*.

MATERIALS AND METHODS

H. compressa Plant Materials

A total of 96 *H. compressa* accessions collected between February and August of 2018 were used in this study. These accessions were collected from different ASALs of Kenya; Eastern (Tharaka Nithi County), Northern (Turkana County), Coastal (Kwale and Tana River counties) as shown in **Figure 1** and **Supplementary Table S1**. Leaf samples of the selected plants were collected using sterile

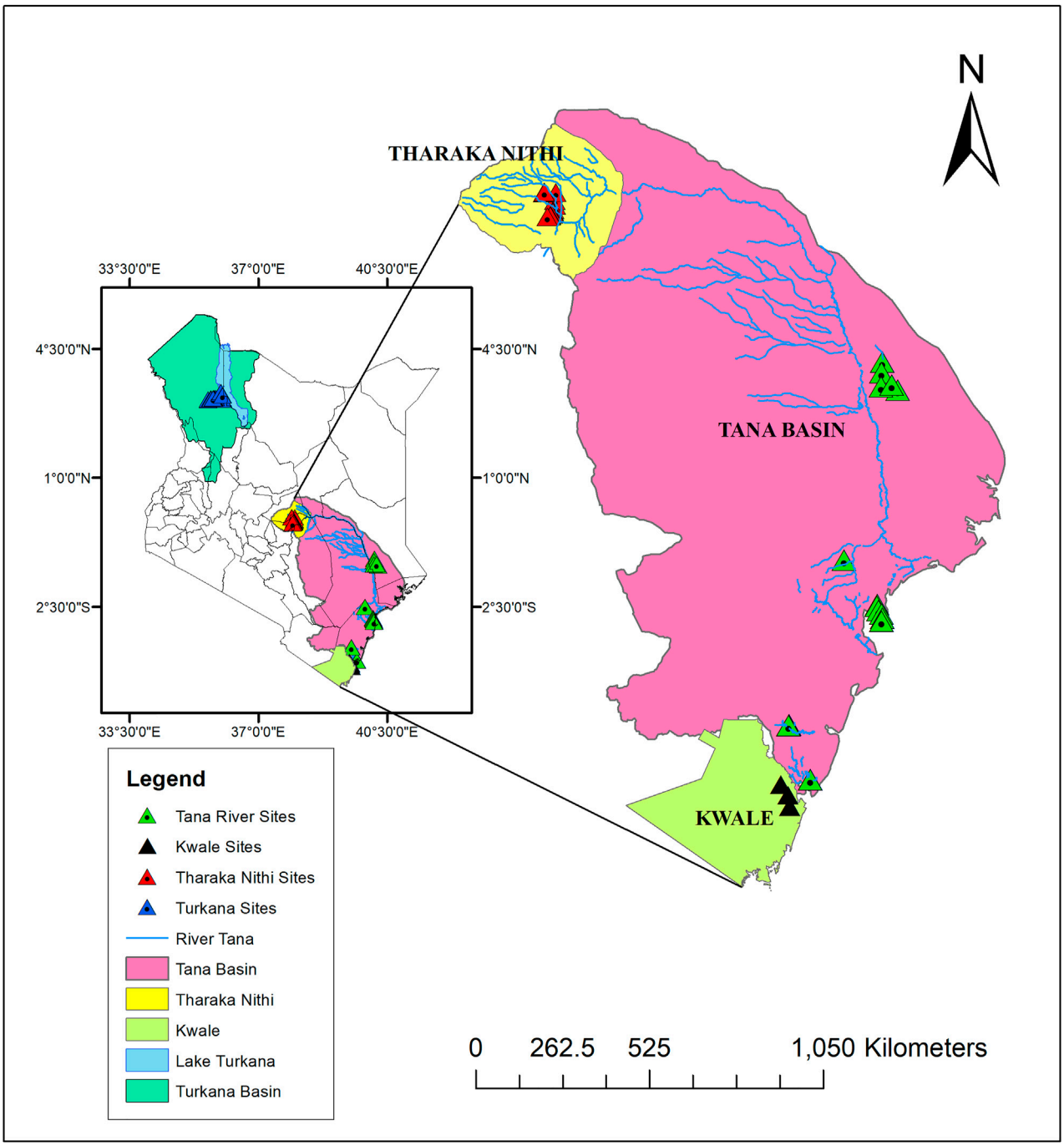

FIGURE 1 | Map illustrating *H. compressa* sampling points along the Turkana and Tana basins. Map was created using ArcMap version 10.8.

blades and placed in sterile tubes containing 10 g of silica gel for DNA extraction. Accessions within approximately the same age group and located as distantly as possible from each other were sampled.

Preparation of Libraries and Sequencing

Genomic DNA was isolated from *H. compressa* leaves using DNeasy® Plant Mini Kit (Qiagen, Germany). The purity and quantity of the DNA were determined using Qubit Fluorometer (Invitrogen) or microplate reader (DR-200B, Diatek), while a 1% agarose gel was used to confirm its integrity. Commercial GBS sequencing was done at Beijing Genomics Institute (BGI, China). A total of 96 samples; Tharaka (27), Turkana (21), Tana River (20) and Kwale (28), passed the sample quality check (QC) and proceeded to library preparation.

Library preparation was done following the method previously described by Elshire et al. (2011). Essentially, the DNA samples were barcoded and adapter pairs plated. The restriction enzyme ApeK1 (GCWGC as the recognition site) was used, followed by adapter ligation to the DNA fragments. This was followed by pooling and purification. PCR was then performed using primers with adapter binding sites. Sample clean-up of the PCR products, fragment size selection and sequencing on a Hiseq X10 platform as paired-end 100 bp (Illumina PE 100) was done. Adapter sequences, sequences with low-quality reads, and those lacking barcodes were discarded from the raw reads.

The data was processed using the *de_novo* and reference-based approaches. In the *de_novo* assembly, ipyrad version 0.9.74 (Eaton and Overcast, 2020) was used to assemble sequences without a reference genome using the following parameters; assembly method *de_novo*, datatype pairgbs, mindepth_statistical 6, mindepth_majrule 6, min_samples_locus 4 and other parameters set to default. In the reference-based approach, paired read ends were mapped to the *Phoenix dactylifera* (date palm) genome (Hazzouri et al., 2019). A confamilial (same family) reference genome was used for SNP calling (Galla et al., 2019) since *H. compressa* had no assembled genome at the time of this study. These two palms belong to the same subfamily Coryphoideae. *Phoenix dactylifera* genome was the only available genome in this subfamily. Alignment of the sequence reads against the date palm reference genome was done using the Burrows-Wheeler Aligner (BWA) using the parameters 'mem–t4 –k32 –M' (Li and Durbin, 2009). SNP filtering was done using VCFtools version 0.1.16 (Danecek et al., 2011) with the following parameters; biallelic SNPs, min meanDp 2, removing indels, Minor Allele Frequency (MAF) 0.05, minDP 2, max-missing 0.8.

Data Analysis

The quality of the filtered VCF files were assessed using the R package tidyverse (Wickham et al., 2019). The read depth per site, heterozygosity, read depth per individual and read quality were determined.

Population Structure Analysis and Genetic Diversity

Population structure was determined by STRUCTURE software version 2.3.4 using the admixture model (Pritchard et al., 2000). Populations of K (K = 1–10) were run with three replications using a burn-in of 100, 000 generations and 100, 000 Markov Chain Monte Carlo (MCMC) iterations. The software STRUCTURE HARVESTER web Version 0.6.94 (Earl and VonHoldt, 2012) available at http://taylor0.biology.ucla.edu/structureHarvester/ was used to determine the optimal K value using the *adhoc delta K* (Evanno et al., 2005). To plot the structure results, the POPHELPER version 2.3.1 package in R was used (Francis, 2017). Genotypes that had ≥0.80 membership proportion and those with less than this value were assigned to pure and admixture populations, respectively (Nkhoma et al., 2020).

Discriminant Analysis of Principal Components (DAPC) was also used to evaluate the population structure of *H. compressa* using the package adegenet version 2.1.3 (Jombart, 2008) in R. To visualize each sample's assignment, a composite stacked bar plot illustrating the probability of population membership on the *Y*-axis was generated. Principal component analysis (PCA) was constructed using the R software package SNPrelate (Zheng et al., 2012) to determine the genetic relationships of *H. compressa* accessions.

Observed heterozygosity (H_o), Expected heterozygosity (H_e) fixation index (F_{ST}), inbreeding coefficient (F_{IS}), Analysis of Molecular Variance (AMOVA) and pairwise F_{ST} values of the population were determined using Arlequin version 3.5.2.2 (Excoffier and Lischer, 2010).

Phylogenetic Analysis

To construct a splitstree, the filtered VCF file was converted to a nexus file using vcf2phylip.py script (Ortiz, 2019). The nexus file was then used to generate an unrooted splitstree using the neighbor net method in SplitsTree software, version 4.17.0 (Huson and Bryant, 2006). An UPGMA distance tree was also constructed using R software to represent the genetic clustering of *H. compressa* accessions.

Migration Rates Between the Eastern and Coastal Populations Along the Tana Basin

To determine if the population structure observed along the Tana basin is influenced by seed dispersal along the river, gene flow was estimated using MIGRATE-n software version 3.6.11. A Bayesian inference strategy was used with constant mutation rates among all loci. Burn in was set at 5,000 iterations at each locus. Static heating at four different temperatures (1, 1.5, 3 and 6) was used to improve the MCMC searches. One gene flow model was designed with direct migration from Tharaka to Tana River and to Kwale. The drainage of the Tana basin was used to design this model. Turkana accessions were excluded from this model since structure analysis and PCA demonstrated little historical gene flow. To judge whether the runs converged on good answers, the histograms and the effective population sizes were checked.

RESULTS

Paired-end sequencing of 96 *H. compressa* accessions yielded an average of 2.4 million reads per sample. The *de_novo*-based assembly using ipyrad software resulted in 3,941 raw loci. After filtering, a total of 2,096 SNPs with a mean depth of 35.7 (minimum 10.47, maximum 217.45) were retained using the *de_novo* based assembly. On the other hand, reference-based assembly using date palm as a reference genome resulted in 3.4 million loci. After filtering, 23,146 biallelic SNPs with a mean depth of 3.5 (minimum 2, maximum 47.49) were obtained using the reference-based assembly.

The SNPs obtained from *de_novo* based assembly had higher depths than the reference-based assembly, as shown by the individual sequencing depth and the mean depth. The proportion of missing data per accession was low for both the *de_novo* based and reference-based assemblies, with a maximum

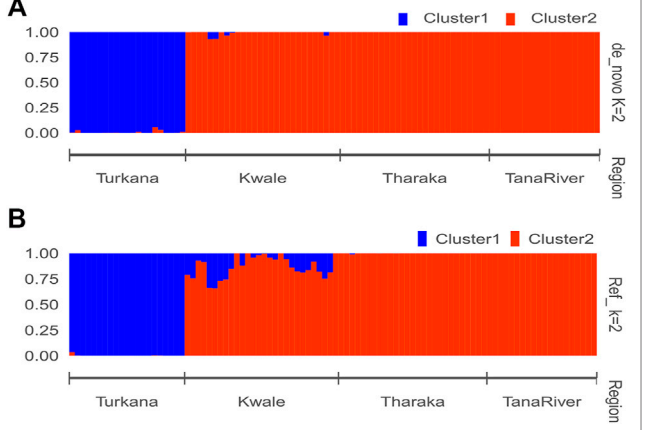

FIGURE 2 | STRUCTURE bar plot of admixture model population assignment of 96 *Hyphaene compressa* accessions from Kenya based on 2,096 SNPs for the *de_novo* assembly **(A)** and 23,416 SNPs based on reference assembly **(B)**. The accessions are divided into two clusters. A combination of different colors represents admixed populations.

of 0.04 and 0.4, respectively. These VCF quality statistics, including the mean depth, observed heterozygosity, depth per individual and missing data per individual, are presented for both the *de_novo* based assembly (**Supplementary Figure S1**) and reference-based assembly (**Supplementary Figure S2**).

In the *de_novo* based assembly there were 1,283 (61.2%) transition SNPs and 813 (38.8%) transversion SNPs with the following types: A↔G type (651, 31.1%), C↔T type (632, 30.2%), A↔C type (192, 9.2%), A↔T type (174, 8.3%), C↔G type (222, 10.6%), G↔T type (225, 10.7%). There were 16,598 (70.9%) transition SNPs and 6,818 (29.1%) transversion SNPs with the following types: A↔G type (8,332, 35.6%), C↔T type (8,266, 35.3%), A↔C type (1,684, 7.2%), A↔T type (1825, 7.8%), C↔G type (1,636, 7%), and G↔T type (1,673, 7.1%) in the reference based assembly. The A↔G and C↔T transition SNPs had the highest counts for both assemblies (**Supplementary Table S2**). The transition SNPs versus transversion SNPs (Ts/Tv) ratio was 1.6 in the *de_novo* based assembly and 2.4 in the reference based assembly.

Population Structure and Genetic Diversity

The optimal delta *K* was detected at $K = 2$ for both the *de_novo*-based assembly (**Supplementary Figure S3A**) and reference-based assembly (**Supplementary Figure S3B**), which inferred two genetic clusters of *H. compressa* (**Figure 2**). Cluster 1 consisted of accessions from Turkana while cluster 2 consisted of accessions from Tharaka, Kwale and Tana River for both reference-based and *de_novo*-based assemblies. The accessions in Cluster 2 were sampled along the Tana basin and Kwale county as shown in **Figure 1**. The expected heterozygosity was lower for the *de_novo*-based assembly for cluster 1 ($H_e = 0.14$) than cluster 2 ($H_e = 0.23$). However, similar expected heterozygosity values were observed in the reference-based assembly for the two clusters ($H_e = 0.30$). The genetic variation among populations in Cluster 1 was higher (*de_novo* $F_{ST} = 0.68$ and reference-based $F_{ST} = 0.17$) than Cluster 2 (*de_novo* $F_{ST} = 0.3$ and Reference-based $F_{ST} = 0.06$). A total of seven accessions from Kwale had admixed ancestry between Cluster 1 and Cluster 2 using the reference based assembly (**Table 1**). There were no accessions in the *de_novo* assembly that had admixed ancestry values less than 88%. Structure results indicate that two gene pools best describe the population structure of *H. compressa*. However, a smaller peak observed at $K = 3$ might be another informative *H. compressa* population clustering (**Supplementary Figures S3, S4**). Similar grouping of accessions using PCA for the *de_novo* and reference-based assemblies was observed in this study (**Figure 3**). In both PCA plots, Tharaka Nithi, Kwale and Tana River accessions were closely clustered. PCA results were congruent with structure results.

DAPC analysis grouped *H. compressa* accessions into two clusters, with samples from Turkana falling to the right side of the DAPC vertical axis while the rest fell on the left side. There was moderate overlap between Kwale and Tana River accessions, while Tharaka accessions were distinctly separate (**Figure 4**). DAPC analysis, composite plot and genetic diversity results are shown for only the reference based assembly SNP data since both assembly methods had shown congruent results in structure and PCA analysis. Population membership assignment using DAPC composite plot confirmed structure and PCA results. All the accessions along the Tana basin exhibited admixture with profoundly shared ancestry between Tharaka and Tana River. Kwale had the highest level of admixture (**Figure 5**). DAPC results also confirmed no admixture between Turkana and accessions from the other regions.

The four sampled regions of Tharaka, Tana River, Kwale and Turkana were assessed for the number of polymorphic sites, expected heterozygosity (H_e) or gene diversity, observed heterozygosity (H_o), F_{IS} and F_{ST}. The genetic variation among the *H. compressa* populations was moderate ($F_{ST} = 0.074$, $p \leq 0.001$). The observed heterozygosity was higher than the expected heterozygosity in all the populations (**Table 2**). Negative F_{IS} values were obtained in all the populations, with Turkana having the lowest F_{IS} value (−0.45). Kwale had the highest polymorphic sites (11,932) followed by Turkana (10,698) as shown in **Table 2**. Tana River had the lowest diversity ($H_e = 0.23$, Polymorphic sites = 8,370) of all the sampled regions (**Table 2**).

Pairwise F_{ST} values ranged between 0.025 (Tharaka and Tana River) and 0.105 (Turkana and Tana River). High pairwise F_{ST} was recorded among Turkana and Tana River samples (**Table 3**). The lowest pairwise F_{ST} was between Tharaka and Tana River, suggesting gene flow between the two regions. AMOVA showed that populations from each of the four regions of Turkana, Tharaka, Tana River and Kwale were slightly different from each other ($p \leq 0.001$, **Table 4**). The variation within populations was higher (92.7%) than among populations (7.3%).

Phylogenetic Analysis

The neighbor net network was able to group *H. compressa* accessions by region. Tana River, Kwale and Tharaka samples clustered closely compared to the Turkana, which was separated from the rest (**Figure 6**). Samples from Turkana clustered together.

TABLE 1 | STRUCTURE analysis of *Hyphaene compressa* from Kenya using reference-based and *de_novo* based assembly.

Assembly method	Tharaka	Tana river	Kwale	Turkana	*He*	F_{ST}
Reference assembly	—	—	—			
Cluster 1				21	0.29	0.17
Cluster 2	27	20	21	—	0.30	0.06
Admixed	—	—	7	—	—	—
***De_novo* assembly**						
Cluster 1	—	—	—	21	0.14	0.68
Cluster 2	27	20	28	—	0.23	0.30

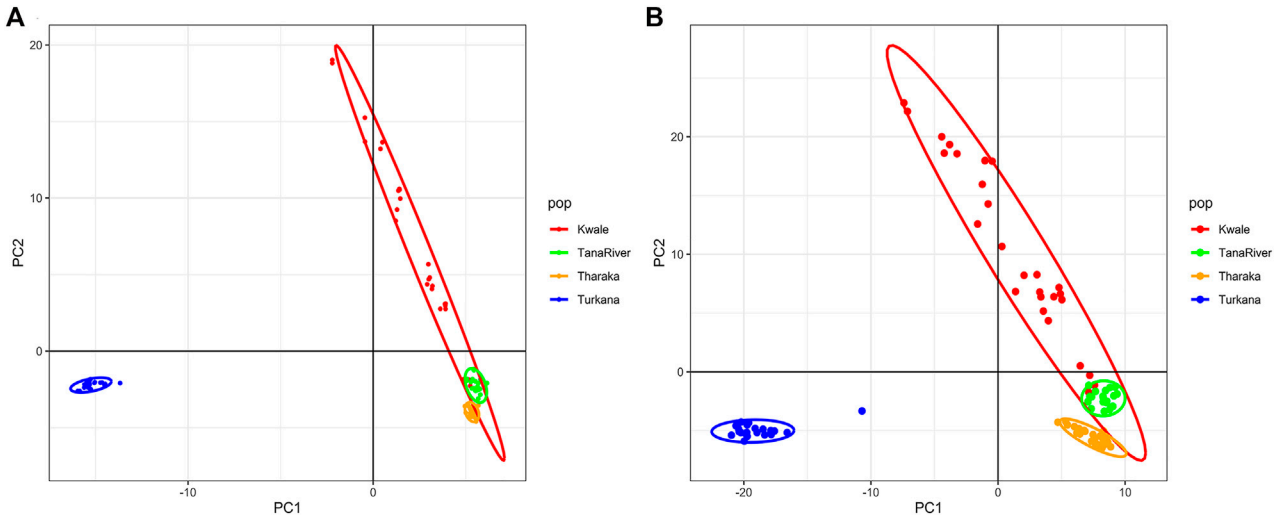

FIGURE 3 | Principal Component Analysis (PCA) of *Hyphaene compressa* Kenyan accessions using 2096 SNPs obtained from the *de novo* assembly **(A)** and 23,416 SNPs obtained from the date palm reference based assembly **(B)**.

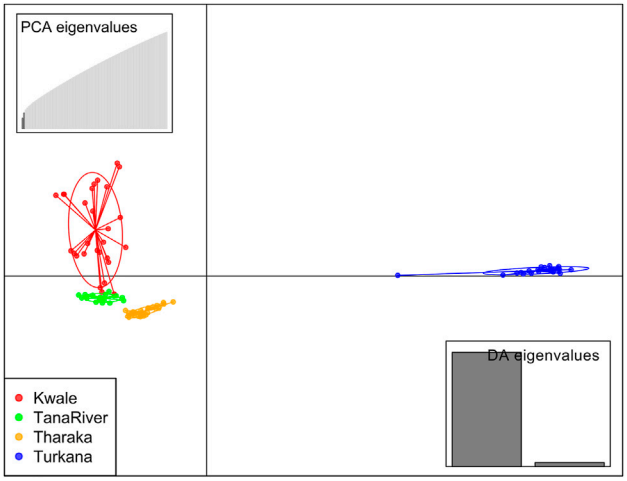

FIGURE 4 | Discriminant analysis of principal components (DAPC) of 96 *Hyphaene compressa* Kenyan accessions using 23,416 SNPs derived from GBS analysis using reference-based assembly. Inertia ellipses represent clusters in different colors. Each dot on the circle represents an accession.

Some of the Kwale accessions clustered closely with Tana River while other Kwale accessions clustered more closely with Tharaka accessions. The UPGMA phylogenetic tree showed two main clusters with Turkana accessions clustering in one cluster and the rest of the accessions in the other cluster. Kwale populations also revealed reciprocal monophyly (**Supplementary Figure S5**).

Migration Rates Among *H. compressa* Accessions Along the Tana Basin

The highest gene flow was observed from Tharaka and Tana River samples (m = 139.1), followed by Kwale to Tharaka (102.7), Tana River to Kwale (63.1), Tana River to Tharaka (59.9), Tharaka to Kwale (50.3) and Kwale to Tana River (57.7). These results indicate that the gene flow along the Tana basin was mostly asymmetrical (**Supplementary Figure S6**).

DISCUSSION

This study is the first to report the use of SNPs through the GBS approach to characterize *H. compressa* accessions in Kenya. SNP markers are very stable, frequent and specific to regions of the

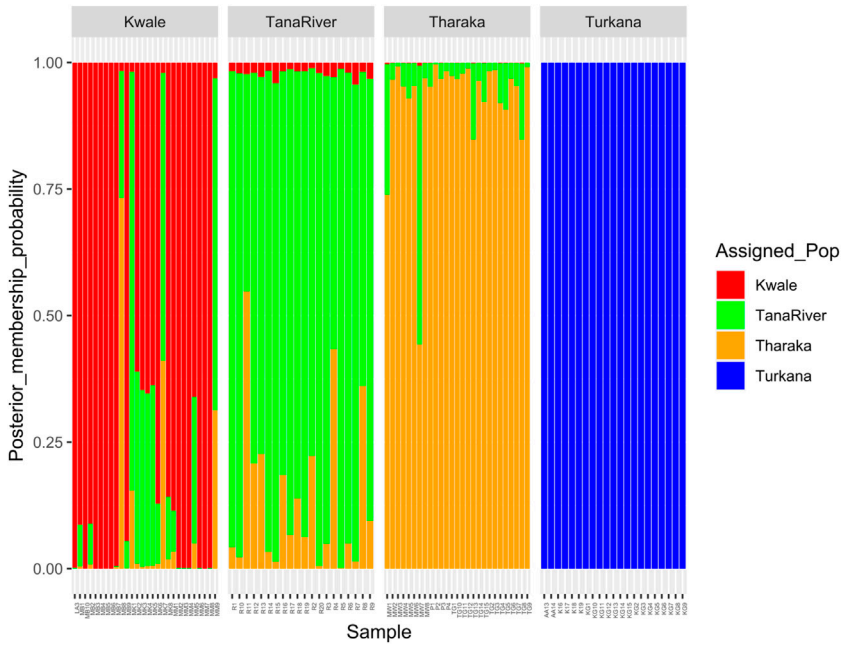

FIGURE 5 | Composite plot of *Hyphaene compressa* Kenyan accessions using reference-based assembly showing mixed ancestry between Kwale, Tana River and Tharaka. Each accession is a stacked bar chart with populations shown in colors.

TABLE 2 | Mean values of genetic diversity indices determined for *Hyphaene compressa* accessions in the sampled populations.

Genetic index	Region				Overall
	Tharaka	Turkana	Tana river	Kwale	
Number of polymorphic sites	9,277	10,698	8,370	11,932	23,416
Observed heterozygosity (H_o)	0.45	0.47	0.46	0.44	0.404
Expected heterozygosity (H_e)	0.32	0.33	0.23	0.33	0.31
F_{IS}	−0.40	−0.45	−0.42	−0.37	−0.040
F_{ST}					0.074

TABLE 3 | Pairwise F_{ST} values of Kenyan populations of *Hyphaene compressa*.

Population	Turkana	Kwale	Tharaka	Tana river
Turkana				
Kwale	0.07952			
Tharaka	0.09795	0.03629		
Tana River	0.10541	0.03329	0.02505	0.00

genome which makes them ideal for use in marker assisted selection (MAS) and diversity studies to aid future germplasm conservation.

Two comparative methods (reference-based and *de_novo*-based approaches) were used to infer the population structure and genetic diversity of *H. compressa*. In the reference-based assembly, *Phoenix dactylifera* was used as a reference genome. This confamilial genome was used because *H. compressa* had no assembled genome at the time of this study. In the absence of a reference genome of the same species (conspecific) or genus (congeneric), a confamilial reference genome can be used to provide similar estimates of diversity (Brandies et al., 2019; Galla et al., 2019). Galla et al. (2019) further recommends using a confamilial reference genome as the most distant genome ideal for diversity studies. There were differences in the two methods concerning abundance, quality scores and the TS/TV ratios of the SNPs obtained. For example, the highest number of SNPs was observed from the reference-based assembly (23,416) compared to the *de_novo*-based assembly (2096). The reference-based assembly has also been previously demonstrated to outperform *de_novo* assembly in determining the number of SNPs in olive cultivars (D'Agostino et al., 2018). Elsewhere, it has been reported that parameters set during assembly and the type of assembly influence the number and depth of SNPs obtained (Bohling, 2020). Besides, more stringent parameters are normally used for *de_novo* assemblies. GBS of *H. compressa* accessions showed considerable SNP variations with transition SNPs (purine-purine or pyrimidine-pyrimidine) being the most frequent

TABLE 4 | Analysis of molecular variance (AMOVA) in genetic variation of *Hyphaene compressa* among and within populations of Kwale, Tharaka, Turkana and Tana River counties in Kenya.

Source of variation	Sum of squares	Variance components	Percentage variation	*p*-value
Among populations	42728.129	266.91865	7.32597	0.00
Within populations	558905.557	3376.54026	92.67403	
Total	**601633.686**	**3643.45891**	**100**	

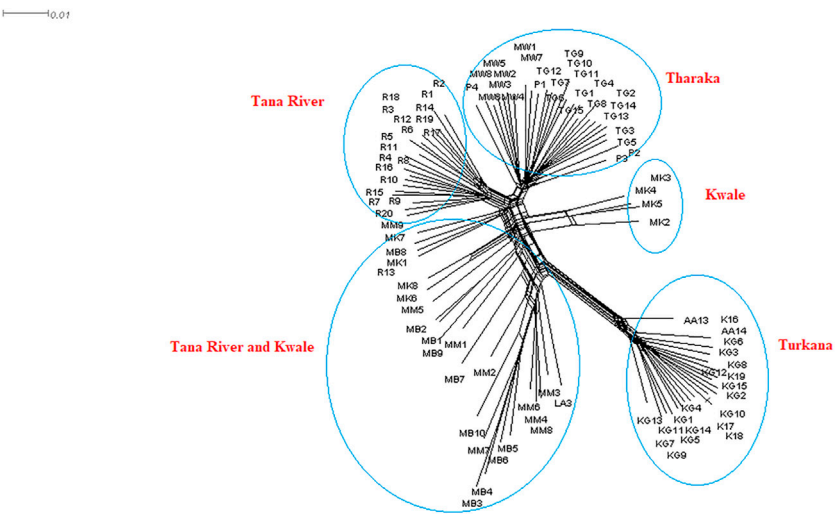

FIGURE 6 | NeighbourNet network generated using 23,416 SNPs derived from reference-based GBS analysis of Kenyan *Hyphaene compressa* using SplitsTree version 4.17.0.

mutations. This high frequency in transition SNPs has also been observed in many plants such as *Capsicum annuum*, *Vigna unguiculata*, *Elaeis guineensis*, and *Camelia sativa* (Pootakham et al., 2015; Taranto et al., 2016; Xiong et al., 2016; Luo et al., 2019; Hyun et al., 2020), of which the C-T transitions are the most frequent (Edwards et al., 2007). However, low TS/TV ratio was observed in *de_novo* based assembly. This phenomenon has also been observed elsewhere and was attributed to the differences in the SNP calling methods (Shafer et al., 2017). Despite the difference between the two methods, structure analysis and PCA produced congruent results. A similar study that used a *de_novo* approach, a confamilial reference and a congeneric reference to determine the phylogenetic relationship of the *Amaranthus* genus, produced differing SNP counts but similar phylogenetic trees (Stetter and Schmid, 2017). Other studies have also reported differing SNP abundance between reference-based and *de_novo* assemblies but consistent population clustering (D'Agostino et al., 2018; Shafer et al., 2017; Stetter and Schmid, 2017).

STRUCTURE analysis grouped the *H. compressa* accessions into two gene pools. PCA and DAPC were consistent with these structure results. Accessions from Turkana (Northern Kenya) were pooled into cluster 1. Cluster 2 had the highest number of genotypes that included accessions collected along the Tana basin sites (Tharaka Nithi and Tana River) and Kwale. In addition, an admixture of accessions was reported for accessions from Kwale that showed mixed ancestry from Tharaka Nithi and Tana River. This admixture may be due to genetic exchange between Kwale, Tharaka Nithi and Tana river. These accessions also exhibited reciprocal monophyly based on phylogenetic analysis. This corroborates the phenotypic diversity studies on *H. compressa* where accessions from Kwale exhibited the highest phenotypic diversity compared to other regions (Omire et al., 2020b).

The fixation index (F_{ST}) is an informative method for measuring population differentiation among populations (Nassiry et al., 2009). An overall F_{ST} value of 0.074 was reported for *H. compressa* populations in this study. According to Nassiry et al. (2009), an F_{ST} of 0–0.05 is considered small, 0.05–0.15 is moderate and 0.15 and above is considered very high. Therefore, *H. compressa* accessions had moderate genetic differentiation. However, F_{ST}, values obtained from STRUCTURE population clustering indicate high genetic differentiation within cluster 1 (accessions from the northern part of Kenya) than cluster 2 (accessions along the Tana Basin). On the other hand, cluster two had higher levels of heterozygosity (*He*) compared to cluster 1. This indicates high diversity in accessions along the Tana basin. The high heterozygosity and lower genetic differentiation between *H. compressa* accessions along the Tana Basin (cluster 2) might

be due to genetic exchange arising from gene flow. Although Tharaka Nithi is found approximately 163 and 391 miles away from Tana River and Kwale respectively, gene flow between these three counties seems high. This could be due to the flow of the River Tana (**Figure 1**), which traverses both Tharaka and Tana River counties and possibly serves as means of germplasm dispersion. This could explain why Tharaka samples are close to the Tana River accessions on the PCA and the high mixed ancestry as demonstrated by DAPC and STRUCTURE analysis. River Tana is the longest river and the most important drainage basin in Kenya. The river drains from the Kenyan highlands to the Eastern ASAL plateaus and coastal Kenya (Kitheka and Ongwenyi, 2002). Since *H. compressa* grows in riverine areas, seed dispersal through the river is an important factor influencing the population structure of *H. compressa* at the Kenyan Coast. Seed dispersal is essential for biodiversity conservation by driving plant gene flow, population dynamics and functional connectivity between regions (Traveset and Rodríguez-Pérez, 2018). Systematic seed dispersal favours gene flow, increases genetic diversity and lowers the genetic differentiation among populations (Paschoa et al., 2018). Migration rates using MIGRATE-n indicate that there is asymmetrical gene flow along the Tana basin. This supports the hypothesis that seed dispersal along the Tana River drives the population structure of *H. compressa* along the Coast. In addition, high migration rates were observed between Kwale and Tharaka an observation that is confirmed by phylogenetic analysis whereby some Kwale accessions clustered with Tharaka accessions.

There is restricted gene flow into or out of Turkana, which may cause differentiation of its population from the other populations. This was supported by STRUCTURE analysis, PCA, DAPC and neighbor net network, which clustered Turkana distinctly from the rest of the populations. This differentiation may be attributed to the physical distance between Turkana and the other populations. Isolation of Turkana populations inhibits them from mating with the other populations. Turkana is found in the far-flung northern part of Kenya and is considered 100% dryland with scarce rain fed agriculture (Barrow and Mogaka, 2007). In addition, the selection pressures in Turkana differ from those present in the other regions.

The negative F_{IS} values obtained for *H. compressa* populations indicate low levels of inbreeding, high diversity and moderate connectivity between the populations. This may be influenced by the mating system. *Hyphaene compressa* is a dioecious plant (Stauffer et al., 2014), a condition that favors obligate cross pollination which inturn increases intrapopulation genetic diversity (Paschoa et al., 2018; Muyle et al., 2020). Dioecy is one of the adaptations in plants that promote outbreeding (Charlesworth, 2006). High genetic diversity and low inbreeding in *H. compressa* was also supported by AMOVA results which showed higher (92.7%) within population diversity than among population diversity (7.3%).

The understanding of the genetic diversity and population structure within *H. compressa* provides useful information for future selection and appropriate conservation strategies. High priority should be given to the conservation of all populations with high genetic diversity. The conservation of *H. compressa* must consider the two identified clusters to ensure that the high diversity within populations is retained. This can be achieved by maximum collection and *ex situ* conservation of germplasm especially for cluster 2 which had the most diversity.

CONCLUSION

This study was able to show the genetic diversity and population structure of *H. compressa* using the GBS approach. *Hyphaene compressa* in Kenyan ASALs is delineated into two gene pools. Cluster 1 comprising accessions in the north of Kenya while cluster 2 comprising accessions found along the River Tana basin. Further, accessions from the Tana basin are more diverse than those found in the northern part of Kenya. In addition, the results indicate that *H. compressa* accessions are interconnected with high gene flow and moderate genetic differentiation, evidenced by high within-population variation than among population variation. The high within population diversity can be harnessed for future breeding and improvement programs for various adaptive traits in *H. compressa.*

DATA AVAILABILITY STATEMENT

The sequence data generated from this study are archived in the NCBI SRA under BioProject accession number PRJNA756042 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA756042/)

AUTHOR CONTRIBUTIONS

CM, JN, NB, and LW conceived the study, CM, JN, NB, SO, and AO designed the analysis, AO, CM, JN, and NB collected the data, CM, AO, SO, NB, and JN analyzed the data, AO drafted the manuscript with significant contributions from all the authors. All authors contributed to data interpretation and approval of the paper.

FUNDING

This work was supported by a grant from The National Research Fund (NRF) Kenya, Grant Number NRF\1\MMC\285.

ACKNOWLEDGMENTS

Special gratitude goes to Kenya Forestry Service, National Museums of Kenya, Nuts and Oil Crops Directorate and Anglican Development Services in Lodwar Kenya for their guidance during sampling.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.762202/full#supplementary-material

Supplementary Figure S1 | Single Nucleotide Polymorphism (SNP) qualities after filtering the VCF file obtained from *de_novo*-based assembly of GBS data showing the depth per individual. **(A)**, observed heterozygosity **(B)**, Mean depth **(C)** and frequency of missing data per individual **(D)** for *Hyphaene compressa* accessions from Kenya.

Supplementary Figure S2 | Single Nucleotide Polymorphism (SNP) qualities after filtering the VCF file obtained from reference-based assembly of GBS data showing the depth per individual. **(A)**, observed heterozygosity **(B)**, Mean depth **(C)** and frequency of missing data per individual **(D)** for *Hyphaene compressa* accessions from Kenya.

Supplementary Figure S3 | Optimal Delta k values for different k values inferred during STRUCTURE analysis of Kenyan Hyphaene compressa showing the optimal delta k at k = 2 for the reference-based assembly **(A)** and the de_novo based assembly **(B)**.

Supplementary Figure S4 | STRUCTURE bar plot of admixture model population assignment of 96 *Hyphaene compressa* accessions from Kenya showing three clusters based on 2096 Single Nucleotide Polymorphisms (SNPs) for the de-novo assembly **(A)** and 23416 SNPs based on reference assembly **(B)**. A combination of different colors represents admixed populations. In these structure plots, Tharaka and Tana River accessions have been placed in one cluster.

Supplementary Figure S5 | Unrooted UPGMA distance tree inferred using 23416 SNPs based on the reference based assembly of *H. compressa* accessions.

Supplementary Figure S6 | Migration rates between Tharaka, Tana River and Kwale using MIGRATE-n software. The direction of the arrow indicate the direction of gene flow.

REFERENCES

Amwatta, C. J. M. (2004). Diversity of Use of Doum (Hyphaene Compressa) in Kenya. *Palms* 48 (4), 184–190.

Andersen, J. R., and Lübberstedt, T. (2003). Functional Markers in Plants. *Trends Plant Sci.* 8 (11), 554–560. doi:10.1016/j.tplants.2003.09.010

Barrow, E., and Mogaka, H. (2007). Kenya's Drylands - Wastelands or an Undervalued National Economic Resource. Available at: https://www.iucn.org/sites/dev/files/import/downloads/kenya_dryland_value_2007.pdf (Accessed June 24, 2021).

Bohling, J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecol. Evol.* 10, 7585–7601. doi:10.1002/ece3.6483

Brandies, P., Peel, E., Hogg, C. J., and Belov, K. (2019). The Value of Reference Genomes in the Conservation of Threatened Species. *Genes (Basel)* 10, 846. doi:10.3390/genes10110846

Brunet, J., Larson-Rabin, Z., and Stewart, C. M. (2012). The Distribution of Genetic Diversity within and Among Populations of the Rocky Mountain Columbine: The Impact of Gene Flow, Pollinators, and Mating System. *Int. J. Plant Sci.* 173 (5), 484–494. doi:10.1086/665263

Burghardt, L. T., Young, N. D., and Tiffin, P. (2017). A Guide to Genome-wide Association Mapping in Plants. *Curr. Protoc. Plant Biol.* 2, 22–38. doi:10.1002/cppb.20041

Charlesworth, D. (2006). Evolution of Plant Breeding Systems. *Curr. Biol.* 16 (17), R726–R735. doi:10.1016/j.cub.2006.07.068

Cosiaux, A., Gardiner, L., and Couvreur, T. L. (2017). *Hyphaene Compressa, the IUCN Red List of Threatened Species*. IUCN Red List. Available at: https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T95317478A95317481.en (Accessed September 2, 2017).

Cosiaux, A., Gardiner, L. M., Stauffer, F. W., Bachman, S. P., Sonké, B., Baker, W. J., et al. (2018). Low Extinction Risk for an Important Plant Resource: Conservation Assessments of continental African Palms (Arecaceae/Palmae). *Biol. Conservation* 221, 323–333. doi:10.1016/j.biocon.2018.02.025

D'Agostino, N., Taranto, F., Camposeo, S., Mangini, G., Fanelli, V., Gadaleta, S., et al. (2018). GBS-derived SNP catalogue unveiled wide genetic variability and geographical relationships of Italian olive cultivars. *Sci. Rep.* 10, 1–13. doi:10.1038/s41598-018-34207-y

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide Genetic Marker Discovery and Genotyping Using Next-Generation Sequencing. *Nat. Rev. Genet.* 12 (7), 499–510. doi:10.1038/nrg3012

Earl, D. A., and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: A Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conservation Genet. Resour.* 4, 359–361. doi:10.1007/s12686-011-9548-7

Eaton, D. A. R., and Overcast, I. (2020). Ipyrad: Interactive Assembly and Analysis of RADseq Datasets. *Bioinformatics* 36 (8), 2592–2594. doi:10.1093/bioinformatics/btz966

Edwards, D., Forster, J. W., Chagné, D., and Batley, J. (2007). "What Are SNPs," in *Association Mapping in Plants*. Editors N. Oraguzie, E. Rikkerink, S. Gardiner, and H. de Silva (New York: Springer), 41–52. doi:10.1007/978-0-387-36011-9_3

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A Robust, Simple Genotyping-By-Sequencing (GBS) Approach for High Diversity Species. *Plos One* 6 (5), e19379. doi:10.1371/journal.pone.0019379

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the Number of Clusters of Individuals Using the Software STRUCTURE: A Simulation Study. *Mol. Ecol.* 14 (8), 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Francis, R. M. (2017). Pophelper: an R Package and Web App to Analyse and Visualize Population Structure. *Mol. Ecol. Resour.* 17 (1), 27–32. doi:10.1111/1755-0998.12509

Galla, S. J., Forsdick, N. J., Brown, L., Hoeppner, M., Knapp, M., Maloney, R. F., et al. (2019). Reference Genomes from Distantly Related Species Can Be Used for Discovery of Single Nucleotide Polymorphisms to Inform Conservation Management. *Genes* 10 (9), 9. doi:10.3390/genes10010009

Ganie, S. H., Upadhyay, P., Das, S., and Prasad Sharma, M. (2015). Authentication of Medicinal Plants by DNA Markers. *Plant Gene* 4, 83–99. doi:10.1016/j.plgene.2015.10.002

Hazzouri, K. M., Gros-Balthazard, M., Flowers, J. M., Copetti, D., Lemansour, A., Lebrun, M., et al. (2019). Genome-wide Association Mapping of Date palm Fruit Traits. *Nat. Commun.* 10, 1–14. doi:10.1038/s41467-019-12604-9

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an Ultimate Marker-Assisted Selection (MAS) Tool to Accelerate Plant Breeding. *Front. Plant Sci.* 5 (484), 1–8. doi:10.3389/fpls.2014.00484

Huson, D. H., and Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* 23 (2), 254–267. doi:10.1093/molbev/msj030

Hyun, D. Y., Sebastin, R., Lee, K. J., Lee, G.-A., Shin, M.-J., Kim, S. H., et al. (2020). Genotyping-by-Sequencing Derived Single Nucleotide Polymorphisms Provide the First Well-Resolved Phylogeny for the Genus Triticum (Poaceae). *Front. Plant Sci.* 11 (688), 1–15. doi:10.3389/fpls.2020.00688

Johnson, D. L. (1993). Nomadism and Desertification in Africa and the Middle East. *GeoJournal* 31 (1), 51–66. doi:10.1007/bf00815903

Jombart, T. (2008). Adegenet: a R Package for the Multivariate Analysis of Genetic Markers. *Bioinformatics* 24 (11), 1403–1405. doi:10.1093/bioinformatics/btn129

Kahn, F., and Luxereau, A. (2008). Doum palm Habit and Leaf Collecting Practices in Niger. *Palms* 52 (1), 23–29.

Khan, S., Al-qurainy, F., and Nadeem, M. (2012). Biotechnological Approaches for Conservation and Improvement of Rare and

Endangered Plants of Saudi Arabia. *Saudi J. Biol. Sci.* 19 (1), 1–11. doi:10.1016/j.sjbs.2011.11.001

Kigomo, N. (2001). "State of forest Genetic Resources in Kenya," in Forest Genetic Resources Working Papers FGR/18E; Issue The sub-regional workshop FAO/IPGRI/ICRAF on the conservation, management, sustainable utilization and enhancement of forest genetic resources in Sahelian and North-Sudanian Africa, Ouagadougou, Burkina Faso, 22-24 September 1998. Available at: http://www.fao.org/3/ab396e/ab396e.pdf.

Kitheka, J. U., and Ongwenyi, G. S. (2002). "The Tana River Basin and the Opportunity for Research on the Land-Ocean Interaction in the Tana Delta," in *Aquadocs*. Nairobi, Kenya: University of Nairobi. Available at: http://hdl.handle.net/1834/7842.

Klimova, A., Ortega-Rubio, A., Vendrami, D. L. J., and Hoffman, J. I. (2018). Genotyping by Sequencing Reveals Contrasting Patterns of Population Structure, Ecologically Mediated Divergence, and Long-Distance Dispersal in North American Palms. *Ecol. Evol.* 8 (11), 5873–5890. doi:10.1002/ece3.4125

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Luo, Z., Brock, J., Dyer, J. M., Kutchan, T., Schachtman, D., Augustin, M., et al. (2019). Genetic Diversity and Population Structure of a Camelina Sativa spring Panel. *Front. Plant Sci.* 10 (February), 1–12. doi:10.3389/fpls.2019.00184

Maundu, P., and Tengnas, B. (2005). "Useful Trees and Shrubs for Kenya," in *Technical Handbook Number 35* (Nairobi, Kenya: World Agroforestry Centre-Eastern and Central Africa Regional Programme).

Mokhtar, M. M., Adawy, S. S., El-assal, S. E.-D. S., and Hussein, E. H. A. (2016). Genic and Intergenic SSR Database Generation, SNPs Determination and Pathway Annotations, in Date Palm (Phoenix Dactylifera L.). *Plos One* 11 (7), e0159268. doi:10.1371/journal.pone.0159268

Muyle, A., Martin, H., Zemp, N., Mollion, M., Gallina, S., Tavares, R., et al. (2020). Dioecy Is Associated with High Genetic Diversity and Adaptation Rates in the Plant Genus Silene. *Mol. Biol. Evol.* 38 (3), 805–818. doi:10.1093/molbev/msaa229

Nassiry, M. R., Javanmard, A., and Tohidi, R. (2009). Application of Statistical Procedures for Analysis of Genetic Diversity in Domestic Animal Populations. *Am. J. Anim. Vet. Sci.* 4 (4), 136–141. doi:10.3844/ajavsp.2009.136.141

Nkhoma, N., Shimelis, H., Laing, M. D., Shayanowako, A., and Mathew, I. (2020). Assessing the Genetic Diversity of Cowpea [Vigna Unguiculata (L.) Walp.] Germplasm Collections Using Phenotypic Traits and SNP Markers. *BMC Genet.* 21 (110), 1–16. doi:10.1186/s12863-020-00914-7

Omire, A., Budambula, N. L. M., Neondo, J., Gituru, R., and Mweu, C. (2020a). Phenotypic Diversity of Doum Palm (Hyphaene Compressa), a Semi-Domesticated Palm in the Arid and Semi-Arid Regions of Kenya. *Scientifica* 2020, 1–13. doi:10.1155/2020/4920830

Omire, A., Neondo, J., Budambula, N. L., Gituru, R., and Mweu, C. (2020b). Hyphaene Compressa, an Important palm in the Arid and Semi-arid Regions of Kenya. *Eth Res. Appl.* 20, 1–15. doi:10.32859/era.20.4.1-15

Ortiz, E. M. (2019). vcf2phylipV. 2.0: Convert a VCF Matrix into Several Matrix Formats for Phylogenetic Analysis. doi:10.5281/zenodo.2540861

Paschoa, R. P. d., Christ, J. A., Valente, C. S., Ferreira, M. F. d. S., Miranda, F. D. d., Garbin, M. L., et al. (2018). Genetic Diversity of Populations of the Dioecious Myrsine Coriacea (Primulaceae) in the atlantic forest. *Acta Bot. Bras.* 32, 376–385. doi:10.1590/0102-33062017abb0355

Pootakham, W., Jomchai, N., Ruang-areerate, P., Shearman, J. R., Sonthirod, C., Sangsrakru, D., et al. (2015). Genome-wide SNP Discovery and Identification of QTL Associated with Agronomic Traits in Oil palm Using Genotyping-By-Sequencing (GBS). *Genomics* 105 (5–6), 288–295. doi:10.1016/j.ygeno.2015.02.002

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155 (2), 945–959. doi:10.1007/s10681-008-9788-010.1093/genetics/155.2.945

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., et al. (2017). Bioinformatic Processing of RAD-seq Data Dramatically Impacts Downstream Population Genetic Inference. *Methods Ecol. Evol.* 8, 907–917. doi:10.1111/2041-210X.12700

Stauffer, F., Ouattara, D., and Stork, A. (2014). Monocotyledons 2 (Anthericaceae - Palmae). *Tropical African Flowering Plants: ecology and distribution* 8 (1), 326–354.

Stetter, M. G., and Schmid, K. J. (2017). Analysis of phylogenetic relationships and genome size evolution of the Amaranthus genus using GBS indicates the ancestors of an ancient crop. *Methods Phylogenet. Evol.* 109, 80–92. doi:10.1016/j.ympev.2016.12.029

Taranto, F., D'Agostino, N., Greco, B., Cardi, T., and Tripodi, P. (2016). Genome-wide SNP Discovery and Population Structure Analysis in Pepper (Capsicum Annuum) Using Genotyping by Sequencing. *BMC Genomics* 17 (1), 1–13. doi:10.1186/s12864-016-3297-7

Traveset, A., and Rodríguez-Pérez, J. (2019). "Seed Dispersal," in *Encyclopedia of Ecology*. Editor B. Fath. 2nd ed. (New York: Elsevier), 592–599. Issue October 2017. doi:10.1016/B978-0-12-409548-9.10950-9

Uhl, N. W., and Moore, H. (2019). palm. *Encyclopedia Britanica*. Available at: https://www.britannica.com/plant/palm-tree (Accessed July 27, 2021).

Wallace, J. G., and Mitchell, S. E. (2017). Genotyping-by-Sequencing. *Curr. Protoc. Plant Biol.* 2 (March), 64–77. doi:10.1002/cppb.20042

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *Joss* 4 (43), 1686. doi:10.21105/joss.01686

Xiong, H., Shi, A., Mou, B., Qin, J., Motes, D., Lu, W., et al. (2016). Genetic Diversity and Population Structure of Cowpea (Vigna Unguiculata L. Walp). *PLoS ONE* 11 (8), e0160941. doi:10.1371/journal.pone.0160941

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* 28 (24), 3326–3328. doi:10.1093/bioinformatics/bts606

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 March 2022
doi: 10.3389/fgene.2022.828238

Comprehensive Analysis of RNA-Seq in Endometriosis Reveals Competing Endogenous RNA Network Composed of circRNA, lncRNA and mRNA

Meichen Yin[1†], Lingyun Zhai[2†], Jianzhang Wang[1], Qin Yu[1], Tiantian Li[1], Xinxin Xu[1], Xinyue Guo[1], Xinqi Mao[1], Jianwei Zhou[2] and Xinmei Zhang[1*]*

[1]Department of Obstetrics and Gynecology, Women's Hospital, School of Medicine, Zhejiang University, Hangzhou, China, [2]Department of Gynecology, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, China

OPEN ACCESS

Edited by:
Jaira Ferreira de Vasconcellos, James Madison University, United States

Reviewed by:
Ugur Sezerman, Sabancı University, Turkey
Xijun Zhang, Uniformed Services University, United States

***Correspondence:**
Jianwei Zhou
2195045@zju.edu.cn
Xinmei Zhang
zhangxinm@zju.edu.cn

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: 03 December 2021
Accepted: 21 February 2022
Published: 22 March 2022

Citation:
Yin M, Zhai L, Wang J, Yu Q, Li T, Xu X, Guo X, Mao X, Zhou J and Zhang X (2022) Comprehensive Analysis of RNA-Seq in Endometriosis Reveals Competing Endogenous RNA Network Composed of circRNA, lncRNA and mRNA. Front. Genet. 13:828238. doi: 10.3389/fgene.2022.828238

Although long non coding RNAs (lncRNAs) and circular RNAs (circRNAs) play important roles in the pathogenesis of diseases, endometriosis related lncRNAs and circRNAs are still rarely reported. This study focused on the potential molecular mechanism of endometriosis related competitive endogenous RNA (ceRNA) composed of lncRNAs and circRNAs. We performed high-throughout sequencing of six normal endometria, six eutopic endometria and six ectopic endometria for the first time to describe and analyze the expression profile of lncRNA, circRNA and mRNA. Our results showed that 140 lncRNAs, 107 circRNAs and 1,206 mRNAs were differentially expressed in the ectopic group, compared with the normal and eutopic groups. We established an lncRNA/circRNA-mRNA co-expression network using pearson correlation test. Meanwhile, the results of Gene set enrichment analysis analysis showed that the 569 up-regulated differentially expressed mRNA (DEmRNA) were mainly related to the epithelial-mesenchymal transition, regulation of immune system process and immune effector process. Subsequently, we established a DElncRNA-miRNA and DEcircRNA-miRNA network using the starbase database, identified the common miRNAs and constructed DElncRNA/DEcircRNA-miRNA pairs. miRDB, Targetscan, miRwalk and circRNA/lncRNA-mRNA pairs jointly determined the miRNA-mRNA portion of the circRNA/lncRNA-miRNA-mRNA co-expression network. RT-qPCR results of 15 control samples and 25 ectopic samples confirmed that circGLIS2, circFN1, LINC02381, IGFL2-AS1, CD84, LYPD1 and FAM163A were significantly overexpressed in ectopic tissues. In conclusion, this is the first study to illustrate ceRNA composed of differentially expressed circRNA, lncRNA and mRNA in endometriosis. We also found that lncRNA and circRNA exerted a pivotal function on the pathogenesis of endometriosis, which can provide new insights for further exploring the pathogenesis of endometriosis and identifying new targets.

Keywords: endometriosis, non coding RNAs, competing endogenous RNA, circular RNA, long non coding RNA

INTRODUCTION

Endometriosis (EMS), the presence of extrauterine endometrial glands and stroma, is a common chronic disease affecting 10% of women in reproductive age. It is also the leading cause of subfertility or infertility in premenopausal women. However, its precise prevalence in the population is difficult to determine because it is asymptomatic or subclinical in most cases (El-Toukhy, 2020). Numerous hypotheses of EMS exist, including: embryonic stem cell origin, retrograde menstruation, implantation or coelomic metaplasia. Evidence accrued in recent years indicates that changes in the immune response can contribute to EMs. As we all know, endometriosis is an inflammatory disease. One pathogenesis of EMS is alterations in cell-mediated and humoral immunity. Immune responses are not only involved in the early immune escape of endometriotic cells, but also enhance the establishment and growth of endometriotic lesions (Zhang et al., 2018). Although in-depth and extensive studies have taken place, the specific pathogenesis of EMS has not been clarified. In terms of treatment of EMs, despite several available treatment options to relive, no real cure exists.

The recent application of next-generation sequencing has revealed thousands of non coding RNAs(ncRNAs). ncRNAs are functional RNAs transcribed from DNA, but not translated into proteins(Nothnick et al., 2015). The human genome project found that about 80% of human DNA is transcribed into RNA, of which only 2% of messenger RNA (mRNA) is translated into proteins, and most of the remainder is called non-coding RNA, including miRNA, lncRNA, circRNA. ncRNAs are important regulators of cell function and are widely involved in a variety of disease processes. Functional experiments show that ncRNAs are an important part of the pathogenesis of endometriosis and are related to the genetic risk associated with endometriosis (Sapkota et al., 2017). Moreover, lncRNAs regulate the expression of protein coding genes through cis acting on adjacent genes and trans on distal genes. Therefore, the expression level of lncRNAs is related to the expression level of the target gene mRNA (Liao et al., 2011).

Currently, little is known about the effect of lncRNA and circRNA in EMS. Therefore, it is necessary to comprehensively analyze lncRNA and circRNA to explore the function of lncRNA/circRNA-miRNA-mRNA ceRNA network in EMS. In this study, we analyzed the expression profiles of lncRNA, circRNA and mRNA using RNA sequencing and predicted the related functions of upregulated DEmRNA by GSEA. The results of GSEA indicated that DEmRNA were mainly enriched in epithelial-mesenchymal transition, positive regulation of immune system process and immune effector process. Subsequently, we established the circRNA/lncRNA-miRNA-mRNA ceRNA network based on database prediction and expression correlation analysis. The RT-qPCR results of clinical samples confirmed that the expression trend and correlation of important molecules in the co-expression network were consistent with the hypothesis.

This is the first study to construct circRNA/lncRNA-miRNA-mRNA co-expression network by analyzing DElncRNA, DEcircRNA and DEmRNA from sequencing data in EMS. Our findings may lead to the discovery of a new pathogenesis of EMS and offer new theories for treatment.

METHODS

Sample Collection and RNA Extraction

This research proposal was approved by the Women's Hospital of Zhejiang University School of Medicine. In accordance with the Declaration of Helsinki, all patients received written informed consent prior to enrollment. Six paired ectopic endometrium and eutopic endometrium samples were selected and prepared for RNA sequencing from six patients with ovarian EMs cysts diagnosed by laparoscopic surgery. Meanwhile, six normal endometrial samples were obtained from six patients who underwent hysteroscopy for endometrial polyps.

Fresh tissue specimens were immediately frozen in liquid nitrogen and stored at −80°C, Samples were sent to Genenergy Biotechnology (http://www.genenergy.cn/) for RNA extraction. According to the manufacturer's instructions, a sequencing bank was created using the TruSeq RNA Sample Preparation Kit (Illumina). Illumina HiSeq X Ten was used to analyze 151-BP paired sequences of the library.

Identification of DERNAs

In order to identify the differentially expressed lncRNA and mRNA in ectopic endometrium, eutopic endometrium and normal endometrium, differential expression profiles were analyzed. DElncRNA and DEmRNA were distinguished according to |log2 (fold change)| > 3.0 and FDR <0.0001. DEcircRNA and were distinguished according to |log2 (fold change)| > 1 and p value <0.05. According to the above criteria, DERNA was screened from the expression profiles of the ectopic endometrium vs. normal endometrium and ectopic endometrium vs. eutopic endometrium respectively, and the common part of the two was regarded as the DERNA of this study. Venn diagrams were used to visualize the share DERNA between the two datasets for further analysis.

GSEA

Gene set enrichment analysis (GSEA) was performed to identify significantly enriched groups of DEmRNA (Subramanian et al., 2005). In this study, GSEA software was applied to analyze biological pathway divergences, KEGG and hallmarks between ectopic and non-ectopic samples. $p < 0.05$ was considered the threshold value for statistical significance.

Establishment of DEcirc/lncRNA-DEmRNA

The interaction between DEmRNA and DElncRNA was identified by the lncRNA-mRNA co-expression network. The basis of this construct is the normalized signal intensities of specific expression levels of mRNAs and lncRNAs. To construct the LncRNA-mRNA co-expression network, pearson correlation analysis was used to calculate statistically significant associations. The lncRNA-mRNA pairs with a rho ≥0.6 and $p < 0.05$ were selected, as these parameters indicated that the lncRNA-mRNA

pairs were significantly co-expressed. Next, we conducted pearson correlation analysis between DEmRNA and DEcircRNA, and these DEmRNA were newly obtained DEmRNA related to DElncRNA expression ($|rho| \geq 0.6$, $p < 0.05$). DEcircRNA/DElncRNA-mRNA interactions were mapped using Cytoscape software (3.8.2) (Janet et al., 2021).

DEcircRNA/lncRNA-miRNA

According to the ceRNA hypothesis, matching circRNA/lncRNA, miRNA and mRNA is crucial. Thus, this network may highlight a new molecular mechanism involved in the development of endometriosis. Pairs of miRNA-DElncRNA were established using the starbase database(PaciP et al., 2014; Li et al., 2014)

DEcircRNA-miRNA was constructed with miRanda software. The miRanda algorithm comprehensively predicts miRNA target genes through two steps of miRNA-circRNA sequence matching and energy stability evaluation. The algorithm uses a dynamic programming algorithm to search the region where the sequences of miRNA and circRNA are complementary and stable to form double strands. Threshold parameters used in predicting miRNA target genes are: S > 150, ΔG<−20 kcal/mol and Demand strict 5′ seed pairing, where S refers to single-legs-pair match scores in the matching area; ΔG is the free energy of double strand formation that constructs a global network of miRNA and target circRNA. The final DEcircRNA-miRNA pairs were the common part of the top 50 DEcircRNA-miRNA in normal vs. ectopic datasets and the top 50 DEcircRNA-miRNA in eutopic vs. ectopic datasets.

Construction of lncRNA/circRNA-miRNA-mRNA Network

A circRNA/lncRNA–miRNA–mRNA ceRNA network was constructed based on the targeted relationships. Three miRNAs in this network were the intersection of predicted miRNA in both lncRNA-miRNA and circRNA-miRNA pairs. The target mRNAs of these three miRNAs were obtained from the TargetScan, miRDB, and miRwalk2.0 databases (Wang, 2008; AgarwalV et al., 2015; Dweep and Gretz, 2015). In order to improve the reliability of the results, we only selected those miRNA-mRNA relationship pairs that overlapped in all three databases for further research. Furthermore, the mRNAs predicted in the three data sets do not all belong to our ceRNA network, only mRNA closely related to the expression of lncRNA and circRNA could be considered as mRNA in the ceRNA network.

Clinical Specimen Collection

This study included patients who had surgical treatments at Zhejiang University's Women's Hospital's Department of Gynaecology during October and November of 2021. The endometrial samples were taken from women who had regular menstrual cycles and had not been treated with steroid hormones in the previous 3 months. Controls ($n = 15$) were endometrial samples from patients without endometriosis, adenomyosis, or other malignant illnesses. Laparoscopy and histological examination of patients with endometriosis (ectopic, $n = 25$) were used to confirm the diagnosis.

TABLE 1 | Primer sequences for quantitative real-time polymerase chain reaction.

Name	Primer type	Primer sequence
circGLIS2	Forward	CAGCAGCTCGCTGTCCCCCGAGCG
circGLIS2	Reverse	GTTGGAGGTGGCAGCAGGCAGTGG
circFN1	Forward	GGAGAAGTATGTGCATGGTGTCA
circFN1	Reverse	TGCAGATTTCCTCGTGGGTTG
LINC02381	Forward	CCCTGCCCATAAGCTACTCA
LINC02381	Reverse	AACTTTGACCCCCAAATGCC
IGFL2-AS1	Forward	AGTTCCTGATTTCAGCCCCA
IGFL2-AS1	Reverse	TCCTGGGTTGACAGGGTAGAA
CD84	Forward	GGAGAAGAGGGTAATGTCCTTCA
CD84	Reverse	CCATTGCGATGTCTGCACA
LYPD1	Forward	GGCAACTTTTTGCGGATTGTT
LYPD1	Reverse	CGTTCACCGTGCAATTCACA
FAM163A	Forward	ATGACAGCGGGAACGGTTG
FAM163A	Reverse	GTAGCACAGGACGGCAATGAT

RT-qPCR

The results of RNA-seq need to be verified by RT-qPCR. Trizol was used to extract the total RNA in the endometrial tissues of the control group and ectopic group. The concentration and purity of RNA were evaluated using a Nanodrop spectrophotometer (Thermo Fisher, United States). When the 260/280 ratio of RNA was >1.8, it was reverse transcribed to cDNA using the PrimeScriptTM RT reagent Kit with gDNA Eraser (Takara, Japan). The reverse transcription products were amplified using the Applied Biosystems ViiA™ seven system (ABI, United States) with the SYBR® Premix Ex TaqTM kit 8 (Takara, Japan). Specific primers were synthesized by Generay (Shanghai, China), and the sequences are presented in **Table 1**. The relative expression was normalized by GAPDH, and the 2-ΔΔCt method was used to calculate the relative expression (Liu et al., 2020).

Statistical Analysis

Statistical analyses were performed by the GraphPad Prism 7 (GraphPad, CA, United States) and SPSS 22.0 software packages (SPSS, IL, United States). Statistically significant differences between groups were estimated by Independent-Sample *t* test. The results were evaluated using Spearman's correlation coefficient test. All values are expressed as the mean ± standard error of the mean; $p < 0.05$ was considered statistically significant.

RESULTS

Differential Expression Profile of mRNAs, lncRNAs and circRNAs

According to the cut off criteria (FDR<0.0001 and |logFC|>3), a total of 140 DElncRNAs (68 up-regulated and 72 down-regulated) and 1,206 DEmRNAs(568 upregulated and 6,388 down regulated) were identified respectively. Owing to the different expression levels in the samples, differentially expressed circRNAs were screened according to |logFC| > 1,

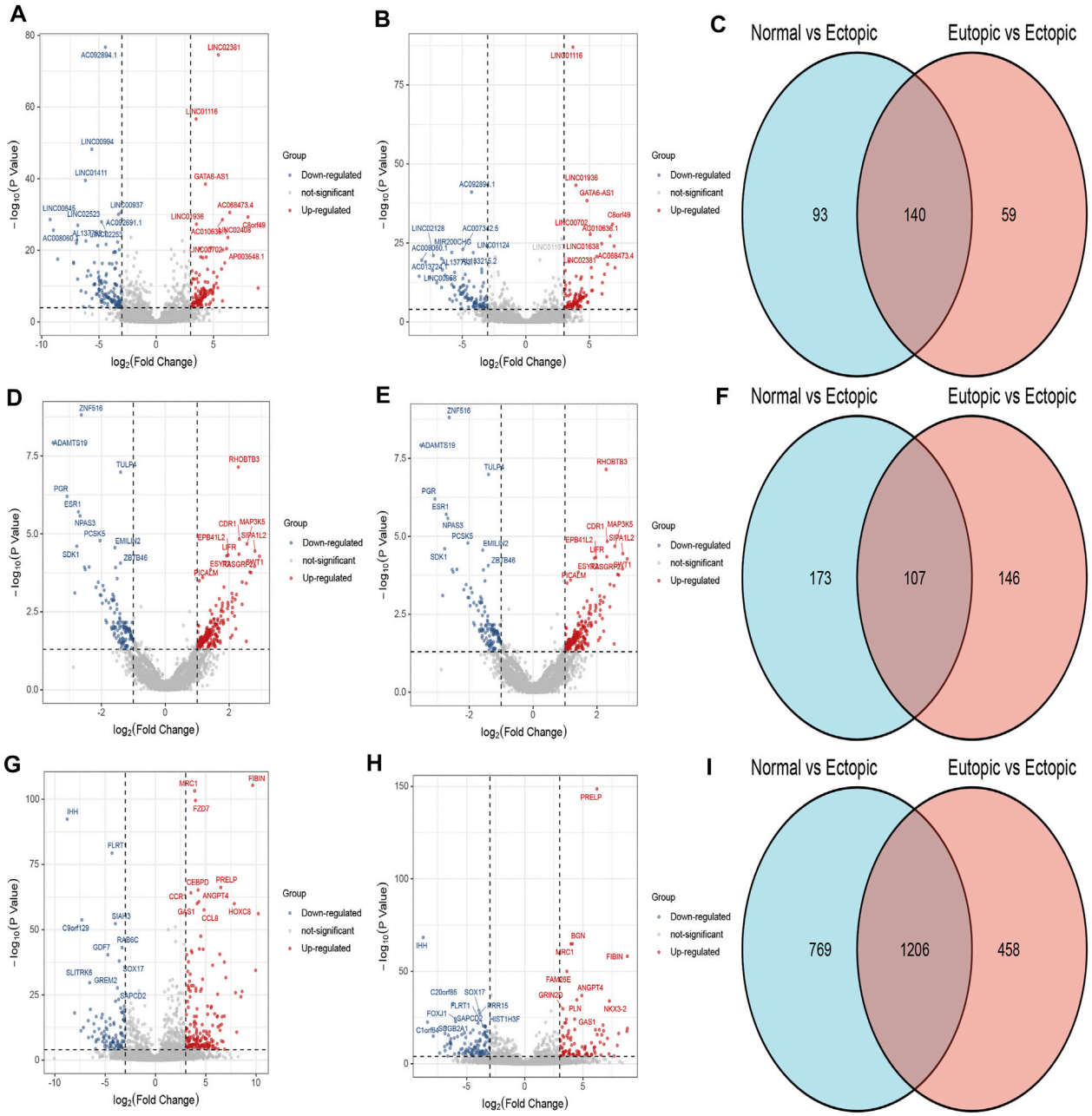

FIGURE 1 | DElncRNA, DEcircRNA and DEmRNA in normal endometrium vs. ectopic endometrium and eutopic endometrium vs. ectopic endometrium datasets. **(A)** Volcano plot of DElncRNA in normal endometrium vs. ectopic endometrium dataset. **(B)** Volcano plot of DElncRNA in eutopic endometrium vs. ectopic endometrium dataset. **(C)** Venn diagram of common part of DElncRNA in normal endometrium vs ectopic endometrium and eutopic endometrium vs. ectopic endometrium datasets. **(D)** Volcano plot of DEcircRNA in normal endometrium vs. ectopic endometrium dataset. **(E)** Volcano plot of DEcircRNA in eutopic endometrium vs. ectopic endometrium. **(F)** Venn diagram of common part of DEcircRNA in normal endometrium vs ectopic endometrium and eutopic endometrium vs. ectopic endometrium datasets. **(G)** Volcano plot of DEmRNA in normal endometrium vs. ectopic endometrium dataset. **(H)** Volcano plot of DEmRNA in eutopic endometrium vs. ectopic endometrium dataset. **(I)** Venn diagram of common part of DEmRNA in normal endometrium vs ectopic endometrium and eutopic endometrium vs. ectopic endometrium datasets.

$p < 0.05$. A total of 107 circRNAs were differentially expressed in the two expression profiles(**Figure 1**). Heatmaps of DEcircRNA, DEmRNA and DElncRNA in their respective expression profiles are shown in **Figure 2**. In the following study, we mainly focused on the highly expressed DElncRNA and DEcircRNA.

GSEA of DEmRNA

To further investigate the role of DEmRNA expression in the endometriosis microenvironment. Gene set enrichment analysis was conducted by utilizing the gene expression profiles of 568 overexpressed DEmRNA (**Supplementary Table**). The gene

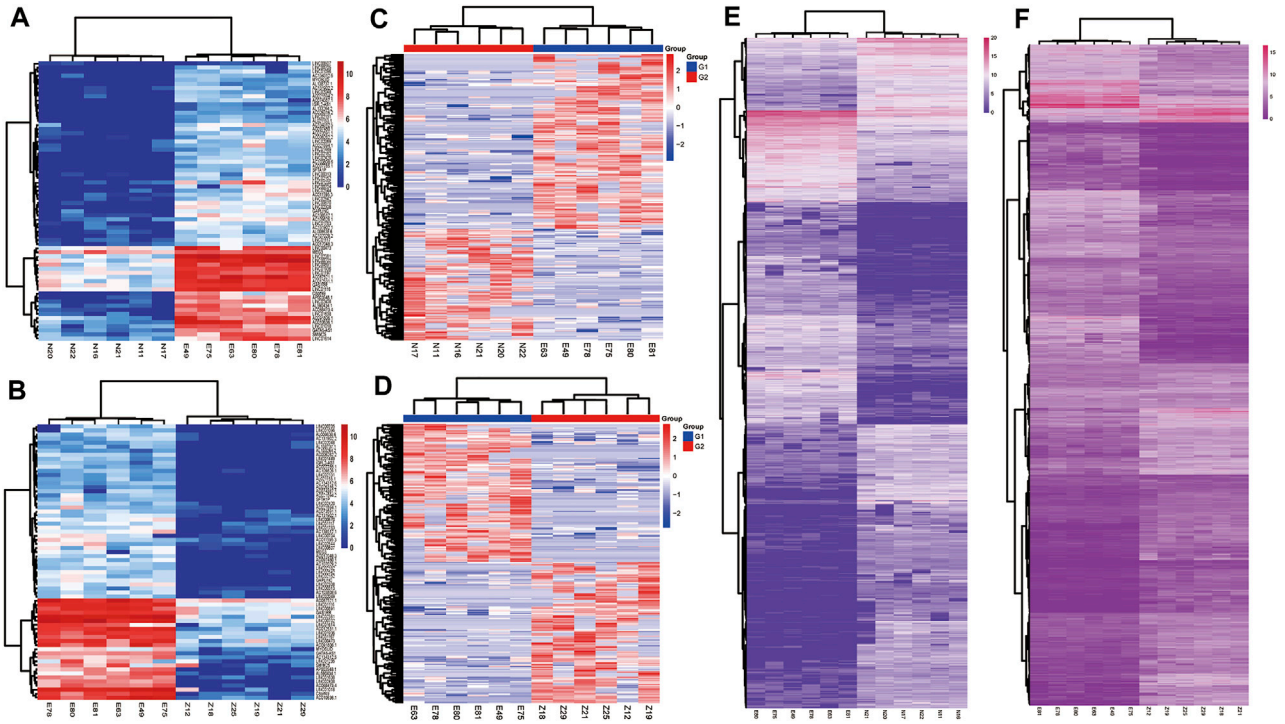

FIGURE 2 | Heatmaps of DElncRNA, circRNA and mRNA in normal endometrium vs. ectopic endometrium and eutopic endometrium vs. ectopic endometrium datasets. **(A)** DElncRNA in normal endometrium vs. ectopic endometrium **(B)** DElncRNA in eutopic endometrium vs. ectopic endometrium datasets **(C)** DEcircRNA in normal endometrium vs. ectopic endometrium **(D)** DEcircRNA in eutopic endometrium vs. ectopic endometrium datasets. **(E)** DEmRNA in normal endometrium vs. ectopic endometrium **(F)** DEmRNA in eutopic endometrium vs. ectopic endometrium datasets.

signatures implied enrichment in many categories, such as epithelial-mesenchymal transition, regulation of immune system process and immune effector process(**Figure 3**).

Circ/lncRNA-mRNA Co-Expression Network

Pearson correlation analysis was performed on DEcirc/DElncRNAs and DEmRNAs to select DEcirc/DElncRNA–DEmRNA pairs ($p < 0.05$ and rho ≥ 0.6). Meanwhile, by constructing pearson correlation analysis to establish ceRNA, circ/lnc/mRNA with unrelated expression were excluded. In total, 68 lncRNAs, 94 circRNA and 546 mRNAs were included in the co-expression network. circ/lncRNA-mRNA pairs are depicted via cytoscape software in **Figures 4A,C**, circRNA-mRNA network was composed of 627 nodes and 1817 edges. lncRNA-mRNA network was composed of 523 nodes and 1,203 edges. Correlation heatmaps of circRNA/lncRNA-mRNA are shown in **Figures 4B,D**.

Prediction and Identification miRNAs Targeted by Both DElncRNA and DEcircRNA

The starbase database was employed to predict the targeting miRNA of 68 DElncRNAs. Since lncRNAs and miRNAs are not one-to-one correspondences and some lncRNA do not predict miRNAs in the database, we obtained the 452 interaction between 31 lncRNAs and 294 miRNAs (**Figure 5A**).

There were 550 interacting circRNA-miRNA pairs predicted between 50 DECs and 50 miRNAs by the miRanda database in normal vs. ectopic datasets. Moreover there were 486 circRNA-miRNA pairs between 50 DECs and 50 miRNAs according to the miRanda database in eutopic vs. ectopic datasets. The circRNA-miRNA pairs mentioned above all met the conditions of S > 150, $\Delta G < -20$ kcal/mol and Demand strict 5′ seed pairing. Among them, 120 pairs of circRNA-miRNAs were shared by the two data sets(**Figure 5B**).

miRNAs shared by circRNA-miRNA and lncRNA-miRNA were selected as common miRNA in the co-expression network for further analysis. As we can seen in **Figure 6A**, hsa-miR-138-5p, hsa-miR-3619-5p and hsa-miR-1301-3p can bind to miRNA response elements of lncRNA and circRNA.

Prediction of miRNA-mRNA Targeting Relationship

miRDB, TargetScan and miRwalk2.0 were simultaneously used to predict putative miRNA–mRNA interactions. In order to further improve the accuracy of prediction, we took the intersection of the predicted results of each miRNA in the three databases. Besides, the miRNA-mRNA pairs predicted by the three databases were not the final mRNAs that our co-expression

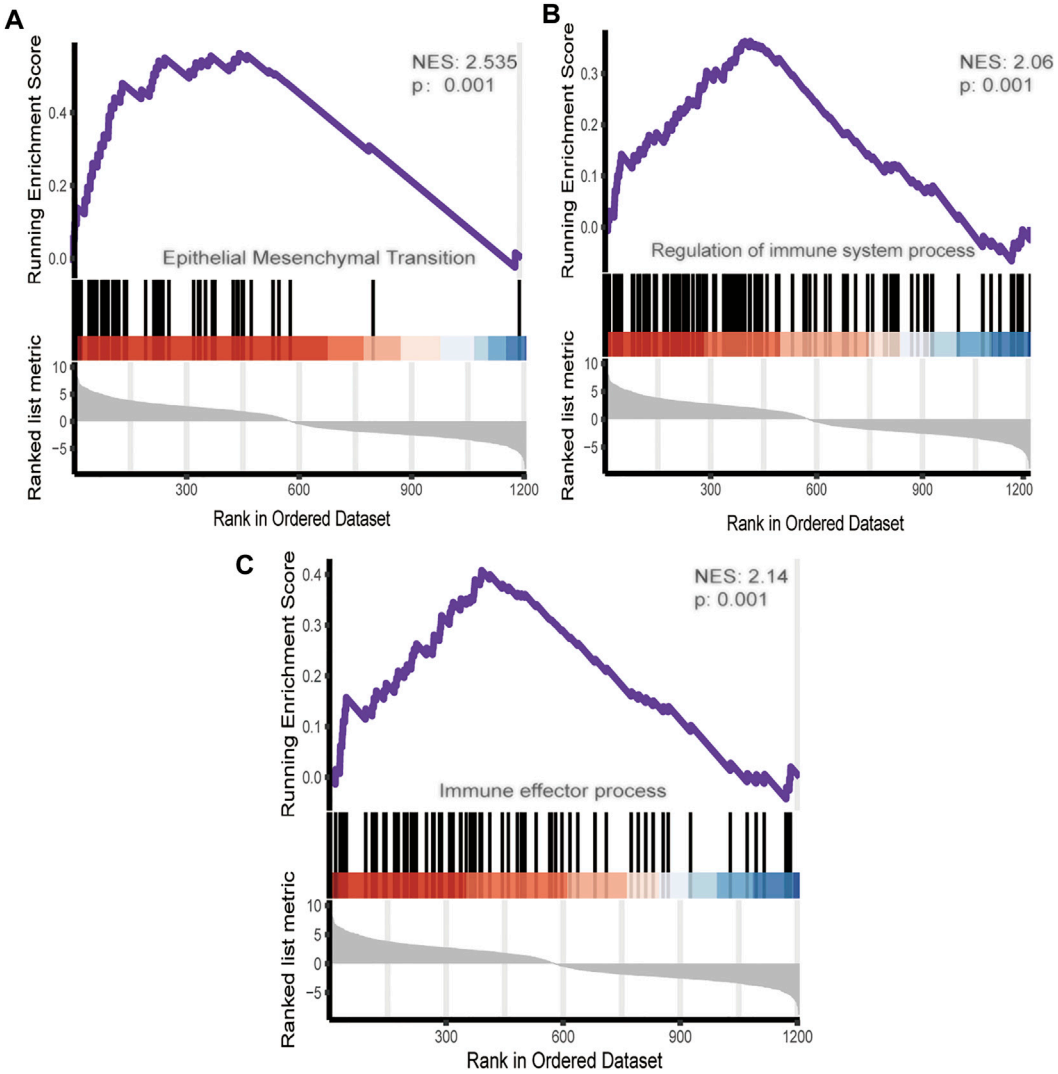

FIGURE 3 | GSEA of DEmRNA.

network wanted to study. Only the predicted mRNAs presented in DEmRNA that we screened previously and related to circRNA/lncRNA expression were included in this co-expression network (**Figures 6B–D**). hsa-miR-1301-3p could bind to the 3 ′UTR of 16 DEmRNAs based on database prediction and expression correlation analysis. Similarly, hsa-miR-138-5p and hsa-miR-3619-5p could bind to 7 and 32 DEmRNAs respectively.

Construction of ceRNA Network

According to the above analysis and prediction, there was a circRNA/lncRNA-miRNA-mRNA co-expression network among DEcircRNA, DElncRNA and DEmRNA in our sequencing results. The co-expression network consisted of three lncRNAs, 13 circRNAs, three miRNAs and 49 mRNAs. Among them, lncRNAs and mRNA were significantly differentially expressed in normal endometrium vs. ectopic endometrium and eutopic vs. ectopic endometrium (logFC > 3, FDR < 0.0001). circRNA met the condition of $|\log FC| > 1$, $p < 0.05$ in the same two datasets as lncRNA and mRNA. The pearson correlation coefficient of circRNA/lncRNA-mRNA expression was greater than 0.6($p < 0.05$) (**Figure6E**) According to starbase database, LINC02381-hsa-miR-1301-5p and GAS1RR-hsa-miR-3619-5p exhibited negative correlation in 1,085 breast invasive carcinoma samples. This is consistent with the ceRNA theory (**Figure 7**). S and ΔG of circRNA-miRNA pairs in co-expression network are shown in **Table 2**. The feasibility of this network was demonstrated from the perspective of expressing relevance and database prediction. However, further experimental verification is needed.

Reverse Transcription Quantitative PCR

As mentioned above, circRNA, lncRNA and mRNA in the co-expression network were significantly differentially expressed in ectopic endometrial samples. RT-qPCR was utilized to detect the expression of DERNAs in 25 ectopic endometrial tissues and 15

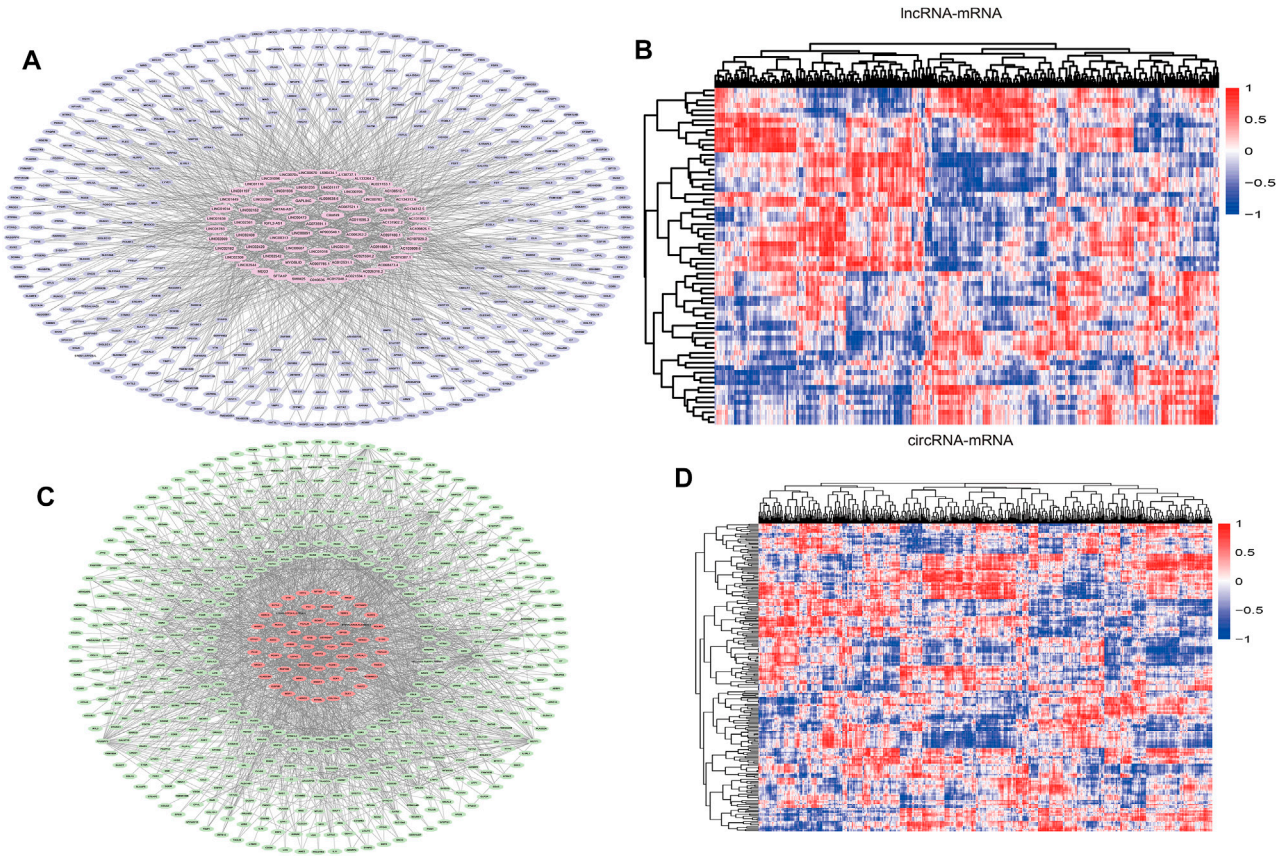

FIGURE 4 | DElncRNA-DEmRNA pairs and DEcircRNA-DEmRNA pairs. **(A)** Networks of DElncRNA-DEmRNA, The pink hexagon in the middle represents DElncRNA, the outer purple oval represents DEmRNA. **(B)** Pearson correlation heatmaps of DElncRNA-DEmRNA. The color of each block represents the pearson correlation coefficient, red is positive, blue is negative. **(C)** Networks of DEcircRNA-DEmRNA. The red oval in the middle represents DEcircRNA, the outer green hexagon represents DEmRNA. **(D)** Pearson correlation heatmaps of DEcircRNA-DEmRNA.

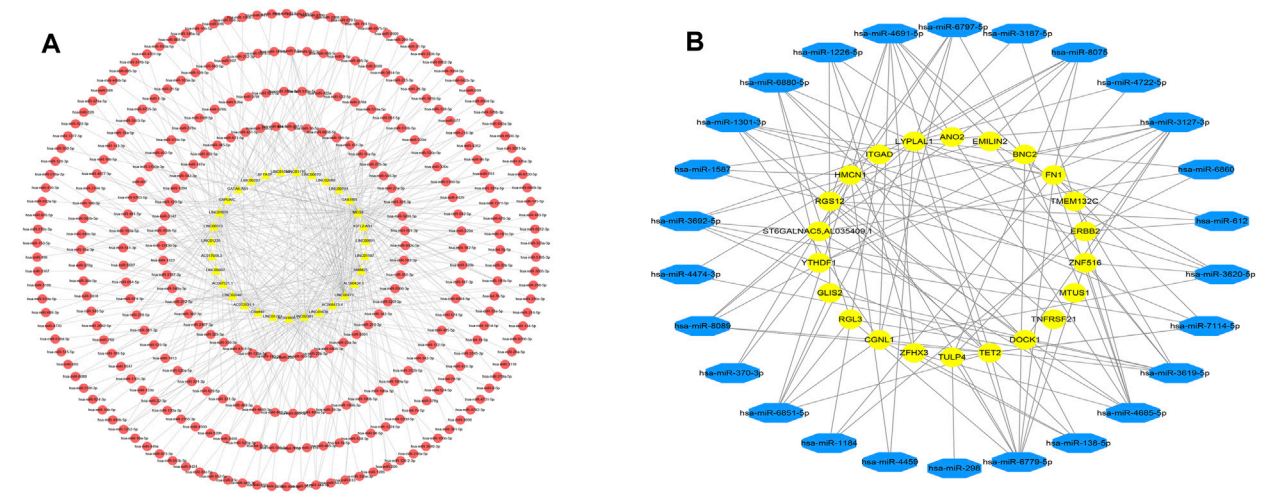

FIGURE 5 | DElncRNA-miRNA and DEcircRNA-miRNA pairs. **(A)** Networks of DElncRNA-miRNA, yellow triangles represent lncRNAs and red circles represent miRNAs. **(B)** Networks of DEcircRNA-miRNA, yellow circle represents circRNA and the red octagon represents miRNA.

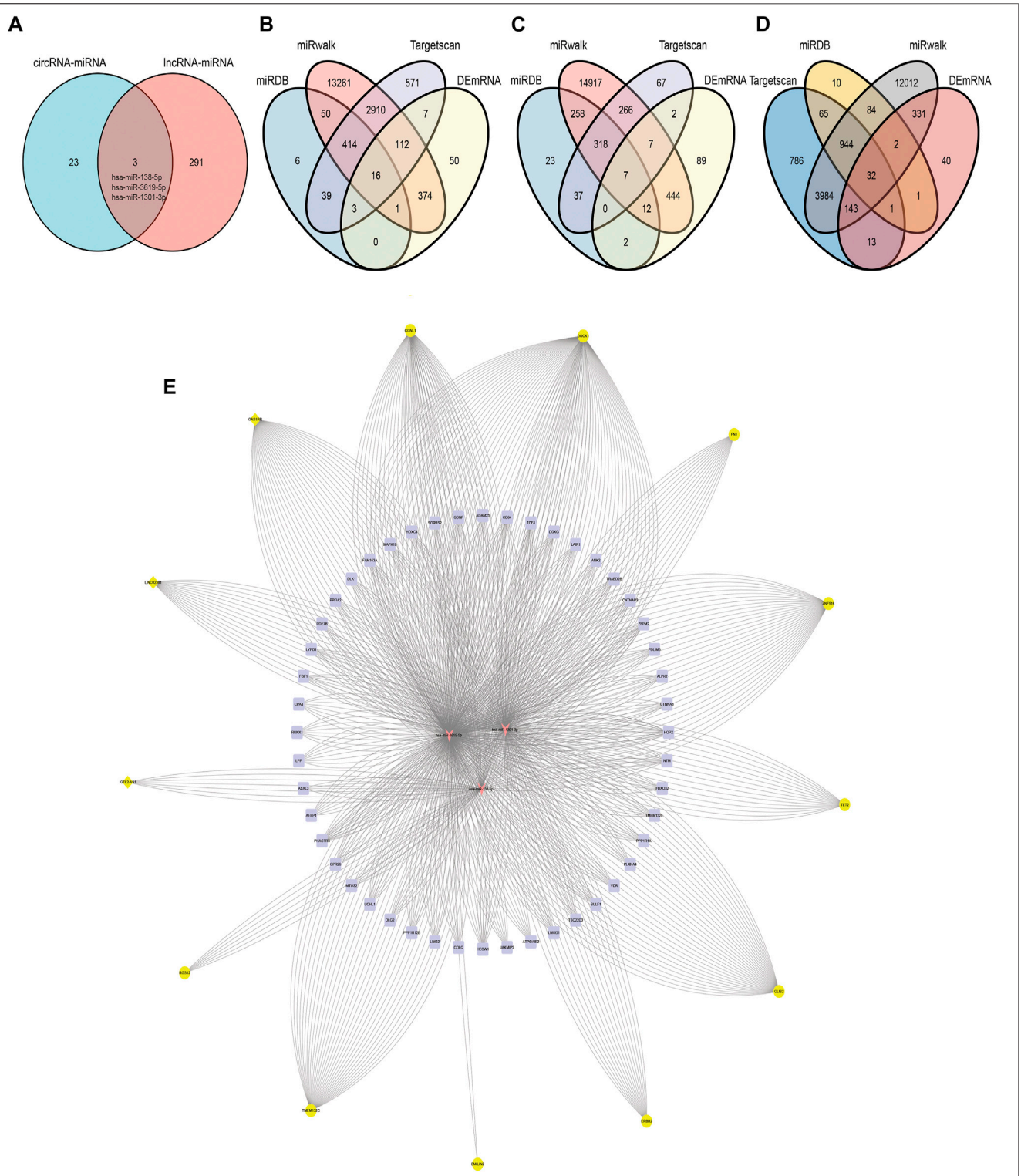

FIGURE 6 | circRNA/lncRNA-miRNA-mRNA. **(A)** Common miRNA of circRNA-miRNA pairs and lncRNA-miRNA pairs. **(B)** Target mRNA of hsa-miR-1301-3p in miRDB, miRwalk, Targetscan and DEmRNA. **(C)** Target mRNA of hsa-miR-138-5p in miRDB, miRwalk, Targetscan and DEmRNA. **(D)** Target mRNA of hsa-miR-3619-5p in miRDB, miRwalk, Targetscan and DEmRNA. **(E)** Co-expression network of circRNA/lncRNA-miRNA-mRNA, 16 yellow circles in the outermost circle represent ncRNAs, the three orange arrows in the innermost circle represent miRNAs, and the 49 purple squares in the middle are mRNAs.

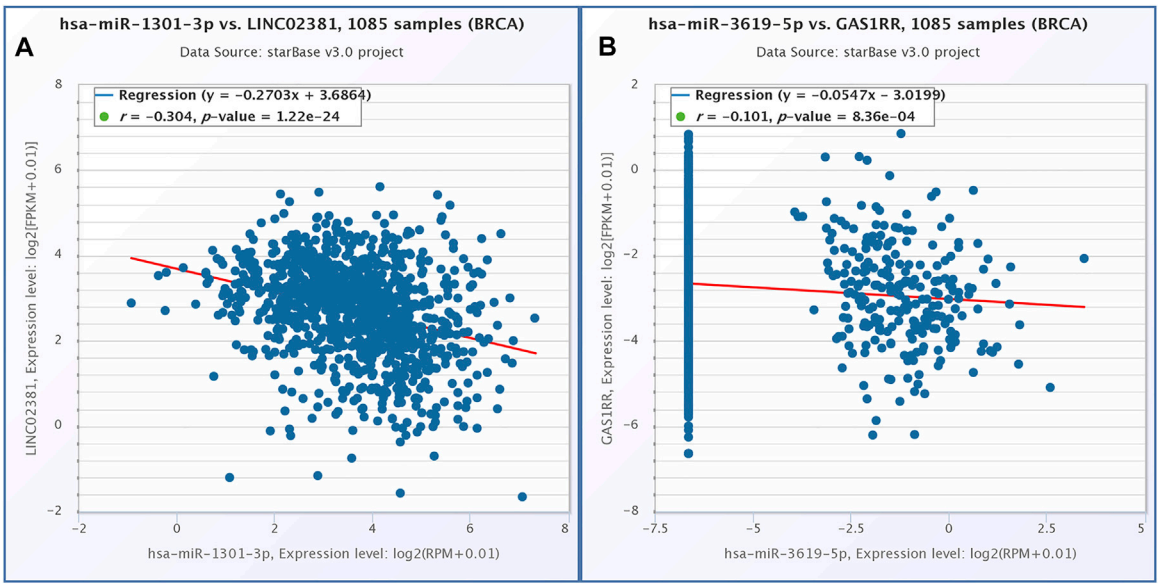

FIGURE 7 | Correlation expression analysis of BRCA in starbase database. **(A)** LINC02381 and hsa-miR-1301-3p. **(B)** GAS1RR and hsa-miR-3619-5p.

TABLE 2 | Single-legs-pair match scores and free energy of double strand formation of circRNA-miRNA pairs in circRNA/lncRNA-miRNA-mRNA network.

circRNA	miRNA	Tot score	Tot energy
RGS12	hsa-miR-138-5p	164	−23.59
ZFHX3	hsa-miR-138-5p	317	−48.12
CGNL1	hsa-miR-138-5p	161	−21.78
DOCK1	hsa-miR-138-5p	156	−24.32
FN1	hsa-miR-1301-3p	157	−22.74
ITGAD	hsa-miR-1301-3p	163	−26.58
DOCK1	hsa-miR-1301-3p	155	−20.91
TET2	hsa-miR-1301-3p	161	−25.18
ERBB2	hsa-miR-1301-3p	150	−25.62
CGNL1	hsa-miR-1301-3p	153	−22.39
CGNL1	hsa-miR-3619-5p	155	−21.89
ZNF516	hsa-miR-3619-5p	306	−48.61
GLIS2	hsa-miR-3619-5p	320	−50.62
DOCK1	hsa-miR-3619-5p	317	−52.08
EMILIN2	hsa-miR-3619-5p	153	−24.32
RGL3	hsa-miR-3619-5p	174	−37.37
TMEM132C	hsa-miR-3619-5p	152	−22.05

negative control endometrial tissues(**Figure 8**). We focused on the two up-regulated lncRNA (IGFL2-AS1, LINC02381) and two up-regulated circRNA (circGLIS2, circFN1) for further research. Among the mRNA in the ceRNA network composed of LINC02381-hsa-miR-1301-3p pairs in co-expression networks, the mRNA with the highest pearson correlation coefficient with LINC02381 was selected for RT-qPCR validation (LYPD1). The mRNA with the strongest correlation with circGLIS2 and circFN1 was also found by the same method (CD84, FAM163A) Compared with the control group, the ncRNA, including circGLIS2, circFN1, LINC02381, IGFL2-AS1, were significantly overexpressed and consistent with the RNA sequencing results. CD84, LYPD1, and FAM163A were also significantly overexpressed in ectopic tissues. Spearman correlation analysis based on our 40 samples indicated that LINC02381 was positively associated with LYPD1 (r = 0.592, p = 0.000057), circFN1 was positively associated with CD84 (r = 0.533, p = 0.001), and circGLIS2 was positively associated with FAM163A (r = 0.572, p = 0.000114). These results were consistent with our above analysis above.

DISCUSSION

Using the forefront technology in microarray analysis, we demonstrated the expression profiles of human lncRNAs, circRNAs and mRNAs in patients with EMS for the first time. Compared to matched controls, these EMS patients expressed 140 lncRNAs, 107 circRNAs and 1,206 mRNAs that did not appear in the control group. Moreover, we identified potential functions of these differentially expressed mRNAs with GSEA and established circRNA/lncRNA-miRNA-mRNA in EMS.

The results of this study showed that the expressions of four ncRNAs (circGLIS2, circFN1, IGFL2-AS1 and LINC02381) and three mRNA (CD84, LYPD1, FAM163A) were markedly different between EMS tissues and control endometrial tissues. Many of these are still incompletely studied.

circFN1-CD84 pair was confirmed to be highly expressed in EMS patients in our analysis. circFN1 has been shown in previous studies to be related to the occurrence of drug resistance to sorafenib in hepatocellular carcinoma and cisplatin in gastric cancer via ceRNA (Chen C et al., 2020; Huang et al., 2020). CD84 (SLAMF5) is a member of the signaling lymphocyte activation molecule (SLAM) family, which are cell surface proteins committed to regulating immune response (Silvia et al., 2008).

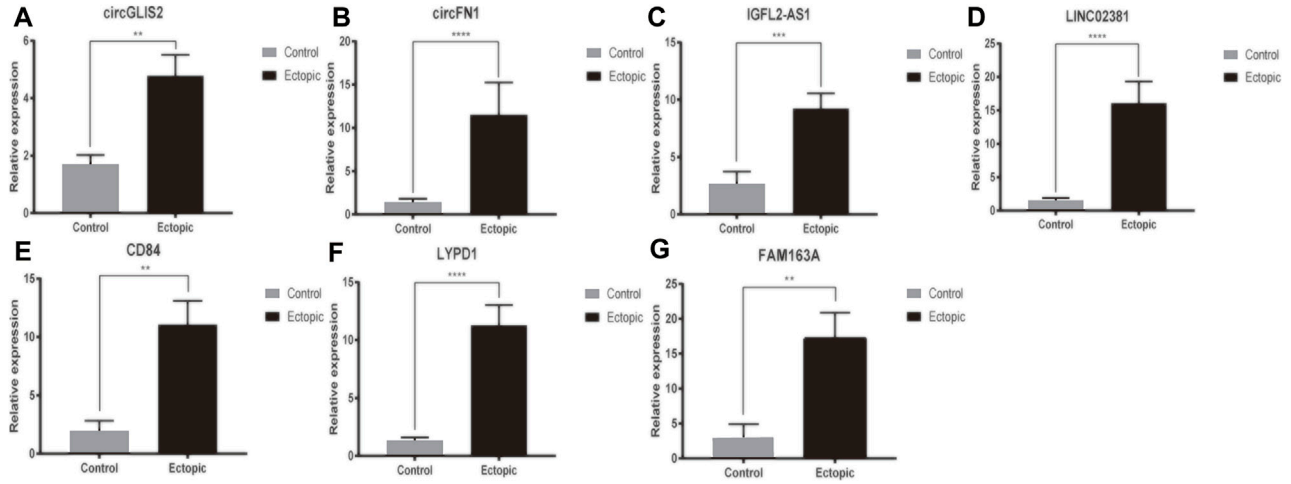

FIGURE 8 | RT-qPCR of DEcircRNA, DElncRNA and DEmRNA in co-expression network.

CD84 mediated signaling regulates a variety of immune processes, including natural killer cytotoxicity, T cytokine secretion, monocyte activation, autophagy, homologous T, B interaction and B cell tolerance at germinal center checkpoints (Marta et al., 2019). Recently, endometriosis has been considered as an autoimmune disease (AID) in view of the presence of autoantibodies, high cytokine levels, therapeutic sensitivity to immunomodulators and co-currence with other AID (Hila et al., 2021). Studies in the 1980s found that the dysregulation of host immune system is the main process leading to endometriosis. Since then, changes in the immune system including cell-mediated immunity and humoral immunity in endometriosis have been identified (Symons et al., 2018). B lymphocyte activation in EMS may result in the formation of a variety of autoantibodies The deposition of immunoglobulin and complement in endometrium indicates the role of immune complex formation in the pathogenesis of EMS. In summary, a disordered immune response plays an integral role in the pathogenesis of EMS. Consistent with the previous research results, we found that circFN1 affected the expression of CD84 by sponging hsa-miR-1301-3p in EMS, thus affecting the immune related activities of EMS.

Bioinformatics analyses indicated that LINC02381 as a ceRNA can sponge hsa-miR-1301-3p to modulate expression level of LYPD1. Several functional experiments have shown that LINC02381 can affect molecular pathogenesis of wide a variety of diseases as ceRNA. In rheumatoid arthritis, upregulation of LINC02381 can complement with miR-590-5p to reduce its level in RA tissues and inhibit the expression of mitogen-activated protein kinase 3 (MAP2K3) at the post-transcription level (Sun et al., 2021). In colorectal carcinoma (CRC), LINC02381 exhibit down-regulated expression in CRC tissues and different cell lines. This differential expression affects the growth and apoptosis of CRC cells through PI3K-Akt signaling pathway (Meisam et al., 2008). In gastric cancer, downregulated LINC02381 resulted in the increase of free miR-21, miR-590 and miR-27a, enhanced cell proliferation and ascension of EMT related markers. However, LINC02381 has never been reported in EMS. In our analysis, the 3' UTR of CD84 and LYPD1 could combine with hsa-miR-1301-3p. This indicatesd that the expression of CD84 and LYPD1 can be regulated by LINC02381 (Wang and Zhao, 2020). For LYPD1, it is reported that LY6/PLAUR Domain containing 1 (LYPD1) is a novel therapeutic antibody target for ovarian cancer. LYPD1 is extensively expressed in both primary and metastatic ovarian cancer. Anti-LYPD1/CD3 T-cell-dependent bispecific antibody (TDB) can lead polyclonal T cells to activate and target ovarian cancer cells with LYPD1 expression (Amy et al., 2020). Moreover, LYPD1 has been recognized as an oncogenic driver in hepatocellular carcinoma (Chen J et al., 2020). LYPD1 can directly inhibit the formation of endothelial cell network, and regulate anti-angiogenic properties of cardiac fibroblasts. Interestingly, both circFN1 and LINC02381 can bind to hsa-miR-1301-3p, which can integrate with both CD84 and LYPD1. All these ncRNA and mRNA have been shown to be overexpressed in EMS. Therefore, according to this, we can infer that circFN1 and LINC02381 can trigger immune response dysregulation of EMS via manipulating expression of CD84 and LYPD1. Nevertheless, further experimental validation is urgently needed.

circGLIS2 is situated on the plus strand of chromosome 16p13.3 and derived from the known protein coding gene GLIS2. Chen et al. found that circGLIS2 was differentially expressed in colorectal cancer and it can activate NF-κB pathway by sponge miR-671 and further promote the production of pro-inflammatory chemokine (Chen Y et al., 2020). Except for this, circGLIS2 is poorly understood in other diseases. In our research, we demonstrated that circGLIS2 was differentially expressed in EMS tissues. Similarly, circGLIS2 can also regulate the occurrence and development of EMS diseases through ceRNA mechanism. circGLIS2 can also regulate the expression of FAM163A and binding with hsa-miR-3619-5p.

Since miRNA and mRNA combinations are not one-to-one correspondence, a single miRNA may bind to multiple mRNAs. FAM163A, located on chromosome 1q25.2 and encoding a 167-amino acid protein, is also recognized as neuroblastoma-derived secretory protein (NDSP) or C1ORF76. In squamous cell lung carcinoma, Liu et al. found that FAM163A interacts with 14-3-3β to promote ERK phosphorylation and thus affect lung cancer cell proliferation (Liu et al., 2019). This is the first time that up-regulated expression of FAM163A in EMS tissues has been shown under the regulation of circGLIS2. After reviewing previous studies, we speculated that FAM163A can also promote the proliferation of EMS cells, but further experiments are needed to demonstrate this.

lncRNA IGFL2-AS1 is an antisense transcript of IGF like family member 2 (IGFL2) gene. It is located on chromosome 19 with three exons. Accumulating evidence suggests that its aberrant expression is associated with regulation of cellular and pathological processes of several cancer through ceRNA, including gastric cancer (Ma et al., 2020), basal-like breast cancer (Wang et al., 2021). In EMS, we utilize next-generation sequencing to reveal differential expression of lncRNA IGFL2-AS1 and determined the ceRNA mechanism can combine with hsa-miR-138-5p to promote NTM expression. NTM, one of the target gene of lncRNA IGFL2-AS1, was shown in our GSEA analysis to participate in the epithelial-mesenchymal transition of EMS (**Supplementary Material**) Increasing numbers of experiments have improved that EMT is involved in inducing the invasion and migration of endometrial epithelial cells, and this process is of great significance in establishment of endometriosis (Matsuzaki and Darcha, 2012). Epithelial-mesenchymal transition (EMT) is the loss of polarity and change of epithelial cells to a mesenchymal phenotype. Epithelial cell invasiveness was increased during EMT. Endometriosis tissues were also shown to have more e-cadherin negative cells than healthy endometrium, while N-cadherin, Twist, Slug, and Snail were all elevated in endometriosis tissues (Bartley et al., 2014).

In summary, this work provided, for the first time, a preliminary overview of differentially expressed lncRNA, circRNA, and mRNA in EMS. This research adds to our understanding of the pathophysiology of EMS and offers clinical therapeutic options. However, *in vivo* and *in vitro* investigations are needed to confirm the precise function of lncRNA and circRNA in EMS.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/ **Supplementary Material**. For further details, please contact the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Woman Hospital of Zhejiang University School of Medicine (NO. 20190012). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ZXM designed and analyzed the research study. ZJW revised the article. YMC and ZLY conducted experiments. WJZ, YQ, LTT, and XXX wrote the article, GXY and MXQ collected and analyzed the data and all authors have read and approved the final manuscript.

FUNDING

This study was funded by the National Key R&D Program of China (grant number 2017YFC1001202), the National Natural Science Foundation of China (grant numbers 81974225 and 82171636) and the Natural Science Foundation of Zhejiang Province (grant number Y20H160278).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.828238/full#supplementary-material

REFERENCES

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting Effective microRNA Target Sites in Mammalian mRNAs. *Elife* 4, e05005. Aug 12;4. doi:10.7554/eLife.05005

Bartley, J., Jülicher, A., Hotz, B., Mechsner, S., and Hotz, H. (2014). Epithelial to Mesenchymal Transition (EMT) Seems to Be Regulated Differently in Endometriosis and the Endometrium. *Arch. Gynecol. Obstet.* 289 (4), 871–881. doi:10.1007/s00404-013-3040-4

Calpe, S., Wang, N., Romero, X., BergerBerger, S. B., Lanyi, A., Engel, P., et al. (2008). The SLAM and SAP Gene Families Control Innate and Adaptive Immune Responses. *Adv. Immunol.* 97, 177–250. doi:10.1016/S0065-2776(08)00004-7

Chen, C., Dong, Z., Hong, H., Dai, B., Song, F., Geng, L., et al. (2020). circFN1 Mediates Sorafenib Resistance of Hepatocellular Carcinoma Cells by Sponging miR-1205 and Regulating E2F1 Expression. *Mol. Ther. - Nucleic Acids* 22, 421–433. Sep 2. doi:10.1016/j.omtn.2020.08.039

Chen, J., Yang, X., Liu, R., Wen, C., Wang, H., Huang, L., et al. (2020). Circular RNA GLIS2 Promotes Colorectal Cancer Cell Motility via Activation of the NF-Kb Pathway. *Cell Death Dis* 11 (9), 788. Sep 23. doi:10.1038/s41419-020-02989-7

Chen, Y., Zhao, Y., Chen, J., Peng, C., Zhang, Y., Tong, R., et al. (2020). ALKBH5 Suppresses Malignancy of Hepatocellular Carcinoma via m6A-Guided Epigenetic Inhibition of LYPD1. *Mol. Cancer* 19 (1), 123. Aug 10. doi:10.1186/s12943-020-01239-w

Cuenca, M., Sintes, J., Lányi, Á., and Engel, P. (2019). CD84 Cell Surface Signaling Molecule: An Emerging Biomarker and Target for Cancer and Autoimmune Disorders. *Clin. Immunol.* 204, 43–49. doi:10.1016/j.clim.2018.10.017

Dweep, H., and Gretz, N. (2015). miRWalk2.0: a Comprehensive Atlas of microRNA-Target Interactions. *Nat. Methods* 12 (8), 697. doi:10.1038/nmeth.3485

El-Toukhy, T. (2020). Prevalence of Endometriosis: How Close Are We to the Truth? *Bjog: Int. J. Obstet. Gy* 128 (4), 666. doi:10.1111/1471-0528.16466

Greenbaum, H., Galper, B.-E. L., Decter, D. H., and Eisenberg, V. H. (2021). Endometriosis and Autoimmunity: Can Autoantibodies Be Used as a Non-invasive Early Diagnostic Tool? *Autoimmun. Rev.* 20 (5), 102795. doi:10.1016/j.autrev.2021.102795

Huang, X. x., Zhang, Q., Hu, H., Jin, Y., Zeng, A. l., Xia, Y. b., et al. (2020). A Novel Circular RNA circFN1 Enhances Cisplatin Resistance in Gastric Cancer via Sponging miR-182-5p. *J. Cel Biochem* 122, 1009–1020. Jan 2. doi:10.1002/jcb.29641

Jafarzadeh, M., and SoltaniSoltani, B. M. (2020). Long Noncoding RNA LOC400043 (LINC02381) Inhibits Gastric Cancer Progression through Regulating Wnt Signaling Pathway. *Front. Oncol.* 10, 562253. Oct 23;10. doi:10.3389/fonc.2020.562253

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucl. Acids Res.* 42 (Database issue), D92–D97. doi:10.1093/nar/gkt1248

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., et al. (2011). Large-scale Prediction of Long Non-coding RNA Functions in a Coding-Non-Coding Gene Co-expression Network. *Nucleic Acids Res.* 39, 3864–3878. doi:10.1093/nar/gkq1348

Liu, N., Zhou, H., Zhang, X., Cai, L., Li, J., Zhao, J., et al. (2019). FAM163A, a Positive Regulator of ERK Signaling Pathway, Interacts with 14-3-3β and Promotes Cell Proliferation in Squamous Cell Lung Carcinoma. *Ott* Vol. 12, 6393–6406. Aug 13;12. doi:10.2147/OTT.S214731

Liu, Y., Ma, J., Cui, D., Fei, X., Lv, Y., and Lin, J. (2020). LncRNA MEG3-210 Regulates Endometrial Stromal Cells Migration, Invasion and Apoptosis through P38 MAPK and PKA/SERCA2 Signalling via Interaction with Galectin-1 in Endometriosis. *Mol. Cell Endocrinol.* 513, 110870. Aug 1. doi:10.1016/j.mce.2020.110870

Lo, A. A., Johnston, J., Li, J., Mandikian, D., Hristopoulos, M., Clark, R., et al. (2021). Anti-LYPD1/CD3 T-cell-dependent Bispecific Antibody for the Treatment of Ovarian Cancer. *Mol. Cancer Ther.* 20 (4), 716–725. doi:10.1158/1535-7163.MCT-20-0490

Ma, Y., Liu, Y., Pu, Y. S., Cui, M. L. Zhi-Jun., Mao, Z. J., Li, Z. Z., et al. (2020). LncRNA IGFL2-AS1 Functions as a ceRNA in Regulating ARPP19 through Competitive Binding to miR-802 in Gastric Cancer. *Mol. Carcinog* 59 (3), 311–322. doi:10.1002/mc.23155

Matsuzaki, S., and Darcha, C. (2012). Epithelial to Mesenchymal Transition-like and Mesenchymal to Epithelial Transition-like Processes Might Be Involved in the Pathogenesis of Pelvic Endometriosis†. *Hum. Reprod. Mar.* 27 (3), 712–721. doi:10.1093/humrep/der442

Nothnick, W. B., Al-Hendy, A., and Lue, J. R. (2015). Circulating Micro-RNAs as Diagnostic Biomarkers for Endometriosis: Privation and Promise. *J. Minimally Invasive Gynecol.* 22, 719–726. doi:10.1016/j.jmig.2015.02.021

Paci, P., Colombo, T., and Farina, L. (2014). Computational Analysis Identifies a Sponge Interaction Network between Long Non-coding RNAs and Messenger RNAs in Human Breast Cancer. *BMC Syst. Biol.* 8, 83. Jul 17. doi:10.1186/1752-0509-8-83

Piñero, J., Saüch, J., Sanz, F., and Furlong, L. I. (2021). The DisGeNET Cytoscape App: Exploring and Visualizing Disease Genomics Data. *Comput. Struct. Biotechnol. J.* 19, 2960–2967. May 11. doi:10.1016/j.csbj.2021.05.015

Sapkota, Y., Steinthorsdottir, V., Steinthorsdottir, V., Morris, A. P., Fassbender, A., Rahmioglu, N., et al. (2017). Meta-analysis Identifies Five Novel Loci Associated with Endometriosis Highlighting Key Genes Involved in Hormone Metabolism. *Nat. Commun.* 8, 15539. doi:10.1038/ncomms15539

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Sun, Y., Wang, X., and Bu, X. (2021). LINC02381 Contributes to Cell Proliferation and Hinders Cell Apoptosis in Glioma by Transcriptionally Enhancing CBX5. *Brain Res. Bull.* 176, 121–129. doi:10.1016/j.brainresbull.2021.07.009

Symons, L. K., Miller, J. E., Kay, V. R., Marks, R. M., Liblik, K., Koti, M., et al. (2018). The Immunopathophysiology of Endometriosis. *Trends Mol. Med.* 24 (9), 748–762. doi:10.1016/j.molmed.2018.07.004

Wang, H., Shi, Y., Chen, C.-H., Wen, Y., Zhou, Z., Yang, C., et al. (2021). KLF5-induced lncRNA IGFL2-AS1 Promotes Basal-like Breast Cancer Cell Growth and Survival by Upregulating the Expression of IGFL1. *Cancer Lett.* 515, 49–62. Sep 1. doi:10.1016/j.canlet.2021.04.016

Wang, J., and Zhao, Q. (2020). Linc02381 Exacerbates Rheumatoid Arthritis through Adsorbing miR-590-5p and Activating the Mitogen-Activated Protein Kinase Signaling Pathway in Rheumatoid Arthritis-fibroblast-like Synoviocytes. *Cel Transpl.* 29, 096368972093802. doi:10.1177/0963689720938023

Wang, X. (2008). miRDB: A microRNA Target Prediction and Functional Annotation Database with a Wiki Interface. *Rna* 14 (6), 1012–1017. doi:10.1261/rna.965408

Zhang, T., De Carolis, C., Man, G. C. W., and Wang, C. C. (2018). The Link between Immunity, Autoimmunity and Endometriosis: a Literature Update. *Autoimmun. Rev.* 17 (10), 945–955. doi:10.1016/j.autrev.2018.03.017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

CASE REPORT
published: 04 April 2022
doi: 10.3389/fgene.2022.832582

Genetic Analysis of a Pedigree With Antithrombin and Prothrombin Compound Mutations and Antithrombin Heterozygotes

*Haiyue Zhang, Yiling Hu, Dongli Pan, Yuehua Xv and Weifeng Shen**

Department of Clinical Laboratory, The First Hospital of Jiaxing, The Affiliated Hospital of Jiaxing University, Jiaxing, China

Background and Aims: Antithrombin (AT) is the most important physiological inhibitor *in vivo*, and coagulation factor II (FII) or prothrombin is a coagulation factor vital to life. The purpose of our research was to illustrate the connection between gene mutations and the corresponding deficiencies of AT and FII.

Methods: Functional and molecular analyses were performed. The possible impact of the mutation was analyzed by online bioinformatics software. ClustalX-2.1-win and PyMol/Swiss-Pdb Viewer software were used for conservative analyses and to generate molecular graphic images, respectively.

Results: The proband showed a lower limb venous thrombosis and acute pulmonary embolism infarction with reduced AT activity (50%). His mother, with subcutaneous ecchymosis, had reduced activities of AT and FII, of 44 and 5%, respectively. Molecular analysis showed that both the proband and his mother carried c.964A > T (p.Lys322stop) heterozygotes in *SERPINC1*. The difference was that his mother carried homozygous c.494C > T (p.Thr165Met) in *F2*, while the proband was wild type. Bioinformatics and model analysis indicated that mutations may destroy the function and structure of AT and FII protein.

Conclusion: This study identified a novel mutation of *SERPINC1* and a missense mutation of *F2*, which may be the molecular mechanism leading to AT and FII deficiency in this family. It will help genetic diagnosis and counseling for thrombotic families.

Keywords: *F2*, *SERPINC1*, deep vein thrombosis, acute pulmonary embolism, subcutaneous ecchymosis, novel mutation

OPEN ACCESS

Edited by:
Rana Dajani,
Hashemite University, Jordan

Reviewed by:
María Eugenia De La Morena-Barrio,
University of Murcia, Spain
Nadia Akawi,
United Arab Emirates University,
United Arab Emirates

****Correspondence:***
Weifeng Shen
jyzhy526@163.com

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *10 December 2021*
Accepted: *03 March 2022*
Published: *04 April 2022*

Citation:
Zhang H, Hu Y, Pan D, Xv Y and Shen W (2022) Genetic Analysis of a Pedigree With Antithrombin and Prothrombin Compound Mutations and Antithrombin Heterozygotes. Front. Genet. 13:832582. doi: 10.3389/fgene.2022.832582

INTRODUCTION

Venous thromboembolism (VTE) encompasses deep vein thrombosis (DVT) and pulmonary embolism (PE), caused by a variety of factors (Caspers et al., 2012). The pathogenesis of VTE is multifactorial, involving the interaction between clinical risk factors and thrombotic tendency, mainly including two types: hereditary and acquired. Surgery, trauma, sedentary, pregnancy, and cancer are considered acquired risk factors of VTE. Studies have demonstrated that genetic factors are responsible for more than 60% of common thrombotic susceptibility (Yue et al., 2019).

Antithrombin (AT) is a physiological anticoagulant, mainly synthesized by the liver, with a half-life of about 2.4 days (Liu et al., 2021). The mature AT molecule has 432 amino acids, including six

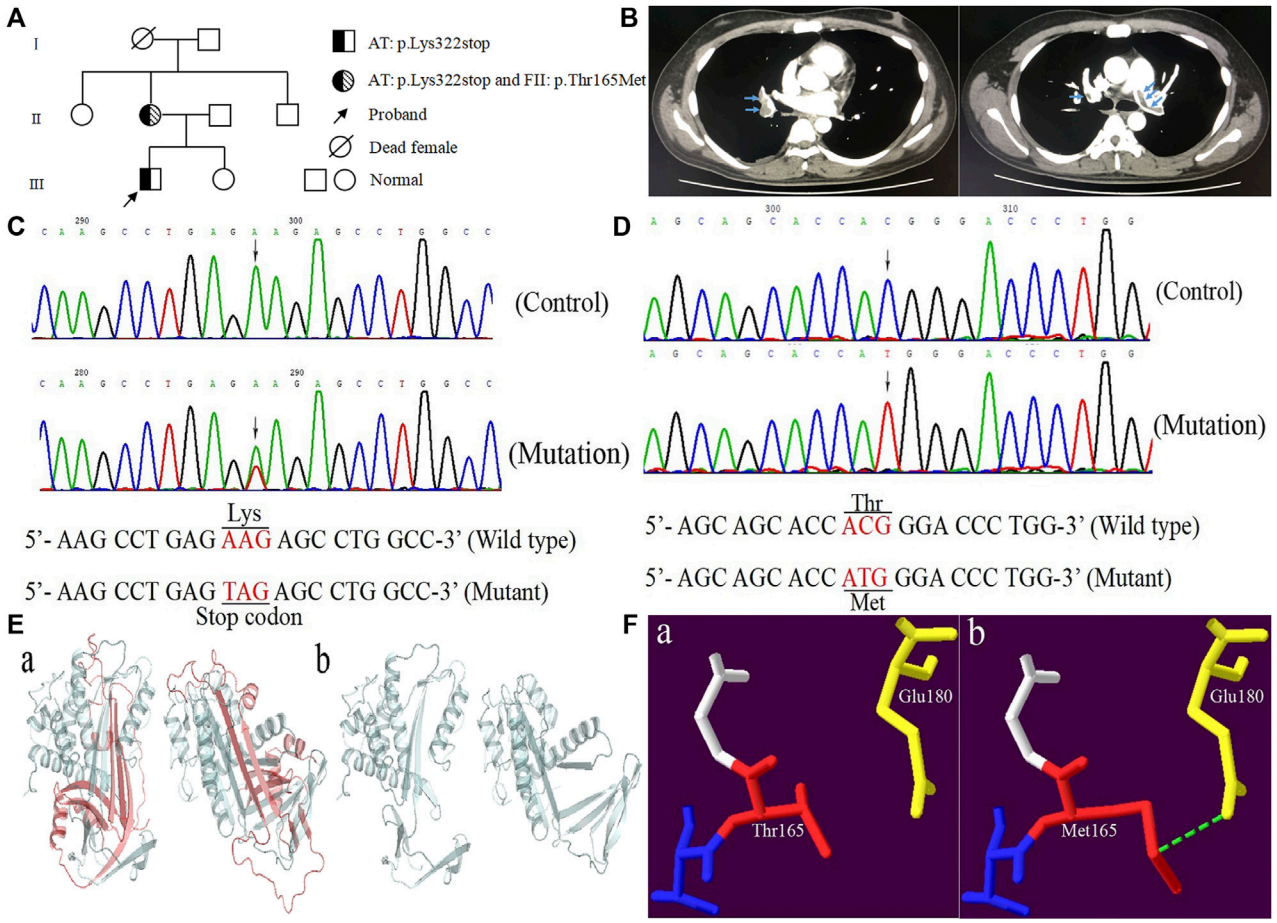

FIGURE 1 | The clinical and genetic data of the proband. **(A)** Pedigree chart of the family. **(B)** The computed tomography angiography of the proband. **(C)** Sequence diagrams of *SERPINC1*:p.Lys322stop. **(D)** Sequence diagrams of *F2*:p.Thr165Met. **(E, F)** Model analysis diagrams: **(A, B)** AT:p.Lys322stop and **(A, B)** FII: p.Thr165Met before and after mutations.

cysteine residues that form three intramolecular disulfide bonds: Cys8-Cys128, Cys21-Cys95, and Cys247-Cys430. There are also four asparagine residues (Asn95, Asn135, Asn155, and Asn192) (Kottke-Marchant and Duncan, 2002). As a serine protease inhibitor belonging to the serine protease inhibitor superfamily, AT can inhibit activated coagulation factors II and X, and to a lesser extent activated factors IX, XI, and XII (Bafunno and Margaglione, 2010). Inherited antithrombin deficiency was first described by Egeberg in 1965 and is the main genetic factor for thrombosis, leading to a 20-fold increase in the risk of venous thromboembolism. It is found in 2–5% of patients with VTE (Rossi et al., 2008).

Prothrombin (FII, coagulation factor II) is a multidomain glycoprotein that is vital to life and an attractive target for anticoagulation therapy (Chinnaraj et al., 2018). Due to bleeding complications, mice lacking prothrombin die prematurely during the embryonic stage (Sun et al., 1998). FII is an allosteric enzyme regulated by sodium binding, controlled by five amino acid residues (Thr540, Arg541, Glu592, Arg596, and Lys599). Mutations in these residues may prevent FII from being inhibited by antithrombin, leading to continuous activation of FII, prone to thrombotic events (Tang et al., 2020). FII is synthesized by hepatocytes into a single polypeptide precursor composed of 622 amino acids. After extensive post-translational modification, FII is secreted into the plasma in its mature form and circulates in the plasma at a concentration of 0.1 mg/ml, with a half-life of about 60 h (Vostal and McCauley, 1991). Hereditary FII deficiency is an autosomal recessive inheritance with an estimated prevalence of 1:2,000,000 people. Heterozygotes with a normal *F2* gene are rarely detected clinically as FII activity (FII:C) is usually within the normal range and hardly results in any bleeding symptoms (Lefkowitz et al., 2003; Kuijper et al., 2013).

In this paper, we recruited a Chinese patient with lower limb venous thrombosis and acute pulmonary embolism infarction. Gene mutation analysis was performed to detect the patient's genetic lesions, and finally a novel heterozygous nonsense mutation was found in the *SERPINC1* gene. It is worth noting that his mother carried the heterozygous nonsense mutation in *SERPINC1* and a homozygous missense mutation in *F2*, with subcutaneous ecchymosis.

TABLE1 | The laboratory data of the proband.

	Measure parameters	Data (proband)	Data (mother)	Reference ranges
Peripheral blood	White blood cells	7.62×10^9/L	5.17×10^9/L	$3.97\sim9.15\times10^9$/L
	Red blood cells	5.11×10^{12}/L	4.54×10^{12}/L	$4.09\sim5.74\times10^{12}$/L
	Hemoglobin	149 g/L	137 g/L	131~172 g/L
	Platelets	373×10^9/L	208×10^9/L	$85\sim303\times10^9$/L
Blood chemistry	Na	139 mmol/L	131 mmol/L	130~147 mmol/L
	K	3.98 mmol/L	3.56 mmol/L	3.5~5.1 mmol/L
	Cl	105 mmol/L	106 mmol/L	96~108 mmol/L
	Total protein	82 g/L	78 g/L	60~83 g/L
	Aspartate aminotransferase	30 u/L	34 u/L	8~40 u/L
	Alanine aminotransferase	53 u/L	44u/L	10~64 u/L
	Blood urea nitrogen	5.5 mmol/L	4.3 mmol/L	2.5~7.1 mmol/L
	Creatinine	63 umol/L	65 umol/L	62~115 umol/L
	Uric acid	478 umol/L	330 umol/L	160~430umol/L
	Total cholesterol	6.15 mmol/L	4.5 mmol/L	2.33~5.7 mmol/L
	Triglyceride	3.2 mmol/L	1.57 mmol/L	0.56~1.7 mmol/L
	LDL	4.13 mmol/L	2.33 mmol/L	1.3~4.3 mmol/L
	HDL	0.85 mmol/L	1.6 mmol/L	0.8~1.8 mmol/L
	Apolipoprotein A1	1.3 g/L	1.58 g/L	1.06~1.88 g/L
	Apolipoprotein B	1.39 g/L	1.03 g/L	0.46~1.13 g/L
	Lipoprotein (a)	0.15 g/L	0.13 g/L	<0.3 g/L
	Apolipoprotein E	5.8 mg/dl	3.3 mg/dl	2.9~5.3 mg/dl
	sdLDL-C	1.9 mmol/L	0.8 mmol/L	0.246~1.393 mmol/L
	Free fatty acid	0.75 mmol/L	0.25 mmol/L	0.1~0.45 mmol/L
	Troponin I	2.8 pg/ml	1.3 pg/ml	<30 pg/ml
	BNP	58.9 pg/ml	55.1 pg/ml	5~115 pg/ml
	Glucose	5.1 mmol/L	4.1 mmol/L	3.9~6.1 mmol/L
	C-reactive protein	8.23 mg/L	1.16 mg/L	0~6 mg/L
Coagulation study	PT -INR	0.92	0.82	0.8~1.2
	APTT	28.6 S	98 S	22.3~38.7 S
	Fibrinogen	2.1 g/L	3.3 g/L	1.8~3.5 g/L
	D-dimer	3.92 mg/L	0.23 mg/L	<0.55 mg/L
	Protein C activity	116%	110%	70~140%
	Protein S activity	88.5%	92.2%	63~130%
	Antithrombin activity	50%	44%	85~120%
	Prothrombin activity	85.3%	5%	50~150%
	FDP	12 mg/L	0~5 mg/L	0~5 mg/L
	LAC	1.1	<1.2	<1.2
Anti-cardiolipin antibody	IgG	—	—	
	IgA	—	—	
	IgM	—	—	

LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; sdLDL-C, small dense low density lipoprotein cholesterol; BNP, N-terminal B-type natriuretic peptide precursor; PT-INR, prothrombin time international normalized ratio; APTT, activated partial thromboplastin time; FDP, fibrinogen degradation products; LAC, lupus anticoagulant.

CASE PRESENTATION

A Chinese patient with lower limb venous thrombosis and acute pulmonary embolism infarction was enrolled from southeast China (**Figure 1A**). The proband, a 24-year-old man, presented to our hospital because of chest tightness for 8 h, feeling weak, walking unsteadily, and left thigh being thicker than before. His B-ultrasound showed enlargement of the right ventricle, moderate pulmonary hypertension, a small amount of pericardial effusion, and thrombosis in the left common iliac vein, external iliac vein, superficial femoral vein, deep femoral vein, popliteal vein, and peroneal vein. The computed tomography angiography (CTA) showed filling defects in both lung lobes and part of the arteries, leading to the consideration of pulmonary embolism (**Figure 1B**). To identify the possible cause of thrombosis in this patient, we conducted screening for genetic risk factors predispose to DVT, and the results showed that the proband's antithrombin activity (AT:A) was reduced to 50% (reference range: 85~120%), parallel decrease in antithrombin antigen content (AT:Ag) was the same as AT:A, remaining at 49 mg/dl (reference range: 80~120 mg/dl), the anticardiolipin antibody was negative, the serum homocysteine and coagulation factor levels were normal, the activities of PS and PC were within the normal range, and blood lipids were higher (**Table 1**). The other secondary risk factors of thrombophilia were also ruled out. Finally, the patient was successfully treated with pulmonary

angiography, inferior cavity arteriography, vascular thrombolysis, thrombus aspiration, and inferior vena cava filter implantation.

The proband's mother was a 45-year-old female, who discovered coincidently during the pedigree study of the proband an AT:A of 44%, AT:Ag of 46 mg/dl, and FII:C of 5%. Other parameters were normal. She was prone to subcutaneous ecchymosis, and had no thrombosis symptoms.

LABORATORY INVESTIGATIONS

Subjects

The study protocol was approved by the Review Board of The First Hospital of Jiaxing and The Affiliated Hospital of Jiaxing University and the study participants gave informed consent. Whole family members (proband and six members) were enrolled and diagnosed by B-ultrasound, CTA, and laboratory examinations.

Genetic Analysis

Genomic DNA was isolated from peripheral blood mononuclear cells using the TIANamp Genomic DNA Kit (TIANGEN, Beijing, China). All exons of *SERPINC1* and *F2* gene along with their intron-exon boundaries and untranslated regions of 3′ and 5′ were amplified by PCR with primers designed on the genomic sequences of AT and FII (GenBank accession numbers are X68793.1 and M17262.1) on a thermal cycler (ABI Thermocycler 2720; ABI, Foster City, California, United States). The PCR products were identified by 1.2% agarose gel electrophoresis, and the positive products were purified and sent to Personal Gene Technology Corporation (Shanghai, China) for direct sequencing. Sanger sequencing revealed that the proband and his mother took c.964A > T (p.Lys322stop) in exon five of *SERPINC1* (NM_000488.4) (**Figure 1C**). The difference was that his mother also carried c.494C > T (p.Thr165Met) in exon six of *F2* (NM_000506.5) (**Figure 1D**), while the proband was wild type. The novel variant was checked in 120 normal individuals.

In-Silico and Protein Structural Analysis

Homologous sequence alignment results showed that Lys322 was not highly conserved among the homologous species. However, 43 of the 143 amino acids deleted by p.Lys322stop were highly conserved among homologous species (*Pan troglodytes*, *Macaca mulatta*, *Canis lupus* familiaris, *Bos taurus*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, and *Oryza sativa* Japonica Group). Conservative analysis showed that Thr165 was located in the highly conserved residues in the conserved region between residues 145 and 185 (Rungroj et al., 2012). The forecasting results of AT: p.Lys322stop was "disease causing" corresponding to "MutationTaster", and the consequence of FII: p.Thr165Met was "polymorphism". Model analysis showed that the 143 amino acid residues deletion caused by p.Lys322stop mutation had an obvious change compared with the previous protein structure (**Figure 1E**). For p.Thr165Met, the Thr165 was located in the kringle one domain of FII. Once substituted by Met165, the extended side chain formed another hydrogen bond with Glu180 (**Figure 1F**).

DISCUSSION

Hereditary AT deficiency is an autosomal dominant thrombotic disease that is associated with potential risk factors for the development of DVT. Even small changes in the wild-type sequence can alter the function of the gene and cause clinical manifestations (Luxembourg et al., 2011). Hereditary FII deficiency is an autosomal recessive inheritance that is related to the lower procoagulant activity. The HGMD database (http://www.hgmd.cf.ac.uk/ac/a11.php) contains more than 480 *SERPINC1* gene mutations and 72 *F2* gene mutations have been identified. According to differences in plasma activity and antigen levels, defects can be divided into two types: quantitative (type I) synthetic protein deficiency or qualitative (type II) defects.

The mutation (p.Arg197stop) of *SERPINC1* can lead to recurrent DVT, leg vein insufficiency, varicose vein resection, crural ulcers, and a family history of venous thrombosis (Michiels et al., 1995). The mutation of p.Glu271stop is associated with recurrent DVT, cerebral artery thrombosis and pulmonary embolism (Tarantino et al., 1999). In the present study, we identified a novel mutation (c.964A > T/p.Lys322stop) of *SERPINC1* in a Chinese young man with lower limb venous thrombosis and acute pulmonary embolism infarction.

The nonsense mutation (c.964A > T/p.Lys322stop) which causes Lys322 was replaced by a stop codon (UAG), resulting in the production of truncated proteins, the disappearance of the glycosylation site Asn 192, and the disulfide bond site Cys247-Cys430. The study by Michiels JJ et al. pointed out that the absence of cross-reactive substances in the patients' plasma indicated that p.Arg197stop either prevented the formation of stable mRNA or the translated peptide was rapidly degraded (Michiels et al., 1995). It has been reported that the 13387-9delG mutation resulted in the loss of the disulfide bond between Cys247 and Cys430, impairing the secretion and stability of the truncated AT protein associated with intracellular degradation (Wang et al., 2005). Since both c.964A > T (p.Lys322stop) and 13387-9delG mutations will cause the loss of the disulfide bond between Cys247 and Cys430, we assumed that c.964A > T (p.Lys322stop) caused the reduction of AT:A and AT:Ag by the same mechanism as 13387-9delG. In addition, the truncation of the AT protein caused by c.964A > T (p.Lys322stop) resulted in the loss of the P1-P1' (Arg393-Ser394) bond, which could not play the role of inactivating the protease.

According to the results of our family study, the proband's mother carried c.494C > T (p.Thr165Met), which is thought to be related to Xinjiang Kazakh thrombotic disease (Ge et al., 2014) and may play a role in kidney stone disease (Rungroj et al., 2012). However, the association of this mutation with Xinjiang Kazakh thrombotic disease may be the result of the interaction of genes and complex environmental factors. The proband's mother had a

FII:C level of 5%, the association with thrombus, if there is any, is relatively weak.

The protein model analysis showed that Thr165 (an amino acid with an uncharged and polar side chain) was replaced by Met165 (an amino acid with a non-polar side chain), which resulted in the formation of another hydrogen bond with Glu180, the change in hydrogen bond-forming was likely to consequently alter protein structure and function. The c.494C > T (p.Thr165Met) substitution may affect kringle one domain glycosylation by destroying an O-glycan site (Webber et al., 2006). The kringle one domain is important in the interaction of proteins with clotting factors, and it is believed to play a role in binding mediators and regulating proteolytic activity (Patthy et al., 1984). Thus, we considered c.494C > T (p.Thr165Met) as leading to a decrease FII:C in this family. Since the pathogenicity of c.494C > T (p.Thr165Met) was not explicitly mentioned in previous reports, and some gene defects may only show functional consequences under specific conditions. We do not rule out the existence of other mechanisms that may be involved in the reduction of FII:C in this family. It should be verified by more basic experiments in the future. A somewhat puzzling finding was that the mother was homozygous while the proband was wild type. Therefore, we decided to investigate the potential cause. We found that the mother of the proband, who worked in a tannery while pregnant, was at high risk for exposure to metal salts, mainly chromates, in the tannery. Hexavalent chromium can be taken up into cells via nonspecific ionophores, causing DNA damage by generating reactive intermediates (Arslan et al., 1987). This may be the reason why the proband was wild type.

Antithrombin is an important protein that inhibits the conversion of fibrinogen by thrombin, and the reduced activity caused by gene mutation provides conditions for thrombosis. FII plays a key role in the activation of the agglutination pathway. The reduction or lack of its activity weakens the activation of the coagulation system, and the demand for antithrombin activity is no longer prominent. The commonly used anticoagulant (dabigatran) exerts an anticoagulant effect by reducing the activity of FII. Furthermore, patients with FII deficiency have clinical manifestations ranging from life-threatening spontaneous bleeding to epistaxis (Kuijper et al., 2013). AT deficiency is a high-risk factor for thrombophilia (Corral et al., 2018), and it may reduce the risk of bleeding due to FII deficiency. The simultaneous decline of FII:C and AT:A may allow physiological coagulation and anticoagulation homeostasis to be maintained.

In conclusion, in the present study, we analyzed a pedigree with antithrombin and prothrombin compound mutations and antithrombin heterozygotes, the proband had AT deficiency, and his mother had compound AT and FII deficiencies. For antithrombin deficiency, it is necessary to evaluate the blood clotting factor levels of the patient and his relatives. This reduction in antithrombin levels may give the patient an age-independent risk of thrombosis.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the First Hospital of Jiaxing and the Affiliated Hospital of Jiaxing University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

HZ wrote the manuscript, YH collected and cleaned the data, DP reviewed the medical records, YX helped in data collection, and WS designed the study. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the Zhejiang Medicine and Health under Grant 2022RC077, the "Venus" Talent Training of the First Hospital of Jiaxing of Zhejiang Province of China under Grant 2020-QMX-25, and Clinical Laboratory Medical Diagnostics Fund of the First Hospital of Jiaxing of Zhejiang Province of China under Grant 2019-cx-03.

REFERENCES

Arslan, P., Beltrame, M., and Tomasi, A. (1987). Intracellular Chromium Reduction. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 931 (1), 10–15. doi:10.1016/0167-4889(87)90044-9

Bafunno, V., and Margaglione, M. (2010). Genetic Basis of Thrombosis. *Clin. Chem. Lab. Med.* 48 Suppl 1, S41–S51. doi:10.1515/CCLM.2010.361

Caspers, M., Pavlova, A., Driesen, J., Harbrecht, U., Klamroth, R., Kadar, J., et al. (2012). Deficiencies of Antithrombin, Protein C and Protein S - Practical Experience in Genetic Analysis of a Large Patient Cohort. *Thromb. Haemost.* 108 (2), 247–257. doi:10.1160/TH11-12-0875

Chinnaraj, M., Planer, W., and Pozzi, N. (2018). Structure of Coagulation Factor II: Molecular Mechanism of Thrombin Generation and Development of Next-Generation Anticoagulants. *Front. Med.* 5, 281. doi:10.3389/fmed.2018.00281

Corral, J., de la Morena-Barrio, M. E., and Vicente, V. (2018). The Genetics of Antithrombin. *Thromb. Res.* 169, 23–29. doi:10.1016/j.thromres.2018.07.008

Ge, X.-h., Zhu, F., Wang, B.-l., Wang, C.-m., Zhu, B., Guan, S., et al. (2014). Association between Prothrombin Gene Polymorphisms and Hereditary Thrombophilia in Xinjiang Kazakhs Population. *Blood Coagul. Fibrinolysis Int. J. Haemost. Thromb.* 25 (2), 114–118. doi:10.1097/MBC.0b013e328364ba00

Kottke-Marchant, K., and Duncan, A. (2002). Antithrombin Deficiency. *Arch. Pathol. Lab. Med.* 126 (11), 1326–1336. doi:10.5858/2002-126-1326-AD

Kuijper, P. H. M., Schellings, M. W. M., Van de Kerkhof, D., Nicolaes, G. A. F., Reitsma, P., Halbertsma, F., et al. (2013). Two Novel Mutations in the Prothrombin Gene Identified in a Patient with Compound Heterozygous

Type 1/2 Prothrombin Deficiency. *Haemophilia* 19 (5), e304–e306. doi:10.1111/hae.12180

Lefkowitz, J. B., Weller, A., Nuss, R., Santiago-Borrero, P. J., Brown, D. L., and Ortiz, I. R. (2003). A Common Mutation, Arg457Gln, Links Prothrombin Deficiencies in the Puerto Rican Population. *J. Thromb. Haemost.* 1 (11), 2381–2388. doi:10.1046/j.1538-7836.2003.00420.x

Liu, S., Wang, H., Xu, Q., Luo, S., Jin, Y., Yang, L., et al. (2021). Type II Antithrombin Deficiency Caused by a Novel Missense Mutation (p.Leu417Gln) in a Chinese Family. *Int. J. Haemost. Thromb.* 32 (1), 57–63. doi:10.1097/MBC.0000000000000973

Luxembourg, B., Delev, D., Geisen, C., Spannagl, M., Krause, M., Miesbach, W., et al. (2011). Molecular Basis of Antithrombin Deficiency. *Thromb. Haemost.* 105 (4), 635–646. doi:10.1160/TH10-08-0538

Michiels, J. J., van der Luit, L., van Vliet, H. H., Jochmans, K., and Lissens, W. (1995). Nonsense Mutation Arg197stop in a Dutch Family with Type 1 Hereditary Antithrombin (AT) Deficiency Causing Thrombophilia. *Thromb. Res.* 78 (3), 251–254. doi:10.1016/0049-3848(95)90875-g

Patthy, L., Trexler, M., Váli, Z., Bányai, L., and Váradi, A. (1984). Kringles: Modules Specialized for Protein Binding. *FEBS Lett.* 171 (1), 131–136. doi:10.1016/0014-5793(84)80473-1

Rossi, E., Za, T., Ciminello, A., Leone, G., and De Stefano, V. (2008). The Risk of Symptomatic Pulmonary Embolism Due to Proximal Deep Venous Thrombosis Differs in Patients with Different Types of Inherited Thrombophilia. *Thromb. Haemost.* 99 (6), 1030–1034. doi:10.1160/TH08-02-0069

Rungroj, N., Sudtachat, N., Nettuwakul, C., Sawasdee, N., Praditsap, O., Jungtrakoon, P., et al. (2012). Association between Human Prothrombin Variant (T165M) and Kidney Stone Disease. *PloS one* 7 (9), e45533. doi:10.1371/journal.pone.0045533

Sun, W. Y., Witte, D. P., Degen, J. L., Colbert, M. C., Burkart, M. C., Holmbäck, K., et al. (1998). Prothrombin Deficiency Results in Embryonic and Neonatal Lethality in Mice. *Proc. Natl. Acad. Sci. U.S.A.* 95 (13), 7597–7602. doi:10.1073/pnas.95.13.7597

Tang, Y., Zhang, L., Xie, W., Jin, J., Luo, Y., Deng, M., et al. (2020). A Novel Heterozygous Variant in F2 Gene in a Chinese Patient with Coronary Thrombosis and Acute Myocardial Infarction Leads to Antithrombin Resistance. *Front. Genet.* 11, 184. doi:10.3389/fgene.2020.00184

Tarantino, M. D., Curtis, S. M., Johnson, G. S., Waye, J. S., and Blajchman, M. A. (1999). A Novel and De Novo Spontaneous point Mutation (Glu271STOP) of the Antithrombin Gene Results in a Type I Deficiency and Thrombophilia. *Am. J. Hematol.* 60 (2), 126–129. doi:10.1002/(sici)1096-8652(199902)60:2<126::aid-ajh7>3.0.co;2-l

Vostal, J. G., and McCauley, R. B. (1991). Prothrombin Plasma Clearance Is Not Mediated by Hepatic Asialoglycoprotein Receptors. *Thromb. Res.* 63 (3), 299–309. doi:10.1016/0049-3848(91)90133-h

Wang, W.-B., Fu, Q.-H., Ding, Q.-L., Zhou, R.-F., Wu, W.-M., Hu, Y.-Q., et al. (2005). Characterization of Molecular Defect of 13387-9delG Mutated Antithrombin in Inherited Type I Antithrombin Deficiency. *Int. J. Haemost. Thromb.* 16 (2), 149–155. doi:10.1097/01.mbc.0000161570.04883.25

Webber, D., Radcliffe, C. M., Royle, L., Tobiasen, G., Merry, A. H., Rodgers, A. L., et al. (2006). Sialylation of Urinary Prothrombin Fragment 1 Is Implicated as a Contributory Factor in the Risk of Calcium Oxalate Kidney Stone Formation. *FEBS J.* 273 (13), 3024–3037. doi:10.1111/j.1742-4658.2006.05314.x

Yue, Y., Sun, Q., Xiao, L., Liu, S., Huang, Q., Wang, M., et al. (2019). Association of SERPINC1 Gene Polymorphism (Rs2227589) with Pulmonary Embolism Risk in a Chinese Population. *Front. Genet.* 10, 844. doi:10.3389/fgene.2019.00844

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 11 April 2022
doi: 10.3389/fgene.2022.762834

Sex-Specific Differences in MicroRNA Expression During Human Fetal Lung Development

*Nancy W. Lin[1,2†], Cuining Liu[2,3†], Ivana V. Yang[2,4], Lisa A. Maier[1,5], Dawn L. DeMeo[6], Cheyret Wood[3], Shuyu Ye[2], Margaret H. Cruse[2], Vong L. Smith[2], Carrie A. Vyhlidal[7], Katerina Kechris[3] and Sunita Sharma[2]**

[1]Division of Environmental and Occupational Health, National Jewish Health, Denver, CO, United States, [2]Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado School of Medicine, Aurora, CO, United States, [3]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado-Denver Anschutz Medical Campus, Aurora, CO, United States, [4]Division of Bioinformatics and Personalized Medicine, Department of Medicine, University of Colorado School of Medicine, Aurora, CO, United States, [5]Environmental and Occupational Health, Colorado School of Public Health, Aurora, CO, United States, [6]Channing Division of Network Medicine, Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States, [7]Children's Mercy Hospital and Clinics, Kansas City, MO, United States

OPEN ACCESS

Edited by:
Jaira Ferreira de Vasconcellos, James Madison University, United States

Reviewed by:
Pragnya Das, Cooper University Hospital, United States
Tong Zhou, University of Nevada, Reno, United States

***Correspondence:**
Sunita Sharma
sunita.sharma@cuanschutz.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: 08 October 2021
Accepted: 05 January 2022
Published: 11 April 2022

Citation:
Lin NW, Liu C, Yang IV, Maier LA, DeMeo DL, Wood C, Ye S, Cruse MH, Smith VL, Vyhlidal CA, Kechris K and Sharma S (2022) Sex-Specific Differences in MicroRNA Expression During Human Fetal Lung Development. Front. Genet. 13:762834. doi: 10.3389/fgene.2022.762834

Background: Sex-specific differences in fetal lung maturation have been well described; however, little is known about the sex-specific differences in microRNA (miRNA) expression during human fetal lung development. Interestingly, many adult chronic lung diseases also demonstrate sex-specific differences in prevalence. The developmental origins of health and disease hypothesis suggests that these sex-specific differences in fetal lung development may influence disease susceptibility later in life. In this study, we performed miRNA sequencing on human fetal lung tissue samples to investigate differential expression of miRNAs between males and females in the pseudoglandular stage of lung development. We hypothesized that differences in miRNA expression are present between sexes in early human lung development and may contribute to the sex-specific differences seen in pulmonary diseases later in life.

Methods: RNA was isolated from human fetal lung tissue samples for miRNA sequencing. The count of each miRNA was modeled by sex using negative binomial regression models in DESeq2, adjusting for post-conception age, age^2, smoke exposure, batch, and RUV factors. We tested for differential expression of miRNAs by sex, and for the presence of sex-by-age interactions to determine if miRNA expression levels by age were distinct between males and females.

Results: miRNA expression profiles were generated on 298 samples (166 males and 132 females). Of the 809 miRNAs expressed in human fetal lung tissue during the pseudoglandular stage of lung development, we identified 93 autosomal miRNAs that were significantly differentially expressed by sex and 129 miRNAs with a sex-specific pattern of miRNA expression across the course of the pseudoglandular period.

Conclusion: Our study demonstrates differential expression of numerous autosomal miRNAs between the male and female developing human lung. Additionally, the expression of some miRNAs are modified by age across the pseudoglandular stage in

a sex-specific way. Some of these differences in miRNA expression may impact susceptibility to pulmonary disease later in life. Our results suggest that sex-specific miRNA expression during human lung development may be a potential mechanism to explain sex-specific differences in lung development and may impact subsequent disease susceptibility.

Keywords: microRNA, lung development, pulmonary disease, sex-specific, gene expression, human

INTRODUCTION

Sex-specific differences in human lung maturation begin *in utero*. Using histologic staging, the fetal lungs of human females are more mature than males from 20 to 32 weeks of gestation (Naeye et al., 1974). Measurements of the lecithin-sphingomyelin ratio in fetuses between 28 and 40 weeks of gestation also suggest that human female lungs are about 1.2–2.5 weeks more mature than male lungs (Torday et al., 1981). However, the time point at which sexual differences emerge in human lung development, the biological pathways differing by sex, and the regulatory mechanisms underlying these differences are not well understood. Using transcriptomic profiling in human fetal lung tissue, we have previously shown that sexual dimorphism in human lung development occurs as early as the pseudglandular stage (Kho et al., 2016a). In part, these differences are hypothesized to be driven by sex hormones and sex-specific differences in endogenous corticosteroid handling (Boucher et al., 2014). For example, estrogens are known to alter surfactant production (Seaborn et al., 2010). In addition, glucocorticoids, which are known to be essential to lung development, are influenced by endogenous androgens during development in a sex-specific manner (Provost et al., 2013). While these mechanisms are known to impact the structural development of the lung in the later stages of gestation, the impact of sex-specific differences in early lung development have not been fully elucidated. Thus, better characterization of sex-specific differences in early lung development, a critical time period of airway growth and branching morphogenesis, may offer insight into the biological mechanisms that underlie sex-based differences in respiratory disease.

Epigenetic regulation of human development is well recognized and may provide insights into additional mechanisms that explain sex-specific differences in lung maturation. Furthermore, an understanding of the sex-specific epigenetic changes that regulate human lung development may also help to understand the developmental origins of diseases that exhibit sex-specific differences in prevalence. For example, the ability to characterize sex-specific differences in epigenetic processes may offer insight into the biological mechanisms that enhance susceptibility to childhood respiratory diseases among males, including respiratory distress syndrome, bronchopulmonary dysplasia, and childhood asthma (Naeye et al., 1971).

Of particular interest are microRNAs (miRNAs), small RNAs (~21–24 nucleotides) that are important regulators of gene expression. miRNAs target protein-coding genes by cleavage of target messenger RNA (mRNA), inhibition of translation, and/or mRNA deadenylation (Bhaskaran et al., 2009). A single miRNA can have many gene targets, with 30% of the human genome thought to be miRNA targets (Lewis et al., 2005). As such, miRNAs are essential regulators of many critical biological processes. Animal models have previously demonstrated the importance of miRNAs in lung development, including airway growth and branching morphogenesis. Proteins that are essential for miRNA processing appear to have a prominent role in murine lung development. For example, a seminal study showed that the inactivation of Dicer, a protein required to generate mature miRNAs, leads to significant branching defects in mouse embryonic lung tissue resulting in large epithelial pouches (Harris et al., 2006). In addition, Ago1 and Ago2, members of the Argonaute family which regulate small RNAs, were shown to be enriched in the distal epithelium and mesenchyme of the early developing lung, suggesting miRNA mediated gene regulation occurs in a localized manner (Lü et al., 2005). Specific miRNAs have also been shown to have a significant impact on the developing lung, including the overexpression of miR-127 in early fetal rat lung development, which causes defective branching and terminal bud formation (Bhaskaran et al., 2009).

Animal models have also established that there are sex-specific differences in miRNA expression within the developing lung (Mujahid et al., 2013) and that these sex-specific differences may impact subsequent disease risk. Given this existing body of evidence, it is plausible that there are sex differences in miRNA levels during early human lung development, and that these sex-specific differences in miRNA expression may impact subsequent disease risk. However, to-date sex-specific differences in miRNA expression have not been explored in human biospecimens, particularly using high-throughput profiling methods. To address this existing knowledge gap, we generated the first miRNA sequencing profiles of early human fetal lung tissue samples to investigate differential miRNA expression between developing male and female lungs during the pseudoglandular stage of lung development. We hypothesize that differences in miRNA expression are present in early human fetal lung development between sexes, and this may be a mechanism to explain sex-specific differences in susceptibility to pulmonary diseases later in life.

METHODS

Sample Acquisition and Phenotypic Characteristics

Human fetal lung tissue samples were collected as part of a prenatal tissue retrieval program sponsored by the National Institute of National Child Health and Development, the University of Maryland Brain and Tissue Bank for Developmental Disorders (Baltimore, MD), and the Center for Birth Defects Research (University of Washington; Seattle, WA). The study was designated an institutional review board (IRB) exempt protocol by the University of Missouri-Kansas City Pediatric IRB, Partners Human Research Committee IRB, and the Colorado Multiple Institutional Review Board (COMIRB). Due to sample de-identification, limited maternal and fetal phenotypic characteristics were available for each sample including gestational age and sex. Sample sex was previously confirmed based on paired gene expression data (Kho et al., 2016a) by classifying samples as female or male based on their expression of X- and Y-chromosome genes. The age of samples was determined using estimated days post-conception. Intrauterine cigarette smoke exposure (based on placental cotinine concentration) (Vyhlidal et al., 2013) is a well-known confounder of fetal lung development and was directly measured in the samples using the Cotinine Direct ELISA kit (Calbiotech, Spring Valley, CA). Unmeasured confounders were accounted for using the RUV method described below in miRNA profiling.

miRNA Profiling

We extracted total RNA from 30 mg of homogenized prenatal lung tissue with the miRNeasy Mini Kit per manufacturer instructions (Qiagen; Valencia, CA, United States). Samples were block-randomized by age, sex, and smoke exposure status to four batches of miRNA library preparation (Small RNA Sequencing Kit v3 for Illumina Platforms; Bio Scientific) and sequencing (HiSeq2500; Illumina; San Diego, CA, United States). The resulting reads were trimmed for low-quality base calls and Illumina adaptor sequence using cutadapt (Martin, 2011), and then mapped to counts of known miRNAs using miR-MaGiC (Russell et al., 2018) with reference to the miRbase (Kozomara and Griffiths-Jones, 2014) v22.1 database. Samples passing quality-control ($\geq 1 \times 10^5$ mapped reads, having available phenotypic information, and being within the 7–17 weeks pseudoglandular period) and autosomal miRNAs passing pre-filtering criteria (non-zero counts in at least 25% of samples; passes DESeq2 independent filtering algorithm (Bourgon et al., 2010); and not mapped to X or Y chromosome) were tested for differences by sex. Using RUV-Seq (Risso et al., 2014), we also obtained four RUVr factors representing unmeasured miRNA expression heterogeneity for inclusion in regression models.

Statistical Modeling of Sex Differences

We evaluated if miRNA features differed by sex in two respects. First, we identified autosomal miRNAs with varying average expression levels between male and female samples regardless of sample estimated post-conception age (i.e., effect present across the entire pseudoglandular developmental stage). Using DESeq2 (Love et al., 2014), we modeled the count of each miRNA (outcome) by sample sex (explanatory variable of interest; indicator variable for male or female), adjusting for age and age squared (age^2) (Torday et al., 1981) (age is measured in days post-conception; age^2 is a quadratic term to capture potential non-linear effects), smoke exposure (≥or <7.5 ng cotinine/g placenta), technical batch (indicator variable for batch 1, 2, 3, or 4), and four inferred covariates to capture additional unmeasured confounders (k = 4 RUVr components). A statistically significant difference in mean miRNA levels by sex was defined by a likelihood ratio test at a multiple testing corrected q-value (Storey, 2002) < 0.05.

Second, we screened for autosomal miRNAs with sex-specific "age-trajectories" using a likelihood ratio test to evaluate whether adding an age-by-sex interaction significantly improved model fit. A significant interaction (q-value < 0.05) implies that the pattern of miRNA expression levels by age were distinct between male and female participants. For example, a miRNA may increase in male samples, yet decrease or remain the same in female samples.

Functional Interpretation of Sex-Differing miRNAs. To interpret the functional impact of miRNAs differentially expressed by sex and with sex-specific age trajectories, we used miRNAtap (Pajak and Simpson, 2021) to identify predicted gene expression targets regulated by each miRNA. In addition, we sorted the list of miRNAs tested by descending p-value and used pre-ranked gene set enrichment analyses in miEAA to conduct pathway analyses (Subramanian et al., 2005; Kern et al., 2020). Statistically significant enrichment in a pathway (q-value < 0.05) indicates that miRNAs associated with the pathway appear at the top of the list (lower p-values) more frequently than would be expected by random chance. Annotations of miRNAs to pathways were based on the miRWalk 2.0 database, as curated by the miEAA developers (Dweep et al., 2011).

TABLE 1 | Characteristics of human fetal lung samples analyzed.

	Female	Male	All
N	132	166	298
Age (dpc)	87.0 (76.0, 96.5)	89.0 (76.0, 96.0)	87.0 (76.0, 96.0)
IUS (exposed)	62 (47.0%)	77 (46.4%)	139 (46.6%)

Study Reproducibility

Additional methodological details are available online (**Supplementary File S1**: Detailed Methods). Code for the statistical analyses and processed miRNA-sequencing data is available at github. com/chooliu/miRNASexDimorphismFetalLung. Raw and processed miRNA-sequencing data are pending deposition approval at the NCBI GEO database.

RESULTS

Characteristics of the Human Fetal Lung Tissue Samples

miRNA-sequencing profiles were generated on human fetal lung tissue samples ranging in gestational age from 54 to 117 days within the pseudoglandular histological stage of human lung development.

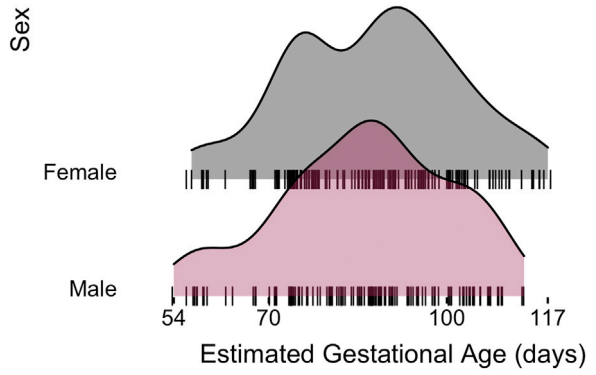

FIGURE 1 | Human fetal lung tissue samples used for this analysis demonstrates a similar distribution of estimated gestational ages observed in each sex.

The final set of 298 samples was comprised of 166 male (56%) and 132 (44%) female samples (**Table 1**). The same approximate distribution of age ranges was observed in each sex (**Figure 1**; no significant difference in distribution based on Wilcoxon signed-rank test, $p = 0.41$). Based on placental cotinine, our previously validated biomarker of intrauterine smoke exposure, 139 (47%) of the samples had been exposed to intrauterine smoke.

Lung miRNA Expression Profiles During the Pseudgolandular Stage of Human Lung Development

After pre-processing and quality control, our final dataset consisted of 809 miRNAs measured in 298 human lung samples. Of the known sample physical characteristic variables, gestational age appeared to explain the largest variability in miRNA expression, with age explaining anywhere

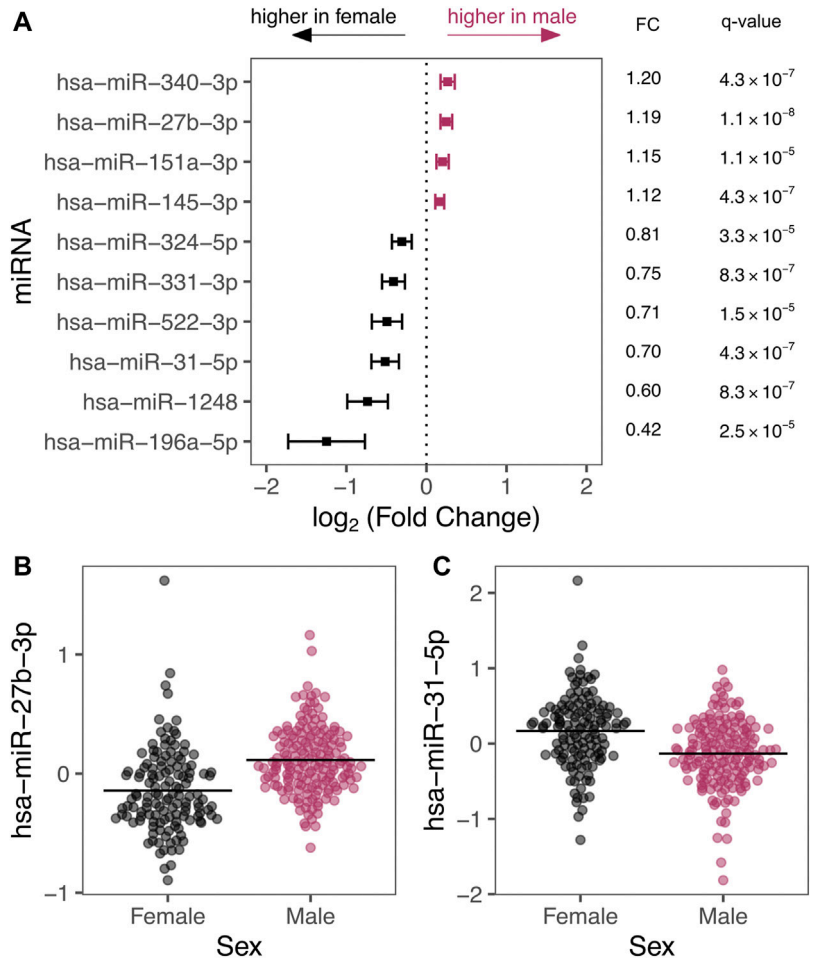

FIGURE 2 | (A) Top 10 miRNAs by p-value are depicted with their effect size expressed as $\log_2$ (fold change) with 95% confidence intervals. Estimated fold changes (miRNA counts in males compared to female samples) and q-values are also presented. **(B,C)** show example scatter plots of residualized miRNA levels (y-axis) plotted by sex (x-axis). Residuals are displayed to illustrate sex-differences after adjusting for covariates including, age, age (Torday et al., 1981), smoke exposure, technical batch, and the four RUV-inferred covariates to emphasize effects of sex.

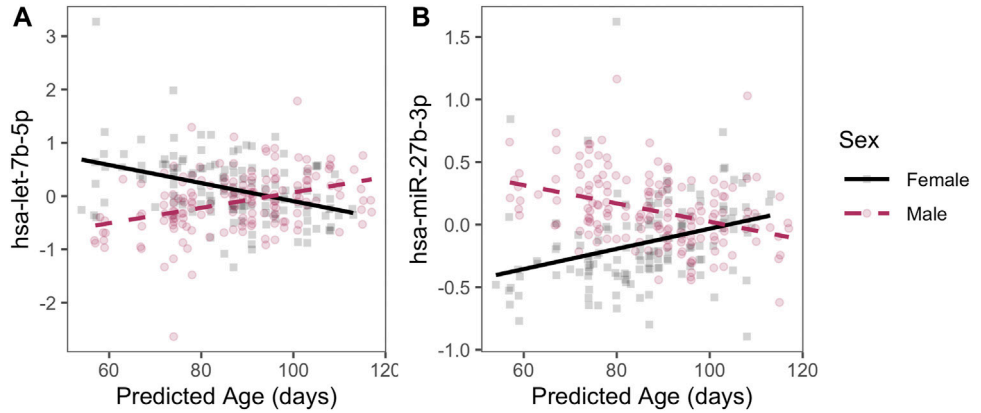

FIGURE 3 | Examples of sex-specific trajectories. Two miRNAs with statistically significant age-by-sex interactions are plotted with residualized expression (*y*-axis) across age (*x*-axis) during the pseudoglandular period differs by sex (indicated by color). Residuals are displayed to illustrate sex-by-age differences after adjusting for smoke exposure, technical batch, and the four RUV-inferred covariates to emphasize patterns by age and sex.

from 0 to 52% of the variance in the expression of each miRNA. In comparison, sex tended to explain only up to 8% of each miRNA's variance.

Differential Expression of miRNAs by Sex

We detected 93 autosomal miRNAs with significant differential expression by sex after correcting for covariates. A subset of differentially expressed miRNAs are shown in **Figure 2** (full list in **Supplementary Table S1**). The predicted mRNA targets for each of these miRNAs are also provided in **Supplementary Table S1**. Interestingly, these miRNAs differentially expressed by sex include miRNAs predicted to regulate gene expression targets with known roles in pseudoglandular lung development: For example, *hsa-miR-27b-3p* is predicted to regulate peroxisome proliferator-activated receptor-γ (PPARγ) and *hsa-miR-196a-5p* is predicted to regulate multiple Hox transcription factors including homeobox B7 (HOXB7).

Pathway enrichment analyses also support the possibility that sex-specific miRNAs may collectively regulate biological pathways relevant to pseudoglandular lung development. Sex-associated miRNAs were significantly enriched in numerous pathways (gene set enrichment q-value < 0.05; full results in **Supplementary Table S2**), including pathways with previously identified roles in lung development, such as Wnt-signaling, vascular endothelial growth factor (VEGF), fibroblast growth factor (FGF), and transforming growth factor beta (TGFβ) signaling pathways. Sex-associated miRNAs were also significantly enriched with genes associated with androgen and estrogen signaling.

Sex-Specific miRNA Expression Trajectories During Human Lung Development

In addition, we detected 129 autosomal miRNAs with significant linear age-sex interactions; that is, a sex-specific age trajectory of miRNA expression over the course of the pseudoglandular period. Two examples of miRNAs with a significant age-sex interaction—*hsa-let-7b-5p* and *hsa-miR-27b-3p*—have their expression levels by age plotted for illustration (**Figure 3**; q-value < 0.05, full list of significant results in **Supplementary Table S3**). Sex-associated miRNAs identified in this manner also were significantly enriched in multiple pathways, including epidermal growth factor (EGF) and muscarinic acetylcholine receptor (mAChR) signaling (full results in **Supplementary Table S4**).

DISCUSSION

While sex-specific differences that exist during *in utero* lung development are well-recognized by histological measurements, much remains to be understood about the differences between males and females during early lung development and their impact on subsequent disease risk. Our study applies high-throughput, untargeted miRNA-sequencing to rare human fetal lung tissue samples and characterizes differences in miRNA expression by sex for the first time, highlighting novel regulatory features and biological pathways that differ by sex in early prenatal lung development. Sex-specific differences in these miRNAs may potentially explain differences in lung maturation and in subsequent respiratory disease susceptibility later in life that differs between males and females.

We detected 93 miRNAs with significantly different average expression levels by sex during the pseudoglandular stage of human lung development, as well as 129 miRNAs with significant age-by-sex interactions indicative of sex-specific miRNA expression age-trajectories across the pseudoglandular stage. The sex-specific miRNAs include some miRNAs previously shown to impact lung development. For example, *hsa-let-7b-5p* was both identified as significantly differentially expressed (higher levels in males on average throughout pseudoglandular

period) and identified as a miRNA with a sex-specific trajectory (expression increasing by age in males, but decreasing by age in females). The *let-7* family of miRNAs are well-known regulators of human lung cell proliferation (Johnson et al., 2007), and in mouse models appear to play crucial roles in regulating both the timing and the cell differentiation processes associated with airway branching, possibly mediated by let-7's downstream impacts on TGFβ signaling and epithelial-mesenchymal interactions (Gulman et al., 2019).

We also identified other miRNAs with differences by sex that have not been as richly characterized but are predicted to regulate gene expression targets essential to lung development. For example, *hsa-miR-27b-3p* was differentially expressed (higher in males) and had a sex-specific age trajectory (decreasing by age in males, increasing in females). A predicted regulatory target of hsa-miR-27b-3p is the transcription factor PPARγ. PPARγ has a multifaceted role in lung development, including the modulation of inflammatory and cell differentiation processes in the lung (Simon et al., 2006; Lecarpentier et al., 2019) and its negative feedback with the Wnt/β-catenin and TGFβ signaling pathways (Lecarpentier et al., 2019), two processes essential to branching morphogenesis (De Langhe and Reynolds, 2008; Saito et al., 2018). Disruption of PPARγ in mice models alters lung volume and airway resistance (Simon et al., 2006). Consequently, the differences in the levels and age-trajectories of *let-7b-5p* and *miR-27b-3p* by sex observed here in human lung samples are one plausible mechanism underlying differences in lung histology and maturation rates between males and females during early human lung development.

Given the importance of pseudoglandular stage in airway development, it has long been hypothesized that insults to lung development during this period contribute to the developmental origins of chronic airway diseases (Stocks and Sonnappa, 2013). Impaired gene expression during this period likely alters susceptibility to chronic obstructive pulmonary disease (COPD) (Warburton et al., 2006) and asthma (Melén et al., 2011). These two diseases differ in etiology and pathophysiology by sex. Because sex-differing pseudoglandular miRNAs appear to regulate known airway development genes and pathways, our findings support the potential for sex-differing miRNA regulation very early in lung development to contribute to later sex differences in airway disease. Our group has previously shown that miRNA are biomarkers for abnormal lung function in asthmatic children in a sex-specific way (Kho et al., 2016b). In our current study, we note that our sex-differing miRNA results overlapped with miRNA markers derived from lung samples in COPD and asthma cases (Cañas et al., 2021), such as *hsa-miR-31-5p*, *hsa-miR-338*-3p, and again the *let-7* family with the same direction of effect. Importantly, lung development miRNAs that are differentially expressed by sex may be associated with the pathogenesis of several different pulmonary diseases by playing a key role in the inflammation that characterizes many chronic lung diseases such as cystic fibrosis, asthma, and emphysema (Sessa and Hata, 2013).

Our findings also have links to other diseases with known sex differences. For example, the higher levels of *miR-27b-3p* in pseudoglandular male lung is notable because PPARγ agonists have been proposed for prevention and treatment of bronchopulmonary dysplasia (BPD) (Simon et al., 2006), a disease more common in males. We also detected sex-specific pseudoglandular trajectories (increase by age at greater rates in males) of *miR-29a-3p* and *mir-150*, two miRNAs with putative BPD-associations but elusive causal impacts on BPD severity in hypoxia-based experimental models (Nardiello and Morty, 2016). Interestingly, the respective predicted gene targets for these genes include collagen type 1 alpha 1 (*COL1A1*)/tropoelastin (*ELN*) and homeobox A5 (*HOXA5*)/C-Myc are known to have structural implications for early lung development (Schmidt et al., 2020).

To our knowledge, sex differences in miRNA expression in early in human lung development have not been previously profiled using high-throughput small RNA sequencing. Previous miRNA-sequencing studies in mice have focused on differential expression of miRNAs by sex during the later canalicular and saccular stages (Mujahid et al., 2013). Importantly, regulatory networks associated with murine miRNAs that are differentially expressed by sex during lung development have been shown to include androgen signaling (Mujahid et al., 2013). Furthermore, mouse models also show that miRNAs are modulated by androgens in developing lungs (Bouhaddioui et al., 2016). Importantly, many of the impacts of sex hormones established from mouse models in later lung development may still apply to earlier lung development. These observations include the finding that androgens alter fetal lung fibroblast maturation in murine lung *via* EGF and TGFβ signaling events (Dammann et al., 2000). Additionally, androgens appear to block endogenous glucocorticoids, which are essential to normal lung development and are promoters of surfactant production (Provost et al., 2013). More broadly, androgens are thought to exert an inhibitory effect, while estrogens exert a stimulatory effect on lung development (Carey et al., 2007). Our results in early human lung development mirror the animal data in the later stages of gestation. Notably, pathway enrichment analysis of the predicted gene expression targets of our sex-specific miRNAs demonstrate enrichment in several sex-hormone related pathways, including both estrogen metabolism and androgen receptor signaling pathways.

We also demonstrate that the expression of several human fetal lung miRNAs was modified by age, demonstrating changes in gene expression across the pseudoglandular stage that varied by sex. These results also suggest a link to sex-steroid regulation of early lung development. The expression of *hsa-let-7b* increases across the pseudoglandular stage in males, while its expression does not change in females. *Hsa-let-7b*, a tumor-suppressor miRNA, has been previously shown to be important in both health and disease. In addition to its functions noted above, the let-7 family has also been shown to be an essential regulator of several endocrine systems. Transgenic mouse models have demonstrated that in combination with Lin28, the let-7 family of miRNAs are crucial to the timing of puberty and that overexpression of this dyad may result in the delayed onset of puberty (Zhu et al., 2010). While there are myriad mechanisms that regulate the expression of this system including dietary manipulation, it is well recognized that hormonal changes

have been shown to have a significant impact (Sangiao-Alvarellos et al., 2015). Animal studies suggest that sex hormones may be partly responsible for the sex-specific differences seen in lung development. While these results suggest that sex-specific lung development miRNAs may influence sex hormone metabolism and signaling during the pseudoglandular stage of development, additional confirmatory studies are necessary to delineate the impact of lung development miRNAs on sex hormone levels.

Although these results provide a comprehensive evaluation of the sex-specific miRNA expression profile of early human lung development, there are several limitations to this study. Our samples were obtained from a fetal tissue biorepository with limited maternal and fetal information on each sample. Therefore, we could not comment on confounders such as maternal and fetal comorbidities. We attempted to address potential unmeasured confounders (i.e., unknown intrauterine exposures, maternal comorbidities, etc.) using the RUV method to adjust for unexplained miRNA expression heterogeneity, which includes possible unmeasured technical and biological confounders. We also limited our current study to investigation of the autosomal microRNAs. While there is a high density of miRNA on the X chromosome, the functional and statistical analyses of these loci remain complex and incomplete (Di Palo et al., 2020). We were unable to do validation of miRNA expression analyses due to the limitations in the quantity of RNA available from these extremely rare pseudoglandular samples. However, many of the predicted gene expression targets have been previously identified as demonstrated differential expression by sex in our earlier work suggesting the biologic plausibility of our current work (Kho et al., 2016a). In addition, our sex-specific miRNA expression analysis of lung development was limited to fetal lung tissue samples from the pseudoglandular stage of development. Therefore, we are not able to determine the changes in the sex-specific miRNA profiles that occur during the later stages of gestation, which may have implications for our understanding of subsequent disease risk. Additionally, the fetal sample age was estimated from days post-conception and not confirmed by histology, which could have introduced classification bias based on errors in estimation. Finally, although our results suggest that sex-specific miRNAs are important in lung development and may impact future disease risk, additional functional validation in animal models would be necessary to confirm the impact of altering these miRNAs on subsequent disease risk.

In conclusion, our study demonstrates sex-specific differences in miRNA expression between the male and female developing human lung and establishes their role in branching morphogenesis and airway development during the pseudoglandular time period. This study suggests that miRNAs may regulate the sex differences seen in lung development and that these differences in miRNA expression may be potential mechanisms to explain sex-specific differences in disease susceptibility to pulmonary disease later in life. Furthermore, our findings provide evidence that sex-specific miRNA expression profiles of lung development using human fetal lung tissues can be used to elucidate novel biological mechanisms that regulate the sex differences of developmental processes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/geo/, accession number: GSE200153.

ETHICS STATEMENT

The requirement for ethical review and approval was waived by the University of Missouri-Kansas City Pediatric IRB, Partners Human Research Committee IRB, and the Colorado Multiple Institutional Review Board (COMIRB). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

NL wrote the first draft of the manuscript, CL performed the statistical analysis and contributed to manuscript writing. IY, LM, DD, CV, KK, CW, AK, and SS contributed to conception and design of the study. SY, MC, and VS helped with regulatory and programmatic support. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by NIH grants R01 HL125734 (PI: SS), R01 HL 097144 (PI: Weiss), P01HL132825 (PI: Weiss, DD), IGNITE First in Women Precision Medicine Award from BWH R01 HG011393 (PI:DD).

ACKNOWLEDGMENTS

We would like to acknowledge Scott T. Weiss (Channing Division of Network Medicine, Brigham and Women's Hospital, Boston MA), Robert P. Chase Channing Division of Network Medicine, Brigham and Women's Hospital, Boston MA), for his support of this work. We would also like to thank J. Steven Leeder and Roger Gaedigk (Children's Mercy Hospital and Clinics, Kansas City, MO) for their support of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.762834/full#supplementary-material

REFERENCES

Bhaskaran, M., Wang, Y., Zhang, H., Weng, T., Baviskar, P., Guo, Y., et al. (2009). MicroRNA-127 Modulates Fetal Lung Development. *Physiol. Genomics* 37, 268–278. doi:10.1152/physiolgenomics.90268.2008.-MicroRNAs

Boucher, E., Provost, P. R., and Tremblay, Y. (2014). Ontogeny of Adrenal-like Glucocorticoid Synthesis Pathway and of 20α-Hydroxysteroid Dehydrogenase in the Mouse Lung. *BMC Res. Notes* 7 (1), 1–10. doi:10.1186/1756-0500-7-119

Bouhaddioui, W., Provost, P. R., and Tremblay, Y. (2016). Expression Profile of Androgen-Modulated microRNAs in the Fetal Murine Lung. *Biol. Sex. Differ.* 7 (1), 1–13. doi:10.1186/s13293-016-0072-z

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent Filtering Increases Detection Power for High-Throughput Experiments. *Proc. Natl. Acad. Sci.* 107 (21), 9546–9551. doi:10.1073/pnas.0914005107

Cañas, J. A., Rodrigo-Muñoz, J. M., Sastre, B., Gil-Martinez, M., Redondo, N., and del Pozo, V. (2021). MicroRNAs as Potential Regulators of Immune Response Networks in Asthma and Chronic Obstructive Pulmonary Disease. *Front. Immunol.* 11 (January), 1–19. doi:10.3389/fimmu.2020.608666

Carey, M. A., Card, J. W., Voltz, J. W., Germolec, D. R., Korach, K. S., and Zeldin, D. C. (2007). The Impact of Sex and Sex Hormones on Lung Physiology and Disease: Lessons from Animal Studies. *Am. J. Physiology-Lung Cell Mol. Physiol.* 293 (2), L272–L278. doi:10.1152/ajplung.00174.2007

Dammann, C. E. L., Ramadurai, S. M., Mccants, D. D., Pham, L. D., and Nielsen, H. C. (2000). Androgen Regulation of Signaling Pathways in Late Fetal Mouse Lung Development1. *Endocrinology* 141 (8), 2923–2929. doi:10.1210/endo.141.8.7615

De Langhe, S. P., and Reynolds, S. D. (2008). Wnt Signaling in Lung Organogenesis. *Organogenesis* 4 (2), 100–108. doi:10.4161/org.4.2.5856

Di Palo, A., Siniscalchi, C., Salerno, M., Russo, A., Gravholt, C. H., and Potenza, N. (2020). What microRNAs Could Tell Us about the Human X Chromosome. *Cell. Mol. Life Sci.* 77 (20), 4069–4080. doi:10.1007/s00018-020-03526-7

Dweep, H., Sticht, C., Pandey, P., and Gretz, N. (2011). miRWalk - Database: Prediction of Possible miRNA Binding Sites by "walking" the Genes of Three Genomes. *J. Biomed. Inform.* 44 (5), 839–847. doi:10.1016/j.jbi.2011.05.002

Gulman, N. K., Armon, L., Shalit, T., and Urbach, A. (2019). Heterochronic Regulation of Lung Development *via* the Lin28-Let-7 Pathway. *FASEB j.* 33 (11), 12008–12018. doi:10.1096/fj.201802702R

Harris, K. S., Zhang, Z., McManus, M. T., Harfe, B. D., and Sun, X. (2006). Dicerfunction Is Essential for Lung Epithelium Morphogenesis. *Pnas* 103 (7), 2208–2213. doi:10.1073/pnas.0510839103

Johnson, C. D., Esquela-Kerscher, A., Stefani, G., Byrom, M., Kelnar, K., Ovcharenko, D., et al. (2007). The Let-7 microRNA Represses Cell Proliferation Pathways in Human Cells. *Cancer Res.* 67 (16), 7713–7722. doi:10.1158/0008-5472.CAN-07-1083

Kern, F., Fehlmann, T., Solomon, J., Schwed, L., Grammes, N., Backes, C., et al. (2020). miEAA 2.0: Integrating Multi-Species microRNA Enrichment Analysis and Workflow Management Systems. *Nucleic Acids Res.* 48 (W1), W521–W528. doi:10.1093/nar/gkaa309

Kho, A. T., Chhabra, D., Sharma, S., Qiu, W., Carey, V. J., Gaedigk, R., et al. (2016). Age, Sexual Dimorphism, and Disease Associations in the Developing Human Fetal Lung Transcriptome. *Am. J. Respir. Cell Mol Biol* 54 (6), 814–821. doi:10.1165/rcmb.2015-0326OC

Kho, A. T., Sharma, S., Davis, J. S., Spina, J., Howard, D., McEnroy, K., et al. (2016). Circulating microRNAs: Association with Lung Function in Asthma. *PLoS One* 11 (6), e0157998–18. doi:10.1371/journal.pone.0157998

Kozomara, A., and Griffiths-Jones, S. (2014). MiRBase: Annotating High Confidence microRNAs Using Deep Sequencing Data. *Nucl. Acids Res.* 42 (D1), D68–D73. doi:10.1093/nar/gkt1181

Lecarpentier, Y., Gourrier, E., Gobert, V., and Vallée, A. (2019). Bronchopulmonary Dysplasia: Crosstalk between PPARγ, WNT/β-Catenin and TGF-β Pathways; the Potential Therapeutic Role of PPARγ Agonists. *Front. Pediatr.* 7, 176. doi:10.3389/fped.2019.00176

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes Are microRNA Targets. *Cell* 120 (1), 15–20. doi:10.1016/j.cell.2004.12.035

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 1–21. doi:10.1186/s13059-014-0550-8

Lü, J., Qian, J., Chen, F., Tang, X., Li, C., and Cardoso, W. V. (2005). Differential Expression of Components of the microRNA Machinery during Mouse Organogenesis. *Biochem. Biophysical Res. Commun.* 334 (2), 319–323. doi:10.1016/j.bbrc.2005.05.206

Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet j.* 17 (1), 10. doi:10.14806/ej.17.1.200

Melén, E., Kho, A. T., Sharma, S., Gaedigk, R., Leeder, J. S., Mariani, T. J., et al. (2011). Expression Analysis of Asthma Candidate Genes during Human and Murine Lung Development. *Respir. Res.* 12 (1), 86. doi:10.1186/1465-9921-12-86

Mujahid, S., Logvinenko, T., Volpe, M. V., and Nielsen, H. C. (2013). MiRNA Regulated Pathways in Late Stage Murine Lung Development. *BMC Dev. Biol.* 13 (1), 13. doi:10.1186/1471-213X-13-13

Naeye, R. L., Burt, L. S., Wright, D. L., Blanc, W. A., and Tatter, D. (1971). Neonatal Mortality, the Male Disadvantage. *Pediatrics* 48 (6), 902–906. doi:10.1542/peds.48.6.902

Naeye, R. L., Freeman, R. K., and Blanc, W. A. (1974). Nutrition, Sex, and Fetal Lung Maturation. *Pediatr. Res.* 8 (3), 200–204. doi:10.1203/00006450-197403000-00008

Nardiello, C., and Morty, R. E. (2016). MicroRNA in Late Lung Development and Bronchopulmonary Dysplasia: the Need to Demonstrate Causality. *Mol. Cell Pediatr* 3 (1), 19. doi:10.1186/s40348-016-0047-5

Pajak, M., and Simpson, T. I. (2021). miRNAtap: miRNAtap: microRNA Targets - Aggregated Predictions. R Package Version 1.28.0.

Provost, P. R., Boucher, E., and Tremblay, Y. (2013). Glucocorticoid Metabolism in the Developing Lung: Adrenal-like Synthesis Pathway. *J. Steroid Biochem. Mol. Biol.* 138, 72–80. doi:10.1016/j.jsbmb.2013.03.004

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples. *Nat. Biotechnol.* 32 (9), 896–902. doi:10.1038/nbt.2931

Russell, P. H., Vestal, B., Shi, W., Rudra, P. D., Dowell, R., Radcliffe, R., et al. (2018). MiR-MaGiC Improves Quantification Accuracy for Small RNA-Seq. *BMC Res. Notes* 11 (1), 1–8. doi:10.1186/s13104-018-3418-2

Saito, A., Horie, M., and Nagase, T. (2018). TGF-β Signaling in Lung Health and Disease. *Ijms* 19 (8), 2460. doi:10.3390/ijms19082460

Sangiao-Alvarellos, S., Manfredi-Lozano, M., Ruiz-Pino, F., León, S., Morales, C., Cordido, F., et al. (2015). Testicular Expression of the Lin28/let-7 System: Hormonal Regulation and Changes during Postnatal Maturation and after Manipulations of Puberty. *Sci. Rep.* 5 (September), 1–13. doi:10.1038/srep15683

Schmidt, A. F., Kannan, P. S., Bridges, J., Presicce, P., Jackson, C. M., Miller, L. A., et al. (2020). Prenatal Inflammation Enhances Antenatal Corticosteroid-Induced Fetal Lung Maturation. *JCI insight* 5 (24). doi:10.1172/jci.insight.139452

Seaborn, T., Simard, M., Provost, P. R., Piedboeuf, B., and Tremblay, Y. (2010). Sex Hormone Metabolism in Lung Development and Maturation. *Trends Endocrinol. Metab.* 21 (12), 729–738. doi:10.1016/j.tem.2010.09.001

Sessa, R., and Hata, A. (2013). Role of microRNAs in Lung Development and Pulmonary Diseases. *Pulm. Circ.* 3 (2), 315–328. doi:10.4103/2045-8932.114758

Simon, D. M., Arikan, M. C., Srisuma, S., Bhattacharya, S., Tsai, L. W., Ingenito, E. P., et al. (2006). Epithelial Cell PPARγ Contributes to normal Lung Maturation. *FASEB j.* 20 (9), 1507–1509. doi:10.1096/fj.05-5410fje

Stocks, J., and Sonnappa, S. (2013). Early Life Influences on the Development of Chronic Obstructive Pulmonary Disease. *Ther. Adv. Respir. Dis.* 7 (3), 161–173. doi:10.1177/1753465813479428

Storey, J. D. (2002). A Direct Approach to False Discovery Rates. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* 64 (3), 479–498. doi:10.1111/1467-9868.00346

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Torday, J. S., Nielsen, H. C., Fencl, Mde. M., and Avery, M. E. (1981). Sex Differences in Fetal Lung Maturation. *Am. Rev. Respir. Dis.* 123 (2), 205–208. doi:10.1164/arrd.1981.123.2.205

Vyhlidal, C. A., Riffel, A. K., Haley, K. J., Sharma, S., Dai, H., Tantisira, K. G., et al. (2013). Cotinine in Human Placenta Predicts Induction of Gene Expression in Fetal Tissues. *Drug Metab. Dispos* 41 (2), 305–311. doi:10.1124/dmd.112.049999

Warburton, D., Gauldie, J., Bellusci, S., and Shi, W. (2006). Lung Development and Susceptibility to Chronic Obstructive Pulmonary Disease. *Proc. Am. Thorac. Soc.* 3 (8), 668–672. doi:10.1513/pats.200605-122SF

Zhu, H., Shah, S., Shyh-Chang, N., Shinoda, G., Einhorn, W. S., Viswanathan, S. R., et al. (2010). Lin28a Transgenic Mice Manifest Size and Puberty Phenotypes Identified in Human Genetic Association Studies. *Nat. Genet.* 42 (7), 626–630. doi:10.1038/ng.593

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 20 May 2022
doi: 10.3389/fgene.2022.847646

Digital Cell Atlas of Mouse Uterus: From Regenerative Stage to Maturational Stage

Leyi Zhang[1,2,3,4], *Wenying Long*[1], *Wanwan Xu*[1], *Xiuying Chen*[1], *Xiaofeng Zhao*[1] *and Bingbing Wu*[1]*

[1]The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, China, [2]Key Laboratory of Tumor Microenvironment and Immune Therapy of Zhejiang Province, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [3]Cancer Institute (Key Laboratory of Cancer Prevention & Intervention, National Ministry of Education), Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [4]Department of Breast Surgery, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

Endometrium undergoes repeated repair and regeneration during the menstrual cycle. Previous attempts using gene expression data to define the menstrual cycle failed to come to an agreement. Here we used single-cell RNA sequencing data of C57BL/6J mice uteri to construct a novel integrated cell atlas of mice uteri from the regenerative endometrium to the maturational endometrium at the single-cell level, providing a more accurate cytological-based elucidation for the changes that occurred in the endometrium during the estrus cycle. Based on the expression levels of proliferating cell nuclear antigen, differentially expressed genes, and gene ontology terms, we delineated in detail the transitions of epithelial cells, stromal cells, and immune cells that happened during the estrus cycle. The transcription factors that shaped the differentiation of the mononuclear phagocyte system had been proposed, being *Mafb*, *Irf7*, and *Nr4a1*. The amounts and functions of immune cells varied sharply in two stages, especially NK cells and macrophages. We also found putative uterus tissue-resident macrophages and identified potential endometrial mesenchymal stem cells (high expression of *Cd34*, *Pdgfrb*, *Aldh1a2*) *in vivo*. The cell atlas of mice uteri presented here would improve our understanding of the transitions that occurred in the endometrium from the regenerative endometrium to the maturational endometrium. With the assistance of a normal cell atlas as a reference, we may identify morphologically unaffected abnormalities in future clinical practice. Cautions would be needed when adopting our conclusions, for the limited number of mice that participated in this study may affect the strength of our conclusions.

Keywords: endometrium, single-cell RNA sequencing, estrus cycle, mononuclear phagocyte system, endometrial mesenchymal stem cells

OPEN ACCESS

Edited by:
Aparna Vasanthakumar, AbbVie, United States

Reviewed by:
Fan Zhou, Tsinghua University, China
Ameya Kulkarni, AbbVie, United States

***Correspondence:**
Bingbing Wu
0012865@zju.edu.cn

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *03 January 2022*
Accepted: *03 May 2022*
Published: *20 May 2022*

Citation:
Zhang L, Long W, Xu W, Chen X, Zhao X and Wu B (2022) Digital Cell Atlas of Mouse Uterus: From Regenerative Stage to Maturational Stage. Front. Genet. 13:847646. doi: 10.3389/fgene.2022.847646

Abbreviations: ESC, Endometrial stromal cells; PCA, principal component analysis; t-SNE, t-distributed stochastic neighbor embedding; GO, Gene Ontology; MMP, Metalloproteinase; NK, natural killer; IFN-γ, interferon-γ; DCs, dendritic cells; Treg, regulatory T cells; scRNA-seq, single-cell RNA sequencing; Pcna, proliferating cell nuclear antigen.

INTRODUCTION

Proper implantation of the embryo in the maternal endometrium is critical for a normal pregnancy, and the sophisticated transformation of the endometrium in the three different menstrual states is regulated by the collaboration of cell populations under the influence of hormones (Kumar et al., 2011).

The attempt to use gene expression to define the menstrual cycle has been seen in previous studies, which reported a strong relationship existing between histopathology and transcriptional profiles of the samples and validated the importance of using molecular profiles to evaluate the endometrial status (Haber et al., 2017). Other than humans, the endometria of mice, rats, and cows have also been analyzed to a certain extent (Reese et al., 2001; Naciff et al., 2002). However, there was little consistency between these microarray-based studies, for the differentially expressed genes reported in each study showed large variability. These molecular profiles remained at the tissue level. The fact that the proportions of the epithelium, stroma, immune cells, and blood vessels in individual specimens were different may cause the variability. However, single-cell analysis can solve this obstacle, as it allows us to detect cell-to-cell variability, possible subpopulations, and rare cell types.

The estrous cycle in mice averages 4–5 days and is a repetitive but dynamic process, reflecting changes in the levels of estradiol and progesterone secreted by the ovarian follicles (Cora et al., 2015). There are different criteria for defining the estrus cycle of mice (Vidal and Filgo, 2017), but no matter how the cycle is defined, mice uteri undergo the same hormone change patterns and the recurrences of regeneration and maturation as humans. In order to match with the recurring physiologic changes in humans, we divided the estrous cycle of mice into two stages: the regeneration stage and the maturation stage. The regenerative stage amounted to the proliferative phase in humans and proestrus and estrus phases in mice, the maturational stage amounted to the secretory phase in humans and the metestrus and the diestrus phases in mice. In doing so, we could seek insights and inspirations from mice single-cell RNA sequencing (scRNA-seq) data to shed light on human endometrium research.

The cognition of the immune conditions in the menstrual cycle has been updated rapidly, yet the amounts or the functional traits of multiple immune cell types are still under debate. The immune cells in decidua have been analyzed thoroughly (Jiang et al., 2018). In sharp contrast, the features of immune cells in the regenerative stage and the maturational stage have not gained enough attention. Up to today, the complicated hormone-induced immune regulation has left a huge riddle for us to solve (Kumar et al., 2014).

Based on the research gaps mentioned above, we aimed to construct a cell atlas of mice uteri including multiple cell types at the single-cell level. We used published scRNA-seq data of two C57BL/6J mice uteri to build an integrated cell atlas from the regenerative endometrium to the maturational endometrium, and elucidated the transitions that happened in epithelial cells, stromal cells, and immune cells during the estrous cycle, hoping to provide new insights into cell dynamics in the uterus and provide a normal reference for future studies under pathologic conditions.

MATERIALS AND METHODS

Datasets Selection and Data Processing

Two datasets from public databases were enrolled in this study. GSE108097 from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/, RRID: SCR_005012), containing scRNA-seq data of 3,756 cells from two 6-to-10-week-old female C57BL/6J mice uteri, was selected to build the digital mouse cell atlas of mouse uterus (Han et al., 2018). Mouse 1 contained 2041 cells, and mouse 2 contained 1715 cells. Based on the expression levels of proliferating cell nuclear antigen, differentially expressed genes of two samples, and gene ontology terms, we concluded that mouse 1 was in the maturational stage and mouse 2 was in the regeneration stage. Another scRNA-seq data of endometrial tissue was obtained from a previous study to compare the cultured endometrial stromal cells (human endometrial cells in the secretory stage were collected and cultured for 8 days in the cell culture medium consisting of a 1:1 mixture of DMEM (Cat No E15-892, GE Healthcare, United States)/Ham's F12 (Cat No E15-890, GE Healthcare), supplemented with 10% FBS and Antibiotic-Antimycotic solution (GE Healthcare) and uncultured ones (endometrial biopsies collected from women in the secretory stage) (Krjutškov et al., 2016). We used R software (version 3.6.3; http://www.Rproject.org, RRID: SCR_001905) and Seurat package of R (version 2.3, https://satijalab.org/seurat/, RRID: SCR_016341) (Satija et al., 2015) to process the data. Seurat is a computational strategy to process and explore scRNA-seq data. The filter criteria of cells were determined as default. It is not necessary to obtain ethical approval because we used published online datasets.

Computational Bioinformatics Analyses of Single-Cell RNA Sequencing Data

Principal component analysis (PCA), uniform manifold approximation and projection (UMAP), t-distributed stochastic neighbor embedding (tSNE), and visualization were performed by the Seurat package of R (Satija et al., 2015). Seurat is a computational strategy to process and explore scRNA-seq data. The fate decisions and pseudotime trajectories of endometrium stromal cells and the mononuclear phagocyte system were reconstructed by the Monocle package of R (version 2.10.1, http://cole-trapnell-lab.github.io/monocle-release/docs/, RRID: SCR_016339) (Cole et al., 2014; Qiu et al., 2017; Wu et al., 2017). In short, Monocle performs differential expression and time-series analysis for single-cell expression experiments, which orders individual cells according to progress through a biological process, without knowing ahead of time which genes define progress through that process. Possible stem cells identification was performed by RaceID3 (https://github.com/dgrun/RaceID, RRID: SCR_017045) and StemID2 (https://github.com/dgrun/StemID, RRID: SCR_017242) (Grün et al., 2015). RaceID3 is an algorithm for rare cell type identification in complex

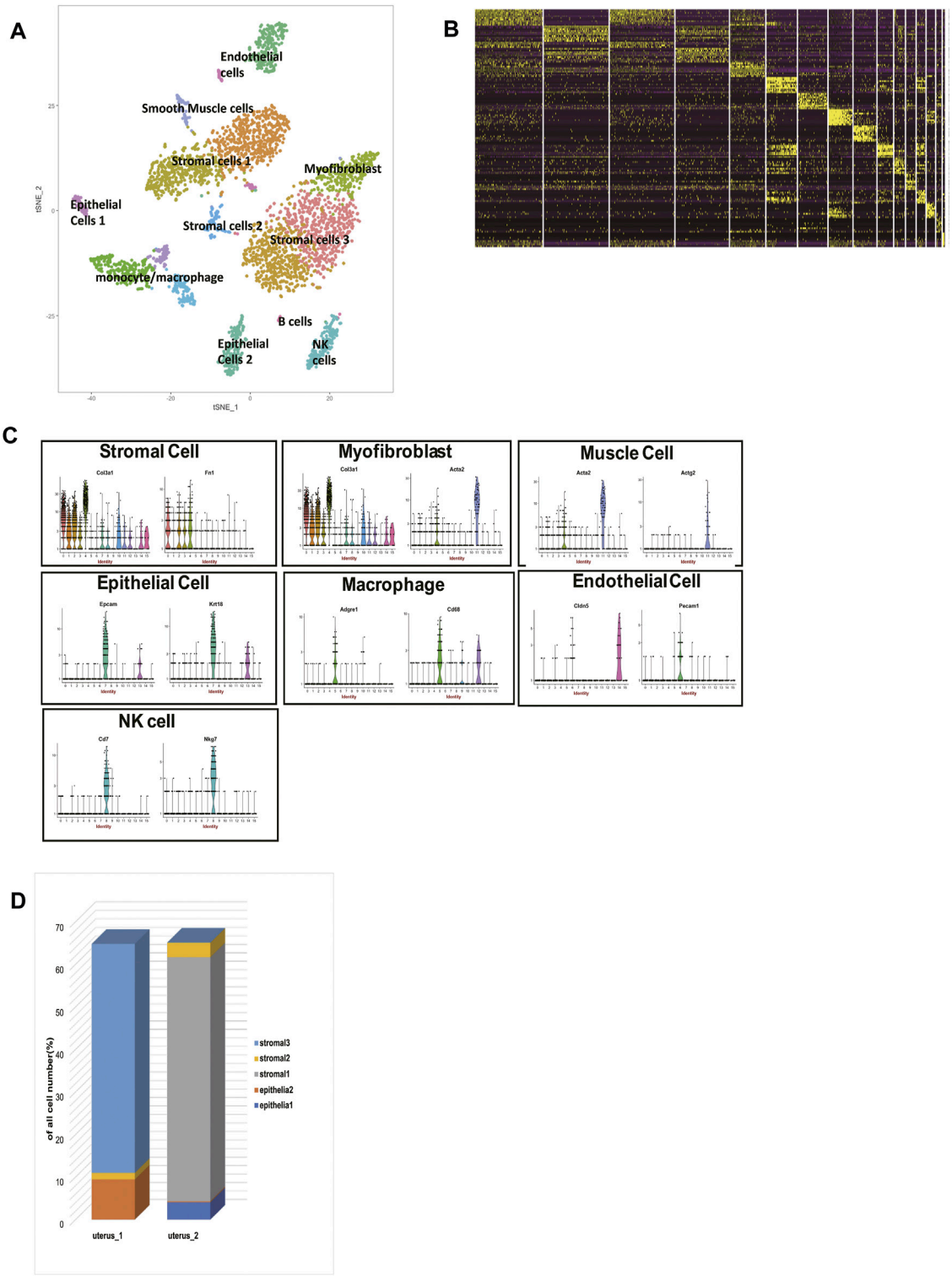

FIGURE 1 | single-cell RNA sequencing analysis of two mice uteri. **(A)** t-distributed stochastic neighbor embedding (t-SNE) diagram of two mice uteri. **(B)** Heatmap of single cells from the two uteri revealed 16 populations. **(C)** Violin plots indicating the expression of marker genes of each cell cluster. **(D)** The cell composition of the two uteri was significantly different. T-SNE: t-distributed stochastic neighbor embedding.

populations of single cells, while StemID2 is an algorithm based on RaceID3 for the inference of differentiation trajectories and the prediction of the stem cell identity.

Functional Enrichment Analyses

Gene Ontology (GO) analyses were conducted through clusterProfiler (clusterProfiler, RRID: SCR_016884) (Yu et al., 2012; Walter et al., 2015) and online GO resource website (Ashburner et al., 2000; 2019). Statistically significant GO terms ($p < 0.05$) were identified.

Histology and Immunofluorescence

Histology and immunofluorescence were conducted according to our previous procedures (Wu et al., 2019). Briefly, vagina smears were firstly used to determine the mouse estrous cycle (**Supplementary Figure S1**) (Bertolin and Murphy, 2014). Then mouse uterine tissues were collected and fixed in 4% (w/v) paraformaldehyde, and then dehydrated in an ethanol gradient. Then the paraffin sections of 10 μm thickness were stained with hematoxylin and eosin. Immunostaining was carried out as follows: The series of 10 μm-thick sections were rehydrated, fixed with 4% (w/v) paraformaldehyde for 30 min, antigen retrieval was conducted by incubating in citrate antigen retrieval solution at 65°C overnight, rinsed three times with PBS, and treated with blocking solution (1% BSA) for 30 min, prior to incubation with primary antibodies at 4°C overnight. The primary antibodies rabbit anti-mouse antibodies against KRT7 (Abcam, ab181598), the primary antibodies mouse anti-mouse antibodies against PCNA (Abcam, ab29), the primary antibodies rat anti-mouse antibodies against CD34 (Biolegend, 119307) were used to detect the expression of selected proteins within the uterine tissues. The goat anti-rat-cy3 secondary antibody (Beyotime Biotechnology, A0507), goat anti-rabbit-488 secondary antibody (Invitrogen, A11008), donkey anti-mouse 546 secondary antibody (Invitrogen, A10036), and DAPI (Beyotime Biotechnology, C1002) were used to visualize the respective primary antibodies and the cell nuclei. All procedures were carried out according to the manufacturer's instructions.

Statistical Analysis

All statistical analyses were performed by Graphpad Prism 8.0 software (https://www.graphpad.com/, RRID: SCR_002798) and R software (version 3.6.3; http://www.Rproject.org, RRID: SCR_001905). A two-sided probability value of $p < 0.05$ was considered being statistically significant.

RESULTS

Single-Cell RNA Sequencing Analysis of Two Mice Uteri

The scRNA-seq data of two 6-to-10-week-old female C57BL/6J mice were processed using published Seurat pipelines (Satija et al., 2015; Han et al., 2018). In total, we analyzed 3,756 single cells and identified 16 cell clusters, which were grouped into eight major cell types (**Figures 1A,B**, **Supplementary Table S1**). Then we clarified the identities of each cluster (**Figure 1C**, **Supplementary Table S2**). Cluster 0, cluster 1, cluster 2, cluster 3, cluster 10 all highly expressed *Col3a1* and *Fn1*, so we clarified them into stromal cells. Cluster 4, in the meantime, specifically expressed *Acta2*, which made us clarify them into myofibroblasts. Cluster 11, however, expressed *Acta2* and *Actg2* without *Col3a1*, so cells in cluster 11 belonged to muscle cells. Cluster 5, cluster 9, and cluster 12 all highly expressed *Cd68* and *Adgre1*, so they were macrophages/monocytes. *Cd7* and *Nkg7* could be found highly expressed in cluster 8, which made cluster 8 natural killer (NK) cells. *Ly6d* and *Cd79a* were highly expressed in cluster 15, so we identified cluster 15 as B cells. Cluster 13 and cluster 7 both had high expression levels of *Krt8*, *Krt18,* and *Epcam,* making them epithelial cells. Cluster 6 and cluster 14 had high expression levels of *Cldn5* and *Pecam*, making them endothelial cells. In this way, we clarified the identities of 16 cell clusters, building a solid foundation for further analysis.

We identified 5 sub-clusters of stromal cells (cluster 0, cluster 1, cluster 2, cluster 3, cluster 10) and 2 sub-clusters of epithelial cells (cluster 13 and cluster 7). According to their distances in the t-SNE, we divided the 5 sub-clusters of stromal cells into three groups: stromal cells 1, stromal cells 2, and stromal cells 3. The proportions of each cell type in two mice were shown (**Figure 1D**). Surprisingly, we found that stromal cells 1 were exclusively in mouse 2, and stromal cells 3 were exclusively in mouse 1. Stromal cells 2 could be found both in mouse 1 and mouse 2. This phenomenon was also seen in epithelial cells. Epithelial cells 1 were exclusively in mouse 2, while epithelial cells 2 were exclusively in mouse 1.

Comparison of Epithelial Cells and Stromal Cells Revealed Two Uteri of Distinct Estrus Stages

We believed the different cell contents of two mice had further biological significance. We found that proliferating cell nuclear antigen (*Pcna*) was highly expressed in epithelial cells 1 (mouse 2) (**Figure 2A**). The protein encoded by this gene is a cofactor of DNA polymerase delta, which is regarded as the marker of proliferation (Celis and Celis, 1985). Previous studies have reported that for humans experiencing estrous cycles, the divergence of the expression level of *Pcna* was mainly due to the states of the estrous cycle rather than age. It has been reported that the expression of *Pcna* showed a high peak in the proliferative phase, and then decreased sharply in the secretory phase (Li et al., 1993; Noci et al., 1995; Hamid et al., 2002). In mice, the treatment with estrogen to ovariectomized mice upregulates PCNA expression, while the co-treatment with estrogen and progesterone downregulates PCNA expression (Annie et al., 2019). In addition, we found that the expression levels of metalloproteinase (MMPs) were significantly suppressed in epithelial cells 1 (mouse 2) (**Supplementary Table S2**). So we preliminarily speculated that the uterus of mouse 2 was in the regenerative stage, and the uterus of mouse1 was in the maturational stage.

To further investigate this speculation, functional enriched terms of the differentially expressed genes in each epithelial cell

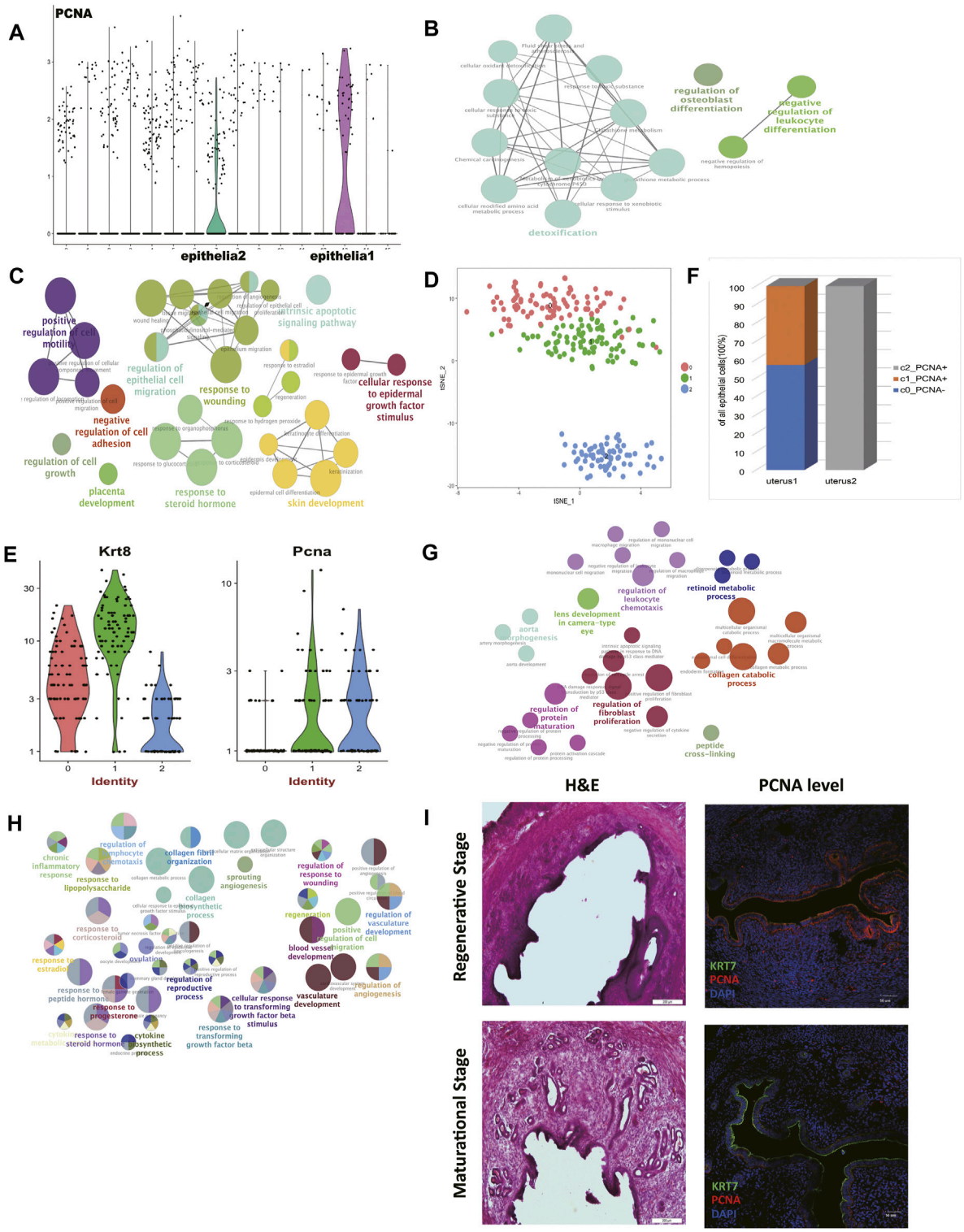

FIGURE 2 | Comparison of epithelial cells and stromal cells revealed two uteri of distinct estrus stages. **(A)** *Pcna* was highly expressed in epithelial cells 1 compared to epithelial cells 2. Gene ontology (GO) analysis of differentially upregulated genes in **(B)** epithelial cells 1 and **(C)** epithelial cells 2. **(D)** t-distributed stochastic neighbor embedding (t-SNE) diagram of epithelial cells in two mice uteri. **(E)** *Pcna* was highly expressed in sub-cluster 1 and sub-cluster 2. **(F)** The epithelial cell composition of the two uteri was significantly different. Gene ontology (GO) analysis of differentially upregulated genes of stromal cells in **(G)** the regenerative stage and **(H)** the maturational stage. **(I)** Representative H&E and IF staining of PCNA expression levels in mice endometrium. (DAPI, blue; PCNA, red; KRT7, green). Scale bars, 200 μm (H&E) and 50 μm (IF). GO: Gene ontology; t-SNE: t-distributed stochastic neighbor embedding; PCNA: proliferating cell nuclear antigen; IF: immunofluorescence; H&E: hematoxylin and eosin.

cluster were shown (**Figures 2B,C**, **Supplementary Table S3**). Terms including "response to steroid hormone," "skin development," "cellular response to epidermal growth factor stimulus," "placenta development," and "response to wounding" were enriched in epithelial cells 2 (mouse 1). These terms showed the active rebuilt of the uterus epithelium to prepare for subsequent fertilization in the maturational status. In this period, we also found terms that indicated "intrinsic apoptotic signaling pathway," "increased cell mobility," and "suppressed cell adhesion," which were all consistent with the changes that happened in the maturational status as previously reported (Yip et al., 2013).

The epithelial cells from two uteri were then clustered exclusively to further verify our hypothesis. We gained 3 clusters (**Figure 2D**, **Supplementary Table S4**). According to their *Pcna* expression levels, cluster 1 and cluster 2 showed proliferating characteristics (**Figure 2E**). We concluded that mouse 2 was in the regenerative stage, for it was entirely composed of proliferating epithelial cells. As for mouse 1, only part of the epithelial cells showed proliferating characteristics (**Figure 2F**). In addition, the expression level of *Esr1* was suppressed in the epithelial cells of cluster 0 in the mouse 1 (**Supplementary Table S4**).

Endometrial stromal cells (ESCs) perform a multitude of functions and undertake different functions in different estrous stages (Cottrell et al., 2017). We further explored the characteristics of stromal cells of two mice to further confirm the stages of two mice. The expression levels of MMPs were significantly upregulated in epithelial cells 2 (mouse 1) (**Supplementary Table S2**), which possibly linked with the active matrix degradation in the maturational stage.

We performed functional enrichment analyses of the differentially expressed genes in each stromal cell cluster to see their functional traits. The term "artery morphogenesis" was enriched in stromal cells in mouse 2 (**Figure 2G**, **Supplementary Table S3**). One of the key features of the regeneration stage is angiogenesis. Besides, "protein maturation" and "collagen catabolic process" were activated in this period.

The stromal cells in mouse 1 presented more dynamic cell communications (**Figure 2H**, **Supplementary Table S3**). The term "response to progesterone and other steroid hormones" further indicated that mouse 1 was in the maturational stage. The term "response to transforming growth factor beta" was enriched too. Also in the maturational stage, the accelerated cell movement was again seen in this stage, as enrichment analysis suggested the term "positive regulation of cell migration" was significantly enriched. We believed the elevated levels of cell communications and material transportation found in mouse 1 constituted a prosperous metabolism network in stromal cells.

In order to validate the data analysis results, vagina smear and hematoxylin and eosin (H&E) staining were used to determine the mouse estrous cycle (**Figure 2I**, **Supplementary Figure S1**) (Bertolin and Murphy, 2014). The expression levels of PCNA in the regenerative mice regenerative endometrium and maturational endometrium were presented (**Figure 2I**). We found that, like humans, expression level of PCNA in the regenerative endometrium was significantly higher than in the maturational endometrium, indicating the practicability of using PCNA expression level to define the estrus stage.

Taking into account the sheer differences in proliferation and apoptosis, cell adhesion and movement, angiogenesis, and extracellular matrix remodeling, we concluded that mouse 1 was in the maturational stage and mouse 2 was in the regeneration stage.

The Molecular Trajectory of Stromal Cells in the Two Estrus Stages

ESCs have been reported to perform a multitude of functions including hormonal regulation, decidualization, maternal-fetal communications, and embryo receptivity (Cottrell et al., 2017). After we clarified the estrous stages of two mice, we further explored the characteristics of stromal cells in each estrus period given their significances in the uterus.

Firstly, functional enrichment analysis of upregulated genes of myofibroblasts was performed (**Figure 3A**, **Supplementary Table S3**). Terms like "muscle structure development" and "regulation of smooth muscle cell proliferation" confirmed our previous classification of its identity.

The analyses of two stromal subsets in the last section showed that stromal cells in different physiological stages had huge differences in their functions. Therefore, we wanted to further explore how these two cell groups gained their unique traits by differentially gene expression. The pseudotime trajectory of stromal cells in two physiological stages was constructed (**Figure 3B**). The results showed that there were two differentiation branches: the regenerative stromal cells would transform into myofibroblasts or maturational stromal cells at the first decisional point, and a small part of the maturational cells still had the ability to differentiate into myofibroblasts at the second decisional point. In this way, an estrus cycle was completed. All the significantly expressed branch 1-dependent genes were clustered into three categories by unsupervised clustering (**Figure 3C**). Among these genes, the myofibroblasts-branch genes had high expression levels of *Acta2* and *Adamts1* (**Figure 3D**). In the meantime, the maturational_stromal-branch genes had high expression levels of *Col6a4*, *Fbln2*, *Tcf4*, and *Wnt5a* (**Figure 3E**). These divergent expression patterns of representative genes further confirmed our previous classification. We listed other significant branch-dependent genes, which may also play key roles in the stromal cell differentiation and need further validation (**Table 1**).

The Immune Landscape of the Uterus During the Two Estrus Stages

Embryo implantation and tumor progression are similar to some extent (Holtan et al., 2009). The maternal immune system needs to find the balance and provides an appropriate environment for the fetus to grow. The current findings of NK, monocytes, and dendritic cells (DCs) during the menstrual cycle are limited. For example, findings concerning NK cell number and cytotoxic activity are

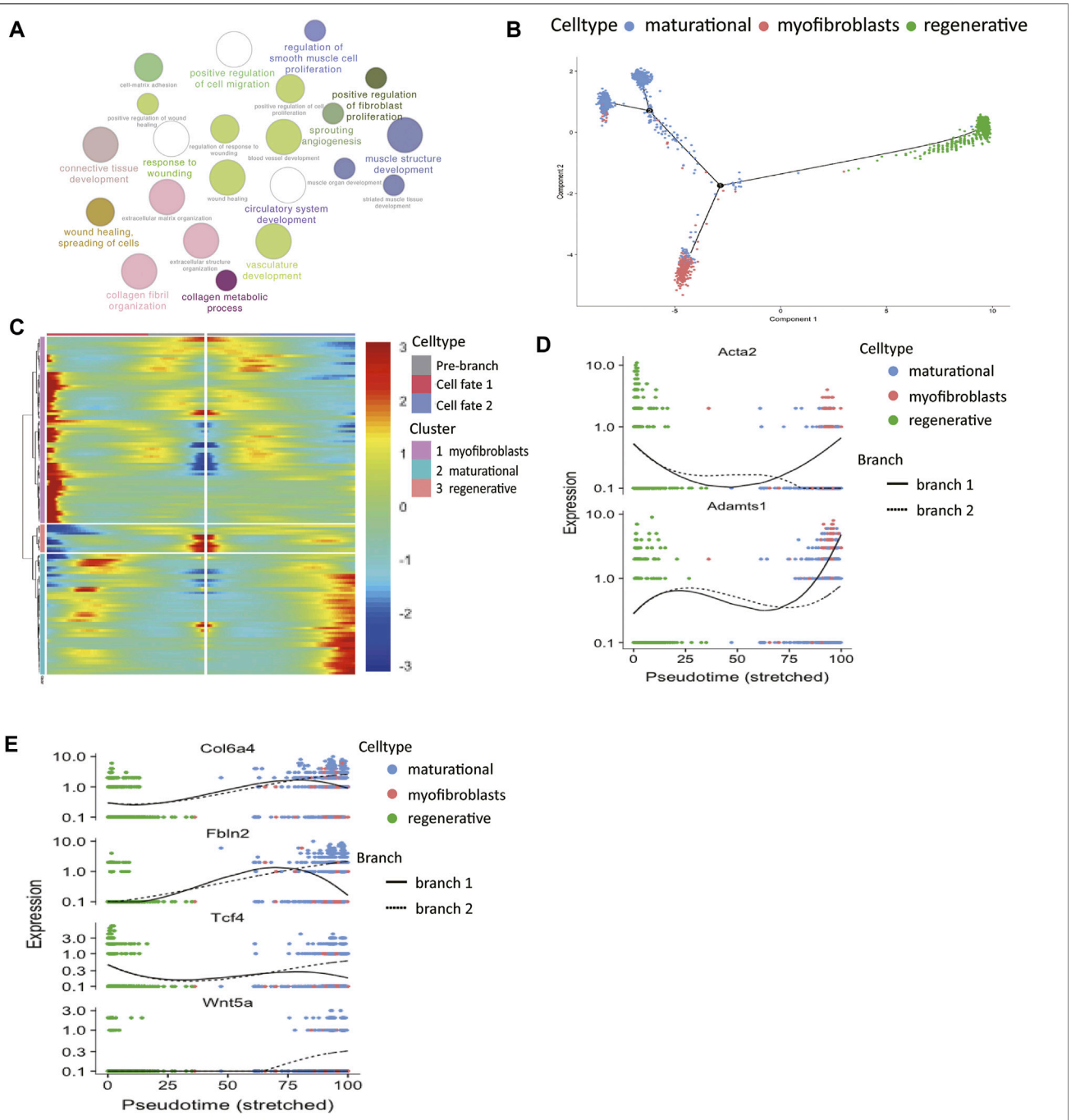

FIGURE 3 | The molecular trajectory of stromal cells in the two estrus stages. **(A)** Gene ontology analysis of differentially upregulated genes in myofibroblasts. **(B)** Monocle generated the pseudotemporal trajectory of stromal cells in different physiological stages. **(C)** Heat map for clustering the significantly branch-1 dependent genes that affected cell fate decisions into three clusters. q < 1E-06. The expression levels of representative genes of **(D)** myofibroblasts and **(E)** maturational stromal cells were shown in the line plots.

conflicting (Oertelt-Prigione, 2012). In order to gain reliable results with regard to uterine immune cells at the single-cell level, all immune cells were clustered exclusively into 4 clusters (**Figure 4A**, **Supplementary Table S5**). The heatmap of the top 50 markers for each cluster showed that we gained a convincing clustering result (**Figure 4B**).

Using known markers, the identities of each cluster were clarified (**Figure 4C**). Cluster 0 expressed *Adgre1* and *Mrc1*, so we clarified them into macrophages. Cluster 1 expressed *Nkg7* and *Cd7*, which made us clarify them into NK cells. Cluster 2 expressed *Cd83* and *Cd209a*, so cells in cluster 2 belonged to DC cells. For their high expression levels of *Ly6c2*, we clarified cluster

TABLE 1 | Significant branch-dependent genes of mice stromal cells in different physiological stages.

Regenerative stromal cells	Maturational stromal cells	Myofibroblast
mt-Cytb	Aldh1a2	Adh1
Igfbp7	Hsd11b2	Col3a1
mt-Nd4	Ccl11	Mgp
Igfbp4	Ramp3	Lum
Hba-a1	P2ry14	Cxcl1
Mfap4	Fbln2	Col1a2
mt-Nd2	Crispld2	Fgl2
Id3	Htra1	Has1
Rpl18a	Ramp2	Sparcl1
Itm2b	Jun	Adamts1
Hbb-bs	Atp8a1	Ccl2
	A2m	Mt1
	Dio2	Hk2
	Cd164	Sparc
	Srgn	Acta2
	Sdc1	Il6
	Emb	Col14a1
	Aqp1	Clec3b
	Col6a4	Vcam1
	Mfap5	Pcolce
	Mettl7a1	Ifi205
	Sptssa	Efemp1
	Tcf4	Ly6c1
	Smoc2	Rbp1
	Laptm4a	Gpx3
	Wnt5a	Sqstm1
	Rhob	Dpt
	Hspa1a	Gsn
	Fos	Cxcl16
	Scube1	

3 into monocytes. We found a clear difference in immune cell composition in these two estrus stages. The proportions of macrophages and NK cells in the two stages varied considerably (**Figure 4D**). Macrophages dominated in the regenerative stage, while NK cells dominated in the maturational stage.

The mechanism of monocytes differentiating into DCs or macrophages is poorly understood. The developmental trajectory of the mononuclear phagocyte system was constructed (**Figure 4E**). The branch point led to two differentiation paths: macrophages or DCs. Branch-dependent genes were clustered into three categories according to their expression patterns (**Figure 4F**). The expression levels of marker genes of each branch over pseudotime verified the identities of two differentiation paths (**Figure 4G**). We listed other significant branch-dependent genes (**Table 2**). Among them, we detected three transcription factors: *Mafb*, *Irf7*, and *Nr4a1*. *Mafb* was detected in the macrophage branch, *Irf7* expression was seen in the monocyte branch, and *Nr4a1* expressed highly in the DC branch (**Figure 4G**).

As for NK cells (**Table 3**), NK cells in the regenerative stage showed intense inflammatory responses with high expression levels of *Tnf*, *Il1b*, and *S100a8*. It also expressed *Cxcl2* and *Ccrl2*, which respectively promoted the recruitment of neutrophils and themselves (Regan-Komito et al., 2017). In contrast, upregulated genes of NK cells in the maturational stage mostly were immune-suppressive genes like *Serpinb9*, *Stat3*, *Cd96*, and *Cd55*. They also secreted substances that were advantageous for follow-up pregnancy like *Ccl2*.

The number of macrophages was higher in the regenerative stage. In this stage, macrophages were pro-inflammatory by secreting resistin, which was a systemic pro-inflammatory cytokine targeting both leukocytes and adipocytes (**Table 3**) (Nagaev et al., 2006). However, it also showed high expression levels of selenoprotein *Msrb1* and *Hes1* to inhibit inflammatory responses (Lee et al., 2017; Zhang et al., 2019). As for macrophages in the maturational stage, it mainly showed the M2 phenotype with high expression levels of *Il10*, *Cd206*, and *Ccl7*, which were consisted of the characteristics of decidual macrophages (Xuan et al., 2015).

In the regenerative stage, the functions of DC were mainly embodied in influencing other cells. It secreted *Ccl22* to recruit regulatory T (Treg) cells and *Ccl5* to recruit *CCR5*-positive cytokine-induced killer cells (Lee et al., 2016; Yashiro et al., 2019). Its pro-inflammatory role was also embodied in the high expression of *Ccrl2*, which played a significant role in inflammation (Regan-Komito et al., 2017). DCs in the maturational stage were immune-suppressive (**Table 3**). DCs expressed *PD-L1*, *Jchain*, and *Anxa1*, which were correlated with immune tolerance and refrained the secretion of pro-inflammatory cytokines (Källberg and Leanderson, 2008; Sena et al., 2016).

Monocytes in the maturational stage secreted multiple chemokines like *Thbs1*, *Ccl2*, *Ccl7*, *Ccl8*, *Lyn*, and *Anxa1* to facilitate the migration of themselves (**Table 3**) (Asano et al., 2015; Liu et al., 2015). Monocytes also expressed M1 markers like *Cd86* and *Lgals8* to enhance inflammatory responses (Patel et al., 2017). By contrast, monocytes in the regenerative stage showed more intense cytolytic ability by producing *S100a8*, *Tnf*, and *Il1b*. It also secreted *Cxcl16* and *Ccl3* to facilitate the recruitment of monocytes and *Mmp14*, *Trem2*, and *Mif* to induce inflammatory responses (Buck et al., 2013; Ruiz-Rosado et al., 2016).

Taken together, we found that the immune responses were suppressed in the maturational stage compared to the regenerative stage by reducing inflammatory cytokines secretion and facilitating differentiation into immune-suppressive cells.

Identification of the Molecular Feature of Endometrial Mesenchymal Stem Cells *In Vivo*

The mesenchymal-to-epithelial transition has been proposed as a possible reason to explain the periodically endometrial epithelial tissue regeneration (Huang et al., 2012). A previous study has reported a group of cells expressed both the epithelial cell marker, pan-cytokeratin, and the stromal cell marker, vimentin as well (Patterson et al., 2013). We tried to identify multipotent ESCs in the scRNA-seq data using published pipelines (Grün et al., 2016). We failed to find the cell groups that expressed all the proposed stem cell surface molecules like *Cd73*, *Cd90*, *and Cd105*. We hypothesized that expressions of *Cd73*, *Cd90*, and *Cd105* may be induced during *in-vitro* culture. Another scRNA-seq dataset containing cultured ESCs and uncultured ones was used to show the diversity of two culture

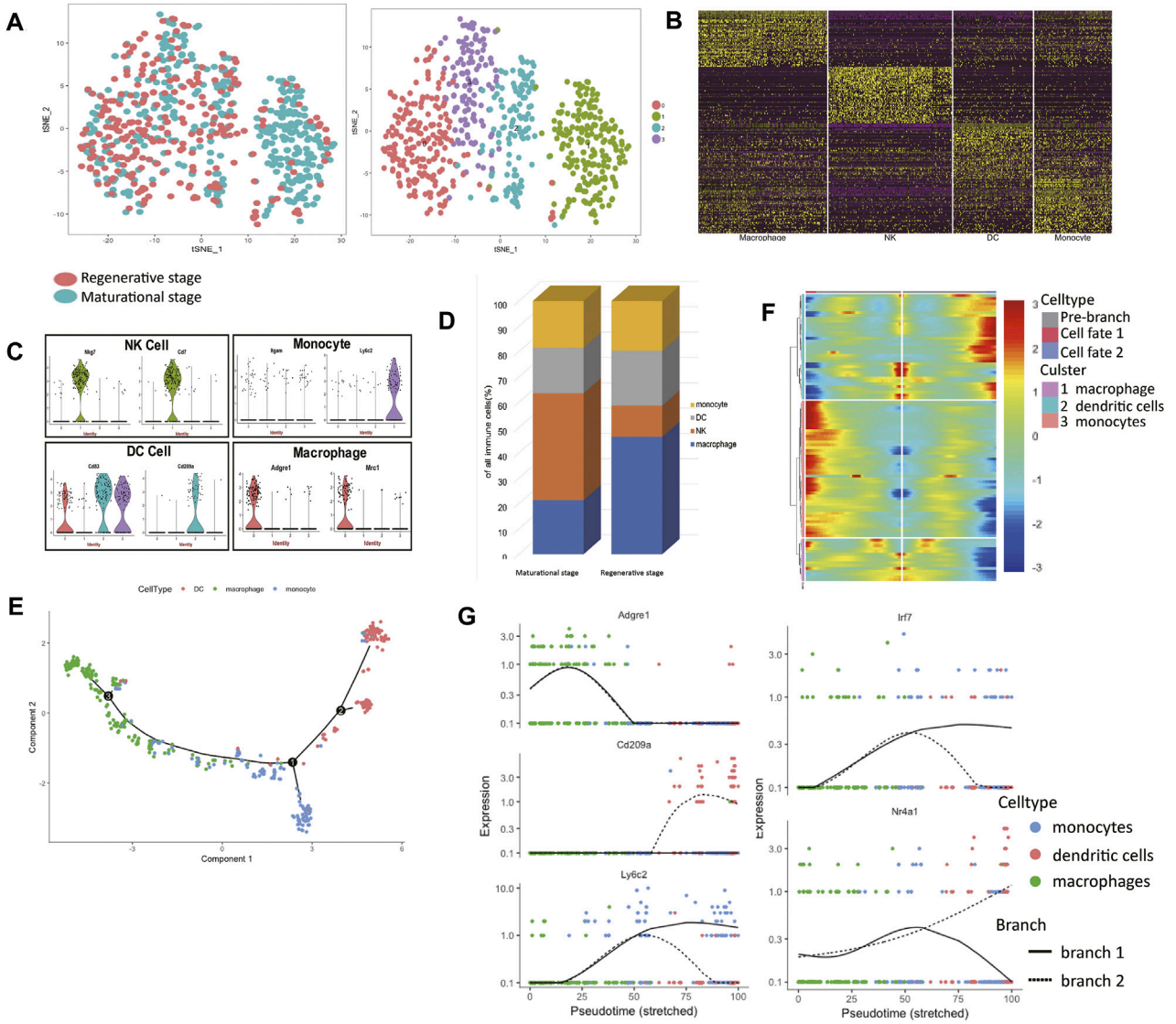

FIGURE 4 | The immune landscape of the uterus during the two estrus stages. **(A)** t-distributed stochastic neighbor embedding (t-SNE) diagram of immune cells in two mice uteri. **(B)** Heatmap of immune cells from the two uteri revealed four populations. **(C)** Violin plots indicating the expression of marker genes of each cell cluster. **(D)** The immune cell composition of the two uteri was different, macrophages dominated in the regenerative stage, while natural killer (NK) cells dominated in the maturational stage. **(E)** Monocle generated the pseudotemporal trajectory of the phagocyte system. **(F)** Heat map for clustering the significantly branch-1 dependent genes that affected cell fate decisions into three clusters. q < 1E-06. **(G)** The expression levels of representative genes of macrophages/dendritic cells (DCs)/ monocytes and three transcription factors were shown in the line plots. t-SNE: t-distributed stochastic neighbor embedding; NK: natural killer; DC: dendritic cells.

environments (Krjutškov et al., 2016). The cells from two different culture environments clustered separately (**Figure 5A**). The stem cell surface markers mentioned above were expressed differently in the two cell groups. We found that stem cell makers like *Cd90* (*Thy1*) and *Cd44* had significant higher expressions in the cultured group (**Figure 5B**). We found that cell group had a higher entropy score and more connections with other cell types (**Figures 5C,D,F**, **Supplementary Table S6**), had the potential to differentiate into stromal cells, epithelial cells, and immune cells (**Figure 5E**), which make it the potential multipotent ESCs *in vivo*. This cell subcluster expressed higher expression of *Cd34*, *Pdgfrb*, and *Aldh1a2* (**Figures 5F–I**). The representative immunofluorescence staining of CD34 in mice endometrium was shown, indicating the existence of this endometrial mesenchymal stem cells (**Figure 5J**).

To conclude, we found the transcriptional signature of endometrial mesenchymal stem cells *in vivo*. And the cell group 5 that we presented here showed a great possibility to be the cells responsible for re-epithelialization.

DISCUSSION

There is still a huge void in uterus-related research at the single-cell level. In humans, the sequence data of $Cd13^{+}$ stromal and $Cd9^{+}$

TABLE 2 | Significant branch-dependent genes of the mononuclear phagocyte system in two mice uteri.

Monocytes	Macrophages	Dendritic cells
Ninj1	Snx2	Cd151
Pilra	C5ar1	Rpl18
Rnh1	Cd83-ps	H2-Eb1
Gm6977	Lgmn	H2-Aa
Card19	Ccl2	Cd74
Ccr1	Csf1r	H2-Ab1
Clec4e	Lst1	Retnlg
Chil3	Mafb	Cxcl16
Thbs1	Lyz2	Rps27
Cd44	Sat1	Rpl18a
Ly6c2	Fcgr3	Cst3
Cd300lf	Ctsb	Rps11
Spp1	C1qb	Rpl32
Mmp19	Apoe	Rps9
Ifitm3	Ctsd	Rps7
Ms4a4c		Rgs1
Lgals3		Nr4a1
Cdkn1a		Bcl2a1d
H3f3b		Cd209a
F10		Tnlp3
Psap		Ccr7
Pkm		Il1b
Lilrb4a		Lsp1
Mdm2		Napsa
Cd14		Mt-Rnr2
Emilin2		Mt-Rnr1
Retnla		Hba-a2
Msr1		Hba-a1
Ms4a6d		Hbb-bs
Clec4d		Ccl4
Ccl9		Hspa1a
Lilr4b		Bcl2a1b
Igkc		Rps21
Hmox1		Rps26
Plin2		Tmsb4x
Fcgr1		Mgp
Cxcl2		Rps29
Ftl1		
Ifi27l2a		
*F*th1		
Irf7		
Ms4a6b		
Cstb		
Fabp5		
Msrb1		
Osm		
Slc11a1		

TABLE 3 | Significantly differentially expressed genes of immune cells in different physiological stages.

	Upregulated in the regenerative stage	Upregulated in the maturational stage
NK	Tnf	Serpinb9
	Il1b	Stat3
	S100a8	Cd96
	Cxcl2	Cd55
	Ccrl2	Rps6
	Cd74	Gzmb
	H2-Ab1	Lck
	H2-Eb1	Igha
	H2-Aa	Ptpn22
	Ifitm3	Stat3
	Ccl2	Tnfaip3
	Klrg1	Trbc1
	Ly6c1	Ptprc
	Rarres2	Ccl7
	Tgfb1	Tnfrsf1b
Macrophage	Resistin	Il10
	Msrb1	Cd206
	Hes1	Ccl7
	Il1b	Ccl2
	Il18bp	Mrc1
	Mpeg1	Rps6
	S100a8	Cd36
	Msrb1	
	Sox4	
	Retnlg	
	Ccl11	
DC	Ccl22	PD-L1
	Ccl5	Jchain
	Ccrl2	Anxa1
	Itga4	Tnfrsf1b
	Mmp14	Igha
	Irf8	Ccr7
	C1qa	Igkc
	Ccl3	
	Ier3	
Monocyte	S100a8	Thbs1
	Tnf	Ccl2
	Il1b	Ccl7
	Cxcl16	Ccl8
	Ccl3	Lyn
	Mmp14	Anxa1
	Trem2	Cd86
	Mif	Lgals8
	C1qc	Igkc
	H2-K1	Ccl7
	Ccl3	Cd86
	Il1rn	Thbs1
	Hes1	Spp1
	Ccrl2	Cd14
	C1qa	Mif
	Trib1	Jchain
	Cxcl16	

epithelial cells from endometrial tissues have been generated (Krjutškov et al., 2016). Researchers also used scRNA-seq to map the temporal transcriptomic changes in cultured primary ESCs along a decidual time-course and in response to the withdrawal of differentiation signals (Lucas et al., 2020). In mice, the transcriptional profiles of uterine epithelial cells at five developmental stages, ranging from neonatal to mature stages were analyzed (Wu et al., 2017). These studies contained a limited number of cell types. A scMCA covering major cell types was completed (Han et al., 2018). It included the uterus and mice other major organs as well, yet without an in-depth analysis of the changes that happened in the endometrium at different estrus stages and left opportunities for later researchers to delineate the dynamic endometrial cell transformation of all cell types in the estrus cycle at the single-cell level. Recently, mesenchymal cells in the adult mouse endometrium have been characterized and five subpopulations have been identified (Kirkwood et al., 2021). Our research is a systematical single-cell level study with a special focus on

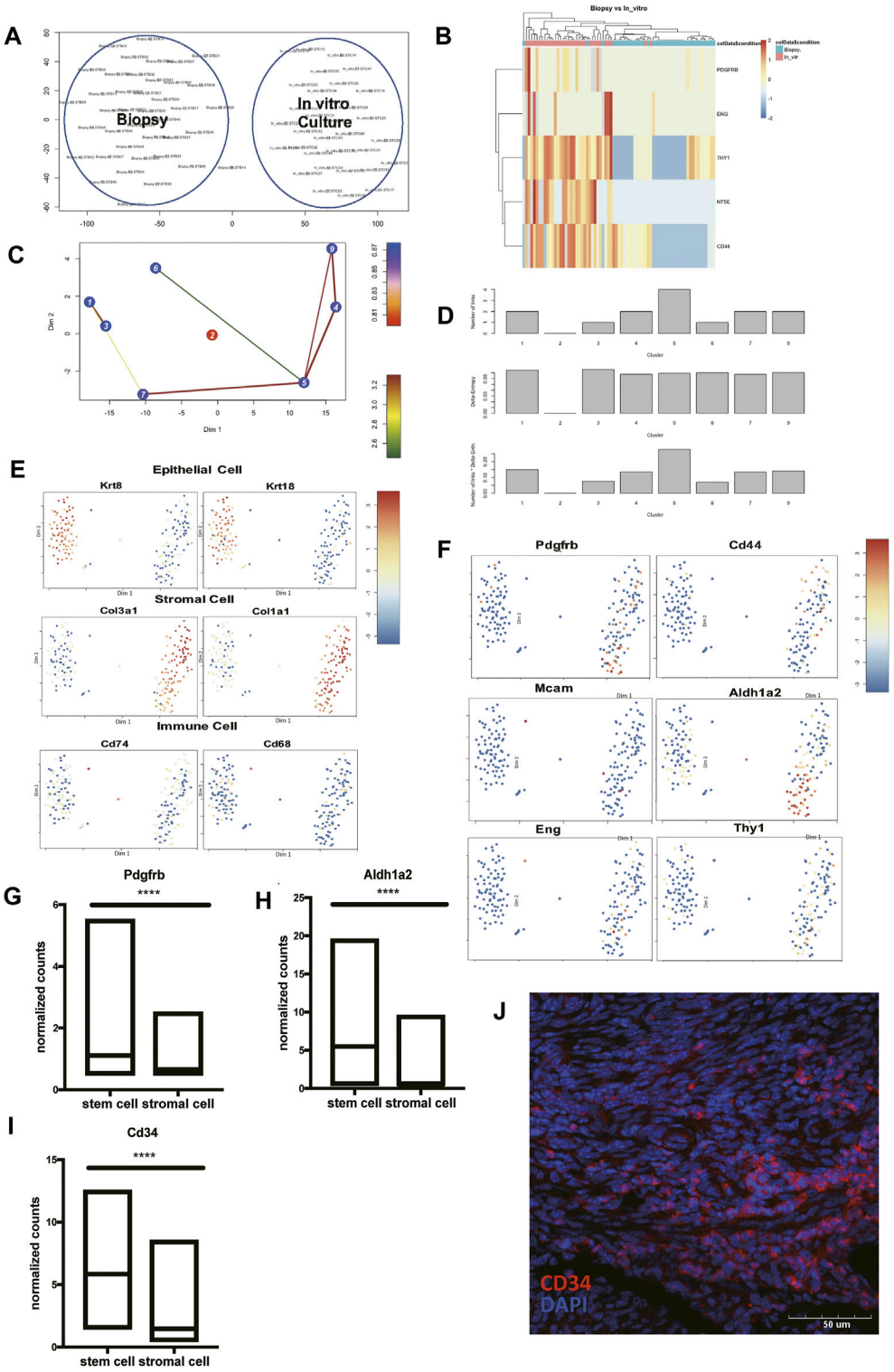

FIGURE 5 | Identify *Pdgfrb*$^{+}$ *Aldh1a2*$^{+}$ *Cd34*$^{+}$ endometrial mesenchymal stem cells *in vivo*. **(A)** Cells directly harvested from biopsy clustered separately with cells after *in-vitro* culture. **(B)** The heat map presented the differentially expressed *Cd73* (*Nt5e*), *Cd90* (*Thy1*), *Cd44*, *Pdgfrb*, and *Eng* in cells directly harvested from biopsy and after *in-vitro* culture. **(C)** Lineage tree generated by StemID2. **(D)** Histograms for the number of links, the delta-entropy, and the StemID2 score generated by StemID2. A high score indicates a higher likelihood that the cluster is an actual stem cell cluster. The t-SNE map with color code representation of log-transformed expression across marker genes of **(E)** epithelial cells, stromal cells, immune cells, and **(F)** stem cells. **(G–I)** The expression level of Pdgfrb, Aldh1a2, and Cd34 in cluster 5 and the rest of the uterine stromal cells. **(J)** Representative IF staining of CD34 expression levels in mice endometrium. (DAPI, blue; CD34, red). Scale bars, 50 μm. IF: immunofluorescence; *p*-value<0.0001.

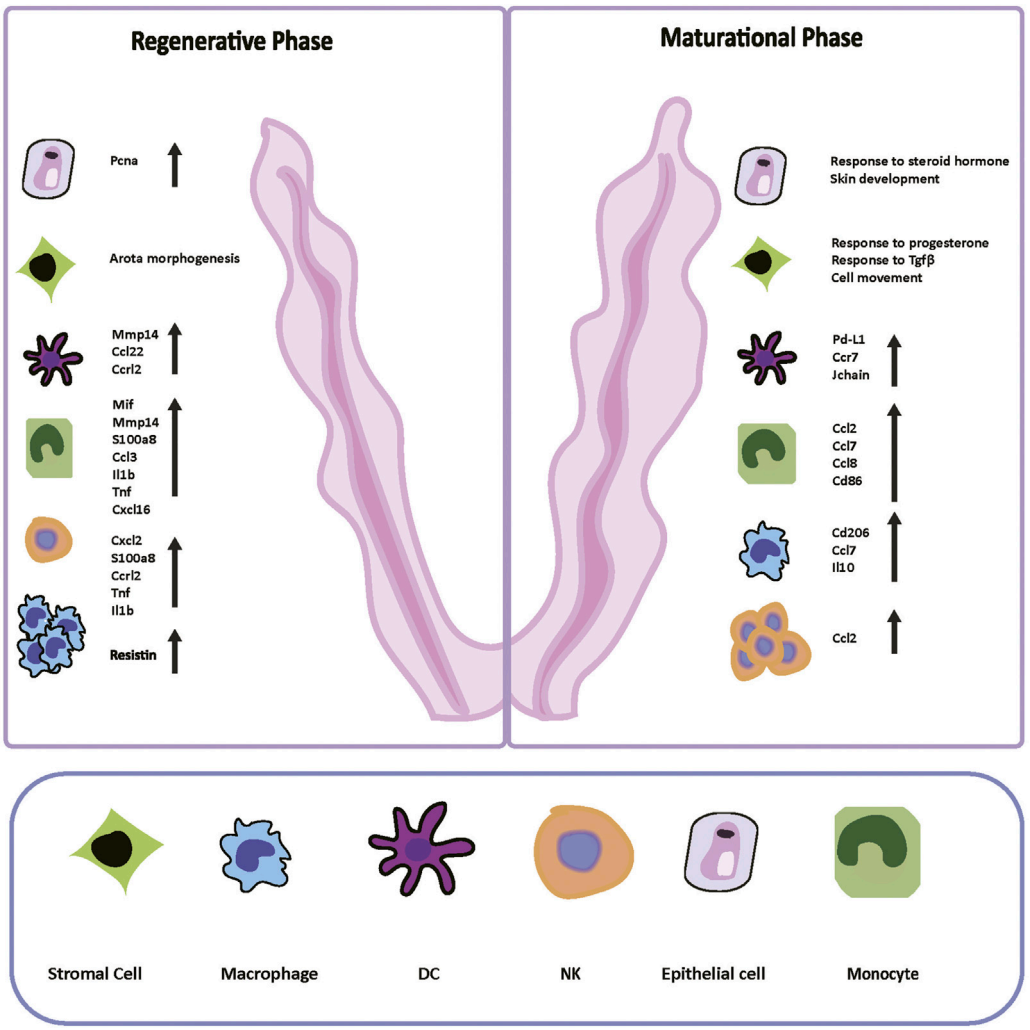

FIGURE 6 | Digital Cell Atlas of Mouse Uterus: from regenerative stage to maturational stage. Based on the *Pcna* expression, differentially expressed genes, and gene ontology terms of two stages, we built up the cell atlas of the mouse uterus from the regenerative to maturational endometrium. The number and function of immune cells varied sharply in two periods. In the regenerative stage, mice epithelial cells expressed Pcna highly, and the term "aorta morphogenesis" was enriched in stromal cells. Macrophages dominated in this stage. The immune system was boosted with intensely secreted cytokines and chemokines. In the maturational stage, cells presented more dynamic cell communication instead of proliferating. Terms "response to progesterone and other steroid hormones" and "response to transforming growth factor beta" were enriched. NK cells dominated in this stage. The immune responses were suppressed compared to the regenerative stage by reducing inflammatory cytokines secretion and facilitating differentiation into immune-suppressive cells. *Pcna*: proliferating cell nuclear antigen; NK: natural killer; DC: dendritic cells.

constructing a mice cell atlas at different estrus stages and showed that different estrus stages had sheer different cell composition.

We identified three transcription factors in the differentiation path of the mononuclear phagocyte system. *Mafb* was reported to be essential for monocyte-macrophage differentiation in the previous study (Goudot et al., 2017). *C1qb*, as the complement component of C1q complex, was one of the *Mafb* target genes and conformably upregulated in the macrophage branch (**Table 2**) (Hamada et al., 2020).

Irf7 has been reported to be involved in the regulatory pathway initiated in DCs during their response to microbial stimuli but dispensable in DC development (Owens et al., 2012). *Nr4a1* has been reported to be the target for modulating the inflammatory phenotypes of monocytes and macrophages (Hanna et al., 2012). In our analyses, *Irf7* expression was seen in the monocyte branch, and *Nr4a1* expressed highly in the DC branch, so the relations between these two transcription factors and their branches needed further experiments. Likewise, understanding the decisional mechanism of fibroblast-to-myofibroblast differentiation may be the key to tackle endometriosis and adenomyosis. Among the highly expressed genes that determined the myofibroblast differentiation (**Table 1**), there were a lot of them have not been reported, which shed light on future research.

Generally, the immune cells in the uterus endometrium showed a pro-inflammatory tendency in the regenerative stage and an anti-inflammatory tendency in the maturational stage.

Yet the macrophages in the regenerative stage also expressed certain immune-suppressive genes, indicating the possibility that tissue-resident macrophages existed and contributed to endometrium repair. The findings concerning uterus tissue-resident macrophages are rare. Putative tissue-resident macrophages have been reported to be spatially restricted and in association with areas of repaired, re-epithelialized endometrium (Cousins et al., 2016). In our pseudotime analysis (**Figure 4E**), we found a branch in the path of macrophage differentiation, which indicated the existence of a subtype of macrophages and might be the uterus tissue-resident macrophages. The molecular profiles of mouse uterus tissue-resident macrophages remained unknown, which needed further research. Our results may provide insights into mechanisms regarding tissue remodeling and aid in tackling endometrium abnormalities like endometritis.

We reported a novel group of markers for *in vivo* endometrial mesenchymal stem cells: *Pdgfrb*, *Aldh1a2*, and *Cd34*, which has not been reported before. $Cd34^{+}Klf4^{+}$ stromal-resident stem cells have been reported to directly contribute to endometrial regeneration (Yin et al., 2019). $PDGFR\alpha^{+}/CD34^{+}$ Celll group 5 highly expressed *Cd34*, but without detectable expression of *Klf4*. Many proposed stem cell surface molecules like *Cd73*, *Cd90*, and *Cd105* failed to express highly in *in vivo* endometrial mesenchymal stem cells (Kyurkchiev et al., 2010).

However, several limitations of our study should be acknowledged. Firstly, this is an RNA-seq based bioinformatics study and caution must be taken when further extrapolating these results *in vivo*. Future study is needed to evaluate the presence of a transcript corresponds to its expression at the protein level. Experiments such as further functional verification of reported endometrial mesenchymal stem cells are crucial in the future. Secondly, the limited number of mice that participated in this study may affect the statistical power and the strength of our conclusions. Thirdly, one of our goals in this study is to seek insights from mice scRNA-seq data and shed light on human endometrium research. However, species-specificities in uterine physiology exist, which may dampen the practicability of our study.

We thoroughly compared the functional traits and molecular profiles of each cell type from the regenerative stage to the maturational stage (**Figure 6**). With a template for the interactions of different cell types at the normal condition, researchers can better identify morphologically unaffected abnormalities in future studies. We also depicted the transitions and transcription factors that shaped the differentiation of ESCs and the mononuclear phagocyte system, and many of them have not been reported before. The genes that have not been reported before can inspire subsequent basic research. The cell atlas of mice uteri presented here would improve our understanding of the functional changes that occurred in the endometrium during the estrus cycle.

CONCLUSION

This study is a systematical single-cell level study to construct a mice cell atlas from the regenerative stage to the maturational stage, including epithelial cells, stromal cells, and immune cells. The functional traits and molecular profiles for each cell type in these two stages are thoroughly compared. The mice cell atlas also delineates the transitions that shape the differentiation of endometrial stromal cells and the mononuclear phagocyte system. This study found putative uterus tissue-resident macrophages and $Pdgfrb^{+}$ $Aldh1a2^{+}$ $Cd34^{+}$ endometrial mesenchymal stem cells *in vivo*.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The animal study was reviewed and approved by the National Cancer Institute Animal Care and Use Committee (ACUC).

AUTHOR CONTRIBUTIONS

LZ and BW initiated and organized the study; LZ performed bioinformatics analyses, statistical analyses, drew figures, and drafted the manuscript; WL and WX performed wet experiments; XC and XZ confirmed the clinical relevance of the study; BW edited the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC) (81871127), Zhejiang Medical and Health Science and Technology Program (2013KYB080), and the Science and Technology program of Jinhua Science and Technology Bureau (Grant No. 2021-3-001).

ACKNOWLEDGMENTS

The datasets of this study are obtained from previous studies. We are grateful to them for the generously shared data and inspiration.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.847646/full#supplementary-material

Supplementary Figure S1 | (A) Scheme of collecting vaginal smears in 6 consecutive days before and at harvest day. Representative images of vaginal smears of **(B)** regenerative endometrium and **(C)** maturational endometrium. Scale bars, 200 μm.

REFERENCES

Annie, L., Gurusubramanian, G., and Roy, V. K. (2019). Estrogen and Progesterone Dependent Expression of Visfatin/NAMPT Regulates Proliferation and Apoptosis in Mice Uterus during Estrous Cycle. *J. Steroid Biochem. Mol. Biol.* 185, 225–236. doi:10.1016/j.jsbmb.2018.09.010

Asano, K., Takahashi, N., Ushiki, M., Monya, M., Aihara, F., Kuboki, E., et al. (2015). Intestinal CD169(+) Macrophages Initiate Mucosal Inflammation by Secreting CCL8 that Recruits Inflammatory Monocytes. *Nat. Commun.* 6 (undefined), 7802. doi:10.1038/ncomms8802

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Bertolin, K., and Murphy, B. D. (2014). "Monitoring Mouse Estrous Cycles," in *The Guide to Investigation of Mouse Pregnancy*. Editors B. A. Croy, A. T. Yamada, F. J. DeMayo, and S. L. Adamson (Boston: Academic Press), 475–477. doi:10.1016/b978-0-12-394445-0.00039-4

Carbon, S., Douglass, E., Dunn, B., Good, N., and Harris, N. L. (2019). The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Res.* 47 (D1), D330–d338. doi:10.1093/nar/gky1055

Celis, J. E., and Celis, A. (1985). Cell Cycle-dependent Variations in the Distribution of the Nuclear Protein Cyclin Proliferating Cell Nuclear Antigen in Cultured Cells: Subdivision of S Phase. *Proc. Natl. Acad. Sci. U.S.A.* 82 (10), 3262–3266. doi:10.1073/pnas.82.10.3262

Cole, T., Davide, C., Jonna, G., Prapti, P., Shuqiang, L., Michael, M., et al. (2014). The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells. *Nat. Biotechnol.* 32 (4), 381–386. doi:10.1038/nbt.2859

Cora, M. C., Kooistra, L., and Travlos, G. (2015). Vaginal Cytology of the Laboratory Rat and Mouse: Review and Criteria for the Staging of the Estrous Cycle Using Stained Vaginal Smears. *Toxicol. Pathol.* 43 (6), 776–793. doi:10.1177/0192623315570339

Cottrell, H. N., Wu, J., Rimawi, B. H., Duran, J. M., Spencer, J. B., Sidell, N., et al. (2017). Human Endometrial Stromal Cell Plasticity: Reversible sFlt1 Expression Negatively Coincides with Decidualization. *Hypertens. pregnancy* 36 (2), 204–211. doi:10.1080/10641955.2017.1299172

Cousins, F. L., Kirkwood, P. M., Saunders, P. T., and Gibson, D. A. (2016). Evidence for a Dynamic Role for Mononuclear Phagocytes during Endometrial Repair and Remodelling. *Sci. Rep.* 6 (undefined), 36748. doi:10.1038/srep36748

Buck, M. D., Gouwy, M., Proost, P., Struyf, S., and Van Damme, J. (2013). Identification and Characterization of MIP-1α/CCL3 Isoform 2 from Bovine Serum as a Potent Monocyte/dendritic Cell Chemoattractant. *Biochem. Pharmacol.* 85 (6), 789–797. doi:10.1016/j.bcp.2012.11.027

Goudot, C., Coillard, A., Villani, A.-C., Gueguen, P., Cros, A., Sarkizova, S., et al. (2017). Aryl Hydrocarbon Receptor Controls Monocyte Differentiation into Dendritic Cells versus Macrophages. *Immunity* 47 (3), 582–596. e586. doi:10.1016/j.immuni.2017.08.016

Grün, D., Muraro, M. J., Boisset, J. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., et al. (2016). De Novo Prediction of Stem Cell Identity Using Single-Cell Transcriptome Data. *Cell stem Cell* 19 (2), 266–277. doi:10.1016/j.stem.2016.05.010

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell Messenger RNA Sequencing Reveals Rare Intestinal Cell Types. *Nature* 525 (7568), 251–255. doi:10.1038/nature14966

Haber, A. L., Biton, M., Rogel, N., Herbst, R. H., Shekhar, K., Smillie, C., et al. (2017). A Single-Cell Survey of the Small Intestinal Epithelium. *Nature* 551 (7680), 333–339. doi:10.1038/nature24489

Hamada, M., Tsunakawa, Y., Jeon, H., Yadav, M. K., and Takahashi, S. (2020). Role of MafB in Macrophages. *Exp. Anim.* 69 (1), 1–10. doi:10.1538/expanim.19-0076

Hamid, A. A., Mandai, M., Konishi, I., Nanbu, K., Tsuruta, Y., Kusakari, T., et al. (2002). Cyclical Change of hMSH2 Protein Expression in Normal Endometrium during the Menstrual Cycle and its Overexpression in Endometrial Hyperplasia and Sporadic Endometrial Carcinoma. *Cancer* 94 (4), 997–1005. doi:10.1002/cncr.10341

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172 (5), 1091–1107. doi:10.1016/j.cell.2018.02.001

Hanna, R. N., Shaked, I., Hubbeling, H. G., Punt, J. A., Wu, R., Herrley, E., et al. (2012). NR4A1 (Nur77) Deletion Polarizes Macrophages toward an Inflammatory Phenotype and Increases Atherosclerosis. *Circ. Res.* 110 (3), 416–427. doi:10.1161/circresaha.111.253377

Holtan, S. G., Creedon, D. J., Haluska, P., and Markovic, S. N. (2009). Cancer and Pregnancy: Parallels in Growth, Invasion, and Immune Modulation and Implications for Cancer Therapeutic Agents. *Mayo Clin. Proc.* 84 (11), 985–1000. doi:10.4065/84.11.985

Huang, C.-C., Orvis, G. D., Wang, Y., and Behringer, R. R. (2012). Stromal-to-epithelial Transition during Postpartum Endometrial Regeneration. *PloS one* 7 (8), e44285. doi:10.1371/journal.pone.0044285

Jiang, X., Du, M.-R., Li, M., and Wang, H. (2018). Three Macrophage Subsets Are Identified in the Uterus during Early Human Pregnancy. *Cell Mol. Immunol.* 15 (12), 1027–1037. doi:10.1038/s41423-018-0008-0

Källberg, E., and Leanderson, T. (2008). A Subset of Dendritic Cells Express Joining Chain (J-Chain) Protein. *J. Immunol.* 123 (4), 590–9. doi:10.1111/j.1365-2567.2007.02733.x

Kirkwood, P. M., Gibson, D. A., Smith, J. R., Wilson-Kanamori, J. R., Kelepouri, O., Esnal-Zufiaurre, A., et al. (2021). Single-cell RNA Sequencing Redefines the Mesenchymal Cell Landscape of Mouse Endometrium. *FASEB J.* 35 (4), e21285. doi:10.1096/fj.202002123R

Krjutškov, K., Katayama, S., Saare, M., Vera-Rodriguez, M., Lubenets, D., Samuel, K., et al. (2016). Single-cell Transcriptome Analysis of Endometrial Tissue. *Hum. Reprod.* 31 (4), 844–853. doi:10.1093/humrep/dew008

Kumar, R., Clerc, A.-C., Gori, I., Russell, R., Pellegrini, C., Govender, L., et al. (2014). Lipoxin A4 Prevents the Progression of De Novo and Established Endometriosis in a Mouse Model by Attenuating Prostaglandin E2 Production and Estrogen Signaling. *PLoS One* 9 (2), e89742. doi:10.1371/journal.pone.0089742

Kumar, R., Vicari, M., Gori, I., Achtari, C., Fiche, M., Surbeck, I., et al. (2011). Compartmentalized Secretory Leukocyte Protease Inhibitor Expression and Hormone Responses along the Reproductive Tract of Postmenopausal Women. *J. Reproductive Immunol.* 92 (1-2), 88–96. doi:10.1016/j.jri.2011.06.103

Kyurkchiev, S., Shterev, A., and Dimitrov, R. (2010). Assessment of Presence and Characteristics of Multipotent Stromal Cells in Human Endometrium and Decidua. *Reprod. Biomed. online* 20 (3), 305–313. doi:10.1016/j.rbmo.2009.12.011

Lee, B. C., Lee, S.-G., Choo, M.-K., Kim, J. H., Lee, H. M., Kim, S., et al. (2017). Selenoprotein MsrB1 Promotes Anti-inflammatory Cytokine Gene Expression in Macrophages and Controls Immune Response *In Vivo*. *Sci. Rep.* 7 (1), 5119. doi:10.1038/s41598-017-05230-2

Lee, H. K., Kim, Y. G., Kim, J. S., Park, E. J., Kim, B., Park, K. H., et al. (2016). Cytokine-induced Killer Cells Interact with Tumor Lysate-Pulsed Dendritic Cells via CCR5 Signaling. *Cancer Lett.* 378 (2), 142–149. doi:10.1016/j.canlet.2016.05.020

Li, S.-f., Nakayama, K., Masuzawa, H., and Fujii, S. (1993). The Number of Proliferating Cell Nuclear Antigen Positive Cells in Endometriotic Lesions Differs from that in the Endometrium. *Vichows Arch. A Pathol. Anat.* 423 (4), 257–263. doi:10.1007/bf01606888

Liu, Z., Morgan, S., Ren, J., Wang, Q., Annis, D. S., Mosher, D. F., et al. (2015). Thrombospondin-1 (TSP1) Contributes to the Development of Vascular Inflammation by Regulating Monocytic Cell Motility in Mouse Models of Abdominal Aortic Aneurysm. *Circ. Res.* 117 (2), 129–141. doi:10.1161/circresaha.117.305262

Lucas, E. S., Vrljicak, P., Muter, J., Diniz-da-Costa, M. M., Brighton, P. J., Kong, C.-S., et al. (2020). Recurrent Pregnancy Loss Is Associated with a Pro-senescent Decidual Response during the Peri-Implantation Window. *Commun. Biol.* 3 (1), 37. doi:10.1038/s42003-020-0763-1

Naciff, J. M., Jump, M., Torontali, S., Carr, G., Tiesman, J., Overmann, G., et al. (2002). Gene Expression Profile Induced by 17alpha-Ethynyl Estradiol, Bisphenol A, and Genistein in the Developing Female Reproductive System of the Rat. *Toxicol. Sci.* 68 (1), 184–199. doi:10.1093/toxsci/68.1.184

Nagaev, I., Bokarewa, M., Tarkowski, A., and Smith, U. (2006). Human Resistin Is a Systemic Immune-Derived Proinflammatory Cytokine Targeting Both Leukocytes and Adipocytes. *PloS one* 1 (undefined), e31. doi:10.1371/journal.pone.0000031

Noci, I., Borri, P., Chieffi, O., Scarselli, G., Biagiotti, R., Moncini, D., et al. (1995). I. Aging of the Human Endometrium: a Basic Morphological and Immunohistochemical Study. *Eur. J. Obstetrics Gynecol. Reproductive Biol.* 63 (2), 181–185. doi:10.1016/0301-2115(95)02244-9

Oertelt-Prigione, S. (2012). Immunology and the Menstrual Cycle. *Autoimmun. Rev.* 11 (6-7), A486–A492. doi:10.1016/j.autrev.2011.11.023

Owens, B. M. J., Moore, J. W. J., and Kaye, P. M. (2012). IRF7 Regulates TLR2-Mediated Activation of Splenic CD11chi Dendritic Cells. *Plos One* 7 (7), e41050. doi:10.1371/journal.pone.0041050

Patel, V. K., Williams, H., Li, S. C. H., Fletcher, J. P., and Medbury, H. J. (2017). Monocyte Inflammatory Profile Is Specific for Individuals and Associated with Altered Blood Lipid Levels. *Atherosclerosis* 263 (undefined), 15–23. doi:10.1016/j.atherosclerosis.2017.05.026

Patterson, A. L., Zhang, L., Arango, N. A., Teixeira, J., and Pru, J. K. (2013). Mesenchymal-to-epithelial Transition Contributes to Endometrial Regeneration Following Natural and Artificial Decidualization. *Stem Cells Dev.* 22 (6), 964–974. doi:10.1089/scd.2012.0435

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed Graph Embedding Resolves Complex Single-Cell Trajectories. *Nat. Methods* 14 (10), 979–982. doi:10.1038/nmeth.4402

Reese, J., Das, S. K., Paria, B. C., Lim, H., Song, H., Matsumoto, H., et al. (2001). Global Gene Expression Analysis to Identify Molecular Markers of Uterine Receptivity and Embryo Implantation. *J. Biol. Chem.* 276 (47), 44137–44145. doi:10.1074/jbc.m107563200

Regan-Komito, D., Valaris, S., Kapellos, T. S., Recio, C., Taylor, L., Greaves, D. R., et al. (2017). Absence of the Non-signalling Chemerin Receptor CCRL2 Exacerbates Acute Inflammatory Responses *In Vivo*. *Front. Immunol.* 8 (undefined), 1621. doi:10.3389/fimmu.2017.01621

Ruiz-Rosado, J. d. D., Olguín, J. E., Juárez-Avelar, I., Saavedra, R., Terrazas, L. I., Robledo-Avila, F. H., et al. (2016). MIF Promotes Classical Activation and Conversion of Inflammatory Ly6C(high) Monocytes into TipDCs during Murine Toxoplasmosis. *Mediat. Inflamm.* 2016 (5), 1–18. doi:10.1155/2016/9101762

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial Reconstruction of Single-Cell Gene Expression Data. *Nat. Biotechnol.* 33 (5), 495–502. doi:10.1038/nbt.3192

Sena, A. A. S., Glavan, T., Jiang, G., Sankaran-Walters, S., Grishina, I., Dandekar, S., et al. (2016). Divergent Annexin A1 Expression in Periphery and Gut Is Associated with Systemic Immune Activation and Impaired Gut Immune Response during SIV Infection. *Sci. Rep.* 6 (1), 31157. doi:10.1038/srep31157

Vidal, J. D., and Filgo, A. J. (2017). Evaluation of the Estrous Cycle, Reproductive Tract, and Mammary Gland in Female Mice. *Curr. Protoc. mouse Biol.* 7 (4), 306–325. doi:10.1002/cpmo.35

Walter, W., Sánchez-Cabo, F., and Ricote, M. (2015). GOplot: an R Package for Visually Combining Expression Data with Functional Analysis. *Bioinformatics* 31 (17), 2912–2914. doi:10.1093/bioinformatics/btv300

Wu, B., An, C., Li, Y., Yin, Z., Gong, L., Li, Z., et al. (2017). Reconstructing Lineage Hierarchies of Mouse Uterus Epithelial Development Using Single-Cell Analysis. *Stem Cell Rep.* 9 (1), 381–396. doi:10.1016/j.stemcr.2017.05.022

Wu, B., Li, Y., Nie, N., Xu, J., An, C., Liu, Y., et al. (2019). Nano Genome Altas (NGA) of Body Wide Organ Responses. *Biomaterials* 205, 38–49. doi:10.1016/j.biomaterials.2019.03.019

Xuan, W., Qu, Q., Zheng, B., Xiong, S., and Fan, G.-H. (2015). The Chemotaxis of M1 and M2 Macrophages Is Regulated by Different Chemokines. *J. Leukoc. Biol.* 97 (1), 61–69. doi:10.1189/jlb.1a0314-170r

Yashiro, T., Nakano, S., Nomura, K., Uchida, Y., Kasakura, K., and Nishiyama, C. (2019). A Transcription Factor PU.1 Is Critical for Ccl22 Gene Expression in Dendritic Cells and Macrophages. *Sci. Rep.* 9 (1), 1161. doi:10.1038/s41598-018-37894-9

Yin, M., Zhou, H. J., Lin, C., Long, L., Yang, X., Zhang, H., et al. (2019). CD34+KLF4+ Stromal Stem Cells Contribute to Endometrial Regeneration and Repair. *Cell Rep.* 27 (9), 2709–2724. doi:10.1016/j.celrep.2019.04.088

Yip, K. S., Suvorov, A., Connerney, J., Lodato, N. J., and Waxman, D. J. (2013). Changes in Mouse Uterine Transcriptome in Estrus and Proestrus1. *Biol. reproduction* 89 (1), 13. doi:10.1095/biolreprod.112.107334

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, X., Li, X., Ning, F., Shang, Y., and Hu, X. (2019). TLE4 Acts as a Corepressor of Hes1 to Inhibit Inflammatory Responses in Macrophages. *Protein Cell* 10 (4), 300–305. doi:10.1007/s13238-018-0554-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 25 May 2022
doi: 10.3389/fgene.2022.879974

Identification of Biomarkers for Predicting Ovarian Reserve of Primordial Follicle *via* Transcriptomic Analysis

Li Liu[1,2], Biting Liu[1], Ke Li[3], Chunyan Wang[1], Yan Xie[1], Ning Luo[1,2], Lian Wang[1,2], Yaoqi Sun[1,2], Wei Huang[1,2], Zhongping Cheng[1,2] and Shupeng Liu[1,2]**

[1]Department of Obstetrics and Gynecology, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China, [2]Institute of Gynecological Minimally Invasive Medicine, School of Medicine, Tongji University, Shanghai, China, [3]Department of Clinical Laboratory Medicine, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China

Ovarian reserve (OR) is mainly determined by the number of primordial follicles in the ovary and continuously depleted until ovarian senescence. With the development of assisted reproductive technology such as ovarian tissue cryopreservation and autotransplantation, growing demand has arisen for objective assessment of OR at the histological level. However, no specific biomarkers of OR can be used effectively in clinic nowadays. Herein, bulk RNA-seq datasets of the murine ovary with the biological ovarian age (BOA) dynamic changes and single-cell RNA-seq datasets of follicles at different stages of folliculogenesis were obtained from the GEO database to identify gene signature correlated to the primordial follicle pool. The correlations between gene signature expression and OR were also validated in several comparative OR models. The results showed that genes including *Lhx8, Nobox*, *Sohlh1*, *Tbpl2*, *Stk31*, and *Padi6* were highly correlated to the OR of the primordial follicle pool, suggesting that these genes might be used as biomarkers for predicting OR at the histological level.

Keywords: ovarian reserve, biomarkers, bioinformatics, assisted reproductive technology, transcriptome

OPEN ACCESS

Edited by:
Zodwa Dlamini, SAMRC Precision Oncology Research Unit (PORU), South Africa

Reviewed by:
Katja Hummitzsch, University of Adelaide, Australia
Solomon O. Rotimi, Covenant University, Nigeria

****Correspondence:***
Zhongping Cheng
mdcheng18@tongji.edu.cn
Shupeng Liu
lshup@tongji.edu.cn

Specialty section:
This article was submitted to Genetics of Aging, a section of the journal Frontiers in Genetics

Received: *20 February 2022*
Accepted: *04 May 2022*
Published: *25 May 2022*

Citation:
Liu L, Liu B, Li K, Wang C, Xie Y, Luo N, Wang L, Sun Y, Huang W, Cheng Z and Liu S (2022) Identification of Biomarkers for Predicting Ovarian Reserve of Primordial Follicle via Transcriptomic Analysis. Front. Genet. 13:879974. doi: 10.3389/fgene.2022.879974

INTRODUCTION

Ovarian reserve (OR) is used to describe the ovarian ability to provide viable oocytes, mainly determined by the number of primordial follicles in the ovary, which continuously depleted until ovarian senescence (Moolhuijsen and Visser, 2020; Ruth et al., 2021). Therefore, OR declined along with ovary aging, and the biological ovarian age (BOA) in humans has a lifespan of about 50 years (Alviggi et al., 2009; Moolhuijsen and Visser, 2020; Ruth et al., 2021). Several pathological factors, including chromosomal abnormalities, autoimmune disorders, and iatrogenic injuries, can accelerate the depletion of OR, leading to the diminished ovarian reserve (DOR) (Steiner et al., 2017; Spears et al., 2019; Takahashi et al., 2021).

With the development of assisted reproductive technology (ART), such as ovarian tissue cryopreservation (OTC) (Anderson et al., 2017) and transplantation (OTP) (Dolmans et al., 2021), *In vitro* activation (IVA), and growth (IVG) of primordial follicles (Telfer and Andersen, 2021), there is an increasing need for the technical operations directly in ovarian tissue obtained by surgical resection, especially the evaluation of OR at the histological level. In the OTC

procedure, for example, ovarian tissue from each patient needs to be processed into several fixed-size ovarian tissue slices (e.g., 0.5 cm*0.5 cm*0.1 cm), whereas the evaluation of OR in ovarian tissue slices is not yet standardized. Traditionally, the gold standard for assessing OR counts the number of primordial follicles on serial H&E/IHC-stained sections slice by slice. However, the time-consuming and highly subjective method makes it inconceivable and ineffective in the clinic (Myers et al., 2004; Youm et al., 2014; Terren et al., 2019; Mahmoudi Asl et al., 2021). Serological hormone testing and ultrasonography are currently used for evaluating OR in the clinic. These methods are based on assessing the quality of the growing follicles but not the number of primordial follicles (Steiner et al., 2017; Lew, 2019). To date, Anti-Müllerian hormone (AMH) is a promising serum marker for assessing OR at the serological level (Visser et al., 2012). However, since AMH is not an oocyte-specific marker, its applicability in histological assessing OR is bound to be limited by many factors, such as the patient's age and the pathological abnormalities of polycystic ovary syndrome (PCOS) (Dewailly et al., 2014; Moolhuijsen and Visser, 2020). Therefore, an effective and easily used method is still needed for the clinical assessment of OR at the histological level.

Currently, increasing transcriptomic data makes it possible to find out genes specifically expressed in unique types of cells. The present study aimed to identify potential biomarkers for predicting OR by combinational analysis of published transcriptomic data from bulk RNA-seq and single-cell RNA-seq of humans and mice ovarian tissues. We hoped our findings will provide new insights into the clinical evaluation of OR and targeted interventions for fertility preservation.

MATERIALS AND METHODS

Animals and Sample Preparation

The mice were raised in an environment with a temperature of between 18 and 23°C and humidity of between 40 and 60% under 12-h light/dark cycles. Animal experiments were approved by the Animal Ethics Committee of Shanghai Tenth People's Hospital (No. SHDSYY-2020-Y0688). 2-month-old (n = 3) and 8-month-old (n = 3) female C57BL/6L mice under specific pathogen-free conditions were provided by Shanghai SLAC Laboratory Animal Corp, Shanghai, China. Bilateral ovariectomy was performed, and ovarian bursas were removed under the microscope according to the previous study (Souza et al., 2019). After three washes with sterile PBS solution, samples were collected from the bilateral ovaries of each mouse for RNA extraction.

RNA-Seq and Microarray Profiles Acquisition

Raw files of five registered datasets used in this study were downloaded from NCBI GEO (http://www.ncbi.nlm.nih.gov/geo) (Barrett et al., 2013). The GSE154890 and GSE179888 datasets include bulk RNA-seq data of murine ovary with BOA dynamic changes. The GSE107746 dataset includes single-cell RNA-seq data of human oocytes and granulosa cells of follicles during dynamic folliculogenesis. The GSE7502 and GSE109473 containing microarray profiles of murine ovary were retained for subsequent analyses. For better understanding, the characteristics and processing procedures of datasets were described in **Table 1** and the flowchart (**Figure 1A**).

Identification of Differentially Expressed Genes

The process of raw data and analysis of differentially expressed genes (DEGs) were performed using the Sangerbox tools, a free online platform for data analysis (http://www.sangerbox.com/tool). DEGs were screened for GSE154890 and GSE179888 datasets with the threshold criterion of p-value < 0.05 and log2 |FC| >1. DEGs of GSE107746 were screened for gene expression of log2 (FPKM+1) with the threshold criterion of adj. p-value < 0.05 and log2 |FC| >2. Overlapping DEGs were identified using the Venn diagram web tool (https://bioinformatics.psb.ugent.be/webtools/Venn/).

Analysis of Human and Mouse Homologous Genes

Human and mouse homologous genes were downloaded from the Vertebrate Homology Database (http://www.informatics.jax.org/homology.shtml). Mouse candidate genes were overlapped with human homologous genes to find homologous genes expressed in humans and mice.

Total RNA Extraction and Quantitative Real-Time PCR

The extraction of the total RNA was completed with RNAiso Plus (TaKaRa, Japan) according to the manufacturer's instructions. RNA quality and quantity were measured by NanoDrop2000 (Thermo Scientific, Wilmington, DE, United States). Then RNA was reverse-transcribed into complementary DNA using PrimeScript RTMaster Mix (TaKaRa, Japan). Quantitative real-time PCRs (qRT-PCR) were performed on a QuantStudio Dx (ABI, America) using the SYBR Premix ExTaq kit (Takara, Shiga, Japan). The thermal cycler conditions were as follows: 30 s at 95.0°C for cDNA denatured, followed by 40 cycles of 15 s at 95°C and 60°C for the 30 s. Verification of specific product amplification was performed by dissociation curve analysis. The mRNA relative expression was calculated by the 2^{-}-ΔΔCt method with *Gapdh* as an internal control (Livak and Schmittgen, 2001). The experiment was repeated in triplicate, and all primer sequences were listed in **Table 2**.

Statistical Analysis

Data analysis was performed using SPSS 20.0 statistical software (IBM, New York, NY, United States). Quantitative

TABLE 1 | Characteristics of datasets used in this study.

GSE series	Data type	Platforms	Application	Sample used category
GSE154890	Bulk RNA-seq	GPL16417	Identification	3/6/9/12months
GSE179888	Bulk RNA-seq	GPL17021	Identification	p3/7/14/21/60/y1/y2
GSE107746	scRNA-seq	GPL20795	Identification	ALL
GSE7502	microarray	GPL2552	Validation	O_AL_1/6/16/24 months
GSE109473	microarray	GPL6887	Validation	Lhx8_P7_WT/KO

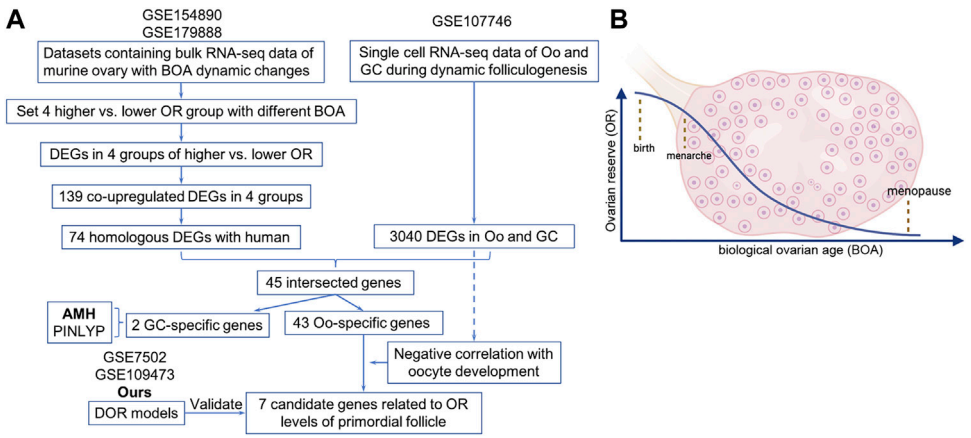

FIGURE 1 | The flowchart of the design and analysis of this study.

TABLE 2 | Primers used for qPCR validation.

Gene (Mouse)	Forward sequence	Reverse sequence
Lhx8	AGCACAGTTCGCTCAGGACAAC	GCTGAGGAAGAATGGTTGGGAC
Nobox	CGTTCCTGGCAGTGACAGCATA	GGAATGAACCCAACTGGCTGCT
Sohlh1	GCCAAACCATCTGCTGTGTCTC	AAGGTCTCTCCAGCAGCTCTGA
Tbpl2	ACTCCAATGCCTTACCTGTGGC	GCCAGATTTGCAGTGGAAACTAC
Padi6	CTGAGCGAGAAGAGCAAAGTGC	ATGACACCGTCTTGTGAGGAGC
Stk31	CGTGTGTAGGAACCAGGCTGAA	GGACCCTTCATCCAACACTTGC
Vrtn	ACCAAGAGCACCTTCTACCGCT	GAACTGCTGCAATGGCACAAAGC

data were expressed as the standard deviation of the mean. Analysis of quantitative parametric data between two groups was assessed with Student's t-test. Correlational statistical analysis was completed with the Spearman correlation test in this article. p values <0.05 were considered statistically significant.

RESULTS

Identification of Differentially Expressed Genes Correlated With Biological Ovarian Age

Since the natural decline of OR changes and BOA (**Figure 1B**), genes differentially expressed with BOA were first identified. Two bulk RNA-seq datasets (GSE154890 and GSE179888) containing transcriptome data of mice ovary tissues at different ages were obtained from the GEO database and sent for further differentially expressed genes (DEGs) analysis. Four pairs of higher vs. lower OR groups, namely "3 vs. 9 m", "6 vs. 12 m", "p14 vs. p60" and "p14 vs. y1" were respectively conducted simulating young vs. old BOA comparison according to murine age referenced to human as previously reported (Dutta and Sengupta, 2016). DEGs of each group were screened with the threshold criterion of Log2|FC| > 1 and p value < 0.05. The cluster analysis showed that ovarian tissues at different ages had unique gene expression profiles (**Figure 2A**). Multiple DEGs were identified in each group, with 1,640 DEGs in 3 vs. 9 m, 894 DEGs in 6 vs. 12 m, 5,620 DEGs in p14 vs. p60, and 6,485 DEGs in p14 vs. y1, respectively (**Figure 2B**). The

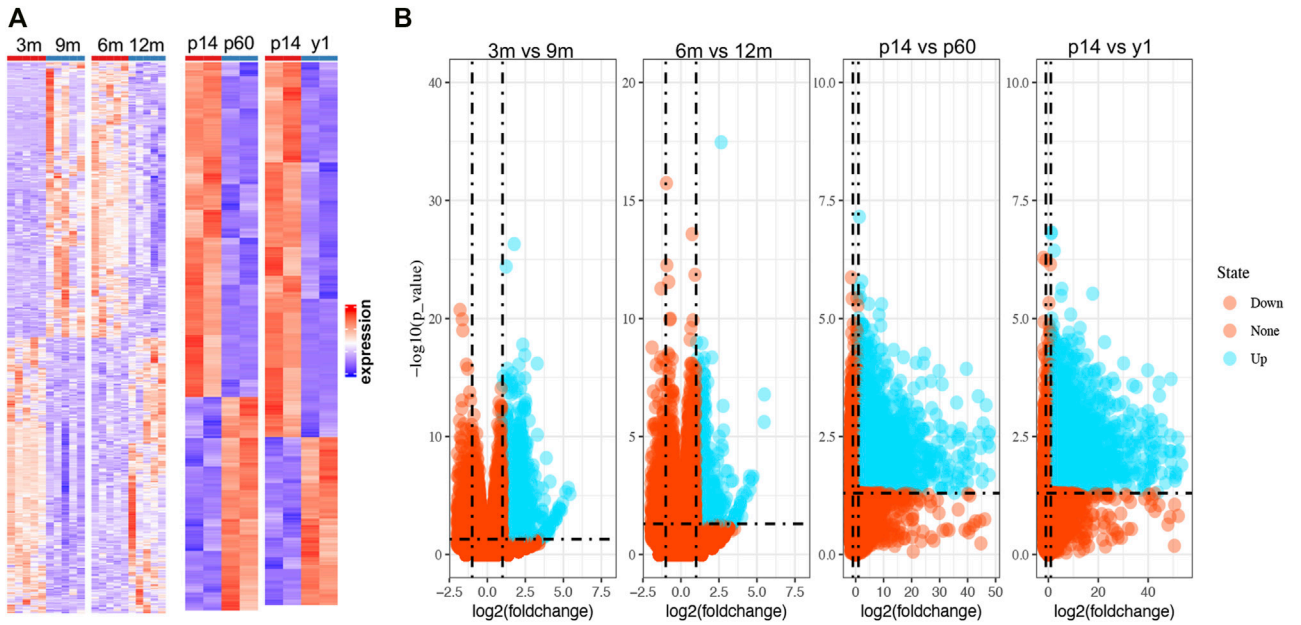

FIGURE 2 | The heatmaps and volcano plots of DEGs in multiple OR comparison groups. **(A)** Heatmaps for DEGs in four OR comparison groups from GSE154890 and GSE179888. The red shading represents the upregulated genes, while the red shading represents the down-regulated genes. **(B)** Volcano plots for DEGs. The blue points represent the upregulated genes, and the red points represent the non-upregulated genes in high OR. DEGs were screened on the criterion of log2|FC| > 1.0 and $p < 0.05$.

results showed that these DEGs might be involved in murine BOA.

Screening the Differentially Expressed Genes Related to Human Oocyte

Among the DEGs identified above, upregulated DEGs in each younger group were selected for further analysis considering higher OR in the more immature ovaries. To make it more reliable, intersection analysis was firstly performed and found a total of 139 DEGs upregulated in all four sets of groups, including 74 human homologous genes (**Figures 3A,B**). Considering the uniqueness of follicular structure and folliculogenesis, we hypothesized that potential biomarkers should be expressed explicitly in oocytes. Then the expression of 74 human homologous genes in human oocytes (Oo) and granulosa cells (GC) was investigated using the single-cell sequencing dataset of human Oo and GC (GSE107746). The analysis of the single-cell sequencing dataset identified 3,040 DEGs between human Oo and GC (**Figure 3C**), including 45 out of 74 human homologous genes identified in murine transcriptomic data (**Figure 3D**). Among 45 DEGs from both human and murine transcriptomic data, 43 genes were specifically highly expressed in Oo, and two were specifically highly expressed in GC (**Figure 3D**). It was suggested that 43 Oo specifically expressed genes may be correlated with human OR. Interestingly, our results revealed that the highest fold change gene of these 2 GC-specific genes was *Amh*, the well-known serum marker for OR (**Supplymentary Figure S1**), validating the feasibility of our screening process.

Identification of Potential Genes Related to the Follicular Developmental Stages

To further investigate the relationship of 43 Oo-specific genes during folliculogenesis, these gene expression patterns and follicular developmental stages were evaluated using the single-cell sequencing dataset (GSE107746). The heatmap showed that these Oo-specific candidate genes have different expression patterns during folliculogenesis (**Figure 4**). As the volume and genetic abundance of oocytes increase with the progressive development of oocytes from primordial follicles to preovulatory follicles, the best potential biomarkers should most likely represent the early stages (primordial and primary stages) during folliculogenesis. Our analysis showed that seven genes, including *Lhx8*, *Sohlh1*, *Nobox*, *Stk31*, *Tbpl2*, *Padi6*, and *Vrtn*, were significantly negatively correlated with the follicular developmental stages (**Figure 4**). Close investigation showed that these seven genes expressed in Oo declined along with folliculogenesis and remained low in GC (**Figure 5A**). Moreover, the expression of these genes in the murine ovary at different age were also investigated (GSE179888). It showed that expression of these genes declined significantly and correlated negatively with age (**Figure 5B**).

In addition, the expression of oocyte-specific genes that were significantly positively correlated with follicular developmental stages was also investigated. Increased expression of these genes was observed in Oo and folliculogenesis and remained low in GC (**Supplymentary Figure S2A**). In the murine ovary, the expression of these genes was similar to *Amh*, rising to a peak at a one-time point and declining to baseline after that, which was

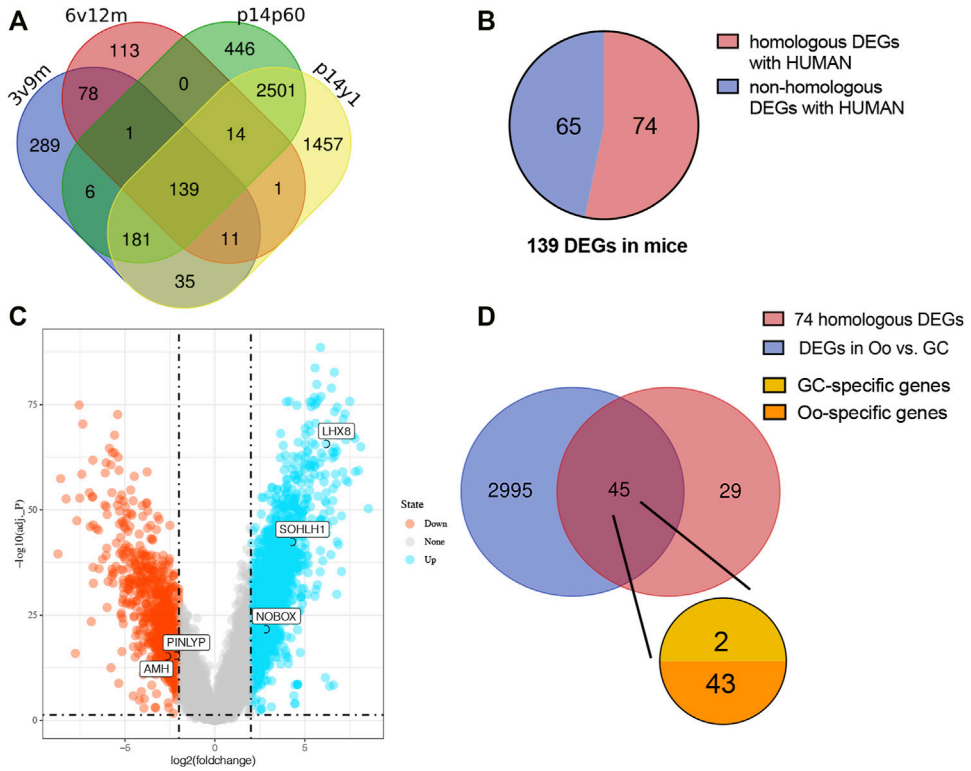

FIGURE 3 | Screening the intersected upregulate DEGs related to human oocyte **(A)** Venn diagram demonstrates the overlap of the murine upregulated DEGs in four higher vs. lower OR groups of BOA comparison. **(B)** The pie chart represents the 139 intersected upregulated DEGs homologous with humans (red) or not (blue). **(C)** Volcano plots for DEGs of human oocytes (Oo) and granulosa cells (GC) from dataset GSE107746. The blue points represent the upregulated genes in Oo, and the red points represent the upregulated genes in GC. DEGs were screened on the criterion of log2|FC| > 2.0 and adj. $p < 0.05$. **(D)** Venn diagram of 74 upregulated murine the homologous DEGs and the DEGs in human Oo and GC (blue) demonstrates the overlap of the upregulated murine DEGs related to human Oo- (orange) or GC-specific (yellow) expression.

inconsistent with the natural decline of OR (**Supplymentary Figure S2B**). Taken together, these data showed that *Lhx8*, *Sohlh1*, *Nobox*, *Stk31*, *Tbpl2*, *Padi6*, and *Vrtn* strongly negatively correlated with BOA were potentially biomarkers for evaluating OR at the histological level.

Validation of the Potential Ovarian Reserve-Related Biomarkers

To further investigate the relationship between the expression of these seven genes and OR, we further validated the expression levels of these genes with comparative OR models from the GEO database and our own established. The RNA expression of the entire murine ovary with different murine BOA was obtained by analyzing the RNA microarray dataset GSE7502, and only a probe for *Lhx8* was available for these seven genes mentioned above. The data showed that *Lhx8* expression in mouse ovaries decreased time-dependent and was significantly lower in the low OR group than in younger ovaries (**Figure 6A**). Then the RNA microarray dataset from *Lhx8* knockout mice was included (GSE109473). It has been reported that *Lhx8* knockout leads to massive depletion of the primordial follicle and induces OR depletion in mice (Ren et al., 2015; D'Ignazio et al., 2018). The results showed that the expression of six genes, including *Lhx8*, *Sohlh1*, *Nobox*, *Stk31*, *Tbpl2*, and *Padi6*, was significantly reduced in ovary tissues from *Lhx8* knockout mice, while no probe of *Vrtn* was observed (**Figure 6B**).

Moreover, we evaluated the expression of these seven genes in our own established comparative OR models by using quantitative real-time PCR. It showed that six out of the seven genes were expressed at a lower level in low OR ovaries than in young ovaries, while *Vrtn* expression showed no difference (**Figure 6C**). These data further revealed that the expression of six genes, including *Lhx8*, *Sohlh1*, *Nobox*, *Stk31*, *Tbpl2*, and *Padi6*, was higher in high OR ovaries. And it suggested that these genes expression might be correlated with OR.

DISCUSSION

The present study identified that seven homologous genes of humans and mice, including *Vrtn*, were highly correlated to OR by combinational analysis of bulk RNA-seq and single-cell RNA-seq data. Further verification consolidated the correlation of six genes with OR based on the published transcriptomic data and quantitative real-time PCR analysis, including *Lhx8*, *Sohlh1*, *Nobox*, *Stk31*, *Tbpl2*, and *Padi6*. Our findings suggested that

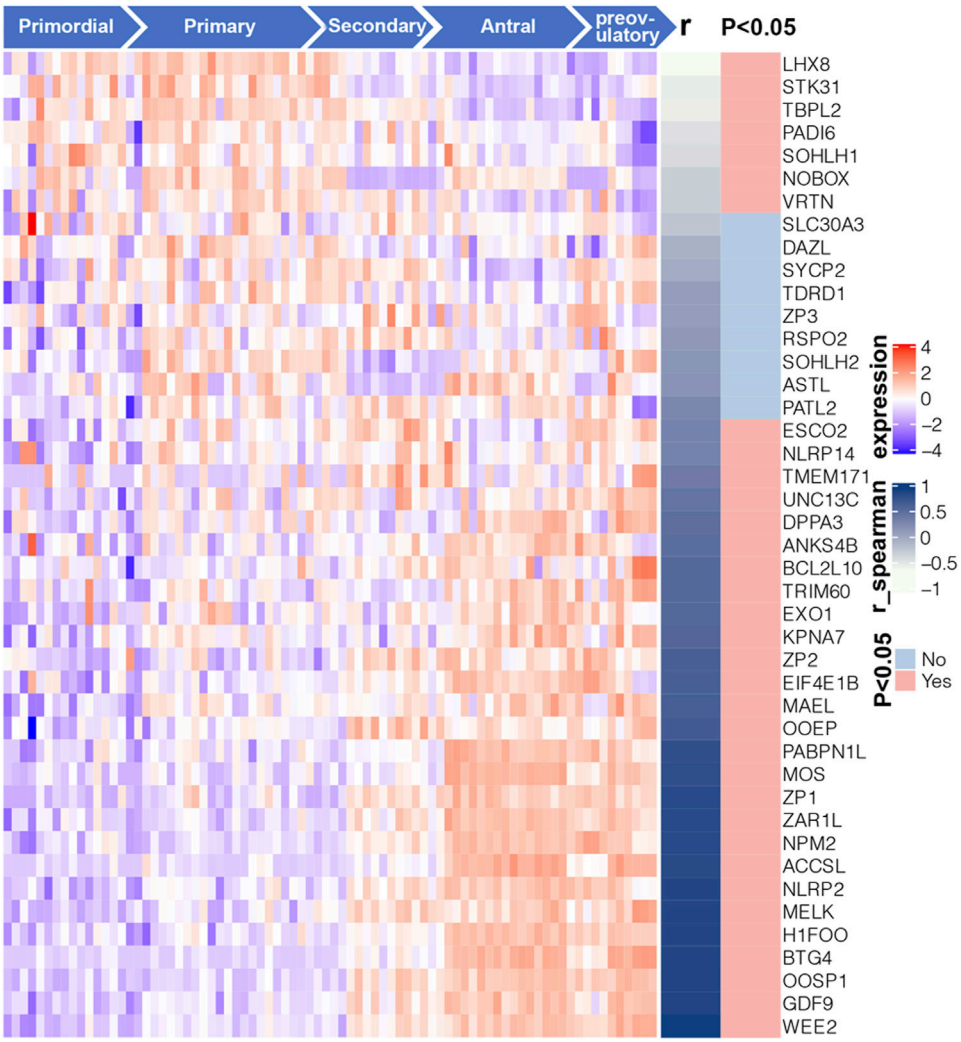

FIGURE 4 | Expression profile and correlation with follicular developmental of 43 candidate Oo-specific genes. Heatmap displays the expression levels of 43 candidate Oo-specific genes in the oocytes at five stages of folliculogenesis in the GSE107746 dataset. The right panel of the heatmap shows the correlation between the expression levels of each gene with the follicular developmental stages in the GSE107746 dataset.

these six genes might be used as potential biomarkers for evaluating OR at the histological level in mice and humans.

Traditionally, the histological approach to assess the entire ovary's OR is time-consuming and somewhat statistically subjective by counting the number of primordial follicles in serial H&E/IHC staining sections (Youm et al., 2014; Terren et al., 2019; Mahmoudi Asl et al., 2021). In current clinical practice, the preferred choice for assessing OR is the non-invasive blood detection of serum anti-Müllerian hormone (AMH) (Di Clemente et al., 2021). AMH is secreted by granulosa cells of small growing follicles in the ovary as a potent inhibitor of primordial follicle recruitment. Serum AMH levels strongly correlate with the number of growing follicles, and therefore AMH has received increasing attention as an indirect marker to assess OR (Moolhuijsen and Visser, 2020; Di Clemente et al., 2021). However, the specificity was affected by several factors, such as the age category analyzed and abnormal expression in PCOS (Moolhuijsen and Visser, 2020). A previous study reported that AMH levels increased to plateau at the approximate age of 25 years. Only from this age onward, serum AMH levels start to decline to undetectable levels at menopause, and a negative correlation between serum AMH levels and ages can be observed (Lie Fong et al., 2012). We also observed a similar expression pattern of AMH gene expression with BOA at the histological level (**Supplymentary Figure S1**), suggesting that it is not appropriate for AMH to represent the natural decline of the primordial follicle pool.

In the present study, we found that the mRNA expression level of six genes was positively correlated with OR in both humans and mice. And the expression of these genes declined linearly along with aging (**Figure 5**). Additionally, previous studies have documented that *Lhx8*, *Sohlh1*, and *Nobox* have

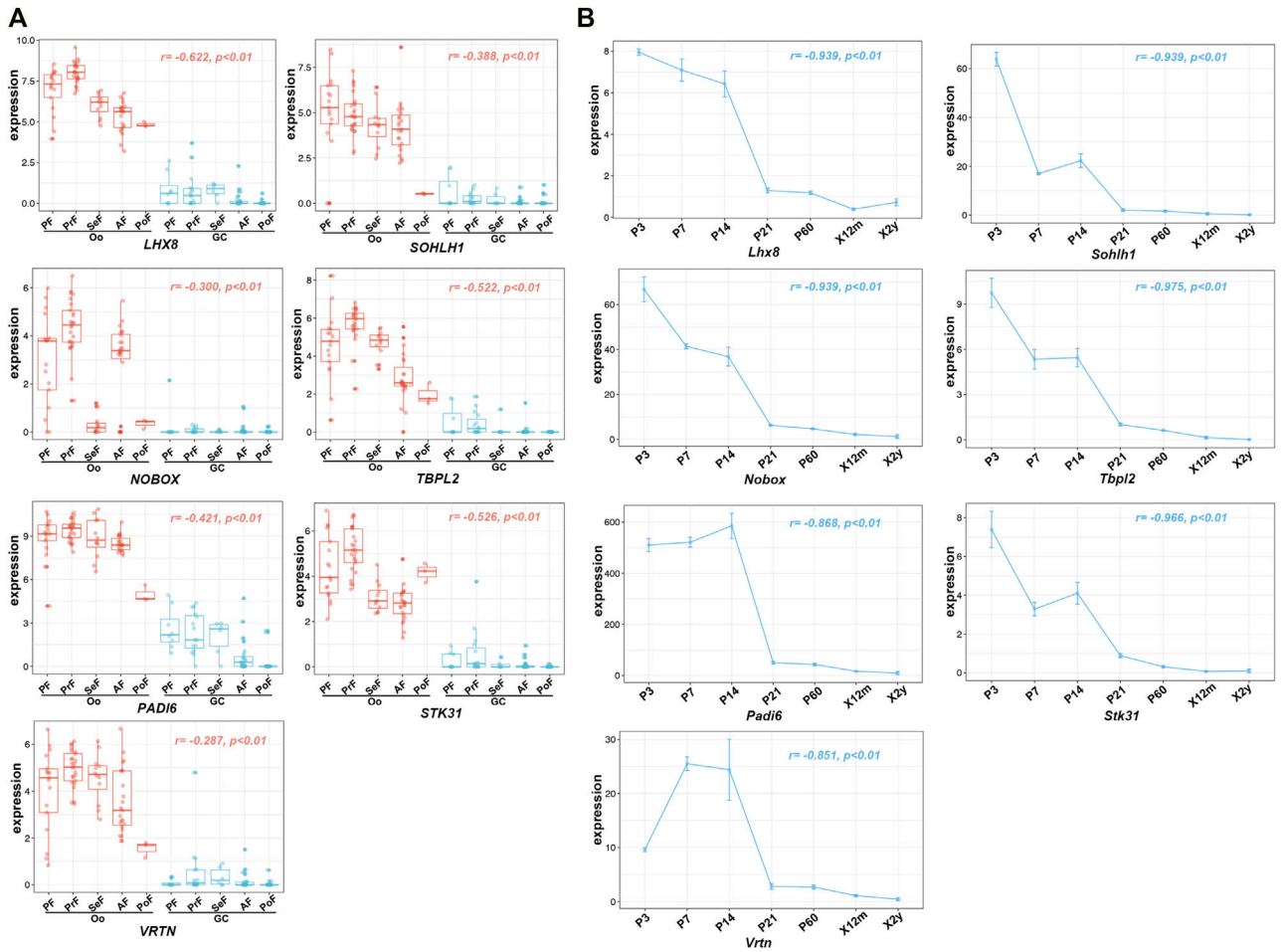

FIGURE 5 | Expression of seven identified potential genes in GSE107746 and GSE179888 datasets. **(A)** The expression levels of seven identified potential genes, *Lhx8*, *Sohlh1*, *Nobox*, *Stk31*, *Tbpl2*, *Padi6*, and *Vrtn* in the human Oo or GC at five stages of folliculogenesis in the GSE107746 dataset. Oo, oocyte; GC, granulosa cells; PF, primordial follicle; PrF, primary follicle; SeF, secondary follicle; AF, antral follicle; PoF, preovulatory follicle; r, correlation coefficient. **(B)** The expression trends with BOA changes in murine ovaries from the GSE179888 dataset. P, postnatal day; X12m, postnatal 12 months; X2y, postnatal 2 years; r, correlation coefficient; p, *p*-value.

critical roles in regulating primordial follicle activation. Knockout of any of these transcriptional factors causes rapid oocyte loss and ovarian failure (Ren et al., 2015; D'Ignazio et al., 2018; Wang et al., 2020). *Padi6* encodes an enzyme of the peptidyl arginine deiminase family and is uniquely expressed in male and female germ cells. The absence of PADI6 protein results in the dispersal of cytoskeletal sheets in oocytes, which ultimately leads to female infertility (Xiong et al., 2019; Bebbere et al., 2020). Even though *Stk31*, *Tbpl2*, and *Vrtn* are poorly studied in follicle development at the current stage, emerging evidence suggests that these genes are specifically expressed in oocytes and play a crucial role during folliculogenesis (Olesen et al., 2007; Di Pietro et al., 2008; Duan et al., 2018; Yu et al., 2020; Yang et al., 2021). These findings indicated that evaluating OR will be more specific *via* assessing the six gene expression.

Nowadays, growing demand has arisen for objective assessment of OR at the histological level with the development of ART. In the OTC procedure, for example, ovarian tissue from each patient needs to be processed into several fixed-size ovarian tissue slices (e.g., 0.5 cm*0.5 cm*0.1 cm), whereas the evaluation of OR in ovarian tissue slices is not yet standardized (Anderson et al., 2017). The traditional histological assessment of OR is challenging to apply in clinics, while our findings might provide a simple and effective method by detecting these potential biomarkers to evaluate the OR level of ovarian tissue slices routinely collected during OTC without any additional tissue to be collected.

However, several limitations remain in the present study. Firstly, the correlation between the exact primordial follicle numbers and mRNA expression levels needs to be verified. Due to the absence of samples for human DOR in public databases, the expression pattern of these genes and the correlation with OR in human ovarian tissue remains unclear. The sensitivity and specificity of these genes used as

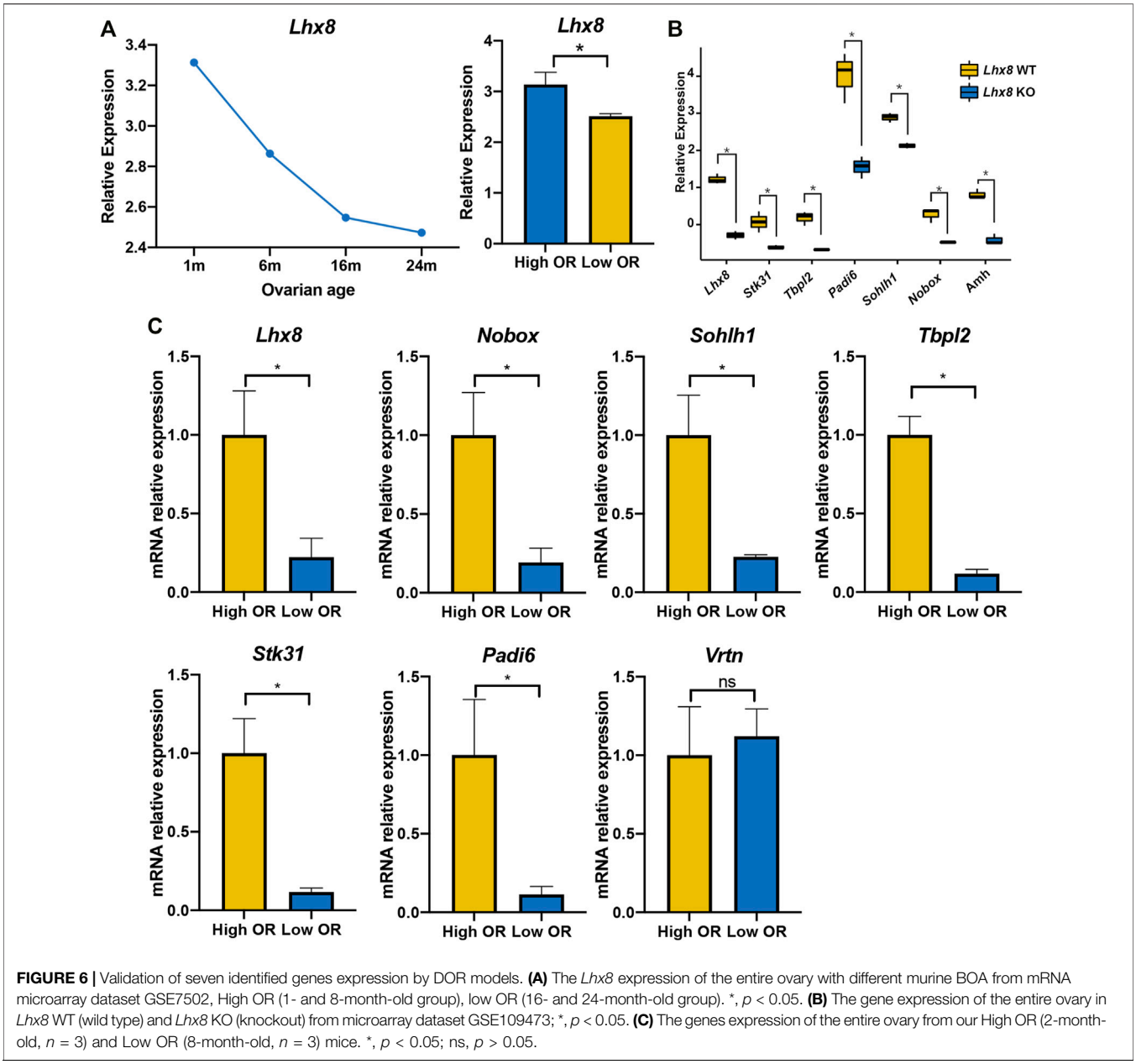

FIGURE 6 | Validation of seven identified genes expression by DOR models. **(A)** The *Lhx8* expression of the entire ovary with different murine BOA from mRNA microarray dataset GSE7502, High OR (1- and 8-month-old group), low OR (16- and 24-month-old group). *, $p < 0.05$. **(B)** The gene expression of the entire ovary in *Lhx8* WT (wild type) and *Lhx8* KO (knockout) from microarray dataset GSE109473; *, $p < 0.05$. **(C)** The genes expression of the entire ovary from our High OR (2-month-old, $n = 3$) and Low OR (8-month-old, $n = 3$) mice. *, $p < 0.05$; ns, $p > 0.05$.

biomarkers for OR evaluation need to be validated at histological levels in the large cohort.

In conclusion, the expression level of six genes, including *Lhx8*, *Sohlh1*, *Nobox*, *Stk31*, *Tbpl2*, and *Padi6*, were highly correlated to OR of the primordial follicle pool, suggesting these genes might be used as potential biomarkers for evaluating OR of ovarian tissue both in humans and mice. With the development of ART such as OTC, OTP, growing demand has arisen for objective assessment of OR at the histological level. Our findings might provide a new perspective on clinical assessment of OR and targeted interventions for fertility preservation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Ethics Committee of Shanghai Tenth People's Hospital.

AUTHOR CONTRIBUTIONS

Conceptualization, ZC and SL; methodology, LL, CW, and ZC; software, YS, WH, and BL; validation, LL and KL; formal analysis, NL and YX; data curation, BL and LL; writing—original draft preparation, LL.; writing—review and editing, SL; funding acquisition, LW. All authors have read and agreed to the published version of the manuscript.

FUNDING

This project was supported by the National Nature Science Foundation of China (NO. 31900522).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.879974/full#supplementary-material

REFERENCES

Alviggi, C., Humaidan, P., Howles, C. M., Tredway, D., and Hillier, S. G. (2009). Biological versus Chronological Ovarian Age: Implications for Assisted Reproductive Technology. *Reprod. Biol. Endocrinol.* 7, 101. doi:10.1186/1477-7827-7-101

Anderson, R. A., Wallace, W. H. B., and Telfer, E. E. (2017). Ovarian Tissue Cryopreservation for Fertility Preservation: Clinical and Research Perspectives. *Hum. Reprod. Open* 2017, hox001. doi:10.1093/hropen/hox001

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for Functional Genomics Data Sets-Update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Bebbere, D., Abazari-Kia, A., Nieddu, S., Melis Murgia, B., Albertini, D. F., and Ledda, S. (2020). Subcortical Maternal Complex (SCMC) Expression during Folliculogenesis is Affected by Oocyte Donor Age in Sheep. *J. Assist. Reprod. Genet.* 37, 2259–2271. doi:10.1007/s10815-020-01871-x

Dewailly, D., Andersen, C. Y., Balen, A., Broekmans, F., Dilaver, N., Fanchin, R., et al. (2014). The Physiology and Clinical Utility of Anti-müllerian Hormone in Women. *Hum. Reprod. Update* 20, 370–385. doi:10.1093/humupd/dmt062

Di Clemente, N., Racine, C., Pierre, A., and Taieb, J. (2021). Anti-Müllerian Hormone in Female Reproduction. *Endocr. Rev.* 42, 753. doi:10.1210/endrev/bnab012

Di Pietro, C., Vento, M., Ragusa, M., Barbagallo, D., Guglielmino, M., Maniscalchi, T., et al. (2008). Expression Analysis of TFIID in Single Human Oocytes: New Potential Molecular Markers of Oocyte Quality. *Reprod. Biomed. Online* 17, 338–349. doi:10.1016/s1472-6483(10)60217-9

D'Ignazio, L., Michel, M., Beyer, M., Thompson, K., Forabosco, A., Schlessinger, D., et al. (2018). Lhx8 Ablation Leads to Massive Autophagy of Mouse Oocytes Associated with DNA Damage. *Biol. Reprod.* 98, 532–542. doi:10.1093/biolre/iox184

Dolmans, M.-M., von Wolff, M., Poirot, C., Diaz-Garcia, C., Cacciottola, L., Boissel, N., et al. (2021). Transplantation of Cryopreserved Ovarian Tissue in a Series of 285 Women: A Review of Five Leading European Centers. *Fertil. Steril.* 115, 1102–1115. doi:10.1016/j.fertnstert.2021.03.008

Duan, Y., Zhang, H., Zhang, Z., Gao, J., Yang, J., Wu, Z., et al. (2018). VRTN is Required for the Development of Thoracic Vertebrae in Mammals. *Int. J. Biol. Sci.* 14, 667–681. doi:10.7150/ijbs.23815

Dutta, S., and Sengupta, P. (2016). Men and Mice: Relating Their Ages. *Life Sci.* 152, 244–248. doi:10.1016/j.lfs.2015.10.025

Lew, R. (2019). Natural History of Ovarian Function Including Assessment of Ovarian Reserve and Premature Ovarian Failure. *Best Pract. Res. Clin. Obstetrics Gynaecol.* 55, 2–13. doi:10.1016/j.bpobgyn.2018.05.005

Lie Fong, S., Visser, J. A., Welt, C. K., de Rijke, Y. B., Eijkemans, M. J. C., Broekmans, F. J., et al. (2012). Serum Anti-müllerian Hormone Levels in Healthy Females: A Nomogram Ranging from Infancy to Adulthood. *J. Clin. Endocrinol. Metabolism* 97, 4650–4655. doi:10.1210/jc.2012-1440

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* 25, 402–408. doi:10.1006/meth.2001.1262

Mahmoudi Asl, M., Rahbarghazi, R., Beheshti, R., Alihemmati, A., Aliparasti, M. R., and Abedelahi, A. (2021). Effects of Different Vitrification Solutions and Protocol on Follicular Ultrastructure and Revascularization of Autografted Mouse Ovarian Tissue. *Cell J.* 22, 491–501. doi:10.22074/cellj.2021.6877

Moolhuijsen, L. M. E., and Visser, J. A. (2020). Anti-Müllerian Hormone and Ovarian Reserve: Update on Assessing Ovarian Function. *J. Clin. Endocrinol. Metab.* 105, 3361–3373. doi:10.1210/clinem/dgaa513

Myers, M., Britt, K. L., Wreford, N. G. M., Ebling, F. J. P., and Kerr, J. B. (2004). Methods for Quantifying Follicular Numbers within the Mouse Ovary. *Reproduction* 127, 569–580. doi:10.1530/rep.1.00095

Olesen, C., Nyeng, P., Kalisz, M., Jensen, T. H., Møller, M., Tommerup, N., et al. (2007). Global Gene Expression Analysis in Fetal Mouse Ovaries with and without Meiosis and Comparison of Selected Genes with Meiosis in the Testis. *Cell Tissue Res.* 328, 207–221. doi:10.1007/s00441-006-0205-5

Ren, Y., Suzuki, H., Jagarlamudi, K., Golnoski, K., McGuire, M., Lopes, R., et al. (2015). Lhx8 Regulates Primordial Follicle Activation and Postnatal Folliculogenesis. *BMC Biol.* 13, 39. doi:10.1186/s12915-015-0151-3

Ruth, K. S., Day, F. R., Hussain, J., Martínez-Marchal, A., Aiken, C. E., Azad, A., et al. (2021). Genetic Insights into Biological Mechanisms Governing Human Ovarian Ageing. *Nature* 596, 393–397. doi:10.1038/s41586-021-03779-7

Souza, V. R., Mendes, E., Casaro, M., Antiorio, A. T. F. B., Oliveira, F. A., and Ferreira, C. M. (2019). Description of Ovariectomy Protocol in Mice. *Methods Mol. Biol.* 1916, 303–309. doi:10.1007/978-1-4939-8994-2_29

Spears, N., Lopes, F., Stefansdottir, A., Rossi, V., De Felici, M., Anderson, R. A., et al. (2019). Ovarian Damage from Chemotherapy and Current Approaches to its Protection. *Hum. Reprod. Update* 25, 673–693. doi:10.1093/humupd/dmz027

Steiner, A. Z., Pritchard, D., Stanczyk, F. Z., Kesner, J. S., Meadows, J. W., Herring, A. H., et al. (2017). Association between Biomarkers of Ovarian Reserve and Infertility Among Older Women of Reproductive Age. *JAMA* 318, 1367–1376. doi:10.1001/jama.2017.14588

Takahashi, A., Yousif, A., Hong, L., and Chefetz, I. (2021). Premature Ovarian Insufficiency: Pathogenesis and Therapeutic Potential of Mesenchymal Stem Cell. *J. Mol. Med.* 99, 637–650. doi:10.1007/s00109-021-02055-5

Telfer, E. E., and Andersen, C. Y. (2021). *In Vitro* growth and Maturation of Primordial Follicles and Immature Oocytes. *Fertil. Steril.* 115, 1116–1125. doi:10.1016/j.fertnstert.2021.03.004

Terren, C., Fransolet, M., Ancion, M., Nisolle, M., and Munaut, C. (2019). Slow Freezing versus Vitrification of Mouse Ovaries: from *Ex Vivo* Analyses to Successful Pregnancies after Auto-Transplantation. *Sci. Rep.* 9, 19668. doi:10.1038/s41598-019-56182-8

Visser, J. A., Schipper, I., Laven, J. S. E., and Themmen, A. P. N. (2012). Anti-Müllerian Hormone: An Ovarian Reserve Marker in Primary Ovarian Insufficiency. *Nat. Rev. Endocrinol.* 8, 331–341. doi:10.1038/nrendo.2011.224

Wang, Z., Liu, C.-Y., Zhao, Y., and Dean, J. (2020). FIGLA, LHX8 and SOHLH1 Transcription Factor Networks Regulate Mouse Oocyte Growth and Differentiation. *Nucleic Acids Res.* 48, 3525–3541. doi:10.1093/nar/gkaa101

Xiong, J., Wu, M., Zhang, Q., Zhang, C., Xiong, G., Ma, L., et al. (2019). Proteomic Analysis of Mouse Ovaries during the Prepubertal Stages. *Exp. Cell Res.* 377, 36–46. doi:10.1016/j.yexcr.2019.02.016

Yang, P., Chen, T., Wu, K., Hou, Z., Zou, Y., Li, M., et al. (2021). A Homozygous Variant in TBPL2 Was Identified in Women with Oocyte Maturation Defects and Infertility. *Hum. Reprod.* 36, 2011–2019. doi:10.1093/humrep/deab094

Youm, H. W., Lee, J. R., Lee, J., Jee, B. C., Suh, C. S., and Kim, S. H. (2014). Optimal Vitrification Protocol for Mouse Ovarian Tissue Cryopreservation: Effect of Cryoprotective Agents and *In Vitro* Culture on Vitrified-Warmed Ovarian Tissue Survival. *Hum. Reprod.* 29, 720–730. doi:10.1093/humrep/det449

Yu, C., Cvetesic, N., Hisler, V., Gupta, K., Ye, T., Gazdag, E., et al. (2020). TBPL2/TFIIA Complex Establishes the Maternal Transcriptome through Oocyte-specific Promoter Usage. *Nat. Commun.* 11, 6439. doi:10.1038/s41467-020-20239-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 20 June 2022
doi: 10.3389/fgene.2022.844684

Association of Vitamin D Anabolism-Related Gene Polymorphisms and Susceptibility to Uterine Leiomyomas

*Shangdan Xie, Mengying Jiang, Hejing Liu, Fang Xue, Xin Chen and Xueqiong Zhu**

Department of Obstetrics and Gynecology, The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

Background: Uterine leiomyomas (ULs) is the most common gynecological benign tumor in women. Our previous study showed that the phenomenon of vitamin D deficiency existed in patients with ULs. However, the association of vitamin D anabolism-related gene polymorphisms and susceptibility to ULs was unclear.

Methods: Vitamin D anabolism-related gene polymorphisms in 110 patients with ULs and 110 healthy controls were detected by sequencing and the differences of the 92 SNPs were analyzed in the two groups *via* chi-square test. To verify the association between the significantly different SNPs and the risk of ULs, the SNPs were genotyped in another 340 patients and 340 healthy controls. Additionally, an unconditional logistic regression model was conducted to calculate the odds ratio (OR) of ULs occurrence and the 95% confidence interval (CI), adjusting for age and BMI.

Findings: In sequencing samples, there were differences in DHCR7 rs1044482 C > T (p = 0.008) and NADSYN1 rs2276360 G > C (p = 0.025) between patients with ULs and healthy controls. DHCR7 rs1044482 was related to the susceptibility to ULs in validation samples (heterogeneous: adjusted OR = 1.967, p = 0.002; homogenous: adjusted OR = 2.494, p = 0.002; additive: adjusted OR = 1.485, p < 0.041; and dominant: adjusted OR = 2.084, p < 0.001). Stratified analysis further showed that the DHCR7 rs1044482 polymorphisms were associated with ULs risks in women over 40 and with 18.5–25.0 BMI. In contrast to the wild-type CG haplotype vectors, individuals with TC haplotypes had a higher risk of developing ULs.

Interpretation: The vitamin D anabolism-related gene DHCR7 rs1044482 C > T polymorphism was a risk factor of ULs, especially in patients over 40 with 18.5–25.0 BMI, while the relationship between NADSYN1 rs2276360 and ULs risk was not clear.

Keywords: vitamin D, uterine leiomyomas, single nucleotide polymorphisms, Dhcr7, NADSYN1

OPEN ACCESS

Edited by:
Zodwa Dlamini, SAMRC Precision Oncology Research Unit (PORU), South Africa

Reviewed by:
Ines Zidi, Tunis El Manar University, Tunisia
Jaqueline De Azevêdo Silva, Federal Rural University of Pernambuco, Brazil

****Correspondence:***
Xueqiong Zhu
zjwzzxq@163.com

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *28 December 2021*
Accepted: *05 May 2022*
Published: *20 June 2022*

Citation:
Xie S, Jiang M, Liu H, Xue F, Chen X and Zhu X (2022) Association of Vitamin D Anabolism-Related Gene Polymorphisms and Susceptibility to Uterine Leiomyomas. Front. Genet. 13:844684. doi: 10.3389/fgene.2022.844684

INTRODUCTION

Uterine leiomyomas (ULs) are the most common gynecological benign tumor and characterized by the hyperplasia of uterine smooth muscle tissues (Stewart et al., 2017). Although ULs is not a lethal disease for most women cases, some patients will experience heavy menstrual abnormalities, abdominal lumps, increased leucorrhea and abdominal distension (Machado-Lopez et al., 2021). At present, the main treatment for symptomatic ULs is myomectomy or hysterectomy (Kotani et al., 2018). Despite such a high incidence of ULs, the etiology of ULs is not clear, which may explain the frequent recurrence of ULs after surgery. Thus, it is still necessary to further explore the pathogenesis of ULs for developing novel effective therapy.

Vitamin D (VitD) is a lipid-soluble steroid and is an important factor in regulating bone metabolism (Narvaez et al., 2020). In addition, VitD plays a central role in maintaining basic cell functions such as proliferation and differentiation (Samuel and Sitrin, 2008). Therefore, low levels of VitD represents a risk factor for several human diseases, including autoimmune, neurodegenerative, diabetes, and cancer (Marsh and Bulun, 2006; Goltzman et al., 2018; Bivona et al., 2019; Akutsu et al., 2020; Pittas et al., 2020; Ismailova and White, 2021). Of note, various studies have showed that vitamin D (VitD) deficiency may be closely bound up with occurrence of ULs (Singh et al., 2019; Srivastava et al., 2020). Our group has found that the levels of serum VitD3 in patients with ULs were significantly decreased compared with that in healthy controls (Li et al., 2020). Consistently, one study has found the expression level of VitD receptor in ULs was lower than that in nonneoplastic myometrial tissue (Lima et al., 2021). Several clinical trials have revealed that VitD intake in ULs cases with VitD deficiency prevented the growth of fibroids and related symptoms (Al-Hendy et al., 2015; Oskovi Kaplan et al., 2018). Contrarily, a recent randomized clinical trial reveals that VitD supplementation did not decrease the volume of ULs (Arjeh et al., 2020). Therefore, further elucidating the role of VitD deficiency in the development of ULs will provide evidence for the use of oral VitD in treating ULs.

VitD deficiency is a prevalence worldwide and can be caused by environmental factors such as diet, Sun exposure and stress (Dimakopoulos et al., 2019). Recently, accumulating evidence have reported a family cluster to VitD deficiency, which suggest the importance of genetic factors (Wang et al., 2010; Bahrami et al., 2018). This is mainly due to genetic variants involved in VitD metabolic pathways identified by whole-exome sequencing analysis (Alharazy et al., 2021). Therefore, it is proposed that mutation or loss of VitD metabolism-related genes may lead to VitD deficiency and subsequently promote the development of ULs. CYP27A1, GC, RXRA, CYP2R1, DHCR7, NADSYN1, VDR, CYP27B1, METTL1, ASIP, and CYP24A1 are important and functional genes in VitD anabolic pathway (Saponaro et al., 2020). A study shows that upregulation of VitD metabolic enzyme CYP24A1 level probably maintains the low VitD level in leiomyoma (Othman et al., 2018). However, there are few studies focused on the causes of decreased VitD expression in ULs patients (Saponaro et al., 2020). Therefore, the study aims to explore the relationship between single-nucleotide polymorphisms (SNPs) of these genes and susceptibility to ULs, which may provide a direction for exploring the causes of low VitD level in patients with ULs and the different effects of VitD anabolism-related gene polymorphisms on ULs risks by age and body mass index (BMI) stratification.

MATERIALS AND METHODS

Subjects

The case-control study recruited 450 patients with ULs and 450 healthy married controls aged 27–58 years from The Second Affiliated Hospital of Wenzhou Medical University (WMU). The patients were confirmed to have ULs by pelvic ultrasound and the healthy controls experienced physical examination and showed no ULs by pelvic ultrasound. The exclusive criteria: 1) adnexal mass or endometrial polyps shown by pelvic ultrasound; 2) various severe diseases, including malignant tumors, cardiovascular diseases (myocardial infarction, cerebral infarction), endocrine diseases (abnormal parathyroid gland and type 2 diabetes), infectious diseases (tuberculosis), autoimmune disorders (Type 1 diabetes mellitus, systemic lupus erythematosus), hepatic or renal diseases; 3) other low vitD related diseases; 4) using vitD or calcium supplements within 6 months before study enrollment; 5) a past history of myomectomy or hysterectomy. This study was approved by ethics committee of The Second Affiliated Hospital of Wenzhou Medical University. Informed consent for involvement in the study was obtained from all participants. The following data were extracted for each case: age, BMI, white blood cells (WBC), red blood cells (RBC), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total protein, carbamide, uric acid, triglyceride and total cholesterol. The BMI calculation formula was listed: weight (kg)/height $(m)^2$. The stratification criteria for BMI referred to WHO criteria and was as follow: BMI <18.5 was underweight; 18.5 ≤ BMI <25.0 was normal weight; 25.0 ≤ BMI <30.0 was pre-obesity; 30.0 ≤ BMI <40.0 was obesity (Weir and Jan, 2021).

DNA Isolation and Genotyping

TIANamp Blood DNA Kit [(TianGen Biotech, China) was utilized to collect genomic DNA from peripheral blood. The DNA purity and concentration was determined by The AUV absorption spectrophotometer (NanoDrop Technologies Inc., Thermo Fisher, United States)].

The DNA of 110 patients and 110 healthy controls was detected the bases at 92 sites of CYP27A1, GC, RXRA, CYP2R1, DHCR7, NADSYN1, VDR, CYP27B1, METTL1, ASIP, and CYP24A1 by NovaSeq6000 Sequencer (Illumina, United States). The probes for DHCR7 rs1044482 and NADSYN1 rs2276360 were purchased from Thermo Fisher (United States). The Genotyping qPCR PreMix was obtained from TianGen Biotech (China). The DNA of remaining 340 patients and remaining 340 healthy controls was applied in the genotyping assays for the two significant different SNPs between 110 patients and 110 controls. Genotyping analysis

TABLE 1 | The sequencing results of 33 SNPs of DHCR7 and NADSYN1 in 110 ULs and 110healthy controls.

Number	SNP	Position	Alleles	Gene	Case (N = 110)			Control (N = 110)			HWE	*P*
					0/0	0/1	1/1	0/0	0/1	1/1		
1	rs199506852	chr11: 71146468	G > A	DHCR7	109	1	0	109	1	0	0.962	1.000
2	rs781687341	chr11: 71146521	C > T	DHCR7	110	0	0	110	0	0	NA	NA
3	NA	chr11: 71146535	G > T	DHCR7	110	0	0	109	1	0	0.962	1.000
4	rs909217	chr11: 71146577	G > A	DHCR7	29	68	13	45	55	10	0.237	0.073
5	rs544442568	chr11: 71146681	G > A	DHCR7	110	0	0	109	1	0	0.962	1.000
6	rs760241	chr11: 71146691	A > G	DHCR7	1	41	68	6	37	67	0.690	0.151
7	NA	chr11: 71146810	C > G	DHCR7	109	1	0	110	0	0	NA	1.000
8	rs72954276	chr11: 71146837	C > T	DHCR7	110	0	0	109	1	0	0.962	1.000
9	rs75225632	chr11: 71146841	G > A	DHCR7	98	12	0	102	8	0	0.692	0.483
10	rs145901607	chr11: 71146862	G > A	DHCR7	106	4	0	100	9	1	0.146	0.212
11	rs1792268	chr11: 71146952	G > A	DHCR7	1	41	68	6	37	67	0.765	0.151
12	rs143811340	chr11: 71147010	G > A	DHCR7	109	1	0	110	0	0	NA	1.000
13	rs770925697	chr11: 71149986	G > A	DHCR7	110	0	0	110	0	0	NA	NA
14	rs949177	chr11: 71152461	A > G	DHCR7	0	30	80	4	34	72	0.969	0.107
15	NA	chr11: 71153351	C > T	DHCR7	109	1	0	110	0	0	NA	1.000
16	rs1790334	chr11: 71155153	A > G	DHCR7	0	30	80	4	34	72	0.996	0.097
17	rs1044482	chr11: 71155171	C > T	DHCR7	29	69	12	51	51	8	0.321	**0.008**
18	NA	chr11: 71155278	A > G	DHCR7	110	0	0	109	1	0	0.962	1.000
19	rs2276360	chr11: 71169547	G > C	NADSYN1	27	70	13	46	54	10	0.296	**0.025**
20	rs7950441	chr11: 71184678	A > C	NADSYN1	0	0	110	0	0	110	NA	NA
21	rs2276354	chr11: 71185479	T > C	NADSYN1	0	29	81	3	34	73	0.683	0.149
22	rs2186778	chr11: 71185518	T > C	NADSYN1	0	29	81	4	33	73	0.910	0.097
23	rs3819215	chr11: 71185582	G > A	NADSYN1	99	11	0	102	8	0	0.692	0.632
24	NA	chr11: 71188468	T > C	NADSYN1	110	0	0	109	1	0	0.962	1.000
25	rs2276353	chr11: 71189436	T > C	NADSYN1	0	29	81	3	34	73	0.683	0.149
26	rs149812928	chr11: 71189473	C > T	NADSYN1	110	0	0	109	1	0	0.962	1.000
27	rs138969547	chr11: 71191891	G > A	NADSYN1	110	0	0	108	2	0	0.923	0.498
28	rs147585323	chr11: 71193059	G > A	NADSYN1	109	1	0	110	0	0	NA	1.000
29	rs184748544	chr11: 71193951	C > A	NADSYN1	109	1	0	110	0	0	NA	1.000
30	rs765545198	chr11: 71194033	G > A	NADSYN1	109	1	0	110	0	0	NA	1.000
31	rs182956982	chr11: 71202892	G > A	NADSYN1	109	1	0	110	0	0	NA	1.000
32	rs371669981	chr11: 71209469	C > T	NADSYN1	109	1	0	110	0	0	NA	1.000
33	rs12282060	chr11: 71212387	G > A	NADSYN1	110	0	0	109	1	0	0.962	1.000

0/0 was homozygote without mutation; 0/1 was heterozygote with 1 mutation; 1/1 was homozygote with 2 mutations.
HWE, Hardy-Weinberg equilibrium; NA, not applicable.

was conducted by the Taqman real-time polymerase chain reaction (PCR) method using a 7900 HT sequence detector system (Applied Biosystems, United States). The volume of each composition in total reaction system and amplification procedures required in PCR were listed in **Supplementary Table S1**. Besides, to evaluate the accuracy of genotyping outcomes, two positive controls and two negative controls were added in every 384-well plate.

Statistical Analysis

The chi-square test was performed to calculate departure from Hardy-Weinberg equilibrium (HWE) for the all polymorphisms in 110 healthy controls. The measurement data with normal distribution were showed with mean ± standard deviation (SD) and compared by independent-samples t test, while those with abnormal distribution were described with quartile value and the differences were compared by nonparametric test. The relationship strength between the two SNPs (DHCR7 rs1044482 and NADSYN1 rs2276360) and ULs risk was evaluated via an unconditional logistic regression model, computed as crude and adjusted odds ratio (OR) and 95% confidence interval (CI). A p value < 0.05 indicated a statistical difference. Statistical analysis was conducted utilizing SPSS26.0 software.

RESULTS

The Polymorphisms of Vitamin D Related Metabolic Genes in 110 Patients With Uterine Leiomyomas and 110 Healthy Controls

There was no difference in age between 110 patients with ULs (43.74 ± 5.15 years) and 110 healthy controls (43.17 ± 4.48 years, p = 0.334). The sequencing results of 59 loci of CYP27A1, GC, RXRA, CYP2R1, VDR, CYP27B1, METTL1, ASIP, and CYP24A1 in 220 subjects were shown in **Supplementary Table S2** and there was no significant difference in the genotype distributions of 59 SNPs between patients and controls. The sequencing results of 33 SNPs of DHCR7 and NASDYN1 in 220 participants were shown in **Table 1** and there were 11 loci with no mutations in healthy

TABLE 2 | The information of selected variables of validated ULs patients and controls.

Variables	Case (N = 340)			Control (N = 340)			P	
	Medium ±SD			Medium ±SD				
Age	39.90 ± 4.80			39.99 ± 4.38			0.770	
BMI	22.46 ± 2.93			21.93 ± 2.61			0.282	
Variables	**Medium**	**P25**	**P75**	**Medium**	**P25**	**P75**	***Z***	***P***
WBC	5.72	4.85	6.84	5.72	4.81	6.72	–0.005	0.996
RBC	4.37	4.15	4.61	4.40	4.19	4.62	–1.052	0.293
ALT	13.00	11.00	17.00	14.00	11.00	19.00	–1.645	0.100
AST	17.00	15.00	20.00	18.00	15.00	20.00	–0.697	0.486
Total protein	73.50	71.20	76.20	73.70	70.90	76.40	–0.667	0.505
Carbamide	4.40	3.80	5.20	4.50	3.90	5.10	–1.312	0.190
Uric acid	269.00	231.00	313.50	277.00	240.00	310.00	–1.039	0.299
Triglyceride	0.91	0.69	1.30	0.98	0.73	1.33	–0.963	0.335
Total cholesterol	4.63	4.06	5.07	4.62	4.19	5.09	–0.487	0.627

BMI, body mass index; WBC, white blood cell; RBC, red blood cell; ALT, alanine aminotransferase; AST, aspartate aminotransferase.
Normally distributed data is represented by Medium ± SD, non-normally distributed data is represented by quartile, and Z is the statistical value of nonparametric test.

TABLE 3 | Association between the two SNPs and ULs by logistic regression analyses.

Allele	Case (N = 340)	Control (N = 340)	Crude OR (95% CI)	*p*	Adjusted OR (95% CI)[a]	*p*[a]
DHCR7 rs1044482 C > T						
C	349 (51.32%)	402 (59.12%)	References		References	
T	293 (43.09%)	238 (35.00%)	1.418 (1.134–1.773)	**0.002**	1.651 (1.251–2.178)	< **0.001**
NA	38 (5.59%)	40 (5.88%)		NA		NA
NADSYN1 rs2276360 G > C						
G	325 (47.79%)	382 (56.18%)	References		References	
C	253 (37.21%)	218 (32.06%)	1.364 (1.080–1.723)	**0.009**	1.430 (1.073–1.906)	**0.015**
NA	102 (15.00%)	80 (11.76%)		NA		NA
Genotype	Case (N = 340)	Control (N = 340)	Crude OR (95% CI)	*p*	Adjusted OR (95% CI)[a]	*p*[a]
DHCR7 rs1044482 C > T, HWE = 0.106						
CC	92 (27.06%)	133 (39.12%)	References		References	
CT	165 (48.53%)	136 (40.00%)	1.754 (1.237–2.488)	**0.001**	1.967 (1.289–3.001)	**0.002**
TT	64 (18.82%)	51 (15.00%)	1.814 (1.152–2.856)	**0.010**	2.494 (1.389–4.477)	**0.002**
NA	19 (5.59%)	20 (5.88%)		NA		NA
Additive			1.431 (1.048–1.954)	**0.024**	1.485 (1.016–2.171)	**0.041**
Dominant	229 (67.35%)	187 (55.00%)	1.770 (1.275–2.459)	**0.001**	2.084 (1.396–3.110)	< **0.001**
Recessive	257 (75.59%)	269 (78.12%)	1.313 (0.875–1.971)	0.188	1.689 (0.994–2.870)	0.053
NADSYN1 rs2276360 G > C, HWE = 0.178						
GG	95 (27.94%)	127 (37.35%)	References		References	
GC	135 (39.71%)	128 (37.65%)	1.410 (0.984–2.020)	0.061	1.336 (0.864–2.067)	0.193
CC	59 (17.35%)	45 (13.24%)	1.753 (1.095–2.805)	**0.019**	2.020 (1.121–3.640)	**0.019**
NA	51 (15.00%)	40 (11.76%)		NA		NA
Additive			1.178 (0.851–1.631)	0.324	1.079 (0.726–1.606)	0.706
Dominant	194 (57.06%)	173 (50.88%)	1.499 (1.072–2.097)	**0.018**	1.494 (0.993–2.247)	0.054
Recessive	230 (67.65%)	255 (75.00%)	1.454 (0.949–2.228)	0.086	1.727 (1.008–2.960)	**0.047**
Combined effect of risk genotypes[b]						
0	78 (22.94%)	111 (32.65%)	References		References	
1	19 (5.59%)	25 (7.35%)	1.082 (0.557–2.099)	0.817	1.198 (0.529–2.716)	0.665
2	181 (53.24%)	151 (44.41%)	1.706 (1.189–2.448)	**0.004**	1.856 (1.197–2.877)	**0.006**
NA	62 (18.24%)	53 (15.59%)		NA		NA

[a]*Adjusted for age and BMI.*
[b]*Risk genotype was with DHCR7 rs1044482 CT/TT, and NADSYN1 rs2276360 GC/CC.*
The results were in bold if p *value <0.05.*
OR, odd ratio; CI, confidence interval; HWE, Hardy-Weinberg equilibrium; NA, not applicable.

TABLE 4 | Stratification analysis of risk alleles and genotypes with ULs susceptibility.

Characteristics	rs1044482 (Case/Control)		OR (95%CI)	*p*	rs2276360 (Case/Control)		OR (95%CI)	*p*
	C	T			G	C		
Age (year)								
<40	190/203	140/113	1.324 (0.964–1.818)	0.083	175/190	117/106	1.198 (0.859–1.673)	0.288
≥40	159/199	153/125	1.532 (1.118–2.099)	**0.008**	150/192	136/112	1.554 (1.119–2.160)	**0.009**
BMI								
<18.5	15/27	11/7	2.829 (0.906–8.832)	0.073	14/23	10/9	1.825 (0.596–5.590)	0.292
18.5 ≤ BMI <25.0	194/220	154/110	1.588 (1.163–2.168)	**0.004**	181/207	133/105	1.449 (1.047–2.004)	**0.025**
25.0 ≤ BMI <30.0	41/31	37/21	1.332 (0.655–2.710)	0.429	36/27	30/23	0.978 (0.468–2.045)	0.953
30.0 ≤ BMI <40.0	1/3	3/221	9.000 (0.367–220.927)	0.178	0/3	0/1	NA	NA
Characteristics	**rs1044482 (Case/Control)**		**OR (95% CI)**	***p***	**rs2276360 (Case/Control)**		**OR (95% CI)**	***p***
	CC	**CT + TT**			**GG**	**GC + CC**		
Age (year)								
<40	56/67	109/91	1.433 (0.913–2.250)	0.118	54/63	92/85	1.226 (0.684–2.196)	0.494
≥40	36/66	120/96	2.292 (1.408–3.729)	**0.001**	41/64	102/88	1.809 (1.114–2.938)	**0.017**
BMI								
<18.5	6/10	7/7	1.667 (0.388–7.153)	0.492	6/8	6/8	1.000 (0.224–4.468)	1.000
18.5 ≤ BMI <25.0	51/78	123/87	2.162 (1.383–3.382)	**0.001**	54/72	103/84	1.635 (1.037–2.578)	**0.034**
25.0 ≤ BMI <30.0	10/8	29/18	1.289 (0.429–3.872)	0.651	10/6	23/19	0.726 (0.223–2.365)	0.595
30.0 ≤ BMI <40.0	0/1	2/1	NA	1.000	0/1	0/1	NA	NA

The results were in bold if p *value <0.05.*
OR, odd ratio; CI, confidence interval.

controls, so the relationship between the 11 SNPs and ULs occurrence was not analyzed. The genotype distributions of the remaining 22 SNPs in ULs-free controls followed HWE ($p > 0.05$). There was no significant difference in the genotype distribution of 20 loci between leiomyoma patients and controls, while there were differences in the genotype distributions of DHCR7 rs1044482 ($p = 0.008$) and NADSYN1 rs2276360 ($p = 0.025$) between the two groups. Therefore, this study enlarged the sample content to investigate the relationship between the polymorphisms of DHCR7 rs1044482 and NADSYN1 rs2276360 and the risk of ULs.

Association Between DHCR7 rs1044482 and NADYSYN1 rs2276360 and Uterine Leiomyomas Risk

340 patients and 340 healthy controls were recruited to verify the link strength between mutations of DHCR7 rs1044482 and NADSYN1 rs2276360 and occurrence of ULs. The demographic characteristics and some common indicators of ULs patients and controls were listed in **Table 2**. There was no difference in age, BMI, WBC, RBC, ALT, AST, total protein, carbamide, uric acid, triglyceride and total cholesterol between two groups ($p > 0.05$). As shown in **Table 3**, The values of HWE of DHCR7 rs1044482 and NADSYN1 rs2276360 demonstrated genetic balance in selected population. It was suggested that DHCR7 rs1044482 C > T was closely related to the occurrence of ULs (heterogeneous: adjusted OR = 1.967, 95% CI = 1.289–3.001, $p = 0.002$; homogenous: adjusted OR = 2.494, 95% CI = 1.389–4.477, $p = 0.002$; additive: adjusted OR = 1.485, 95% CI = 1.016–2.171, $p < 0.041$; and dominant: adjusted OR = 2.084, 95% CI = 1.396–3.110, $p < 0.001$). In addition, NADSYN1 rs2276360 G > C might be a risk factor of ULs (homogenous: adjusted OR = 2.020, 95% CI = 1.121–3.640, $p = 0.019$; additive: adjusted OR = 1.079, 95% CI = 0.726–1.606, $p = 0.706$; and recessive: adjusted OR = 1.727, 95% CI = 1.008–2.960, $p = 0.047$). When the two SNPs were analyzed in combination, the vectors of two risk genotypes could raise the susceptibility to ULs (adjusted OR: 1.856, 95% CI = 1.197–2.877, $p = 0.030$).

Stratified Analysis and Haplotype Analysis

As shown in **Table 4**, the relationship of DHCR7 rs1044482 C > T and NADSYN1 rs2276360 G > C and ULs susceptibility was further studied via stratified analysis. The DHCR7 rs1044482 CT/TT genotypes ($p = 0.001$) and the NADSYN1 rs2276360 GC/CC genotypes ($p = 0.017$) could increase the risk of ULs among the women over 40 years old, respectively. Besides, among the participants with 18.5–25.0 BMI, the susceptibility to ULs was raised in the women with DHCR7 rs1044482 CT/TT genotypes ($p = 0.001$) or NADSYN1 rs2276360 GC/CC genotypes ($p = 0.034$). Furthermore, the haplotypes of the two SNPs were investigated in **Table 5**. In comparison of the reference haplotype CG, CC (adjusted OR = 1.640, 95% CI =

TABLE 5 | The frequency of inferred haplotypes of the two SNPs based on observed genotypes and their association with the risk of ULs.

rs1044482	rs2276360	Case (N = 340)	Control (N = 340)	Crude OR (95% CI)	*p*	Adjusted OR (95% CI)[a]	*p*[a]
C	G	231 (41.25%)	301 (50.50%)	References		References	
C	C	80 (14.29%)	77 (12.92%)	1.456 (1.128–1.880)	**0.004**	1.640 (1.193–2.256)	**0.002**
T	G	82 (14.64%)	76 (12.75%)	1.401 (1.085–1.810)	**0.010**	1.374 (1.009–1.872)	**0.012**
T	C	167 (29.82%)	142 (23.83%)	1.587 (1.297–1.942)	< **0.001**	1.851 (1.441–2.379)	< **0.001**

[a]*Adjusted for age and BMI.*
The results were in bold if p *value <0.05.*
OR, odd ratio; CI, confidence interval.

1.193–2.256, $p = 0.002$, TG (adjusted OR = 1.374, 95% CI = 1.009–1.872, $p = 0.012$) and TC (adjusted OR = 1.851, 95% CI = 1.441–2.379, $p < 0.001$) all increased ULs risk, respectively.

DISCUSSION

ULs are a kind of disease with high incidence and unclear causes, which affects the normal life of women (McWilliams and Chennathukuzhi, 2017). Various genes may have been mutated in ULs via genome sequencing (Hodge et al., 2014; Mehine et al., 2014; Ajabnoor et al., 2018; Bray et al., 2019). Hence, there are more and more studies focused on SNPs of the genes in ULs, including polymorphisms of MED12, folk1, FANCA, genes for age at menarche (Yatsenko et al., 2017; Gulec Yilmaz et al., 2018; Ha et al., 2020; Ponomarenko et al., 2020). Many studies found that VitD deficiency might increase ULs risk (Davari Tanha et al., 2021; Islam et al., 2021; Vergara et al., 2021). Our previous study also showed that the levels of VitD in patients with ULs were lower than those in healthy controls (Li et al., 2020). It is speculated that there may be some mutations occurred in VitD anabolism-related genes and then influence the susceptibility to ULs. Therefore, this study investigated the relationship between polymorphisms of vitamin D anabolism related genes and ULs risks.

DHCR7 encodes 7-dehydrocholesterol reductase and the enzyme is involved in the conversion procedure of 7-dehydrocholesterol to cholesterol (Prabhu et al., 2016). 7-dehydrocholesterol is a precursor of VitD and can be changed into VitD3 under sunlight or ultraviolet radiation on the skin. NADSYN1 catalyzes the synthesis of NAD, a significant cofactor in multiple redox reactions, and may indirectly take part in the formation of VitD (Prabhu et al., 2016). The polymorphisms of DHCR7/NADSYN1 participate in susceptibility to many diseases, including deficiency of VitD, acute coronary syndrome, Alzheimer's disease and ULs (Lu et al., 2012; Wise et al., 2014; Elbehairy et al., 2021; Liu et al., 2021). Three variants in DHCR7/NADSYN1 (rs11606033, rs3829251 and rs1790349) loci were tightly correlated to low levels of serum VitD (Lu et al., 2012; Elbehairy et al., 2021). Notably, it was reported that NADSYN1 rs2276360 G > C was related to serum 25(OH)D3 deficiency (Elbehairy et al., 2021). However, DHCR7 rs1044482 has not been found to be associated with VitD deficiency. At present, there are few studies on DHCR7/NADSYN1 polymorphisms in ULs. Liu et al. demonstrated that rs12800438 near DHCR7 was linked closely to the risk of ULs (Wise et al., 2014). For rs1044482, the allele frequency of T was 0.324 in the Asian population of NCBI's dbSNP database, 0.350 in the control group and 0.431 in the ULs group in this study. For rs2276360, the allele frequency of C was 0.341 in the Asian population in the dbSNP database of NCBI, 0.321 in the control group and 0.372 in the ULs group in this study. DHCR7 rs1044482 was related to the occurrence of ULs and haploid analysis of DHCR7 rs1044482 and NADSYN1 rs2276360 showed that the combination of the two SNPs was significantly associated with susceptibility to ULs.

A study found that the incidence of ULs reached the peak in women aged 40–45 years (Yu et al., 2018). It was demonstrated that the incidence of ULs gradually increased with age (Pavone et al., 2018). Besides, aging is correlated with the reduction in intake, synthesis and function of VitD3 (Biesalski, 2021). In our study, although there was no difference in age between ULs group and control group, both of the two SNPs were related closely to susceptibility to ULs in older women (age ≥40 years). Therefore, age and site mutation may affect the level or function of vitamin D and thus induce the occurrence of hysteromyoma. In addition, many studies found that obesity is also a risk factor of ULs and women with higher BMI more tended to develop ULs (Giri et al., 2017; Qin et al., 2021). Compared with women with normal BMI, obese women produced more estrogen, making them prone to ULs (Cleary and Grossmann, 2009). However, DHCR7 rs1044482 was only statistically associated with the occurrence of ULs in women with normal BMI, not in women with high BMI in our study. The reason for the phenomenon might partly be the small number of women with high BMI. The proportion of participants in validation samples with >25.0 BMI was only 10.38%.

There are still some expectations in our study. In the future, this study will continue to recruit a large number of patients with ULs and normal women, and test the VitD level of each participant while analyzing the differences in DHCR7 rs1044482 and NADSYN1 rs2276360 between the two groups. Then, the source of participants will be enriched and more subjects will be recruited from multiple centers. Finally, potential mechanisms of DHCR7 rs1044482 increasing ULs risk should also be addressed in the following studies.

CONCLUSION

In a word, the study suggested that DHCR7 rs1044482 C > T and NADSYN1 rs2276360 G > C might be related to the susceptibility to uterine leiomyomas in the Chinese population, especially in patients over 40 with 18.5–25.0 BMI.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: Bioproject, accession number PRJNA807129; SRA database, accessions SRR18114348-SRR18114567.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Research Ethics Committee of the Second Affiliated Hospital of Wenzhou Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SX and XZ participated in the design of the study, statistical analysis, and manuscript drafting. SX, MJ, HL, FX, and XC gathered the samples, performed experiments and revised manuscript. XZ instructed and supervised the whole study and revised manuscript. All authors read and approved the final manuscript.

FUNDING

This work was funded by the clinical trial centre of the Second Affiliated Hospital of Wenzhou Medical University (No: SAHoWMU-CR2017-07-101).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.844684/full#supplementary-material

REFERENCES

Ajabnoor, G. M. A., Mohammed, N. A., Banaganapalli, B., Abdullah, L. S., Bondagji, O. N., Mansouri, N., et al. (2018). Expanded Somatic Mutation Spectrum of MED12 Gene in Uterine Leiomyomas of Saudi Arabian Women. *Front. Genet.* 9, 552. doi:10.3389/fgene.2018.00552

Akutsu, T., Kitamura, H., Himeiwa, S., Kitada, S., Akasu, T., and Urashima, M. (2020). Vitamin D and Cancer Survival: Does Vitamin D Supplementation Improve the Survival of Patients with Cancer? *Curr. Oncol. Rep.* 22 (6), 62. doi:10.1007/s11912-020-00929-4

Alharazy, S., Naseer, M. I., Alissa, E., Robertson, M. D., Lanham-New, S., and Chaudhary, A. G. (2021). Whole-Exome Sequencing for Identification of Genetic Variants Involved in Vitamin D Metabolic Pathways in Families with Vitamin D Deficiency in Saudi Arabia. *Front. Genet.* 12, 677780. doi:10.3389/fgene.2021.677780

Al-Hendy, A., Diamond, M. P., El-Sohemy, A., and Halder, S. K. (2015). 1,25-dihydroxyvitamin D3 Regulates Expression of Sex Steroid Receptors in Human Uterine Fibroid Cells. *J. Clin. Endocrinol. Metabolism* 100 (4), E572–E582. doi:10.1210/jc.2014-4011

Arjeh, S., Darsareh, F., Asl, Z. A., and Azizi Kutenaei, M. (2020). Effect of Oral Consumption of Vitamin D on Uterine Fibroids: A Randomized Clinical Trial. *Complementary Ther. Clin. Pract.* 39, 101159. doi:10.1016/j.ctcp.2020.101159

Bahrami, A., Sadeghnia, H. R., Tabatabaeizadeh, S. A., Bahrami-Taghanaki, H., Behboodi, N., Esmaeili, H., et al. (2018). Genetic and Epigenetic Factors Influencing Vitamin D Status. *J. Cell Physiol.* 233 (5), 4033–4043. doi:10.1002/jcp.26216

Biesalski, H. K. (2021). Obesity, Vitamin D Deficiency and Old Age a Serious Combination with Respect to Coronavirus Disease-2019 Severity and Outcome. *Curr. Opin. Clin. Nutr. Metab. Care* 24 (1), 18–24. doi:10.1097/MCO.0000000000000700

Bivona, G., Gambino, C. M., Iacolino, G., and Ciaccio, M. (2019). Vitamin D and the Nervous System. *Neurol. Res.* 41 (9), 827–835. doi:10.1080/01616412.2019.1622872

Bray, M. J., Davis, L. K., Torstenson, E. S., Jones, S. H., Edwards, T. L., and Velez Edwards, D. R. (2019). Estimating Uterine Fibroid SNP-Based Heritability in European American Women with Imaging-Confirmed Fibroids. *Hum. Hered.* 84 (2), 73–81. doi:10.1159/000501335

Cleary, M. P., and Grossmann, M. E. (2009). Obesity and Breast Cancer: The Estrogen Connection. *Endocrinology* 150 (6), 2537–2542. doi:10.1210/en.2009-0070

Davari Tanha, F., Feizabad, E., Vasheghani Farahani, M., Amuzegar, H., Moradi, B., and Samimi Sadeh, S. (2021). The Effect of Vitamin D Deficiency on Overgrowth of Uterine Fibroids: A Blinded Randomized Clinical Trial. *Int. J. Fertil. Steril.* 15 (2), 95–100. doi:10.22074/IJFS.2020.134567

Dimakopoulos, I., Magriplis, E., Mitsopoulou, A.-V., Karageorgou, D., Bakogianni, I., Micha, R., et al. (2019). Association of Serum Vitamin D Status with Dietary Intake and Sun Exposure in Adults. *Clin. Nutr. ESPEN* 34, 23–31. doi:10.1016/j.clnesp.2019.09.008

Elbehairy, M. M., Abdelnasser, H. Y., Hanafi, R. S., Hassanein, S. I., and Gad, M. Z. (2021). An Intronic DHCR7 Genetic Polymorphism Associates with Vitamin D Serum Level and Incidence of Acute Coronary Syndrome. *Steroids* 169, 108825. doi:10.1016/j.steroids.2021.108825

Giri, A., Edwards, T. L., Hartmann, K. E., Torstenson, E. S., Wellons, M., Schreiner, P. J., et al. (2017). African Genetic Ancestry Interacts with Body Mass Index to Modify Risk for Uterine Fibroids. *PLoS Genet.* 13 (7), e1006871. doi:10.1371/journal.pgen.1006871

Goltzman, D., Mannstadt, M., and Marcocci, C. (2018). Physiology of the Calcium-Parathyroid Hormone-Vitamin D Axis. *Front. Horm. Res.* 50, 1–13. doi:10.1159/000486060

Güleç Yılmaz, S., Gül, T., Attar, R., Yıldırım, G., and İşbir, T. (2018). Association between Fok1 Polymorphism of Vitamin D Receptor Gene with Uterine Leiomyoma in Turkish Populations. *J. Turk. Ger. Gynecol. Assoc.* 19 (3), 128–131. doi:10.4274/jtgga.2018.0002

Ha, E., Lee, S., Lee, S. M., Jung, J., Chung, H., Choi, E., et al. (2020). FANCA Polymorphism is Associated with the Rate of Proliferation in Uterine Leiomyoma in Korea. *J. Pers. Med.* 10 (4), 228. doi:10.3390/jpm10040228

Hodge, J. C., Pearce, K. E., Clayton, A. C., Taran, F. A., and Stewart, E. A. (2014). Uterine Cellular Leiomyomata with Chromosome 1p Deletions Represent a Distinct Entity. *Am. J. Obstetrics Gynecol.* 210 (6), e1–572. doi:10.1016/j.ajog.2014.01.011

Islam, M. S., Akhtar, M. M., and Segars, J. H. (2021). Vitamin D Deficiency and Uterine Fibroids: An Opportunity for Treatment or Prevention? *Fertil. Steril.* 115 (5), 1175–1176. doi:10.1016/j.fertnstert.2021.02.040

Ismailova, A., and White, J. H. (2021). Vitamin D, Infections and Immunity. *Rev. Endocr. Metab. Disord.* 23, 265–277. doi:10.1007/s11154-021-09679-5

Kotani, Y., Tobiume, T., Fujishima, R., Shigeta, M., Takaya, H., Nakai, H., et al. (2018). Recurrence of Uterine Myoma after Myomectomy: Open Myomectomy versus Laparoscopic Myomectomy. *J. Obstet. Gynaecol. Res.* 44 (2), 298–302. doi:10.1111/jog.13519

Li, S., Chen, B., Sheng, B., Wang, J., and Zhu, X. (2020). The Associations between Serum Vitamin D, Calcium and Uterine Fibroids in Chinese Women: A Case-Controlled Study. *J. Int. Med. Res.* 48 (5), 030006052092349. doi:10.1177/0300060520923492

Lima, M. S. O., da Silva, B. B., de Medeiros, M. L., Dos Santos, A. R., do Nascimento Brazil, E. D., Filho, W. M. N. E., et al. (2021). Evaluation of Vitamin D Receptor Expression in Uterine Leiomyoma and Nonneoplastic Myometrial Tissue: A

Cross-sectional Controlled Study. *Reprod. Biol. Endocrinol.* 19 (1), 67. doi:10.1186/s12958-021-00752-x

Liu, X., Wu, P., Shen, L., Jiao, B., Liao, X., Wang, H., et al. (2021). DHCR7 Rs12785878 T>C Polymorphism is Associated with an Increased Risk of Early Onset of Alzheimer's Disease in Chinese Population. *Front. Genet.* 12, 583695. doi:10.3389/fgene.2021.583695

Lu, L., Sheng, H., Li, H., Gan, W., Liu, C., Zhu, J., et al. (2012). Associations between Common Variants in GC and DHCR7/NADSYN1 and Vitamin D Concentration in Chinese Hans. *Hum. Genet.* 131 (3), 505–512. doi:10.1007/s00439-011-1099-1

Machado-Lopez, A., Simón, C., and Mas, A. (2021). Molecular and Cellular Insights into the Development of Uterine Fibroids. *Int. J. Mol. Sci.* 22 (16), 8483. doi:10.3390/ijms22168483

Marsh, E. E., and Bulun, S. E. (2006). Steroid Hormones and Leiomyomas. *Obstetrics Gynecol. Clin. N. Am.* 33 (1), 59–67. doi:10.1016/j.ogc.2005.12.001

McWilliams, M., and Chennathukuzhi, V. (2017). Recent Advances in Uterine Fibroid Etiology. *Semin. Reprod. Med.* 35 (2), 181–189. doi:10.1055/s-0037-1599090

Mehine, M., Mäkinen, N., Heinonen, H.-R., Aaltonen, L. A., and Vahteristo, P. (2014). Genomics of Uterine Leiomyomas: Insights from High-Throughput Sequencing. *Fertil. Steril.* 102 (3), 621–629. doi:10.1016/j.fertnstert.2014.06.050

Narvaez, J., Maldonado, G., Guerrero, R., Messina, O. D., and Ríos, C. (2020). Vitamin D Megadose: Definition, Efficacy in Bone Metabolism, Risk of Falls and Fractures. *Open Access Rheumatol.* 12, 105–115. doi:10.2147/OARRR.S252245

Oskovi Kaplan, Z. A., Taşçi, Y., Topçu, H. O., and Erkaya, S. (2018). 25-Hydroxy Vitamin D Levels in Premenopausal Turkish Women with Uterine Leiomyoma. *Gynecol. Endocrinol.* 34 (3), 261–264. doi:10.1080/09513590.2017.1391774

Othman, E. R., Ahmed, E., Sayed, A. A., Hussein, M., AbdelaalII, Fetih, A. N., et al. (2018). Human Uterine Leiomyoma Contains Low Levels of 1, 25 Dihdroxyvitamin D3, and Shows Dysregulated Expression of Vitamin D Metabolizing Enzymes. *Eur. J. Obstetrics Gynecol. Reproduct. Biol.* 229, 117–122. doi:10.1016/j.ejogrb.2018.08.018

Pavone, D., Clemenza, S., Sorbi, F., Fambrini, M., and Petraglia, F. (2018). Epidemiology and Risk Factors of Uterine Fibroids. *Best Pract. Res. Clin. Obstetrics Gynaecol.* 46, 3–11. doi:10.1016/j.bpobgyn.2017.09.004

Pittas, A. G., Jorde, R., Kawahara, T., and Dawson-Hughes, B. (2020). Vitamin D Supplementation for Prevention of Type 2 Diabetes Mellitus: To D or Not to D? *J. Clin. Endocrinol. Metab.* 105 (12), 3721–3733. doi:10.1210/clinem/dgaa594

Ponomarenko, I., Reshetnikov, E., Polonikov, A., Verzilina, I., Sorokina, I., Yermachenko, A., et al. (2020). Candidate Genes for Age at Menarche are Associated with Uterine Leiomyoma. *Front. Genet.* 11, 512940. doi:10.3389/fgene.2020.512940

Prabhu, A. V., Luu, W., Li, D., Sharpe, L. J., and Brown, A. J. (2016). DHCR7: A Vital Enzyme Switch between Cholesterol and Vitamin D Production. *Prog. Lipid Res.* 64, 138–151. doi:10.1016/j.plipres.2016.09.003

Qin, H., Lin, Z., Vásquez, E., Luan, X., Guo, F., and Xu, L. (2021). Association between Obesity and the Risk of Uterine Fibroids: A Systematic Review and Meta-Analysis. *J. Epidemiol. Community Health* 75 (2), jech–2019. doi:10.1136/jech-2019-213364

Samuel, S., and Sitrin, M. D. (2008). Vitamin D's Role in Cell Proliferation and Differentiation. *Nutr. Rev.* 66 (10 Suppl. 2), S116–S124. doi:10.1111/j.1753-4887.2008.00094.x

Saponaro, F., Saba, A., and Zucchi, R. (2020). An Update on Vitamin D Metabolism. *Int. J. Mol. Sci.* 21 (18), 6573. doi:10.3390/ijms21186573

Singh, V., Barik, A., and Imam, N. (2019). Vitamin D3 Level in Women with Uterine Fibroid: An Observational Study in Eastern Indian Population. *J. Obstet. Gynecol. India* 69 (2), 161–165. doi:10.1007/s13224-018-1195-4

Srivastava, P., Gupta, H. P., Singhi, S., Khanduri, S., and Rathore, B. (2020). Evaluation of 25-hydroxy Vitamin D3 Levels in Patients with a Fibroid Uterus. *J. Obstetrics Gynaecol.* 40 (5), 710–714. doi:10.1080/01443615.2019.1654986

Stewart, E., Cookson, C., Gandolfo, R., and Schulze-Rath, R. (2017). Epidemiology of Uterine Fibroids: A Systematic Review. *BJOG Int. J. Obstet. Gy* 124 (10), 1501–1512. doi:10.1111/1471-0528.14640

Vergara, D., Catherino, W. H., Trojano, G., and Tinelli, A. (2021). Vitamin D: Mechanism of Action and Biological Effects in Uterine Fibroids. *Nutrients* 13 (2), 597. doi:10.3390/nu13020597

Wang, T. J., Zhang, F., Richards, J. B., Kestenbaum, B., van Meurs, J. B., Berry, D., et al. (2010). Common Genetic Determinants of Vitamin D Insufficiency: A Genome-wide Association Study. *Lancet* 376 (9736), 180–188. doi:10.1016/S0140-6736(10)60588-0

Weir, C. B., and Jan, A. (2021). "BMI Classification Percentile and Cut Off Points," in *StatPearls* (Treasure Island, FL: StatPearls).

Wise, L. A., Ruiz-Narváez, E. A., Haddad, S. A., Rosenberg, L., and Palmer, J. R. (2014). Polymorphisms in Vitamin D-Related Genes and Risk of Uterine Leiomyomata. *Fertil. Steril.* 102 (2), 503–510. doi:10.1016/j.fertnstert.2014.04.037

Yatsenko, S. A., Mittal, P., Wood-Trageser, M. A., Jones, M. W., Surti, U., Edwards, R. P., et al. (2017). Highly Heterogeneous Genomic Landscape of Uterine Leiomyomas by Whole Exome Sequencing and Genome-wide Arrays. *Fertil. Steril.* 107 (2), 457–466. doi:10.1016/j.fertnstert.2016.10.035

Yu, O., Scholes, D., Schulze-Rath, R., Grafton, J., Hansen, K., and Reed, S. D. (2018). A US Population-Based Study of Uterine Fibroid Diagnosis Incidence, Trends, and Prevalence: 2005 through 2014. *Am. J. Obstetrics Gynecol.* 219 (6), e1 e591–591 e598. doi:10.1016/j.ajog.2018.09.039-

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 28 June 2022
doi: 10.3389/fgene.2022.770569

Identification of Alternative Splicing-Related Genes CYB561 and FOLH1 in the Tumor-Immune Microenvironment for Endometrial Cancer Based on TCGA Data Analysis

Dan Sun, Aiqian Zhang, Bingsi Gao, Lingxiao Zou, Huan Huang, Xingping Zhao and Dabao Xu**

Department of Gynecology, Third Xiangya Hospital of Central South University, Changsha, China

OPEN ACCESS

Edited by:
Zodwa Dlamini, SAMRC Precision Oncology Research Unit (PORU), South Africa

Reviewed by:
Carine Le Goff, Institut National de la Santé et de la Recherche Médicale (INSERM), France
Camelia Alexandra Coada, University of Bologna, Italy

***Correspondence:**
Xingping Zhao
zxp8846@126.com
Dabao Xu
dabaoxu2022@163.com

Specialty section:
This article was submitted to Genetics of Common and Rare Diseases, a section of the journal Frontiers in Genetics

Received: *04 September 2021*
Accepted: *13 June 2022*
Published: *28 June 2022*

Citation:
Sun D, Zhang A, Gao B, Zou L, Huang H, Zhao X and Xu D (2022) Identification of Alternative Splicing-Related Genes CYB561 and FOLH1 in the Tumor-Immune Microenvironment for Endometrial Cancer Based on TCGA Data Analysis. Front. Genet. 13:770569. doi: 10.3389/fgene.2022.770569

Background: Advanced and recurrent endometrial cancer EC remains controversial. Immunotherapy will play a landmark role in cancer treatment, and alternative splicing (AS) of messenger RNA (mRNA) may offer the potential of a broadened target space.

Methods: We downloaded the clinical information and mRNA expression profiles from The Cancer Genome Atlas (TCGA) database. Hub genes were extracted from 11 AS-related genes to analyze the correlation between clinical parameters and the tumor-immune microenvironment. We also analyzed the correlations between the copy numbers, gene expressions of hub genes, and immune cells. The correlation between the risk score and the six most important checkpoint genes was also investigated. The ESTIMATE algorithm was finally performed on each EC sample based on the high- and low-risk groups.

Results: The risk score was a reliable and stable independent risk predictor in the Uterine Corpus Endometrial Carcinoma (UCEC) cohort. CYB561|42921|AP and FOLH1|15817|ES were extracted. The expression of CYB561 and FOLH1 decreased gradually with the increased grade and International Federation of Gynecology and Obstetrics (FIGO) stage ($p < 0.05$). Gene copy number changes in CYB561 and FOLH1 led to the deletion number of myeloid DC cells and T cell $CD8^+$. Low expression of both CYB561 and FOLH1 was associated with poor prognosis ($p < 0.001$). The checkpoint genes, CTLA-4 and PDCD1, exhibited a negative correlation with the risk score of AS in UCEC.

Conclusion: AS-related gene signatures were related to the immune-tumor microenvironment and prognosis. These outcomes were significant for studying EC's immune-related mechanisms and exploring novel prognostic predictors and precise therapy methods.

Keywords: alternative splicing, tumor-immune microenvironment, endometrial cancer, CYB561, FOLH1

INTRODUCTION

Uterine corpus endometrial carcinoma (UCEC) is one of the three major malignancies of the female reproductive system (Bray et al., 2018). Although most patients are diagnosed at an early stage and have a good prognosis, there are still some patients who are at an advanced stage at first diagnosis or have recurrence and metastasis after treatment, with a 5-year survival rate of only 20–26% (Morice et al., 2016). Therefore, further studies on therapeutic monitoring and prognostic assessment of UCEC are crucial for both clinicians and patients. In 2013, tumor immunotherapy was regarded as an important scientific breakthrough and suggested that immunotherapy will play a landmark role in the field of cancer treatment (Couzin-Frankel, 2013). In recent years, immune checkpoint inhibitors have made breakthrough progress and have been written into the treatment guidelines for endometrial cancer (EC).

Alternative splicing (AS) is a universal mechanism to produce mRNA isomers using a limited set of genes, resulting in structurally and functionally different protein isoforms, modifying more than 95% of human genes (Gilbert, 1978; Buratti et al., 2006). Studies have shown that aberrant AS is closely associated with the occurrence, development, metastasis, and drug resistance of various cancers (de Necochea-Campion et al., 2016; Climente-González et al., 2017; Wang and Lee, 2018; Bonnal et al., 2020). AS has also become a hot topic in tumor immunotherapy and attracted the attention of researchers. Several forms of mRNA processing are dysregulated in cancer and offer promise concerning immunotherapy target expansion (Venables, 2004; Baralle and Giudice, 2017; Kahles et al., 2018). Although significant progress has been achieved in expanding the immunotherapy target space using tumor-specific mRNA processing events, much work is still needed (Frankiw et al., 2019).

What's more, splice factors might play a vital carcinogenic role in EC (Dou et al., 2020a; Li et al., 2020; Popli et al., 2020). By analyzing the whole genome of AS events in EC, studies have found several candidate splicing factors that may become therapeutic targets and predict patients' prognosis by constructing gene signatures (Wang C. et al., 2019; Wang Q. et al., 2019), which further demonstrates the importance of AS events in EC. A study found that with the increase of the ESTIMATE score and the infiltration of immune cells in UCEC patients, the prognosis would be better (Liu et al., 2021); however, further discussion was lacking.

Given the importance of immunotherapy in UCEC, characterization of immune infiltrating features is essential for further understanding the oncogenesis of UCEC and the development of a novel prognostic signature and therapeutic response classifier. In the present study, whole-genome analysis and prognostic model construction were firstly used to determine prognosis-related genes involved in the AS prognostic model. The characteristics of two AS-related genes in the tumor-immune microenvironment were then analyzed. Finally, the correlation between the six most important immune checkpoint genes and the risk score was also investigated. Our research provides a more comprehensive insight into precise immunotherapy for UCEC.

TABLE 1 | The key demographic, clinical, and pathological characteristics of the 524 patients with UCEC.

Variables	Count	Percentage (%)
Age (mean ± SD)	63.88 ± 11.20	
Follow-up (mean ± SD) (y)	3.05 ± 2.47	
Status		
Alive	436	83.21
Dead	88	16.79
Histological type		
Adenomas and adenocarcinomas	390	74.43
Cystic, mucinous, and serous neoplasms	134	25.57
FIGO stage		
I	330	62.98
II	47	8.97
III	119	22.71
IV	28	5.34
Grade		
G1	93	17.75
G2	118	22.52
G3	313	59.73
Race		
White	361	68.89
Black or African American	104	19.85
Asian	19	3.63
Other	11	2.10
Not reported	29	5.53

UCEC: uterine corpus endometrial carcinoma; FIGO, international federation of gynecology and obstetrics.

METHODS

Data Acquisition and Curation Processing

We downloaded the mRNA expression profiles and corresponding clinical data of the UCEC cohort from the TCGA database (June 2021, https://portal.gdc.cancer.gov/). The AS event data for UCEC were obtained from the https://bioinformatics.mdanderson.org/TCGASpliceSeq/(Ryan et al., 2016). Since these data are publically available, there was no requirement for approval by an ethics committee. We fully assessed the availability of clinical information. In our research, a few patients were excluded because they met the following criteria: 1) Epithelial neoplasm disease type, nos in TCGA; and 2) incomplete clinical data (e.g., age, grade, FIGO stage, and survival data). The percent spliced in (PSI) value can be used to quantify each AS event, which is the ratio of normalized reads indicating the presence of a transcript element versus the total normalized reads for that event, with a rating from 0 to 1. PSI = splice in/splice in+splice out. We screened the AS data for PSI value >0.75, representing the association between gene expression and AS events. We then merged the gene expression and clinical profiles using Perl (v5.30.0, https://www.perl.org/), establishing genomics and clinical databases for further research. A total of 524 patients with complete AS events and clinical data were included in our analysis. The clinical features of the patients are summarized in **Table 1**.

Screening for Prognostic AS Events in UCEC

TCGA SpliceSeq is a database based on TCGA RNA sequencing (RNA-seq) data. Seven types of selective splicing events were

analyzed, including Alternate Acceptor site (AA), Alternate Donor site (AD), Alternate Promoter (AP), Alternate Terminator (AT), Exon Skip (ES), Mutually Exclusive Exons (ME), and Retained Intron (RI). We analyzed the distributions of all encoded genes using the UpSetR package in each of the seven different types of AS events and survival-related AS events in UCEC.

Construction of Prognostic Models and Survival Analysis

Different AS events in genes led to diversity in outcomes, and changes in gene expression affected survival time. To further understand the prognostic value of AS events in UCEC patients, univariate Cox regression analysis with R package "survival" was performed to determine the survival-related different expressed alternative splicing (DEAS) events, including overall survival (OS)-related DEAS events. Next, the least absolute shrinkage and selection operator (LASSO) regression was applied to identify the final elimination of potential predictors with non-zero coefficients using the R package "glmnet", which can avoid model overfitting to obtain a better fitting model. Furthermore, based on the results of LASSO Cox regression, predictive models were constructed using multivariate Cox regression analysis. Based on the PSI values and multivariate Cox analysis, we calculated the risk scores of each patient and obtained the corresponding coefficients, respectively. The following formula obtained the risk score:

$$\text{Risk score} = \sum_{i=0}^{n} \text{PSI} \times \beta_i$$

where β is the regression coefficient of the AS events. A total of 524 EC patients were divided into high- and low-risk groups bound by the median of risk score, and Kaplan-Meier survival analysis was performed to determine whether they had completely different prognoses. Furthermore, receiver operating characteristic (ROC) curves of 1, 3, and 5 years were generated using the survival ROC package in R to show the discrimination of the predictive signatures (Heagerty et al., 2000).

Establishment and Validation of a Predictive Nomogram

All clinical factors, including the risk score, age, FIGO stage, and grade, were incorporated to construct a nomogram to evaluate the probability of 1-, 3-, and 5-year OS of UCEC patients in the entire set. Validation of the nomogram was evaluated by calibration plot using the "rms" package. The calibration curve of the nomogram was plotted to assess the nomogram-predicted probabilities against the actual rates.

Immune Score Estimate and Immune Cell Infiltrating Proportion Inference

Normalized RNA expression data were used to infer the Immune Score using the estimate package (Yoshihara et al., 2013) and quantify the infiltrating proportions of 22 types of immune cells using the "CIBERSORT" package (Newman et al., 2015). The infiltrating percentage of 22 types of immune cells was equal to 100%. Single sample gene set enrichment analysis (ssGSEA) was used to quantify and classify the immunity stage based on immune-related gene (IRGs) sets (He et al., 2018). Next, 47 immune checkpoint genes were analyzed, and 16 of them that differed from the tumor and normal samples were screened. The differences between the 16 hub immune checkpoints among the high- and low-risk groups were analyzed, and the correlations between the six most important immune checkpoint genes (CD274, PDCD1, PDCD1LG2, CTLA4, HAVCR2, and IDO1) and the risk score were determined.

Extraction of AS-Related DEGs in UCEC Samples

Using R package "limma" with the threshold of $|\log2FC|>1$ and $p < 0.05$, the 11 genes (**Table 2**) involved in the model construction were analyzed to observe whether their expression differed between the UCEC and normal samples.

Integration of AS-Related DEGs With Clinical Characteristics and Prognosis

High- and low-expression groups of gub genes were obtained according to the gene expression. These were then used to analyze the difference in clinical indicators, including age, grade, and FIGO stage. Finally, the prognosis of the hub genes in the two groups was judged using the "survival" and "survminer" packages.

Analysis of the Relationship Between Stromal/Immune Scores and AS-Related DEGs in the EC Immune Microenvironment

The ESTIMATE algorithm was applied to analyze the Stromal Score, Immune Score, ESTIMATE Score, and Tumor Purity based on the transcriptome profiles of UCEC to determine the effect ssGSEA grouping. We further compared the Stromal Score, Immune Score, ESTIMATE Score, and Tumor Purity in the high- and low-expression groups of hub genes using the Limma.R and ggpubr.R packages. The relationships between the copy number of hub genes and the quantity of six immune cells (B cell, myeloid DC cell, macrophage, neutrophil, T cell $CD4^+$, and T cell $CD8^+$) were evaluated using the Tumor Immune Estimation Resource (TIMER) database.

Construction of a Potential SF-AS Regulatory Network

Splicing factors (SF) are protein factors involved in the splicing process of RNA precursors, which are closely related to the development and treatment of cancer (Dvinge et al., 2016; Obeng et al., 2019). A total of 404 SFs data downloaded from the SpliceAid2 database were used to analyze the correlation between the expression level of SFs and the PSI values of OS-associated AS events by R packages (BiocManager, limma). An

TABLE 2 | Eleven AS events associated with the OS of UCEC patients.

ID	Coefficient	HR	HR.95L	HR.95H	*p* value
MAST1|47878|AT	1.756261	5.790747	2.208989	15.180133	0.000355
CYB561|42921|AP	2.851763	17.318280	3.689684	81.286860	0.000301
MAGED1|89145|AP	2.280885	9.785341	0.965560	99.168280	0.053569
PCYT2|44230|ES	1.310632	3.708517	1.054692	13.039925	0.041056
SULT1A3|94136|AP	−1.203415	0.300167	0.070887	1.271038	0.102203
FOLH1|15817|ES	−5.303092	0.004976	0.000452	0.054736	0.000015
ZNF706|84749|ES	3.998052	54.491877	2.482313	1196.208995	0.011185
CCNL2|162|ES	1.824693	6.200888	0.853256	45.063858	0.071366
RPLP0|24731|ES	−7.978477	0.000343	0.000003	0.043102	0.001218
STK32C|13483|AP	2.122228	8.349723	0.979867	71.150383	0.052215
C4orf29|70557|AT	1.696728	5.456064	0.611076	48.715079	0.128758

absolute value of the correlation coefficient >0.6 and $p < 0.001$ were considered statistically significant. Finally, Cytoscape software (v3.7.2, https://cytoscape.org/) was used to visualize the potential SF-AS regulatory network.

Statistical Analysis

All statistical analyses were performed using R version 4.1.0 (R packages: survival, survminer, UpSetR, glmnet, estimate, ggpubr, e1071, rms, preprocessCore, vioplot, ggExtra, GSVA, GSEABase, reshape2, pheatmap, corrplot, ggplot, ggplot2, and BiocManager). For all analyses, a two-tailed $p < 0.05$ was regarded as statistically significant if not noted.

RESULTS

Overview of AS Events in TCGA UCEC Cohort

A total of 524 UCEC patients were identified, and the baseline characteristics of these patients are summarized in **Table 1**. The mRNA splicing data included in this study contains 28,281 AS events in 8,141 genes. Given the possibility of multiple splicing modes for a single gene, we created UpSet plots to analyze interactive sets of seven types of AS events quantitatively. As shown in **Figure 1A**, a single gene could have up to five different splicing modes, and most genes had more than one AS event. Exon skip (ES) was the most frequent splice type among the seven AS types (34.4%), followed by an alternate terminator (AT) (27.5%) and alternate promoter (AP) (15.7%).

Prognostic Index Models Featured by AS Events for UCEC

To explore the prognostic utility of an AS signature in EC, AS events associated with OS were identified by fitting univariate Cox proportional hazard regression models after merging the clinical data in the training cohort using Perl. In total, 1,108 AS events were determined with $p < 0.05$, including 633 high-risk survival-associated AS events (hazard ratio HR > 1) and 475 low-risk survival-associated AS events (HR < 1). The AS events can be counted through the UpSet plot. An UpSet plot was generated to visualize the intersecting sets between different genes and AS events. The bar charts on the left showed the number of genes with some kind of AS events. The upper bar charts showed the number of intersecting genes, indicating the number of genes with a certain type or types of AS events (**Figure 1A**). **Figure 1A** indicates that one gene might have more than one survival-associated AS event. It is noteworthy that the three highest frequency survival-associated AS events were still ES, AT, and AP in the UCEC cohort.

After conducting univariate regression analysis, LASSO regression was performed to select the optimal survival-related AS events to construct the prediction models to avoid model overfitting based on OS. First, 15 AS events were screened out by LASSO regression, and then the AS events with the same contribution were optimized (**Figure 1B**). Finally, an 11-AS event signature was identified as a predictor of survival in EC through the Cox proportional hazards regression model (**Table 2**). Besides, the minimum adjusted estimate of cross-validation prediction error was 0.020, and the *p* value of bootstrap ($K = 1000$) was 4.82E-305.

Kaplan-Meier curves and log-rank tests were plotted to explore the relationship between risk score and survival status. The survival probability of low-risk patients was higher than that of high-risk patients; in other words, high-risk patients had a higher mortality rate, as illustrated in **Figure 1C** ($p < 0.001$). We then applied ROC analysis to compare the predictive power of these prognostic models. The larger the area under the curve, the higher the accuracy of the model to predict the prognosis of patients. **Figure 1D** showed a robust and significantly improved performance; the areas under the ROC curve (AUC) in 1, 2, and 3 years were all greater than 0.800. The result illustrated that the accuracy of using the model to predict the 1-, 2- and 3-year survival rate of patients was relatively high. Moreover, the AUC of the risk score model predicting the 1-year survival rate was larger than that of the age, grade, and FIGO stage (**Figure 1D**). It means that the accuracy of predicting the 1-year survival rate of patients by the model is better than that of using other clinical parameters (age, grade, stage) to predict the prognosis.

Meanwhile, the risk scores of each UCEC patient were calculated, and all patients were divided into low- and high-risk groups bound by the median risk score. The distribution diagram of survival risk score (**Figure 2A**), survival status of EC patients (**Figure 2B**), and clustering heatmap of the PSI levels of eleven-AS

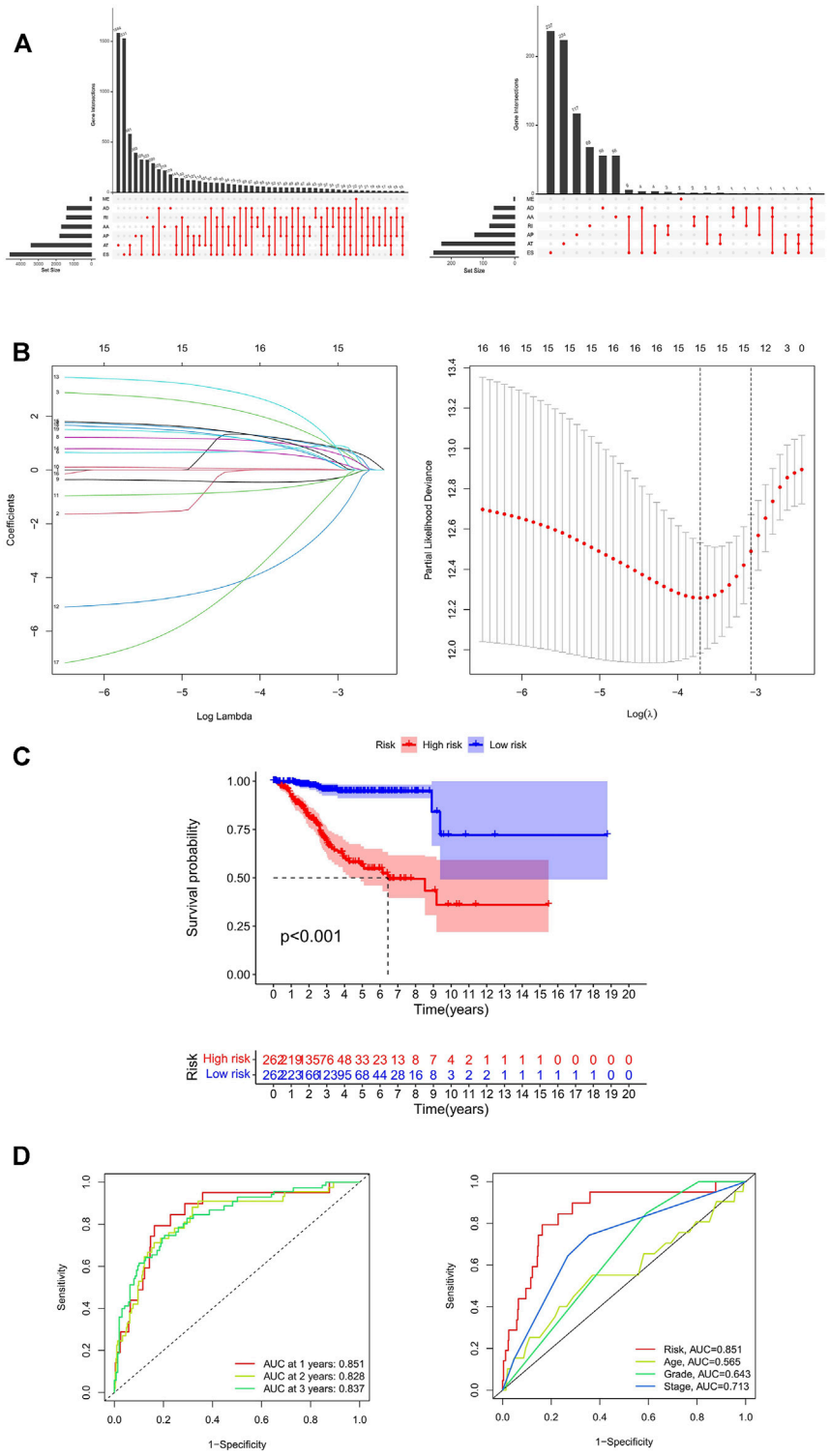

FIGURE 1 | Identification and prognosis of AS markers in UCEC. **(A)** The upSet plot of intersections and aggregates among diverse types of AS (left) and survival-associated AS events (right) in UCEC. **(B)** LASSO coefficient profiles of survival-associated AS events and 10-time cross-validation for tuning parameter selection in the LASSO model. **(C)** Kaplan-Meier analysis for OS of UCEC patients. **(D)** ROC curve in the predicted groups (high- and low-risk groups) by the 11-AS events signature in the UCEC cohort. AA, alternate acceptor; AD, alternate donor; AP, alternate promoter; AT, alternate terminator; ES, exon skip; ME, mutually-exclusive exons; RI, retained intron; UCEC, Uterine Corpus Endometrial Carcinoma.

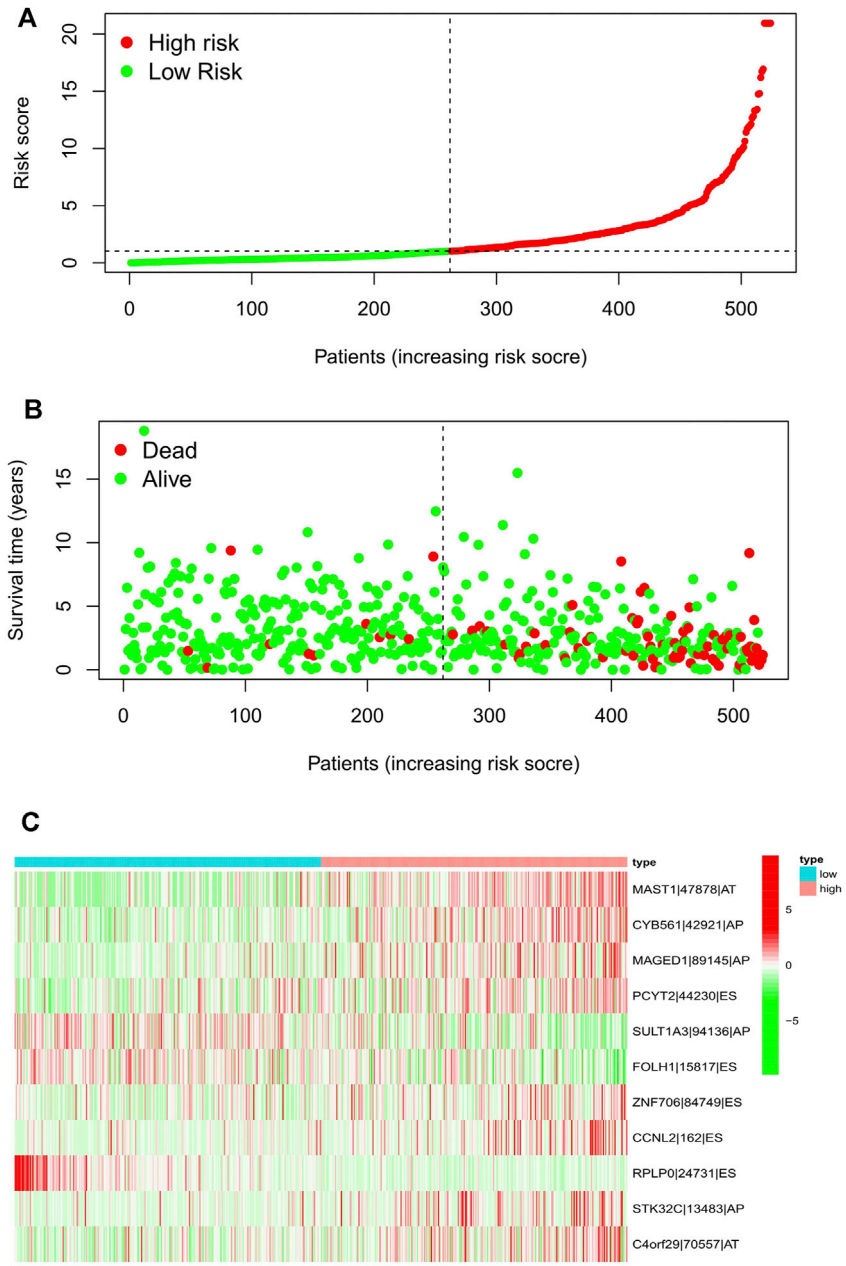

FIGURE 2 | Risk score distribution of the 11-AS events signature in the TCGA cohort. Risk scores **(A)**, survival status **(B)**, and heatmap **(C)** of the 11-AS events PSI profiles were shown from top to bottom. AA, alternate acceptor; AD, alternate donor; AP, alternate promoter; AT, alternate terminator; ES, exon skip; ME, mutually-exclusive exons; RI, retained intron.

markers (**Figure 2C**) are shown. The horizontal axis displays the patients' order of risk score from low to high (**Figure 2**).

Construction and Evaluation of the Nomogram

The calibration curve demonstrated that the predicted values are satisfactorily consistent in the prediction of the 1-, 3-, and 5-year OS because the red lines in three pictures are almost overlap with the 45° dashed lines (**Figure 3B**). The box charts in **Figure 3E** show whether there are differences in patients' risk score among different clinical index (age, grade, stage). The risk score of patients >65 years old was higher than that of patients ≤65 years old. With the increase of grade and stage of UCEC, the risk score increased gradually (**Figure 3B**).

Univariate and multivariate Cox regression methods were used and combined with patient clinical characteristics (age, grade, and FIGO stage) to analyze whether the 11-AS event signature could be an independent predictor of survival in patients with UCEC. When the *p* values of univariate analysis and multivariate analysis were both less

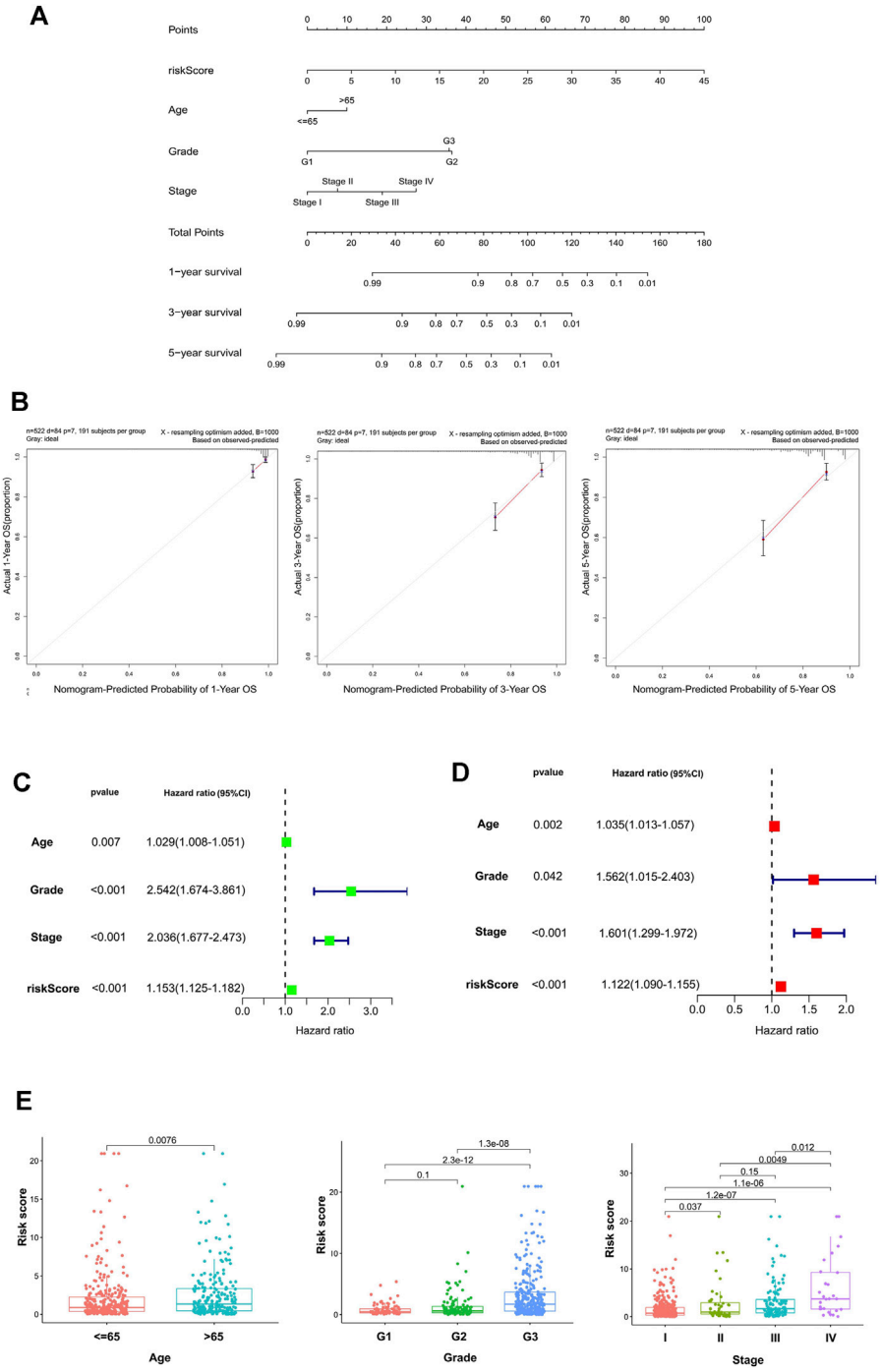

FIGURE 3 | The establishment and validation of the nomogram. **(A)** The nomogram consisted of age, gender, FIGO stage, and risk score, and was used to predict the 1-, 3-, and 5-year survival probability of EC patients. **(B)** Calibration plots of the AS-clinical nomogram are in agreement between the nomogram-predicted and observed 1-, 3-, and 5-year outcomes of the UCEC cohort. The nomogram-predicted survival probability is plotted on the *x*-axis, and the actual survival is plotted on the *y*-axis. The 45° dashed line represents the ideal performance. The red lines represent the actual performances of the model, and the figures from left to right depict the 1-, 3-, and 5-year results. Univariate analysis **(C)** and multivariate analysis **(D)** of risk scores and clinical characteristics that were simultaneously associated with OS. **(E)** Differences in the risk score in terms of age, grade, and FIGO stage groups. The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range).

than 0.05, it was considered that the model could be an independent prognostic factor. As depicted in **Figures 3C,D**, the results showed that the risk score could still be used as a reliable and stable independent risk predictor in the UCEC cohort ($p < 0.001$; **Figures 3C,D**). We then constructed a predictive nomogram based on the multivariate analysis (**Figure 3A**) that included risk scores and

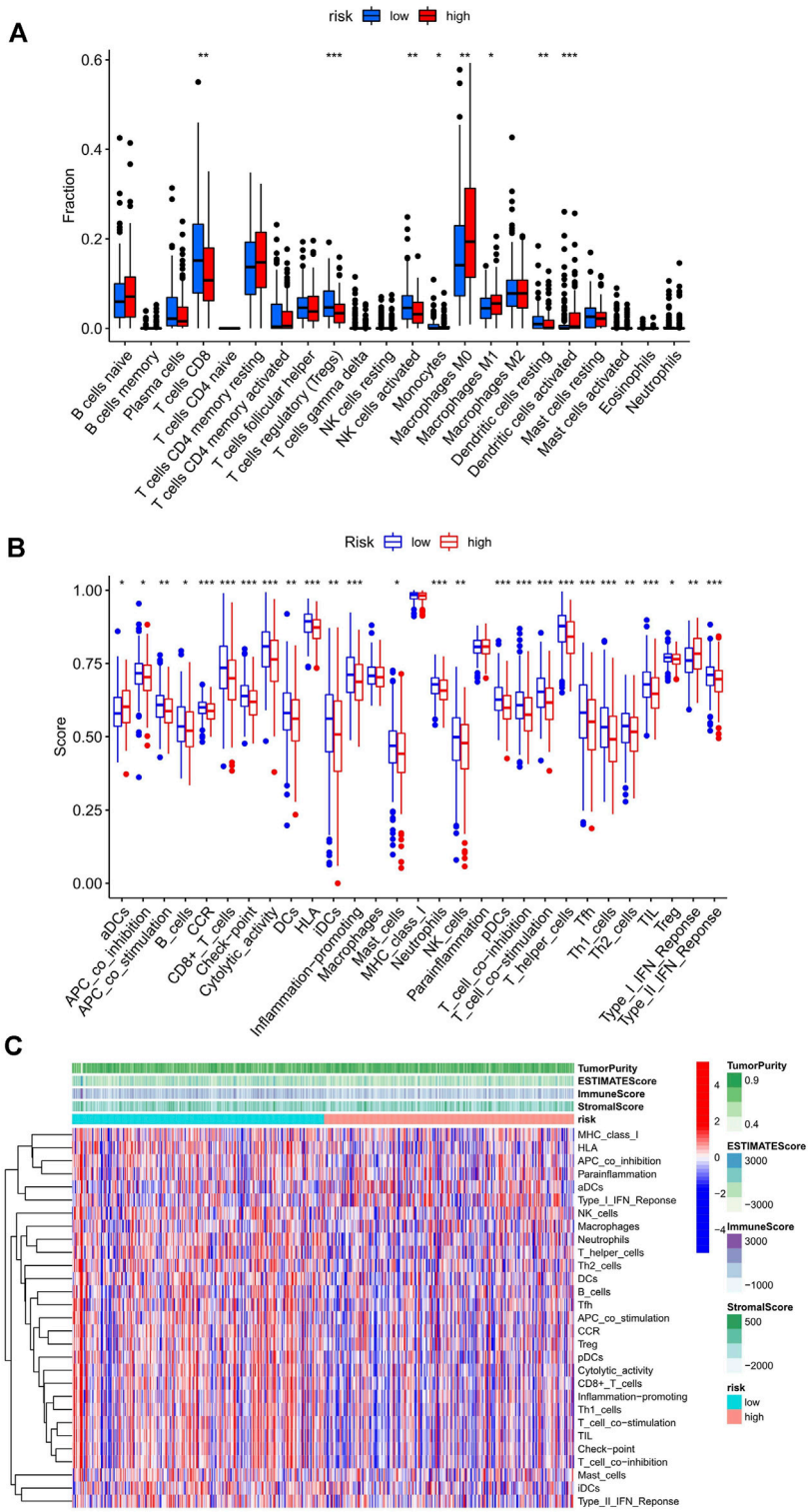

FIGURE 4 | Relationship between the risk score and infiltrating immune cells in the UCEC tumor-immune microenvironment. **(A)** The landscape of 22 types of infiltrating immune cells in the low-risk score ($n = 262$) and high-risk score ($n = 262$) groups. **(B)** The landscape of 29 types of infiltrating immune cells and immune function in the two groups. The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range). Blue: low risk, red: high risk. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. **(C)** The heatmap showed a difference in the infiltrating immune cells between the two groups in the UCEC tumor-immune microenvironment.

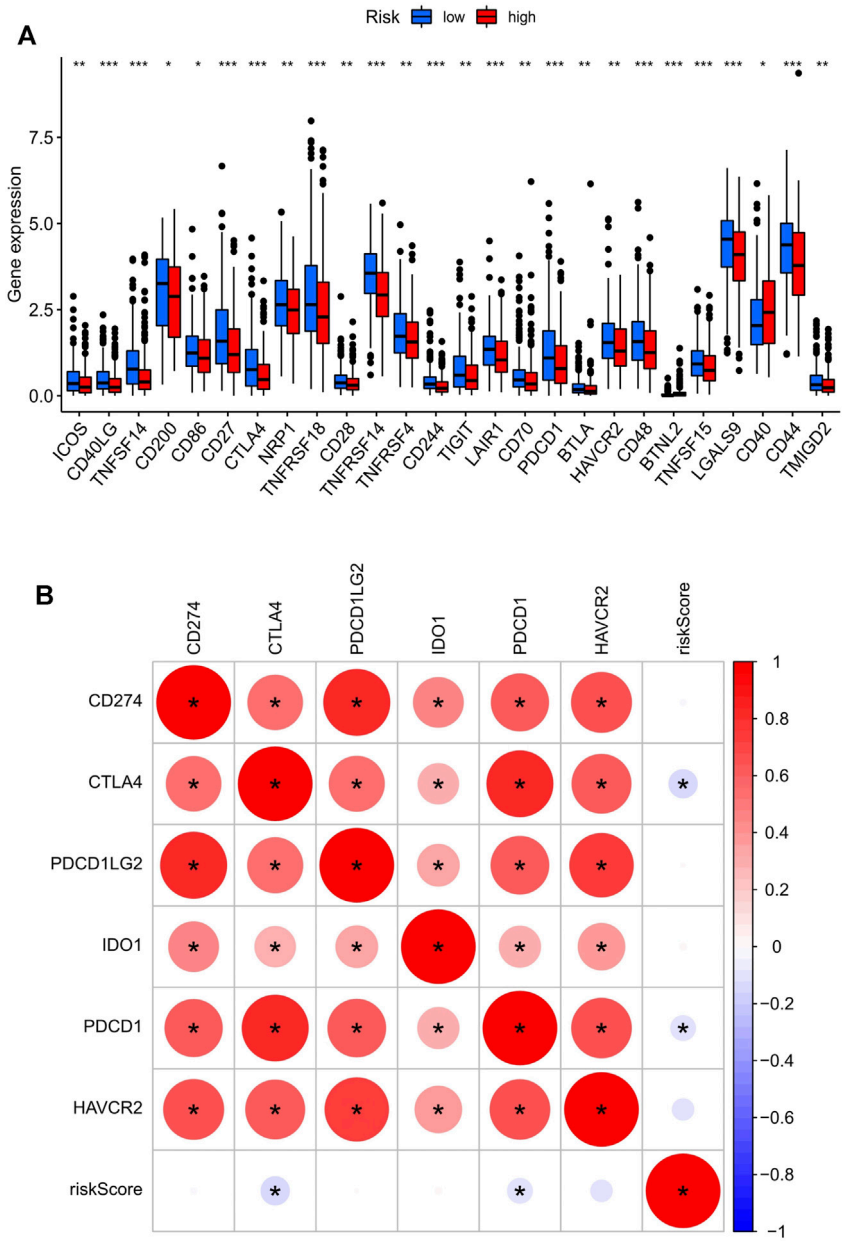

FIGURE 5 | The key immune checkpoint genes are related to the risk score in the UCEC tumor-immune microenvironment. **(A)** The landscape of 26 types of immune checkpoint genes in low- and high-risk score groups. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. **(B)** The correlation between the risk score and the six most important checkpoint genes (CD274, PDCD1, PDCD1LG2, CTLA4, HAVCR2, and IDO1). The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range). *: statistically significant; red: positive correlation, blue: negative correlation.

clinical characteristics. The results demonstrated that the risk score had satisfactory diagnostic ability and clinical characteristics ($p < 0.05$).

The Risk Score and AS Events Are Associated With the Infiltration of Immune Cells in the UCEC Microenvironment

First, the immune score in 29 types of infiltrating immune cells and immune function was assessed by the ssGSEA method (He et al., 2018). **Figures 4B,C** show the immune score differences of each immune cell in the low and high-risk score groups (**Figures 4B,C**). We further explored the impact of the risk score on the infiltration of 22 types of immune cells in the tumor microenvironment using the CIBERSORT algorithm. The landscape of 22 types of infiltrating immune cells in the low and high-risk score groups is shown in **Figure 4A**. Differential analysis results showed that eight types of immune cells [CD8 T cells, regulatory T cells (Tregs), activated natural killer (NK)

cells, monocytes, M0 macrophages, M1 macrophages, resting dendritic cells, and activated dendritic cells] were significantly different between the two groups ($p < 0.05$).

The Risk Score is Associated With the Key Immune Checkpoint Genes in the UCEC Tumor-Immune Microenvironment

The difference in the expression level of 47 immune checkpoint genes in the low- and high-risk score groups was assessed, and 26 genes were found to have significant differences (**Figure 5A**). Next, R packages (limma, corrplot, ggpubr, and ggExtra) were used to screen the risk scores related to the six most important checkpoint genes (CD274, PDCD1, PDCD1LG2, CTLA4, HAVCR2, and IDO1). Two immune checkpoint genes, PDCD1 and CTLA4, with negative correlation with risk score were identified ($p < 0.001$; **Figure 5B**). The scatter plot displaying the correlation of these two genes and the risk score were plotted separately. Although two of the correlation coefficients did not reach 0.3, the scatter plot showed a negative correlation (**Supplementary Figure S1**). At the same time, we can find that the expression of PDCD1 and CTLA4 in the high-risk score group was lower than that in the low-risk score group (**Figure 5A**).

Extraction of IRGs Depending on AS Events and Their Correlation With Clinical Parameters

The expressions of 11 genes (**Table 2**) were identified to analyze the difference between UCEC and normal cohorts by the "limma" package (with the threshold of $|\log2FC| > 1$ and $p < 0.05$), and two genes, CYB561 ($|\log FC| = 1.1892$, $p < 0.001$) and FOLH1 ($|\log FC| = 1.0862$, $p < 0.001$), were extracted for further analysis. Next, we divided the tumor patients into high- and low-expression groups according to the optimal cut-offs in CYB561 and FOLH1 (4.67 in CYB561, 2.64 in FOLH1) for clinical prognostic analysis.

The correlations between the expression of the two hub genes and the clinicopathological parameters were evaluated using R packages (limma, survival, and survminer). CYB561 and FOLH1 expression levels were significantly associated with grade and FIGO stage ($p < 0.05$). The expression of CYB561 and FOLH1 decreased gradually with increases in the grade and stage. However, no notable association between the two genes and age was observed ($p \geq 0.05$) (**Figures 6A,B**). Survival analysis revealed that the low expressions of both CYB561 and FOLH1 were associated with poor prognosis ($p < 0.001$) (**Figure 6C**).

Associations Between CYB561 Expression and Immune Cell Infiltration

The landscape of 22 types of infiltrating immune cells in the low- and high- CYB561 expression groups are shown in **Figure 7A**. The two groups differed between resting dendritic cells, neutrophils, activated memory CD4 T cells, resting memory CD4 T cells, and Tregs. **Figure 7B** shows the immune score difference of each immune cell in the two groups (**Figure 7B**). We investigated the association between CYB561 expression and the tumor-infiltrating immune cells in UCEC using the TIMER database. The results demonstrated that CYB561 expression was positively correlated with B cell, $CD4^+$ T cells, and $CD8^+$ T cells, and was negatively correlated with myeloid dendritic cells, macrophages, and neutrophils ($p < 0.05$) (**Supplementary Figure S2**). However, we found no strong correlation between immune cell infiltration and CYB561 expression. Given that the risk score was related to tumor immunity, we finally appraised the correlation between the gene signature and the expression of immune checkpoints. **Figure 7C** shows the 23 immune checkpoints with differential expression in the low- and high-CYB561 expression groups. The gene expression of CD40LG, TNFSF14, TNFRSF14, CD276, VTCN1, HHLA2, TNFSF15, LGALS9, and CD44 was lower in the low-CYB561 expression groups (**Figure 7C**).

Correlations Between FOLH1 Expression and Immune Cell Infiltration

The landscape of 22 types of infiltrating immune cells in the low- and high- FOLH1 expression groups are shown in **Figure 8A**. Resting memory CD4 T cells, gamma delta T cells, resting NK cells, resting dendritic cells, activated dendritic cells, and neutrophils were different in the two groups. **Figure 8B** shows the immune score difference of each immune cell in the two groups. We further investigated the association between FOLH1 expression and the tumor-infiltrating immune cells in UCEC. The results showed that FOLH1 expression was positively correlated with macrophages, $CD4^+$ T cells, and $CD8^+$ T cells, and was negatively correlated with B cells, myeloid dendritic cells, and neutrophils ($p < 0.05$) (**Supplementary Figure S3**). However, we also found no strong correlation between immune cell infiltration and FOLH1 expression. **Figure 8C** shows the immune checkpoints with differential expression in the low- and high-FOLH1 expression groups.

The Tumor Purity, ESTIMATE Score, Immune Score, Stromal Score and Between the Low- and High-Expression Groups of CYB561 and FOLH1

First, the violin plot assessed the differences in Tumor Purity, ESTIMATE Score, Immune Score, and Stromal Score between the two groups, calculated using the ESTIMATE algorithm (**Figure 9A**). ESTIMATE Score and Immune Score were higher in the low-risk score group, while Tumor Purity in the low-risk score group was lower than that in the high-risk score group ($p < 0.05$). The Stromal Score showed no difference ($p \geq 0.05$).

In order to determine the effectiveness of the grouping strategy between the low- and high-expression groups of CYB561 and FOLH1, the ESTIMATE method was applied to evaluate Tumor Purity, ESTIMATE Score, Immune Score, and Stromal Score. Compared with the high-CYB561 expression group, the low expression group had a higher Stromal Score ($p < 0.05$) (**Figure 9B**). The other parameters had no differences between the two groups in CYB561 and FOLH1 ($p \geq 0.05$) (**Figures 9B,C**).

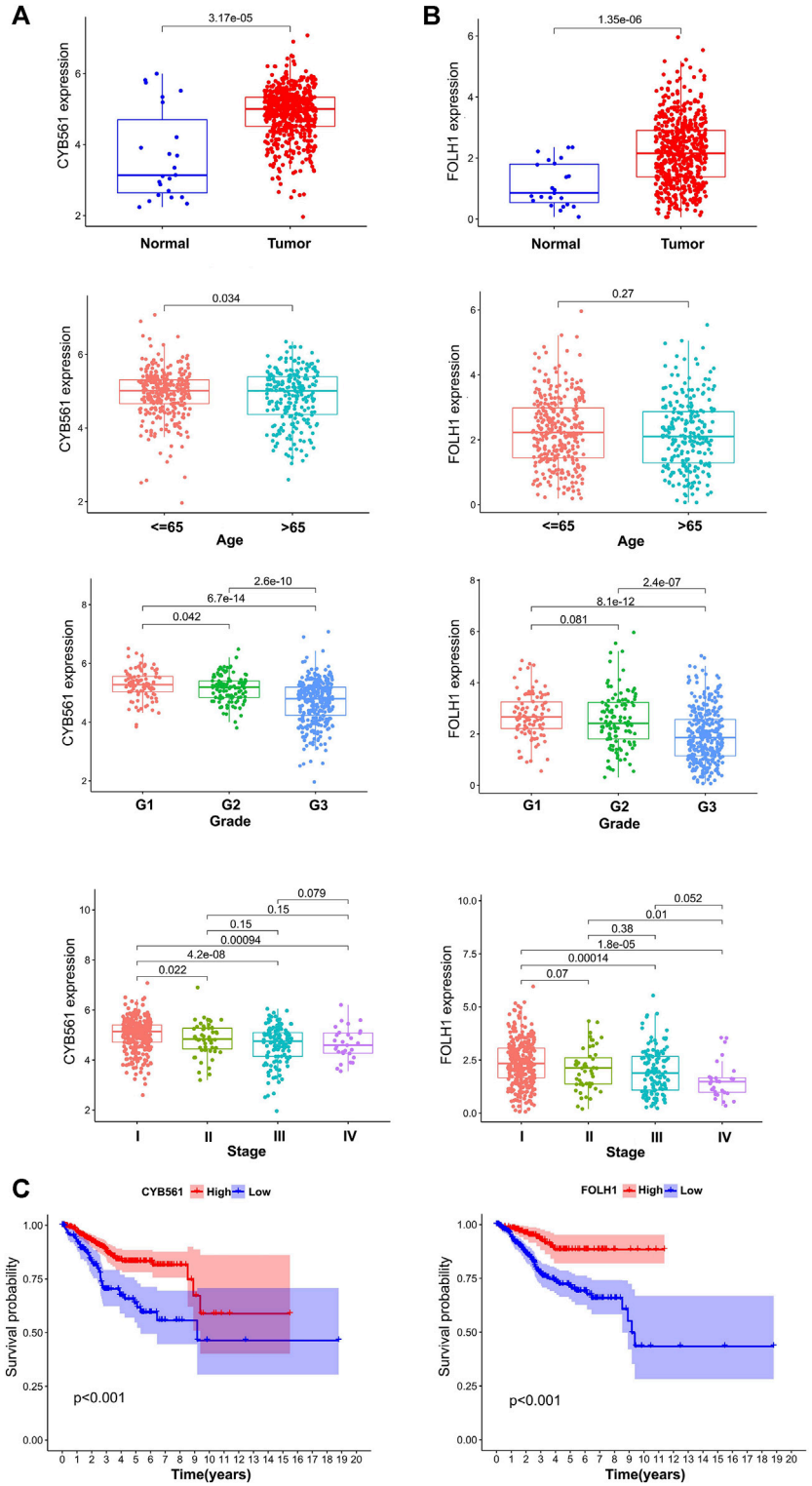

FIGURE 6 | The prognostic signature of CYB561 and FOLH1 expression. The expression of CYB561 **(A)** and FOLH1 **(B)** in age, grade, and stage groups. **(C)** Kaplan-Meier survival curve of CYB561 and FOLH1 in high- and low-expression groups. The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range).

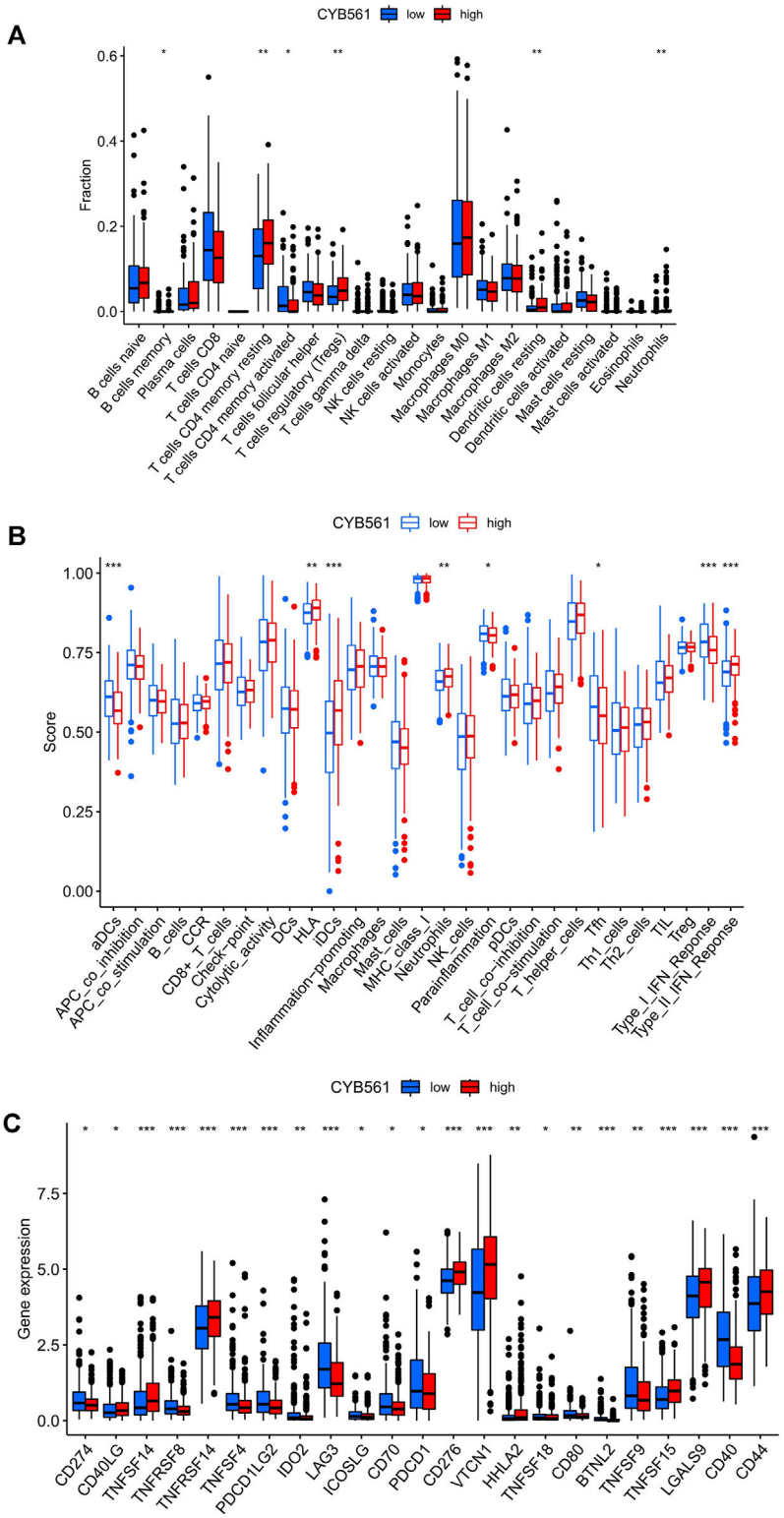

FIGURE 7 | Relationship between CYB561 expression and infiltrating immune cells in the UCEC tumor-immune microenvironment. **(A)** The box plot shows the proportion difference of each immune cell between the low- and high- CYB561 expression groups. **(B)** The landscape of infiltrating immune cells and immune function in both groups. **(C)** The expression of CYB561 was associated with the key immune checkpoint genes in the UCEC microenvironment. The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range). Blue: low risk, red: high risk. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

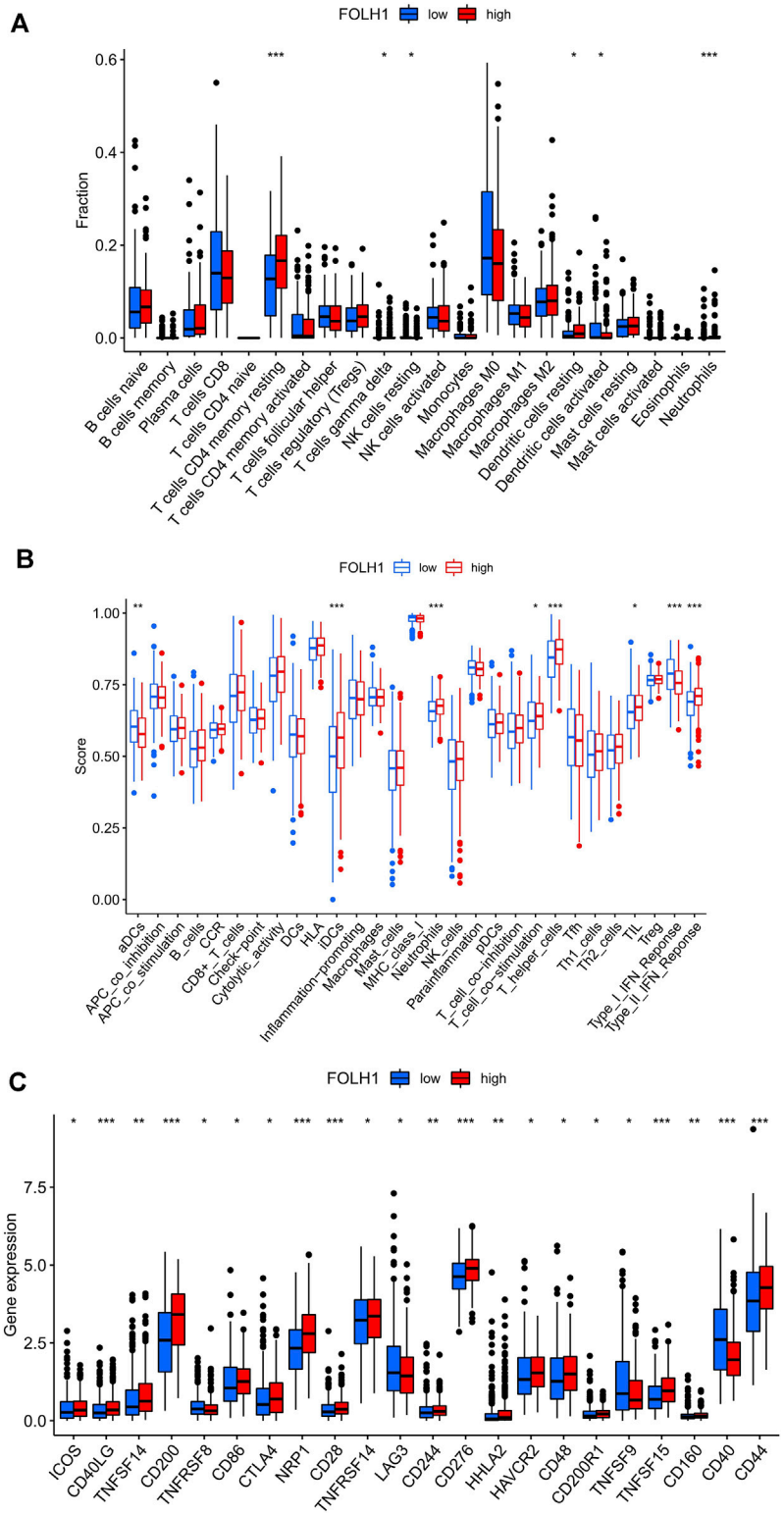

FIGURE 8 | Relationship between FOLH1 expression and infiltrating immune cells in the UCEC tumor-immune microenvironment. **(A)** The box plot shows the proportion difference of each immune cell between the low- and high-FOLH1 expression groups. **(B)** The landscape of infiltrating immune cells and immune function in the two groups. **(C)** The expression of FOLH1 was associated with key immune checkpoint genes in the tumor microenvironment (TME). The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range). Blue: low risk, red: high risk. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

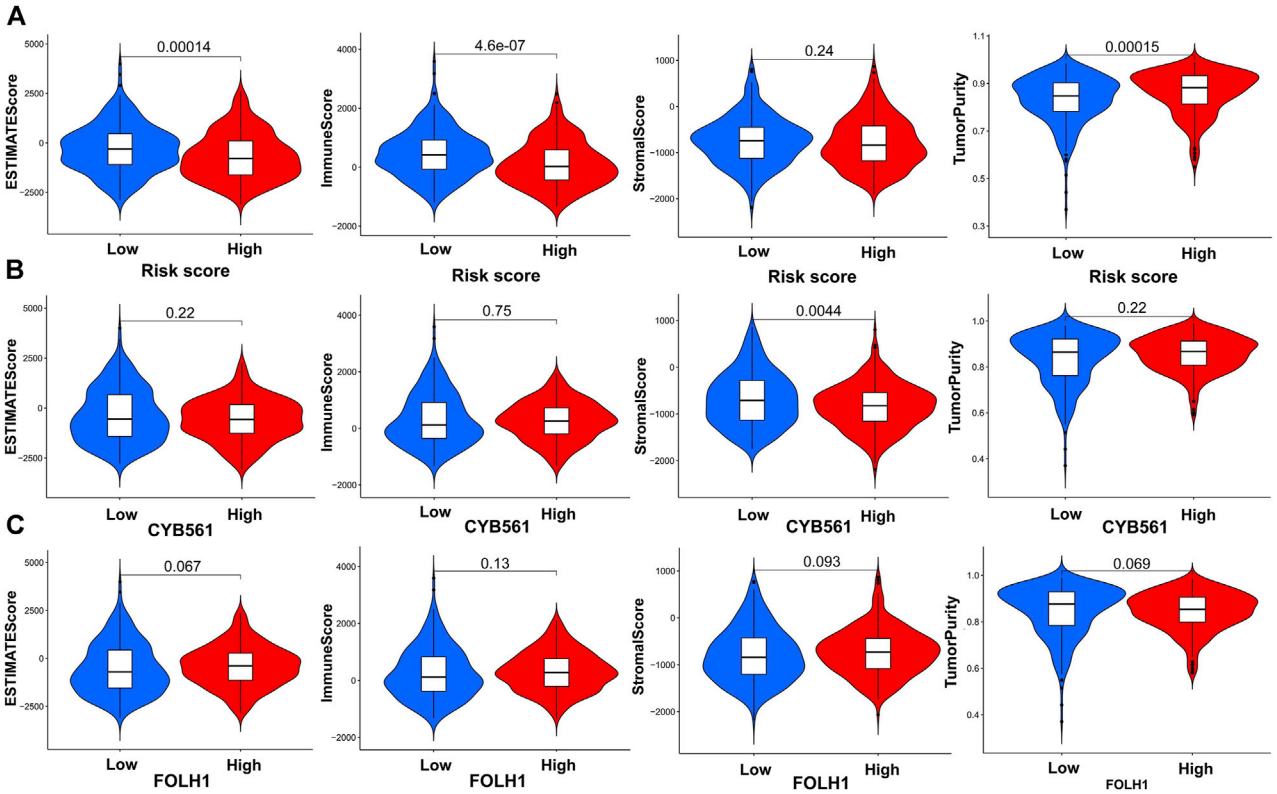

FIGURE 9 | Construction and verification of the immune-related UCEC subgroups. **(A)** The violin plot shows the differences in Tumor Purity, ESTIMATE Score, Immune Score, and Stromal Score between the high- and low-risk score groups, calculated using the ESTIMATE algorithm. Comparison of Stromal Scores, Immune Scores, ESTIMATE Scores, and Tumor Purity between two groups of CYB561 **(B)** and FOLH1**(C)**.

The Quantity of Six Immune Cells in Gene Copy Number of CYB561 and FOLH1

The correlations between the CYB561 copy number and six immune cells were also analyzed using the TIMER database. The number of cells was found to decrease with the increase in the gene copy number in myeloid DC cells and $CD8^+$ T cells, and was found to decrease with the decrease in the gene copy number in myeloid DC cells and $CD8^+$ T cells ($p < 0.05$) (**Supplementary Figure S4**). Next, we analyzed the correlations between the FOLH1 copy number and six immune cells. The number of cells was found to decrease with the increase in the gene copy number in myeloid DC cells, macrophage, $CD8^+$ T cells, and the cells number was found to decrease with the reduce in the gene copy number in myeloid DC cells and $CD8^+$ T cells ($p < 0.05$) (**Supplementary Figure S5**). Notably, we observed that the change in gene copy numbers in the two hub genes led to the deletion number of both myeloid DC cells and $CD8^+$ T cells.

The Potential Regulatory Network Between SFs and AS Events

Thirty-six SFs (blue) were found to be significantly related to 19 survival-associated AS events, consisting of nine low-risk AS events (RPS24|12297|AA, ZNF169|86927|AT, MAT2A|54280|AT, C22orf39|61055|AT, VDAC1|73335|AP, FAM72A|9577|AT, PPP2R5C|29315|AT, TMEM33|69133|RI, SYTL1|1321|AP; purple) and 10 high-risk AS events (LEPROTL1|83274|AT, ANKHD1|73652|AT, MAT2A|54281|AT, BUB3|13390|AA, ATP5D|46401|RI, FAM72A|9578|AT, VDAC1|73334|AP, MARVELD3|37467|AT, NDUFA3|51776|AT,SYTL1|1320|AP; red). The majority of low-risk AS events were negatively correlated with SF expression (blue lines), and all of the high-risk AS events were positively correlated with SF expression (red lines) (**Figure 10**).

DISCUSSION

Dysregulation of AS can affect essential biological processes and thus drive disease-associated pathophysiology (Gamazon and Stranger, 2014). Emerging data have demonstrated that aberrant AS events are closely associated with cancer progression, metastasis, therapeutic resistance, and other oncogenic processes (Climente-González et al., 2017). Cancer cells have general and cancer type-specific and subtype-specific alterations in the splicing process, which can have prognostic value and contribute to every hallmark of cancer progression, including the cancer-immune responses (Bonnal et al., 2020). Moreover, substantial preclinical work has identified a variety of

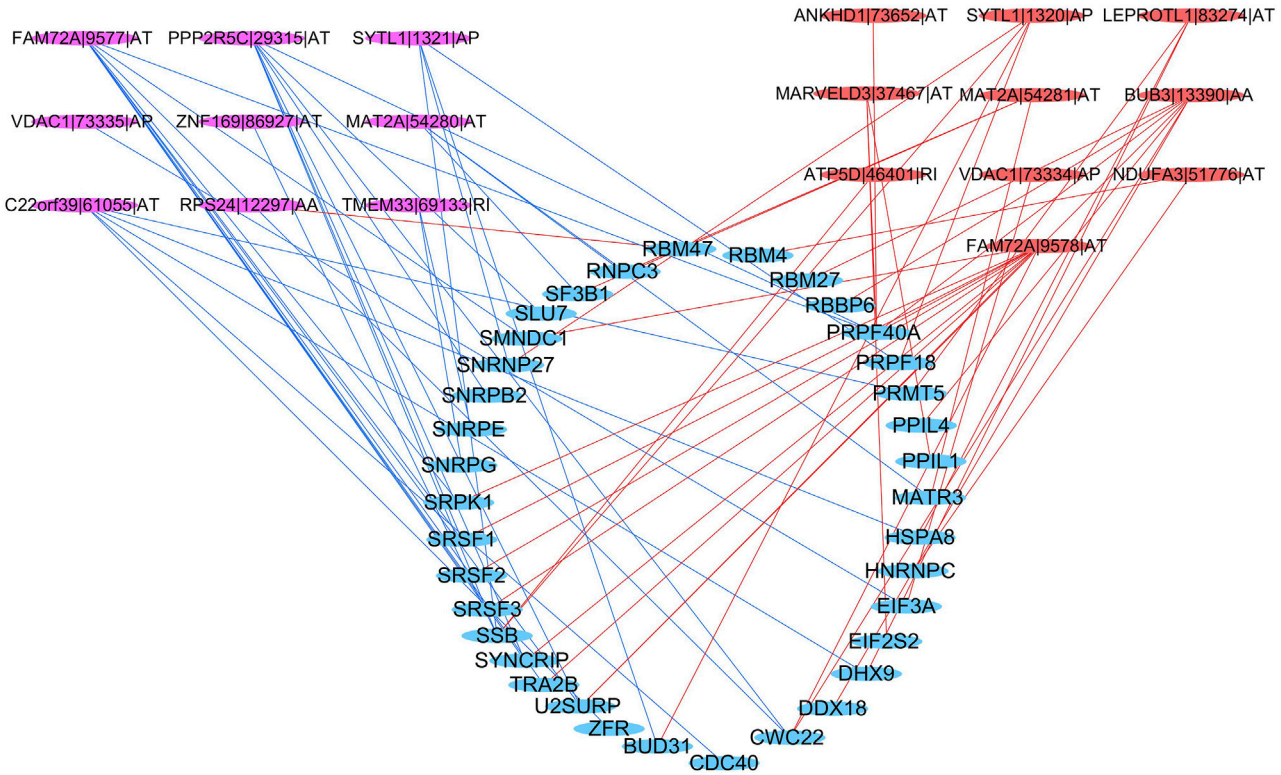

FIGURE 10 | Regulatory network between SFs and AS events. Blue ellipse: spliced factors, purple ellipse: low-risk AS events, red ellipse: high-risk AS events, blue line: negative correlation, red line: a positive correlation.

small molecule compounds and genetic and other approaches to target the spliceosome or its products with potential therapeutic effects (Bonnal et al., 2020). Therefore, it is of great importance to further study the characteristics of AS in the immune microenvironment for UCEC immunotherapy. In recent years, the relationship between AS and UCEC has been studied. In endometrial cancer, AS of vascular endothelial growth factor A (VEGF-A) is regulated by RBM10 (Dou et al., 2020a). Popli et al. found that SF3B1 plays a crucial oncogenic role in the tumorigenesis of EC and hence may support the development of SF3B1 inhibitors to treat this disease (Dou et al., 2020a). XQ et al. showed that miR-335 modulates Numb AS *via* targeting RBM10 in EC (Dou et al., 2020b).

We extracted IRGs depending on AS events and examined their correlation with clinical parameters. Finally, two AS-related genes, CYB561|42921|AP and FOLH1|15817|ES, were extracted from the 11 genes involved in the AS prognostic model. CYB561 encodes the protein CYB561, named as such because of its optical absorbance at 561 nm CYB561 is a heme-containing enzyme that is necessary for the continuous regeneration of semidehydroascorbate to ascorbate inside chromaffin granules and neuropeptide secretory vesicles (van den Berg et al., 2018). It is widely expressed in the adrenal glands, prostate, and 23 other tissues, including the endometrium. However, data on the role of the CYB561 gene in human cancers are very limited. A meta-analysis showed that low mRNA expression of CYB561 was prognostic of a poor outcome in ovarian cancer (Willis et al., 2016). CYB561 serves as a potential prognostic biomarker and target for breast cancer (Yang et al., 2021). We found the expression of CYB561 decreased gradually with the increased grade and FIGO stage which indicated a lower survival probability ($p < 0.001$) in UCEC. Our results were consistent with the previous ones. Besides, alternate promoter of CYB561 was associated with the OS of UCEC patients ($p = 0.0003$) in our research. The changes in transcription is regarded as a defining feature of cancer. Most human protein-coding genes are regulated by multiple, distinct promoters, suggesting that the selection of promoter is closely related to the expression of target gene (Demircioğlu et al., 2019). How the AP contributes the low expression of CYB561 in endometrial carcinoma remains to be further explored.

FOLH1 is also known as prostate-specific membrane antigen (PSMA), which encodes a transmembrane glycoprotein that acts as a glutamate carboxypeptidase on different alternative substrates. In the prostate, this protein is up-regulated in cancerous cells and is used as an effective diagnostic and prognostic indicator of prostate cancer (Date et al., 2017). PSMA is highly and specifically expressed in the neovasculature of ovarian, endometrial, and cervical squamous carcinomas (Wernicke et al., 2017). Mhawech-Fauceglia et al. showed that PSMA is under-expressed in advanced stage endometrial adenocarcinoma (Mhawech-Fauceglia et al., 2008),

which is consistent with our findings. Their research indicated that the loss of PSMA expression can be considered a prognostic marker in patients with endometrial adenocarcinoma and could be due to epigenetic silencing (Mhawech-Fauceglia et al., 2008). FOLH1 likely arose from a duplication event of a nearby chromosomal region. Alternative splicing gives rise to multiple transcript variants encoding several different isoforms (Watt et al., 2001; Zink et al., 2020). Our research found ES in FOLH1 was associated with the OS of UCEC patients (p = 0.0000). What's more, ES was also the most frequent splice type among the seven AS types (34.4%) in UCEC. If the normal exon can be restored into the exon of ES occurred, it will bring hope to the treatment of many diseases (Verhaart and Aartsma-Rus, 2019). However, the mechanism of ES in FOLH1 leading to a high stage and poor prognosis of UCEC is unknown.

Next, two immune checkpoint genes, Cytotoxic Lymphocyte Antigen 4 (CTLA-4) and Programmed Cell Death 1 (PDCD1), showed negative correlations with the risk score of AS in UCEC. CTLA-4 is expressed on the surface of naive effector T cells and Tregs (Billeskov et al., 2017; Menéndez-Menéndez et al., 2019). Based on its role as a negative regulator of T cell activation, CTLA-4 has become an attractive target for therapies aiming to enhance the effector activity of T lymphocytes. The first targeted drug for CTLA-4, ipilimumab, was approved by the Food and Drug Administration (FDA) in 2011 to treat melanoma (Lipson and Drake, 2011). At present, both nivolumab and ipilimumab are undergoing phase II clinical trials in UCEC (Grywalska et al., 2019). In our study, the CTLA-4 gene expression, the number and immune score of Tregs all decreased in the high-risk score group, which predicted a worse prognosis. Therefore, we supposed that a high-risk score of AS might be related to the decreased immune activity of Treg cells and the low expression of CTLA-4. It is possible that the targeted regulation of AS can improve the immune activity of Treg cells and increase the expression of CTLA-4, which may be valuable in improving the survival rate of UCEC patients, although further confirmation is needed.

PDCD1, also known as PD-1, functions primarily in peripheral tissues. It is expressed on the surface of activated T cells, Tregs, activated B cells, and NK cells (Page et al., 2014). In 2014, the first FDA-approved immune checkpoint inhibitor targeting PD-1 was nivolumab (Grywalska et al., 2019). During the 2015 annual meeting of the Society of Gynecologic Oncology, Herzog et al. reported that the highest PD-1 expression rates among studied cancer types were in EC (75.2%) (Page et al., 2014). We found that PDCD1 expression was suppressed in the high-risk score group, and the 5-year survival rate was lower than that in the low-risk score group. It has been confirmed that there are variable splicing events in the PD1 gene (Nielsen et al., 2005; Wang et al., 2021). Another research considered the AS events in PD-1 may be a novel source for diagnostic and therapeutic target on celiac disease (Ponce de León et al., 2021). Why the high-risk AS events in PDCD1 lead to a worse prognosis of endometrial cancer needs further study.

The current study also has several limitations that should be noted. Firstly, this study is based on bioinformatics analysis, and there are no recruited cohorts for prognostic verification. Secondly, the values of the two-gene signatures for immunotherapeutic drugs prediction have not been verified in patient cohorts.

CONCLUSION

This study assessed the heterogeneity of tumor-infiltrating immune cells in UCEC and identified two AS-related genes, CYB561 and FOLH1, from the 11 genes involved in the AS prognostic model. Two immune checkpoint genes, CTLA4 and PDCD1, were negatively correlated with the risk score. The outcomes of this study are significant for investigating the immune-related mechanisms of tumor progression and exploring novel prognostic predictors and precise therapy methods.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

DX conceived and designed the study with XZ. DS drafted the manuscript and analyzed the data. AZ and BG handled the picture and article format. LZ and HH reviewed the data. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Hunan Science and Technology Department (Grant No. 2020 SK4017), the National Key Research and Development Program of China (Grant No. 2018YFC1004800), and the Hunan Provincial Clinical Medical Technology Innovation Guiding Project (Grant Nos 2020SK53605 and 2020SK53606). The results of this study are based upon data generated by the TCGA database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.770569/full#supplementary-material

Supplementary Figure S1 | The correlation between the risk score and the checkpoint genes PDCD1 and CTLA4.

Supplementary Figure S2 | The association between CYB561 expression and the tumor-infiltrating immune cells in UCEC.

Supplementary Figure S3 | The association between FOLH1 expression and the tumor-infiltrating immune cells in UCEC.

Supplementary Figure S4 | Relationship between the gene copy number of CYB561 and the quantity of six immune cells using the TIMER database.

Supplementary Figure S5 | Relationship between the gene copy number of FOLH1 and the quantity of six immune cells using the TIMER database.

REFERENCES

Baralle, F. E., and Giudice, J. (2017). Alternative Splicing as a Regulator of Development and Tissue Identity. *Nat. Rev. Mol. Cell. Biol.* 18, 437–451. doi:10.1038/nrm.2017.27

Billeskov, R., Wang, Y., Solaymani-Mohammadi, S., Frey, B., Kulkarni, S., Andersen, P., et al. (2017). Low Antigen Dose in Adjuvant-Based Vaccination Selectively Induces CD4 T Cells with Enhanced Functional Avidity and Protective Efficacy. *J. Immunol.* 198 (9), 3494–3506. doi:10.4049/jimmunol.1600965

Bonnal, S. C., López-Oreja, I., and Valcárcel, J. (2020). Roles and Mechanisms of Alternative Splicing in Cancer - Implications for Care. *Nat. Rev. Clin. Oncol.* 17 (8), 457–474. doi:10.1038/s41571-020-0350-x

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J. Clin.* 68 (6), 394–424. doi:10.3322/caac.21492

Buratti, E., Baralle, M., and Baralle, F. E. (2006). Defective Splicing, Disease and Therapy: Searching for Master Checkpoints in Exon Definition. *Nucleic Acids Res.* 34 (12), 3494–3510. doi:10.1093/nar/gkl498

Climente-González, H., Porta-Pardo, E., Godzik, A., and Eyras, E. (2017). The Functional Impact of Alternative Splicing in Cancer. *Cell Rep.* 20 (9), 2215–2226. doi:10.1016/j.celrep.2017.08.012

Couzin-Frankel, J. (2013). Breakthrough of the Year 2013. Cancer Immunotherapy. *Science* 342 (6165), 1432–1433. doi:10.1126/science.342.6165.1432

Date, A. A., Rais, R., Babu, T., Ortiz, J., Kanvinde, P., Thomas, A. G., et al. (2017). Local Enema Treatment to Inhibit FOLH1/GCPII as a Novel Therapy for Inflammatory Bowel Disease. *J. Control. Release* 263, 132–138. doi:10.1016/j.jconrel.2017.01.036

de Necochea-Campion, R., Shouse, G. P., Zhou, Q., Mirshahidi, S., and Chen, C.-S. (2016). Aberrant Splicing and Drug Resistance in AML. *J. Hematol. Oncol.* 9 (1), 85. doi:10.1186/s13045-016-0315-9

Demircioğlu, D., Cukuroglu, E., Kindermans, M., Nandi, T., Calabrese, C., Fonseca, N. A., et al. (2019). A Pan-Cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell* 178 (6), 1465–1477.e17. doi:10.1016/j.cell.2019.08.018

Dou, X. Q., Chen, X. J., Wen, M. X., Zhang, S. Z., Zhou, Q., and Zhang, S. Q. (2020). Alternative Splicing of VEGFA Is Regulated by RBM10 in Endometrial Cancer. *Kaohsiung J. Med. Sci.* 36 (1), 13–19. doi:10.1002/kjm2.12127

Dou, X. Q., Chen, X. J., Zhou, Q., Wen, M. X., Zhang, S. Z., and Zhang, S. Q. (2020). miR-335 Modulates Numb Alternative Splicing via Targeting RBM10 in Endometrial Cancer. *Kaohsiung J. Med. Sci.* 36 (3), 171–177. doi:10.1002/kjm2.12149

Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R. K. (2016). RNA Splicing Factors as Oncoproteins and Tumour Suppressors. *Nat. Rev. Cancer* 16 (7), 413–430. doi:10.1038/nrc.2016.51

Frankiw, L., Baltimore, D., and Li, G. (2019). Alternative mRNA Splicing in Cancer Immunotherapy. *Nat. Rev. Immunol.* 19 (11), 675–687. doi:10.1038/s41577-019-0195-7

Gamazon, E. R., and Stranger, B. E. (2014). Genomics of Alternative Splicing: Evolution, Development and Pathophysiology. *Hum. Genet.* 133 (6), 679–687. doi:10.1007/s00439-013-1411-3

Gilbert, W. (1978). Why Genes in Pieces? *Nature* 271 (5645), 271501–501. doi:10.1038/271501a0

Grywalska, E., Sobstyl, M., Putowski, L., and Roliński, J. (2019). Current Possibilities of Gynecologic Cancer Treatment with the Use of Immune Checkpoint Inhibitors. *Int. J. Mol. Sci.* 20 (19), 4705. doi:10.3390/ijms20194705

He, Y., Jiang, Z., Chen, C., and Wang, X. (2018). Classification of Triple-Negative Breast Cancers Based on Immunogenomic Profiling. *J. Exp. Clin. Cancer Res.* 37 (1), 327. doi:10.1186/s13046-018-1002-1

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* 56 (2), 337–344. doi:10.1111/j.0006-341x.2000.00337.x

Kahles, A., Lehmann, K. V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., et al. (2018). Comprehensive Analysis of Alternative Splicing across Tumors from 8,705 Patients. *Cancer Cell* 34, 211–224. doi:10.1158/0008-5472.can-04-1910

Li, H., Liu, J., Shen, S., Dai, D., Cheng, S., Dong, X., et al. (2020). Pan-cancer Analysis of Alternative Splicing Regulator Heterogeneous Nuclear Ribonucleoproteins (hnRNPs) Family and Their Prognostic Potential. *J. Cell. Mol. Med.* 24 (19), 11111–11119. doi:10.1111/jcmm.15558

Lipson, E. J., and Drake, C. G. (2011). Ipilimumab: an Anti-CTLA-4 Antibody for Metastatic Melanoma. *Clin. Cancer Res.* 17 (22), 6958–6962. doi:10.1158/1078-0432.CCR-11-1595

Liu, X., Liu, C., Liu, J., Song, Y., Wang, S., Wu, M., et al. (2021). Identification of Tumor Microenvironment-Related Alternative Splicing Events to Predict the Prognosis of Endometrial Cancer. *Front. Oncol.* 11, 645912. doi:10.3389/fonc.2021.645912

Menéndez-Menéndez, J., Hermida-Prado, F., Granda-Díaz, R., González, A., García-Pedrero, J. M., Del-Río-Ibisate, N., et al. (2019). Deciphering the Molecular Basis of Melatonin Protective Effects on Breast Cells Treated with Doxorubicin: TWIST1 a Transcription Factor Involved in EMT and Metastasis, a Novel Target of Melatonin. *Cancers* 11 (7), 1011. doi:10.3390/cancers11071011

Mhawech-Fauceglia, P., Smiraglia, D. J., Bshara, W., Andrews, C., Schwaller, J., South, S., et al. (2008). Prostate-specific Membrane Antigen Expression Is a Potential Prognostic Marker in Endometrial Adenocarcinoma. *Cancer Epidemiol. Biomarkers Prev.* 17 (3), 571–577. doi:10.1158/1055-9965.EPI-07-0511

Morice, P., Leary, A., Creutzberg, C., Abu-Rustum, N., and Darai, E. (2016). Endometrial Cancer. *Lancet* 387 (10023), 1094–1108. doi:10.1016/S0140-6736(15)00130-0

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Nielsen, C., Ohm-Laursen, L., Barington, T., Husby, S., and Lillevang, S. T. (2005). Alternative Splice Variants of the Human PD-1 Gene. *Cell. Immunol.* 235 (2), 109–116. doi:10.1016/j.cellimm.2005.07.007

Obeng, E. A., Stewart, C., and Abdel-Wahab, O. (2019). Altered RNA Processing in Cancer Pathogenesis and Therapy. *Cancer Discov.* 9 (11), 1493–1510. doi:10.1158/2159-8290.cd-19-0399

Page, D. B., Postow, M. A., Callahan, M. K., Allison, J. P., and Wolchok, J. D. (2014). Immune Modulation in Cancer with Antibodies. *Annu. Rev. Med.* 65, 185–202. doi:10.1146/annurev-med-092012-112807

Ponce de León, C., Lorite, P., López-Casado, M. Á., Barro, F., Palomeque, T., and Torres, M. I. (2021). Significance of PD1 Alternative Splicing in Celiac Disease as a Novel Source for Diagnostic and Therapeutic Target. *Front. Immunol.* 12, 678400. doi:10.3389/fimmu.2021.678400

Popli, P., Richters, M. M., Chadchan, S. B., Kim, T. H., Tycksen, E., Griffith, O., et al. (2020). Splicing Factor SF3B1 Promotes Endometrial Cancer Progression via Regulating KSR2 RNA Maturation. *Cell Death Dis.* 11 (10), 842. doi:10.1038/s41419-020-03055-y

Ryan, M., Wong, W. C., Brown, R., Akbani, R., Su, X., Broom, B., et al. (2016). TCGASpliceSeq a Compendium of Alternative mRNA Splicing in Cancer. *Nucleic Acids Res.* 44 (D1), D1018–D1022. doi:10.1093/nar/gkv1288

van den Berg, M. P., Almomani, R., Biaggioni, I., van Faassen, M., van der Harst, P., Silljé, H. H. W., et al. (2018). Mutations in CYB561 Causing a Novel Orthostatic Hypotension Syndrome. *Circ. Res.* 122 (6), 846–854. doi:10.1161/CIRCRESAHA.117.311949

Venables, J. P. (2004). Aberrant and Alternative Splicing in Cancer. *Cancer Res.* 64, 7647–7654. doi:10.1093/jmcb/mjz033

Verhaart, I. E. C., and Aartsma-Rus, A. (2019). Therapeutic Developments for Duchenne Muscular Dystrophy. *Nat. Rev. Neurol.* 15 (7), 373–386. doi:10.1038/s41582-019-0203-3

Wang, B.-D., and Lee, N. (2018). Aberrant RNA Splicing in Cancer and Drug Resistance. *Cancers* 10 (11), 458. doi:10.3390/cancers10110458

Wang, C., Weng, M., Xia, S., Zhang, M., Chen, C., Tang, J., et al. (2021). Distinct Roles of Programmed Death Ligand 1 Alternative Splicing Isoforms in Colorectal Cancer. *Cancer Sci.* 112 (1), 178–193. doi:10.1111/cas.14690

Wang, C., Zheng, M., Wang, S., Nie, X., Guo, Q., Gao, L., et al. (2019). Whole Genome Analysis and Prognostic Model Construction Based on Alternative Splicing Events in Endometrial Cancer. *BioMed Res. Int.* 2019, 2686875. doi:10.1155/2019/2686875

Wang, Q., Xu, T., Tong, Y., Wu, J., Zhu, W., Lu, Z., et al. (2019). Prognostic Potential of Alternative Splicing Markers in Endometrial Cancer. *Mol. Ther. - Nucleic Acids* 18, 1039–1048. doi:10.1016/j.omtn.2019.10.027

Watt, F., Martorana, A., Brookes, D. E., Ho, T., Kingsley, E., O'Keefe, D. S., et al. (2001). A Tissue-Specific Enhancer of the Prostate-Specific Membrane Antigen Gene, FOLH1. *Genomics* 73 (3), 243–254. doi:10.1006/geno.2000.6446

Wernicke, A. G., Kim, S., Liu, H., Bander, N. H., and Pirog, E. C. (2017). Prostate-specific Membrane Antigen (PSMA) Expression in the Neovasculature of Gynecologic Malignancies: Implications for PSMA-Targeted Therapy. *Appl. Immunohistochem. Mol. Morphol.* 25 (4), 271–276. doi:10.1097/PAI.0000000000000297

Willis, S., Villalobos, V. M., Gevaert, O., Abramovitz, M., Williams, C., Sikic, B. I., et al. (2016). Single Gene Prognostic Biomarkers in Ovarian Cancer: A Meta-Analysis. *PLoS One* 11 (2), e0149183. doi:10.1371/journal.pone.0149183

Yang, X., Zhao, Y., Shao, Q., and Jiang, G. (2021). Cytochrome B561 Serves as a Potential Prognostic Biomarker and Target for Breast Cancer. *Int. J. Gen. Med.* 14, 10447–10464. doi:10.2147/IJGM.S338878

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Zink, C. F., Barker, P. B., Sawa, A., Weinberger, D. R., Wang, M., Quillian, H., et al. (2020). Association of Missense Mutation in FOLH1 with Decreased NAAG Levels and Impaired Working Memory Circuitry and Cognition. *Am. J. Psychiatry* 177 (12), 1129–1139. doi:10.1176/appi.ajp.2020.19111152

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 12 August 2022
doi: 10.3389/fgene.2022.850892

The Causal Evidence of Birth Weight and Female-Related Traits and Diseases: A Two-Sample Mendelian Randomization Analysis

Renke He[1†], *Rui Liu*[2,3†], *Haiyan Wu*[2,3], *Jiaen Yu*[2,3], *Zhaoying Jiang*[1] *and Hefeng Huang*[2,3,4,5*]

[1]International Institutes of Medicine, The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, China, [2]Department of Reproductive Endocrinology, Women's Hospital, School of Medicine, Zhejiang University, Hangzhou, China, [3]Key Laboratory of Reproductive Genetics, Ministry of Education, School of Medicine, Zhejiang University, Hangzhou, China, [4]Shanghai Frontiers Science Center of Reproduction and Development, Shanghai, China, [5]Research Units of Embryo Original Diseases, Chinese Academy of Medical Sciences, Shanghai, China

Objectives: A large meta-analysis indicated a more pronounced association between lower birth weight (BW) and diseases in women but less concern about the causality between BW and female-related phenotypes and diseases.

Methods: Mendelian randomization (MR) analysis was used to estimate the causal relationship between two traits or diseases using summary datasets from genome-wide association studies. Exposure instrumental variables are variants that are strongly associated with traits and are tested using four different statistical methods, including the inverse variance weighting, MR-Egger, weighted median, and weighted mode in MR analysis. Next, sensitivity analysis and horizontal pleiotropy were assessed using leave-one-out and MR-PRESSO packages.

Results: The body mass index (BMI) in adulthood was determined by BW (corrected β = 0.071, p = 3.19E-03). Lower BW could decrease the adult sex hormone-binding globulin (SHBG) level (β = –0.081, p = 2.08E-06), but it resulted in increased levels of bioavailable testosterone (bio-T) (β = 0.105, p = 1.25E-05). A potential inverse effect was observed between BW and menarche (corrected β = –0.048, p = 4.75E-03), and no causal association was confirmed between BW and the risk of endometriosis, leiomyoma, and polycystic ovary syndrome.

Conclusion: Our results suggest that BW may play an important role and demonstrates a significant direct influence on female BMI, SHBG and bio-T levels, and menarche.

Keywords: birthweight, reproductive hormones, body mass index, menarche, leiomyoma, Mendelian randomization

OPEN ACCESS

Edited by:
Aparna Vasanthakumar, AbbVie, United States

Reviewed by:
Shixiong Zhang, Xidian University, China
Ameya S. Kulkarni, AbbVie, United States

***Correspondence:**
Hefeng Huang
hhf57@zju.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *08 January 2022*
Accepted: *16 June 2022*
Published: *12 August 2022*

Citation:
He R, Liu R, Wu H, Yu J, Jiang Z and Huang H (2022) The Causal Evidence of Birth Weight and Female-Related Traits and Diseases: A Two-Sample Mendelian Randomization Analysis. Front. Genet. 13:850892. doi: 10.3389/fgene.2022.850892

INTRODUCTION

The hypothesis of "developmental origins of adult disease" was first stated by Barker (Barker and Osmond, 1986) in the 20th century, which mainly explained that the adverse influences in the early developmental period could cause permanent changes in physiology and metabolism, which finally results in an increased risk of disease in adulthood. Thus, birth weight (BW) is widely used as an indicator of exposure during the intrauterine period and early life development

(Peck et al., 2003; Scharf et al., 2016). Numerous observational studies provided evidence for the correlation between reduced BW and increased risk of adult diseases, such as type 2 diabetes mellitus (T2DM) (Carlsson et al., 1999; Whincup et al., 2008), coronary heart disease (CHD) (Ferrie et al., 2006; Morley et al., 2006), hypertension (Eriksson et al., 2000a; Tamakoshi et al., 2006), and stroke (Eriksson et al., 2000b). In particular, it is worth noting that only women demonstrated an increased risk of T2DM and CHD with a raised BW in a recent sex-specific binary meta-analysis (Knop et al., 2018), indicating that BW is more acceptable in predicting the correlation between several traits and diseases in women. Indeed, early observational studies provided controversial evidence supporting the association between BW and female-related traits, including female-only body mass index (BMI) (Zhao et al., 2012; Jelenkovic et al., 2017), reproductive hormones (estradiol [E_2] (Jasienska et al., 2006; Espetvedt Finstad et al., 2009), testosterone (Ruder et al., 2011), anti-Mullerian hormone [AMH] (Dior et al., 2021)), menarche (Juul et al., 2017; Fan et al., 2018), menopause (Tom et al., 2010; Bjelland et al., 2020), and female-specific diseases (polycystic ovaries syndrome [PCOS] (Cresswell et al., 1997), endometriosis (Olšarová and Mishra, 2020), and leiomyomata (Wise et al., 2012)), which can influence women's reproductive health and life expectancy. However, whether the identified correlation between BW and these female-related traits represents a truly causal relationship remains uncertain because of bias, pleiotropy, or common confounders during prenatal life (Ruiz-Narváez et al., 2014; Kahn et al., 2017; Lawlor et al., 2017).

Two-sample Mendelian randomization (TSMR), a novel and popular analysis tool, was used to estimate the causal inference in observational studies, which avoided all possible and potential biases from confounding factors. The fundamental theory of TSMR is that during the period when gametes were formatted and combined, the alleles of genetic variants were segregated randomly based on Mendel's law, which led to their independence with confounding factors such as the environment, age, and sex. To some extent, this implies that the TSMR results are stable and convincing.

In recent years, Mendelian randomization (MR) studies provided evidence of a positive association between lower birth weight (LBW) and T2DM (Huang et al., 2019) and stroke (Wang et al., 2020), a negative association with chronic kidney disease (Yu et al., 2020), and no relationship with asthma (Zeng et al., 2019). Furthermore, no related or specific reports focused on women's health and diseases exist. Here, a large TSMR analysis was conducted to comprehensively estimate the causality of BW on eight related traits and three common reproductive endocrine diseases in adulthood. Our results remained statistically significant and robust after validating the heterogeneity, sensitivity, and horizontal pleiotropy.

MATERIALS AND METHODS

Data Sources

First, a genome-wide association studies (GWAS) of female BMI was obtained from a large genome-wide meta-analysis combining summary data from the United Kingdom Biobank and the GIANT consortium (European ancestry, n = 143,677) (Pulit et al., 2019). The summary datasets of reproductive hormones in women were identified from the United Kingdom Biobank (European ancestry, total testosterone (TT) level, n = 230,454; bioavailable testosterone (bio-T), n = 188,507; sex hormone-binding globulin (SHBG) level, n = 189,473; E_2 level, n = 163,985) (Ruth et al., 2020; Schmitz et al., 2021). The summarized statistics for the AMH was collected from a genome-wide meta-analysis including five cohorts and 3,344 premenopausal women (Ruth et al., 2019). In addition, other traits closely related to the female sex were age at menarche (AAM) and menopause, which were sourced from the MER-IEU Consortium and included 243,944 and 211,114 women, respectively. To conclude, we decided on three common female-specific reproductive endocrine diseases—endometriosis, leiomyoma, and PCOS—to evaluate their causal relationship. The GWAS outcome of endometriosis and leiomyoma was obtained from the FinnGen biobank, recruiting 6,502 cases and 57,407 controls and 14,569 cases and 72,789 controls, respectively. The population of PCOS patients was determined through a large-scale meta-analysis, including six studies and 24,267 samples (Day et al., 2018). The detailed information and characteristics of the GWAS outcomes are listed in **Supplementary Table S1**.

Selection of Instrumental Variables

First, the plinked version of 47 independent single-nucleotide polymorphisms (SNPs) were identified as instrumental variables (IVs) representing interest exposure (e.g., BW) to perform MR analysis, which showed a strong association with statistical significance (p < 5.0E-08) based on early growth genetics (EGG) consortium research (Zeng and Zhou, 2019) (Table 1). Up to now, the EGG consortium study is the largest GWAS on BW (a continuous trait) and contains 16,245,523 imputed SNPs based on 153,781 infants collected from more than 30 studies (**Supplementary Table S2**). Another different version of the 48 SNPs was used to validate the robustness of the results (Horikoshi et al., 2016) (**Supplementary Table S3**). Then, all IVs were independent after performing the clumping procedure (R^2 = 0.001, kb = 10,000) and removing the linkage disequilibrium between SNPs. Third, the F-statistics has been applied to ensure the sufficient power of IVs in the MR analysis, and the results proved strong effect sizes with overall F-statistics > 10. The SNP of IVs for lower BW was presented by supplying a negative sign on the estimated BW effect (Zeng and Zhou, 2019). To conclude, all the above procedures were run in the R software (version 4.0.3) using the "TwoSampleMR" package to automatically prune SNPs with linkage dependence.

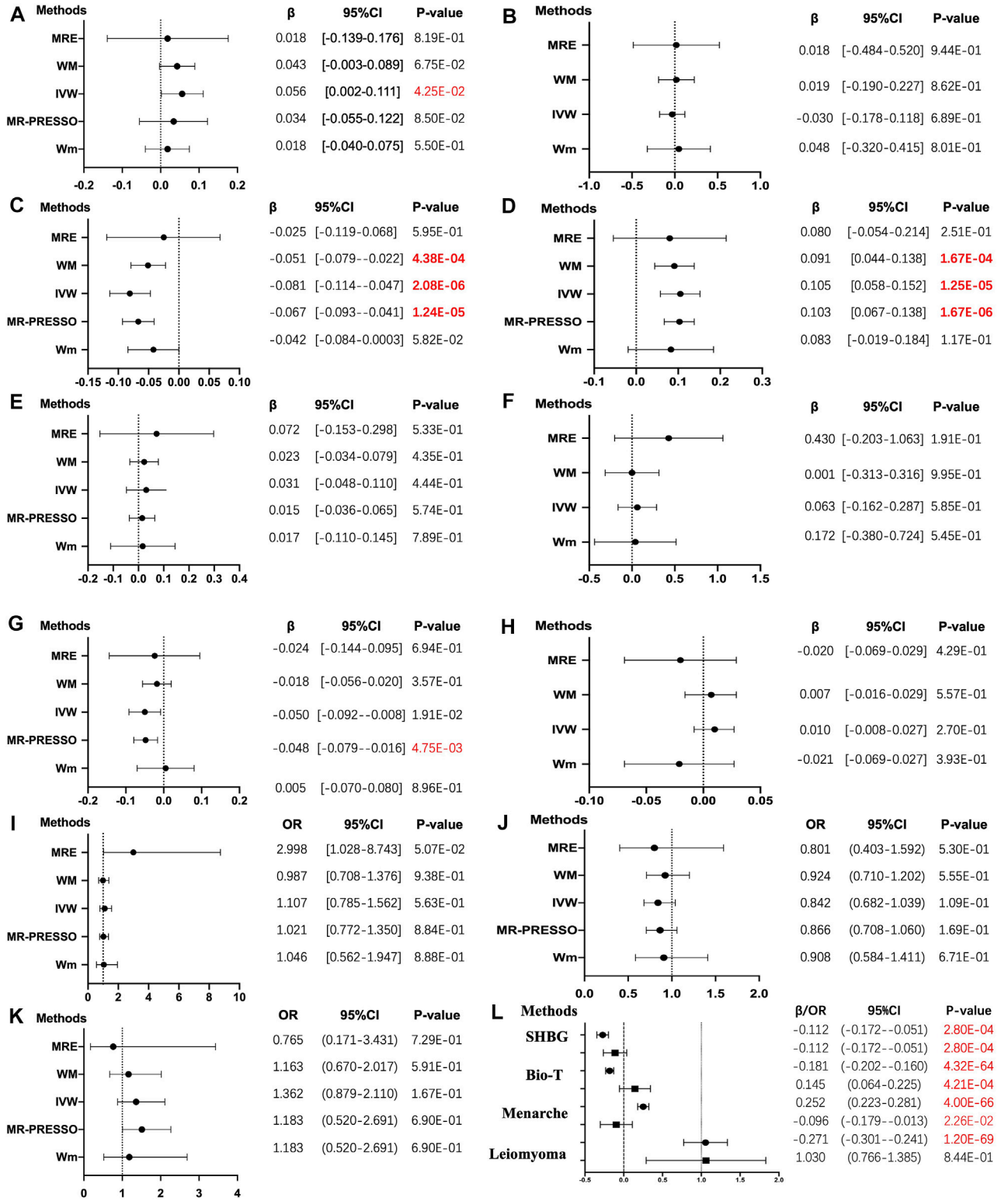

FIGURE 1 | The results of four different methods of Mendelian randomization (MR) analysis. (The MR analysis showing the effect of the exposure SNPs on the outcomes. **(A–K)**: **(A)** body mass index, BMI; **(B)** estradiol, E_2; **(C)** sex hormone-binding globulin, SHBG; **(D)** bioavailable testosterone, bio-T; **(E)** total testosterone, TT; **(F)** anti-Mullerian hormone, AMH; **(G)** menarche; **(H)** menopause; **(I)** endometriosis; **(J)** leiomyoma; **(K)** polycystic ovarian syndrome, PCOS. **(L)** in the multivariable MR analysis, each trait with two results is presented. The solid dot means the causal effects of LBW on traits, whereas the square means BMI on traits in the MVMR. MRE, MR-Egger; WM, weighted median; Wm, weighted mode; IVW, inverse variance weighting; MR-PRESSO, MR-Egger and Mendelian Randomization Pleiotropy RESidual Sum and Outlier. The results of the continuous outcomes are presented by β [95% CI], whereas the binary outcomes are shown by OR [95% CI]. Numbers in red mean *p*-values < 5.00E-02 and red and bold font means *p*-values < 4.55E-03)

Multivariable Mendelian Randomization

The inverse variance weighting method was applied in the two-sample multivariable MR (MVMR), which fits multiple risk factors as exposures (e.g., fetal body weight and BMI in our study), to simultaneously estimate their genetically predicted effects on an outcome (e.g., concentration of SHBG and bio-T, menarche). This analysis allowed us to estimate the direct effect of LBW (i.e., the effect after accounting for adult BMI) and its indirect effect (i.e., the effect mediated by BMI in adulthood) on each female trait. To evaluate the causal effects of BMI-adjusted LBW in our study, MVMR analysis was performed, which included SNPs that reached genome-wide significance ($p <$ 5.00E-8) in both GWAS of LBW and BMI. For these two exposures, we used nonoverlapping populations. After excluding SNPs with a pairwise $R^2 > 0.001$, 966 independent SNPs were used as IVs in the analysis. Then, MVMR analysis was conducted using both the MVMR and TSMR packages in the R software.

Statistical Analysis

Two-sample MR was applied to the GWAS data in our study. We chose the IVW random-effects model as the main tool to estimate causal associations based on GWAS data for BW and female-related traits and diseases. Next, we performed an estimation using three other methods—MR-Egger (MRE), weighted median (WM), and weighted mode—ensuring the stability and reliability of the results. Also, we measured the causal effect heterogeneity using Cochran's Q test and I^2 statistics, and the "leave-one-out" sensitivity analysis was performed to ascertain whether the heterogeneity was caused by specific SNPs. The MRE and Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MR-PRESSO) analysis (Verbanck et al., 2018) were conducted to eliminate the bias caused by horizontal pleiotropy, outlier SNPs were identified using the MR-PRESSO analysis, and the results were corrected. Further, all results are presented in forest plots, scatterplots, leave-one-out plots, and funnel plots. All procedures were repeated using another version of the exposure SNPs. In general, p-values < 0.05 were considered statistically significant, but in multiple testing, the p-value threshold was adjusted through Bonferroni correction ($p <$ 0.05/11 = 4.55E-03). If the outcomes are continuous variables, the estimated effects are exhibited as a beta effect (β), standard error (se), and p-value. They are presented as odds ratios (Ors) with 95% confidence intervals (Cis). Also, the R software and "TwoSampleMR" package were used for all analyses (Yavorska and Burgess, 2017).

RESULTS

Higher Birth Weight May Determine Higher Body Mass Index in Women

The primary 47-SNP IVW analysis provided suggestive evidence for a positive causal relationship between BW and BMI (β = 0.056, p = 4.25E-02). In addition, similar but more significant results were identified in the 48-SNP IVW analysis (β = 0.071, p = 7.63E-03) (**Figure 1**, **Supplementary Figure S1**).

The Causality Between Lower Birth Weight and Reproductive Hormones

As for reproductive hormones, lower BW demonstrated a positive effect on bio-T levels (β = 0.105, p = 1.25E-05) in the 47-SNP version and the same causality in the validated 48-SNP version (β = 0.103, p = 1.42E-04), but it demonstrated an inverse effect on SHBG concentration (β = −0.081, p = 2.08E-06 versus β = −0.075, p = 9.36E-05). However, no evidence showed an association between lower BW and levels of E_2 (β = −0.030, p = 6.89E-01), TT (β = 0.031, p = 0.444), and AMH (β = 0.063, p = 5.85E-01) (**Figure 1**, **Supplementary Figure S1**).

Lower Birth Weight May Result in Higher Risk of Early Age at Menarche

The results of the IVW analyses showed that lower BW tended to exhibit a negative causal effect on AAM, but it did not reach the corrected p-value of strong significance (β = −0.048, p = 1.90E-02 versus β = −0.053, p = 9.82E-03), whereas no relationship was observed between LBW and age at natural menopause (ANM) (β = 0.010, p = 2.70E-01) (**Figure 1**, **Supplementary Figure S1**).

No Association Was Identified Between Lower Birth Weight and Three Reproductive Endocrine Diseases

No evidence of causal effects was found between a unit lower BW and endometriosis, leiomyoma, and PCOS, even after the heterogeneity and horizontal pleiotropy were eliminated (OR = 1.107; 95% CI = [0.785–1.562], p = 5.63E-01; OR = 0.842; 95% CI = [0.682–1.039], p = 109E-01; OR = 1.362; 95% CI = [0.879–2.110], p = 1.67E-01). However, 48 LBW SNPs showed potential causality with leiomyoma (OR = 0.791; 95% CI = [0.629–0.994], p = 4.46E-02) (**Figure 1**, **Supplementary Figure S1**).

Multivariable Mendelian Randomization

Applying MVMR resulted in the majority of effect estimates identified in the previous analysis being strengthened to include the adjustment for adult BMI. In the MVMR analysis controlling for BMI, more robust evidence was found for a direct and negative causal effect of LBW on SHBG concentration (β = −0.112, 95% CI = [−0.172–0.051]) and AAM (β = −0.096, 95% CI = [−0.179–0.013]) and a positive effect of LBW on bio-T levels (β = 0.145, 95% CI = [0.064–0.225]). Moreover, the weak relationship between LBW and leiomyoma was eliminated in MVMR (OR = 1.030, 95% CI = [0.766–1.385]). The causal relationships estimated from MVMR (including LBW and BMI) were consistent with the univariable IVW analysis (LBW) for SHBG, bio-T, and menarche, except for leiomyoma (**Supplementary Table S6**, **Figure 1L**).

Sensitivity Analysis

Also, the measurement of WM was used to test sensitivity. Similar results proved the negative association between lower BW and SHBG (β= −0.051, p = 4.38E-04) and positive causality

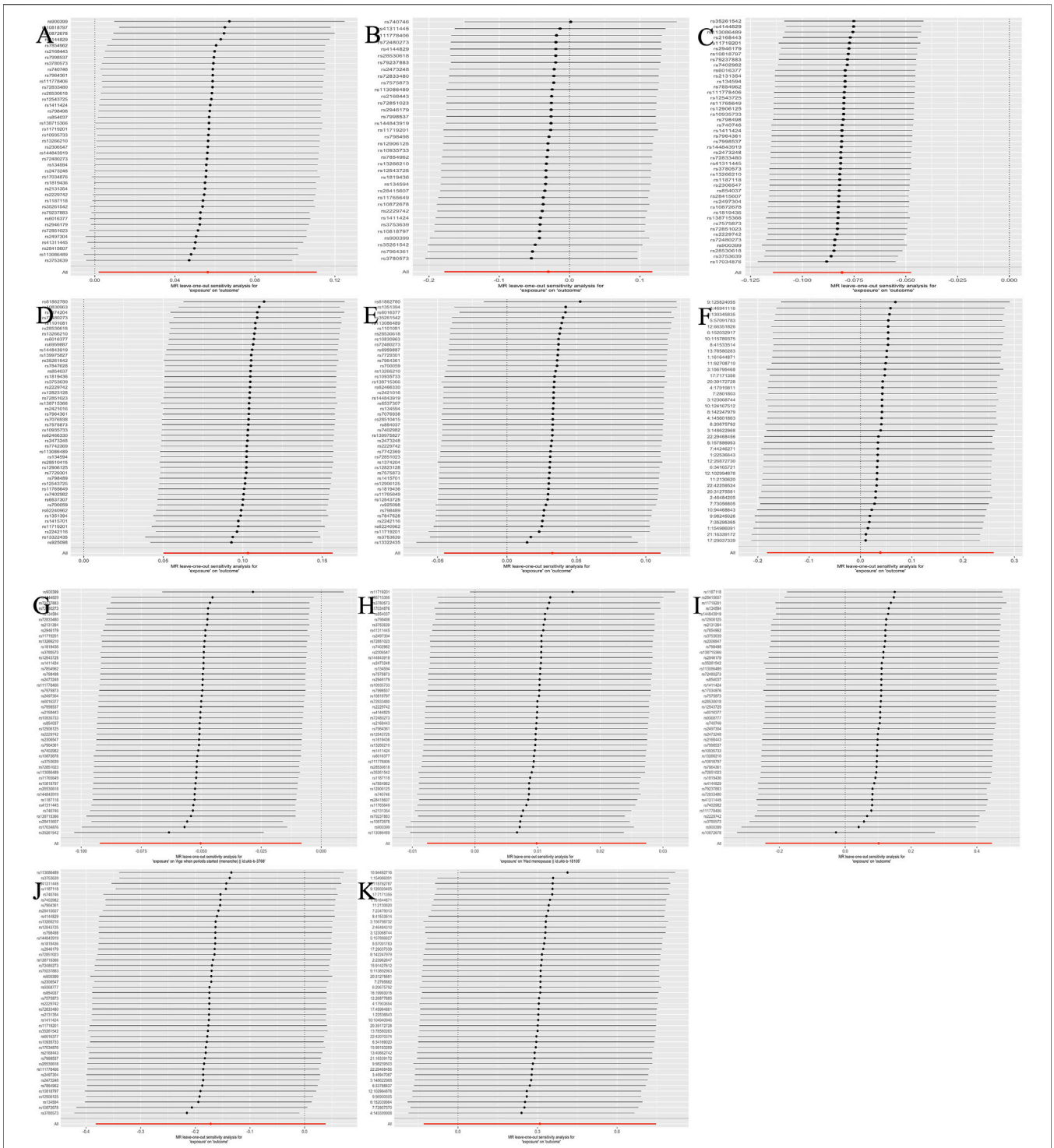

FIGURE 2 | The leave-one-out analysis plot (The estimation effects are reported per SD increase in the exposure, and error bars represent 95% confidence intervals).

with bio-T (β= 0.091, p = 1.67E-04). Next, conversely, the correlation between BW and BMI was not consistent with our previous findings. All results of the MRE intercept were close to zero and $p > 0.05$, which suggested no horizontal pleiotropy. Owing to the existing heterogeneity, a leave-one-out analysis was applied and presented in the plots (**Figure 2**, **Supplementary Figure S2**). Next, the horizontal line and black points in the leave-one-out plot of TT, E_2, AMH, menopause,

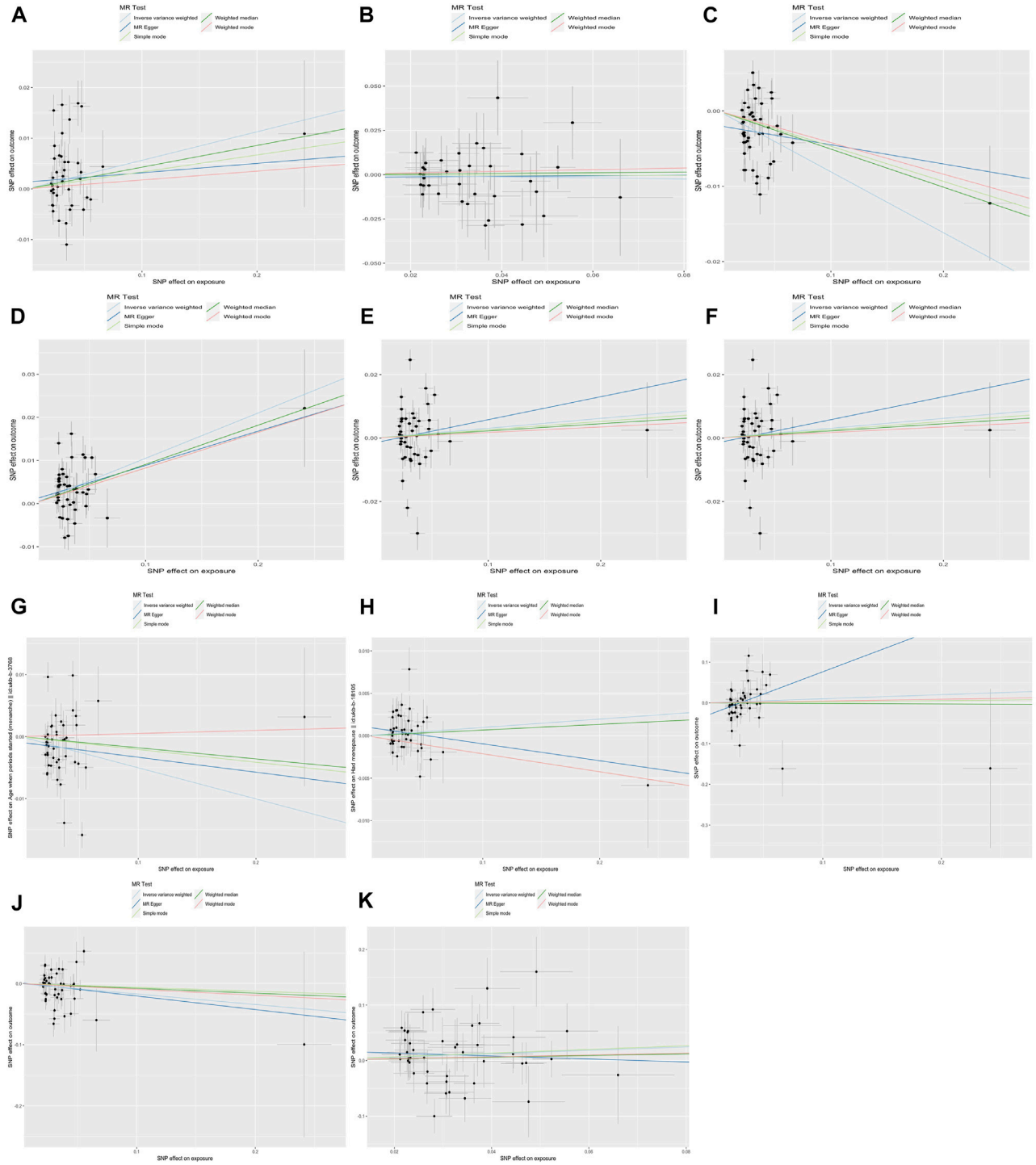

FIGURE 3 | The results of the scatter plot.

endometriosis, leiomyoma, and PCOS crossed the zero line, suggesting potential heterogeneity. The scatterplot, forest plot, and funnel plot are shown in **Figures 3–5** and **Supplementary Figures S3–S5**. Then, we performed MR-PRESSO analysis to identify outlier SNPs and corrected the primary results. In the causal relationship between BW and BMI, rs1374204, rs2150052, rs12823128, and rs2229742 were identified as outliers, and the corrected results reached statistical significance (p = 3.19E-03). After removing rs17034876, rs1187118, rs11765649, rs1411424, rs10818797, rs2497304,

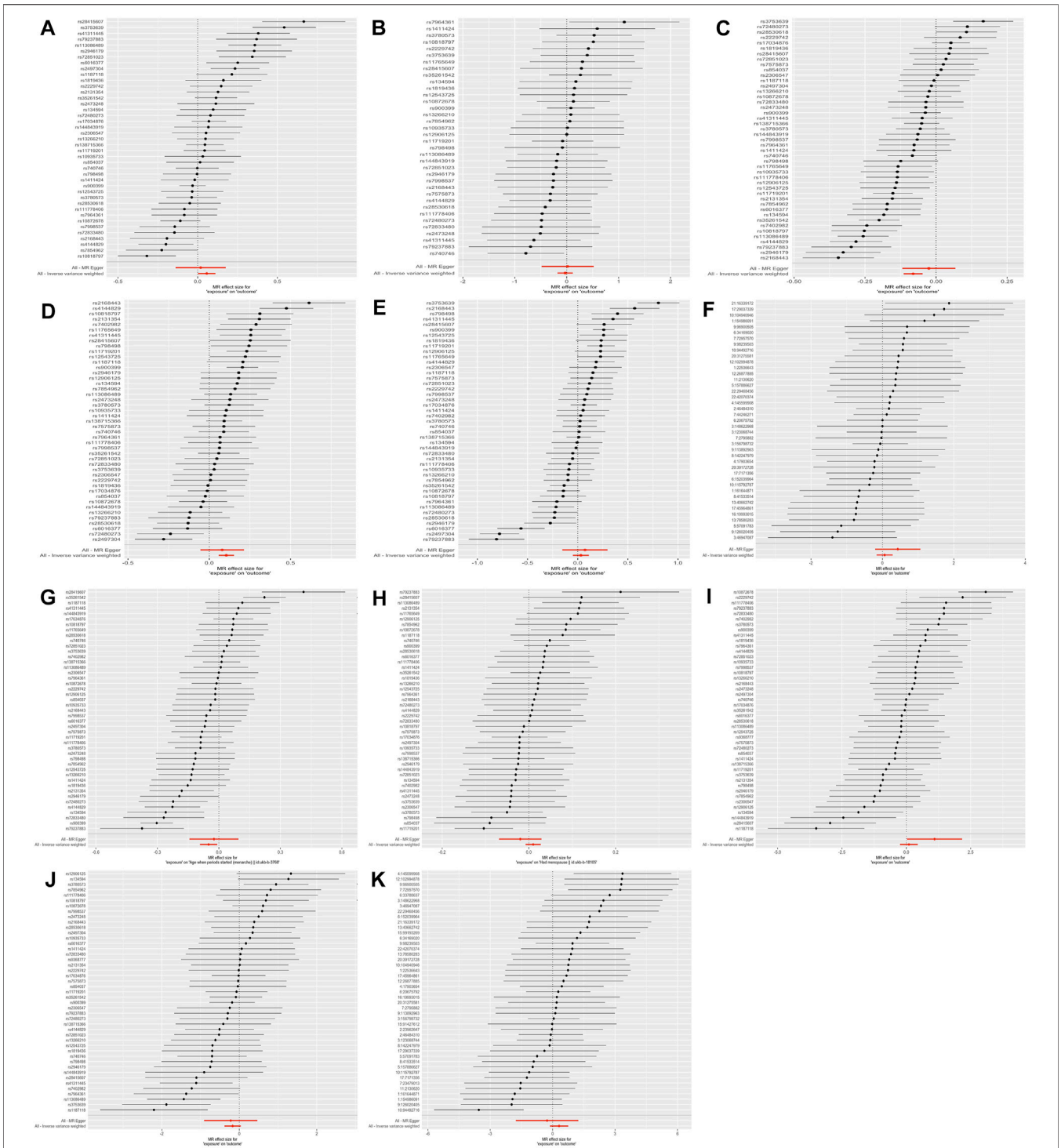

FIGURE 4 | The results of the forest plot.

rs72851023, rs7964361, and rs144843919, a negative relationship was observed between lower BW and SHBG (p = 1.24E-05). Outlier SNPS (rs11765649, rs12543725, rs72851023, and rs7964361) were deleted in the MR analysis of lower BW and bio-T, and the results were not altered (p = 1.67E-06). The rest of the sensitivity and MR-PRESSO analyses are shown in **Supplementary Tables S3–S6**. The causal effect of each SNP on the outcome is presented in **Supplementary Tables S7–S17**.

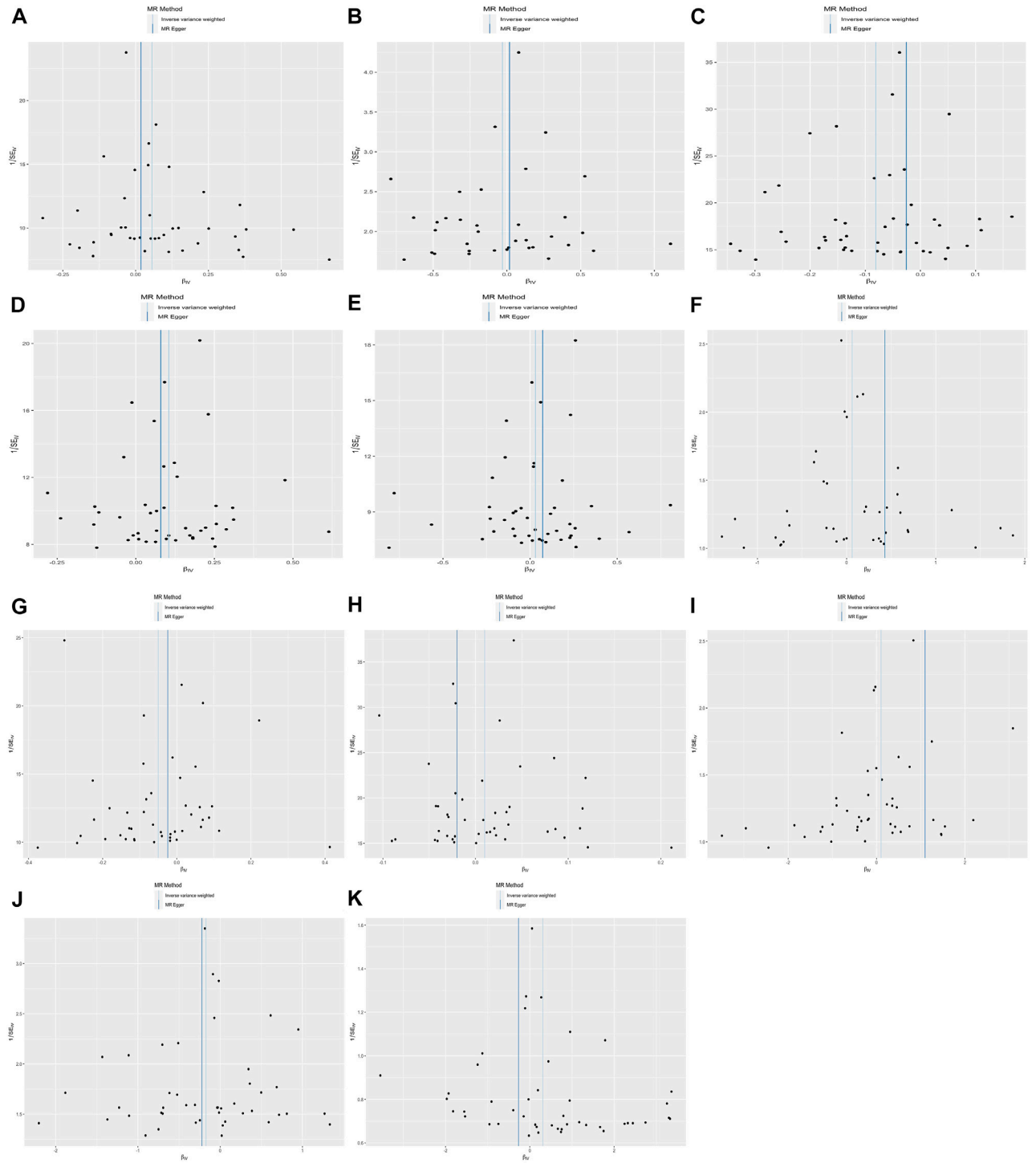

FIGURE 5 | The results of the funnel plot.

DISCUSSION

The present MR study clarified the genetic association between BW and female-related traits in the largest sample size of the European population. In this study, we found a positive effect of lower BW on bio-T, whereas LBW demonstrated an adverse effect on SHBG level. In contrast, we also found a causal effect of BW on BMI and lower BW on menarche but no detrimental effects of LBW on female-specific diseases. To the best of our knowledge, this is the first study to examine the likely causal

relationship between BW and female traits and diseases based on hereditary information.

Our research results on infant BW and adult BMI are similar to those of most existing studies. Rogers (2003) and Zhao et al. (2012) provided good evidence of an association between high BW and subsequent BMI and an increased risk of overweight in young adults. In addition, Jelenkovic et al. (2017) and Liao et al. (2020) suggested a positive association between high BW and a later high BMI. For each individual, a 1.0 kg of BW increased, accompanied with a 0.33 or 0.9 kg/m^2 of BMI increase in adulthood, respectively ($p < 0.001$). Moreover, factors such as genetics, development, and environment could result in individual variations in the concentrations of reproductive hormones. The association between lower BW and hyperandrogenism has been confirmed by almost all published evidence. Petraitiene et al. (2020) and Ruder et al. (2011) stated that small for gestational age or reduced BW girls demonstrated lower SHBG levels ($p < 0.05$) but higher concentrations of androstenedione, testosterone (T) ($p < 0.05$), dehydroepiandrosterone sulfate, and free androgen index ($p < 0.01$). Some studies showed that premature adrenarche results in an increased insulin response and hyperandrogenism in later adulthood (Szathmári et al., 2001; Schulte et al., 2016; Novello and Speiser, 2018). However, the relationship between lower BW and E_2 levels remains controversial. Ruder et al. (2011) and Espetvedt Finstad et al. (2009) found an inverse association between BW and levels of E_2, while Tworoger et al. (2006) and Sydsjö et al. (2019) demonstrated no differences in E_2 levels between LBW women and controls. Although no direct evidence exists to prove the positive relationship between BW and E_2, we might estimate an association between the ponderal index at birth/birth size and E_2 (Jasienska et al., 2006; Espetvedt Finstad et al., 2009). Furthermore, our research was the same with that of Kerkhof et al. (2010) and Sydsjö et al. (2019), concluding that BW did not affect AMH concentrations. However, Dior et al. (2021) reported a significant association between lower BW and reduced AMH levels in 32-year-old women who were identified after adjusting for confounders ($\hat{1}^2 = 0.18$, $p < 0.05$).

It is known that both genetic and environmental factors, such as smoking, body fat content, exposure to endocrine-disrupting chemicals, and BW, may result in early AAM (Behie and O'Donnell, 2015; Epplein et al., 2010; Żelaźniewicz et al., 2020; Adair, 2001). In our MR study, a surprising result was confirmed as well as in numerous observational studies. Moreover, Fan et al. (2018), Juul et al. (2017), and Morris et al. (2010) discovered that lower BW in infancy may increase the risk of early menarche ($p < 0.001$). However, other studies found that no or an inverse relationship was found between BW and AAM (Sorensen et al., 2013; Sydsjö et al., 2019). Moreover, conclusions on LBW and ANM have not yet been unified, and positive (Alexander et al., 2014; Bjelland et al., 2020; Goldberg et al., 2020), or inverse (Tom et al., 2010), and even no relationship (A.Treloar et al., 2000) exist.

PCOS, endometriosis, and leiomyoma were regarded as the main female endocrine diseases that may affect reproduction. Although, we did not identify any causal effects of LBW on these three diseases. Few studies considered BW as an independent risk factor related to PCOS (Fulghesu et al., 2015), and they suggested that the high risk of PCOS and related traits are because of high BW (Cresswell et al., 1997; Michelmore et al., 2001). Almost all studies confirmed a correlation between LBW and endometriosis (Olšarová and Mishra, 2020). The studies of Gao et al. (2019), Gao et al. (2020) and Borghese et al. (Borghese et al., 2015) supported the fetal origins hypothesis of endometriosis (hazard ratio [HR] = 1.35, 95% CI = 1.08–1.67; OR = 1.5, 95% CI = 1.0–2.3, $p < 0.05$); even after adjusting for confounding factors, the results still remained (risk ratio = 1.3, 95% CI = 1.0–1.8, $p < 0.05$) (Missmer et al., 2004). Furthermore, the results of Aarestrup et al. (2020) and Wolff et al. (2013) studies did not reach statistical significance and confirmed this relationship. Last, limited evidence exists of an association between leiomyomas and BW (Wise et al., 2012).

The current study exhibited a few strengths. The major preponderance was the MR design, which cut down remaining confounders and reverse causality and, thereby, improved the causal inference in associations of lower BW with female-related traits. To avoid false-positive results and bias, we selected different sourced populations to eliminate overlapping, and two-sample MR analysis was performed *via* SNPs to analyze the causal relationship from exposure to outcomes. Next, the random allocation of individual genetic variation during gamete binding is used as an IV; thereby, MR analysis can largely avoid the influence of confounders, artificial errors, and bias and provide high-quality evidence. This MR analysis included sufficient samples and only included European participants to improve the dependability of the results. To conclude, the MVMR analysis was vital for exploring the direct correlation between LBW and female outcomes.

Next, inevitably, the present study demonstrates several limitations. First, GWAS were obtained only from European individuals in this study, whose results are not representative of other races or geographic areas. Second, parts of the summary dates were incomplete because of the privacy policy and long application period, which caused a lost partial population; in addition, the BW datasets obtained contained both males and females, which may lead to collider bias (Fry et al., 2017). To conclude, the present study mainly focused on the causal role of LBW on female-specific traits, but the underlying mechanisms remain to be elucidated.

CONCLUSION

In summary, this analysis demonstrated that BW is positively associated with BMI in adulthood. In addition, LBW exhibits causal effects on decreased SHBG levels, increased bio-T levels, and early AAM.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, and further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

RH and RL envisaged and designed the study. RH, HW, and JY obtained and analyzed the datasets. RH finalized the main manuscript, whereas RL, JY, and ZJ received funding and revised the manuscript. The final version of the manuscript has been reviewed and approved by all authors.

FUNDING

The National Natural Science Foundation of China (82088102), Collaborative Innovation Program of Shanghai Municipal Health Commission (2020CXJQ01), Shanghai Frontiers Science Center of Reproduction and Development, CAMS Innovation Fund for Medical Sciences (2019-I2M-5-064), Research Units of Embryo Original Diseases, Chinese Academy of Medical Sciences (No.2019RU056), and a project supported by Scientific Research Fund of Zhejiang Provincial Education Department (Y202148357) provided financial support.

ACKNOWLEDGMENTS

First, we thank the United Kingdom Biobank Consortium and the ReproGen Consortium for providing access to their datasets. Moreover, we acknowledge the participants and investigators of the FinnGen study. Lastly, we thank Chen YX from the First Affiliated Hospital, School of Medicine, Zhejiang University, for help with the R code.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.850892/full#supplementary-material

REFERENCES

Aarestrup, J., Jensen, B. W., Ulrich, L. G., Hartwell, D., Trabert, B., and Baker, J. L. (2020). Birth Weight, Childhood Body Mass Index and Height and Risks of Endometriosis and Adenomyosis. *Ann. Hum. Biol.* 47, 173–180. doi:10.1080/03014460.2020.1727011

Adair, L. S. (2001). Size at Birth Predicts Age at Menarche. *Pediatrics* 107, E59. doi:10.1542/peds.107.4.e59

Alexander, B. T., Henry Dasinger, J., and Intapad, S. (2014). Effect of Low Birth Weight on Women's Health. *Clin. Ther.* 36, 1913–1923. doi:10.1016/j.clinthera.2014.06.026

A.Treloar, S., Sadrzadeh, S., Do, K.-A., Martin, N., and Lambalk, C. B. (2000). Birth Weight and Age at Menopause in Australian Female Twin Pairs: Exploration of the Fetal Origin Hypothesis. *Hum. Reprod.* 15, 55–59. doi:10.1093/humrep/15.1.55

Barker, D., and Osmond, C. (1986). Infant Mortality, Childhood Nutrition, and Ischaemic Heart Disease in England and Wales. *Lancet* 327, 1077–1081. doi:10.1016/s0140-6736(86)91340-1

Behie, A. M., and O'Donnell, M. H. (2015). Prenatal Smoking and Age at Menarche: Influence of the Prenatal Environment on the Timing of Puberty. *Hum. Reprod.* 30, 957–962. doi:10.1093/humrep/dev033

Bjelland, E. K., Gran, J. M., Hofvind, S., and Eskild, A. (2020). The Association of Birthweight with Age at Natural Menopause: a Population Study of Women in Norway. *Int. J. Epidemiol.* 49, 528–536. doi:10.1093/ije/dyz207

Borghese, B., Sibiude, J., Santulli, P., Lafay Pillet, M.-C., Marcellin, L., Brosens, I., et al. (2015). Low Birth Weight Is Strongly Associated with the Risk of Deep Infiltrating Endometriosis: Results of a 743 Case-Control Study. *PLoS One* 10, e0117387. doi:10.1371/journal.pone.0117387

Carlsson, S., Persson, P. G., Alvarsson, M., Efendic, S., Norman, A., Svanström, L., et al. (1999). Low Birth Weight, Family History of Diabetes, and Glucose Intolerance in Swedish Middle-Aged Men. *Diabetes Care* 22, 1043–1047. doi:10.2337/diacare.22.7.1043

Cresswell, J., Barker, D., Osmond, C., Egger, P., Phillips, D., and Fraser, R. (1997). Fetal Growth, Length of Gestation, and Polycystic Ovaries in Adult Life. *Lancet* 350, 1131–1135. doi:10.1016/s0140-6736(97)06062-5

Day, F., Karaderi, T., Jones, M. R., Meun, C., He, C., Drong, A., et al. (2018). Large-scale Genome-wide Meta-Analysis of Polycystic Ovary Syndrome Suggests Shared Genetic Architecture for Different Diagnosis Criteria. *PLoS Genet.* 14, e1007813. doi:10.1371/journal.pgen.1007813

Dior, U. P., Karavani, G., Soloveichick, V., Friedlander, Y., and Hochner, H. (2021). Early-life Factors and Adult Anti-müllerian Hormone Levels. *J. Assist. Reprod. Genet.* 38, 3019–3025. doi:10.1007/s10815-021-02281-3

Epplein, M., Novotny, R., Daida, Y., Vijayadeva, V., Onaka, A. T., and Le Marchand, L. (2010). Association of Maternal and Intrauterine Characteristics with Age at Menarche in a Multiethnic Population in Hawaii. *Cancer Causes Control* 21, 259–268. doi:10.1007/s10552-009-9457-1

Eriksson, J., Forse´n, T., Osmond, C., and Barker, D. (2000). Fetal and Childhood Growth and Hypertension in Adult Life. *Hypertension* 36, 790–794. doi:10.1161/01.hyp.36.5.790

Eriksson, J. G., Forse´n, T., Tuomilehto, J., Osmond, C., and Barker, D. J. P. (2000). Early Growth, Adult Income, and Risk of Stroke. *Stroke* 31, 869–874. doi:10.1161/01.str.31.4.869

Espetvedt Finstad, S., Emaus, A., Potischman, N., Barrett, E., Furberg, A.-S., Ellison, P. T., et al. (2009). Influence of Birth Weight and Adult Body Composition on 17β-Estradiol Levels in Young Women. *Cancer Causes Control* 20, 233–242. doi:10.1007/s10552-008-9238-2

Fan, H.-Y., Huang, Y.-T., Hsieh, R.-H., Chao, J. C.-J., Tung, Y.-C., Lee, Y. L., et al. (2018). Birthweight, Time-Varying Adiposity Growth and Early Menarche in Girls: A Mendelian Randomisation and Mediation Analysis. *Obes. Res. Clin. Pract.* 12, 445–451. doi:10.1016/j.orcp.2018.07.008

Ferrie, J. E., Langenberg, C., Shipley, M. J., and Marmot, M. G. (2006). Birth Weight, Components of Height and Coronary Heart Disease: Evidence from the Whitehall II Study. *Int. J. Epidemiol.* 35, 1532–1542. doi:10.1093/ije/dyl184

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., et al. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *Am. J. Epidemiol.* 186, 1026–1034. doi:10.1093/aje/kwx246

Fulghesu, A. M., Manca, R., Loi, S., and Fruzzetti, F. (2015). Insulin Resistance and Hyperandrogenism Have No Substantive Association with Birth Weight in Adolescents with Polycystic Ovary Syndrome. *Fertil. Steril.* 103, 808–814. doi:10.1016/j.fertnstert.2014.12.109

Gao, M., Allebeck, P., Mishra, G. D., and Koupil, I. (2019). Developmental Origins of Endometriosis: a Swedish Cohort Study. *J. Epidemiol. Community Health* 73, 353–359. doi:10.1136/jech-2018-211811

Gao, M., Scott, K., and Koupil, I. (2020). Associations of Perinatal Characteristics with Endometriosis: a Nationwide Birth Cohort Study. *Int. J. Epidemiol.* 49, 537–547. doi:10.1093/ije/dyz140

Goldberg, M., Tawfik, H., Kline, J., Michels, K. B., Wei, Y., Cirillo, P., et al. (2020). Body Size at Birth, Early-Life Growth and the Timing of the Menopausal Transition and Natural Menopause. *Reprod. Toxicol.* 92, 91–97. doi:10.1016/j.reprotox.2019.02.013

Horikoshi, M., Beaumont, R. N., Day, F. R., Warrington, N. M., Kooijman, M. N., Fernandez-Tajes, J., et al. (2016). Genome-wide Associations for Birth Weight and Correlations with Adult Disease. *Nature* 538, 248–252. doi:10.1038/nature19806

Huang, W., Huang, T., WangGao, T., Zheng, Y., Ellervik, C., Li, X., et al. (2019). Association of Birth Weight with Type 2 Diabetes and Glycemic Traits: A Mendelian Randomization Study. *JAMA Netw. Open* 2, e1910915. doi:10.1001/jamanetworkopen.2019.10915

Jasienska, G., Ziomkiewicz, A., Lipson, S. F., Thune, I., and Ellison, P. T. (2006). High Ponderal Index at Birth Predicts High Estradiol Levels in Adult Women. *Am. J. Hum. Biol.* 18, 133–140. doi:10.1002/ajhb.20462

Jelenkovic, A., Yokoyama, Y., Sund, R., Pietiläinen, K. H., Hur, Y.-M., Willemsen, G., et al. (2017). Association between Birthweight and Later Body Mass Index: an Individual-Based Pooled Analysis of 27 Twin Cohorts Participating in the CODATwins Project. *Int. J. Epidemiol.* 46, 1488–1498. doi:10.1093/ije/dyx031

Juul, F., Chang, V. W., Brar, P., and Parekh, N. (2017). Birth Weight, Early Life Weight Gain and Age at Menarche: a Systematic Review of Longitudinal Studies. *Obes. Rev.* 18, 1272–1288. doi:10.1111/obr.12587

Kahn, L. G., Buka, S. L., Cirillo, P. M., Cohn, B. A., Factor-Litvak, P., Gillman, M. W., et al. (2017). Evaluating the Relationship between Birth Weight for Gestational Age and Adult Blood Pressure Using Participants from a Cohort of Same-Sex Siblings, Discordant on Birth Weight Percentile. *Am. J. Epidemiol.* 186, 550–554. doi:10.1093/aje/kwx126

Kerkhof, G. F., Leunissen, R. W. J., Willemsen, R. H., de Jong, F. H., Visser, J. A., Laven, J. S. E., et al. (2010). Influence of Preterm Birth and Small Birth Size on Serum Anti-müllerian Hormone Levels in Young Adult Women. *Eur. J. Endocrinol.* 163, 937–944. doi:10.1530/eje-10-0528

Knop, M. R., Geng, T. T., Gorny, A. W., Ding, R., Li, C., Ley, S. H., et al. (2018). Birth Weight and Risk of Type 2 Diabetes Mellitus, Cardiovascular Disease, and Hypertension in Adults: A Meta-Analysis of 7 646 267 Participants from 135 Studies. *Jaha* 7, e008870. doi:10.1161/jaha.118.008870

Lawlor, D. A., Richmond, R., Warrington, N., McMahon, G., Smith, G. D., Bowden, J., et al. (2017). Using Mendelian Randomization to Determine Causal Effects of Maternal Pregnancy (Intrauterine) Exposures on Offspring Outcomes: Sources of Bias and Methods for Assessing Them. *Wellcome Open Res.* 2, 11. doi:10.12688/wellcomeopenres.10567.1

Liao, C. X., Gao, W. J., Sun, L. L., Gao, Y., Cao, W. H., Lyu, J., et al. (2020). Birth Weight Predicts Physical Indicators in Adulthood: a Large Population-Based Study in Chinese Twins. *Zhonghua Liu Xing Bing Xue Za Zhi* 41, 310–314. doi:10.3760/cma.j.issn.0254-6450.2020.03.006

Michelmore, K., Ong, K., Mason, S., Bennett, S., Perry, L., Vessey, M., et al. (2001). Clinical Features in Women with Polycystic Ovaries: Relationships to Insulin Sensitivity, Insulin Gene VNTR and Birth Weight. *Clin. Endocrinol. (Oxf)* 55, 439–446. doi:10.1046/j.1365-2265.2001.01375.x

Missmer, S. A., Hankinson, S. E., Spiegelman, D., Barbieri, R. L., Michels, K. B., and Hunter, D. J. (2004). In Utero exposures and the Incidence of Endometriosis. *Fertil. Steril.* 82, 1501–1508. doi:10.1016/j.fertnstert.2004.04.065

Morley, R., McCalman, J., and Carlin, J. B. (2006). Birthweight and Coronary Heart Disease in a Cohort Born 1857-1900 in Melbourne, Australia. *Int. J. Epidemiol.* 35, 880–885. doi:10.1093/ije/dyl032

Morris, D. H., Jones, M. E., Schoemaker, M. J., Ashworth, A., and Swerdlow, A. J. (2010). Determinants of Age at Menarche in the UK: Analyses from the Breakthrough Generations Study. *Br. J. Cancer* 103, 1760–1764. doi:10.1038/sj.bjc.6605978

Novello, L., and Speiser, P. W. (2018). Premature Adrenarche. *Pediatr. Ann.* 47, e7–e11. doi:10.3928/19382359-20171214-04

Olšarová, K., and Mishra, G. D. (2020). Early Life Factors for Endometriosis: a Systematic Review. *Hum. Reprod. Update* 26, 412–422. doi:10.1093/humupd/dmaa002

Peck, J. D., Hulka, S., Baird, P., and Richardson, B. (2003). Accuracy of Fetal Growth Indicators as Surrogate Measures of Steroid Hormone Levels during Pregnancy. *Am. J. Epidemiol.* 157, 258–266. doi:10.1093/aje/kwf183

Petraitiene, I., Valuniene, M., Jariene, K., Seibokaite, A., Albertsson-Wikland, K., and Verkauskiene, R. (2020). Sex Hormones, Gonad Size, and Metabolic Profile in Adolescent Girls Born Small for Gestational Age with Catch-Up Growth. *J. Pediatr. Adolesc. Gynecol.* 33, 125–132. doi:10.1016/j.jpag.2019.11.001

Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., et al. (2019). Meta-analysis of Genome-wide Association Studies for Body Fat Distribution in 694 649 Individuals of European Ancestry. *Hum. Mol. Genet.* 28, 166–174. doi:10.1093/hmg/ddy327

Rogers, I. (2003). The Influence of Birthweight and Intrauterine Environment on Adiposity and Fat Distribution in Later Life. *Int. J. Obes.* 27, 755–777. doi:10.1038/sj.ijo.0802316

Ruder, E. H., Hartman, T. J., Rovine, M. J., and Dorgan, J. F. (2011). Birth Characteristics and Female Sex Hormone Concentrations during Adolescence: Results from the Dietary Intervention Study in Children. *Cancer Causes Control* 22, 611–621. doi:10.1007/s10552-011-9734-7

Ruiz-Narváez, E. A., Palmer, J. R., Gerlovin, H., Wise, L. A., Vimalananda, V. G., Rosenzweig, J. L., et al. (2014). Birth Weight and Risk of Type 2 Diabetes in the Black Women's Health Study: Does Adult BMI Play a Mediating Role? *Diabetes Care* 37, 2572–2578. doi:10.2337/dc14-0731

Ruth, K. S., Soares, A. L. G., Borges, M.-C., Eliassen, A. H., Hankinson, S. E., Jones, M. E., et al. (2019). Genome-wide Association Study of Anti-müllerian Hormone Levels in Pre-menopausal Women of Late Reproductive Age and Relationship with Genetic Determinants of Reproductive Lifespan. *Hum. Mol. Genet.* 28, 1392–1401. doi:10.1093/hmg/ddz015

Ruth, K. S., Tyrrell, T., Day, F. R., Tyrrell, J., Thompson, D. J., Wood, A. R., et al. (2020). Using Human Genetics to Understand the Disease Impacts of Testosterone in Men and Women. *Nat. Med.* 26, 252–258. doi:10.1038/s41591-020-0751-5

Scharf, R. J., Stroustrup, A., Conaway, M. R., and DeBoer, M. D. (2016). Growth and Development in Children Born Very Low Birthweight. *Arch. Dis. Child. Fetal Neonatal Ed.* 101, F433–F438. doi:10.1136/archdischild-2015-309427

Schmitz, D., Ek, W. E., Berggren, E., Höglund, J., Karlsson, T., and Johansson, Å. (2021). Genome-wide Association Study of Estradiol Levels and the Causal Effect of Estradiol on Bone Mineral Density. *J. Clin. Endocrinol. Metab.* 106, e4471–e4486. doi:10.1210/clinem/dgab507

Schulte, S., Wölfle, J., Schreiner, F., Stoffel-Wagner, B., Peter, M., Bartmann, P., et al. (2016). Birthweight Differences in Monozygotic Twins Influence Pubertal Maturation and Near Final Height. *J. Pediatr.* 170, 288–294. e1-2. doi:10.1016/j.jpeds.2015.12.020

Sorensen, K., Christensen, K., Juul, A., Skytthe, A., Scheike, T., and Kold Jensen, T. (2013). Birth Size and Age at Menarche: a Twin Perspective. *Hum. Reprod.* 28, 2865–2871. doi:10.1093/humrep/det283

Sydsjö, G., Törnblom, P., Gäddlin, P.-O., Finnström, O., Leijon, I., Nelson, N., et al. (2019). Women Born with Very Low Birth Weight Have Similar Menstrual Cycle Pattern, Pregnancy Rates and Hormone Profiles Compared with Women Born at Term. *BMC Women's Health* 19, 56. doi:10.1186/s12905-019-0753-y

Szathmári, M., Vásárhelyi, B., and Tulassay, T. (2001). Effect of Low Birth Weight on Adrenal Steroids and Carbohydrate Metabolism in Early Adulthood. *Horm. Res. Paediatr.* 55, 172–178. doi:10.1159/000049991

Tamakoshi, K., Yatsuya, H., Wada, K., Matsushita, K., Otsuka, R., Yang, P. O., et al. (2006). Birth Weight and Adult Hypertension Cross-Sectional Study in a Japanese Workplace Population. *Circ. J.* 70, 262–267. doi:10.1253/circj.70.262

Tom, S. E., Cooper, R., Kuh, D., Guralnik, J. M., Hardy, R., and Power, C. (2010). Fetal Environment and Early Age at Natural Menopause in a British Birth Cohort Study. *Hum. Reprod.* 25, 791–798. doi:10.1093/humrep/dep451

Tworoger, S. S., Eliassen, A. H., Missmer, S. A., Baer, H., Rich-Edwards, J., Michels, K. B., et al. (2006). Birthweight and Body Size throughout Life in Relation to Sex Hormones and Prolactin Concentrations in Premenopausal Women. *Cancer Epidemiol. Biomarkers Prev.* 15, 2494–2501. doi:10.1158/1055-9965.Epi-06-0671

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50, 693–698. doi:10.1038/s41588-018-0099-7

Wang, T., Tang, Z., Yu, X., Gao, Y., Guan, F., Li, C., et al. (2020). Birth Weight and Stroke in Adult Life: Genetic Correlation and Causal Inference with Genome-wide Association Data Sets. *Front. Neurosci.* 14, 479. doi:10.3389/fnins.2020.00479

Whincup, K., Owen, H., and Cook, A. (2008). Birth Weight and Risk of Type 2 Diabetes. *Jama* 300, 2886–2897. doi:10.1001/jama.2008.886

Wise, L. A., Radin, R. G., Palmer, J. R., and Rosenberg, L. (2012). Association of Intrauterine and Early Life Factors with Uterine Leiomyomata in Black Women. *Ann. Epidemiol.* 22, 847–854. doi:10.1016/j.annepidem.2012.09.006

Wolff, E. F., Sun, L., Hediger, M. L., Sundaram, R., Peterson, C. M., Chen, Z., et al. (2013). In Utero exposures and Endometriosis: the Endometriosis, Natural History, Disease, Outcome (ENDO) Study. *Fertil. Steril.* 99, 790–795. doi:10.1016/j.fertnstert.2012.11.013

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R Package for Performing Mendelian Randomization Analyses Using Summarized Data. *Int. J. Epidemiol.* 46, 1734–1739. doi:10.1093/ije/dyx034

Yu, X., Yuan, Z., Lu, H., Gao, Y., Chen, H., Shao, Z., et al. (2020). Relationship between Birth Weight and Chronic Kidney Disease: Evidence from Systematics Review and Two-Sample Mendelian Randomization Analysis. *Hum. Mol. Genet.* 29, 2261–2274. doi:10.1093/hmg/ddaa074

Żelaźniewicz, A., Nowak, J., and Pawłowski, B. (2020). Birth Size and Morphological Femininity in Adult Women. *BMC Evol. Biol.* 20, 102. doi:10.1186/s12862-020-01670-z

Zeng, P., Yu, X., and Zhou, X. (2019). Birth Weight Is Not Causally Associated with Adult Asthma: Results from Instrumental Variable Analyses. *Sci. Rep.* 9, 7647. doi:10.1038/s41598-019-44114-5

Zeng, P., and Zhou, X. (2019). Causal Association between Birth Weight and Adult Diseases: Evidence from a Mendelian Randomization Analysis. *Front. Genet.* 10, 618. doi:10.3389/fgene.2019.00618

Zhao, Y., Wang, S.-F., Mu, M., and Sheng, J. (2012). Birth Weight and Overweight/obesity in Adults: a Meta-Analysis. *Eur. J. Pediatr.* 171, 1737–1746. doi:10.1007/s00431-012-1701-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership