frontiers | Research Topics

High-throughput sequencing-based investigation of chronic disease markers and mechanisms, volume II

Edited by
Hua Li, Wen-Lian Chen, Yuriy L. Orlov and Guoshuai Cai

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-6454-7
DOI 10.3389/978-2-8325-6454-7

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

High-throughput sequencing-based investigation of chronic disease markers and mechanisms, volume II

Topic editors

Hua Li — Shanghai Jiao Tong University, China

Wen-Lian Chen — Shanghai University of Traditional Chinese Medicine, China

Yuriy L. Orlov — I.M. Sechenov First Moscow State Medical University, Russia

Guoshuai Cai — University of Florida, United States

Citation

Li, H., Chen, W.-L., Orlov, Y. L., Cai, G., eds. (2025). *High-throughput sequencing-based investigation of chronic disease markers and mechanisms, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6454-7

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 29 May 2025
DOI 10.3389/fgene.2025.1627976

OPEN ACCESS

EDITED AND REVIEWED BY
Jared C. Roach,
Institute for Systems Biology (ISB), United States

*CORRESPONDENCE
Yuriy L. Orlov,
orlov@d-health.institute

RECEIVED 13 May 2025
ACCEPTED 19 May 2025
PUBLISHED 29 May 2025

CITATION
Li H, Cai G, Chen W-L, Zhao X and Orlov YL (2025) Editorial: High-throughput sequencing-based investigation of chronic disease markers and mechanisms, volume II.
Front. Genet. 16:1627976.
doi: 10.3389/fgene.2025.1627976

Editorial: High-throughput sequencing-based investigation of chronic disease markers and mechanisms, volume II

Hua Li[1,2], Guoshuai Cai[3], Wen-Lian Chen[4], Xiaodong Zhao[1,2] and Yuriy L. Orlov[5,6]*

[1]Jiangsu Province Engineering Research Center of Development and Translation of Key Technologies for Chronic Disease Prevention and Control, Suzhou Vocational Health College, Suzhou, China, [2]Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China, [3]Department of Surgery, College of Medicine, University of Florida, Gainesville, FL, United States, [4]Longhua Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [5]The Digital Health Center, I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), Moscow, Russia, [6]Systems Biology Department, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

KEYWORDS

high-throughput sequencing, biomarker development, chronic disease, omics study, disease mechanism

Editorial on the Research Topic
High-throughput sequencing-based investigation of chronic disease markers and mechanisms, volume II

Introduction

Second-generation (short-read, massively parallel) sequencing and third-generation (long-read, single-molecule) sequencing technologies have matured rapidly, irreversibly altering how we interrogate human health and disease. A series of *Frontiers in genetics* Research Topics highlight this area (Orlov and Baranova, 2020; Anashkina et al., 2023). Building on the inaugural 2022 Research Topic (Orlov et al., 2022), this second volume of "High-throughput Sequencing-based Investigation of Chronic Disease Markers and Mechanisms" (https://www.frontiersin.org/research-topics/53085/high-throughput-sequencing-based-investigation-of-chronic-disease-markers-and-mechanisms-volume-ii/articles) again harnesses deep sequencing technologies, sophisticated analytics, and cross-scale validation to illuminate biomarkers and mechanisms that underlie a spectrum of chronic conditions - from inflammatory bowel disease to neuromuscular degeneration and pandemic infection. Together, the nine articles accepted in this Research Topic exemplify three converging trends: (i) omics integration across genome, epigenome, transcriptome and proteome; (ii) fast sequencing applications that translate into clinically actionable diagnostics; and (iii) mechanistic dissection of how candidate markers shape or signal pathophysiology.

Gastrointestinal and metabolic diseases: decoding tissue-specific signatures

Crohn's disease remains clinically heterogeneous and therapeutically stubborn. Yang et al. performed bulk RNA-seq of intact bowel walls and revealed two strikingly upregulated transcripts, PDE1A [OMIM 171890] and SEMA3D [OMIM 609907], associated with smooth muscle cell apoptosis and autonomic dysregulation, respectively, providing a plausible axis for the distinctive neuromuscular complications of this disease.

Turning to metabolic syndromes, Yao et al. mined public expression datasets, intersected them with ER-stress gene sets, and narrowed 49 differentially expressed genes down to three diagnostic markers - CLGN [OMIM 601858], ILF2 [OMIM 603181], IMPA1 [OMIM 602064] - that were subsequently validated in mouse models and patient sera. The study underscores how *in silico* LASSO feature selection, when wedded to wet-lab confirmation, can yield serum-accessible biomarkers for type 2 diabetes mellitus.

Oncology: multi-omics and precise mutation discovery

Two cancer-focused articles showcase complementary high-throughput strategies. Wang et al. isolated a circulating bio-active peptide (YG-22) generated when adjuvant chemotherapy was combined with the traditional Chinese Jianpi formula; multi-layer omics (transcriptome, metabolome, chromatin accessibility, H3K4me3 ChIP-seq, NF-κB ChIP-seq) revealed that YG-22 reprograms epigenetic states and lysosomal pathways to suppress colorectal cancer cell viability. This study demonstrates the added value of peptide therapeutics derived from phytochemical regimens.

At the single-gene end of the spectrum, Zhang et al. applied targeted next-generation sequencing to four myeloproliferative-neoplasm cases that were "triple negative" by canonical testing, unmasking novel driver lesions and arguing for routine targeted sequencing in ambiguous myeloid diagnoses.

Neuromuscular and neurodevelopmental research: from modifiers to toxicants

By pairing bulk and single-nucleus RNA-seq in healthy vastus lateralis versus tibialis anterior, Nieves-Rodriguez et al. identified >3,400 genes - including those related to calcium release and collagen-containing extracellular matrix transcripts - that may dictate differential vulnerability of muscles in Duchenne muscular dystrophy, supplying an invaluable reference for stratified gene-therapy design.

Li et al. then leveraged whole-exome sequencing of 113 patients with intellectual disability to uncover a novel *de novo* SYNGAP1 [OMIM 603384] splice-site variant (c.664-2A>G). Minigene assays confirmed exon 7 skipping, emphasizing that modest intronic changes that are detectable by high-depth sequencing can produce profound neurodevelopmental phenotypes.

Complementing human genetics, Lyu et al. used comparative transcriptomics in zebrafish embryos to show that extremely small iron-oxide nanoparticles (ESIONPs) perturb neuro-muscular development and trigger ferroptosis. Weighted gene co-expression network analysis (WGCNA) pinpointed stage-specific hub genes (highly connected nodes in the network), such as neurodevelopmental regulators and oxidative-stress mediators, whose dysregulation, together with elevated apoptosis markers, signals potential health risks of nanoparticle biomedical imaging.

Infection and immunity: from viral alternative polyadenylation to host GWAS loci

The interface between host gene regulation and pathogen assault is another recurring theme. Tan et al. profiled grass-carp cells during early grass carp reovirus infection and uncovered extensive shifts in alternative polyadenylation (APA) despite stable DNA methylation patterns, particularly affecting cytoskeletal and microtubule genes - an underappreciated layer of post-transcriptional control in fish viral pathogenesis.

On the human front, Loktionov et al. genotyped 10 GWAS-significant loci in nearly 800 Russians and confirmed that the SLC6A20-LZTFL1 rs17713054 risk allele magnifies severe COVID-19 particularly in obese, low-activity, or low-dietary-fruit subgroups, with concordant effects on thrombodynamics. Network analyses further highlighted interactive SNP constellations linking coagulation and immune genes. Such population-targeted validation of multi-locus risk underlines the translational scope of sequencing even after the acute pandemic phase.

Methodological cross-talk and shared biological threads

The field of gene expression regulation including chronic disease markers has been covered in a Frontiers in Genetics Research Topic (Anashkina et al., 2023) based on omics data integration. Sequencing technologies give background for gene expression regulation studies at genome scale (Anashkina et al., 2021; Orlov et al., 2023).

Across the current Research Topic, several common methodological themes emerge. First, multi-omics integration - whether combining peptidomics with chromatin readouts, or pairing methylome and APA maps - magnifies biological signals and reveals underlying mechanisms. Second, targeted or panel-based sequencing continues to sharpen genetic diagnosis where standard assays falter. Third, bioinformatics methods (WGCNA, LASSO, SNP-SNP interaction models) distill high-dimensional data into clinically tractable results.

Biologically, six recurrent pathways unite otherwise disparate studies: ER stress, calcium homeostasis, apoptotic regulation, extracellular-matrix remodeling, innate immune activation, and ferroptosis. This convergence reinforces the idea that chronic diseases, despite tissue specificity, share certain conserved response architectures that are captured by high-throughput sequencing.

Outlook

Together, the nine articles in this volume broaden the map of chronic-disease markers, bring sequencing into daily clinical applications, and deepen our grasp of how genetic and epigenetic patterns drive long-term illness. Looking forward to this Research Topic development, we may anticipate:

- Single-cell and spatial omics will dissect cell type-restricted marker function within complex tissues such as muscle, gut, and tumor microenvironments.
- Long-read platforms will resolve structural and splice isoform diversity.
- Prospective, multi-center cohorts integrating more data (e.g., diet, exercise) with host genetics, as illustrated in the COVID-19 study, will refine gene-environment risk algorithms.
- In addition, from the current perspective, AI applications will get a more important role in complex disease studies (Koshechkin et al., 2022; Zhang et al., 2024).

Author contributions

HL: Conceptualization, Funding acquisition, Writing – original draft, Writing – review and editing. GC: Writing – review and editing. W-LC: Writing – review and editing. XZ: Writing – review and editing. YO: Conceptualization, Funding acquisition, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by RSF (grant 24-24-00563) for YO. This research was supported by Jiangsu Province Engineering Research Center of Development and Translation of Key Technologies for Chronic Disease Prevention and Control (CDSGK1202503). This research was supported by the Fundamental Research Funds for the Central Universities (KLSB2024KF-02).

Acknowledgments

We thank all authors and reviewers for their pivotal contributions to this Research Topic.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Anashkina, A. A., Leberfarb, E. Y., and Orlov, Y. L. (2021). Recent trends in cancer genomics and bioinformatics tools development. *Int. J. Mol. Sci.* 22, 12146. doi:10.3390/ijms222212146

Anashkina, A. A., Orlova, N. G., Ignatov, A. N., Chen, M., and Orlov, Y. L. (2023). Editorial: bioinformatics of genome regulation and systems biology, Volume III. *Front. Genet.* 14, 1215987. doi:10.3389/fgene.2023.1215987

Koshechkin, K. A., Lebedev, G. S., Fartushnyi, E. N., and Orlov, Y. L. (2022). Holistic approach for artificial intelligence implementation in pharmaceutical products lifecycle: a meta-analysis. *Appl. Sci.* 12 (16), 8373. doi:10.3390/app12168373

Orlov, Y. L., Anashkina, A. A., Kumeiko, V. V., Chen, M., and Kolchanov, N. A. (2023). Research topics of the bioinformatics of gene regulation. *Int. J. Mol. Sci.* 24 (10), 8774. doi:10.3390/ijms24108774

Orlov, Y. L., and Baranova, A. V. (2020). Editorial: bioinformatics of genome regulation and systems biology. *Front. Genet.* 11, 625. doi:10.3389/fgene.2020.00625

Orlov, Y. L., Chen, W. L., Sekacheva, M. I., Cai, G., and Li, H. (2022). Editorial: high-throughput sequencing-based investigation of chronic disease markers and mechanisms. *Front. Genet.* 13, 922206. doi:10.3389/fgene.2022.922206

Zhang, S., Wu, L., Zhao, Z., Fernandez Masso, J. R., and Chen, M. (2024). Artificial intelligence in gerontology: data-driven health management and precision medicine. *Adv. Gerontology* 14, 97–110. doi:10.1134/S2079057024600691

frontiers | Frontiers in Genetics

TYPE Case Report
PUBLISHED 16 June 2023
DOI 10.3389/fgene.2023.1198834

OPEN ACCESS

EDITED BY
Hua Li,
Shanghai Jiao Tong University, China

REVIEWED BY
Kausar Jabbar,
Beaumont Health, United States
Jess Peterson,
Mayo Clinic, United States
Gonzalo Carreño-Tarragona,
University Hospital October 12, Spain

*CORRESPONDENCE
Yaqin Wang,
wangyaq@163.com
Zhenya Hong,
hongzhenya@126.com

†PRESENT ADDRESS
Jin Zhang,
Department of Hematology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China.

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

RECEIVED 02 April 2023
ACCEPTED 31 May 2023
PUBLISHED 16 June 2023

CITATION
Zhang J, Shen K, Xiao M, Huang J, Wang J, Wang Y and Hong Z (2023), Case report: Application of targeted NGS for the detection of non-canonical driver variants in MPN.
Front. Genet. 14:1198834.
doi: 10.3389/fgene.2023.1198834

Case report: Application of targeted NGS for the detection of non-canonical driver variants in MPN

Jin Zhang[1,2†‡], Kefeng Shen[1‡], Min Xiao[1], Jinjin Huang[1], Jin Wang[1], Yaqin Wang[3]* and Zhenya Hong[1]*

[1]Department of Hematology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China, [2]Department of Clinical Immunology, Xijing Hospital, Fourth Military Medical University, Xi'an, China, [3]Department of Pediatric Hematology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

Background: JAK2, CALR, and MPL gene mutations are recognized as driver mutations of myeloproliferative neoplasms (MPNs). MPNs without these mutations are called triple-negative (TN) MPNs. Recently, novel mutation loci were continuously discovered using next-generation sequencing (NGS), along with continued discussion and modification of the traditional TN MPN.

Case presentation: Novel pathogenic mutations were discovered by targeted NGS in 4 patients who were diagnosed as JAK2 unmutated polycythaemia vera (PV) or TN MPN. Cases 1, 2, and 3 were of patients with PV, essential thrombocythemia (ET), and primary myelofibrosis (PMF); NGS detected JAK2 p.H538_K539delinsQL (uncommon), CALR p.E380Rfs*51 (novel), and MPL p.W515_Q516del (novel) mutations. Case 4 involved a patient with PMF; JAK2, CALR, or MPL mutations were not detected by qPCR or NGS, but a novel mutation SH2B3 p.S337Ffs*3, which is associated with the JAK/STAT signal transduction pathway, was found by NGS.

Conclusion: NGS, a more multidimensional and comprehensive gene mutation detection, is required for patients suspected of having MPN to detect non-canonical driver variants and avoid the misdiagnosis of TN MPN. SH2B3 p.S337Ffs*3 can drive MPN occurrence, and SH2B3 mutation may also be a driver mutation of MPN.

KEYWORDS

triple-negative myeloproliferative neoplasm, next-generation sequencing, JAK2, CALR, MPL, SH2B3

Introduction

Myeloproliferative neoplasms (MPNs) are a group of myeloid tumours characterised by relatively normal differentiation but uncontrolled proliferation of myeloid granulocytes, erythroid cells, and/or megakaryocytes. Classic MPNs include polycythaemia vera (PV, Phenotype MIM number 263300), essential thrombocythemia (ET), and primary myelofibrosis (PMF, Phenotype MIM number 254450) (Arber et al., 2016). The main mechanism of MPN is mutations in genes associated with the JAK/STAT signal transduction pathway, driving excessive proliferation of myeloid cells (Grinfeld et al., 2018). According to the 5th edition of

TABLE 1 Common MPN driver mutations.

Gene	Mutation type
JAK2	p.V617F
	p.N542_E543del
	p.E543_D544del
	p.K539L
CALR	p.L367fs*46
	p.K385fs*47
MPL	p.W515K/A/L/R/S
	p.S505N

MPN: myeloproliferative neoplasm. del: delete. fs: frame-shift.

the World Health Organization (WHO) Classification of Haematolymphoid Tumours (Khoury et al., 2022), in addition to blood cell counts and bone marrow biopsy, one of the major diagnostic criteria is the existence of JAK2, CALR, and/or MPL mutations. Approximately 80%–90% of patients with MPN have these driver gene mutations, while the others are patients with triple-negative (TN) MPN and a worse prognosis (Passamonti and Maffioli, 2016; Rumi and Cazzola, 2017).

Currently, there are three commonly used detection methods for MPN gene mutations: fluorescent quantitative PCR (qPCR) (Supplementary Material S1), Sanger sequencing (Supplementary Material S2), and targeted next-generation sequencing (NGS) (Supplementary Material S3). In qPCR, the hybridization probes were designed based on the hotspot mutations of JAK2, CALR, and MPL. qPCR is characterised by its high sensitivity, short detection time, and fair price. However, qPCR also has defects: it requires specific primers, so novel and non-hotspot mutations cannot be detected. Sanger sequencing covers more mutations than qPCR does, but its sensitivity is relatively low (15%–20%), implying that the mutation cannot be detected if the variant allele frequency (VAF) is lower than 15%–20%. More than 98% of mutations can be detected using the high-throughput and high-sensitivity approach of targeted NGS. Both known and novel mutations of MPN are covered with a high sensitivity at a higher cost.

When applied to MPN patients to detect relevant mutations, both qPCR and Sanger sequencing have defects such as incomplete covering loci or low sensitivity. As a result, targeted NGS is especially important for the diagnosis of triple-negative MPN.

Case description

Case 1 (P1) was of a 30-year-old female patient with PV, who exhibited increased haemoglobin level 4 years ago and then underwent bone marrow aspiration and biopsy in a local hospital. The usual MPN-related gene mutations (Table 1) were not detected using qPCR and the bone marrow biopsy result at the initial diagnosis was lost. She was diagnosed with JAK2 unmutated PV at the same hospital and underwent phlebotomy and oral aspirin therapy. The patient visited our hospital in October 2020, and her haemoglobin level was 198.0 g/L. Bone marrow biopsy was conducted again, and the results confirmed diagnosis of MPN (Table 2; Figure 1). Targeted NGS revealed the presence of JAK2 exon12 mutation (p.H538_K539delinsQL) (Figure 2) with a VAF of 20.9%. This mutation is very rare, and was reported only few times in the COSMIC database before May 2023. COSMIC is the catalogue of somatic mutations in cancer, and is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer (https://cancer.sanger.ac.uk/cosmic?genome=37). It was omitted from initial diagnosis because qPCR did not include this locus.

Case 2 (P2) was of a 56-year-old female patient with ET. The patient presented with dizziness when she first visited the hospital, after which an increased platelet count was noted. She underwent routine blood tests regularly, and the platelet count increased progressively to 812.0 × 109/L. Conditions involving increasing number of reactive platelets, such as in infection, bleeding, or tumours, were excluded. The patient then underwent bone marrow aspiration and biopsy. Aspirated bone marrow films showed clues of MPN (Table 2; Figure 1). MPN-related gene

TABLE 2 Bone marrow examinations.

	P1-PV	P2-ET	P3-PMF	P4-PMF
Films of bone marrow	The percentage of erythrocytes increased, and the central pale area of mature erythrocytes disappeared	Increased platelets	Mature erythrocytes varied in size	Mature erythrocytes varied in size
			Nucleated and tear-drop erythrocytes were found	Nucleated erythrocytes were found
Bone marrow biopsy	Significant hypercellularity (~90%), proliferation of erythrocytes, slightly proliferation of megakaryocytes, MF-1. MPN to be determined	Hypercellularity (~50%), hyperlobulated megakaryocytes were found	Hypocellularity (~30%), and there were megakaryocytes with atypical and bare nuclei. MF-3	Hypocellularity (~30%), megakaryocytes with atypical, bare nuclei and cloud-like nuclei were found. MF-2
Flow cytometry	No significant abnormalities were found	—	—	—
Karyotype	46, XX	—	46, XX	—
Next-generation sequencing	JAK2 p.H538_K539delinsQL	CALR p.E380Rfs*51	MPL p.W515_Q516del	SH2B3 p.S337Ffs*3
		NFE2 p.R284C	ASXL1 p.G996Sfs*3	
		TET2 p.Q622*	EZH2 p.C590Y	

—: The patient did not accept the test.

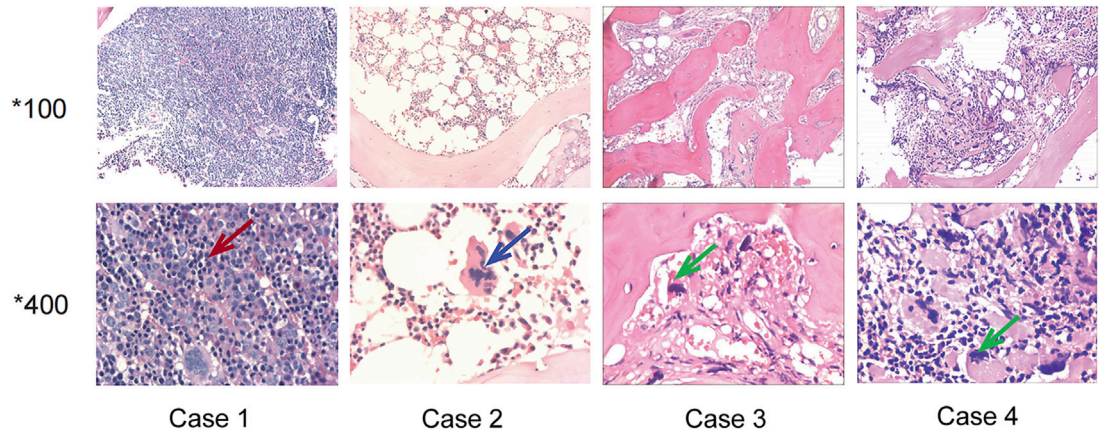


FIGURE 1
Bone marrow sections of the four patients. Case 1: The red arrow showed erythroblast proliferation, supporting the diagnosis of polycythemia vera. Case 2: The blue arrow showed hyperlobulated megakaryocyte, supporting the diagnosis of essential thrombocythemia. Case 3 and Case 4: The green arrow showed megakaryocytes with atypical and bare nucleus, supporting the diagnosis of primary myelofibrosis.

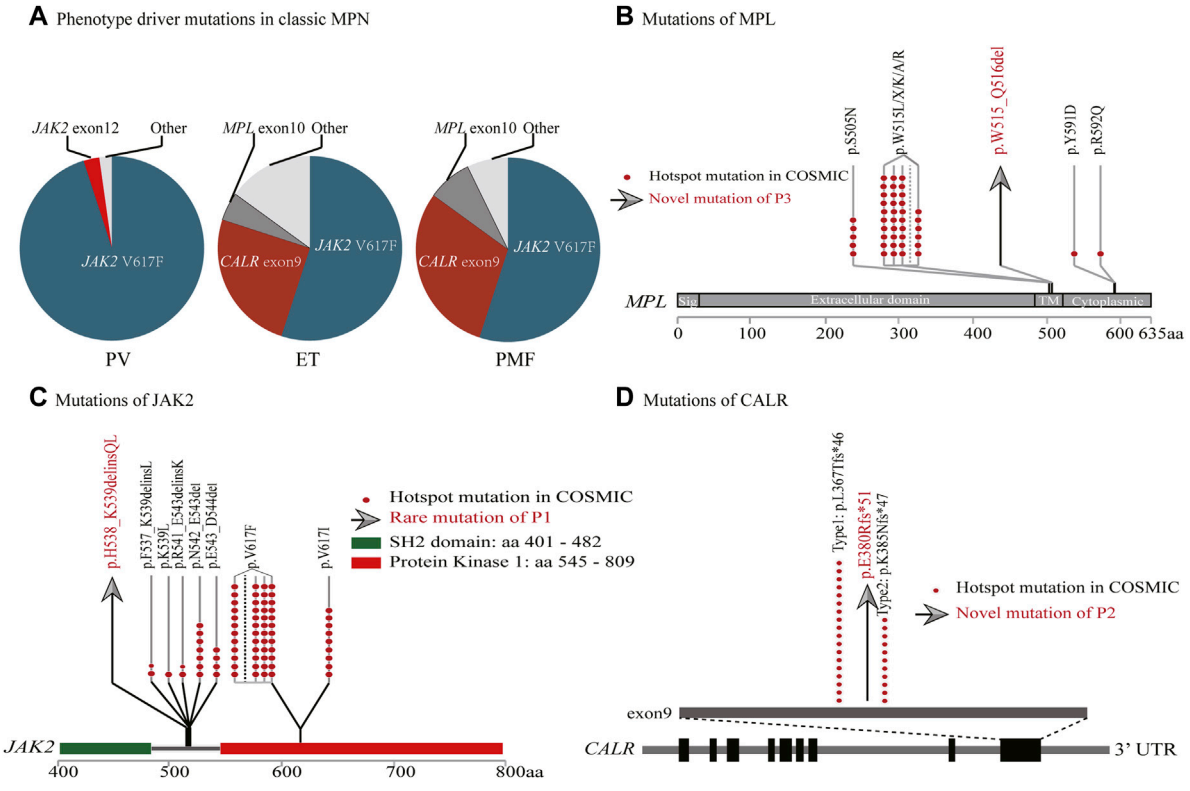


FIGURE 2
Genomic landscape of myeloproliferative neoplasms. **(A)** Driver gene mutation frequencies. **(B–D)** MPL, JAK2 and CALR gene structures. The red dots represent hotspot mutations in COSMIC. The gray arrows represent rare/novel mutations of the cases (MPL p.W515_Q516del, JAK2 p.H538_K539delinsQL, CALR p.E380Rfs*51).

mutations were found to be negative using Sanger sequencing. After 6 months of close monitoring, the platelet count did not decrease. Therefore, targeted NGS was conducted, and CALR p.E380Rfs*51 was detected with a VAF of 12.3% (Figure 2). NFE2 p.R284C with a VAF of 14.9%, and TET2 p.Q622* with a VAF of 9.6% were also detected. CALR p.E380Rfs*51 is a novel driver variant; therefore, any relevant report was not retrieved from the COSMIC. Because of the relatively low sensitivity of Sanger

sequencing (15%–20%), the mutation was eliminated at initial diagnosis. Subsequently, the patient was administered interferon therapy. After 2 months of treatment, the platelet count decreased to 761 × 10⁹/L. The patient then switched to oral hydroxyurea therapy and maintained a stable platelet count that varied in the range of 560–630 × 10⁹/L.

Case 3 (P3) was of a 36-year-old patient with PMF. She was admitted to our hospital because of pleomorphic adenoma, splenomegaly (with a thickness of 6.3 cm), anaemia, increased leukocyte count, platelet count, and lactate dehydrogenase (LDH) level (954 U/L). Additionally, the patient had constitutional symptoms of night sweats and weight loss. She underwent bone marrow aspiration and biopsy. Peripheral blood films and bone marrow tissues confirmed the diagnosis of MPN (Table 2; Figure 1). Common MPN-related mutations were confirmed to be negative using qPCR. Targeted NGS was performed, and the MPL p.W515_Q516del (Figure 2) mutation with a VAF of 68.9% was identified. ASXL1 p.G996Sfs*3 with a VAF of 46.7%, and EZH2 p.C590Y with a VAF of 48.3% were also identified. MPL p.W515_Q516del is a novel driver variant, and has not been reported in the COSMIC. qPCR did not cover this locus; therefore, this mutation was omitted. After 3 months of treatment with JAK2 inhibitor, the platelet count decreased to normal, constitutional symptoms disappeared, and the spleen shrank by more than 50%.

Case 4 (P4) was of a 65-year-old patient with PMF who was diagnosed with triple-negative MPN 3 years ago. He presented to the hospital with fatigue and dyspnoea lasting 8 months. The patient exhibited constitutional symptoms including significant weight loss of 7.5 kg in 20 days. After admission to the hospital, anaemia, increased LDH level (998 U/L), and splenomegaly (with a thickness of 4.7 cm) were observed. Peripheral blood smears and bone marrow tissues revealed the diagnosis the MPN (Table 2; Figure 1). Both qPCR and targeted NGS did not detect JAK2, CALR, or MPL mutations; however, the SH2B3 p.S337Ffs*3 with a VAF of 37.3%, which is associated with JAK/STAT signal transduction pathway, was detected by NGS. The mutation is novel; therefore, any relevant report was not retrieved from the COSMIC. After receiving stimulus to erythrocytes and support treatment, the patient voluntarily left the hospital.

Discussion

JAK2 gene is located on chromosome 9p24 and encodes one of the four non-receptor tyrosine kinases of the Janus kinase (JAK) family which is involved in the JAK/STAT signal transduction pathway. Its abnormalities, such as those due to mutations, loss of heterozygosity (LOH) on the short arm of chromosome 9 (9p LOH), and copy amplification, are common in haematologic tumours, inducing consistent activation of the JAK/STAT pathway and eventual incidence and progression of disease (Kralovics et al., 2005). JAK2 mutations can be found in approximately 98% of PV, 50%–60% of ET, and 50%–60% PMF cases. Common JAK2 mutations are p.V617F (in exon14), whereas a minor number are due to deletion/insertion in exon12 which is clustered in amino acids 535–547, such as p.N542_E543del. Mutations in exon12 can be found in approximately 1%–3% of PV patients, and are very rare in ET and PMF patients (Scott et al., 2007). P1 (a PV patient) possessed JAK2 p.H538_K539delinsQL mutation, and the incomplete covering sites of qPCR led to the omission. We then performed pathogenicity prediction analysis using Mutation taster, a pathogenicity prediction tool; JAK2 p.H538_K539delinsQL is predicted to be deleterious (Schwarz et al., 2014).

Patients who have typical clinical patterns of PV but lack JAK2 V617F or JAK2 exon 12 mutations are extremely rare (Rumi and Cazzola, 2017). Therefore, when facing these patients, clinicians should perform the differential diagnosis again. If the diagnosis is confirmed, then NGS should be employed to determine whether the patient has an unusual JAK2 mutation.

CALR gene is located on chromosome 19p13 and encodes for a multifunctional calreticulin residing in the endoplasmic reticulum and nucleus. Calreticulin cooperates with other molecules to maintain calcium ion homeostasis, and regulate cell proliferation, apoptosis, and migration. CALR mutations mainly include deletion/insertion in exon9, with type 1 (p.L367Tfs*46), and type 2 (p.K385Nfs*47) comprising about 84.7%, while other types are relatively rare. The frame-shift mutation of CALR exon9 conduces a new C-terminal, activating MPL and JAK/STAT pathways that are vital pathogenic factors of MPN. CALR mutations can be found in 20%–30% of ET and 30%–40% of PMF cases, whereas they are rare in PV (Imai et al., 2017; How et al., 2019). The CALR p.E380Rfs*51 of P2 (a ET patient) is a newly occurring type with a low VAF value; as a result, it was omitted by Sanger sequencing. This mutation is predicted to be deleterious using Mutation taster.

Triple-negative ET patients seemed to have better overall survival than driver gene mutated patients. Tefferi A et al. reported that TN ET patients displayed lower incidence of thrombosis compared with JAK2-mutated cases (Tefferi et al., 2014a). In another study, TN ET patients had significantly lower symptom load, and slightly longer survival than mutated cases (Santoro et al., 2022). Considering these differences, it is necessary to apply targeted NGS to detect non-canonical driver variants.

MPL, which is located on chromosome 1p34, encodes thrombopoietin receptor protein (TpoR) and participates in the activation of JAK/STAT signal transduction pathway. MPL mutations are clustered at exon10; W515 and S505 locus missense mutations are the most frequent types, and other types (such as S204, Y591, and R592) are occasionally found. MPL mutations can result in over-activation of JAK/STAT, promoting the occurrence of tumours. MPL mutations can be found in approximately 3%–5% of ET and 5%–10% of PMF cases, while they are rarely seen in PV (Cabagnols et al., 2016; Milosevic Feenstra et al., 2016). The MPL p.W515_Q516del mutation of P3 (a PMF patient) is also a newly occurring mutation type. It was excluded because qPCR did not cover the locus, either. The analysis by Mutation taster showed this mutation leads to amino acid sequence change, whereas it may be benign.

Triple-negative PMF is an aggressive myeloid neoplasm with significantly worse survival than driver gene mutated cases. A study examined the long-term disease outcomes in 428 patients with PMF (Tefferi et al., 2014b). They discovered that TN PMF patients displayed significantly worse survival (median, 2.3 years), compared to that of CALR (15.9 years), JAK2 (5.9 years), or MPL (9.9 years) mutated patients. Leukaemia-free survival (LFS) in PMF was significantly worse in the presence of triple-negative mutational status, either. Therefore, it is important to distinguish between real and pseudo TN PMF. In our study, P3 responded well to the JAK2 inhibitor, also implying that the patient may not be a real TN PMF patient.

To sum up, targeted NGS should be applied to detect non-canonical and low burden driver variants in MPN, such as JAK2 p.H538_K539delinsQL, CALR p.E380Rfs*51, and MPL p.W515_Q516del, to avoid the misdiagnosis of JAK2 unmutated PV and TN MPN.

SH2B3 is located on chromosome 12q24, encoding the LNK protein which can inhibit JAK/STAT signal transduction pathway by directly binding to JAK2 (Tong et al., 2005; Bersenev et al., 2008). SH2B3 mutations occurred in 5%–7% MPN patients, and the majority are frame-shift-truncated and non-sense mutations in the PH and SH2 domains, resulting in loss of function (Lasho et al., 2010; Maslah et al., 2017). SH2B3 p.S337Ffs*3 of P4 leads to the early emergence of stop codon, and shortens the length of mRNA significantly. Consequently, nonsense-mediated mRNA decay occurs, and LNK protein cannot be synthesized. The mutation is clearly deleterious. In summary, SH2B3 mutation is able to relieve the reverse regulating effect of LNK, and over-activate JAK/STAT. As mentioned above, MPN is characterized by elevated JAK/STAT activity, thus SH2B3 mutation may also be a driver mutation of MPN.

In summary, P1 was formerly misdiagnosed as JAK2 unmutated PV, P2 and P3 were formerly misdiagnosed as triple-negative MPN, but all later detected to contain non-canonical driver mutations by targeted NGS. P4 did not have JAK2, CALR or MPL mutations, but was detected to contain a mutation involved in the negative regulation of JAK/STAT pathway by targeted NGS. All these cases imply the important role of NGS in detecting MPN-related mutations. qPCR does not cover the complete loci and the sensitivity of Sanger sequencing is relatively low, consequently, about 5%–10% mutations can be eliminated. NGS can detect both canonical and non-canonical mutations, and the sensitivity of NGS is higher than that of qPCR and Sanger sequencing. Notably, NGS can detect JAK2, CALR, and MPL mutations as well as other mutations (such as SH2B3 and NFE2) that are associated with the JAK/STAT pathway and haematopoiesis regulation, favouring the discovery of new driver mutations in MPN (Jeromin et al., 2016; Zoi and Cross, 2017).

Conclusion

NGS, a more multidimensional and comprehensive gene mutation detection, is required for patients suspected of having MPN to detect non-canonical driver variants and avoid the misdiagnosis of TN MPN. SH2B3 p.S337Ffs*3 can drive MPN occurrence, and SH2B3 mutation may also be a driver mutation of MPN.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JZ and KS contributed equally to this study and should be considered co-first authors. ZH and YW contributed equally to this study and should be considered co-corresponding authors. JZ, KS, MX, and JW analysed and interpreted the data regarding the hematological diseases. KS and JZ wrote the manuscript. ZH and JH managed patients. ZH and YW was responsible for the revision of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from the National Natural Science Foundation of China (grant number: 81873430 to ZH).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1198834/full#supplementary-material

References

Arber, D. A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., et al. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* 127 (20), 2391–2405. doi:10.1182/blood-2016-03-643544

Bersenev, A., Wu, C., Balcerek, J., and Tong, W. (2008). Lnk controls mouse hematopoietic stem cell self-renewal and quiescence through direct interactions with JAK2. *J. Clin. Invest.* 118 (8), 2832–2844. doi:10.1172/JCI35808

Cabagnols, X., Favale, F., Pasquier, F., Messaoudi, K., Defour, J. P., Ianotto, J. C., et al. (2016). Presence of atypical thrombopoietin receptor (MPL) mutations in triple-negative essential thrombocythemia patients. *Blood* 127 (3), 333–342. doi:10.1182/blood-2015-07-661983

Grinfeld, J., Nangalia, J., Baxter, E. J., Wedge, D. C., Angelopoulos, N., Cantrill, R., et al. (2018). Classification and personalized prognosis in myeloproliferative neoplasms. *N. Engl. J. Med.* 379 (15), 1416–1430. doi:10.1056/NEJMoa1716614

How, J., Hobbs, G. S., and Mullally, A. (2019). Mutant calreticulin in myeloproliferative neoplasms. *Blood* 134 (25), 2242–2248. doi:10.1182/blood.2019000622

Imai, M., Araki, M., and Komatsu, N. (2017). Somatic mutations of calreticulin in myeloproliferative neoplasms. *Int. J. Hematol.* 105 (6), 743–747. doi:10.1007/s12185-017-2246-9

Jeromin, S., Kohlmann, A., Meggendorfer, M., Schindela, S., Perglerová, K., Nadarajah, N., et al. (2016). Next-generation deep-sequencing detects multiple clones of CALR mutations in patients with BCR-ABL1 negative MPN. *Leukemia* 30 (4), 973–976. doi:10.1038/leu.2015.207

Khoury, J. D., Solary, E., Abla, O., Akkari, Y., Alaggio, R., Apperley, J. F., et al. (2022). The 5th edition of the world Health organization classification of haematolymphoid tumours: Myeloid and histiocytic/dendritic neoplasms. *Leukemia* 36 (7), 1703–1719. doi:10.1038/s41375-022-01613-1

Kralovics, R., Passamonti, F., Buser, A. S., Teo, S. S., Tiedt, R., Passweg, J. R., et al. (2005). A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N. Engl. J. Med.* 352 (17), 1779–1790. doi:10.1056/NEJMoa051113

Lasho, T. L., Pardanani, A., and Tefferi, A. (2010). LNK mutations in JAK2 mutation-negative erythrocytosis. *N. Engl. J. Med.* 363 (12), 1189–1190. doi:10.1056/NEJMc1006966

Maslah, N., Cassinat, B., Verger, E., Kiladjian, J. J., and Velazquez, L. (2017). The role of LNK/SH2B3 genetic alterations in myeloproliferative neoplasms and other hematological disorders. *Leukemia* 31 (8), 1661–1670. doi:10.1038/leu.2017.139

Milosevic Feenstra, J. D., Nivarthi, H., Gisslinger, H., Leroy, E., Rumi, E., Chachoua, I., et al. (2016). Whole-exome sequencing identifies novel MPL and JAK2 mutations in triple-negative myeloproliferative neoplasms. *Blood* 127 (3), 325–332. doi:10.1182/blood-2015-07-661835

Passamonti, F., and Maffioli, M. (2016). Update from the latest WHO classification of MPNs: A user's manual. *Hematol. Am. Soc. Hematol. Educ. Program* 2016 (1), 534–542. doi:10.1182/asheducation-2016.1.534

Rumi, E., and Cazzola, M. (2017). Diagnosis, risk stratification, and response evaluation in classical myeloproliferative neoplasms. *Blood* 129 (6), 680–692. doi:10.1182/blood-2016-10-695957

Santoro, M., Accurso, V., Mancuso, S., Napolitano, M., Mattana, M., Vajana, G., et al. (2022). Triple-negativity identifies a subgroup of patients with better overall survival in essential thrombocythemia. *Hematol. Rep.* 14 (3), 265–269. doi:10.3390/hematolrep14030037

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: Mutation prediction for the deep-sequencing age. *Nat. Methods* 11 (4), 361–362. doi:10.1038/nmeth.2890

Scott, L. M., Tong, W., Levine, R. L., Scott, M. A., Beer, P. A., Stratton, M. R., et al. (2007). JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. *N. Engl. J. Med.* 356 (5), 459–468. doi:10.1056/NEJMoa065202

Tefferi, A., Guglielmelli, P., Larson, D. R., Finke, C., Wassie, E. A., Pieri, L., et al. (2014a). Long-term survival and blast transformation in molecularly annotated essential thrombocythemia, polycythemia vera, and myelofibrosis. *Blood* 124 (16), 2507–2513. doi:10.1182/blood-2014-05-579136

Tefferi, A., Lasho, T. L., Finke, C. M., Knudson, R. A., Ketterling, R., Hanson, C. H., et al. (2014b). CALR vs JAK2 vs MPL-mutated or triple-negative myelofibrosis: Clinical, cytogenetic and molecular comparisons. *Leukemia* 28 (7), 1472–1477. doi:10.1038/leu.2014.3

Tong, W., Zhang, J., and Lodish, H. F. (2005). Lnk inhibits erythropoiesis and Epo-dependent JAK2 activation and downstream signaling pathways. *Blood* 105 (12), 4604–4612. doi:10.1182/blood-2004-10-4093

Zoi, K., and Cross, N. C. (2017). Genomics of myeloproliferative neoplasms. *J. Clin. Oncol.* 35 (9), 947–954. doi:10.1200/JCO.2016.70.7968

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 27 July 2023
DOI 10.3389/fgene.2023.1216066

OPEN ACCESS

EDITED BY
Yuriy L. Orlov,
I.M.Sechenov First Moscow State Medical University, Russia

REVIEWED BY
Jun Tanihata,
Jikei University School of Medicine, Japan
Kavitha Mukund,
University of California, San Diego, United States
William John Duddy,
Ulster University, United Kingdom
Prech Uapinyoying,
National Institutes of Health (NIH), United States

*CORRESPONDENCE
Stanley F. Nelson,
snelson@mednet.ucla.edu

†PRESENT ADDRESS
Jeremy D. Woods,
Valley Children's Hospital, Madera, CA, United States

RECEIVED 06 May 2023
ACCEPTED 04 July 2023
PUBLISHED 27 July 2023

CITATION
Nieves-Rodriguez S, Barthélémy F, Woods JD, Douine ED, Wang RT, Scripture-Adams DD, Chesmore KN, Galasso F, Miceli MC and Nelson SF (2023), Transcriptomic analysis of paired healthy human skeletal muscles to identify modulators of disease severity in DMD.
Front. Genet. 14:1216066.
doi: 10.3389/fgene.2023.1216066

Transcriptomic analysis of paired healthy human skeletal muscles to identify modulators of disease severity in DMD

Shirley Nieves-Rodriguez[1,2], Florian Barthélémy[2,3], Jeremy D. Woods[4†], Emilie D. Douine[2,5], Richard T. Wang[1,2], Deirdre D. Scripture-Adams[2,3], Kevin N. Chesmore[1,2], Francesca Galasso[1], M. Carrie Miceli[2,3] and Stanley F. Nelson[1,2,5,6]*

[1]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [2]Center for Duchenne Muscular Dystrophy at UCLA, Los Angeles, CA, United States, [3]Department of Microbiology, David Geffen School of Medicine and College of Letters and Sciences, University of California, Los Angeles, Los Angeles, CA, United States, [4]Department of Pediatrics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [5]Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [6]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

Muscle damage and fibro-fatty replacement of skeletal muscles is a main pathologic feature of Duchenne muscular dystrophy (DMD) with more proximal muscles affected earlier and more distal affected later in the disease course, suggesting that different skeletal muscle groups possess distinctive characteristics that influence their susceptibility to disease. To explore transcriptomic factors driving differential gene expression and modulating DMD skeletal muscle severity, we characterized the transcriptome of vastus lateralis (VL), a more proximal and susceptible muscle, relative to tibialis anterior (TA), a more distal and protected muscle, in 15 healthy individuals using bulk RNA sequencing to identify gene expression differences that may mediate their relative susceptibility to damage with loss of dystrophin. Matching single nuclei RNA sequencing data was generated for 3 of the healthy individuals, to infer cell composition in the bulk RNA sequencing dataset and to improve mapping of differentially expressed genes to their cell source of expression. A total of 3,410 differentially expressed genes were identified and mapped to cell type using single nuclei RNA sequencing of muscle, including long non-coding RNAs and protein coding genes. There was an enrichment of genes involved in calcium release from the sarcoplasmic reticulum, particularly in the myofibers and these myofiber genes were higher in the VL. There was an enrichment of genes in "Collagen-Containing Extracellular Matrix" expressed by fibroblasts, endothelial, smooth muscle and pericytes, with most genes higher in the TA, as well as genes in "Regulation Of Apoptotic Process" expressed across all cell types. Previously reported genetic modifiers were also enriched within the differentially expressed genes. We also identify 6 genes with differential isoform usage between the VL and TA. Lastly, we integrate our findings with DMD RNA sequencing data from the TA, and identify "Collagen-Containing Extracellular Matrix" and "Negative Regulation Of Apoptotic Process" as differentially expressed between DMD compared to healthy. Collectively, these findings propose novel

candidate mechanisms that may mediate differential muscle susceptibility in muscular dystrophies and provide new insight into potential therapeutic targets.

KEYWORDS

muscle, transcriptomics, DMD, muscle susceptibility, gene expression, single nuclei RNAseq

1 Introduction

Duchenne muscular dystrophy (DMD) is the most common progressive muscular dystrophy with childhood onset, and is caused by loss of function mutations in *DMD* (Hoffman et al., 1987), leading to profound weakness and premature death, mainly from cardiorespiratory failure. *DMD* encodes dystrophin, which plays a critical structural role in skeletal and cardiac muscle fibers by linking the intra-myofiber F-actin of the Z-disk to the extracellular matrix through binding components of the dystrophin-associated glycoprotein complex at the muscle membrane (Hoffman et al., 1987; Way et al., 1992). Absence of dystrophin in skeletal muscle leads to greater susceptibility to damage from contraction-induced injury (Petrof et al., 1993), resulting in leakage of calcium into the myofiber with a plethora of downstream consequences ultimately leading to myofiber death and replacement with fat and fibrosis. Fibroblasts, immune cells, and muscle stem cells are expanded, changing the extracellular matrix (Scripture-Adams et al., 2022). A large number of other muscular dystrophies have had their genetic basis decoded and many are components of the dystrophin-glycoprotein complex (Cohen et al., 2021), including Limb-girdle muscular dystrophies (LGMDs) with similar patterns of muscle loss from proximal to more distal.

While DMD is always degenerative and leads to premature death, variation in disease progression between individuals in DMD has been used to identify genetic factors correlated with disease severity or progression. Disease severity is mitigated with residual dystrophin expression which usually results in slowing of disease progression (Fanin et al., 1995). However, even in cases of siblings with DMD who have the same *DMD* mutation, there can be discordance in the progression (Pettygrove et al., 2014), indicating that environmental or other genetic factors may modify disease severity. Various studies use variability in age at loss of ambulation (LOA) (Pegoraro et al., 2011; Flanigan et al., 2013; Bello et al., 2016; Weiss et al., 2018; Spitali et al., 2020) to identify variants associated with disease progression.

The overall pattern of sequentially affected muscles in DMD is highly similar across affected individuals and describes a distinctive pattern of progression with more proximal muscles affected earlier than more distal muscles (Rooney et al., 2020), suggesting that constitutive differences in the formation of those muscle groups encode factors that influence relative myofiber susceptibility to damage. Therefore, the study of healthy muscle may provide insights into susceptibility mechanisms in disease. An extreme example of protected striated muscles in DMD across multiple species are the extraocular muscles (EOM) (Karpati et al., 1988; Valentine et al., 1990; Kaminski et al., 1992). However, the functional requirements of EOM are substantially different from limb skeletal muscle. EOM have multiple innervated fibers, compartmentalization of layers with different fiber types, expression of EOM-specific myosin isoforms (encoded by *MYH13*, *MYH15*), and partial retention of embryonic and neonatal myosin expression in mature fibers (Porter, 2002). Differential gene expression studies in mouse (Porter et al., 2001) and rat (Fischer et al., 2002) highlighted calcium homeostasis, mitochondrial genes, lipid catabolism, immune processes, apoptosis, and extracellular matrix.

Here we study healthy muscle tissue from vastus lateralis (VL) and tibialis anterior (TA), to identify genes that alter myofiber susceptibility to fibrofatty replacement in DMD individuals, using paired samples from 15 donors to control for interindividual and age differences. While TA has a much more modest degree of protection from disease progression than EOM, TA is substantially and consistently protected from ongoing muscle damage in DMD relative to VL from longitudinal imaging and spectroscopy data of children with DMD (Rooney et al., 2020). We reasoned that the differential expression analysis of VL and TA in healthy individuals would provide insight into protective mechanisms relevant in the absence of dystrophin. The difference in progression is substantial. VL progresses faster than TA with an about 8.5-year longer time for the TA to attain similar levels of damage as the VL (Rooney et al., 2020). In this transcriptomic study, we sampled VL and TA at a single timepoint from healthy young adult volunteers. We report differentially expressed genes and map differentially expressed genes to specific intra-muscular cell types using single nuclei analyses.

2 Materials and methods

2.1 Muscle biopsies

Fifteen healthy individuals (age range 18–26 years) with no history of muscle or other chronic or acute disease were consented on UCLA protocol IRB#18-001366. Eight ambulatory DMD patients with a confirmed nonsense *DMD* mutation were consented on UCLA protocol IRB#11-001087 (age range 2–7 years). All biopsies were obtained using a Vacora (Bard) vacuum-assisted core needle from the VL and TA as previously described (Barthelemy et al., 2020). In brief, before the biopsies, the participant's leg was observed via ultrasound to ensure that the muscle showed no excess fat or blood vessels nearby. VL sample was obtained from about two-thirds of the muscle length, and the TA from about one-third of the muscle length. We chose muscle pieces that had similar muscle appearance without visible connective tissue to reduce sample variability. Each needle muscle biopsy core (about 125 mg) was dissected into about 25 mg pieces and flash frozen in liquid nitrogen within tissue cassettes within 5 min of excision and stored in liquid nitrogen until RNA extraction or sectioning for histological examination.

2.2 RNA sequencing

Frozen skeletal muscle (8–25 mg) was homogenized on ice in 500 µL of Trizol for RNA extraction using standard protocols (Lee et al., 2020). RNA quality was recorded by the RNA integrity number (RIN) using the Agilent RNA 6000 Nano chips. Healthy muscle RNA samples with RIN above 7 and DMD muscle RNA samples with RIN above 4 were used to prepare cDNA libraries with ribosomal RNA depletion using the KAPA RiboErase Kit (HMR) (Roche). About 50 million 150-151 bp paired-end RNA sequencing (RNAseq) reads were generated per RNA sample using Illumina Novaseq 6000 S4. Sequencing reads were aligned to GRCh38 (Ensembl 105, Gencode v39) using STAR 2.6.0c (Dobin et al., 2012; Lee et al., 2020). Data quality control included alignment metrics (ribosomal and globin RNA, aligned and unmapped reads, sequencing depth), hierarchical clustering, principal component analysis and Pearson correlation.

2.3 Single nuclei isolation and RNA sequencing

Single nuclei were isolated from a subset of 3 paired male healthy VL and TA frozen muscle and sequenced using the 10X Genomics platform as described previously (Scripture-Adams et al., 2022). Six to twelve 40 µM cross sections of frozen muscle biopsies were collected in a sterile tube to estimate a total of 3 mg of skeletal muscle, dounced with two cycles of strokes (one with a loose douncer followed by one with a tight douncer) in 1% bovine serum albumin (BSA) in phosphate-buffered saline (PBS) with 100 U/mL of type IV collagenase and 0.5 U/µL RNAse inhibitor, and stained with 10 µg/mL DAPI. The nuclei were sorted by fluorescence-activated cell sorting (FACS) to separate from debris and create a pure nuclear preparation prior to library preparation. 10X Chromium Single cell 3′ v3 libraries were prepared and sequenced on Illumina Novaseq 6000 S2 (2 × 50 bp) (10X Genomics). Single nuclei RNA sequencing (snRNAseq) reads were aligned to GRCh38 (Ensembl 105, Gencode v39) using Cell Ranger (10X Genomics). Data was aggregated for downstream processing and analysis. Initial cell clustering was performed using k-means within Cell Ranger (10X Genomics). Nuclear doublets were identified using DoubletFinder (version 2.0.3) (McGinnis et al., 2019) with a doublet rate of detection of 15%. Doublets as well as nuclei with 200 or fewer unique molecular identifiers (UMI), were excluded from downstream analysis. Re-clustering was performed after data filtering, and clustered nuclei populations were identified using known cell-type marker genes via Loupe Browser (version 6.0.0) (10X Genomics). Downstream analysis and statistical testing of differentially expressed genes across cell types was performed using the R package Seurat (version 4.0.2) (Hao et al., 2021) and the Wilcoxon statistical test. UMI-normalized average expression across cell types was obtained from Seurat's AverageExpression function, which returns the average number of transcripts per 10,000 transcripts (TP10K).

2.4 Cell deconvolution using single nuclei RNA sequencing

Raw bulk RNAseq read counts for were obtained from the STAR alignment (version 2.6.0c) (Dobin et al., 2012) and batch-corrected for the two sequencing runs using CombatSeq (sva version 3.38.0) (Zhang et al., 2020). Differential gene expression analysis across cell types in the snRNAseq dataset identified statistically significant (adjusted p-value <0.05) marker genes for each cell type. Highly specific markers for a specific cell type were defined as those expressed in less than 10% (for large cell populations) or 1% (for small cell populations) of the other cell types. A list of 69 cell-specific genes was obtained after further manual curation. Estimated cell proportions for each sample were obtained with CIBERSORTx (Newman et al., 2019) using the average expression of these 69 cell-specific genes. The parameters used for CIBERSORTx were: Job type = Impute Cell Fractions, Batch correction = disabled, Disable quantile normalization = true, Run mode = relative, Permutations = 100.

2.5 Differential gene expression analysis

The R package DESeq2 (version 1.30.1) (Love et al., 2014) was used to perform differential gene expression analysis using the raw read counts. The covariates included in the healthy VL *versus* TA analysis design were: participant study ID, RIN, and batch. The covariates included in the DMD *versus* Healthy analysis design were: batch, RIN, age, and sex. Multiple testing adjustment was done within DESeq2 using Benjamini–Hochberg for a false discovery rate (FDR) of less than 0.05.

Functional enrichment analysis of differentially expressed genes was performed for all differentially expressed genes (independent of their direction of highest expression) using EnrichR (https://maayanlab.cloud/Enrichr/, (Chen et al., 2013)), with all expressed genes included in the DESeq2 analysis as background. For the genes differentially expressed between VL and TA, we tested 4,701 terms from GO Biological Process 2023, and 408 terms from GO Cellular Component 2023. For genes differentially expressed between DMD and healthy, we tested 3,133 terms from GO Biological Process 2023, and 272 terms from GO Cellular Component 2023. Significant gene ontology (GO) terms (adjusted p-value <0.05) for the VL *versus* TA analysis were further summarized with ReviGO (http://revigo.irb.hr/, (Supek et al., 2011)) with the following parameters: dispensability threshold = 0.5, GO metric = adjusted p-value (lower value is better), remove obsolete GO terms = yes, species = Whole UniProt database, similarity measure = SimRel.

ENCODE_and_ChEA_Consensus_TFs_from_ChIP-X enrichment category within EnrichR (Chen et al., 2013) was used to identify transcription factors (104 transcription factors tested) that putatively bind to the differentially expressed genes. Pathway enrichment of druggable genes higher in the VL was performed using EnrichR (Chen et al., 2013) KEGG 2021 Human enrichment category (250 terms tested).

2.6 Differential isoform usage analysis

The VL and TA raw data was aligned using the Kallisto app (Kallisto Quantification version 2.0.2, Kallisto 04.46.1) on the DNAnexus platform pipeline (Bray et al., 2016) to obtain counts and relative abundance (TPM) for each transcript (Gencode v39, Ensembl 105).

Differential isoform usage analysis between VL and TA was performed using the IsoformSwitchAnalyzeR (version 1.12.0) R package (Vitting-Seerup and Sandelin, 2017). The design matrix included: sex, sample RIN, and batch. Gencode v39 primary assembly annotation and transcripts were used to generate the switch list. Isoforms were prefiltered before testing for differential isoform usage using the following parameters: gene expression cutoff = 0.1 and isoform expression cutoff = 0, and genes with only one isoform were excluded. Isoform switch testing was done using DEXSeq (version 1.36.0) (Anders et al., 2012) within the IsoformSwitchAnalyzeR package, and correction for confounding factors indicated in the design matrix was performed simultaneously via the limma package (version 3.46.0) (Ritchie et al., 2015). The isoform switch analysis was limited to: switching genes (genes with at least one isoform significantly differentially used), genes with consequence potential (with isoforms differentially used in opposing directions, i.e., one with increased and one with decreased usage), and isoforms with at least two isoforms significantly differentially used (alpha<0.05), and a difference in isoform usage between muscles of at least 0.01 (1%).

2.7 Immunofluorescence

Skeletal muscle tissue was cross-sectioned at 10 µM thickness after equilibration at −22°C in a cryostat and then stored at −80°C until immunofluorescence was performed. Slides were acclimated to room temperature and sections were circled using a hydrophobic barrier pen. For actinin-3 staining, sections were fixed with PFA 4% for 10 min and permeabilized using 0.5% Triton-X for 10 min at room temperature. Sections were treated with TrueBlack Lipofuscin Autofluorescence Quencher before blocking with 3% BSA/10% goat serum in PBS for 1 h at room temperature, and then incubated in primary antibody in blocking solution overnight at 4°C in a humidified chamber. For nebulin staining, unfixed samples were permeabilized with 0.5% Triton-X for 10 min at room temperature. Samples were blocked with 3% BSA for 1 h at room temperature and incubated in primary antibody solution in 3% BSA overnight at 4°C in a humidified chamber. Primary antibodies used were: monoclonal anti-actinin-3 (Abcam, ab68204, 1.61 µg/mL), monoclonal anti-myosin skeletal slow (Sigma, m8421, 4.8 µg/mL), mouse anti-NEB143(3F4) ((Lam et al., 2018), 149 µg/mL), rabbit anti-MYH1 (Sigma, SAB2104768, 5–10 µg/mL), rat anti-MYH2 clone 8F72C8 (EMD Millipore, MABT848, 40 µg/mL). Sections were then incubated in secondary antibody in PBS for 2 h at room temperature. For nebulin staining, secondary antibody for rat and mouse were cross adsorbed to prevent cross-reactivity: Goat anti-Mouse IgG (H + L) Cross-Adsorbed Secondary Antibody, DyLight 550 (Invitrogen, SA5-10173, 1:500), and Cy5 AffiniPure Donkey Anti-Rat IgG (H + L) (Jackson Laboratories, 712-175-153, 1:300). Slides were mounted in Antifade Mounting Medium with DAPI (Vectashield, H-1200-10). Images were obtained using a fluorescent microscope and processed using ImageJ (Schneider et al., 2012) (release 1.53c).

3 Results

3.1 Identification of transcriptional differences between VL and TA

Because of the substantial difference in the rate of progression/damage of VL and TA in DMD (Figure 1A), we characterized the intrinsic transcriptomic profiles of paired healthy VL and TA using RNAseq and snRNAseq to reveal candidate mechanisms that may underlie this differential susceptibility to DMD (Figure 1B). VL and TA biopsies were sampled from each of 15 healthy young adults during the same procedure. Extraction of RNA from frozen skeletal muscle was adequate with an average RIN of 8 across all samples (range 7.1–8.7), and an average of 54 million sequencing reads were obtained per sample (range 45–76 million reads). One sample that had lower sequencing depth, and two samples that were outliers by hierarchical clustering and had relatively lower correlation with the overall dataset were excluded. A total of 27 healthy muscle samples (26 paired VL-TA, 1 unpaired VL) were used for further analysis. Two-dimensional principal component analysis (PCA) on the expression of all 22,414 expressed genes among 27 samples demonstrates that RNAseq data cluster predominantly by muscle type and that muscles from the same individual do not cluster together (Figure 1C). This indicates that there is more expression similarity between unrelated individuals in either the VL or TA than within an individual, or alternatively stated there are more intraindividual gene expression differences between VL and TA than interindividual differences from genetic variation.

Using DESeq2 (Love et al., 2014) with a paired analysis design, we identified a large set of 3,410 significantly differentially expressed genes (Supplementary Table S1), or 15.2% of all genes, demonstrating a substantial number of gene expression differences between the skeletal muscle groups. When we randomize participant IDs within muscle groups such that samples are no longer paired, and test for differential gene expression, we do not observe as many differentially expressed genes as we do with a paired analysis (empirical p-value = 0, n = 1,000 permutations). That is, our paired analysis of both muscles from the same individuals allowed us to identify a larger number of differentially expressed genes than we would have with an unpaired design. The most statistically significant differentially expressed gene was the transcription factor *ZNF385A*, and other top differentially expressed genes (by fold difference or p-value) included *MYH1*, *COL22A1*, the transcription factors *BNC2*, *SIM2* and *ZNF273*, and the non-coding RNAs lnc-HLCS-1, lnc-CES1-7, lnc-APOB-2, and lnc-RORA-1 (Figure 1D). Genes classified as protein coding by the Ensembl automatic annotation system were more likely to be differentially expressed, comprising 82% of all differentially expressed genes, but a substantial number of long noncoding RNAs (lncRNAs) are differentially expressed between the muscles (Supplementary Table S1).

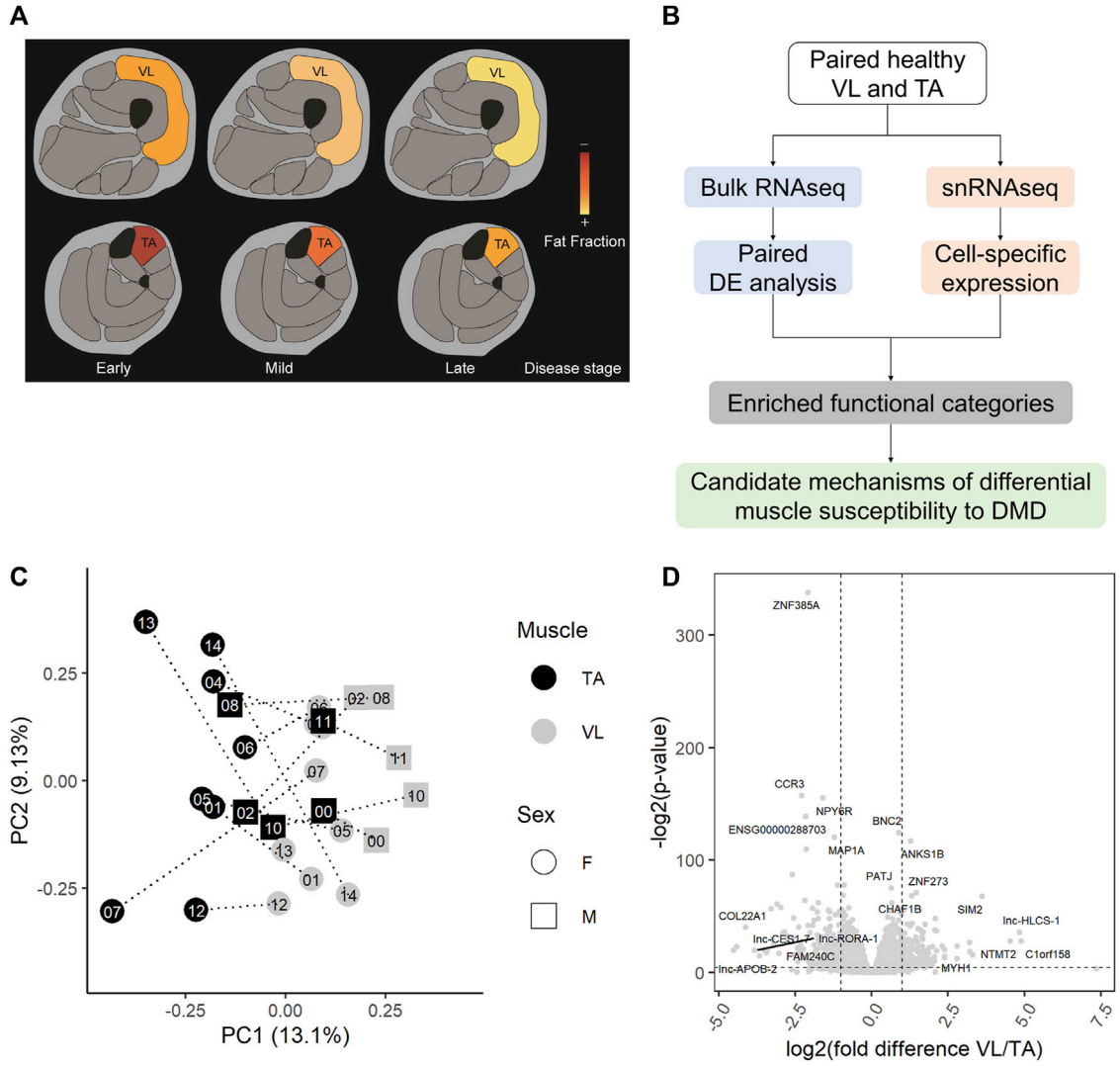


FIGURE 1
Healthy VL and TA transcriptomes are highly different. **(A)** Representation of the differential progression of vastus lateralis (VL) and tibialis anterior (TA) in DMD. The color scale indicates the progression from early stages of DMD where muscle fat fraction is minimal (illustrated in red), to late stages where muscle fibers are completely replaced by fat (illustrated in yellow). **(B)** Workflow for the identification of candidate mechanisms mediating differential muscle susceptibility to DMD. **(C)** The first two principal components (PC1 and PC2) are shown for batch-corrected normalized RNAseq data expression of all expressed genes (n = 22,414) across the 27 muscle samples. **(D)** Volcano plot for all 22,414 genes tested for differential expression. Dashed lines depict a fold change of 2 and a p-value of 0.05. For each muscle, the top 5 differentially expressed genes by fold difference and the top 5 genes by p-value are labeled.

Gene ontology analysis on all differentially expressed genes revealed an enrichment of 619 biological processes and 85 cellular component categories (Supplementary Table S2A). After summarization of redundant terms with ReviGO, the summarized GO term list is comprised of 75 biological processes and 28 cellular component GO terms (Supplementary Table S2B). We further focused on GO terms with over 10 genes, such that we could examine a larger number of genes contributing to the enrichment, resulting in a list of 58 biological processes and 21 cellular components. For each GO category, we ranked them by adjusted p-value, and then selected 6 relevant categories (Table 1) based on their involvement in muscle function and the dystrophic pathology.

3.2 Cell type differences in VL and TA

Fiber type composition varies between skeletal muscles in mice (Terry et al., 2018) and humans (Abbassi-Daloii et al., 2023). In humans, VL has a larger portion of fast myofibers than TA (Edgerton et al., 1975; Jakobsson et al., 1991), whereas in mouse, the TA is composed entirely of fast myofibers (Hämäläinen and Pette, 1993; Scripture-Adams et al., 2022). In DMD, the differential disease susceptibility between different skeletal muscles has been partly attributed to the higher proportion of fast fibers, which are more susceptible to damage in the disease course than slow fibers (Webster et al., 1988). These differences in cell composition may contribute to the differential disease susceptibility and be reflected in

TABLE 1 Differentially expressed genes are enriched within regulation of calcium, extracellular matrix and regulation of apoptosis. Significant gene ontology (GO) terms enriched among all 3,410 differentially expressed genes (independent of the muscle where they are highest expressed) are shown. EnrichR significant GO terms for biological process (BP) and cellular component (CC) were summarized using ReviGO. Selected most relevant and significant terms with over 10 genes are shown. The top 15 genes by average fold difference between VL and TA are shown.

GO	GO term	Count	Adjusted *p*-value	Odds ratio	Top 15 genes by average fold difference
BP	Cytoplasmic Translation	71	9.77E-36	18.28	*RPS15A, RPL21, RPL35, RPL9, RPL39, RPL6, RPL29, RPS3A, RPL7, RPL27, RPLP0, MRPS12, RPS18, RPS13, RPS15*
	Muscle Contraction	48	7.56E-15	6.92	*MYH1, RYR2, MYL6B, MYH11, MYH6, TPM1, MYH2, MYLK, MYLK2, MYH3, MYH4, TPM4, OXTR, MYL1, TPM3*
	Regulation Of Cell Migration	121	2.46E-14	2.75	*TBX5, EPPK1, TNC, FGF9, CCR1, NKD1, PLXNA4, SERPINE1, NTRK3, SH3RF2, STC1, SFRP1, TPM1, TWIST2, PAK1*
	Regulation Of Cell Adhesion	45	3.97E-07	3.41	*TNC, DACT2, PLXNA4, TPM1, PLXNB1, ADAM22, DLL1, SRC, PDE3B, MYADM, EPHA4, EPHA2, TGFBI, PPP3CA, TGM2*
	Regulation Of Apoptotic Process	139	4.96E-06	1.76	*EGR3, COMP, ACTN3, EGR1, SH3RF2, ANGPTL4, FRZB, GATA6, SFRP1, MLLT11, ACTN1, TENT5B, GADD45G, MPO, SMAD6*
	Regulation Of Release Of Sequestered Calcium Ion Into Cytosol By Sarcoplasmic Reticulum	13	9.98E-05	9.05	*RYR2, CASQ1, CASQ2, SLC8A1, GSTO1, ANK2, CACNA1C, DMD, CALM2, CALM1, TRDN, PDE4D, ATP1A2*
CC	Focal Adhesion	156	7.17E-33	4.1	*CNN1, ACTN3, TNC, CSRP1, ACTN1, SLC9A1, SPRY4, FLNA, MCAM, BCAR3, ITGA5, THY1, LAYN, CD9, SRC*
	Large Ribosomal Subunit	41	3.52E-24	28.79	*RPL21, RPL35, RPL9, RPL39, RPL6, RPL29, RPL7, RPL27, RPLP0, RPL38, RPL37A, RPL7A, RPL4, RPL14, RPL24*
	Actin Cytoskeleton	105	8.76E-17	3.18	*CNN1, ACTN3, SORBS2, TPM1, ACTN1, PAK1, FLNA, MYLK, ABLIM1, CD274, MYL2, CNN2, TPM4, MYADM, MYLK3*
	Collagen-Containing Extracellular Matrix	88	2.10E-09	2.43	*ACAN, COMP, TNC, LEFTY2, COL28A1, COL21A1, CCN2, SERPINE1, COLQ, ANGPTL4, INHBE, SLPI, SFRP1, SERPINB8, NCAM1*
	Sarcoplasmic Reticulum	25	3.30E-09	7.76	*RYR2, ATP2A1, SLN, THBS1, ATP2A2, DMPK, CASQ1, ATP2A3, STRIT1, JPH1, CASQ2, JSRP1, ITPR1, S100A1, ITPR3*
	Sarcolemma	21	6.01E-05	4.04	*RYR2, ATP1A1, CAV2, SLC8A1, CACNG1, ANK2, SLC2A5, DMD, CAV3, SGCB, SYNC, CAV1, POPDC3, RYR1, DYSF*

Italic values are the gene names (gene symbols).

the observed transcriptomic differences. To assess the contribution of cell composition to gene expression, we performed snRNAseq of nuclei dissociated from a small subset of the healthy individuals. After excluding doublets and nuclei with less than 200 UMI, the VL and TA dataset consists of 14,887 single nuclei (5,151 VL and 9,736 TA) with a median of 386 genes and 568 UMI per nucleus, and with a total of 25,248 genes detected within all of the nuclei.

Clustering analysis resulted in the identification of 8 known major cell types (Figure 2A) with distinct transcriptomes (refer to Supplementary Table S2C for a list of positive marker genes). We compared the proportion of each major cell type within VL and TA. Consistent with previous reports, the three VL samples had a higher proportion of fast fibers compared to TA (paired *t*-test p = 4.25E-02, average fold difference 1.82) (Figure 2C). The 3 TA samples had 1.29 times as many slow fibers, although this difference was not statistically significant in our snRNAseq dataset (paired *t*-test p = 2.52E-01) (Figure 2C). Performing snRNAseq on a subset of samples allows us to infer cell composition in bulk RNAseq. By integrating bulk RNAseq and snRNAseq, we can explore the transcriptomic profile of our large dataset taking into consideration if the gene is specifically expressed in just 1 cell type, and thus map the differential expression of some genes to cell type. The snRNAseq dataset was also used to infer the percentage of all major cell types across our larger bulk RNAseq dataset. For this, we used CIBERSORTx (Newman et al., 2019) to deconvolute the bulk RNAseq data with 69 marker genes that we identified and define as being uniquely expressed within only one of the 8 major cell types (Figure 2B). Overall, the percentage of cell types inferred by CIBERSORTx agreed with those observed by snRNAseq in the six samples with both data types (Pearson correlation = 0.81), and we can infer from bulk RNAseq that TA has a larger portion of slow fibers than VL (paired *t*-test p = 3.69E-05, average fold difference 1.54) and VL has a higher portion of fast fibers (paired *t*-test p = 1.93E-04, average fold difference 2.07) (Figure 2C). We also observed a slight but significant increase in the percentage of endothelial (paired *t*-test p = 4.40E-03, average fold difference 1.16), pericytes (paired *t*-test p = 7.22E-03, average fold difference 1.14), and immune (paired *t*-test p = 3.63E-03, average fold difference 1.16) cells in TA compared to VL. The higher percentage of endothelial cells and pericytes in the TA is suggestive of a higher capillarity density compared to the VL. Similar differences in capillarity density across leg muscles have been reported previously, with a higher density in the lower leg gastrocnemius lateralis compared to the upper leg semitendinosus muscle (Abbassi-Daloii et al., 2023).

10.3389/fgene.2023.1216066

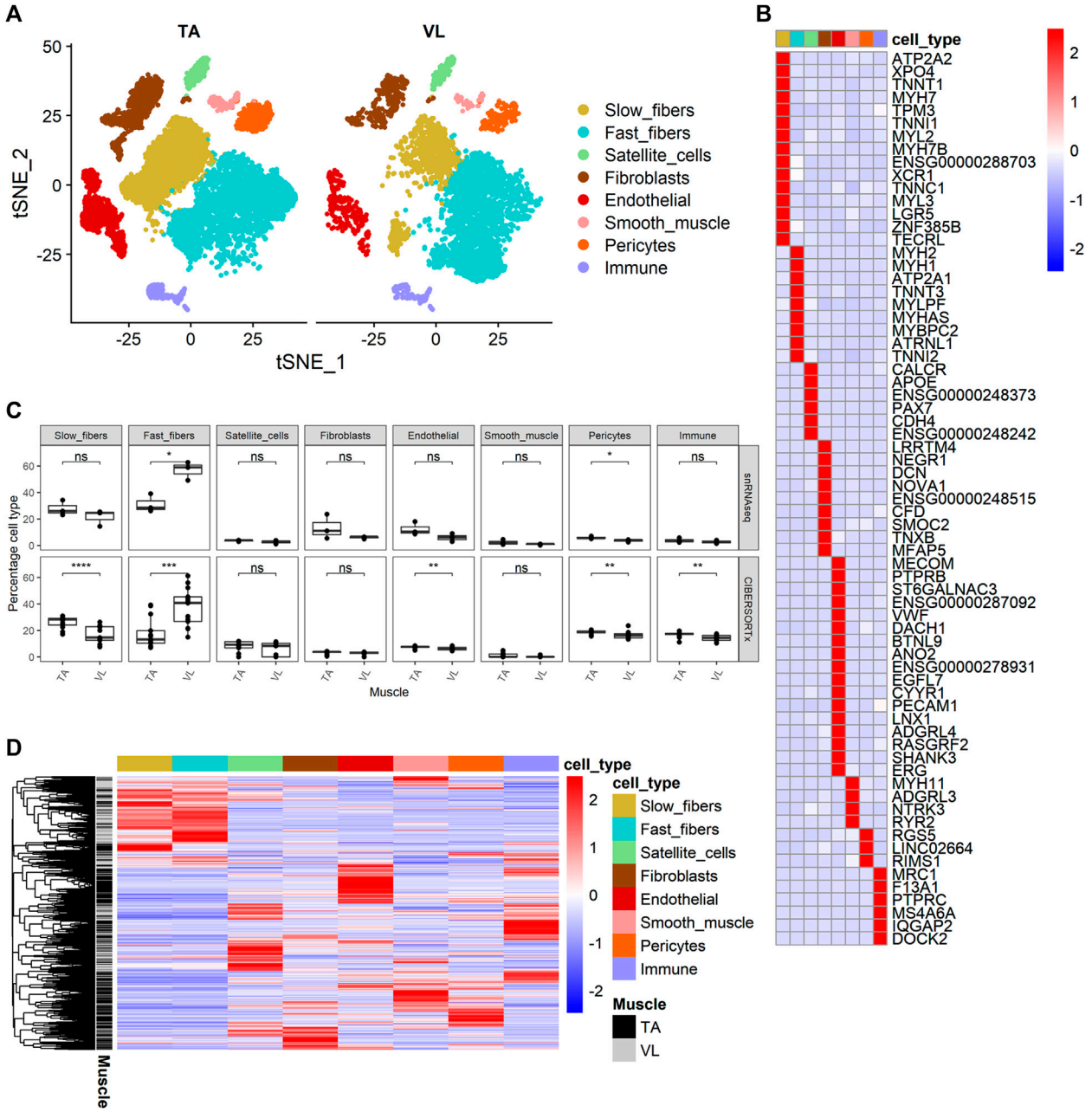


FIGURE 2
Identification of cell composition differences and cell source of differential gene expression. **(A)** t-SNE projection of 14,887 nuclear transcriptomes from 3 TA and 3 VL paired healthy muscles. **(B)** Heatmap of expression of 69 cell-specific marker genes of the major cell types identified. **(C)** Percentage of cell types from snRNAseq (top) and percentage of cell types inferred by CIBERSORTx (bottom) for paired VL and TA samples using the 69 specific markers in **(B)**. Paired *t*-test, *: p ≤ 0.05, **: p≤0.01, ***: $p < 0.001$,****: $p < 0.0001$. **(D)** Heatmap of expression of 3,221 genes differentially expressed between VL and TA in the main cell types identified in (A).

3.3 Mapping of differentially expressed genes to specific cell types within skeletal muscles

We next sought to identify the cellular source of differentially expressed genes identified by bulk RNAseq by interrogating their relative expression across the 8 major cell types identified in healthy human muscle. Out of 3,410 differentially expressed genes observed within the bulk RNAseq data, 3,221 (94.4%) were also observed in our snRNAseq dataset (Figure 2D). Hierarchical clustering of their expression shows that the differentially expressed genes are typically not expressed in all cell types, but rather the vast majority are observed to have much higher expression in 1 cell type.

475 (14.75%) and 327 (10.15%) of the differentially expressed genes have the highest average expression in the fast and slow myofibers, respectively. Considering that the VL has a higher proportion of fast myofibers than TA, we expected to observe many genes that are higher in VL from the bulk RNA analysis to

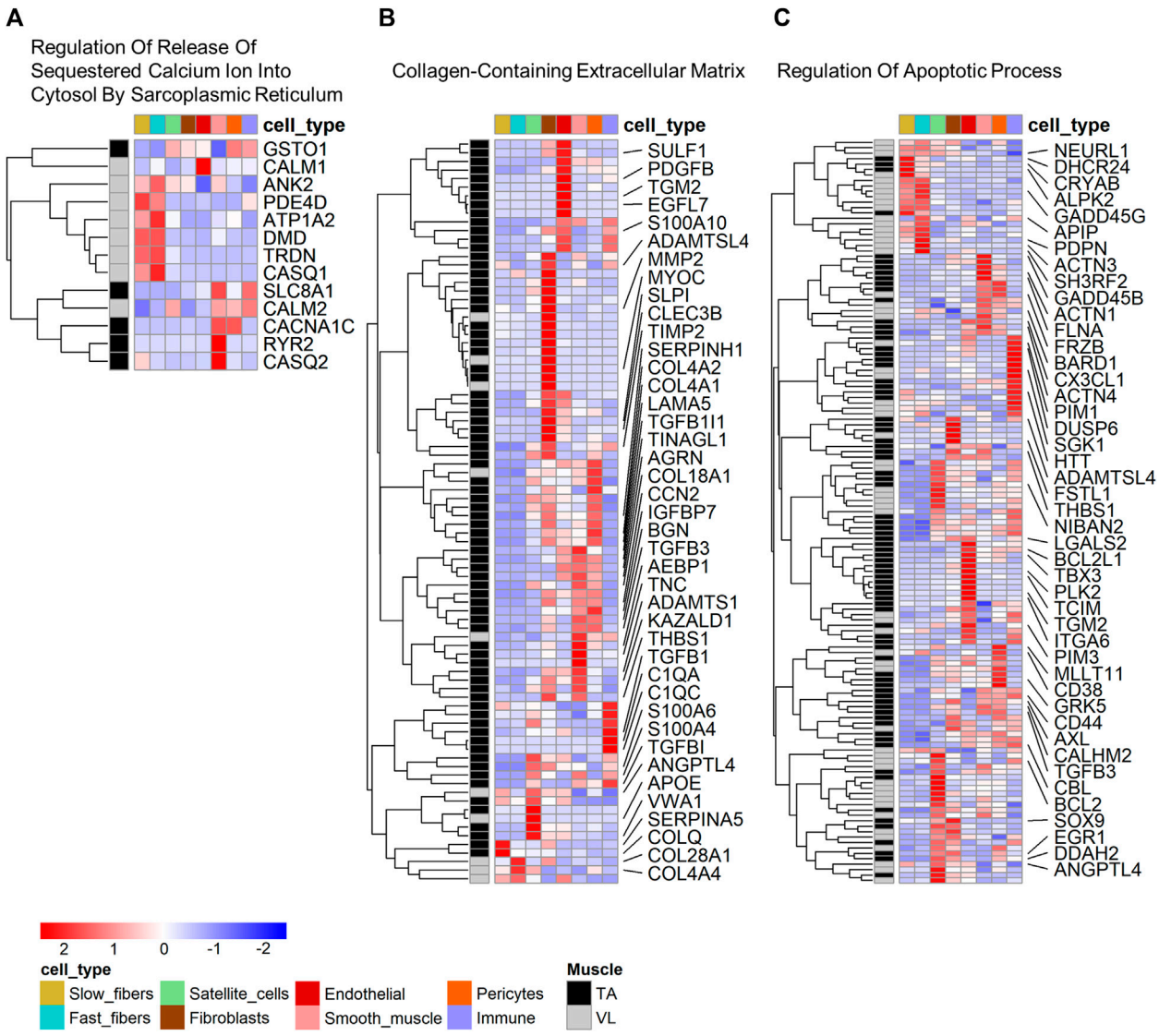


FIGURE 3
Cell type expression of genes within enriched gene ontology categories. Heatmap of expression of genes that are both differentially expressed and identified within select gene ontology categories shown in Table 1: **(A)** Regulation Of Release Of Sequestered Calcium Ion Into Cytosol By Sarcoplasmic Reticulum, **(B)** Collagen-Containing Extracellular Matrix, and **(C)** Regulation Of Apoptotic Process. For B and C, only select genes with a fold difference above 1.3 and average TPM in the muscle where it has highest expression above 5 are named. All data are present within Supplementary Tables S2A, B.

be higher because they are expressed in fast fibers, and those higher in TA to be highest expressed in slow fibers. In line with this, we observed that differentially expressed genes that are higher in VL are more often expressed highest in fast fibers (445 genes, or 80.0% of the 556 genes higher in VL with highest expression in the myofibers) and conversely, differentially expressed genes that are higher in TA are most often restricted in their expression in slow fibers (216 or 87.8% of the 246 genes higher in TA with highest expression in the myofibers). However, there are exceptions to this expected pattern of expression based on the higher proportion of fast fibers in VL compared to TA (Supplementary Figure S1), and these may indicate shifts in metabolic phenotype within each muscle type. For instance, 111 of 1,559 genes higher in VL are most highly expressed in slow fibers (Supplementary Figure S1A) and 30 of 1,851 genes higher in TA have highest expression in fast fibers (Supplementary Figure S1B). Interesting exceptions include *CAMK2A* (encoding CaMKIIα) which is among the genes higher in the VL with higher expression in slow fibers than fast fibers. Conversely, *IRX3*, encoding iroquois homeobox 3, which has been linked to body weight (Gholamalizadeh et al., 2019), has ten-fold higher expression in fast fibers compared to slow fibers, but is higher in the TA muscles.

Remarkably, the remaining 75.1% of the differentially expressed genes have highest expression in other muscle resident cell types that are not the myofibers. Despite satellite cells, endothelial, smooth muscle and fibroblasts accounting for 6.60%, 7.01%, 0.69% and 3.17% of total cell population in both VL and TA, the percentage of

differentially expressed genes with highest expression in these cell types were 15.71%, 14.96%, 10.87% and 8.20%, respectively, demonstrating differences in virtually all cells between skeletal muscle groups. Pericytes accounted for 17.57% and immune cells for 15.79% of all cells but had fewer genes that were detected as differentially expressed, 9.87% and 15.49%, respectively.

To determine which cell types have the highest expression of the differentially expressed genes within functional categories, we mapped cell expression using the single nuclei data (Supplementary Table S1). Eight of 13 (61.54%) genes in "Regulation Of Release Of Sequestered Calcium Ion Into Cytosol By Sarcoplasmic Reticulum" (Figure 3A) were higher in the VL, and these have highest expression predominantly in fast fibers (5 genes, 38.46%). These include *CALM1* and *CASQ1*, encoding calmodulin 1 and calsequestrin 1, respectively. Only *PDE4D* has highest expression in slow fibers, which we infer is differentially expressed independent of the differences in fiber type composition.

Among the genes in "Collagen-Containing Extracellular Matrix", 78 of 88 (88.64%) genes have highest expression in the TA. Genes in this category have mapped highest expression in fibroblasts (26 genes, 29.55%), smooth muscle (14 genes, 15.91%), endothelial (13 genes, 14.77%) and pericytes (13 genes, 14.77%) (Figure 3B). "Collagen-Containing Extracellular Matrix" genes include metalloproteinase-2 (*MMP2*) which is responsible for remodeling the muscle extracellular matrix, a process important for proper satellite cell migration and differentiation (Chen and Li, 2009), along with tissue inhibitors of metalloproteinases, such as *TIMP*1 and *TIMP3*. Only 2 of the genes higher in TA (2.56%) have highest expression in myofibers, specifically in slow fibers (Figure 3B).

Genes in "Regulation Of Apoptotic Process" are more broadly expressed across all cell types (Figure 3C), suggesting that differential regulation of cell death is a characteristic of all cells in the VL and TA due to muscle origin. 22 genes have highest expression in the myofibers, including the heat shock protein *CRYAB* higher in the TA, mapping to the slow fibers and also annotated in the biological process category "Negative Regulation of Apoptotic Process", which is also enriched among the differentially expressed genes (Supplementary Table S2A). Among genes with higher expression in other muscle resident cells, the widely studied anti-apoptotic BCL-xL/*BCL2L1*, higher in the TA, was highest expressed in endothelial cells. Despite not being annotated in "Regulation Of Apoptotic Process", Hzf (encoded by *ZNF385A*) has been linked to negative regulation of apoptosis. In conditions of DNA-damaging stress, Hzf induction and binding to p53 modulates p53-mediated transcription such that the expression of pro-arrest p53 target genes is preferentially activated over pro-apoptotic p53 target genes (Das et al., 2007). *ZNF385A* is the most statistically significant gene with a 4.2X higher expression in TA (p = 1.76E-102) (Figure 1D), and has the highest expression in pericytes (average expression 0.31 TP10K) and similar levels of expression in fast and slow fibers (average expression 0.03 TP10K in both fiber types). These data suggest that the VL and TA have differential regulation of apoptotic signaling, with a potentially superior negative regulation in the TA that may be protective in DMD.

3.4 Search for transcription factors that may underlie the differential gene expression

Using the ENCODE_and_ChEA_Consensus_TFs_from_ChIP-X enrichment category within EnrichR (Chen et al., 2013), we identified 45 transcription factor genes that are reported to bind to multiple differentially expressed genes between VL and TA, and these may thus regulate the differentially expressed genes (Supplementary Table S2D). Because some transcription factors can act as both positive and negative regulators of expression, we searched for transcription factors that bind upstream of all differentially expressed genes, independently of the muscle in which they are highest expressed. Among these 45 transcription factors, 38 were expressed in the VL/TA bulk RNAseq dataset, and 11 were differentially expressed. The top 6 by p-value are: *TP63*, *AR*, *GATA2*, *KLF4*, *SMC3*, and *SMAD4* (Supplementary Table S2D). For each transcription factor, the putative target genes are listed in Supplementary Table S2D. These genes are further categorized in "Regulation Of Release Of Sequestered Calcium Ion Into Cytosol By Sarcoplasmic Reticulum", "Collagen-Containing Extracellular Matrix", and "Regulation Of Apoptotic Process" by the muscle in which they are highest expressed and listed in descending fold difference between the muscles (Supplementary Table S2D).

The 11 differentially expressed transcription factors were detected by snRNAseq (Supplementary Figure S2). The transcription factors higher in TA (*ZMIZ1*, *FOSL2*, *GATA2*, *KLF4* and *EGR1*) are mainly expressed in non-myolineage cell types, and mainly in fibroblasts and endothelial cells, consistent with a potential role regulating the extracellular matrix gene expression in non-myolineage cells. Among the differentially expressed genes in "Collagen-Containing Extracellular Matrix" and higher in TA, the metalloprotease *MMP2* is a target of GATA2 (Supplementary Table S2D). The transcription factors higher in VL are expressed in myolineage and non-myolineage cell types. *TP63* is restricted to the myofibers, and highest in fast fibers, whereas *AR* is highest expressed from satellite cells. Among the differentially expressed genes in "Regulation Of Release Of Sequestered Calcium Ion Into Cytosol By Sarcoplasmic Reticulum" is the TP63 target *ATP1A2* (Supplementary Table S2D), highest expressed in fast fibers (Supplementary Table S1), and the AR target *CALM1* (Supplementary Table S2D) with highest expression in endothelial and fast fibers (Supplementary Table S1).

3.5 Genes that are previously reported as DMD biomarkers and genetic modifiers are enriched among genes differentially expressed between healthy VL and TA

To explore potential relationships between our differential gene expression and reported mechanisms of DMD pathology that can be mapped to individual genes, we analyzed the differential expression of previously reported human serum/blood DMD biomarkers (Hathout et al., 2014; Parolo et al., 2018; Spitali et al., 2018; Al-Khalili Szigyarto, 2020; Grounds et al., 2020; Alonso-Jiménez et al., 2021; Wagner et al., 2021; Lee-Gannon et al., 2022; Wu et al., 2022), and of genes modifying the phenotype of DMD in humans (Pegoraro et al., 2011; Flanigan et al., 2013; Bello et al., 2016;

Hogarth et al., 2017; Li et al., 2018; Weiss et al., 2018; Spitali et al., 2020; Flanigan et al., 2023) or the *mdx* mouse (Deconinck et al., 1997; Wagner et al., 2002; Han et al., 2011; Morales et al., 2013; de Zélicourt et al., 2022), also known as genetic modifiers, across the VL and TA and identified the predominant cell type in which they were expressed.

Among 88 DMD biomarkers expressed in healthy muscle, 41 (46.59%) were differentially expressed between VL and TA (Supplementary Figure S3A), a significant enrichment of this set of genes among differentially expressed genes (only 13 expected, p-value < 1E0-5, χ^2 test). Of this set of differentially expressed biomarkers, 20 of 41 (48.78%) have highest expression in the myofibers, with 10 of these being more expressed in the slow and 10 in the fast fibers. Among these myofiber-derived biomarkers, 13 have the highest expression in VL, and 9 of these (69.2%) have highest expression in the fast fibers (*ALDOA*, *CAMK2B*, *ENO3*, *GAPDH*, *LDHA*, *MSTN*, *MYL1*, *PYGM*, *TNNI2*). The remaining 4 biomarkers higher in VL have either highest expression in the slow fibers (*CAMK2A*, *ACTA1*) or are similarly expressed in fast and slow fibers (*CKM*, *TTN*).

We also culled from the literature 18 genes described as genetic modifiers based on either a human genetic variant association with a DMD phenotype (*SPP1*, *ACTN3*, *THBS1*, *LTBP4*, *HLA-A*, *DYNLT5*, *CD40*, *NCALD*, *ETAA1*, *ADAMTS19*, *MAN1A1*, *GALNTL6*, *PARD6G*) or an *mdx* phenotype modified by concomitant deletion of another gene (*MSTN*, *DYSF*, *CCN2*, *UTRN*, *CD38*) (Deconinck et al., 1997; Wagner et al., 2002; Han et al., 2011; Pegoraro et al., 2011; Flanigan et al., 2013; Morales et al., 2013; Bello et al., 2016; Hogarth et al., 2017; Li et al., 2018; Weiss et al., 2018; Spitali et al., 2020; de Zélicourt et al., 2022; Flanigan et al., 2023). All were observable within the RNAseq and snRNAseq datasets, except *SPP1* (encoding osteopontin), which has a median TPM of 0.01 across VL and TA RNAseq and was not detected in snRNAseq of healthy VL and TA in any cell type. Thus, *SPP1* was below the limits of detection, and it was excluded from the differential gene expression analysis. 11 of 17 (64.7%) remaining DMD genetic modifiers were differentially expressed between VL and TA (Supplementary Figure S3B). This is a larger number than expected from a random sampling of all genes (only 3 expected, p-value < 1E0-5, χ^2 test). Of these 11 genetic modifiers with differential expression detected, *DYSF*, *MSTN*, *ACTN3*, *NCALD*, *ADAMTS19* and *CD38* were higher in the VL, and *HLA-A*, *UTRN*, *LTBP4*, *CCN2*/CTGF, and *THBS1* were higher in the TA (Supplementary Figure S3B). This indirectly supports the relevance of genetic variants indeed contributing to differential disease progression across individuals. Most of the genetic modifiers (10 of 17) had expression mainly within a non-myofiber cell type, consistent with the known role of non-myofiber lineage cells in orchestrating muscle remodeling during regeneration and fibrosis (Mann et al., 2011).

LTBP4 is most expressed in fibroblasts (Supplementary Figure S3B), consistent with prior reports on its ameliorative effect through a reduction in TGF-β signaling in fibroblasts with the IAAM haplotype (Flanigan et al., 2013). In addition to fibroblasts, *LTBP4* also shows high expression in satellite cells, suggesting the potential of a modifying effect acting upon muscle stem cells that has not been studied previously. *DYNLT5* (also known as *TCTEX1D1*) has a median TPM of 0.6 in the bulk RNAseq dataset, and it was not detected in fast or slow fibers by snRNAseq but was rather expressed in endothelial cells. This suggests its modifying mechanism is not due to direct expression within myofibers, or that it could be upregulated in a cell type other than endothelial cells in DMD to exert its modifying mechanism. *NCALD*, encoding the calcium-sensing neurocalcin delta, is most expressed in smooth muscle (7.11 TP10K), and has lower expression in fast (1.90 TP10K) and slow (0.42 TP10K) fibers. Its proposed mechanism is via regulation of a surrogate cGMP pathway that compensates for the defective nitric oxide-induced cGMP production in DMD, with lower expression of *NCALD* being protective (Flanigan et al., 2023). Consistent with this proposed mechanism, *NCALD* is not only higher in the VL in bulk RNAseq, but also in the VL fast (2.51X, p = 2.11E-76) and VL slow (1.11X, p = 3.01E-02) fibers compared to the TA fast and slow fibers, respectively. *HLA-A* is expressed higher in the VL, and class I MHC expression on myofibers may influence immune mediated mechanisms of myofiber damage in dystrophic muscle.

Only four reported genetic modifiers, *DYSF*, *ADAMTS19*, *MSTN*, and *ACTN3* are observed to have the highest expression in myofibers (Supplementary Figure S3C). The expression pattern of *DYSF*, *MSTN* and *ACTN3* is consistent with their described modifying mechanisms (Wagner et al., 2002; Vincent et al., 2007; Han et al., 2011). *DYSF* has similar expression in fast and slow fibers, with a slight 1.2X higher expression in fast fibers. Although the proposed modifying mechanism of *ADAMTS19* is through extracellular matrix (ECM) remodeling and TGF-β signaling (Flanigan et al., 2023), its highest expression in healthy muscle was not in fibroblasts (0.12 TP10K) or vasculature cells that typically produce ECM, but rather in the fast (1.58 TP10K) and slow (1.07 TP10K) fibers. A 13.2X higher expression of *ADAMTS19* in the fast fibers compared to fibroblasts suggests a modifying role in the myofibers that needs further exploration. At the single cell level, *ADAMTS19* is 1.55X higher in the VL slow fibers compared to the TA slow fibers (p = 1.18E-14), which further contributes to its higher expression in the VL. *MSTN* is expressed in both fast and slow fibers, with highest expression in fast fibers (4.4X compared to slow fibers), a pattern of expression that contributes to it being higher in VL by bulk RNAseq, as VL has a higher proportion of fast fibers. *ACTN3* is highly specific to fast fibers (13.3X higher in fast fibers), although not absent in slow fibers. In addition, at the single cell level, *ACTN3* expression is 1.43X higher in the VL fast fibers compared to the TA fast fibers (p = 2.17E-11), indicating that the higher expression of *ACTN3* in VL is influenced by both a higher proportion of fast fibers and by a VL-specific upregulation within the fast fibers.

The well-studied null allele of *ACTN3* (rs1815739, NM_001104.4:c.1729C>T, NP_001095.2:p.Arg577Ter/p.R577X) is a common allele found in the population with a frequency of the X allele of 0.36 (dbSNP). Actinin-3 loss was associated with a reduced DMD severity as measured by a longer 10-min walk test (Hogarth et al., 2017), and this was attributed to a switch to a more protective oxidative metabolism without a shift in fiber type distribution (MacArthur et al., 2008). To further investigate the effects of *ACTN3* expression across muscles, we genotyped rs1815739 in the 15 individuals. We identified 3 null homozygotes (XX), and 3 reference homozygotes (RR), and 9 heterozygotes (RX) among the 15 individuals. The expression of *ACTN3* was significantly differentially expressed dependent on

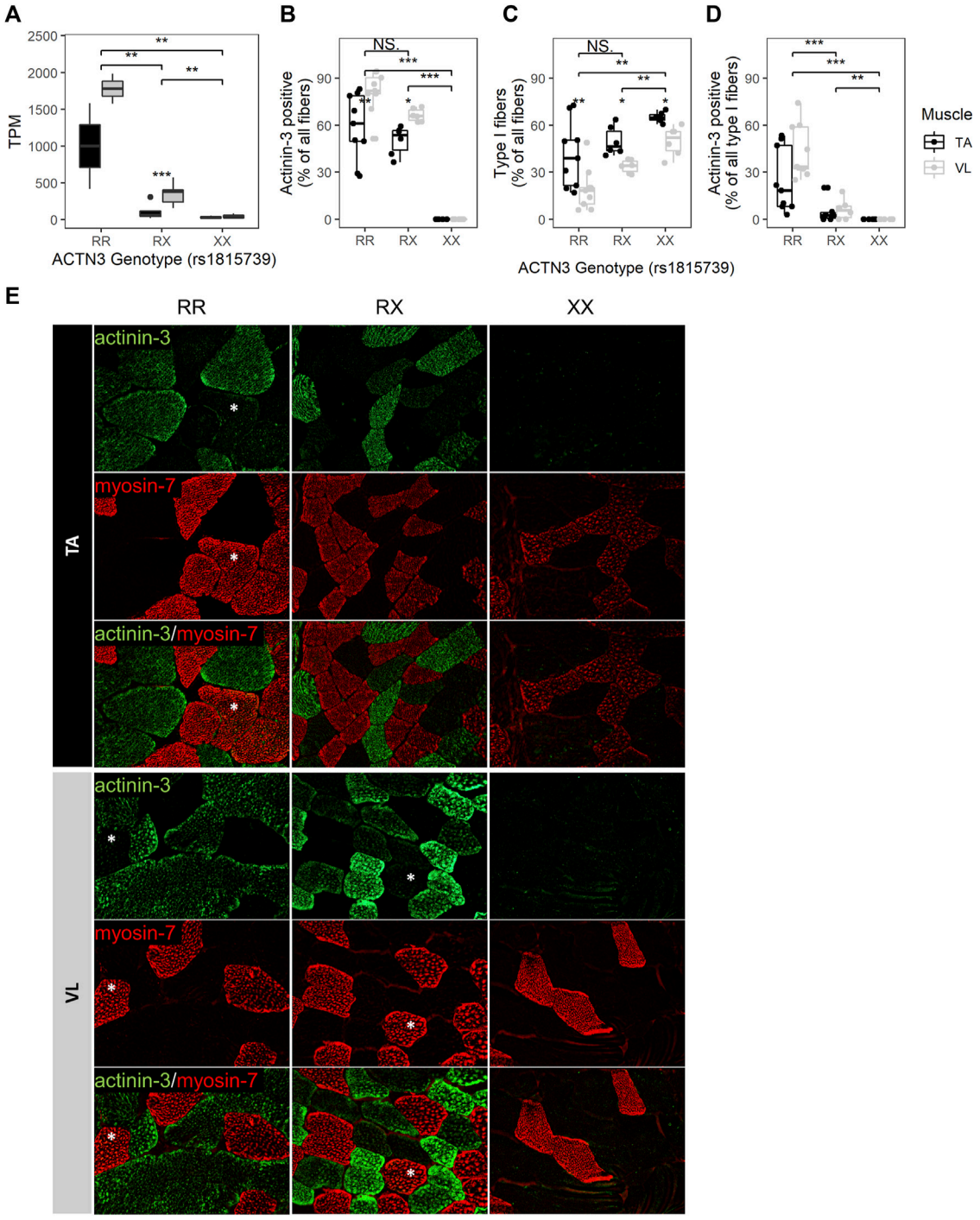


FIGURE 4
ACTN3 genotype correlates with the proportion of slow fibers and with the expression of actinin-3 in slow fibers. **(A)** *ACTN3* TPM by genotype at the rs1815739 polymorphism locus (NM_001104.4:c.1729C>T, NP_001095.2:p.Arg577Ter/p.R577X) for all 27 samples in the bulk RNAseq dataset (RR n = 2, RX = n = 9, XX n = 3). Percentage of all counted fibers that are actinin-3 positive **(B)** and type I (slow) **(C)**. Percentage of all type I fibers that are actinin-3 positive **(D)**. Wilcoxon test, *: p ≤ 0.05, **: p ≤ 0.01, ***: p ≤ 0.001. **(E)** Representative images of actinin-3 and myosin-7 staining. Magnification = 20X. Asterisks indicate actinin-3 positive type I fibers. The bars in the box plots indicate 1.5* IQR, which is the interquartile range, or the distance between the first and third quartiles.

genotype (Kruskal–Wallis p = 7.73E-04), with the RR group showing highest expression, indicating nonsense-mediated decay of the X allele (Figure 4A). XX homozygotes have no *ACTN3* mRNA expression for both VL and TA muscles (Figure 4A). The expression of *ACTN3* RR and RX mRNA was consistently higher in the VL compared to TA, although only statistically significant in the RX genotype (Wilcoxon p = 9.9E-04). For both the VL and TA, the mean level of *ACTN3* mRNA in RX heterozygotes was

substantially lower than the expected 50% of the RR genotype mRNA level, suggesting that the X allele reduces the expression of the R allele through unknown mechanisms.

To validate the RNA findings and assess whether the *ACTN3* genotype groups have differences in fiber type composition, we performed immunofluorescent staining for 3 RR, 2 RX and 2 XX individuals' VL and TA. For each sample and muscle, 3 10 μM tissue sections were stained (total of 42 sections), and fibers were counted across the entire sections. An average of 204 fibers were counted per sample (range 25–758, total 8,577) (Supplementary Figure S4). Observed differences in mRNA expression were consistent with antibody staining for actinin-3. As expected, *ACTN3* XX homozygotes showed no detectable protein (Figures 4B,E). The percentage of actinin-3 positive fibers was consistently higher in the VL for both RR and RX genotypes (Figure 4B). The percentage of type I slow fibers (positive for myosin-7) was consistently higher in the TA across genotype groups (Figure 4C), as expected (Edgerton et al., 1975; Jakobsson et al., 1991). There was a higher percentage of type I slow fibers in the RX and XX groups compared to the RR group across both muscles, although only the XX group reached statistical significance (Figure 4C). There was also a higher percentage of type I fibers in the XX compared to RX group. These findings are supported by observed similar relative expression of the myosin heavy chain genes at the RNA level (Supplementary Figure S5). Interestingly, we also identified slow fibers with low expression of actinin-3 (Figure 4D) in the RR and RX genotypes but not in the XX genotypes, reflective of a low level of expression of actinin-3 in some slow myofibers. This low level of expression is only apparent because of the true null staining revealed in the XX genotype individuals.

3.6 Identification of druggable targets within differentially expressed genes

We place the list of differentially expressed genes into context as potential for disease modification because their RNA or protein products are targeted or 'druggable' with existing drugs documented in the DrugBank database (Wishart et al., 2018). 535 of 3,410 (15.7%) differentially expressed genes are reported targets of 1,812 known drugs (Supplementary Table S1). The protein product of 197 genes higher in the VL are targeted by 984 drugs, and thus may constitute a set of known drugs that may be explored to induce a shift of a VL-like susceptible state towards a TA-like protected state in DMD. Druggable genes expressed higher in VL are enriched in 158 pathways (Supplementary Table S2E). Among the top 5 most significant pathways is calcium signaling pathway, with 22 genes higher in VL that include calmodulin (*CALM1*, *CALM2*), calmodulin-dependent kinases (*CAMK2A*, *CAMK2B*, *CAMK2G*), calsequestrin (*CASQ1*), ryanodine receptor (*RYR1*), and the dihydropyridine receptor alpha 1S subunit (*CACNA1S*). These 8 genes alone are reported to be targeted by 71 drugs and may suggest ways to therapeutically regulate intracellular and sarcoplasmic reticulum (SR) calcium concentration in myofibers. *CD38* (1.7X higher in VL) is highest expressed in the pericytes (1.83 TP10K), but also is expressed in fast (1.28 TP10K) and slow (0.48 TP10K) fibers. Its higher expression in VL is also observed at the single cell level, with 1.24X higher expression in VL fast fibers compared to TA fast fibers (p = 4.71E-04). *CD38* encodes a NAD+ glycohydrolase that produces regulators of Ca2+ signaling, and deletion of *CD38* or treatment with CD38 inhibitors restored the *mdx* heart, diaphragm and limb function, reduced fibrosis and inflammation, and reduced the cycles of degeneration and regeneration (de Zélicourt et al., 2022). DMD myotubes treated with a monoclonal antibody against CD38 (Isatuximab) reduced the frequency of spontaneous Ca2+ waves (de Zélicourt et al., 2022).

3.7 Identification of isoforms with differential abundance between VL and TA

Because extensive alternative splicing is observed in developing and mature muscle (Brinegar et al., 2017; Nakka et al., 2018), including in genes encoding sarcomere structural (Donner et al., 2004; Bowman et al., 2007; Lam et al., 2018; Savarese et al., 2018), and excitation-contraction coupling proteins (Nakka et al., 2018), we hypothesized that the VL and TA transcriptomes are also differentially influenced by alternative splicing leading to significant shifts in the usage of specific isoforms (isoform switch). To identify isoform switching events between VL and TA, we utilized IsoformSwitchAnalyzeR (Vitting-Seerup and Sandelin, 2017), which uses the abundance (TPM) and count data obtained from Kallisto transcript alignment. Prefiltering of the annotated transcripts resulted in 130,664 transcripts to be considered for isoform switch analysis. Further filtering of transcripts for switching genes with at least two significantly switching isoforms and with at least two isoforms preferentially used in opposed directions (higher in one muscle, lower in the other) resulted in 47 transcripts. Among these, 12 transcripts have a significant isoform switch (isoform switch q-value <0.05) between VL and TA, and these are located within 6 genes (gene switch q-value <0.05) (Supplementary Table S2F). Two of the 6 genes with isoform switching, *NPR3* and *TNNT1*, were differentially expressed between VL and TA, whereas the remaining 4 (*NEB*, *ABCC6P2*, *ENSG00000288071* and *MAD1L1*) did not have differential expression at the gene expression level (Figure 5A; Supplementary Table S1). *NEB* is highly and similarly expressed in slow (217.6 TP10K) and fast (185.4 TP10K) myofibers (Supplementary Table S1). *ENSG00000288071* and *TNNT1* are more highly expressed in the slow fibers (6.8X and 16.4X higher expression in the slow compared to fast fibers), and *ABCC6P2* in the fast fibers (0.009 TP10K in fast and not detected in slow fibers) (Supplementary Table S1). *NPR3* and *MAD1L1* are expressed highest in the smooth muscle cells (Supplementary Table S1). Four of 6 genes with isoform switch are protein coding (*NEB*, *NPR3*, *TNNT1* and *MAD1L1*), whereas *ENSG00000288071* is a long non-coding RNA, and *ABCC6P2* is a transcribed unprocessed pseudogene (Supplementary Table S1).

To assess potential functional consequences of the isoform switches, we examined the biotype of each isoform in the switching (Figure 5B). The *NEB*, *TNNT1* and *NPR3* isoform switches are among protein coding transcripts, and that of *ENSG00000288071* is among long noncoding RNAs (lncRNA). The remaining 2 genes (*ABCC6P2* and *MAD1L1*) are switching between isoforms of different biotypes (Figure 5B). ABCC6P2-202,

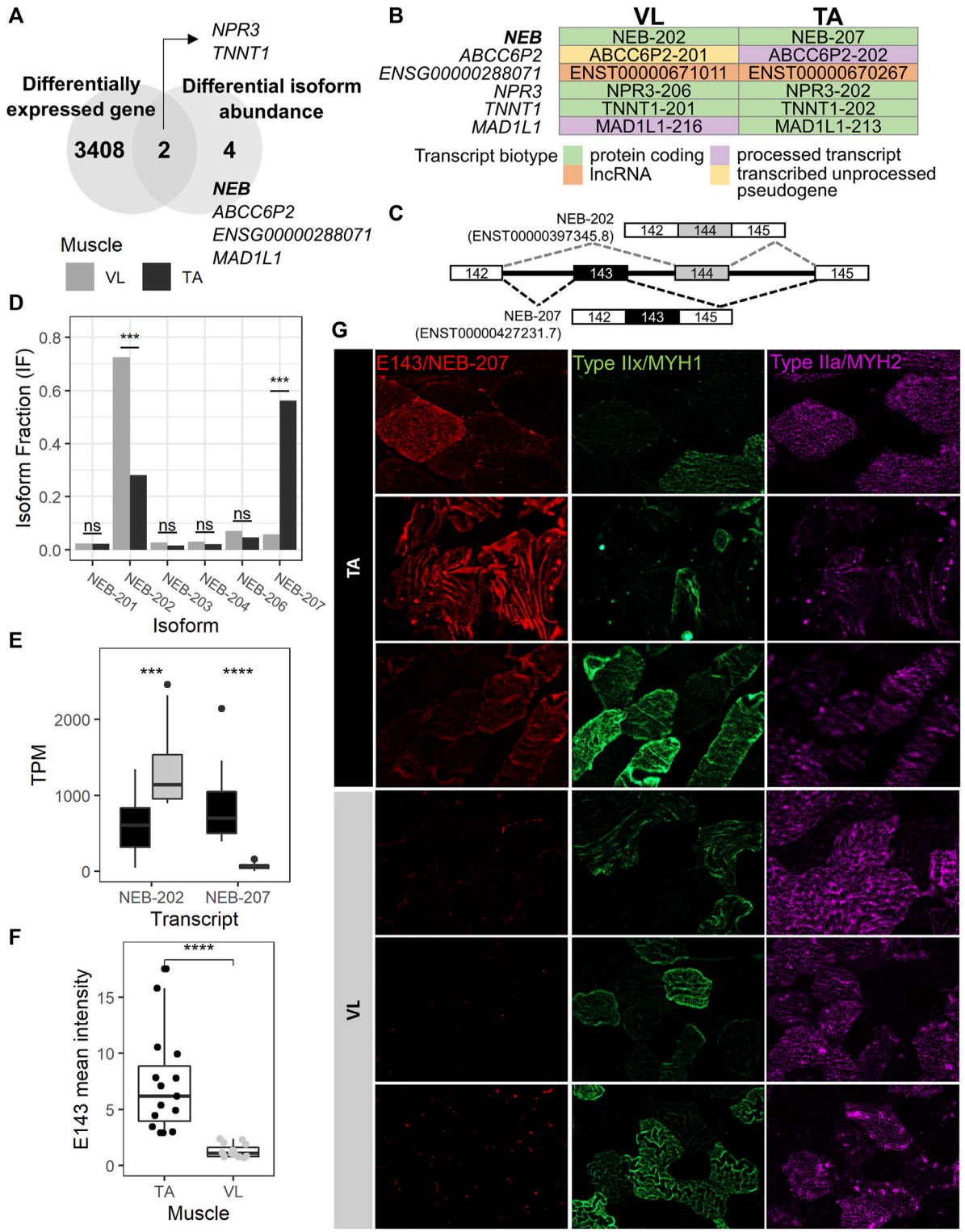


FIGURE 5
NEB-207 is upregulated in the TA across all fiber types. **(A)** Venn diagram showing the overlap of differentially expressed genes and genes with differential isoform abundance (isoform switch) between VL and TA. **(B)** For each muscle, the Ensembl transcript biotype of each preferentially used isoform is shown for each isoform switch event. **(C)** Diagram of the alternative usage of exons 143 and 144 in *NEB*. The resulting isoform name is indicated for each exon usage, and corresponding transcript IDs are in parenthesis. **(D)** Isoform fraction (usage) of expressed *NEB* isoforms obtained from IsoformSwitchAnalyzeR. **(E)** Expression of the NEB-202 and NEB-207 isoforms obtained from the Kallisto alignment. TPM = Transcripts per million. **(F)** Quantification of the overall mean immunofluorescence signal intensity of nebulin exon 143 (NEB E143) in 5 20X images for each VL and TA among 3 healthy individuals. Wilcoxon test p = 1.29E-08; ****: p≤0.0001). **(G)** Representative images of immunofluorescence staining of paired VL and TA sections.

preferentially used in TA, is a processed transcript, whereas ABCC6P2-201, preferentially used in VL, is a transcribed unprocessed pseudogene. MAD1L1-213, with preferential usage in TA, is protein coding, whereas MAD1L1-216, preferentially used in VL, is a processed transcript, which means that it does not have an open reading frame (a start codon followed by an in-frame stop codon (Kute et al., 2022)).

The *MAD1L1* isoform switch comprises MAD1L1-213 and MAD1L1-216 with the former used more in TA (absolute difference isoform fraction = 0.073) and the latter in VL (absolute difference isoform fraction = 0.043) (Supplementary Table S2F). *MAD1L1* has a relatively low expression in muscle. The top 3 isoforms expressed in both VL and TA have an average TPM ranging from 0.80 to 1.65. *MAD1L1* encodes for the mitotic arrest deficient-like protein 1 (also known as MAD1). In *mdx*, *Mad1l1* is most expressed in late activated satellite cells, myoblasts and myocytes (Scripture-Adams et al., 2022) (data not shown), suggesting a potential role in muscle regeneration in wild-type muscle and in DMD, although which isoform is most important is unknown. The *MAD1L1* isoform switch involves a switch from a protein coding isoform in TA to a processed transcript that has no open reading frame in VL, suggesting a potential mechanism of reducing its protein expression in the VL via alternative splicing, an event that cannot be detected by gene expression analysis.

The gene with the most striking isoform switch is *NEB*. This switch comprises the mutually exclusive exon splicing event that occurs between exons 143 and 144 of *NEB*, which has been previously described (Donner et al., 2004; Lam et al., 2018). Exons 143 (E143, included in NEB-207) and 144 (E144, included in NEB-202) are mutually exclusive exons (Figure 5C), such that in the same transcript, only one of either exon is included. NEB-207 has a higher usage in TA, a 0.505 isoform fraction difference compared to VL (Supplementary Table S2F; Figure 5D). NEB-202 has a higher usage in VL, with a 0.443 isoform fraction difference compared to TA (Supplementary Table S2; Figure 5D). This difference in isoform usage is readily observed at the isoform expression level (Figure 5E). NEB-207 has higher expression in TA, with an average TPM of 850, compared to an average TPM of 66 in VL. NEB-202 is more broadly expressed across both muscles but has a preferential expression in the VL with an average TPM of 1,357, compared to an average TPM of 641 in TA. We confirmed this differential alternative splicing event in the RNA by semi-quantitative reverse transcription polymerase chain reaction (RT-PCR) using exon junction-specific primers (data not shown).

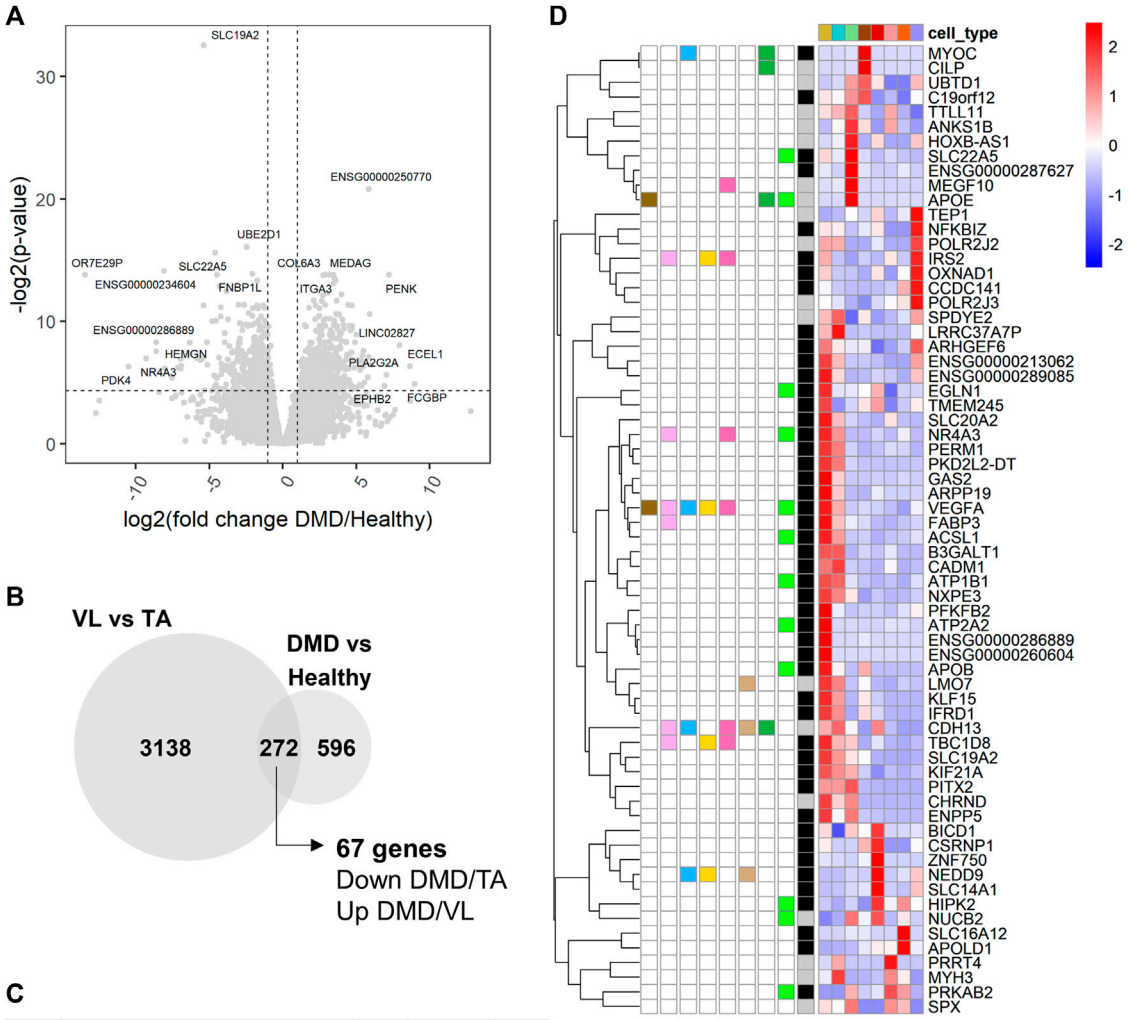
Previous reports sought to determine whether the expression of nebulin exon 143 presents a fiber type-specific pattern in adult human quadriceps (Lam et al., 2018). In this previous study, E143 was found expressed more often in fast fibers compared to slow fibers (Lam et al., 2018), although a distinction between type IIa and type IIx fast fibers was not explored. Consequently, it was concluded that fast fibers usually express E143, and that slow fibers may express either E143 or E144. However, because we observe that the VL mainly includes E144 and not E143, and because VL has in average 1.82 more fast fibers than TA (Figure 2C), we reasoned that the differential pattern of expression of E143 between VL and TA is not solely dependent on fast fiber type. Thus, we assessed the protein expression of E143 across VL and TA in relation to type IIa and IIx fast myosin. We examined overall E143 protein intensity among VL and TA in 3 healthy individuals. We found low to no E143 myofiber intracellular protein expression in the VL among either fiber type (Figure 5G). Consistent with this observation, overall E143 protein intensity was statistically higher in the TA compared to the VL (Wilcoxon test p = 1.29E-08, 95% CI = 2.70-7.04, average fold difference = 5.77) (Figure 5F). These findings suggest that although nebulin including exon 143 is more often expressed in the fast fibers (Lam et al., 2018), and E143 is more consistently observed in the fast type IIx than in the slow type I (data not shown), a fast fiber type is not the sole determinant of its expression in human skeletal muscle. That is, that the association between fast myosin and exon 143 of nebulin, as described previously (Lam et al., 2018) is also muscle-type specific, and might be regulated by specific differentially expressed splicing factors, or their combinations.

To identify potential splicing factors underlying the alternative splicing observed between VL and TA, we looked for splicing factors that are differentially expressed between the muscles. Out of 66 expressed splicing factors obtained from the SpliceAid-F database (Giulietti et al., 2013), 18 (27.3%) were differentially expressed between VL and TA (Supplementary Table S1). Only one splicing factor, *NOVA2*, is higher in TA, and the remaining 17 are higher in VL, with *ESRP2* and *CELF2* being the most differentially expressed.

3.8 Comparison of VL *versus* TA differentially expressed genes with DMD *versus* healthy muscle differentially expressed genes

To assess whether the differentially expressed genes between healthy muscles differentially susceptible to lack of dystrophin are also changed in expression in the context of DMD, we generated bulk RNAseq data of the TA from 8 young ambulatory DMD patients (mean age 4.5 years). An average of 60 million paired end sequencing reads were generated per sample (range 43–72 million reads). To our knowledge, this is the second and largest reported bulk RNAseq dataset of DMD muscle (an existing dataset can be found in SRA ID PRJNA734152), and the first of the TA muscle.

Using DESeq2 (Love et al., 2014), we performed differential gene expression analysis between DMD (n = 8) and healthy TA (n = 13). 868 of 17,183 analyzed genes were differentially expressed between DMD and healthy (Supplementary Table S3; Figure 6A). 272 of these genes were also differentially expressed between VL and TA (Figure 6B). Among these overlapping genes, 67 were downregulated in DMD and higher in the less susceptible TA or upregulated in DMD and higher in the more affected VL (Figure 6B). Next, we assessed functional enrichment in the genes dysregulated in DMD using EnrichR. 49 biological processes and 11 cellular component categories were enriched among genes dysregulated in DMD (adjusted p-value <0.05) (Supplementary Table S2G). Among these, 31 categories were also enriched among genes differentially expressed between VL and TA. "Collagen-Containing Extracellular Matrix" was the most significant shared term and also the most significant term among all enriched in DMD *versus*



C

GO	Term	Size	Adjusted p-value	Odds Ratio
CC	Collagen-Containing Extracellular Matrix	47	2.68E-11	4.1
	Focal Adhesion	37	2.99E-03	2.16
BP	Positive Regulation Of Cell Population Proliferation	40	1.67E-03	2.43
	Positive Regulation Of Cellular Process	42	1.27E-02	2.05
	Regulation Of Cell Migration	36	1.27E-02	2.18
	Regulation Of Cell Population Proliferation	51	1.36E-02	1.88
	Negative Regulation Of Apoptotic Process	34	3.57E-02	1.95

FIGURE 6
Extracellular matrix and regulation of apoptosis are also dysregulated in DMD. **(A)** Volcano plot for all 17,183 genes tested for differential expression between DMD and healthy TA. Dashed lines depict a fold change of 2 and a *p*-value of 0.05. For each up and downregulated genes, the top 5 differentially expressed genes by fold change and the top 5 genes by *p*-value are labeled. **(B)** Overlap of the differentially expressed genes between VL and TA and between DMD and Healthy. The 67 genes are those downregulated in DMD and higher in TA or upregulated in DMD and higher in VL. **(C)** 7 of 9 functional categories with over 30 genes that are shared between the two analyses (sorted by ascending *p*-value). "Nervous System Development" (which had only *VEGFA* in the 67-gene list, and "Cell-Substrate Junction" (which has the same 37 genes as "Focal Adhesion", and a larger *p*-value) were excluded. **(D)** Single cell expression of the 67 overlapping genes. The overlapping GO terms from (C) in which each gene member is categorized among these 7 categories are indicated. "Druggable" indicates which genes have existing drugs documented on DrugBank. "Direction" indicates the direction of expression in both the VL *versus* TA and DMD *versus* Healthy differential gene expression analyses.

Healthy (Supplementary Table S2G). The categories with over 30 gene members include "Collagen-Containing Extracellular Matrix" and "Negative Regulation of Apoptotic Process", supporting the involvement of these gene sets in both the differential susceptibility between VL and TA, and the dystrophic pathology (Figure 6C).

To identify candidate susceptibility factors that are also dysregulated in DMD, we further explored the 67 overlapping genes. All except one gene (*SBK3*) were detected in the healthy muscle snRNAseq (Figure 6D). 13 of these 67 genes are druggable (Figure 6D). Among the overlapping protein coding genes, the 3 most dysregulated genes in DMD (by fold change) are *NR4A3*, *APOB* and *MYOC*, and the 3 most differentially expressed in VL compared to TA are *APOB*, *SBK3* and *NR4A3*, highlighting the relevance of these genes in DMD and consequently, the differential susceptibility of VL and TA (Supplementary Table S3).

Among genes in "Collagen-Containing Extracellular Matrix" are *MYOC* and *CILP* (Figure 6D). *MYOC*, encoding myocilin, is most expressed in fibroblasts (Figure 6D; Supplementary Table S1) and is downregulated 47.5X in DMD (Supplementary Table S3), despite the expansion of fibroblasts within dystrophic muscle (Scripture-Adams et al., 2022), suggesting a downregulation in dystrophic fibroblasts. Myocilin has been widely studied in glaucoma as a secreted protein in the trabecular meshwork (Resch and Fautsch, 2009). Myocilin was also found to be induced during C2C12 myoblast differentiation via regulation of the TGF-β pathway (Zhang et al., 2021), and to interact with the dystrophin-glycoprotein complex via syntrophin (Joe et al., 2012). In the Human Protein Atlas (Uhlén et al., 2015), *MYOC* is highest expressed in fibroblasts in skeletal muscle, and not in myocytes, consistent with our observations. The overexpression of *MYOC* increases muscle mass (Joe et al., 2012), and downregulation of myocilin is observed in cancer cachexia, with its loss inducing muscle fiber atrophy and an increase in fibrotic and fatty tissue (Judge et al., 2020). Because its expression is highest in fibroblasts, and it is found within "Collagen-Containing Extracellular Matrix", we hypothesize that its main role in human skeletal muscle is in the fibroblasts, and not the myolineage, and that its downregulation promotes fibrosis. These data, along with the 1.77X higher expression of *MYOC* in the TA (Supplementary Table S1), support myocilin as a protective factor for the TA.

Conversely, *CILP*, encoding the cartilage intermediate layer protein 1 (CILP-1), is upregulated 4.3X in DMD (Supplementary Table S3), and is 1.27X higher in the more susceptible VL. *CILP* also has highest expression in the fibroblasts in our dataset (Figure 6D; Supplementary Table S1) and in the Human Protein Atlas (Uhlén et al., 2015), but its role is not well understood. Upregulation of CILP-1 occurs upon cardiac injury in fibrotic regions, and there is a decrease in serum of patients with heart failure (Park et al., 2020). Its anti-fibrotic effect in pressure-overload cardiac remodeling (Zhang et al., 2018) suggests that CILP-1 is regulated in relation to processes that involve cardiac fibrosis. Because of its upregulation in DMD and higher expression in VL, and its restricted expression in the fibroblasts, we hypothesize that CILP-1 is pro-fibrotic in skeletal muscle, and a susceptibility factor for the VL.

Among other overlapping genes is *SLC19A2*, which is the most statistically significantly dysregulated gene in DMD compared to healthy TA (p = 1.59E-10), and which is downregulated 41.77X in DMD (Supplementary Table S3). *SLC19A2* encodes the thiamine (vitamin B1) transporter 1 (THT1), which has highest expression in skeletal muscle (GTEx), specifically in slow fibers (Figure 6A; Supplementary Table S1). Thiamine supplementation has been shown to improve muscle strength in myotonic dystrophy type 1 (Costantini et al., 2016). In *mdx*, supplementation with the thiamine precursor benfotiamine ameliorated the dystrophic pathology and increased grip strength (Woodman et al., 2018), supporting a protective role for the TA compared to VL in DMD, and potentially also in the slow fibers compared to fast fibers. Lastly, *KIF21A*, highest expressed in slow fibers, is downregulated 2.91X in DMD (Supplementary Table S3), and is 1.20X higher in the TA (Supplementary Table S1). Heterozygous mutations in *KIF21A* cause autosomal dominant congenital fibrosis of extraocular muscles (EOM) (Yamada et al., 2003). The downregulation of *KIF21A* in DMD skeletal muscle, reduced function leading to pathologic fibrosis in EOM, and its higher expression in the TA suggest a protective role of higher *KIF21A* expression within myofibers that leads to some protection from damage in DMD. *KIF21A* encodes a kinesin, involved in cargo transport between the Golgi apparatus and the endoplasmic reticulum (Hirokawa and Noda, 2008), but its role in skeletal myofibers is not established.

Among the 67 overlapping genes, *APOE* is found in "Negative Regulation of Apoptotic Process" and is upregulated in DMD and higher in the VL. *APOE* is a highly specific satellite cell marker in healthy muscle (Figure 2B), suggesting that a potential differential regulation of apoptotic signaling (Figure 3C) may alter the regenerative capabilities of VL and TA.

Cell type specificity of the 67 genes differentially expressed between VL and TA and also dysregulated in DMD (Figure 6D) was examined within previously published single cell and nuclei RNAseq datasets from the mouse TA (scMuscle) (McKellar et al., 2021) and soleus (myoatlas) (Petrany et al., 2020). Using these datasets, the cell type specificity of 17 protein coding genes was confirmed. These include *MYOC* and *CILP*, highest expressed in the fibroblasts in both scMuscle (Supplementary Figure S6A) and myoatlas (Supplementary Figure S6B) in mouse.

We note *PDK4* as the most dysregulated gene within "Negative Regulation of Apoptotic Process", and the most dysregulated protein coding gene among all 868 genes differentially expressed between DMD and healthy TA in our transcriptome-wide analysis. *PDK4*, encoding pyruvate dehydrogenase kinase 4, is downregulated 1,453X in DMD and only superseded by the unprocessed pseudogene *OR7E29P*, which is downregulated 11,396X (Figure 6A; Supplementary Table S3). Downregulation of *PDK4* in DMD has been previously reported in the slow (type I) fibers (1.74X, p = 2.78E-04), and upregulation in the fast type IIa and IIx (Scripture-Adams et al., 2022). *PDK4* is also downregulated in *mdx* quadriceps and TA compared to age-matched controls (Matsakas et al., 2013). *PDK4* is downregulated in DMD, but higher in the more susceptible VL (1.98X, Supplementary Table S1). Interestingly, *PDK4* is a druggable gene, with tretinoin (a vitamin A derivative) being a known upregulator (Supplementary Table S1).

Lastly, among the 17 known DMD genetic modifiers included in our DMD *versus* Healthy comparison, none were significantly dysregulated in DMD compared to Healthy TA (Supplementary Table S3), but some trended toward a significant upregulation after multiple testing correction (p < 0.30), including (by ascending p-value): *LTBP4* (4.47X, p = 5.37E-02), *MAN1A1* (4.01X, p = 6.21E-02), *THBS1* (24.89X, p = 9.76E-02), *PARD6G* (3.50X, 1.54E-01), *SPP1* (139X, p = 2.80E-01), and *NCALD* (2.39X, p =

2.97E-02). Interestingly, all 6 genes have variants that have been identified to modulate the disease progression in DMD patients (*versus* in *mdx* double knockouts), supporting their relevance in human pathology.

4 Discussion

Our goal in this study was to analyze two different healthy limb muscles with more similar functional roles (VL and TA), which have a consistently observed difference in disease progression, that is more modest in degree than the greater protection from damage of EOM compared to limb muscles in the absence of dystrophin. Because of published longitudinal imaging data (Rooney et al., 2020), we could select comparable muscles amenable to biopsy in healthy adults (Barthelemy et al., 2020). The protection of TA relative to VL is less striking than that of EOM compared to limb but is still substantial with an estimated shift in equivalent damage of 8.5 years in humans (Rooney et al., 2020). VL and TA demonstrate a substantial difference in their susceptibility to lack of dystrophin, and our transcriptomic study of paired samples from the same healthy individuals identifies a large portion of the transcriptome as altered, with 3,410 differentially expressed genes. There is inherent biological variability within each large muscle, and there is some potential variability added to transcriptomic comparisons due to the small sample analyzed, which may not be representative of the whole muscle. The relatively small sampling by biopsy can introduce variability from sampling different parts of each muscle. This could result in a reduction in the number of differentially expressed genes, but should not lead to false expression differences between muscle groups. We try to limit variation by sampling the same relative location of VL and TA, which indeed resulted in highly significant differential gene expression detection. Because the transcriptomic differences between muscles in this study greatly exceeded those driven by genetic variation, we note that future studies may not require a paired design approach that was used here to maximize discovery and control for interindividual differences.

Our study particularly highlights calcium homeostasis, ECM, and regulation of apoptosis, and provides a dataset for exploration to investigate potential protective mechanisms of myofibers to loss of dystrophin in skeletal muscle. By studying muscles that have a substantial difference in their rate of disease progression in DMD, we sought to identify mechanisms of myofiber protection, complement genetic modifier studies and reveal novel therapeutic targets. There is some overlap between prior gene expression work comparing EOM to limb muscles and this study comparing TA to VL, including enrichment of genes with functions related to sarcomere structure, calcium homeostasis, muscle development, metabolic and immune processes, vasculature development, regulation of cell death and extracellular matrix, and thus supports that these pathways are relevant to how muscles are differentially susceptible to damage with lack of dystrophin. Comparison with genes dysregulated in DMD skeletal muscle compared to healthy further highlights the potential role of ECM and negative regulation of apoptosis in the differential susceptibility of VL and TA in DMD. We note that the mean age of our DMD cohort (4.5 years) is younger than the healthy cohort (21.2 years), and we attempted to reduce the effect of age on the identified gene expression differences. However, age could also be contributing to gene expression differences reported here.

Recently, a relatively higher regenerative capacity of EOM muscle stem cells was identified and attributed to upregulation of thyroid-stimulating hormone receptor (TSHR) signaling through upregulation of adenylate cyclase activity in EOM relative to limb muscle (Taglietti et al., 2023). Although *TSHR* is not differentially expressed between VL and TA in our study, "Adenylate Cyclase-Activating G Protein-Coupled Receptor Signaling Pathway" trended toward significance among the biological process GO terms (p = 9.92E-02), with 16 of the 17 genes higher in the TA (data not shown), suggesting a protective role in the TA and consistent with the proposed therapeutic relevance of upregulation of adenylate cyclase in DMD, where adenylate cyclase activation stimulates TSHR signaling, reduces muscle stem cell senescence and improves their proliferation (Taglietti et al., 2023).

Of note, only 24.9% of the differentially expressed genes were highest expressed in the myofibers, indicating that many of non-myofiber cells are likely to play an important role in protecting myofibers from death. A caution of our work is that the healthy muscles are sampled without active degeneration/regeneration or induced muscle damage, which is a chronic state in DMD, and thus our data does not necessarily reveal mechanisms that may be only induced with muscle injury.

ECM deposition is an important component of the muscle structure and function (Loreti and Sacco, 2022), and there is an enrichment of differentially expressed genes that encode "Collagen-Containing Extracellular Matrix". ECM remodeling is necessary to properly activate muscle stem cells during regeneration, and the dysregulation of ECM proteins has been associated with regeneration defects in muscle diseases (Loreti and Sacco, 2022). In addition, the ECM stiffness, which varies depending on ECM composition, can modulate satellite cell activity and myofiber-generated force during contraction, and undergoes changes with age (Sinha et al., 2020). Thus, observed differences in ECM gene expression in VL and TA may contribute to their differential progression in DMD and cause differences in the fibrotic response within each muscle type. We highlight *MYOC* as a potentially protective anti-fibrotic, and *CILP* as a potentially damaging pro-fibrotic gene in DMD.

Myofiber death in DMD has been mainly attributed to necrosis (Bencze, 2023). However, a higher rate of apoptotic nuclei in DMD compared to healthy muscle has been repeatedly observed (Tews and Goebel, 1997; Sandri et al., 1998; Serdaroglu et al., 2002), particularly before necrosis initiates (Tidball et al., 1995). In addition, p53 is one of the most highly induced transcription factors in *mdx* (Dogra et al., 2008), and its inhibition reduced exercise-induced necrosis in the dystrophic mouse (Waters et al., 2010), suggesting an important role in the *mdx* pathology. We identify an enrichment of "Regulation Of Apoptotic Process" genes that are differentially expressed, and a 4.2 fold increase in *ZNF385A* in TA (p-value = 1.76E-102), which is a reported modulator of p53 that reduces pro-apoptotic signaling (Das et al., 2007). This relatively higher expression of *ZNF385A* is also observed in gracilis (Abbassi-Daloii et al., 2023) and EOM (Porter et al., 2001; Terry et al., 2018) which are protected in DMD compared to the VL. Because of the potential impact of ZNF385A to suppress apoptosis, *ZNF385A*

may protect the TA via modulation of p53 signaling towards an anti-apoptotic state. Further studies need to be conducted on 1) what are the direct or indirect targets of *ZNF385A* in human muscle, 2) whether up or downregulating the expression of *ZNF385A* has an effect on the apoptotic rate in myotubes and muscle stem cells exposed to an apoptotic-inducing condition, and 3) whether inducing its expression in dystrophic myotubes protects myofibers and other resident muscle cells from death.

DMD modifier genes were more likely to be differentially expressed between VL and TA, supporting a functional role for several modifiers from this orthogonal transcriptomic study. Identifying novel genetic modifiers of DMD remains a challenge, as studies are limited due to sample size, and thus under-powered to detect genome-wide significance. Thus, augmenting with other data types is relevant to increasing confidence in observed genetic modifiers.

The higher expression of the genetic modifier *LTBP4* in TA and its restriction to fibroblasts is consistent with a role in slowing disease progression, as it binds TGF-β and thus reduces TGF-β signaling (Flanigan et al., 2013), a major driver of fibrosis. *LTBP4* also had high expression in satellite cells, an unexpected finding. Although TGF-β signaling is known to modulate the muscle stem cell function, the specific role of *LTBP4* in satellite cells has not been elucidated. If *LTBP4* participates in regulation of satellite cell function, it may create differences in the muscle-specific regenerative ability that needs further exploration, particularly in the context of the protective IAAM haplotype.

Actinin-3 null allele has been previously reported to be protective in DMD via a shift to a more oxidative metabolism (Hogarth et al., 2017) characteristic of slow fibers, which are more protected from loss in DMD. Various studies have examined whether there is an associated change in fiber type composition in the *ACTN3* null genotype, with some finding no evidence for a fiber type shift (MacArthur et al., 2008; Broos et al., 2016), and others finding significant differences in the fiber type composition across genotype groups (Vincent et al., 2007). The discrepancies could be due to different sampling methods, such as the number of fibers counted. The higher proportion of slow fibers in XX individuals may be protective because slow fibers are protected for longer in DMD (Webster et al., 1988). We detected previously unreported expression of actinin-3 in slow fibers at low levels, particularly in the RR and RX groups. The presence of actinin-3 in slow fibers in the VL may render VL slow fibers more susceptible to damage by increasing glycolytic and reducing oxidative metabolism.

Differential mutual exclusion of exons 143 and 144 of *NEB*, as we observed here, has been observed for another pair of human muscles, and gastrocnemius (GN) preferentially includes exon 144 and TA exon 143 (Donner et al., 2004; Lam et al., 2018), the latter being consistent with this study. Similar to the difference in progression between the VL and TA in DMD where the TA is delayed by about 8.5 years, the TA is delayed by 3.4 years relative to GN (Rooney et al., 2020). Because the more affected VL and GN preferentially include E144 and not E143, we hypothesize that E143 included nebulin could plausibly confer different sarcomere properties that result in protection of the muscle membrane to contraction-induced injury in the absence of dystrophin.

Nebulin has various roles in skeletal muscle. Although the most commonly known role is thin filament length regulation and stabilization, it also has been recently found to have roles in modulating contractile force, calcium handling, and the actin-myosin interaction (Chu et al., 2016). Mutations in *NEB* are the most common cause of autosomal-recessive nemaline myopathy, characterized by Z-disk and thin filament proteins aggregated into nemaline bodies, Z-disk disorganization and consequently, early-onset muscle weakness that mainly affect proximal muscles (Lehtokari et al., 2014). Homozygous intronic mutations in intron 144, which created an alternative donor (5') splice site in exon 144 and a decrease in *NEB* expression, were found causal in a case of a 6-year-old boy with general muscle weakness and nemaline bodies consistent with nemaline myopathy (Laflamme et al., 2021). Exons 143 and 144 encode the super repeat region 21 (S21) of nebulin (Lam et al., 2018) and how they differ functionally has not been extensively studied. The only reported difference is in their charge, hydrophobicity and the predicted presence of a protein kinase C phosphorylation site in the E144 but not in the E143 (Donner et al., 2004). The central super repeat region, which has 22 super repeats in total, has been proposed to interact with KLHL40 (Garg et al., 2014). KLHL40 is located in the sarcomere I and A bands, where it binds to nebulin (Garg et al., 2014). Similar to mutations in the *NEB* exon 143-144 region, KLHL40 deficiency is associated with nemaline myopathy (Garg et al., 2014). These data indicate that the S21 repeat region is critical for proper sarcomere organization, and consequently, muscle function. Nebulin S21 isoforms with different charge and hydrophobicity can potentially modulate the sarcomere organization, structure, and stability and lead to a different susceptibility of the dystrophin-glycoprotein-sarcomere link to damage in the absence of dystrophin.

Previous reports on isoform switching across leg muscles identified 200 switching isoforms among 79 genes (Abbassi-Daloii et al., 2023). However, we did not identify any of these isoforms switch events between VL and TA. These findings could be partially attributed to the different skeletal muscles studied, RNA quality and the sequencing library type. In our study, we utilized ribosomal depletion before cDNA synthesis. However, poly(A) libraries can be 3' end biased (Shi et al., 2021) and this can affect isoform quantification.

This study further provides insights into transcriptomic signatures of differentially affected muscle groups, at both the gene and isoform level, and constitutes the first study, to our knowledge, to augment transcriptomic data from different healthy human skeletal muscles using single nuclei transcriptomics to unravel the complexity of tissue heterogeneity and its contribution to intrinsic transcriptomic signatures. To our knowledge, this study also generated the second and largest reported DMD bulk RNAseq dataset, from young ambulatory patients with the same type of *DMD* mutation (nonsense mutation). An existing dataset of 5 DMD muscle RNAseq (sequenced muscle not specified) can be found in the Sequence Read Archive (SRA) database (PRJNA734152), and RNAseq for four different muscles (1 biceps, 1 quadriceps, 1 gastrocnemius, 1 tibialis anterior) can be found in PRJNA342787. In addition, this is the first throughput dataset of the DMD TA. Although various other datasets of human DMD muscle microarray are found in the Gene Expression Omnibus (GEO) database (GSE3307, GSE109178, GSE6011,

GSE1004, GSE38417, GSE13608), GSE6011 is the microarray dataset at the earliest stage reported, and it corresponds to the quadriceps (at times used to refer to the VL) at less than 2-year of age. Considering that our DMD TA dataset is from muscle at 2–7 years of age, and that the TA is protected for 8.5 years compared to the VL (Rooney et al., 2020), we estimate that the herein generated TA dataset is the DMD whole muscle transcriptome at the earliest stage of the disease reported to date. Furthermore, the healthy snRNAseq and bulk RNAseq datasets provide useful resources for identification of muscle disease genes through transcriptomics, which require healthy reference materials (Lee et al., 2020). In addition, the snRNAseq dataset could be used to identify splicing factors co-expressed in single cells predominantly expressing different isoforms, and various methods have been developed to overcome the challenges of isoform quantification caused by 3' bias, low sequencing depth and dropout (Huang and Sanguinetti, 2017; Song et al., 2017; Hu et al., 2020; Pan et al., 2021). Establishing a single nuclei atlas of healthy human muscles will allow for a better understanding of muscle-specific responses to lack of dystrophin in particular cell types, how genetic modifiers may influence these, whether there is a preferential responsiveness of specific muscle groups to therapeutic approaches and what the cellular underlying mechanisms are, and how to mimic these intrinsic mechanisms to improve the effectiveness of current therapeutics.

Data availability statement

The raw RNAseq and snRNAseq data generated and analyzed in this study can be found in the Sequence Read Archive (SRA) (BioProject ID: PRJNA976807, https://www.ncbi.nlm.nih.gov/sra/PRJNA976807). The Seurat object for the snRNAseq data can be found in https://www.synapse.org/#!Synapse:syn51794252/.

Ethics statement

The studies involving human participants were reviewed and approved by UCLA Institutional Review Board (UCLA IRB, #18-001366, #11-001087). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

SN-R, SN, MM and FB contributed to the conception, execution, and design of the study. SN-R wrote the first draft of the manuscript. SN and JW performed the muscle biopsies. FB and SN-R processed and preserved the muscle biopsies. ED consented the healthy volunteers. SN-R and RW extracted the whole tissue RNA. FB performed the muscle sectioning. DS and KC designed the protocol and extracted and sorted the single nuclei. SN-R performed the RNAseq and snRNAseq analyses. SN-R performed the immunofluorescent staining, microscopy, and image analyses with guidance from FB. FG performed the validation of alternative splicing in *NEB* by RT-PCR. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Center for Duchenne Muscular Dystrophy at UCLA. SN-R was supported by NIH T32HG002536 Genomic Analysis and Interpretation Training Grant and the CDMD Azrieli Graduate Award. KC was supported by NIH T32AR065972 Muscle Cell Biology Pathophysiology and Therapeutics Training Grant.

Acknowledgments

We thank the UCLA Technology Center for Genomics and Bioinformatics (TCGB) core for generating the RNAseq and the snRNAseq data, and Hane Lee, Alden Huang and Lee-kai Wang for designing the RNAseq data processing pipeline. Content within this manuscript has previously appeared online as part of a doctoral dissertation (Nieves Rodríguez, 2023).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1216066/full#supplementary-material. Supplementary table descriptions can be found in Data Sheet 1.

References

Abbassi-Daloii, T., el Abdellaoui, S., Voortman, L. M., Veeger, T. T. J., Cats, D., Mei, H., et al. (2023). A transcriptome atlas of leg muscles from healthy human volunteers reveals molecular and cellular signatures associated with muscle location. *eLife* 12, e80500. doi:10.7554/eLife.80500

Al-Khalili Szigyarto, C. (2020). Duchenne muscular dystrophy: Recent advances in protein biomarkers and the clinical application. *Expert Rev. Proteomics* 17 (5), 365–375. doi:10.1080/14789450.2020.1773806

Alonso-Jiménez, A., Fernández-Simón, E., Natera-de Benito, D., Ortez, C., García, C., Montiel, E., et al. (2021). Platelet derived growth factor-AA correlates with muscle function tests and quantitative muscle magnetic resonance in dystrophinopathies. *Front. Neurol.* 12, 659922. doi:10.3389/fneur.2021.659922

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res.* 22 (10), 2008–2017. doi:10.1101/gr.133744.111

Barthelemy, F., Woods, J. D., Nieves-Rodriguez, S., Douine, E. D., Wang, R., Wanagat, J., et al. (2020). A well-tolerated core needle muscle biopsy process suitable for children and adults. *Muscle and Nerve* 62 (6), 688–698. doi:10.1002/mus.27041

Bello, L., Flanigan, K. M., Weiss, R. B., Dunn, D. M., Swoboda, K. J., Gappmaier, E., et al. (2016). Association study of exon variants in the NF-κB and TGFβ pathways identifies CD40 as a modifier of duchenne muscular dystrophy. *Am. J. Hum. Genet.* 99 (5), 1163–1171. doi:10.1016/j.ajhg.2016.08.023

Bencze, M. (2023). Mechanisms of myofibre death in muscular dystrophies: The emergence of the regulated forms of necrosis in myology. *Int. J. Mol. Sci.* 24 (1), 362. doi:10.3390/ijms24010362

Bowman, A. L., Kontrogianni-Konstantopoulos, A., Hirsch, S. S., Geisler, S. B., Gonzalez-Serratos, H., Russell, M. W., et al. (2007). Different obscurin isoforms localize to distinct sites at sarcomeres. *FEBS Lett.* 581 (8), 1549–1554. doi:10.1016/j.febslet.2007.03.011

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34 (5), 525–527. doi:10.1038/nbt.3519

Brinegar, A. E., Xia, Z., Loehr, J. A., Li, W., Rodney, G. G., and Cooper, T. A. (2017). Extensive alternative splicing transitions during postnatal skeletal muscle development are required for calcium handling functions. *eLife* 6, e27192. doi:10.7554/eLife.27192

Broos, S., Malisoux, L., Theisen, D., van Thienen, R., Ramaekers, M., Jamart, C., et al. (2016). Evidence for ACTN3 as a speed gene in isolated human muscle fibers. *PLOS ONE* 11 (3), e0150594. doi:10.1371/journal.pone.0150594

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* 14 (1), 128. doi:10.1186/1471-2105-14-128

Chen, X., and Li, Y. (2009). Role of matrix metalloproteinases in skeletal muscle: Migration, differentiation, regeneration and fibrosis. *Cell Adhesion Migr.* 3 (4), 337–341. doi:10.4161/cam.3.4.9338

Chu, M., Gregorio, C. C., and Pappas, C. T. (2016). Nebulin, a multi-functional giant. *J. Exp. Biol.* 219 (2), 146–152. doi:10.1242/jeb.126383

Cohen, E., Bonne, G., Rivier, F., and Hamroun, D. (2021). The 2022 version of the gene table of neuromuscular disorders (nuclear genome). *Neuromuscul. Disord.* 31 (12), 1313–1357. doi:10.1016/j.nmd.2021.11.004

Costantini, A., Trevi, E., Pala, M. I., and Fancellu, R. (2016). Can long-term thiamine treatment improve the clinical outcomes of myotonic dystrophy type 1? *Neural Regen. Res.* 11 (9), 1487–1491. doi:10.4103/1673-5374.191225

Das, S., Raj, L., Zhao, B., Kimura, Y., Bernstein, A., Aaronson, S. A., et al. (2007). Hzf determines cell survival upon genotoxic stress by modulating p53 transactivation. *Cell* 130 (4), 624–637. doi:10.1016/j.cell.2007.06.013

de Zélicourt, A., Fayssoil, A., Dakouane-Giudicelli, M., De Jesus, I., Karoui, A., Zarrouki, F., et al. (2022). CD38-NADase is a new major contributor to Duchenne muscular dystrophic phenotype. *EMBO Mol. Med.* 14 (5), e12860. doi:10.15252/emmm.202012860

Deconinck, A. E., Rafael, J. A., Skinner, J. A., Brown, S. C., Potter, A. C., Metzinger, L., et al. (1997). Utrophin-dystrophin-deficient mice as a model for duchenne muscular dystrophy. *Cell* 90 (4), 717–727. doi:10.1016/S0092-8674(00)80532-2

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). Star: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29 (1), 15–21. doi:10.1093/bioinformatics/bts635

Dogra, C., Srivastava, D. S., and Kumar, A. (2008). Protein–DNA array-based identification of transcription factor activities differentially regulated in skeletal muscle of normal and dystrophin-deficient mdx mice. *Mol. Cell. Biochem.* 312 (1), 17–24. doi:10.1007/s11010-008-9716-6

Donner, K., Sandbacka, M., Lehtokari, V.-L., Wallgren-Pettersson, C., and Pelin, K. (2004). Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts. *Eur. J. Hum. Genet.* 12 (9), 744–751. doi:10.1038/sj.ejhg.5201242

Edgerton, V. R., Smith, J. L., and Simpson, D. R. (1975). Muscle fibre type populations of human leg muscles. *Histochem. J.* 7 (3), 259–266. doi:10.1007/BF01003594

Fanin, M., Danieli, G. A., Cadaldini, M., Miorin, M., Vitiello, L., and Angelini, C. (1995). Dystrophin-positive fibers in duchenne dystrophy: Origin and correlation to clinical course. *Muscle and Nerve* 18 (10), 1115–1120. doi:10.1002/mus.880181007

Fischer, M. D., Gorospe, J. R., Felder, E., Bogdanovich, S., Pedrosa-Domellöf, F., Ahima, R. S., et al. (2002). Expression profiling reveals metabolic and structural components of extraocular muscles. *Physiol. Genomics* 9 (2), 71–84. doi:10.1152/physiolgenomics.00115.2001

Flanigan, K. M., Ceco, E., Lamar, K.-M., Kaminoh, Y., Dunn, D. M., Mendell, J. R., et al. (2013). LTBP4 genotype predicts age of ambulatory loss in duchenne muscular dystrophy. *Ann. Neurology* 73 (4), 481–488. doi:10.1002/ana.23819

Flanigan, K. M., Waldrop, M. A., Martin, P. T., Alles, R., Dunn, D. M., Alfano, L. N., et al. (2023). A genome-wide association analysis of loss of ambulation in dystrophinopathy patients suggests multiple candidate modifiers of disease severity. *Eur. J. Hum. Genet.* 2023, 663–673. doi:10.1038/s41431-023-01329-5

Garg, A., O'Rourke, J., Long, C., Doering, J., Ravenscroft, G., Bezprozvannaya, S., et al. (2014). KLHL40 deficiency destabilizes thin filament proteins and promotes nemaline myopathy. *J. Clin. Invest.* 124 (8), 3529–3539. doi:10.1172/jci74994

Gholamalizadeh, M., Jarrahi, A. M., Akbari, M. E., Rezaei, S., Doaei, S., Mokhtari, Z., et al. (2019). The possible mechanisms of the effects of IRX3 gene on body weight: An overview. *Archives Med. Sci. – Atheroscler. Dis.* 4 (1), 225–230. doi:10.5114/amsad.2019.87545

Giulietti, M., Piva, F., D'Antonio, M., D'Onorio De Meo, P., Paoletti, D., Castrignanò, T., et al. (2013). SpliceAid-F: A database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.* 41, D125–D131. Database issue). doi:10.1093/nar/gks997

Grounds, M. D., Terrill, J. R., Al-Mshhdani, B. A., Duong, M. N., Radley-Crabb, H. G., and Arthur, P. G. (2020). Biomarkers for duchenne muscular dystrophy: Myonecrosis, inflammation and oxidative stress. *Dis. Model Mech.* 13 (2), dmm043638. doi:10.1242/dmm.043638

Hämäläinen, N., and Pette, D. (1993). The histochemical profiles of fast fiber types IIB, IID, and IIA in skeletal muscles of mouse, rat, and rabbit. *J. Histochem Cytochem* 41 (5), 733–743. doi:10.1177/41.5.8468455

Han, R., Rader, E. P., Levy, J. R., Bansal, D., and Campbell, K. P. (2011). Dystrophin deficiency exacerbates skeletal muscle pathology in dysferlin-null mice. *Skelet. Muscle* 1 (1), 35. doi:10.1186/2044-5040-1-35

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184 (13), 3573–3587.e29. doi:10.1016/j.cell.2021.04.048

Hathout, Y., Marathi, R. L., Rayavarapu, S., Zhang, A., Brown, K. J., Seol, H., et al. (2014). Discovery of serum protein biomarkers in the mdx mouse model and cross-species comparison to Duchenne muscular dystrophy patients. *Hum. Mol. Genet.* 23 (24), 6458–6469. doi:10.1093/hmg/ddu366

Hirokawa, N., and Noda, Y. (2008). Intracellular transport and kinesin superfamily proteins, KIFs: Structure, function, and dynamics. *Physiol. Rev.* 88 (3), 1089–1118. doi:10.1152/physrev.00023.2007

Hoffman, E. P., Brown, R. H., Jr., and Kunkel, L. M. (1987). Dystrophin: The protein product of the duchenne muscular dystrophy locus. *Cell* 51 (6), 919–928. doi:10.1016/0092-8674(87)90579-4

Hogarth, M. W., Houweling, P. J., Thomas, K. C., Gordish-Dressman, H., Bello, L., Vishwanathan, V., et al. (2017). Evidence for ACTN3 as a genetic modifier of Duchenne muscular dystrophy. *Nat. Commun.* 8 (1), 14143. doi:10.1038/ncomms14143

Hu, Y., Wang, K., and Li, M. (2020). Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers. *PLoS Comput. Biol.* 16 (6), e1007925. doi:10.1371/journal.pcbi.1007925

Huang, Y., and Sanguinetti, G. (2017). Brie: Transcriptome-wide splicing quantification in single cells. *Genome Biol.* 18 (1), 123. doi:10.1186/s13059-017-1248-5

Jakobsson, F., Edström, L., Grimby, L., and Thornell, L. E. (1991). Disuse of anterior tibial muscle during locomotion and increased proportion of type II fibres in hemiplegia. *J. Neurological Sci.* 105 (1), 49–56. doi:10.1016/0022-510X(91)90117-P

Joe, M. K., Kee, C., and Tomarev, S. I. (2012). Myocilin interacts with syntrophins and is member of dystrophin-associated protein complex. *J. Biol. Chem.* 287 (16), 13216–13227. doi:10.1074/jbc.M111.224063

Judge, S. M., Deyhle, M. R., Neyroud, D., Nosacka, R. L., D'Lugos, A. C., Cameron, M. E., et al. (2020). MEF2c-Dependent downregulation of myocilin mediates cancer-induced muscle wasting and associates with cachexia in patients with cancer. *Cancer Res.* 80 (9), 1861–1874. doi:10.1158/0008-5472.Can-19-1558

Kaminski, H. J., Al-Hakim, M., Leigh, R. J., Bashar, M. K., and Ruff, R. L. (1992). Extraocular muscles are spared in advanced duchenne dystrophy. *Ann. Neurology* 32 (4), 586–588. doi:10.1002/ana.410320418

Karpati, G., Carpenter, S., and Prescott, S. (1988). Small-caliber skeletal muscle fibers do not suffer necrosis in mdx mouse dystrophy. *Muscle and Nerve* 11 (8), 795–803. doi:10.1002/mus.880110802

Kute, P. M., Soukarieh, O., Tjeldnes, H., Trégouët, D.-A., and Valen, E. (2022). Small open reading frames, how to find them and determine their function. *Front. Genet.* 12, 796060. doi:10.3389/fgene.2021.796060

Laflamme, N., Lace, B., Thonta Setty, S., Rioux, N., Labrie, Y., Droit, A., et al. (2021). A homozygous deep intronic mutation alters the splicing of nebulin gene in a patient with nemaline myopathy. *Front. Neurol.* 12, 660113. doi:10.3389/fneur.2021.660113

Lam, L. T., Holt, I., Laitila, J., Hanif, M., Pelin, K., Wallgren-Pettersson, C., et al. (2018). Two alternatively-spliced human nebulin isoforms with either exon 143 or exon 144 and their developmental regulation. *Sci. Rep.* 8 (1), 15728. doi:10.1038/s41598-018-33281-6

Lee, H., Huang, A. Y., Wang, L.-k., Yoon, A. J., Renteria, G., Eskin, A., et al. (2020). Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet. Med.* 22 (3), 490–499. doi:10.1038/s41436-019-0672-1

Lee-Gannon, T., Jiang, X., Tassin, T. C., and Mammen, P. P. A. (2022). Biomarkers in duchenne muscular dystrophy. *Curr. Heart Fail. Rep.* 19 (2), 52–62. doi:10.1007/s11897-022-00541-6

Lehtokari, V. L., Kiiski, K., Sandaradura, S. A., Laporte, J., Repo, P., Frey, J. A., et al. (2014). Mutation update: The spectra of nebulin variants and associated myopathies. *Hum. Mutat.* 35 (12), 1418–1426. doi:10.1002/humu.22693

Li, H., Xiao, L., Wang, L., Lin, J., Luo, M., Chen, M., et al. (2018). HLA Polymorphism Affects Risk of de novo Mutation of dystrophin Gene and Clinical Severity of Duchenne Muscular Dystrophy in a Southern Chinese Population. *Front. Neurology* 9, 970. doi:10.3389/fneur.2018.00970

Loreti, M., and Sacco, A. (2022). The jam session between muscle stem cells and the extracellular matrix in the tissue microenvironment. *npj Regen. Med.* 7 (1), 16. doi:10.1038/s41536-022-00204-z

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), 550. doi:10.1186/s13059-014-0550-8

MacArthur, D. G., Seto, J. T., Chan, S., Quinlan, K. G. R., Raftery, J. M., Turner, N., et al. (2008). An Actn3 knockout mouse provides mechanistic insights into the association between alpha-actinin-3 deficiency and human athletic performance. *Hum. Mol. Genet.* 17 (8), 1076–1086. doi:10.1093/hmg/ddm380

Mann, C. J., Perdiguero, E., Kharraz, Y., Aguilar, S., Pessina, P., Serrano, A. L., et al. (2011). Aberrant repair and fibrosis development in skeletal muscle. *Skelet. Muscle* 1 (1), 21. doi:10.1186/2044-5040-1-21

Matsakas, A., Yadav, V., Lorca, S., and Narkar, V. (2013). Muscle ERRγ mitigates Duchenne muscular dystrophy via metabolic and angiogenic reprogramming. *FASEB J.* 27 (10), 4004–4016. doi:10.1096/fj.13-228296

McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019). DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 8 (4), 329–337.e4. doi:10.1016/j.cels.2019.03.003

[DATASET] McKellar, D. W., Walter, L. D., Song, L. T., Mantri, M., Wang, M. F. Z., De Vlaminck, I., et al. (2021). Large-scale integration of single-cell transcriptomic data captures transitional progenitor states in mouse skeletal muscle regeneration (scMuscle). Dryad. v1.1. doi:10.5061/dryad.t4b8gtj34

Morales, M. G., Gutierrez, J., Cabello-Verrugio, C., Cabrera, D., Lipson, K. E., Goldschmeding, R., et al. (2013). Reducing CTGF/CCN2 slows down mdx muscle dystrophy and improves cell therapy. *Hum. Mol. Genet.* 22 (24), 4938–4951. doi:10.1093/hmg/ddt352

Nakka, K., Ghigna, C., Gabellini, D., and Dilworth, F. J. (2018). Diversification of the muscle proteome through alternative splicing. *Skelet. Muscle* 8 (1), 8. doi:10.1186/s13395-018-0152-3

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 37 (7), 773–782. doi:10.1038/s41587-019-0114-2

Nieves Rodríguez, S. (2023). *Transcriptomic analysis of healthy skeletal muscle to identify modulators of differential skeletal muscle susceptibility in Duchenne muscular dystrophy*. Los Angeles: Ph.D., University of California.

Pan, L., Dinh, H. Q., Pawitan, Y., and Vu, T. N. (2021). Isoform-level quantification for single-cell RNA sequencing. *Bioinformatics* 38 (5), 1287–1294. doi:10.1093/bioinformatics/btab807

Park, S., Ranjbarvaziri, S., Zhao, P., and Ardehali, R. (2020). Cardiac fibrosis is associated with decreased circulating levels of full-length CILP in heart failure. *JACC Basic Transl. Sci.* 5 (5), 432–443. doi:10.1016/j.jacbts.2020.01.016

Parolo, S., Marchetti, L., Lauria, M., Misselbeck, K., Scott-Boyer, M. P., Caberlotto, L., et al. (2018). Combined use of protein biomarkers and network analysis unveils deregulated regulatory circuits in Duchenne muscular dystrophy. *PLoS One* 13 (3), e0194225. doi:10.1371/journal.pone.0194225

Pegoraro, E., Hoffman, E. P., Piva, L., Gavassini, B. F., Cagnin, S., Ermani, M., et al. (2011). SPP1 genotype is a determinant of disease severity in Duchenne muscular dystrophy. *Neurology* 76 (3), 219–226. doi:10.1212/WNL.0b013e318207afeb

[DATASET] Petrany, M. J., Swoboda, C. O., Sun, C., Chetal, K., Chen, X., Weirauch, M. T., et al. (2020). snRNA-Seq of multinucleated skeletal myofibers (myoatlas), Synapse. v1. syn21676145 (syn51119242).

Petrof, B. J., Shrager, J. B., Stedman, H. H., Kelly, A. M., and Sweeney, H. L. (1993). Dystrophin protects the sarcolemma from stresses developed during muscle contraction. *Proc. Natl. Acad. Sci. U. S. A.* 90 (8), 3710–3714. doi:10.1073/pnas.90.8.3710

Pettygrove, S., Lu, Z., Andrews, J. G., Meaney, F. J., Sheehan, D. W., Price, E. T., et al. (2014). Sibling concordance for clinical features of Duchenne and Becker muscular dystrophies. *Muscle and Nerve* 49 (6), 814–821. doi:10.1002/mus.24078

Porter, J. D. (2002). Extraocular muscle: Cellular adaptations for a diverse functional repertoire. *Ann. N. Y. Acad. Sci.* 956 (1), 7–16. doi:10.1111/j.1749-6632.2002.tb02804.x

Porter, J. D., Khanna, S., Kaminski, H. J., Rao, J. S., Merriam, A. P., Richmonds, C. R., et al. (2001). Extraocular muscle is defined by a fundamentally distinct gene expression profile. *Proc. Natl. Acad. Sci. U. S. A.* 98 (21), 12062–12067. doi:10.1073/pnas.211257298

Resch, Z. T., and Fautsch, M. P. (2009). Glaucoma-associated myocilin: A better understanding but much more to learn. *Exp. Eye Res.* 88 (4), 704–712. doi:10.1016/j.exer.2008.08.011

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Rooney, W. D., Berlow, Y. A., Triplett, W. T., Forbes, S. C., Willcocks, R. J., Wang, D. J., et al. (2020). Modeling disease trajectory in Duchenne muscular dystrophy. *Neurology* 94 (15), e1622–e1633. doi:10.1212/wnl.0000000000009244

Sandri, M., Minetti, C., Pedemonte, M., and Carraro, U. (1998). Apoptotic myonuclei in human Duchenne muscular dystrophy. *Lab. Invest.* 78 (8), 1005–1016.

Savarese, M., Jonson, P. H., Huovinen, S., Paulin, L., Auvinen, P., Udd, B., et al. (2018). The complexity of titin splicing pattern in human adult skeletal muscles. *Skelet. Muscle* 8 (1), 11. doi:10.1186/s13395-018-0156-z

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH image to ImageJ: 25 years of image analysis. *Nat. Methods* 9 (7), 671–675. doi:10.1038/nmeth.2089

Scripture-Adams, D. D., Chesmore, K. N., Barthélémy, F., Wang, R. T., Nieves-Rodriguez, S., Wang, D. W., et al. (2022). Single nuclei transcriptomics of muscle reveals intra-muscular cell dynamics linked to dystrophin loss and rescue. *Commun. Biol.* 5 (1), 989. doi:10.1038/s42003-022-03938-0

Serdaroglu, A., Gücüyener, K., Erdem, S., Köse, G., Tan, E., and Okuyaz, Ç. (2002). Role of apoptosis in duchenne's muscular dystrophy. *J. Child Neurology* 17 (1), 66–68. doi:10.1177/088307380201700120

Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y., and Ge, Q. (2021). Bias in RNA-seq library preparation: Current challenges and solutions. *Biomed. Res. Int.* 2021, 6647597. doi:10.1155/2021/6647597

Sinha, U., Malis, V., Chen, J. S., Csapo, R., Kinugasa, R., Narici, M. V., et al. (2020). Role of the extracellular matrix in loss of muscle force with age and unloading using magnetic resonance imaging, biochemical analysis, and computational models. *Front. Physiol.* 11, 626. doi:10.3389/fphys.2020.00626

Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., et al. (2017). Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell* 67 (1), 148–161.e5. doi:10.1016/j.molcel.2017.06.003

Spitali, P., Hettne, K., Tsonaka, R., Charrout, M., van den Bergen, J., Koeks, Z., et al. (2018). Tracking disease progression non-invasively in Duchenne and Becker muscular dystrophies. *J. Cachexia Sarcopenia Muscle* 9 (4), 715–726. doi:10.1002/jcsm.12304

Spitali, P., Zaharieva, I., Bohringer, S., Hiller, M., Chaouch, A., Roos, A., et al. (2020). TCTEX1D1 is a genetic modifier of disease progression in Duchenne muscular dystrophy. *Eur. J. Hum. Genet.* 28 (6), 815–825. doi:10.1038/s41431-019-0563-6

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* 6 (7), e21800. doi:10.1371/journal.pone.0021800

Taglietti, V., Kefi, K., Rivera, L., Bergiers, O., Cardone, N., Coulpier, F., et al. (2023). Thyroid-stimulating hormone receptor signaling restores skeletal muscle stem cell regeneration in rats with muscular dystrophy. *Sci. Transl. Med.* 15 (685), 5275. doi:10.1126/scitranslmed.add5275

Terry, E. E., Zhang, X., Hoffmann, C., Hughes, L. D., Lewis, S. A., Li, J., et al. (2018). Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues. *eLife* 7, e34613. doi:10.7554/eLife.34613

Tews, D. S., and Goebel, H. H. (1997). DNA-fragmentation and expression of apoptosis-related proteins in muscular dystrophies. *Neuropathology Appl. Neurobiol.* 23 (4), 331–338. doi:10.1111/j.1365-2990.1997.tb01304.x

Tidball, J. G., Albrecht, D. E., Lokensgard, B. E., and Spencer, M. J. (1995). Apoptosis precedes necrosis of dystrophin-deficient muscle. *J. Cell Sci.* 108 (6), 2197–2204. doi:10.1242/jcs.108.6.2197

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347 (6220), 1260419. doi:10.1126/science.1260419

Valentine, B. A., Cooper, B. J., Cummings, J. F., and de Lahunta, A. (1990). Canine X-linked muscular dystrophy: Morphologic lesions. *J. Neurological Sci.* 97 (1), 1–23. doi:10.1016/0022-510X(90)90095-5

Vincent, B., Bock, K. D., Ramaekers, M., Eede, E. V. d., Leemputte, M. V., Hespel, P., et al. (2007). ACTN3 (R577X) genotype is associated with fiber type distribution. *Physiol. Genomics* 32 (1), 58–63. doi:10.1152/physiolgenomics.00173.2007

Vitting-Seerup, K., and Sandelin, A. (2017). The landscape of isoform switches in human cancers. *Mol. Cancer Res.* 15 (9), 1206–1220. doi:10.1158/1541-7786.Mcr-16-0459

Wagner, K. R., Guglieri, M., Ramaiah, S. K., Charnas, L., Marraffino, S., Binks, M., et al. (2021). Safety and disease monitoring biomarkers in duchenne muscular dystrophy: Results from a phase II trial. *Biomarkers Med.* 15 (15), 1389–1396. doi:10.2217/bmm-2021-0222

Wagner, K. R., McPherron, A. C., Winik, N., and Lee, S.-J. (2002). Loss of myostatin attenuates severity of muscular dystrophy in mdx mice. *Ann. Neurology* 52 (6), 832–836. doi:10.1002/ana.10385

Waters, F. J., Shavlakadze, T., McIldowie, M. J., Piggott, M. J., and Grounds, M. D. (2010). Use of pifithrin to inhibit p53-mediated signalling of TNF in dystrophic muscles of mdx mice. *Mol. Cell. Biochem.* 337 (1), 119–131. doi:10.1007/s11010-009-0291-2

Way, M., Pope, B., Cross, R. A., Kendrick-Jones, J., and Weeds, A. G. (1992). Expression of the N-terminal domain of dystrophin in *E. coli* and demonstration of binding to F-actin. *FEBS Lett.* 301 (3), 243–245. doi:10.1016/0014-5793(92)80249-G

Webster, C., Silberstein, L., Hays, A. P., and Blau, H. M. (1988). Fast muscle fibers are preferentially affected in Duchenne muscular dystrophy. *Cell* 52 (4), 503–513. doi:10.1016/0092-8674(88)90463-1

Weiss, R. B., Vieland, V. J., Dunn, D. M., Kaminoh, Y., Flanigan, K. M., and Project, f. t. U. D. (2018). Long-range genomic regulators of THBS1 and LTBP4 modify disease severity in duchenne muscular dystrophy. *Ann. Neurology* 84 (2), 234–245. doi:10.1002/ana.25283

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–d1082. doi:10.1093/nar/gkx1037

Woodman, K. G., Coles, C. A., Toulson, S. L., Gibbs, E. M., Knight, M., McDonagh, M., et al. (2018). Benfotiamine reduces pathology and improves muscle function in mdx mice. *bioRxiv* 288621. doi:10.1101/288621

Wu, X., Dong, N., Yu, L., Liu, M., Jiang, J., Tang, T., et al. (2022). Identification of immune-related features involved in duchenne muscular dystrophy: A bidirectional transcriptome and proteome-driven analysis. *Front. Immunol.* 13, 1017423. doi:10.3389/fimmu.2022.1017423

Yamada, K., Andrews, C., Chan, W. M., McKeown, C. A., Magli, A., de Berardinis, T., et al. (2003). Heterozygous mutations of the kinesin KIF21A in congenital fibrosis of the extraocular muscles type 1 (CFEOM1). *Nat. Genet.* 35 (4), 318–321. doi:10.1038/ng1261

Zhang, C. L., Zhao, Q., Liang, H., Qiao, X., Wang, J. Y., Wu, D., et al. (2018). Cartilage intermediate layer protein-1 alleviates pressure overload-induced cardiac fibrosis via interfering TGF-β1 signaling. *J. Mol. Cell Cardiol.* 116, 135–144. doi:10.1016/j.yjmcc.2018.02.006

Zhang, Y., Li, S., Wen, X., Tong, H., Li, S., and Yan, Y. (2021). MYOC promotes the differentiation of C2C12 cells by regulation of the TGF-β signaling pathways via CAV1. *Biol. (Basel)* 10 (7), 686. doi:10.3390/biology10070686

Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* 2 (3), lqaa078. doi:10.1093/nargab/lqaa078

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 31 August 2023
DOI 10.3389/fgene.2023.1194882

OPEN ACCESS

EDITED BY
Yuriy L. Orlov,
I.M.Sechenov First Moscow State Medical University, Russia

REVIEWED BY
Yunia Sribudiani,
Padjadjaran University, Indonesia
Jay V Patankar,
University of Erlangen Nuremberg, Germany

*CORRESPONDENCE
Ye Shu,
sy999222@hotmail.com
Xiaoting Wu,
wxt1@medmail.com.cn

RECEIVED 27 March 2023
ACCEPTED 16 August 2023
PUBLISHED 31 August 2023

CITATION
Yang Y, Xia L, Yang W, Wang Z, Meng W, Zhang M, Ma Q, Gou J, Wang J, Shu Y and Wu X (2023), Transcriptome profiling of intact bowel wall reveals that PDE1A and SEMA3D are possible markers with roles in enteric smooth muscle apoptosis, proliferative disorders, and dysautonomia in Crohn's disease.
Front. Genet. 14:1194882.
doi: 10.3389/fgene.2023.1194882

Transcriptome profiling of intact bowel wall reveals that PDE1A and SEMA3D are possible markers with roles in enteric smooth muscle apoptosis, proliferative disorders, and dysautonomia in Crohn's disease

Yun Yang[1,2,3], Lin Xia[1,2], Wenming Yang[4], Ziqiang Wang[1,2], Wenjian Meng[1,2], Mingming Zhang[1,2,3], Qin Ma[3,4], Junhe Gou[5], Junjian Wang[6], Ye Shu[1,2]* and Xiaoting Wu[4,7]*

[1]Department of General Surgery, West China Hospital, Sichuan University, Chengdu, China, [2]Colorectal Cancer Center, West China Hospital, Sichuan University, Chengdu, China, [3]Department of General Surgery, West China Chengdu Shangjin Nanfu Hospital, Sichuan University, Chengdu, China, [4]Division of Gastrointestinal Surgery, Department of General Surgery, West China Hospital, Sichuan University, Chengdu, China, [5]Department of Pathology, West China Hospital, Sichuan University, Chengdu, China, [6]Department of Laboratory Medicine, West China Hospital, Sichuan University, Chengdu, China, [7]Colorectal and Pelvic Floor Center, West China Tianfu Hospital, Sichuan University, Chengdu, China

Background: Inflammatory bowel disease (IBD) is a complex and multifactorial inflammatory condition, comprising Crohn's disease (CD) and ulcerative colitis (UC). While numerous studies have explored the immune response in IBD through transcriptional profiling of the enteric mucosa, the subtle distinctions in the pathogenesis of Crohn's disease and ulcerative colitis remain insufficiently understood.

Methods: The intact bowel wall specimens from IBD surgical patients were divided based on their inflammatory status into inflamed Crohn's disease (iCD), inflamed ulcerative colitis (iUC) and non-inflamed (niBD) groups for RNA sequencing. Differential mRNA GO (Gene Ontology), and KEGG (Kyoto Encyclopedia of Genes and Genomes), and GSEA (Gene Set Enrichment Analysis) bioinformatic analyses were performed with a focus on the enteric autonomic nervous system (ANS) and smooth muscle cell (SMC). The transcriptome results were validated by quantitative polymerase chain reaction (qPCR) and immunohistochemistry (IHC).

Results: A total of 2099 differentially expressed genes were identified from the comparison between iCD and iUC. Regulation of SMC apoptosis and proliferation were significantly enriched in iCD, but not in iUC. The involved gene PDE1A in iCD was 4-fold and 1.5-fold upregulated at qPCR and IHC compared to that in iUC. Moreover, only iCD was significantly associated with the gene sets of ANS abnormality. The involved gene SEMA3D in iCD was upregulated 8- and 5-fold at qPCR and IHC levels compared to iUC.

Conclusion: These findings suggest that PDE1A and SEMA3D may serve as potential markers implicated in enteric smooth muscle apoptosis, proliferative disorders, and dysautonomia specifically in Crohn's disease.

KEYWORDS

Crohn's disease, RNA seq, autonomic nervous system, smooth muscle cell, proliferation, apoptosis, Sema3D, PDE1A

1 Introduction

Inflammatory bowel disease (IBD) is a chronic idiopathic inflammatory disorder characterized by relapsing and remitting symptoms. The two most common forms of IBD are Crohn's disease (CD) and ulcerative colitis (UC) (Assadsangabi et al., 2019). Morphologically, UC primarily affects the rectum and colon, exhibiting superficial inflammation confined to the mucosal and submucosal layers, often accompanied by cryptitis and crypt abscesses. In contrast, CD is characterized by a non-continuous and transmural pattern of inflammation, presenting additional complications such as thickened submucosa and muscularis propria, intestinal fibrosis, strictures, fissuring ulceration, non-caseating granulomas, abscesses, and fistulas (Abraham and Cho, 2009). A notable feature of CD is the presence of fibrostenosis, which contributes to therapeutic challenges and the need for surgical resection. However, a recent histological grading scheme study discovered that smooth muscle hyperplasia/hypertrophy, rather than fibrosis, is the primary change associated with the "fibrostenosis" phenotype in CD. Neuromuscular hyperplasia/hypertrophy was also identified as a significant change (Chen et al., 2017).

Although morphological and histological differences exist between CD and UC, a comprehensive whole-genome gene expression meta-analysis (Granlund et al., 2013) based on 11 available datasets (Wu et al., 2007; Galamb et al., 2008a; Ahrens et al., 2008; Galamb et al., 2008b; Carey et al., 2008; Kugathasan et al., 2008; Noble et al., 2008; Arijs et al., 2009; Olsen et al., 2009; Bjerrum et al., 2010; van Beelen Granlund et al., 2013) did not unveil any significant differences between CD and UC. Interestingly, gene expression in the inflamed mucosa from both UC and CD was remarkably similar. The patterns of antimicrobial peptide (AMP) and T-helper cell-related gene expression were also comparable, except for the higher expression of IL23A observed in UC compared to CD. Another study conducted by the IBD-CHARACTER consortium, which included 323 subjects, found that a comparison of inflamed UC and inflamed CD identified 204 highly differentially expressed upregulated transcripts and 58 downregulated transcripts (Consortium, 2021). These two gene expression signatures were highly correlated, suggesting that inflammation might mask underlying biological differences among the diagnostic groups. Furthermore, when comparing inflamed biopsies from UC and CD on a biological pathway level, the normalized enrichment scores were remarkably similar, irrespective of diagnosis or whether healthy or symptomatic controls were used in the comparison. However, mitochondria-associated pathways exhibited negative normalized enrichment scores in inflamed UC compared to inflamed CD (Vatn et al., 2022).

Despite these findings, previous studies (Wu et al., 2007; Galamb et al., 2008a; Ahrens et al., 2008; Galamb et al., 2008b; Carey et al., 2008; Kugathasan et al., 2008; Noble et al., 2008; Arijs et al., 2009; Olsen et al., 2009; Bjerrum et al., 2010; van Beelen Granlund et al., 2013; Vatn et al., 2022) encountered limitations due to the challenges of obtaining surgical resection specimens. Instead, mucosa-submucosa (SM) specimens from colonoscopy pinch biopsies were commonly used. However, these specimens lack the layers of muscularis propria (MP) and subserosal adventitia (SS), making it difficult to fully elucidate the underlying disease-inducing mechanisms in IBD. Consequently, subtle differences between CD and UC might have been unintentionally overlooked.

Therefore, the present study aims to utilize intact bowel wall specimens obtained during surgical resection. Through RNA-seq, bioinformatics analysis, and validation using quantitative polymerase chain reaction (qPCR) and immunohistochemistry (IHC), we aim to explore the subtle differences between CD and UC, with a primary focus on smooth muscle cells (SMCs) and the enteric autonomic nervous system (ANS).

2 Materials and methods

2.1 Specimen collection

All the intact bowel wall specimens were collected from the Biobank of West China Hospital (WCH), Sichuan province, China. The study was approved (No. 20221470) and supervised by the WCH Ethics Committee. Patients who received bowel resection after being diagnosed with IBD were recruited. Informed consent was obtained from all patients in the study prior to the medical history and collection of specimens. For the inflamed CD (iCD) and inflamed UC (iUC) groups, specimens from the most inflamed segment within the colon were selected. For the non-inflamed (niBD) group, specimens from the uninvolved non-inflamed (niCD/niUC) segment within the colon were selected. The postoperative pathological diagnosis was confirmed by a team of pathologists using the guidelines on the pathological diagnosis of IBD (Shen and Weber, 2017).

2.2 RNA extraction and library preparation

Total RNA was extracted using TRIzol reagent (Cat.# 15596018, Thermo Fisher Scientific, United States of America) according to the manufacturer's protocol. RNA purity and quantification were evaluated on the NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). RNA integrity was evaluated using the Agilent 2100 Bioanalyzer (Agilent Technologies, United States of America). The specimens with RNA integrity number (RIN) ≥7 were subjected to the subsequent analysis. The libraries were constructed using TruSeq Stranded Total RNA with Ribo-Zero Gold (Cat.# RS-122–2301, Illumina, United States of America) according to the manufacturer's instructions and sequenced on the Illumina HiSeq X Ten platform; 150-bp paired-end reads were generated. The sequencing and analyses were performed by OE Biotech Co., Ltd. (Shanghai, China).

TABLE 1 Primer sequences.

Num	Gene symbol	Direction	Primer sequences	Product length (bp)	Tm (°C)
1	SEMA3D	Forward	GTTCATCAGAAGGACTGGATT	89	60
		Reverse	TAGAAAGATGTGGTCTTTGGC		
2	SLC18A2	Forward	GATTTCCATGGCTCATGACA	89	60
		Reverse	TTCTTTGGCAGGTGGACT		
3	PDE1A	Forward	AAGCAAGTGGAGAGCATAG	85	60
		Reverse	ACAGGAATCTTGAAACGGT		
4	TACR1	Forward	GAGAAATAGGAGTTGCAGGC	84	60
		Reverse	AAGAAATTCCACCGGTCAC		
5	SPHK1	Forward	ACCATTATGCTGGCTATGAG	96	60
		Reverse	GCAGGTTCATGGGTGACA		
6	ADRA1A	Forward	GTGAACATTTCCAAGGCCA	81	60
		Reverse	CACTAGGATGTTACCCAGC		
7	GAPDH	Forward	CCTCACAGTTGCCATGTAGA	69	60
		Reverse	TGGTACATGACAAGGTGCG		

2.3 Bioinformatics analysis

Raw reads for each specimen were generated in FASTQ format and processed using the Trimmomatic software (Bolger et al., 2014). Subsequently, clean reads were obtained by removing the adapter and ploy-N or low-quality sequences from raw data. Then, the clean reads for each specimen were mapped to the human genome (GRCh38) using HISAT2(Kim et al., 2015). For mRNAs, FPKM (fragments per kilobase of exon model per million mapped fragments) (Roberts et al., 2011) of each gene was calculated using Cufflinks (Trapnell et al., 2010), and the read counts of each gene were obtained by HTSeq-count (Anders et al., 2015). Differential expression analysis was performed using DESeq (2012) R package (Anders et al., 2012). p-value <0.05 was set as the threshold for a significantly differential expression. The differential mRNA GO (Ashburner et al., 2000, 2021) and KEGG (Kanehisa et al., 2010; Kanehisa et al., 2017) enrichments were analyzed based on selected differential transcripts with p-values <0.05 and fold-change (FC) > 1.5 based on the hypergeometric distribution test. Also, gene set expression analysis (GSEA) of molecular pathways affected by differentially expressed genes (DEGs) was performed by GSEA R (v1.2) with weighted enrichment statistic and Signal2Noise for gene ranking (Mootha et al., 2003; Subramanian et al., 2005).

The data of gene sets analyzed on GSEA are summarized in Supplementary Tables S1 and S2.

2.4 Quantitative polymerase chain reaction (qPCR)

Quantification was performed with a two-step reaction: reverse transcription and PCR. Each 10-μL reaction of reverse transcription consisted of 0.5 μg RNA, 2 μL of 5× TransScript All-in-one SuperMix for qPCR, and 0.5 μL of gDNA Remover. The reactions were performed on a GeneAmp® PCR System 9700 (Applied Biosystems, United States of America) at 42°C for 15 min and 85°C for 5 s. The 10-μL RT reaction mix was then diluted in 90 μL nuclease-free water and held at −20°C. Real-time PCR was performed on LightCycler® 480 II Real-time PCR Instrument (Roche, Swiss) in a 10-μL PCR reaction mixture in a 384-well optical plate (Roche, Swiss), consisting of 1 μL of cDNA, 5 μL of 2× PerfectStartTM Green qPCR SuperMix, 0.2 μL of 10 μM forward primer, 0.2 μL of 10 μM reverse primer, and 3.6 μL of nuclease-free water. The reactions were incubated at 94°C for 30 s, followed by 45 cycles of 94°C for 5 s, and 60°C for 30 s. Each sample was assessed in triplicate. Finally, melting curve analysis was performed to validate the specific qPCR product. The expression levels of mRNAs were normalized to GAPDH. The primer sequences were designed in the laboratory and synthesized by TsingKe Biotech (Beijing, China), based on the mRNA sequences obtained from the NCBI database (Table 1).

2.5 Immunohistochemistry (IHC)

The expression of SEMA3D and PDE1A was assessed by IHC using formalin-fixed paraffin-embedded (FFPE) tissue. The staining antibodies were as follows: SEMA3D (dilution 1/50; Cat.# NBP1-85517, NOVUS, Centennial, United States of America) and PDE1A (dilution 1/200; Cat.# 12442-2-AP, Proteintech, Wuhan, China) (Supplementary Table S3). Antibody detection and visualization were performed using DAB (3,3′-diaminobenzidine) as the chromogenic substrate. The images were captured under BA400 Digital microscope (Motic, China). The percentage of DAB-positive tissue in each image was calculated using the Halo data analysis system (Halo 101-WL-HALO-1, Indica labs, United States of America).

TABLE 2 Demographic characteristics.

	iCD (n = 6)	iUC (n = 6)	niBD (n = 6)	*p*-value
Patient characteristics				
Age (years)	28.83 ± 9.06	64.00 ± 7.80	42.67 ± 18.69	0.0009*
Gender	6M 0F	5M 1F	5M 1F	0.5698
Smoker	1/6	3/6	2/6	0.4724
Alcohol	0/6	4/6	2/6	0.0498*
Preoperative treatment history				
5-ASA	4/6	6/6	4/6	0.2765
Steroids	3/6	4/6	3/6	0.7985
Immunomodulation	2/6	2/6	2/6	0.9999
Anti-TNF	2/6	0/6	0/6	0.1054
Non-anti-TNF biologic treatment	NA	NA	NA	NA
Location involvement				
Ileum	5/6	0/6	NA	0.0152*
Cecum	3/6	0/6	NA	0.1818
Ascending colon	3/6	5/6	NA	0.5455
Transverse colon	3/6	6/6	NA	0.1818
Descending colon	1/6	6/6	NA	0.0152*
Sigmoid	2/6	5/6	NA	0.2424
Rectal	0/6	5/6	NA	0.0152*
Phenotypes				
Depth score of inflammatory infiltration※	3.67 ± 0.52	2.00 ± 1.10	NA	0.0071*
Acute inflammation	2/6	4/6	NA	0.5671
Chronic inflammation	6/6	5/6	NA	0.9999
Ulcers	2/6	5/6	NA	0.2424
Penetrate/fistula	2/6	0/6	NA	0.4545
Stricturing	4/6	0/6	NA	0.0606
Periganglitis	2/6	0/6	NA	0.4545
Postoperative outcomes				
Biologic use	3/6	0/6	1/6	0.1054
Median time to first resection (months)	54.33 ± 45.86	53.08 ± 54.42	53.00 ± 45.29	0.9986
Median time from first resection to second resection (months)	NA	NA	NA	NA

※Depth score of inflammatory infiltration. Mucosa: Score 1; muscularis mucosa (MM): Score 2; submucosa (SM): Score 3; muscularis propria (MP): Score 4; subserosal adventitia (SS): Score 5.

2.6 Statistical analysis

Contingency data were assessed for significant differences using chi-square or Fisher's exact test. The data were expressed as means ± standard deviation (SD). The comparison between the two groups was assessed using the Holm-Šídák test. The multiple comparisons were evaluated using Fisher's LSD (least significant difference) test. *p*-value <0.05 indicated a statistically significant difference. The statistical analyses were performed using GraphPad Prism9 (GraphPad Software, United States of America).

3 Results

3.1 Demographic characteristics

The demographic and clinical information of the individual patient are summarized in Supplementary Table S4, and the grouping design is provided in Supplementary Table S5.

Both CD and UC specimens were divided based on inflammatory status into inflamed (iCD and iUC) and non-inflamed (niBD: niCD + niUC) groups.The total number of specimens primarily included 6 iCD, 6 iUC, and 6 niBD for RNA extraction and quality control. Since bacteria RNA contamination was detected in iCD4, iUC6 and niBD6 (Supplementary Figure S1), these 3 samples were excluded, the final total number of 15 specimens including 5 iCDs (iCD1, iCD2, iCD3, iCD5 and iCD6), 5 iUCs(iUC1, iUC2, iUC3, iUC4 and iUC5), and 5 niBDs (niBD1, niBD2, niBD3, niBD4, niBD5)were used for the subsequent bioinformatics analysis.

Overall, patients with UC had greater left colon (descending colon or rectal) involvement (p = 0.0152). Patients with CD tended to have a young onset age (iCD vs. iUC, 28.83 ± 9.06 vs. 64 ± 7.80 years old, $p < 0.0001$) and ileal involvement ($p = 0.0152$) and required the postoperative biological therapy. Importantly, unlike iUC, iCD had a higher depth score of inflammatory infiltration (iCD vs. iUC, 3.67 ± 0.52 vs. 2.00 ± 1.10, p = 0.0071) (Table 2).

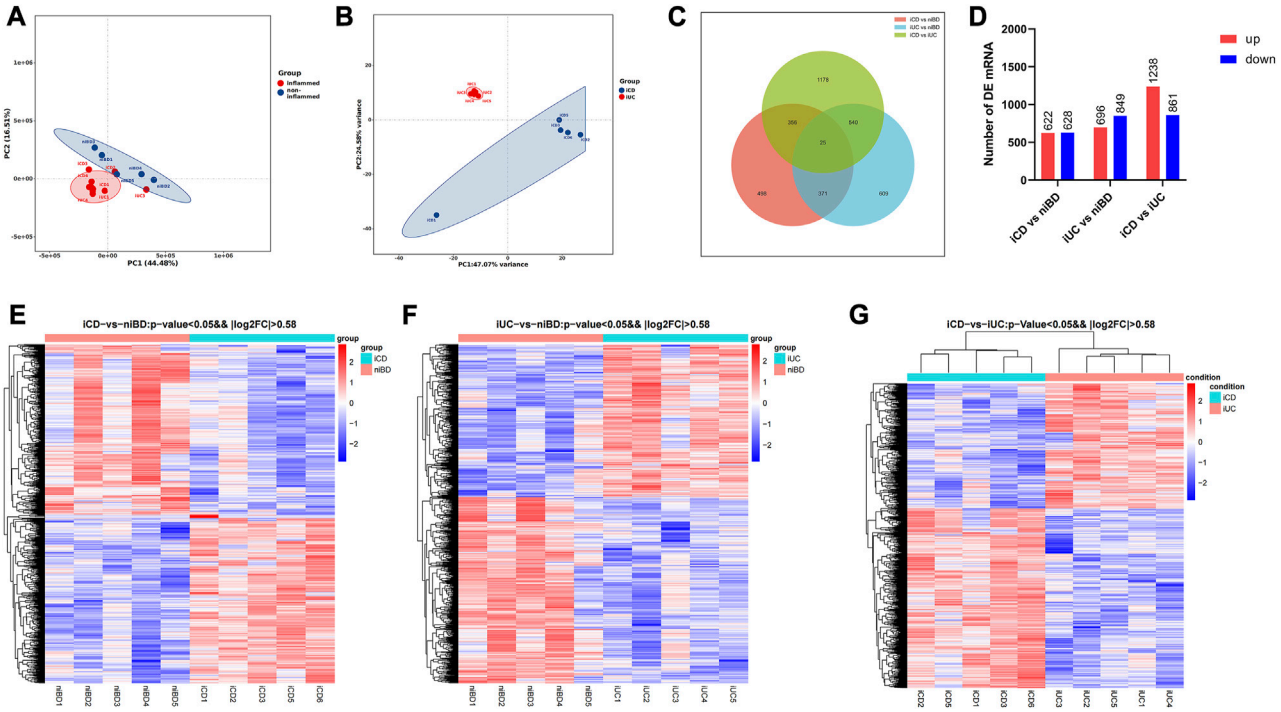


FIGURE 1
PCA and differential expression analysis of transcriptomic profiling. **(A,B)** Principal Component Analysis (PCA) plot displaying mRNA transcriptomic data. Each point represents a specific specimen, while each color represents a different group. Numbers along the perimeter indicate principal components (PC1-PC2), and numbers in parentheses represent the percentage variance accounted for by each PC. **(C)** Venn diagram illustrating the overlaps between differentially expressed mRNAs identified in iCD vs. niBD, iUC vs. niBD, and iCD vs. iUC comparisons. The criteria for differential expression were log2 fold change (log2FC) greater than absolute value of 0.58 and an adjusted *p*-value less than 0.05. **(D)** Bar graph showing the number of dysregulated mRNAs. Upregulated mRNAs are depicted in red, while downregulated mRNAs are shown in blue. The criteria for dysregulation were log2FC greater than absolute value of 0.58 and an adjusted *p*-value less than 0.05. **(E–G)** Heatmap representing differentially expressed mRNAs identified through whole transcriptomic analysis of IBD specimens. Each row corresponds to one mRNA, and each column represents a specimen. Upregulated and downregulated mRNAs are highlighted in red and blue, respectively. The dendrograms on top of the heatmap display hierarchical clustering correlation between the specimens.

3.2 Transcriptome profiling distinguished the differences between iCD and iUC

Based on principal component analysis (PCA) of mRNA expressions, the results showed that approximately 44.48% and 16.51% of the variability in gene expression data were captured by the first and second principal components (PC1 and PC2), respectively (Figure 1A). This indicated that the non-inflamed specimens (niBD) formed a distinct cluster separate from the inflamed specimens (iCD and iUC). Furthermore, on PC1 and PC2, there was clear separation between iCD and iUC, accounting for 47.07% and 24.58% of the variability, respectively (Figure 1B). These findings suggest significant heterogeneity between iCD and iUC.

RNA transcript differential expression analysis of iCD, iUC, and niBD was performed after high-throughput RNA sequencing. The genes with fold-change (FC) > 1.5 and adjusted *p*-value <0.05 were considered differentially expressed genes (DEGs) with statistical significance. A total of 1250 and 1545 DEGs were identified from either iCD or iUC specimens compared to the niBD group (Figures 1C, D). Then, 2099 DEGs were identified when iCD was compared to iUC specimens (Figures 1C, D). To stratify the iCD, iUC, and niBD specimens, the expression profiles of DE mRNA (FC > 1.5) were compared through unsupervised hierarchical clustering. Compared to the niBD group, the heat map of these DE mRNAs showed intra-group similarity in the iCD or the iUC group (Figures 1E, F). Notably, the comparison of the iCD vs. iUC revealed a tight intra-group cluster and distinguished iCD from iUC (Figure 1G), indicating an underlying difference between iCD and iUC.

3.3 CD revealed dysregulation of enteric SMC apoptosis and proliferation

To identify disrupted biological processes and pathways in IBD patients, gene enrichment analysis was conducted to obtain overrepresented gene ontology (GO) terms of biological processes and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways from dysregulated genes. When GO terms were used, the two SMC-related terms, “regulation of SMC apoptotic process” and “regulation of SMC proliferation” were highly enriched in iCD but not in iUC (Figure 2A). Among the identified genes such as AGTR1, PDE1A, RBM10, SIRT1, BMP4, NPPC, NR4A3, PRKDC, TACR1, TCF7L2, XRCC5, and XRCC6, volcano plots revealed that

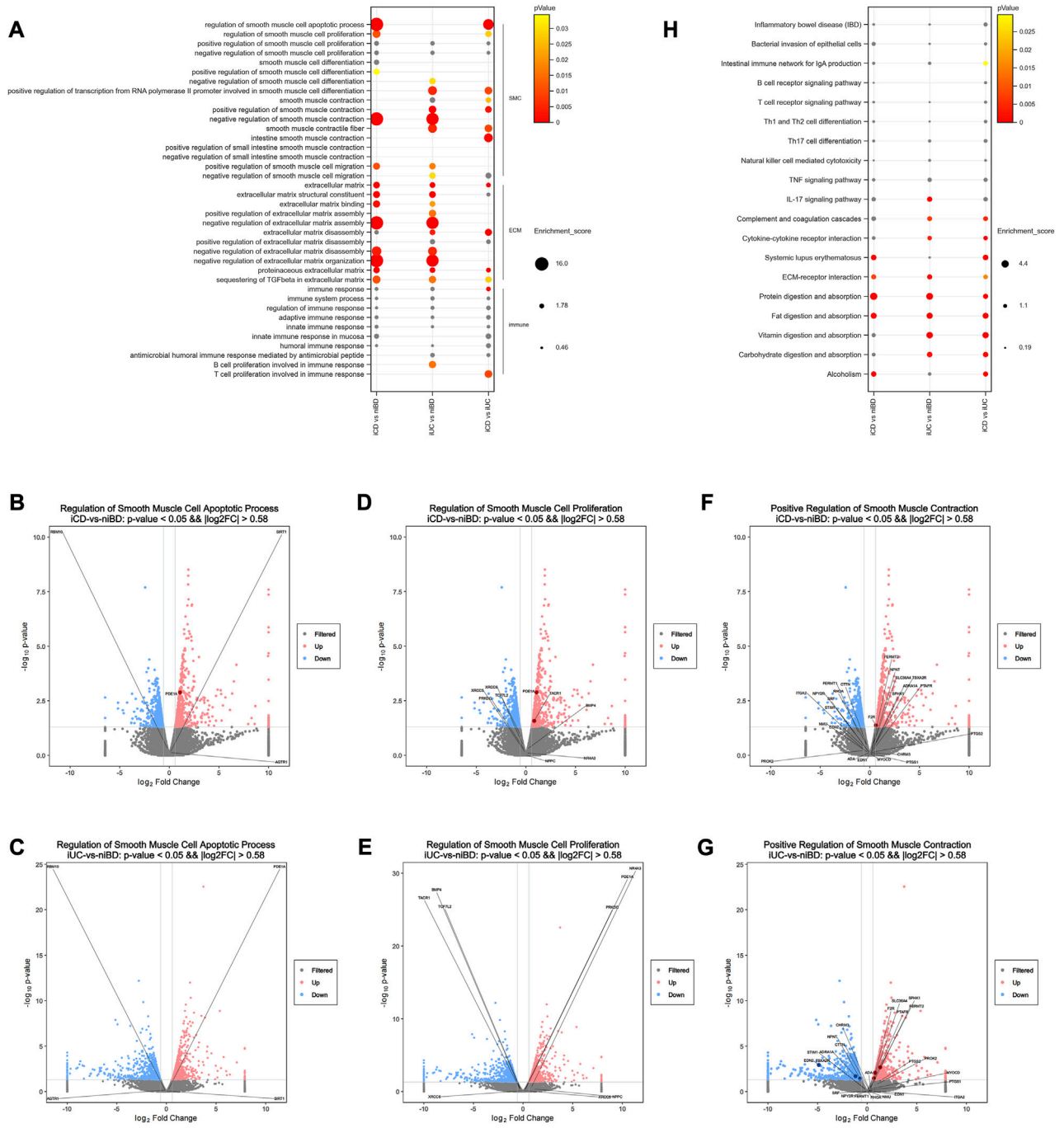


FIGURE 2
Pathway analysis of dysregulated mRNAs reveals dysregulated enteric smooth muscle cell apoptosis and proliferation in CD. **(A)** Selected Gene Ontology (GO) biological processes that exhibit significant enrichment among dysregulated mRNAs(log2FC >|0.58| and adjusted *p*-value <0.05). The enrichment score for each dysregulated gene annotated to the corresponding GO term is indicated. *p*-values greater than 0.05 are labeled in grey. **(B–G)** Volcano plots depicting dysregulated mRNAs. Pink and blue dots represent upregulated and downregulated mRNAs, respectively (log2FC >|0.58| and adjusted *p*-value <0.05). Non-significant mRNAs are displayed in grey. Dysregulated mRNAs associated with selected processes are labeled in dark red or blue. **(H)** Selected KEGG signaling pathways significantly enriched among dysregulated mRNAs (log2FC >|0.58| and adjusted *p*-value <0.05). The enrichment score for each dysregulated gene annotated to the corresponding KEGG pathway is presented. *p*-values greater than 0.05 are labeled in grey.

PDE1A and TACR1 are the only two significantly upregulated mRNAs (log2FC >|0.58| and adjusted *p*-value <0.05) (Figures 2B–E). Similarly, "positive regulation of SMC contraction" was highly enriched in iUC but not in iCD; the genes involved (Figures 2F, G) were ADA, ADRA1A, CHRM3, CTTN, EDN1, EDN2, F2R, FERMT1, FERMT2, ITGA2, MYOCD, NMU, NPNT, NPY2R, PROK2, PTAFR, PTGS1, PTGS2, RHOA, SLC36A4, SPHK1, SRF, STIM1, and TBXA2R. These altered biological processes were overrepresented in direct comparison between the iCD and iUC groups (Figure 2A), which alludes to

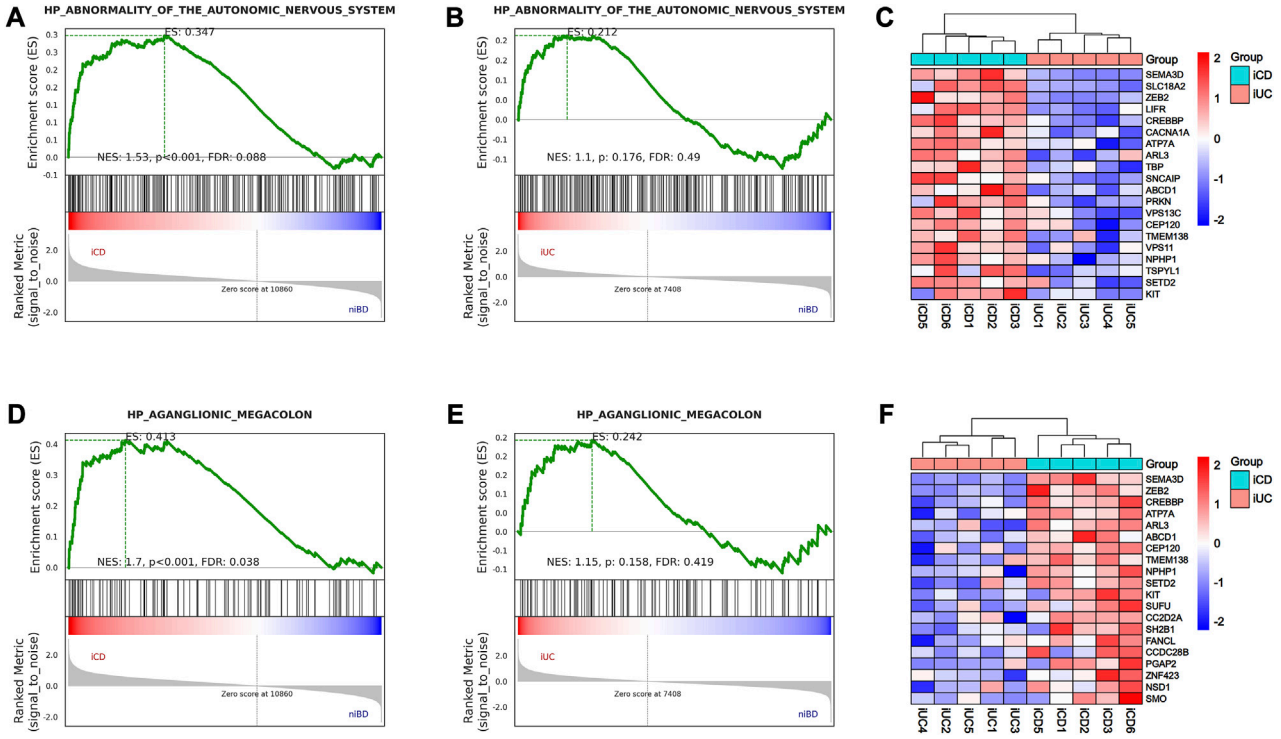


FIGURE 3
GSEA analysis shows activation of abnormality of the autonomic nervous system (ANS) and aganglionic megacolon in iCD. **(A,B)** Gene Set Enrichment Analysis (GSEA) of abnormality of the autonomic nervous system-related gene sets in iCD and iUC specimens vs. niBD. **(C)** Top 20 core enrichment genes associated with abnormality of the autonomic nervous system. **(D,E)** GSEA of aganglionic megacolon-related gene sets in iCD and iUC specimens vs. niBD. **(F)** Top 20 core enrichment genes associated with aganglionic megacolon.

SMC phenotypic and subtle functional divergence between CD and UC. Notably, in extracellular matrix (ECM)- or immune-related terms, CD and UC presented similar enrichment patterns, except slight differences in biological processes of ECM, ECM disassembly, and proteinaceous ECM (Figure 2A).

When KEGG pathways were used, the pathways of bacterial invasion of epithelial cells, B or T cell receptor signaling, Th1/2/17 cell differentiation, natural killer cell-mediated cytotoxicity, TNF signaling, and IL-17 signaling did not show significant differences in iCD or iUC (Figure 2H). Interestingly, the activation of complement and coagulation cascades and cytokine-cytokine receptor interaction was observed significantly in iUC, whereas activation of systemic lupus erythematosus was predominant in iCD (Figure 2H). These activations were overrepresented in the direct comparison between the iCD and iUC groups (Figure 2H). Similar to the data obtained by the enrichment analysis of GO terms, ECM-receptor interaction pathway was activated in both iCD and iUC, albeit with slight differences (Figure 2H). In the pathways of protein/fat digestion and absorption, iCD and iUC presented similar activation but with slight variances; however, vitamin/carbohydrate digestion and absorption pathways were only activated in iUC (Figure 2H), suggesting that varying degrees of gastrointestinal mucosa injury have a differential impact on digestion and absorption between CD and UC. Moreover, the activation of "alcoholism" was enriched in iCD, not iUC (Figure 2H).

3.4 CD exhibited abnormalities in the enteric ANS

The enteric nervous system (ENS) is a part of the ANS located in the digestive tract and innervating SMC with a marked influence on gastrointestinal function. Herein, we explored the ENS deviance in iCD and iUC using GSEA to identify the genes associated with abnormality of the autonomic nervous system and aganglionic megacolon. As a result, only iCD was significantly associated with the gene sets related to abnormality of the autonomic nervous system (Figures 3A, B) and aganglionic megacolon (Figures 3D, E), indicating that CD may involve pathogenic activity resembling ENS abnormalities or gangliopathy. Notably, SEMA3D emerged as the top-ranking gene among the top 20 core enrichment genes in both ANS abnormality (Figure 3C) and aganglionic megacolon (Figure 3F) gene sets.

3.5 Validating dysregulated genes SEMA3D and PDE1A: Implications for SMC and ANS dysfunction in CD

We conducted validation of dysregulated key genes involved in smooth muscle cell (SMC) apoptosis and proliferation, as well as abnormalities in the autonomic nervous system (ANS), using quantitative polymerase chain reaction (qPCR). The results

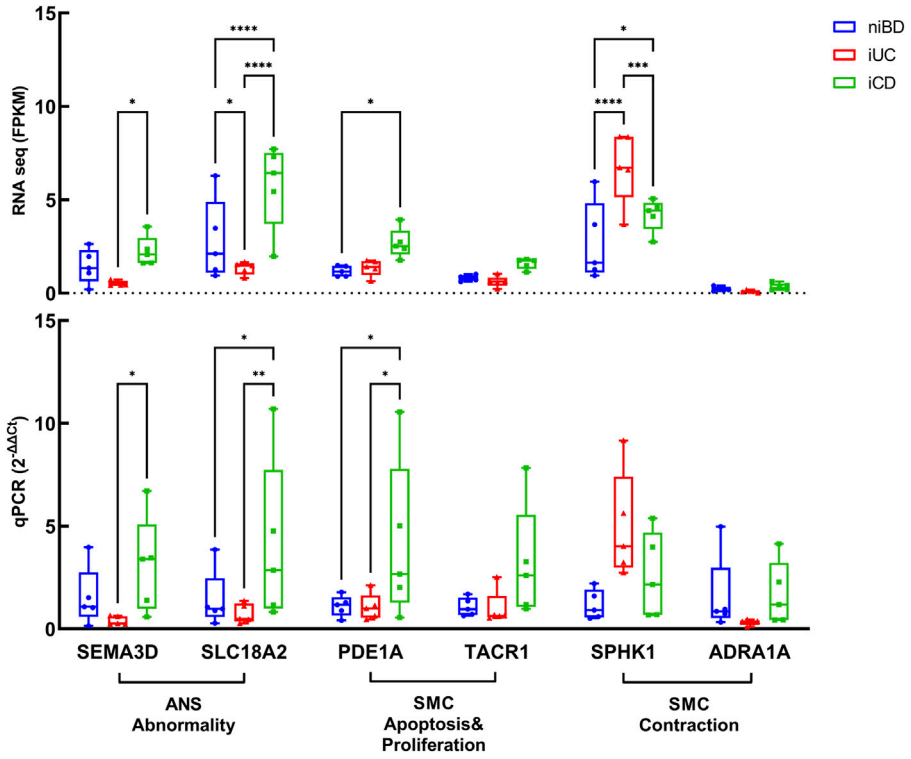


FIGURE 4
Validation by real-time qPCR of genes overactivated or downregulated in colon specimens of iCD (n = 5) and iUC (n = 5) compared to those from niBD (n = 5). Graphs display interleaved box and whisker plots representing the range from minimum to maximum values. For RNA sequence values expressed in FPKM and qPCR values expressed in $2^{-\Delta\Delta Ct}$, statistical analysis was performed using Fisher's Least Significant Difference (LSD) test.Nonsignificant *p*-values (>0.05) are denoted. Asterisks (*) indicate statistical significance levels: *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

confirmed that SEMA3D (iUC vs. iCD, $p = 0.0309$) and SLC18A2 (iUC vs. iCD, $p = 0.0087$), which were identified as core enrichment genes for ANS abnormality, were upregulated by 8-fold and 5-fold respectively in iCD (Figure 4). Additionally, the core enrichment gene PDE1A, implicated in SMC apoptosis and proliferation, exhibited a 4-fold upregulation in iCD (iUC vs. iCD, $p = 0.0144$) (Figure 4). Conversely, SPHK1 was upregulated in iUC for positive regulation of SMC contraction (niBD vs. iUC, $p = 0.0030$) (Figure 4).

The dysregulation of key DEGs at the protein levels detected in iCD and iUC (already confirmed at the mRNA level by RNAseq and real-time qPCR) was assessed by immunohistochemistry (IHC) staining. The clinical information of specimen sections is provided in Supplementary Table S6.

In total, we subjected 15 specimens from iCD, 7 specimens from iUC, and 4 non-inflammatory specimens from niBD (2 niCD +2 niUC) to IHC staining. PDE1A, associated with SMC apoptosis and proliferation, and SEMA3D, associated with ANS abnormality were selected for IHC staining. Consistent with the data presented in Table 2, CD patients tended to have a early onset age (Figure 5A), histologically higher depth score of inflammatory infiltration (Figure 5C), and perigangalionitis (iCD vs. iUC, $p = 0.0167$) (Figure 5D) (Supplementary Figure S2). The IHC data demonstrated that SEMA3D protein levels were upregulated in muscularis propria (iUC vs. iCD, $p = 0.0178$) (Figures 6A, B) and mucosal layer (iUC vs. iCD, $p = 0.0023$) of iCD (Figures 6A, C), and displayed significant aggregation around the ganglia in iCD (Figure 6A). Additionally, the protein level of PDE1A was significantly increased in muscularis propria (iUC vs. iCD, $p = 0.0128$) (Figures 6A, B) and mucosa (iUC vs. iCD, $p = 0.0243$) layers of iCD (Figures 6A, C).

Overall, these findings validated the mRNA data obtained in this study and suggest the role of SEMA3D and PDE1A as key genes involved in the dysregulation of SMC apoptosis and proliferation, as well as in orchestrating the abnormality of the enteric ANS, particularly in the pathogenesis of CD.

4 Discussion

This study revealed a widespread distinguishable dysregulation of mRNA expression between CD and UC in the colon inflammatory region. The regulation of SMC apoptosis and proliferation was significantly enriched in iCD, rather than in iUC. The involved PDE1A gene was upregulated 4-fold and 1.5-fold in iCD, as assessed by qPCR and IHC, respectively. Moreover, iCD was significantly associated with gene sets of ANS abnormality, while SEMA3D gene was upregulated 8-fold and 5-fold, respectively, compared to iUC.

In previous studies, the phenomenon of smooth muscle hyperplasia/hypertrophy in CD has been described briefly

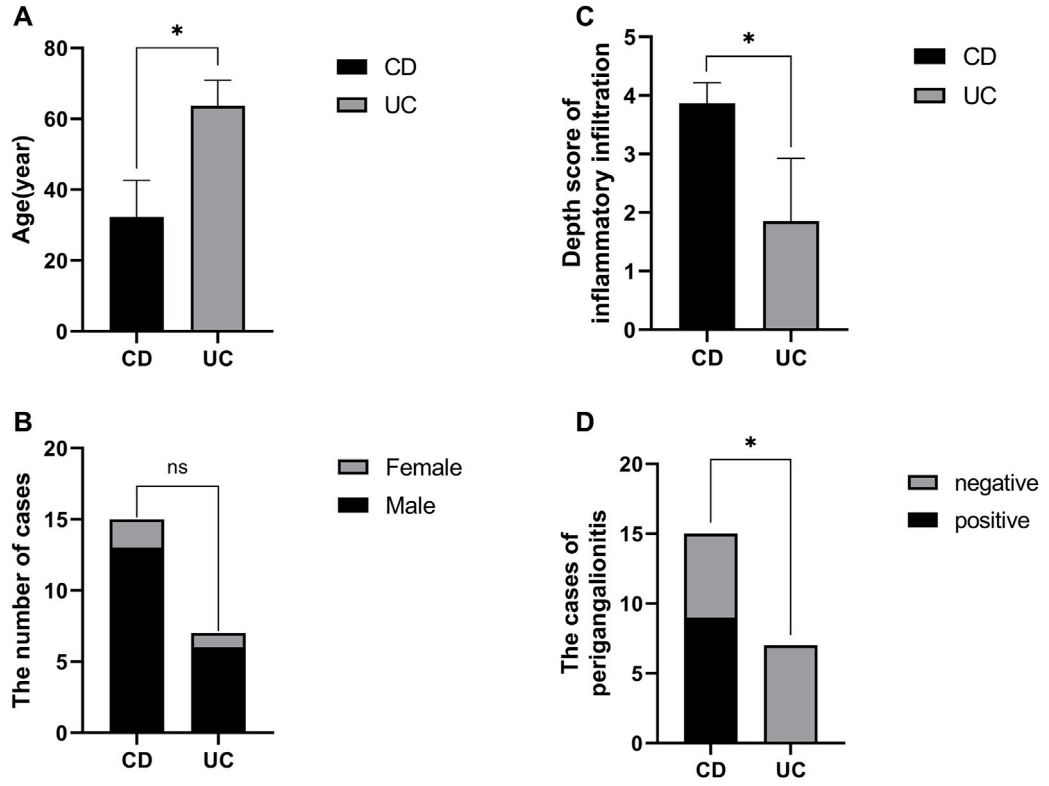


FIGURE 5
Baseline information of specimen sections prepared with non-inflamed and inflamed colon from CD and UC patients. **(A)** Comparison of disease onset ages between CD and UC patients. **(B)** Gender distribution in CD and UC patients. **(C)** Variation in depth scores of inflammatory infiltration observed in CD and UC patients. **(D)** Proportion of CD patients exhibiting periganglionitis.

(Suekane et al., 2010; Flynn et al., 2011; Scirocco et al., 2013). A recent novel histological grading scheme study demonstrated that muscularization, including hypertrophy of the MP and smooth muscle hyperplasia of the SM, is the most prevalent histological change in CD (Chen et al., 2017), accompanied by neuronal hypertrophy in both myenteric (Auerbach's) plexuses and submucosal (Meissner's) plexuses (Chen et al., 2017).

In this study, CD was significantly associated with the biological processes of SMC apoptosis and proliferation regulation. Both mRNA and protein levels of PDE1A involved in enteric SMC apoptosis/proliferation balance increased significantly in iCD. Cyclic nucleotide phosphodiesterases (PDEs) are critical in the homeostasis of cyclic nucleotides that regulate SMC growth by hydrolyzing cAMP or cGMP. Previous findings (Nagel et al., 2006) suggested that cytoplasmic PDE1A is associated with the "contractile" phenotype, whereas nuclear PDE1A is associated with the "synthetic" phenotype. Decreasing the levels of nuclear PDE1A via RNA interference or pharmacological inhibition significantly attenuated SMC growth by reducing proliferation via G1 arrest, induced apoptosis, elevated intracellular cGMP level, and altered gene expression, which was consistent with growth arrest and apoptosis (Nagel et al., 2006). Conversely, cytoplasmic PDE1A regulates myosin light chain phosphorylation with little effect on apoptosis (Nagel et al., 2006). In another study (Rajagopal et al., 2015), PDE1A expression was induced and accompanied by an increase in PDE1A activity in muscle cells isolated from muscle strips cultured with IL-1 β (interleukin-1 beta) or TNF- α (tumor necrosis factor alpha) or obtained from the colon of TNBS (2,4,6-trinitrobenzene sulfonic acid)-treated mice. Also, nitric oxide-induced muscle relaxation was inhibited in longitudinal muscle cells. This inhibition was completely reversed by the combination of both 1400 W dihydrochloride and vinpocetine (a PDE1 inhibitor) (Rajagopal et al., 2015). The inhibition of smooth muscle relaxation during inflammation reflected the combined effects of decreased sGC activity via S-nitrosylation and increased cGMP hydrolysis via PDE1 expression, thereby indicating that PDE1A might be a novel target for relieving altered pathogenesis of enteric smooth muscle in CD (Rajagopal et al., 2015).

Previously, 9 patients with both Hirschsprung disease (HSCR, also called congenital aganglionic megacolon) and IBD were described, suggesting an association between the two conditions (Sherman et al., 1989). HSCR is a neurocristopathy caused by a failure of the ENS progenitors derived from neural crest cells (NCCs) to migrate, proliferate, differentiate, or survive on and within the gastrointestinal tract, resulting in aganglionosis in the colon. This association has been confirmed in a few case reports and small case series (Levin et al., 2012; Kim and Kim, 2017). A recent population-based cohort study showed that individuals with HSCR had a 5-fold higher risk for IBD than those without HSCR (Löf Granström et al., 2018). Also, a follow-up study (Granström et al., 2021) found that the extent of aganglionosis is related to the risk of IBD. This theory was also proposed in a meta-analysis (Nakamura

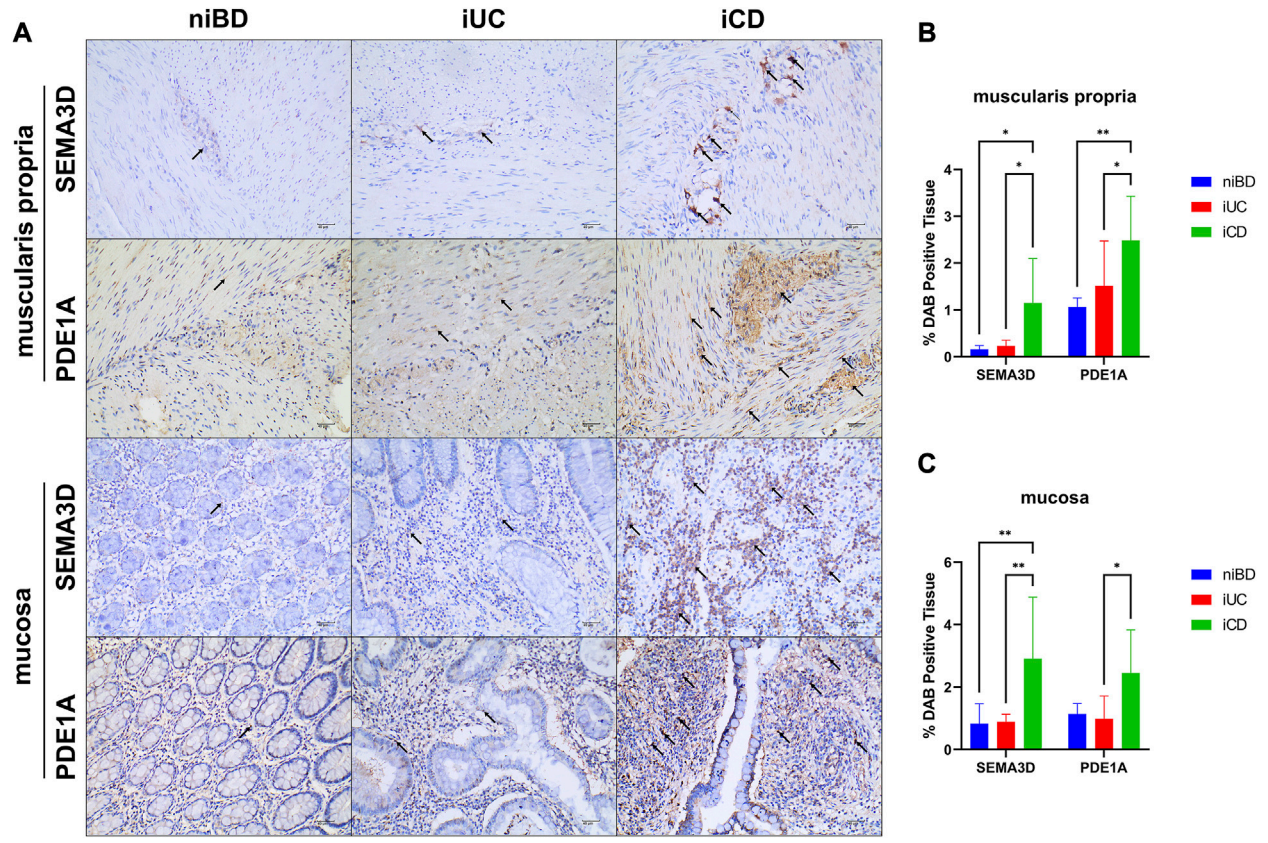


FIGURE 6
Abundant expression of SEMA3D and PDE1A proteins in layers of muscularis propria and mucosa in inflamed CD specimens. **(A)** Immunohistochemistry (IHC) images showing representative colon specimens from patients with CD and UC, including non-inflammatory and inflammatory samples. Staining demonstrates expression of SEMA3D and PDE1A at a magnification of ×20. The scale bar corresponds to 40 μm. **(B,C)** Graphs presenting the quantification of staining (percentage of DAB-positive tissue) in niBD, iUC, and iCD cohorts (niBD, n = 4; iUC, n = 7; iCD, n = 15). Statistical analysis was performed using Fisher's LSD test. Nonsignificant *p*-values (>0.05) are denoted. Asterisks (*) indicate statistical significance levels: $*p \leq 0.05$, $**p \leq 0.01$.

et al., 2018), including 14 studies encompassing a total of 66 patients with HSCR associated with IBD; moreover, the distribution of IBD is in 72.3% of CD patients. Another population-based cohort study (Bernstein et al., 2021) showed that individuals diagnosed with HSCR resulted in a 12-fold increased risk of subsequently diagnosed IBD. Interestingly, IBD can emerge in >2% of patients with HSCR and is more frequently classified as CD rather than UC (Bernstein et al., 2021).

In the present study, CD was significantly associated with gene sets of abnormality of ANS and aganglionic megacolon, indicating that the abnormality of ANS/ENS, such as gangliopathy may show pathogenic activity in CD similar to that in HSCR. Both the mRNA and protein levels of SEMA3D involved in the abnormality of ANS and aganglionic megacolon increased significantly in iCD. SEMA3D encodes a member of the semaphorin III family of secreted signaling proteins involved in axon guidance during neuronal development and is one of the three signaling pathways of HSCR pathogenesis (Luzón-Toro et al., 2013; Jiang et al., 2015; Kapoor et al., 2015); the other two are RET and EDNRB signaling pathways (Amiel et al., 2008; Tilghman et al., 2019).

SEMA3D has been implicated in the development of HSCR and contributes to risk in European (Luzón-Toro et al., 2013; Jiang et al., 2015; Kapoor et al., 2015) and Asian ancestries (Wang et al., 2011; Li et al., 2017; Gunadi et al., 2020). In a previous study (Luzón-Toro et al., 2013), the E198K-SEMA3D, A131T-SEMA3A, and S598G-SEMA3A mutations presented an increased protein level in the smooth muscle layer of ganglionic segments. Moreover, A131T-SEMA3A also maintained high protein levels in the aganglionic muscle layers. The coincident upregulation of SEMA3A expression in aganglionic colons was detected in Chinese patients of HSCR: the circular muscle layer, the submucosa, and the longitudinal muscles layer (Wang et al., 2011). These findings indicated that the SEMA3 variants increase the SEMA3 proteins levels in the HSCR colon tissue, thus supporting the functional implication of SEMA3s as a signaling molecule to influence the phenotype of HSCR patient. Thus, SEMA3D involvement of ANS/ENS abnormality may be a common pathogenesis mechanism in CD and HSCR.

However, it is important to acknowledge several limitations in our study. Firstly, future investigations should consider using a larger sample size to enhance the statistical power of our analysis. While the presence of variation in clinical and demographic characteristics may have constrained our analysis, it is worth noting that the observed alterations in RNA expression most

likely arise from the underlying disease pathophysiology. This inference is supported by the fact that most of the variations in the clinical and demographic characteristics of the specimens were not statistically significant, except for onset age, lesion location, and inflammatory infiltration depth, which have traditionally been considered disease phenotypic features. To evaluate the potential influence of colonic location on RNA expression levels, we compared the expressions of PDE1A, SEMA3D, and SLC18A2 between non-inflamed whole-wall cecum tissues (n = 6) and non-inflamed whole-wall transverse (n = 6) and descending (n = 6) colonic tissues, as depicted in Supplementary Figure S3. Our analysis did not reveal any significant differences in RNA expressions among the different colonic locations. Therefore, it could be cautiously inferred that the disparities in PDE1A, SEMA3D, and SLC18A2 expression levels among iCD, iUC, and niBD may reflect the inflammation status or disease phenotypic features rather than the anatomical location. Secondly, it is important to note that our study samples consisted exclusively of individuals of Chinese ethnicity. Consequently, future investigations should aim to explore the genetic backgrounds of different ethnic groups to obtain a more comprehensive understanding. Thirdly, although our transcriptome profile suggests abnormalities in enteric autonomic nervous system (ANS) and dysregulation of enteric smooth muscle cell (SMC) apoptosis/proliferation in the inflamed colon of CD, further research is necessary to determine whether these biological processes are secondary to the "inflammation-smooth muscle hyperplasia axis," analogous to chronic asthma, or if they involve independent pathways.

Conclusively, this study highlights the presence of ANS abnormality and dysregulation of SMC apoptosis/proliferation in the pathogenesis of CD. The identified genes, including SEMA3D and PDE1A, may serve as potential diagnostic biomarkers for differentiating between CD and UC, as well as therapeutic targets for restoring enteric dysautonomia and SMC proliferative disorders in CD. Future diagnostic and therapeutic strategies could be designed based on the dysregulation of enteric SMC apoptosis and proliferation, as well as enteric dysautonomia.

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/, GSE227747.

Ethics statement

The studies involving humans were approved by Ethics Committee of West China Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

YY contributed to study design, data analysis and interpretation, and drafting and revising the article critically for intellectual content. LX performed data analysis and interpretation. WY revised article critically for intellectual content. ZW and WM performed surgery and specimen collection. MZ did demographics and clinical data acquisition for transcriptome. QM did clinical data acquisition for histology. JG performed IHC staining and data acquisition. JW performed qPCR and data acquisition. YS and XW equally contributed to conception, final approval of the version to be submitted, and agreement to be accountable for all aspects of the work, thereby ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Science and Technology Department of Sichuan Province (2019YFS0261).

Acknowledgments

YY thanks the colleagues from Department of General Surgery, WCH. A special thanks to Zhaohui Yang, a valued researcher, and motivated the medical career of YY and interest in research on enteric neuromuscular pathology.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1194882/full#supplementary-material

References

Abraham, C., and Cho, J. H. (2009). Inflammatory bowel disease. *N. Engl. J. Med.* 361 (21), 2066–2078. doi:10.1056/NEJMra0804647

Ahrens, R., Waddell, A., Seidu, L., Blanchard, C., Carey, R., Forbes, E., et al. (2008). Intestinal macrophage/epithelial cell-derived CCL11/eotaxin-1 mediates eosinophil recruitment and function in pediatric ulcerative colitis. *J. Immunol.* 181 (10), 7390–7399. doi:10.4049/jimmunol.181.10.7390

Amiel, J., Sproat-Emison, E., Garcia-Barcelo, M., Lantieri, F., Burzynski, G., Borrego, S., et al. (2008). Hirschsprung disease, associated syndromes and genetics: A review. *J. Med. Genet.* 45 (1), 1–14. doi:10.1136/jmg.2007.053959

Anders, S., and Huber, W. (2012). *Differential expression of RNA-Seq data at the gene level – the DESeq package*.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31 (2), 166–169. doi:10.1093/bioinformatics/btu638

Arijs, I., De Hertogh, G., Lemaire, K., Quintens, R., Van Lommel, L., Van Steen, K., et al. (2009). Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *PLoS One* 4 (11), e7984. doi:10.1371/journal.pone.0007984

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. The gene ontology consortium. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Assadsangabi, A., Evans, C. A., Corfe, B. M., and Lobo, A. (2019). Application of proteomics to inflammatory bowel disease research: current status and future perspectives. *Gastroenterology Res. Pract.* 2019, 1426954. doi:10.1155/2019/1426954

Bernstein, C. N., Kuenzig, M. E., Coward, S., Nugent, Z., Nasr, A., El-Matary, W., et al. (2021). Increased incidence of inflammatory bowel disease after Hirschsprung disease: A population-based cohort study. *J. Pediatr.* 233, 98–104.e2. doi:10.1016/j.jpeds.2021.01.060

Bjerrum, J. T., Hansen, M., Olsen, J., and Nielsen, O. H. (2010). Genome-wide gene expression analysis of mucosal colonic biopsies and isolated colonocytes suggests a continuous inflammatory state in the lamina propria of patients with quiescent ulcerative colitis. *Inflamm. Bowel Dis.* 16 (6), 999–1007. doi:10.1002/ibd.21142

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi:10.1093/bioinformatics/btu170

Carey, R., Jurickova, I., Ballard, E., Bonkowski, E., Han, X., Xu, H., et al. (2008). Activation of an IL-6:STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflamm. Bowel Dis.* 14 (4), 446–457. doi:10.1002/ibd.20342

Chen, W., Lu, C., Hirota, C., Iacucci, M., Ghosh, S., and Gui, X. (2017). Smooth muscle hyperplasia/hypertrophy is the most prominent histological change in Crohn's fibrostenosing bowel strictures: A semiquantitative analysis by using a novel histological grading scheme. *J. Crohns Colitis* 11 (1), 92–104. doi:10.1093/ecco-jcc/jjw126

Consortium, G. O. (2021). The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 49 (D1), D325–d334. doi:10.1093/nar/gkaa1113

Flynn, R. S., Mahavadi, S., Murthy, K. S., Grider, J. R., Kellum, J. M., Akbari, H., et al. (2011). Endogenous IGFBP-3 regulates excess collagen expression in intestinal smooth muscle cells of Crohn's disease strictures. *Inflamm. Bowel Dis.* 17 (1), 193–201. doi:10.1002/ibd.21351

Galamb, O., Györffy, B., Sipos, F., Spisák, S., Németh, A. M., Miheller, P., et al. (2008a). Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature. *Dis. Markers* 25 (1), 1–16. doi:10.1155/2008/586721

Galamb, O., Sipos, F., Solymosi, N., Spisák, S., Krenács, T., Tóth, K., et al. (2008b). Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results. *Cancer Epidemiol. Biomarkers Prev.* 17 (10), 2835–2845. doi:10.1158/1055-9965.EPI-08-0231

Granlund, A., Flatberg, A., Østvik, A. E., Drozdov, I., Gustafsson, B. I., Kidd, M., et al. (2013). Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between Crohn's disease and ulcerative colitis. *PLoS One* 8 (2), e56818. doi:10.1371/journal.pone.0056818

Granström, A. L., Ludvigsson, J. F., and Wester, T. (2021). Clinical characteristics and validation of diagnosis in individuals with Hirschsprung disease and inflammatory bowel disease. *J. Pediatr. Surg.* 56 (10), 1799–1802. doi:10.1016/j.jpedsurg.2020.11.015

Gunadi, Kalim, A. S., Budi, N. Y. P., Hafiq, H. M., Maharani, A., Febrianti, M., et al. (2020). Aberrant expressions and variant screening of SEMA3D in Indonesian Hirschsprung patients. *Front. Pediatr.* 8, 60. doi:10.3389/fped.2020.00060

Jiang, Q., Arnold, S., Heanue, T., Kilambi, K. P., Doan, B., Kapoor, A., et al. (2015). Functional loss of semaphorin 3C and/or semaphorin 3D and their epistatic interaction with ret are critical to Hirschsprung disease liability. *Am. J. Hum. Genet.* 96 (4), 581–596. doi:10.1016/j.ajhg.2015.02.014

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi:10.1093/nar/gkw1092

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360. doi:10.1093/nar/gkp896

Kapoor, A., Jiang, Q., Chatterjee, S., Chakraborty, P., Sosa, M. X., Berrios, C., et al. (2015). Population variation in total genetic risk of Hirschsprung disease from common RET, SEMA3 and NRG1 susceptibility polymorphisms. *Hum. Mol. Genet.* 24 (10), 2997–3003. doi:10.1093/hmg/ddv051

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: A fast spliced aligner with low memory requirements. *Nat. Methods* 12 (4), 357–360. doi:10.1038/nmeth.3317

Kim, H. Y., and Kim, T. W. (2017). Crohn's disease with ankylosing spondylitis in an adolescent patient who had undergone long ileo-colonic anastomosis for Hirschsprung's disease as an infant. *Intest. Res.* 15 (1), 133–137. doi:10.5217/ir.2017.15.1.133

Kugathasan, S., Baldassano, R. N., Bradfield, J. P., Sleiman, P. M., Imielinski, M., Guthery, S. L., et al. (2008). Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* 40 (10), 1211–1215. doi:10.1038/ng.203

Levin, D. N., Marcon, M. A., Rintala, R. J., Jacobson, D., and Langer, J. C. (2012). Inflammatory bowel disease manifesting after surgical treatment for Hirschsprung disease. *J. Pediatr. Gastroenterol. Nutr.* 55 (3), 272–277. doi:10.1097/MPG.0b013e31824f617a

Li, Q., Zhang, Z., Diao, M., Gan, L., Cheng, W., Xiao, P., et al. (2017). Cumulative risk impact of RET, SEMA3, and NRG1 polymorphisms associated with Hirschsprung disease in han Chinese. *J. Pediatr. Gastroenterol. Nutr.* 64 (3), 385–390. doi:10.1097/MPG.0000000000001263

Löf Granström, A., Amin, L., Arnell, H., and Wester, T. (2018). Increased risk of inflammatory bowel disease in a population-based cohort study of patients with Hirschsprung disease. *J. Pediatr. Gastroenterol. Nutr.* 66 (3), 398–401. doi:10.1097/MPG.0000000000001732

Luzón-Toro, B., Fernández, R. M., Torroglosa, A., de Agustín, J. C., Méndez-Vidal, C., Segura, D. I., et al. (2013). Mutational spectrum of semaphorin 3A and semaphorin 3D genes in Spanish Hirschsprung patients. *PLoS One* 8 (1), e54800. doi:10.1371/journal.pone.0054800

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34 (3), 267–273. doi:10.1038/ng1180

Nagel, D. J., Aizawa, T., Jeon, K. I., Liu, W., Mohan, A., Wei, H., et al. (2006). Role of nuclear Ca2+/calmodulin-stimulated phosphodiesterase 1A in vascular smooth muscle cell growth and survival. *Circ. Res.* 98 (6), 777–784. doi:10.1161/01.RES.0000215576.27615.fd

Nakamura, H., Lim, T., and Puri, P. (2018). Inflammatory bowel disease in patients with hirschsprung's disease: A systematic review and meta-analysis. *Pediatr. Surg. Int.* 34 (2), 149–154. doi:10.1007/s00383-017-4182-4

Noble, C. L., Abbas, A. R., Cornelius, J., Lees, C. W., Ho, G. T., Toy, K., et al. (2008). Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut* 57 (10), 1398–1405. doi:10.1136/gut.2008.148395

Olsen, J., Gerds, T. A., Seidelin, J. B., Csillag, C., Bjerrum, J. T., Troelsen, J. T., et al. (2009). Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflamm. Bowel Dis.* 15 (7), 1032–1038. doi:10.1002/ibd.20879

Rajagopal, S., Nalli, A. D., Kumar, D. P., Bhattacharya, S., Hu, W., Mahavadi, S., et al. (2015). Cytokine-induced S-nitrosylation of soluble guanylyl cyclase and expression of phosphodiesterase 1A contribute to dysfunction of longitudinal smooth muscle relaxation. *J. Pharmacol. Exp. Ther.* 352 (3), 509–518. doi:10.1124/jpet.114.221929

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12 (3), R22. doi:10.1186/gb-2011-12-3-r22

Scirocco, A., Rosati, S., Sferra, R., Vetusti, A., Pallotta, N., Tellan, G., et al. (2013). P004 Smooth muscle cells participate in Crohn's disease intestinal fibrosis. *J. Crohn's Colitis* 7 (1), S12. doi:10.1016/s1873-9946(13)60027-6

Shen, L., and Weber, C. R. (2017). "Pathological diagnosis of inflammatory bowel disease," in *Inflammatory bowel disease: Diagnosis and therapeutics*. Editor R. D. Cohen (Cham: Springer International Publishing), 121–136.

Sherman, J. O., Snyder, M. E., Weitzman, J. J., Jona, J. Z., Gillis, D. A., O'Donnell, B., et al. (1989). A 40-year multinational retrospective study of 880 Swenson procedures. *J. Pediatr. Surg.* 24 (8), 833–838. doi:10.1016/s0022-3468(89)80548-2

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. U. S. A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Suekane, T., Ikura, Y., Watanabe, K., Arimoto, J., Iwasa, Y., Sugama, Y., et al. (2010). Phenotypic change and accumulation of smooth muscle cells in strictures in Crohn's disease: relevance to local angiotensin II system. *J. Gastroenterol.* 45 (8), 821–830. doi:10.1007/s00535-010-0232-6

Tilghman, J. M., Ling, A. Y., Turner, T. N., Sosa, M. X., Krumm, N., Chatterjee, S., et al. (2019). Molecular genetic anatomy and risk profile of hirschsprung's disease. *N. Engl. J. Med.* 380 (15), 1421–1432. doi:10.1056/NEJMoa1706594

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28 (5), 511–515. doi:10.1038/nbt.1621

van Beelen Granlund, A., Østvik, A. E., Brenna, Ø., Torp, S. H., Gustafsson, B. I., and Sandvik, A. K. (2013). REG gene expression in inflamed and healthy colon mucosa explored by *in situ* hybridisation. *Cell tissue Res.* 352 (3), 639–646. doi:10.1007/s00441-013-1592-z

Vatn, S. S., Lindstrøm, J. C., Moen, A. E. F., Brackmann, S., Tannæs, T. M., Olbjørn, C., et al. (2022). Mucosal gene transcript signatures in treatment naïve inflammatory bowel disease: A comparative analysis of disease to symptomatic and healthy controls in the European IBD-character cohort. *Clin. Exp. Gastroenterol.* 15, 5–25. doi:10.2147/CEG.S343468

Wang, L. L., Fan, Y., Zhou, F. H., Li, H., Zhang, Y., Miao, J. N., et al. (2011). Semaphorin 3A expression in the colon of Hirschsprung disease. *Birth Defects Res. A Clin. Mol. Teratol.* 91 (9), 842–847. doi:10.1002/bdra.20837

Wu, F., Dassopoulos, T., Cope, L., Maitra, A., Brant, S. R., Harris, M. L., et al. (2007). Genome-wide gene expression differences in Crohn's disease and ulcerative colitis from endoscopic pinch biopsies: insights into distinctive pathogenesis. *Inflamm. Bowel Dis.* 13 (7), 807–821. doi:10.1002/ibd.20110

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 19 October 2023
DOI 10.3389/fgene.2023.1270175

OPEN ACCESS

EDITED BY
Hua Li,
Shanghai Jiao Tong University, China

REVIEWED BY
Abhijeet Botre,
KEM Hospital Research Centre, India
Yoichi Araki,
Johns Hopkins University, United States

*CORRESPONDENCE
Ping Sun,
sp77223@163.com

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

RECEIVED 03 August 2023
ACCEPTED 09 October 2023
PUBLISHED 19 October 2023

CITATION
Li B, Wang Y, Hou D, Song Z, Zhang L, Li N, Yang R and Sun P (2023), Identification and functional characterization of *de novo* variant in the *SYNGAP1* gene causing intellectual disability.
Front. Genet. 14:1270175.
doi: 10.3389/fgene.2023.1270175

Identification and functional characterization of *de novo* variant in the *SYNGAP1* gene causing intellectual disability

Boxuan Li[1†], Yu Wang[1†], Dong Hou[2,3], Zhen Song[1], Lihua Zhang[1], Na Li[1], Ruifang Yang[1] and Ping Sun[1*]

[1]Center of Prenatal Diagnosis, Department of Obstetrics and Gynecology, Qilu Hospital of Shandong University, Jinan, China, [2]Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Qilu Hospital of Shandong University, Jinan, China, [3]Suzhou Research Institute of Shandong University, Suzhou, China

Background: Intellectual disability (ID) is defined by cognitive and social adaptation defects. Variants in the *SYNGAP1* gene, which encodes the brain-specific cytoplasmic protein SYNGAP1, are commonly associated with ID. The aim of this study was to identify novel *SYNGAP1* gene variants in Chinese individuals with ID and evaluate the pathogenicity of the detected variants.

Methods: Whole exome sequencing (WES) was performed on 113 patients diagnosed with ID. In the study, two *de novo* variants in *SYNGAP1* were identified. Sanger sequencing was used to confirm these variants. Minigene assays were used to verify whether the *de novo* intronic variant in *SYNGAP1* influenced the normal splicing of mRNA.

Results: Two *de novo* heterozygous pathogenic variants in *SYNGAP1,* c.333del and c.664-2A>G, were identified in two ID patients separately. The c.333del variant has been reported previously as a *de novo* finding in a child with ID, while the c.664-2A>G variant was novel *de novo* intronic variant, which has not been reported in the literature. Functional studies showed that c.664-2A>G could cause aberrant splicing, resulting in exon 7 skipping and a 16bp deletion within exon 7.

Conclusion: We identified two *de novo* pathogenic heterozygous variants in *SYNGAP1* in two patients with ID, among which the c.664-2A>G variant was a novel *de novo* pathogenic variant. Our findings further enrich the variant spectrum of the *SYNGAP1* gene and provide a research basis for the genetic diagnosis of ID.

KEYWORDS

SYNGAP1, whole exome sequencing (WES), intellectual disability, minigene, variant

Introduction

Intellectual disability (ID) is characterized by cognitive and social adaptation defects, which typically occur before the age of 18 (Chelly et al., 2006). ID is the most common severe disability in children, affecting approximately 1%–3% of the population (Goldenberg and Saugier-Veber, 2010). ID is classified as either a syndromic or non-syndromic (NSID) form.

The majority of ID patients have the NSID form of the disease, which is mainly characterized by the lack of relevant morphological, radiological, and metabolic features (Hane et al., 1996). ID patients often need lifelong rehabilitation support treatment, causing substantial psychological and economic burdens for families and society. Determining the genetic and molecular basis of ID remains a significant challenge in neuroscience.

The *SYNGAP1* gene encodes the brain-specific RAS GTPase-activating protein SYNGAP1, which is an important component of the N-methyl-D-aspartate receptor (NMDA) complex and plays a pivotal role in neuronal synaptic development, structure, and plasticity (Clement et al., 2012). *De novo* variants of *SYNGAP1* are a common cause of NSID, autism spectrum disorders (ASD), and epilepsy (Berryer et al., 2013; Mignot et al., 2016). *De novo* nonsense variants in *SYNGAP1* cause haploinsufficiency, resulting in a neurodevelopmental disorder known as intellectual developmental disorder (OMIM #612621), with phenotypes including ID, motor disorders, and epilepsy. The effects of these variants demonstrate the importance of SYNGAP1 in developing the nervous system and brain (Agarwal et al., 2019). Currently, approximately 0.7%–1% of ID cases are caused by *SYNGAP1* variants (Mignot et al., 2016).

In this study, we identified two *de novo* heterozygous pathogenic variants in the *SYNGAP1,* c. 333del and c.664-2A>G, in two patients with ID, among which the c.664-2A>G variant was a novel *de novo* pathogenic variant. The patients exhibited generalized developmental delay, motor retardation, hypotonia, and severe language impairment. To assess the impact of c.664-2A>G variant on splicing, we performed a minigene splicing assay. We found that the c.664-2A>G variant causes aberrant splicing, which would result in impaired function of the SYNGAP1 protein and consequently contribute to the occurrence of ID.

Methods and materials

Subjects

Two female patients from unrelated families were diagnosed with intellectual developmental disorder from a clinical cohort of 113 cases with ID from between January 2019 and January 2023 at Qilu Hospital of Shandong University. All the patients were diagnosed by experienced experts of the hospital according to the DSM-5 criteria. Among these patients with ID, 25 cases were combined with epilepsy and 49 cases were combined with other structural anomalies. The age of the children ranged from 12 months to 18 years, with a median age of 8 years. The etiology of these patients was unknown. All of these patients underwent karyotyping and chromosome microarray (CMA) analysis, and these results were inconclusive, following these samples were processed for WES. We collected peripheral blood and clinical information from their families. The families accepted the inheritance consultation and signed the informed consent form before the genetic test. This study was authorized by the Ethics Committee of Qilu Hospital of Shandong University.

DNA extraction and whole exome sequencing (WES)

The genomic DNA for sequencing was obtained from peripheral blood. The extraction steps were conducted according to the instructions of the DNA extraction kit (Tiangen Biotech). WES was performed on the DNA from the affected individual and sequenced on NovaSeq 6000 platforms (Illumina) with 150 bp paired-end reads. Reads data were aligned with the GRCh37/hg19 human reference sequence. The single-nucleotide variants (SNVs) and other variants were called with the Genome Analysis Toolkit (GATK). The variants were annotated using Annovar software. During the annotation, several public databases such as Clinvar, gnomAD, PubMed, HGMD, dbNSFP, etc., were used. Variants with allele frequencies higher than 1% in any public databases (ExAC Browser and gnomAD) were excluded. *De novo* variants were analyzed from sequencing data by DeNovoGear software (Ramu et al., 2013). The candidate variants were confirmed in the patients with ID by Sanger sequencing.

Minigene assay

The *SYNGAP1* c.664-2A>G variant is located at the splice-acceptor site of exon 7. We obtained the *SYNGAP1* fragment [intron6 (192bp)-Exon7 (99bp)-intron7 (547bp)] with restriction sites (KpnI and BamHI) from human genomic DNA by nested PCR amplification and then cloned it into a pcMINI plasmid using nucleic acid endonuclease and DNA ligase. The pcMINI vector contain ExonA-IntronA-multiple cloning site-IntronB-ExonB (Bioeagle Biotech Company). Exon A and Exon B simulate exon 6 and exon 8, respectively. The pcMINI-*SYNGAP1*-MUT (c.664-2A>G) plasmids were produced using a QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent) with pcMINI-*SYNGAP1*-WT (wild-type) plasmid as the template. Both WT and mutant plasmids contained the whole sequence of exon 7 and a portion of the upstream and downstream intron sequences. The recombinant plasmids were transiently transfected into HEK293T and HeLa cells according to the transfection reagent instructions. After the transfected cells were cultured for 48 h, total RNA was extracted with Trizol (TaKaRa), and cDNA was acquired with HifairTM 1st Strand cDNA Synthesis SuperMix (TEASEN). The RT-PCR products was analyzed by electrophoresis on 2% agarose gels containing ethidium bromide and visualized by exposure to ultraviolet light. Each DNA band was purified by DNA Gel Exctration Kit (SIMGEN). Direct sequencing of purified RT-PCR products was performed with the Big Dye Terminator Cycler Sequencing Ready Reaction Kit (Applied Biosystems) on the ABI3730xl Genetic Analyzer (Applied Biosystems). Primers used for minigene assay of *SYNGAP1* were as follows: *SYNGAP1*-F1: 5′-AACTCCTGGGCTCAAGTGAC-3′; *SYNGAP1*-R1: 5′-TGGGTAAAGCTTGGCCAGAT-3′; *SYNGAP1*-F2:5′-AGCACTTTGGGAGGCTGAAT-3′; *SYNGAP1*-R2: 5′-GAGGTTGCAGTGAGCCAAGA-3′; MINI-*SYNGAP1*-KpnI-F: 5′-GGTAGGTACCCTGGGGAGGGCCAAAGGACA-3′; MINI-*SYNGAP1*-BamHI-R: 5′-TAGTGGATCCGAGAATAGCTGACAGAACTG-3′; *SYNGAP1*-c.664-2A>G-F: 5′-TCCACACTCCTTTCTGGGTAACAACTTCATC-3′; *SYNGAP1*-c.664-2A>G-R: 5′-GATGAAGTTGTTACCCAGAAAGGAGTGTGGA -3′.

10.3389/fgene.2023.1270175

TABLE 1 Variations of *SYNGAP1* genes identified in two ID patients and their clinical characteristics.

Patient	Age (m)	IQ	Language delay	Motor delay	Age of walking (m)	Current speech ability	Feeding difficulty	ASD	Seizures	Gait	EEG	MRI	Karyotype	Mutation type	Source of mutation	Nuclotide change	Amino acid change	ACMG	Pathogenicity
Patient 1	67	45	Yes	Yes	16	Two-three words	No	No	No	Unsteady gait	Normal	Normal	46,XX	Frameshift	*De novo*	c.333del	p.Lys114SerfsX20	PVS1+PS1+PS2+PP5	Pathogenic
Patient 2	84	55	Yes	Yes	36	Simple sentences	No	No	No	Wide-based gait	Normal	Normal	46,XX	Splicing	*De novo*	c.664-2A>G	-	PVS1+PS2+PM2	Pathogenic

Note: –, absent; m, month(s); IQ, intelligence quotient; ASD, autism spectrum disorder; EEG, electroencephalogram; MRI, magnetic resonance imaging.

Results

De novo heterozygous variants were identified

Two *de novo* heterozygous pathogenic variants in *SYNGAP1*, NM_006772.3: c. 333del and c.664-2A>G, were identified in two ID patients separately, as detailed in Table 1. The c.333del variant has been reported previously as a *de novo* finding in a child with ID (Carvill et al., 2013). The c.333del variant is predicted to cause loss of normal protein function either through protein truncation or nonsense-mediated mRNA decay. Conversely, the c.664-2A>G variant was a novel *de novo* intronic variant identified in a child with ID, which has not been reported in the literature and databases (ClinVar, DVD, PubMed, HGMD, etc.) (Figure 1).

Clinical characteristics of the NSID affected individual

Both patients in their respective families, carrying *de novo* variants in *SYNGAP1*, exhibit delays in intellectual and motor developmental. These two patients had impaired speaking ability and were speech disabled, along with symptoms such as hypotonia, muscle flaccidity, and a wide-based/unsteady gait. Notably, they didn't exhibit feeding difficulty, autism spectrum disorder (ASD), epilepsy, or microcephaly. We conducted a comprehensive assessment for ASD on these two patients, which included psychological tests, clinical examinations, and consideration of their family medical history. We utilized standardized assessment tools such as the Autism Diagnostic Observation Schedule (ADOS) for ASD evaluation and found no symptoms of ASD in these patients. Additionally, we conducted initial psychological tests on both patients to assess cognitive functioning and identify any coexisting conditions, and the results indicate that both of these patients don't exhibit symptoms of autism. Both patients had normal electroencephalograms (EEGs) and normal karyotypes (Table 1). There is no family history related to developmental disorders. The two patients are girls and are the only children in their respective families. One patient is 5 years old and the other is 7 years old, and their parents are healthy. They have experienced an overall delay in developmental milestones. For example, the patient which carrying the c.664-2A>G variant was independently sitting and walking later than children of the same age. She didn't achieve independent walking until the age of 3 years, and her gait was extensive and unstable. At the age of 7, her intelligence quotient (IQ) was measured at 55 on the Tanaka-Binet IQ Scale V. Furthermore, she exhibits delayed language development and only uses short and simple sentences with limited vocabulary.

Expression of *SYNGAP1* mRNA in transfected cells with recombinant plasmids

The c.333del variant has been reported previously as a *de novo* finding, so we did not perform functional experiments on it. We performed *in vitro* experiments on the novel *de novo* splicing variant in *SYNGAP* gene. To investigate the influence of the c.664-2A>G

10.3389/fgene.2023.1270175

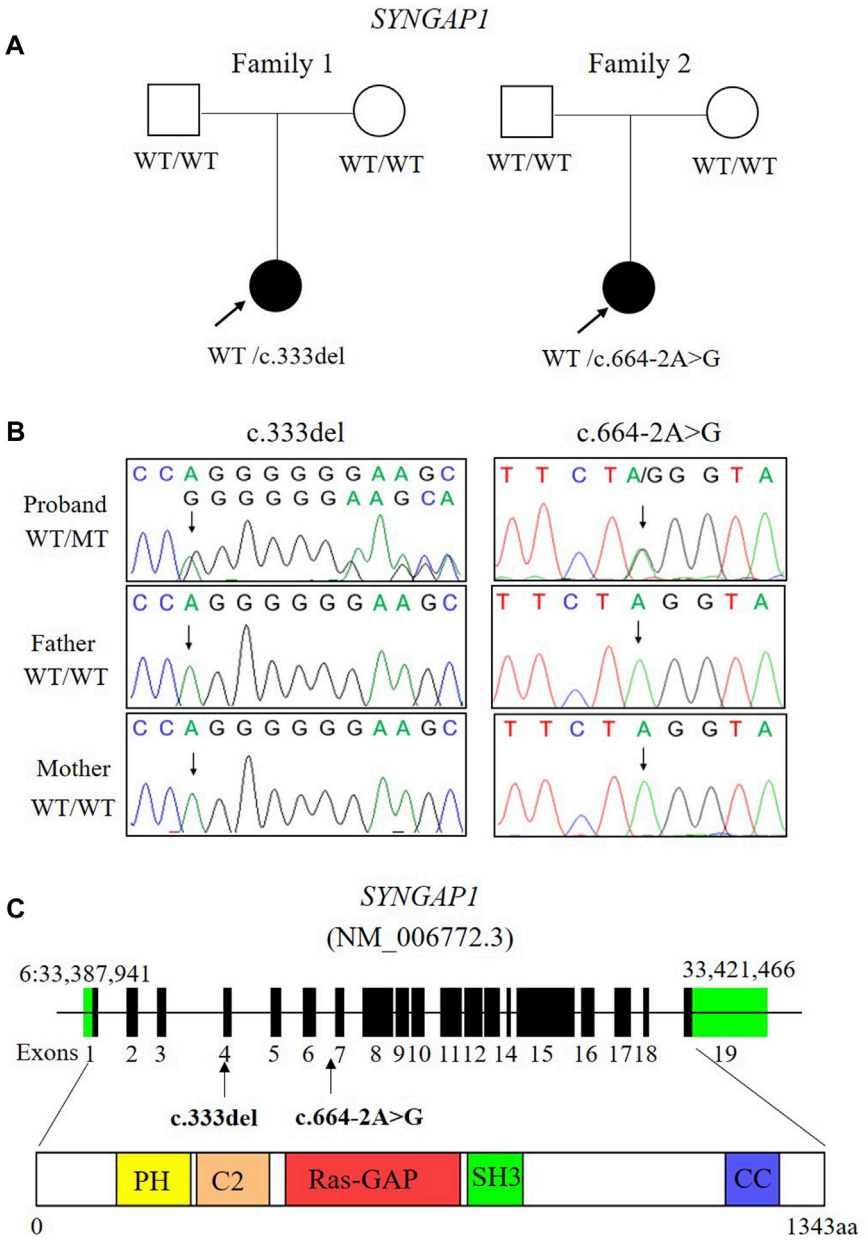


FIGURE 1
Two *de novo* variants of *SYNGAP1* were identified in two patients with ID. **(A)** Families pedigree and genotype are shown. The probands with ID underwent WES. Filled symbols represent affected individuals. **(B)** Sanger sequencing chromatograms of the *SYNGAP1* variants in these families. **(C)** Localization of the *SYNGAP1:* c. 333del and c.664-2A>G variant found in the study. The amino acid (aa) positions are referenced to RefSeq number NM_006772.3 (isoform-1: 1343 aa). Various predicted SYNGAP1 domains are showed: PH, pleckstrin homology domain (amino acid positions 150–251), C2 domain (amino acid positions 263–362), Ras-GAP (amino acid positions 392–729), SH3 (amino acid positions 785–815), coiled coil (CC; amino acid positions 1189–1262).

variant on splicing, we conducted a minigene splicing assay. The pcMINI-*SYNGAP1*-WT and pcMINI-*SYNGAP1*-MUT (c.664-2A>G) plasmids were transiently transfected into 293T and HeLa cells (Figure 2A). The total RNA was extracted and reverse-transcribed into cDNA after transfection over 48 h. The cDNA was amplified by PCR and analyzed by agarose gel electrophoresis. Agarose gel electrophoresis showed that the mutant-type (MUT) had two bands, all bands were smaller than the WT band, and their migrations were relatively faster (Figure 2B). DNA sequencing results showed that the WT minigene (pcMINI-*SYNGAP1*-WT) transcribed normal mRNA composed of exon 7 (Figure 2C, band a), while the c.664-2A>G mutant minigene caused abnormal splicing, resulting in exon 7 skipping (Figures 2C, D, band b) and a 16 bp deletion within exon 7 (Figures 2C, D, band c), which reveals that this variant may be a crucial mechanism for the pathogenesis of ID. Exon 7 skipping could result in the loss of 33 amino acids (c.664_762del p.Val222_Lys254del), and the deletion of 16bp in exon 7 could result in a frameshift of amino acids at position 222 and a premature stop codon (c.664_679del p.Val222Glufs*24). The aberrant splicing is predicted to abolish the pleckstrin homology

10.3389/fgene.2023.1270175

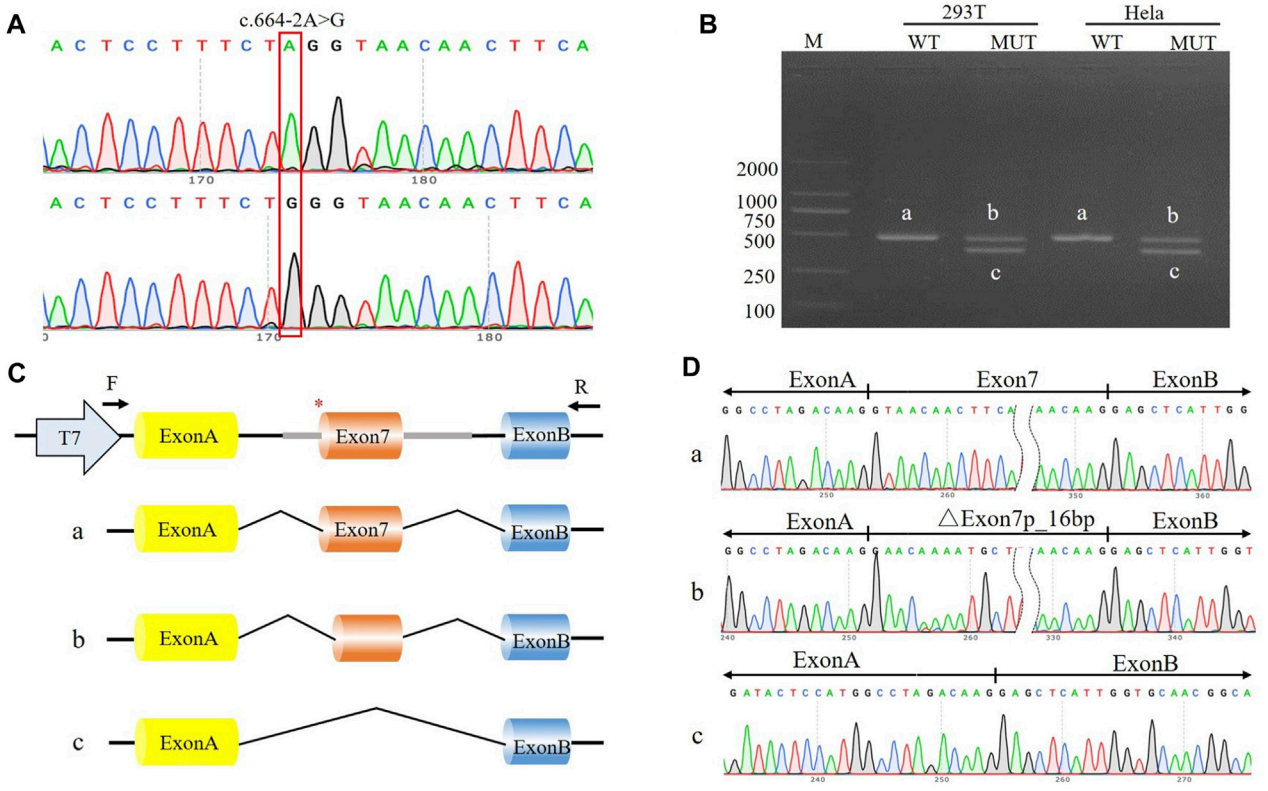


FIGURE 2
The effect of the c.664-2A>G variant on splicing was assessed through a minigene assay. **(A)** Construction of the pcMINI-*SYNGAP1*-WT/MUT vector, which contain exon 7 and flanking intronic sequences of WT or mutant type (c.664-2A>G) of the *SYNGAP1* gene. **(B)** Minigene assay performed in 293T and Hela cells transfected with the pcMINI-*SYNGAP1*-WT/MUT vector. The PCR products were isolated by gel electrophoresis. The *SYNGAP1* splicing products of wild-type (band a) and variant type (band b and c) are shown. **(C, D)** Schematic diagram of minigene construction and sanger sequencing of PCR products.

(PH, pos. 150–251aa) and C-terminal domains, suggesting that it would prevent the SYNGAP1 protein from performing its normal functions. *SYNGAP1* c.664-2A>G may damage cognitive and social adaptation development by impairing maturation of dendritic spine synapses in neurons (Figure 3).

Discussion

In this study, we identified two *de novo* heterozygous pathogenic variants in *SYNGAP1* among 113 patients with ID. The c.333del variant has been previously reported as a *de novo* finding in a child with ID (Carvill et al., 2013), while the other splicing variant c.664-2A>G has not been reported in any literature or databases. The c.333del variant is predicted to cause loss of normal protein function either through protein truncation or nonsense-mediated mRNA decay. The phenotypes of the ID patient, which carrying the c.664-2A>G variant, are similar to the previously published truncating variant in *SYNGAP1*. Although the intronic variants may have more deleterious effects than the exonic variants, they are underexplored (Kallel-Bouattour et al., 2017). To further evaluate the deleterious effects of the intronic variant c.664-2A>G, we conducted a minigene assay to investigate its impact on mRNA splicing. The minigene experiment results showed that the intronic variant c.664-2A>G causes aberrant splicing of *SYNGAP1*. The c.664-2A>G variant would result in exon 7 skipping and partial exon 7 deletion, which would abolish critical functional domains and impair the function of the SYNGAP1 protein.

There is a potential acceptor site located 16 bp upstream of exon 7 in *SYNGAP1*. After c.664-2A>G variant, this site is activated for splicing. The c.664-2A>G variant disrupts the original acceptor site, potentially leading to the recognition of the alternative splicing site that causes a 16 bp deletion on upstream of exon 7. Alternatively, the disruption of the original acceptor site might result in direct skipping of exon 7 during splicing, resulting in the overall deletion of exon 7, much like how a gene in a database might have multiple normal transcripts.

SYNGAP1 is an important gene that is necessary for neuronal development, and its dysfunction is associated with ID (Jeyabalan and Clement, 2016). *SYNGAP1* is located on human chromosome 6, contains 19 exons, and generates approximately a 6 kb transcript, which encodes a brain-specific synaptic Ras GTP-ase activating protein. The impairment of SYNGAP1 function may make patients with ID susceptible to seizures by increasing the recruitment of AMPA receptors at postsynaptic glutamatergic synapses, which leads to increased transmission of excitatory synapses (Hamdan et al., 2009). In large-scale studies, almost all *SYNGAP1* variants associated with NSID, ASD, and epilepsy are loss-of-function and lead to SYNGAP1 haploinsufficiency, resulting in intellectual developmental

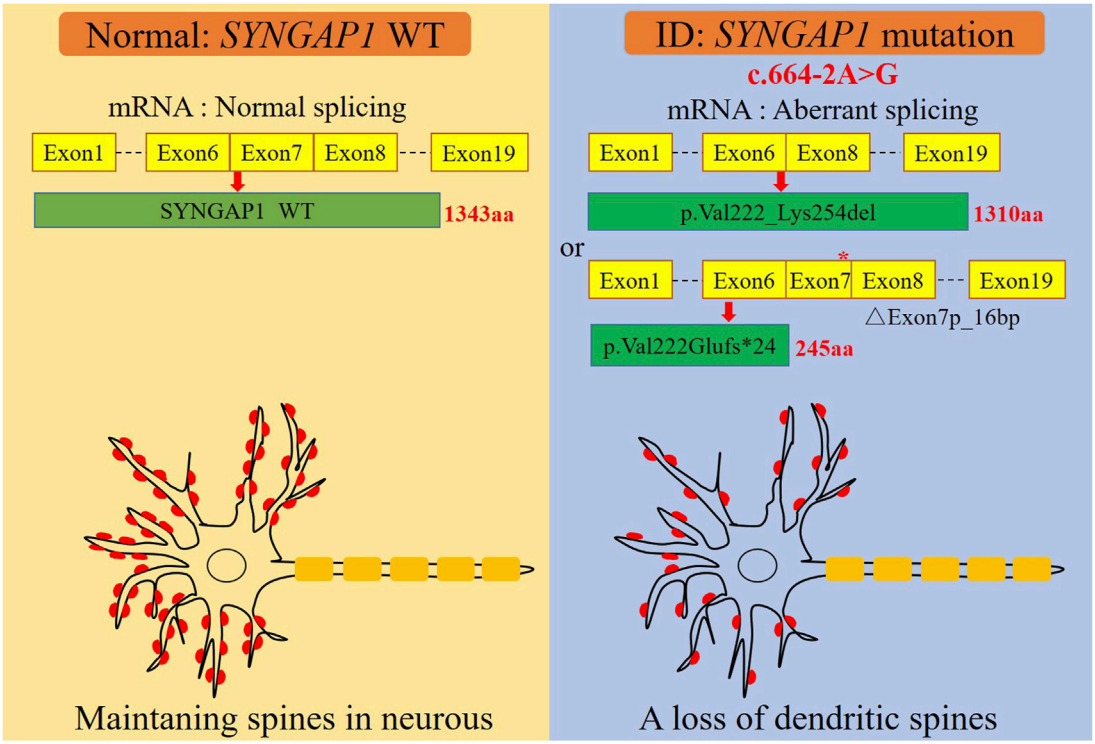


FIGURE 3
A graphical summary of the mechanism of dendritic spine loss caused by the *SYNGAP1* c.664-2A>G variant. The *SYNGAP1* c.664-2A>G variant causing aberrant splicing and dendritic spine loss.

disorder (Hamdan et al., 2011; Berryer et al., 2013; Carvill et al., 2013; Fieremans et al., 2016).In our study, as well as previous observations, suggest that the *SYNGAP1* c.664-2A>G variant would cause ID mainly through a mechanism of haploinsufficiency.

NSID poses a challenge for clinicians because of the absence of specific clinical features to guide them toward an etiological diagnosis. The identification of novel variants in known pathogenic genes or novel ID genes suggests that molecular diagnostic approaches are becoming increasingly significant in unraveling the underlying causes of this condition. In the present study, the two patients presented with comprehensive developmental delays, particularly motor milestones and language development, and exhibited behavioral disorders. We identified pathogenic variants in *SYNGAP1* in both of these patients by WES. Based on clinical and genetic features, the patients were diagnosed with intellectual developmental disorder. *De novo SYNGAP1* variants were initially reported to cause ID, accounting for approximately 0.62% of all the patients in the Deciphering Developmental Disorders (DDD) study (Hamdan et al., 2011; Wright et al., 2015). Six patients with *SYNGAP1* variants exhibited moderate to-severe ID due to severe language impairment (Hamdan et al., 2011). Studies involving rodent models with the deletion of the *SYNGAP1* allele showed abnormal formation and maturation of dendritic spines in neurons, altered excitatory-inhibitory (E/I) balance, and changed the critical period of development, suggesting that heterozygous variants also have the potential to disrupt brain function in humans and lead to ID through the mechanism of haploinsufficiency (Rumbaugh et al., 2006; Guo et al., 2009; Muhia et al., 2010).

In conclusion, we identified two *de novo* pathogenic heterozygous variants in *SYNGAP1,* c. 333del and c.664-2A>G, among which the c.664-2A>G variant was a novel *de novo* pathogenic variant. Based on previous findings from others and our research, nonsense variants in *SYNGAP1* remain the most common variant type leading to ID. This study further enriched the variant landscape of *SYNGAP1* in ID and provided a basis for the clinical diagnosis and genetic counseling of ID.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://ngdc.cncb.ac.cn/search/?dbId=hra&q=HRA005421.

Ethics statement

The studies involving humans were approved by the Ethics Committee of Qilu Hospital, Shandong University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

BL: Data curation, Formal Analysis, Validation, Writing–original draft. YW: Data curation, Formal Analysis, Writing–original draft. DH: Funding acquisition, Methodology, Writing–review and editing. ZS: Data curation, Investigation. LZ: Investigation, Methodology. NL: Investigation, Methodology. RY: Project administration, Resources. PS: Conceptualization, Project administration, Resources, Supervision, Writing–review and editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Natural Science Foundation of China (82001509) and Natural Science Foundation of Jiangsu Province (BK20200233).

Acknowledgments

We thank the families and clinical staff for participation in this study.

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

References

Agarwal, M., Johnston, M. V., and Stafstrom, C. E. (2019). SYNGAP1 mutations: clinical, genetic, and pathophysiological features. *Int. J. Dev. Neurosci.* 78, 65–76. doi:10.1016/j.ijdevneu.2019.08.003

Berryer, M. H., Hamdan, F. F., Klitten, L. L., Moller, R. S., Carmant, L., Schwartzentruber, J., et al. (2013). Mutations in SYNGAP1 cause intellectual disability, autism, and a specific form of epilepsy by inducing haploinsufficiency. *Hum. Mutat.* 34 (2), 385–394. doi:10.1002/humu.22248

Carvill, G. L., Heavin, S. B., Yendle, S. C., McMahon, J. M., O'Roak, B. J., Cook, J., et al. (2013). Targeted resequencing in epileptic encephalopathies identifies *de novo* mutations in CHD2 and SYNGAP1. *Nat. Genet.* 45 (7), 825–830. doi:10.1038/ng.2646

Chelly, J., Khelfaoui, M., Francis, F., Cherif, B., and Bienvenu, T. (2006). Genetics and pathophysiology of mental retardation. *Eur. J. Hum. Genet.* 14 (6), 701–713. doi:10.1038/sj.ejhg.5201595

Chen, T., Chen, X., Zhang, S., Zhu, J., Tang, B., Wang, A., et al. (2021). The genome sequence archive family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinforma.* 19 (4), 578–583. doi:10.1016/j.gpb.2021.08.001

Clement, J. P., Aceti, M., Creson, T. K., Ozkan, E. D., Shi, Y., Reish, N. J., et al. (2012). Pathogenic SYNGAP1 mutations impair cognitive development by disrupting maturation of dendritic spine synapses. *Cell* 151 (4), 709–723. doi:10.1016/j.cell.2012.08.045

Fieremans, N., Van Esch, H., Holvoet, M., Van Goethem, G., Devriendt, K., Rosello, M., et al. (2016). Identification of intellectual disability genes in female patients with a skewed X-inactivation pattern. *Hum. Mutat.* 37 (8), 804–811. doi:10.1002/humu.23012

Goldenberg, A., and Saugier-Veber, P. (2010). Genetics of mental retardation. *Pathol. Biol. Paris.* 58 (5), 331–342. doi:10.1016/j.patbio.2009.09.013

Guo, X., Hamilton, P. J., Reish, N. J., Sweatt, J. D., Miller, C. A., and Rumbaugh, G. (2009). Reduced expression of the NMDA receptor-interacting protein SynGAP causes behavioral abnormalities that model symptoms of Schizophrenia. *Neuropsychopharmacology* 34 (7), 1659–1672. doi:10.1038/npp.2008.223

Hamdan, F. F., Daoud, H., Piton, A., Gauthier, J., Dobrzeniecka, S., Krebs, M. O., et al. (2011). *De novo* SYNGAP1 mutations in nonsyndromic intellectual disability and autism. *Biol. Psychiatry* 69 (9), 898–901. doi:10.1016/j.biopsych.2010.11.015

Hamdan, F. F., Gauthier, J., Spiegelman, D., Noreau, A., Yang, Y., Pellerin, S., et al. (2009). Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. *N. Engl. J. Med.* 360 (6), 599–605. doi:10.1056/NEJMoa0805392

Hane, B., Schroer, R. J., Arena, J. F., Lubs, H. A., Schwartz, C. E., and Stevenson, R. E. (1996). Nonsyndromic X-linked mental retardation: review and mapping of MRX29 to Xp21. *Clin. Genet.* 50 (4), 176–183. doi:10.1111/j.1399-0004.1996.tb02622.x

Jeyabalan, N., and Clement, J. P. (2016). SYNGAP1: mind the Gap. *Front. Cell Neurosci.* 10, 32. doi:10.3389/fncel.2016.00032

Kallel-Bouattour, R., Belguith-Maalej, S., Zouari-Bradai, E., Mnif, M., Abid, M., and Hadj Kacem, H. (2017). Intronic variants of SLC26A4 gene enhance splicing efficiency in hybrid minigene assay. *Gene* 620, 10–14. doi:10.1016/j.gene.2017.03.043

CNCB-NGDC Members and Partners (2023). Database resources of the national genomics data center, China national center for bioinformation in 2023. *Nucleic Acids Res.* 51 (D1), D18–D28. doi:10.1093/nar/gkac1073

Mignot, C., von Stulpnagel, C., Nava, C., Ville, D., Sanlaville, D., Lesca, G., et al. (2016). Genetic and neurodevelopmental spectrum of SYNGAP1-associated intellectual disability and epilepsy. *J. Med. Genet.* 53 (8), 511–522. doi:10.1136/jmedgenet-2015-103451

Muhia, M., Yee, B. K., Feldon, J., Markopoulos, F., and Knuesel, I. (2010). Disruption of hippocampus-regulated behavioural and cognitive processes by heterozygous constitutive deletion of SynGAP. *Eur. J. Neurosci.* 31 (3), 529–543. doi:10.1111/j.1460-9568.2010.07079.x

Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., et al. (2013). DeNovoGear: *de novo* indel and point mutation discovery and phasing. *Nat. Methods* 10 (10), 985–987. doi:10.1038/nmeth.2611

Rumbaugh, G., Adams, J. P., Kim, J. H., and Huganir, R. L. (2006). SynGAP regulates synaptic strength and mitogen-activated protein kinases in cultured neurons. *Proc. Natl. Acad. Sci. U. S. A.* 103 (12), 4344–4351. doi:10.1073/pnas.0600084103

Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., van Kogelenberg, M., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 385 (9975), 1305–1314. doi:10.1016/S0140-6736(14)61705-0

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 03 November 2023
DOI 10.3389/fmicb.2023.1269164

OPEN ACCESS

EDITED BY
Yuriy L. Orlov,
I.M.Sechenov First Moscow State Medical University, Russia

REVIEWED BY
Hong-Yan Wang,
Chinese Academy of Fishery Sciences (CAFS), China
Sheng Liu,
Indiana University Bloomington, United States
Min Wei,
Nankai University, China

*CORRESPONDENCE
Sheng Tan
✉ tansheng@sjtu.edu.cn
Qin Fang
✉ qfang@wh.iov.cn

†These authors have contributed equally to this work

RECEIVED 29 July 2023
ACCEPTED 28 September 2023
PUBLISHED 03 November 2023

CITATION
Tan S, Zhang J, Peng Y, Du W, Yan J and Fang Q (2023) Integrative transcriptome analysis reveals alternative polyadenylation potentially contributes to GCRV early infection. *Front. Microbiol.* 14:1269164. doi: 10.3389/fmicb.2023.1269164

Integrative transcriptome analysis reveals alternative polyadenylation potentially contributes to GCRV early infection

Sheng Tan[1*†], Jie Zhang[2,3†], Yonglin Peng[1†], Wenfei Du[1], Jingxuan Yan[4] and Qin Fang[2*]

[1]Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai, Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China, [2]State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China, [3]Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China, [4]Bio-ID Center, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

Introduction: Grass carp reovirus (GCRV), a member of the *Aquareovirus genus* in the *Reoviridae* family, is considered to be the most pathogenic aquareovirus. Productive viral infection requires extensive interactions between viruses and host cells. However, the molecular mechanisms underlying GCRV early infection remains elusive.

Methods: In this study we performed transcriptome and DNA methylome analyses with *Ctenopharyngodon idellus* kidney (CIK) cells infected with GCRV at 0, 4, and 8h post infection (hpi), respectively.

Results: We found that at early infection stage the differentially expressed genes related to defense response and immune response in CIK cells are activated. Although DNA methylation pattern of CIK cells 8 hpi is similar to mock-infected cells, we identified a considerable number of genes that selectively utilize alternative polyadenylation sites. Particularly, we found that biological processes of cytoskeleton organization and regulation of microtubule polymerization are statistically enriched in the genes with altered 3′UTRs.

Discussion: Our results suggest that alternative polyadenylation potentially contributes to GCRV early infection.

KEYWORDS

grass carp reovirus, Aquareovirus genus, *Ctenopharyngodon idellus*, DNA methylation, alternative polyadenylation

Introduction

Aquareoviruses are nonenveloped viruses and classified within the family Reoviridae, a family of double-stranded RNA virus composed of aquareoviruses, mammalian reoviruses (MRV), and the other 13 genera. Aquareoviruses cause infection in aquatic organisms including bony fish, shellfish, and crustacean worldwide (Lupianni et al., 1995). Although most aquareoviruses are isolated from seemingly healthy fish and do not give rise to high mortalities, grass carp reovirus (GCRV) is recognized to be most pathogenic among the isolated aquareoviruses (Rangel et al., 1999). GCRV can cause serious hemorrhagic disease in aquatic organisms. Our previous studies have shown that GCRV can induce cell–cell fusion and produce characteristic cytopathic effect (CPE) consisting of large syncytia within infected cultures (Fang et al., 1989; Ke et al., 1990), and it has been extensively used to understand aquareovirus molecular and structural biology. Seven structural (VP1–VP7) and six nonstructural proteins

(NS12, NS16, NS26, NS31, NS38, and NS80) of GCRV have been well identified (Guo et al., 2013; Yan et al., 2015). Comparative proteomic analysis of lysine acetylation in fish Ctenopharyngodon idellus kidney (CIK) cells reveals the proteome-wide changes in host cell acetylome with GCRV infection (Guo et al., 2017). MRV can cause chronic infection. It has been revealed that there is a close molecular evolutionary relationship between aquareoviruses and mammalian orthoreoviruses. In addition to morphological similarity, GCRV and MRV share high amino acid conservation. A better knowledge of the interaction during early infection stage between GCRV and host cells will help the understanding molecular pathogenesis of the aquareovirus and other members in the family Reoviridae.

Accumulating evidence has demonstrated that epigenetics is actively involved in host-virus interaction. Epigenetic trait is defined as a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence (Berger et al., 2009). These chromosomal changes include methylation of cytosine in CpG dinucleotides (often referred to as DNA methylation) and other posttranslational covalent modifications to histones, such as methylation, acetylation, and ubiquitylation. The epigenetic modifications are associated with structural organization of chromatin and transcriptional activities of the affected genes. As intracellular parasites, viruses develop various ways of remodeling epigenetic alterations to facilitate their infection and replication. Through inducing DNA methylation changes in host cells viruses epigenetically manipulate host functions upon virus infection. For instance, Epstein–Barr virus (EBV) infection activates cellular DNA methyltransferases and results in aberrant DNA methylation in host cells (Tsai et al., 2006; Hino et al., 2009). HIV infection can also trigger the differential DNA methylation at cis-regulatory regions of host genomic DNA and inhibit the function of T cells (Pion et al., 2013; Youngblood et al., 2013). Nevertheless, the influence on cellular DNA methylation during GCRV infection remains to be further characterized.

In addition to epigenetic modifications, formation of stress granules is also actively involved in the interaction between viruses and host cells. It has been recognized that the innate immune response of host cells is triggered by upon virus infection to prevent pathogen invasion, partially through stress granules. Some components of stress granules have been identified, such as T-cell-restricted intracellular antigen 1 (TIA-1), TIA-1-related protein (TIAR), Ras GTPase-activating protein-binding proteins (G3BPs)and poly(A)-binding proteins (PABPs). PABPs are a family of RNA recognition motif (RRM)-containing proteins that bind poly(A) tail and regulate translation and stability of mRNA. The previous report has demonstrated that alternative polyadenylation (APA) plays an important role in the antiviral innate immune response (Jia et al., 2017). However, it remains unclear whether APA of host cells is involved in GCRV infection. Thus, in this study we carried out integrative analyses of transcriptome, DNA methylome and APA in GCRV-infected CIK cells for understanding the molecular events in GCRV early infection.

Materials and methods

Cells, virus and infection assays

CIK cells, purchased from the China Center for Type Culture Collection (CCTCC, 4201FIS-CCTCC00086), were grown in minimum essential medium (MEM; Gibco-BRL) supplemented with 10% fetal bovine serum (FBS), 100 mg/mL penicillin, and 100 mg/mL streptomycin at 28°C. Grass carp reovirus (strain GCRV-873), previously isolated and stored in the author's laboratory, was propagated in CIK cells with Eagle's MEM supplemented with 2% FBS (MEM-2) as previously described (Fang et al., 1989).

Viral infection, cytopathic effect observation and plaque assay

The infection assays were carried out as we described previously (Zhang et al., 2019). Briefly, the 80% confluent CIK cells in T-25 flask (Corning Inc., Corning, NY, United States) with a concentration of 2×10^6 cells/ml were inoculated with GCRV at a multiplicity of infection (MOI) of 1 in serum-free MEM medium at 28°C for 1 h following the method as previously reported (Guo et al., 2017). For comparison, the mock-infected cells were treated with same amount of medium in the same conditions. Upon adsorption, cells were washed with phosphate-buffered saline (PBS) to remove non-adsorbed virions. The infected cells were maintained in MEM-2 at 28°C and harvested at 0 (mock), 4 and 8 h post infection, respectively. When initial cytopathic effects were observed, the infected cells and mock-infected cells were prepared and harvested for further transcriptome analyses. Three rounds of independent experiments were performed. For MOI determination, plaque assays were done according to our previously described method (Yan et al., 2015; Zhang et al., 2018).

RNA isolation, RNA-seq library construction and deep sequencing

CIK Cells were infected by GCRV for 0, 4 and 8 h, respectively. Total RNA was extracted with Trizol reagent (Invitrogen, United States), which was further treated with RNase-free DNase to remove genomic DNA. mRNA was purified with poly(dT) oligo-attached magnetic beads and broken down into 200 ~ 400 bp fragments. A strand-specific RNA-seq library was constructed with NEBNext Ultra Directional RNA Library Prep Kit (NEB, New England, United States). Briefly, the fragmented mRNA was reversely transcribed into cDNA with random primers and then the second-strand cDNA was generated. The resulting double-strand DNA fragments were purified with AMPure beads (Beckman Coulter, Brea, CA, United States) and ligated with Illumina adapters. The ligation products were purified by agarose gel electrophoresis to remove adapter dimmers, which were subsequently subjected to HiSeq X sequencing (Illumina, San Diego, CA, United States). The raw sequencing data could be obtained in the EMBL database[1] under accession number E-MTAB-13002.[2]

1 http://www.ebi.ac.uk/arrayexpress/

2 https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-13002

MeDIP-seq library construction and deep sequencing

Genomic DNA of CIK cells was extracted using GenElute™ Mammalian Genomic DNA Miniprep Kit (Sigma, United States). DNA was randomly sheared into fragments of 200 ~ 500 bp and used for library preparation with NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB), the resulting libraries were purified with 1 × Agencourt AMPure XP beads (Beckman Coulter). The immunoprecipitation procedure was basically performed according to a previous MeDIP protocol (Taiwo et al., 2012). Briefly, the library DNA was denatured at 95°C for 10 min and immediately placed into an ice for 10 min, 1/10 volume of denatured product was set aside as Input. The Protein A + G magnetic beads (Millipore, United States) were incubated with 5-Methylcytosine (5-mC) monoclonal antibody (Epigentek) at 4°C for 2 h and the library was incubated with antibody-bead complexes at 4°C overnight with a slight rotation. The dynabead-antibody-methylated DNA complexes were washed three times, followed by proteinase K (Thermo scientific) treatment for 3 h at 55°C. The immunoprecipitated DNA was extracted by phenol/chloroform/isoamylalcohol, followed by ethanol precipitation, and resuspended in EB buffer (10 mM Tris–HCl pH 8.0). The enriched methylated DNA and Input DNA were amplified using Q5 High-Fidelity DNA Polymerase (NEB), and subject to Illumina sequencing platforms. The raw sequencing data could be obtained in the EMBL database (see Footnote 1) under accession number E-MTAB-13003.[3]

Bioinformatics analysis

The raw reads with low quality and the adapter sequences of RNA-seq and MeDIP-seq data were removed using Cutadapt v4.1 (Kechin et al., 2017). For RNA-seq data, clean reads were mapped to the grass carp reference genome (Wang et al., 2015) using Hisat v2.2.1 (Kim et al., 2019). The Subread toolkits was used to quantify read counts for genes (Liao et al., 2014), and reads per kilobase of transcript per million mapped reads (RPKM) were calculated as expression levels. Differential expression analysis was performed using the edgeR package in R platform v3.6.3 (Robinson et al., 2010). Those genes with an value of $p < 0.05$ and fold change >1.5 were regarded as differentially expressed genes (DEGs). For MeDIP-seq data, clean reads were mapped to the grass carp reference genome using Bowtie v2.4.5 (Langmead and Salzberg, 2012). PCR duplicate reads were removed with Picard v2.27.4.[4] DNA methylation peaks were called with MACS2 with deduplicated alignments (Zhang et al., 2008) and the differentially methylated regions (DMRs) were identified with DiffBind and DESeq2 packages (Anders and Huber, 2010). Functional enrichment analysis was performed with DAVID.[5]

3 https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-13003
4 https://github.com/broadinstitute/picard
5 https://david.ncifcrf.gov/home.jsp

Analysis of APA with RNA-seq data

The APAs were identified by using DaPars algorithm (Masamha et al., 2014) based on RNA-seq data. Briefly, the observed sequence coverage was represented as a linear combination of annotated 3'UTRs. For each transcript with annotated proximal adenylation site (PAS), a regression model was used to infer the end point of alternative novel PAS within this 3' UTR at single nucleotide resolution, by minimizing the deviation between the observed read coverage and the expected read coverage based on a two-PAS model in both mock-infected and GCRV-infected samples simultaneously. A percentage dPAS usage index (PDUI) was utilized to define shortening (negative index) or lengthening (positive index) of 3'UTR and thus capable of quantifying the degree of difference in 3'UTR usage between mock-infected and GCRV-infected CIK cells. The greater PDUI means that the more distal PAS of a given transcript is used and vice versa.

Results

Grass carp reovirus infection-induced cytopathic effects at early stage

To characterize the interaction of GCRV and host cells for integrative analyses of transcriptome in GCRV-infected CIK cells, we firstly carefully examined the cytopathic effects induced by GCRV infection at early stage. In both mock-infected cells and GCRV-infected cells at 4 h, we did not observe obvious CPE. As infection progressed, we detected an initial characteristic CPE on the monolayers of CIK cells at 8 h post infection (hpi) by comparing to mock-infected cell (Figure 1), which suggests that efficient infection was obtained, and the harvested infected cell lysates were suitable for follow-up transcriptome related assays.

Transcription program associated with grass carp reovirus early infection

To detect the molecular events at the GCRV early infection, we performed RNA-seq analysis of CIK cells 0, 4 h and 8 hpi. Totally we generated 197.5 millions raw sequencing reads in the groups of mock, 4 and 8 hpi. Among these reads 94.7% are mappable and are used for downstream analysis.

Totally we identified 15,255 expressed genes in three groups. We then used gene set enrichment analysis (GSEA) to compare the transcriptome data between CIK cells 8 hpi and the MOCK-infected cells. We found that several gene sets were significantly enriched in cells 8 hpi comparing with the MOCK, such as defense response to virus, immune response, and cholesterol metabolic process (Figure 2A). Comparing with MOCK, we identified 675 differentially expressed genes in CIK cells 8 hpi (Figure 2B). Gene ontology (GO) analysis indicates that the biological processes of defense response to virus, cholesterol metabolic process (Figure 2C) and mitogen-activated protein kinase (MAPK) signaling pathway (Figure 2D) are significantly enriched in these differentially expressed genes.

DNA methylation pattern in grass carp reovirus early infected *Ctenopharyngodon idellus* kidney cells

It has been reported that DNA methylation contributes to resistance to GCRV infection (Shang et al., 2017). Here, we asked whether DNA methylation is involved in GCRV early infection. To address this issue, we performed MeDIP-seq analysis with MOCK and CIK cells 8 hpi. We totally generated 232 million sequence reads, and 96% are mappable. The data sets of biological replicates are highly correlated (Supplementary Figure). Among 9,426 identified methylation sites 37.7% are located in intergenic regions, 32% in exons, 21% in introns and only 9.3% in promoter regions (Figure 3A). We examined DNA methylation signal around transcription starting site (TSS) and found the obviously enriched methylation signal at TSS regions both in MOCK and 8 hpi groups (Figure 3B). It is well recognized that DNA methylation is negatively associated with gene expression. We then examined the correlation between methylated genomic regions and transcription levels. We observed that the methylated regions at promoters, exons and introns are weakly and negatively correlated with transcription (Figure 3C). Compared with the MOCK, we found the DNA methylation of CIK cells 8 hpi is very similar to MOCK (Figure 3D), suggesting that DNA methylation pattern is less functionally involved in early GCRV infection.

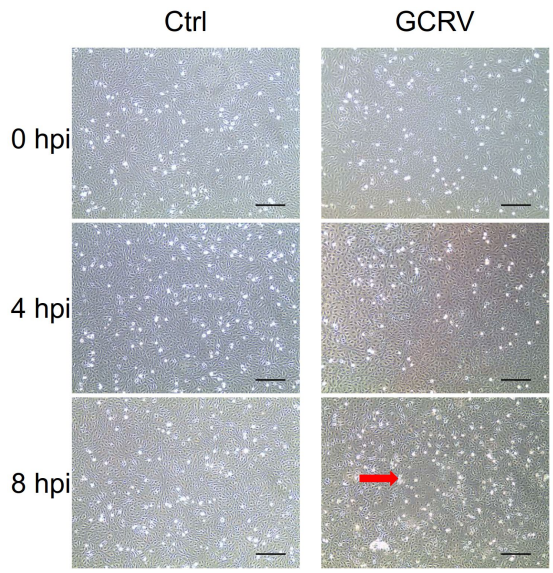


FIGURE 1
Grass carp reovirus (GCRV)-induced cytopathic effect in CIK cells at different time points. CIK cells were mock-infected (Left panel) or infected with GCRV(Right panel), and the images were taken at 0, 4 and 8 hpi, respectively. Red arrow shows the representative CPE. Sale bar: 200 μm.

Alternative polyadenylation profile in grass carp reovirus early infected *Ctenopharyngodon idellus* kidney cells

Since DNA methylation is less involved in GCRV early infection, we next investigated other mechanisms. Alternative polyadenylation (APA) modulates gene expression and has been reported to be involved in antiviral response. We then examined the APA patterns between MOCK and 8 hpi group. Comparing with the MOCK, we identified 404 genes with the APA-derived altered 3'UTRs, including 201 genes with lengthened 3'UTRs and 203 genes

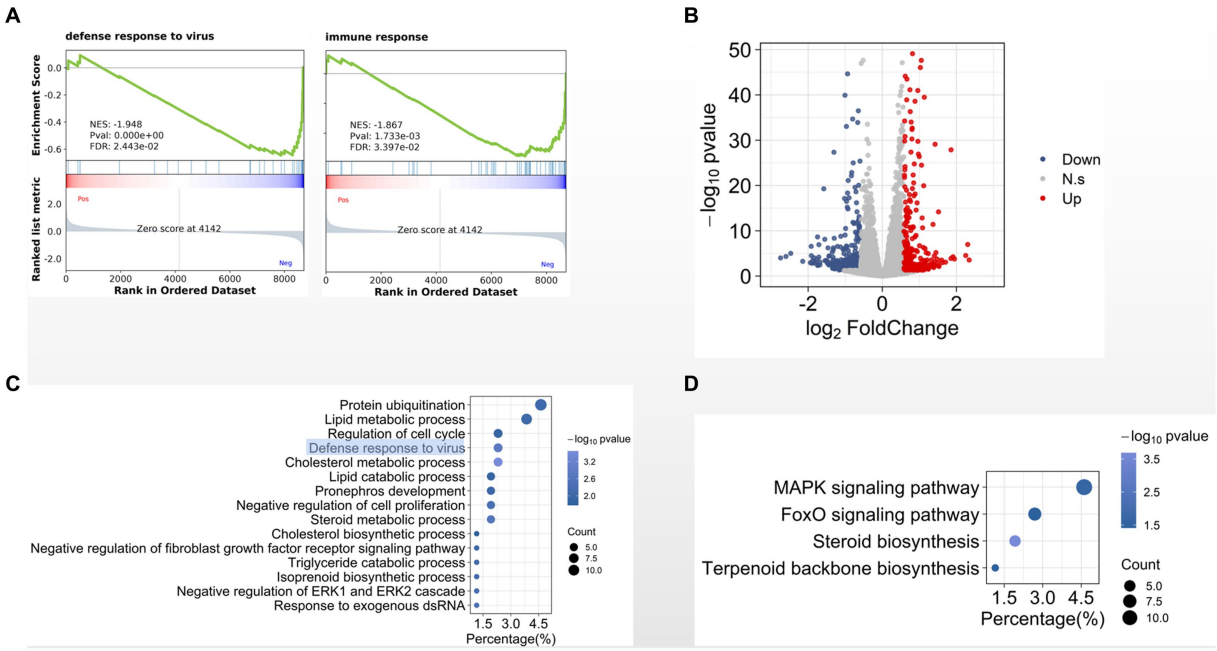


FIGURE 2
Transcriptional programs in GCRV-infected CIK. **(A)** Enrichment plot for defense- and immune-related genes. **(B)** Volcano plot of statistically significant differentially expressed genes in early infection. **(C, D)**. GO **(C)** and KEGG pathway **(D)** analyses of differentially expressed genes.

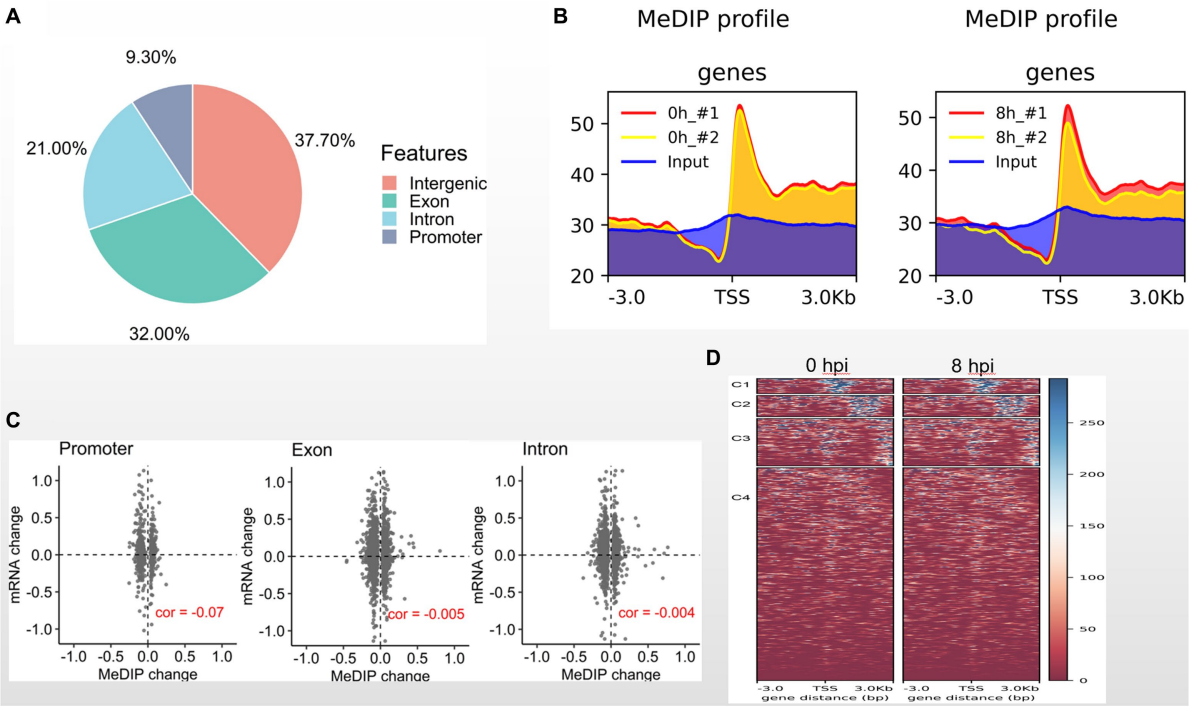


FIGURE 3
DNA methylome in early GCRV infection. **(A)** Genomic distribution of methylated regions in CIK cells. **(B)** DNA methylated signal around TSS regions. **(C)** Correlation between transcriptional signal and methylated regions. **(D)** DNA methylation heatmaps of MOCK and two biological replicates of 8 hpi group.

with shortened 3'UTRs (Figure 4A). When examining the 3'UTR alterations and transcription levels, we observed that the overall transcription levels of the genes with shortened 3'UTRs is higher than the those with lengthened 3'UTRs (Figure 4B). Through GO analysis with the genes containing altered 3'UTRs we found the biological processes of cytoskeleton organization and regulation of microtubule polymerization are statistically enriched (Figure 4C). In particular, we observed that *Camsap1b*, a gene involved in microtubule formation and stability, preferentially utilized the proximal poly(A) sites in GRCV-infected CIK cells when comparing MOCK (Figure 4D). These observations suggest that alternative poly(A) usage is potentially involved in the early infection of CIK cells.

Discussion

Viral infections involve intensive interactions between viruses and host cells. As obligate intracellular parasites, viruses misappropriate host cellular machinery to allow their replication; while host cells also orchestrate the transcriptional programs to repress viral infection. For example, in our previous studies we found that aquareovirus NS38 (the GCRV nonstructural protein expressed in host cells as early as 3 h post infection) interacts with host translation initiation factor eIF3A for virus replication (Shao et al., 2013; Zhang et al., 2019). Meanwhile, host innate immune response would be activated after virus infection. Consist with the reported studies (Chen et al., 2012; Shi et al., 2014; Wan and Su, 2015; Dang et al., 2016; Xu et al., 2016; Chen et al., 2018), we observed that the host genes related to defense response to virus and immune response are differentially expressed in CIK cells 8 hpi (Figures 2A,C). In addition to these immune-related genes, we found the genes involved in cholesterol metabolic process and cholesterol biosynthetic process are also activated (Figure 2C), supporting our previous report that cellular membrane cholesterol is required for GCRV productive infection (Zhang et al., 2018). Activation of MAPK signaling pathway has been reported to be required for cell entry of avian reovirus (Huang et al., 2011). Interestingly, in this study we found that this pathway is most significantly enriched among all identified cellular signaling pathways (Figure 2D), suggesting MAPK signaling pathway is involved in GCRV infection.

DNA methylation has been reported to control the resistance and susceptibility to GCRV infection in CIK cells (Shang et al., 2017). In this study we examined the DNA methylome of CIK cells 8 hpi and found that the DNA methylation pattern of infected cells is very similar to the MOCK (Figure 3D). The statistically enriched biological processes of differentially methylated genes do not include defense response to virus or immune response (data not shown). These findings indicate that DNA methylation is less functionally involved in early GCRV infection.

Alternative polyadenylation functionally contributes to antiviral immune response (Jia et al., 2017). Some poly(A) binding proteins are the components of stress granules, the membrane-less ribonucleoprotein (RNP)-based cellular compartments in the cytoplasm triggering antiviral immune response. Moreover, alternative polyadenylation is involved in chronically infected disease (Su et al.,

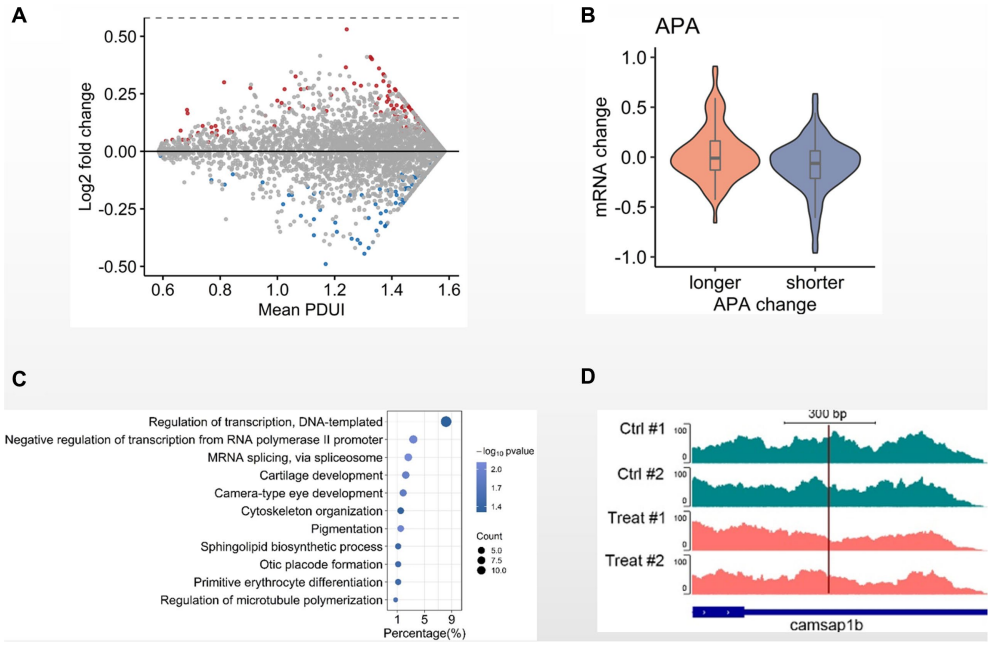


FIGURE 4
Alternative polyadenylation analysis in GCRV-infected CIK cells. **(A)** MA-plot depicted 3′UTRs of transcripts demarcated by DaPars-defined APA sites. The 3′UTRs were significantly shortened (blue) or lengthened (red) in CIK cells 8 hpi when compared the MOCK (value of $p < 0.05$). **(B)** Violin plot showing transcription level change of genes with altered 3′ UTR in GCRV-infected CIK cells vs. MOCK comparison. **(C)** GO items of biological processes enriched in genes with altered 3′UTR. **(D)** Genomic view of APA site usage preference at *Camsap1b* 3′UTR in IGV browser, showing the transcriptional density shifting to the proximal APA upon GCRV infection.

2001). Previously, we have performed extensive alternative polyadenylation analysis to understand its functional relevance in tumorigenesis (Lai et al., 2015; Tan et al., 2018, 2021). In this study we identified a considerable number of genes that selectively utilize alternative poly(A) sites in GCRV-infected CIK cells (Figures 4A,B). Among the genes with altered 3'UTRs we identified the biological processes of cytoskeleton organization, regulation of microtubule polymerization (Figures 4C,D). Interestingly, our recent study reported that microtubules are required for productive GCRV infection (Zhang et al., 2020), which is similar to MRV infection (Mainou et al., 2013).These observations suggest that alternative polyadenylation is potentially involved in GCRV early infection. Taken together, our study provides evidence of molecular events during early infection of dsRNA viruses for understanding their pathogenesis.

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.

Ethics statement

The animal studies were approved by the Research Ethics Committee of the Shanghai Jiao Tong University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

ST: Conceptualization, Data curation, Writing – review & editing. JZ: Data curation, Resources, Writing – review & editing. YP: Data curation, Software, Writing – review & editing, Formal analysis, Investigation. WD: Investigation, Writing – review & editing. JY: Writing – review & editing, Formal analysis. QF: Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China grant (31972838 and 82173231).

Acknowledgments

The authors are grateful to Yuliang Deng of Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University for experimental assistance.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1269164/full#supplementary-material

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Berger, S. L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes Dev.* 23, 781–783. doi: 10.1101/gad.1787609

Chen, G., He, L., Luo, L., Huang, R., Liao, L., Li, Y., et al. (2018). Transcriptomics sequencing provides insights into understanding the mechanism of grass carp Reovirus infection. *Int. J. Mol. Sci.* 19:488. doi: 10.3390/ijms19020488

Chen, J., Li, C., Huang, R., Du, F., Liao, L., Zhu, Z., et al. (2012). Transcriptome analysis of head kidney in grass carp and discovery of immune-related genes. *BMC Vet. Res.* 8:108. doi: 10.1186/1746-6148-8-108

Dang, Y., Xu, X., Shen, Y., Hu, M., Zhang, M., Li, L., et al. (2016). Transcriptome analysis of the innate immunity-related complement system in spleen tissue of *Ctenopharyngodon idella* infected with *Aeromonas hydrophila*. *PLoS One* 11:e0157413. doi: 10.1371/journal.pone.0157413

Fang, Q., Ke, L. H., and Cai, Y. Q. (1989). Growth characteristics and high titer culture of grass carp hemorrhage virus (GCHV)-873 in vitro. *Virol. Sin.* 3, 315–319.

Guo, H., Sun, X., Yan, L., Shao, L., and Fang, Q. (2013). The NS16 protein of aquareovirus-C is a fusion-associated small transmembrane (FAST) protein, and its activity can be enhanced by the nonstructural protein NS26. *Virus Res.* 171, 129–137. doi: 10.1016/j.virusres.2012.11.011

Guo, H., Zhang, J., Wang, Y., Bu, C., Zhou, Y., and Fang, Q. (2017). Comparative proteomic analysis of lysine acetylation in fish CIK cells infected with Aquareovirus. *Int. J. Mol. Sci.* 18:2419. doi: 10.3390/ijms18112419

Hino, R., Uozaki, H., Murakami, N., Ushiku, T., Shinozaki, A., Ishikawa, S., et al. (2009). Activation of DNA methyltransferase 1 by EBV latent membrane protein 2A leads to promoter hypermethylation of PTEN gene in gastric carcinoma. *Cancer Res.* 69, 2766–2774. doi: 10.1158/0008-5472.CAN-08-3070

Huang, W. R., Wang, Y. C., Chi, P. I., Wang, L., Wang, C. Y., Lin, C. H., et al. (2011). Cell entry of avian reovirus follows a caveolin-1-mediated and dynamin-2-dependent endocytic pathway that requires activation of p38 mitogen-activated protein kinase (MAPK) and Src signaling pathways as well as microtubules and small GTPase Rab5 protein. *J. Biol. Chem.* 286, 30780–30794. doi: 10.1074/jbc.M111.257154

Jia, X., Yuan, S., Wang, Y., Fu, Y., Ge, Y., Ge, Y., et al. (2017). The role of alternative polyadenylation in the antiviral innate immune response. *Nat. Commun.* 8:14605. doi: 10.1038/ncomms14605

Ke, L. H., Fang, Q., and Cai, Y. Q. (1990). Characteristics of a novel isolate of grass carp hemorrhagic virus. *Acta Hydrobiologica Sinica* 14, 153–159.

Kechin, A., Boyarskikh, U., Kel, A., and Filipenko, M. (2017). cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing. *J. Comput. Biol.* 24, 1138–1143. doi: 10.1089/cmb.2017.0096

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Lai, D. P., Tan, S., Kang, Y. N., Wu, J., Ooi, H. S., Chen, J., et al. (2015). Genome-wide profiling of polyadenylation sites reveals a link between selective polyadenylation and cancer metastasis. *Hum. Mol. Genet.* 24, 3410–3417. doi: 10.1093/hmg/ddv089

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Lupianni, B., Subramanian, K., and Samal, S. K. (1995). Aquareoviruses. *Annu. Rev. Fish Dis.* 5, 175–208. doi: 10.1016/0959-8030(95)00006-2

Mainou, B. A., Zamora, P. F., Ashbrook, A. W., Dorset, D. C., Kim, K. S., and Dermody, T. S. (2013). Reovirus cell entry requires functional microtubules. *MBio* 4:13. doi: 10.1128/mBio.00405-13

Masamha, C. P., Xia, Z., Yang, J., Albrecht, T. R., Li, M., Shyu, A. B., et al. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* 510, 412–416. doi: 10.1038/nature13261

Pion, M., Jaramillo-Ruiz, D., Martínez, A., Muñoz-Fernández, M. A., and Correa-Rocha, R. (2013). HIV infection of human regulatory T cells downregulates Foxp3 expression by increasing DNMT3b levels and DNA methylation in the FOXP3 gene. *AIDS* 27, 2019–2029. doi: 10.1097/QAD.0b013e32836253fd

Rangel, A. A. C., Rockemann, D. D., Hetrick, F. M., and Samal, S. K. (1999). Identification of grass carp haemorrhage virus as a new genogroup of aquareovirus. *J. Gen. Virol.* 80, 2399–2402. doi: 10.1099/0022-1317-80-9-2399

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Shang, X., Yang, C., Wan, Q., Rao, Y., and Su, J. (2017). The destiny of the resistance/susceptibility against GCRV is controlled by epigenetic mechanisms in CIK cells. *Sci. Rep.* 7:4551. doi: 10.1038/s41598-017-03990-5

Shao, L., Guo, H., Yan, L. M., Liu, H., and Fang, Q. (2013). Aquareovirus NS80 recruits viral proteins to its inclusions, and its C-terminal domain is the primary driving force for viral inclusion formation. *PLoS One* 8:e55334. doi: 10.1371/journal.pone.0055334

Shi, M., Huang, R., Du, F., Pei, Y., Liao, L., Zhu, Z., et al. (2014). RNA-seq profiles from grass carp tissues after reovirus (GCRV) infection based on singular and modular enrichment analyses. *Mol. Immunol.* 61, 44–53. doi: 10.1016/j.molimm.2014.05.004

Su, Q., Wang, S. F., Chang, T. E., Breitkreutz, R., Hennig, H., Takegoshi, K., et al. (2001). Circulating hepatitis B virus nucleic acids in chronic infection: representation of differently polyadenylated viral transcripts during progression to nonreplicative stages. *Clin. Cancer Res.* 7, 2005–2015.

Taiwo, O., Wilson, G. A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., et al. (2012). Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat. Protoc.* 7, 617–636. doi: 10.1038/nprot.2012.012

Tan, S., Li, H., Zhang, W., Shao, Y., Liu, Y., Guan, H., et al. (2018). NUDT21 negatively regulates PSMB2 and CXXC5 by alternative polyadenylation and contributes to hepatocellular carcinoma suppression. *Oncogene* 37, 4887–4900. doi: 10.1038/s41388-018-0280-6

Tan, S., Zhang, M., Shi, X., Ding, K., Zhao, Q., Guo, Q., et al. (2021). CPSF6 links alternative polyadenylation to metabolism adaption in hepatocellular carcinoma progression. *J. Exp. Clin. Cancer Res.* 40:85. doi: 10.1186/s13046-021-01884-z

Tsai, C. L., Li, H. P., Lu, Y. J., Hsueh, C., Liang, Y., Chen, C. L., et al. (2006). Activation of DNA methyltransferase 1 by EBV LMP1 involves c-Jun NH(2)-terminal kinase signaling. *Cancer Res.* 66, 11668–11676. doi: 10.1158/0008-5472.CAN-06-2194

Wan, Q., and Su, J. (2015). Transcriptome analysis provides insights into the regulatory function of alternative splicing in antiviral immunity in grass carp (*Ctenopharyngodon idella*). *Sci. Rep.* 5:12946. doi: 10.1038/srep12946

Wang, Y., Lu, Y., Zhang, Y., Ning, Z., Li, Y., Zhao, Q., et al. (2015). The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation. *Nat. Genet.* 47, 625–631. doi: 10.1038/ng.3280

Xu, B. H., Zhong, L., Liu, Q. L., Xiao, T. Y., Su, J. M., Chen, K. J., et al. (2016). Characterization of grass carp spleen transcriptome during GCRV infection. *Genet. Mol. Res.* 15:gmr6650. doi: 10.4238/gmr.15026650

Yan, S., Zhang, J., Guo, H., Yan, L., Chen, Q., Zhang, F., et al. (2015). VP5 autocleavage is required for efficient infection by in vitro-recoated aquareovirus particles. *J. Gen. Virol.* 96, 1795–1800. doi: 10.1099/vir.0.000116

Youngblood, B., Noto, A., Porichis, F., Akondy, R. S., Ndhlovu, Z. M., Austin, J. W., et al. (2013). Cutting edge: prolonged exposure to HIV reinforces a poised epigenetic program for PD-1 expression in virus-specific CD8 T cells. *J. Immunol.* 191, 540–544. doi: 10.4049/jimmunol.1203161

Zhang, F., Guo, H., Chen, Q., Ruan, Z., and Fang, Q. (2020). Endosomes and Microtubles are required for productive infection in Aquareovirus. *Virol. Sin.* 35, 200–211. doi: 10.1007/s12250-019-00178-1

Zhang, J., Guo, H., Zhang, F., Chen, Q., Chang, M., and Fang, Q. (2019). NS38 is required for aquareovirus replication via interaction with viral core proteins and host eIF3A. *Virology* 529, 216–225. doi: 10.1016/j.virol.2019.01.029

Zhang, F., Guo, H., Zhang, J., Chen, Q., and Fang, Q. (2018). Identification of the caveolae/raft-mediated endocytosis as the primary entry pathway for aquareovirus. *Virology* 513, 195–207. doi: 10.1016/j.virol.2017.09.019

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 17 May 2024
DOI 10.3389/fgene.2024.1402771

OPEN ACCESS

EDITED BY
Hua Li,
Shanghai Jiao Tong University, China

REVIEWED BY
Ping Yuan,
Sun Yat-Sen University, China
Yi-Wen Liu,
Tunghai University, Taiwan
Xiaodong Zhao,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Jing Tian,
tianjing@nwu.edu.cn

RECEIVED 18 March 2024
ACCEPTED 22 April 2024
PUBLISHED 17 May 2024

CITATION
Lyu Z, Kou Y, Fu Y, Xie Y, Yang B, Zhu H and Tian J (2024), Comparative transcriptomics revealed neurodevelopmental impairments and ferroptosis induced by extremely small iron oxide nanoparticles.
Front. Genet. 15:1402771.
doi: 10.3389/fgene.2024.1402771

Comparative transcriptomics revealed neurodevelopmental impairments and ferroptosis induced by extremely small iron oxide nanoparticles

Zhaojie Lyu[1,2], Yao Kou[1], Yao Fu[1], Yuxuan Xie[1], Bo Yang[1], Hongjie Zhu[2] and Jing Tian[1,2]*

[1]Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, College of Life Sciences, Northwest University, Xi'an, China, [2]Center for Automated and Innovative Drug Discovery, School of Medicine, Northwest University, Xi'an, China

Iron oxide nanoparticles are a type of nanomaterial composed of iron oxide (Fe_3O_4 or Fe_2O_3) and have a wide range of applications in magnetic resonance imaging. Compared to iron oxide nanoparticles, extremely small iron oxide nanoparticles (ESIONPs) (~3 nm in diameter) can improve the imaging performance due to a smaller size. However, there are currently no reports on the potential toxic effects of ESIONPs on the human body. In this study, we applied ESIONPs to a zebrafish model and performed weighted gene co-expression network analysis (WGCNA) on differentially expressed genes (DEGs) in zebrafish embryos of 48 hpf, 72 hpf, 96 hpf, and 120 hpf using RNA-seq technology. The key hub genes related to neurotoxicity and ferroptosis were identified, and further experiments also demonstrated that ESIONPs impaired the neuronal and muscle development of zebrafish, and induced ferroptosis, leading to oxidative stress, cell apoptosis, and inflammatory response. Here, for the first time, we analyzed the potential toxic effects of ESIONPs through WGCNA. Our studies indicate that ESIONPs might have neurotoxicity and could induce ferroptosis, while abnormal accumulation of iron ions might increase the risk of early degenerative neurological diseases.

KEYWORDS

high-throughput sequencing, RNA-seq, WGCNA, nervous system, neurotoxicity, ferroptosis, ESIONPs

Introduction

Iron oxide nanoparticles, including magnetite (Fe_3O_4), hematite (α-Fe_2O_3), and maghemite (γ-Fe_2O_3) NPs as well as modified products, have been widely used in drug carriers and imaging (Lee et al., 2015). For better absorption and imaging, extremely small iron oxide nanoparticles (<5 nm in diameter) (ESIONPs) have been synthesized and modified. ESIONPs have shown the great application values in magnetic resonance imaging (MRI) due to their unique properties, such as switchable contrast signals and high biocompatibility (Kim et al., 2011; Shen et al., 2017; Cao et al., 2020). In addition, ESIONPs are also used as highly sensitive probes for detecting tumors and other lesions (Groult et al., 2021; Mishra et al., 2022; Zhang et al., 2022). As a component of nanotechnology, the application of ESIONPs in different fields is rapidly increasing,

and understanding their potential cytotoxicity and mechanism is crucial for the safety of their application (Mohammadinejad et al., 2019).

Related studies have shown that upon ingestion, iron oxide nanoparticles can impact various organs and tissues in the human body. For example, Fe_3O_4 nanoparticles reduced neuronal activity, triggered oxidative stress, and might be related to the development of neurodegenerative diseases (Wu et al., 2013); ultra-small superparamagnetic iron oxide NPs accumulated in the lower digestive tract and induced cellular autophagy (Schütz et al., 2014). In addition, the spleen is the main organ responsible for clearing iron oxide nanoparticles from the systemic circulation. The proteomic analysis results showed that iron oxide nanoparticles could promote autophagy and lysosomal activation of splenic macrophages through the AKT/mTOR/TFEB signaling pathway (Han et al., 2022). Moreover, iron oxide nanoparticles detected in the environment also increased the risk of early neurodegenerative diseases among urban residents (Calderón-Garcidueñas et al., 2022). Compared with iron oxide nanoparticles, the potential toxic effects of ESIONPs with higher adsorption capacity and accumulation on the human body still need to be systematically investigated and summarized (Kim et al., 2011).

Zebrafish has a high degree of genetic homology with humans and is widely used to detect the toxicity of nanomaterials (Yang et al., 2018; Qiu et al., 2019). Previous studies showed that iron oxide nanoparticles could penetrate the chorion and act directly on the zebrafish embryos, leading to death, malformation, developmental delay, hatching failure, oxidative stress, and the alteration of redox homeostasis (Pitt et al., 2018; Chemello et al., 2019; Pereira et al., 2020; Thirumurthi et al., 2022). The accumulation of iron oxide nanoparticles in zebrafish larvae also caused the obvious cardiotoxicity, characterized by slowed heart rate, pericardial edema, and cardiac hemorrhage (Pereira et al., 2020). Therefore, zebrafish is suitable as a model animal to detect the toxicity of ESIONPs.

In this study, for the first time, we dynamically analyzed the toxicity of ESIONPs (~3 nm in diameter) at multiple embryonic development stages by weighted gene co-expression network analysis (WGCNA). The purpose is to explore the impact of ESIONPs on gene expression at different stages of embryonic development, identify the central regulatory genes and related mechanisms affected. It will help evaluate the safety of ESIONPs application, as well as provide valuable insights for the research of other NPs.

Materials and methods

Zebrafish husbandry and embryo collection

Zebrafish (*Danio rerio*) was raised according to standard protocols. The following zebrafish lines were used: AB wild-type (wt) strain, and transgenic *Tg(eef1a1l1:EGFP)* expressing enhanced green fluorescent protein (GFP) in neuron cells. Zebrafish embryos were obtained by natural spawning, collected within 30 min after fertilization, and cultured at 28.5°C (Tian et al., 2019). To evaluate the toxicity of ESIONPs, a dose–response analysis was carried out to determine the median lethal dose (LC50). The zebrafish embryos at 4 h post-fertilization (hpf) were distributed in 6-well plates (30 for each group), and exposed to ESIONPs suspensions at different concentrations (0 mg/L, 10 mg/L, 20 mg/L, 30 mg/L, 40 mg/L, 60 mg/L, 80 mg/L, and 100 mg/L). The medium was changed every 24 h. The survival rate was determined every day by counting the embryos that survived. The exposed embryos were collected at indicated stages for different analysis. Embryos from each group were observed and photographed taken an SMZ25 stereomicroscope with a DS-Ri2 digital camera (Nikon, Japan). All experimental procedures on zebrafish were approved by the Institutional Animal Care and Use Committee of Northwest University and carried out in accordance with the approved guidelines (NWU-AWC-20190103Z).

RNA library preparation and sequencing

Embryos in 40 mg/L ESIONP-exposed group and control group at 48 hpf, 72 hpf, 96hpf, and 120 hpf were collected and used for total RNA extraction. mRNA was isolated using the NEBNext PolyA mRNA Magnetic Isolation Module (New England Biolabs, Ipswich, MA, United States). Libraries were prepared with the NEB Next Ultra Directional RNA Library Prep Kit (New England Biolabs, United States), and subjected to Illumina sequencing with paired end 2 × 150 as the sequencing mode. The clean reads were mapped to reference genome (*D. rerio*: NCBI_GRCz11). Gene expression levels were estimated using FPKM (fragments per kilobase of exon per million fragments mapped) by StringTie v1.3.4d (Pertea et al., 2015). Differential expressed genes (DEG) were measured using R package, edgeR v3.24.2 (Robinson et al., 2010). The false discovery rate (FDR) was used to calculate the adjusted *p*-value in multiple testing in order to evaluate the significance of the differences. Here, only gene with an adjusted q-value < 0.05 and $|\log2FC| \geq 1$ were used for subsequent analysis. The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Chen et al., 2021) in National Genomics Data Center (CNCB-NGDC Members and Partners, 2024), China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA016266) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

Weighted gene co-expression network analysis (WGCNA)

A weighted gene co-expression network analysis was performed using the WCGNA package in R (Langfelder and Horvath, 2008). Samples were clustered by *hclust* to filter outliers (h > 15). In order to construct scale-free network, the optimal soft-thresholding power β was defined by picking Soft Threshold function ($\beta = 6$, $R^2 \geq 0.8$). Based on pairwise correlations between genes, genes with similar expression patterns were clustered into a group through a TOM clustering tree according to the dynamic tree cut method, and similar groups were combined into one module. The key hub genes, which were the node of co-expression network, were defined based on the connectivity by the CytoHubba plugin in Cytoscape v3.9.1.

Enrichment analysis

Gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were carried on DEGs (Gene Ontology Consortium, 2021; Kanehisa et al., 2023). Enrichment analysis was performed using the R package "clusterProfiler" (Wu et al., 2021). GO terms and KEGG analysis with corrected p-value < 0.05 were considered to be significantly enriched (Wang et al., 2021).

Real-time quantitative PCR (qRT-PCR) analysis

Total RNA was isolated from zebrafish embryos in each group using TRIzol™ reagent (Ambion, United States). The cDNA was synthesized using the SuperScriptIII (Invitrogen, United States), according to the manufacturer's protocol. qPCR was conducted using SYBR FAST Universal qPCR kit (KAPA, Germany) and ViiA 7 Real-Time PCR System (ABI, United States), as described previously (Wang et al., 2021). Primer sequences were shown in Supplementary Table S1.

Analysis of skeletal muscle structure by birefringence

Zebrafish embryos at 96 hpf were anesthetized with tricaine (0.04%) and embedded in 5% methylcellulose to score the skeletal muscle lesions. Birefringence was imaged under SMZ25 stereo microscope equipped with a DS-Ri2 digital camera (Nikon, Japan), as previously described (Lu et al., 2021). For quantification analysis, 10 somites between the levels of somite 5 to 15 were imaged per embryo.

Tracking of swimming behavior

For locomotion tracking, single zebrafish larvae (treated with or without ESIONP) developed to 120 hpf was placed in individual wells of 24-well cell culture plate containing approximately 500 μL embryo medium. Swimming behavior was monitored at room temperature using a DanioVision system and EthoVision XT 11.5 locomotion tracking software (Noldus Information, Netherlands), according previously described (Lu et al., 2021).

Oxidative stress detection

The oxidative stress and damage caused by ESIONP was detected by measuring ROS production in zebrafish embryo (Zhu et al., 2022). Briefly, embryos at 72 hpf treated with or without ESIONPs were stained with an oxidation-sensitive fluorescent probe dye, dichloro-dihydro-fluorescein diacetate (DCFH-DA) (Beyotime, China) at a final concentration of 20 μg/mL. Stained embryos were incubated at 28°C for 1 h and then washed with PBS. The photos were taken under a fluorescence microscope with a DS-Ri2 digital camera (Nikon, Japan). The fluorescence intensity of embryos was quantified using ImageJ software (NIH, United States).

Apoptosis analysis

To detect apoptotic cells in zebrafish embryos, acridine orange (AO), a fluorescent dye was used. Zebrafish larvae developed to 96 hpf were incubated with 10 μg/mL AO staining solution (Beyotime, China) at 28.5°C for 30 min in the dark, and rinsed with PBS (Zhu et al., 2022). The zebrafish embryos were observed and recorded under a fluorescence microscope with a DS-Ri2 digital camera (Nikon, Japan). The intensity of the fluorescence signal was measured and analyzed using ImageJ software (NIH, United States).

Statistical analysis

Each experiment was repeated at least three times. All data were presented as the mean ± SD. Student's t-test was applied for comparisons among different groups. p-value < 0.05 was considered significant.

Result

Construction of the stage-specific gene co-expression networks via WGCNA

To assess the toxicity of ESIONPs, zebrafish embryos were exposed to different concentrations of ESIONPs (0, 10, 20, 30, 40, 60, 80, and 100 mg/L), the survival rate was counted at 24, 48, 72, 96, and 120 hpf (Supplementary Figure S1A). The LC50 of ESIONPs was determined at 72 hpf (Supplementary Figure S1B). The degrees of malformations are defined in 4 classes, including dorsal bending, shortened body length, yolk sac swelling, cardiac edema, smaller eyes, and head (Supplementary Figure S1C). A concentration of 40 mg/L was chosen for subsequent experiments, to ensure the maximum survival rate of embryos while including diverse morphological abnormalities.

In order to explore the mechanism of toxicity of ESIONPs on zebrafish embryonic development, the DEGs of the control groups and the ESIONPs-exposed groups at 48 hpf, 72 hpf, 96 hpf, 120 hpf were analyzed by RNA-Seq. To determine the relationship between replicates, the samples were clustered using Principal Component Analysis (PCA) and correlation analysis. The PCA showed high repeatability in duplicate samples (Figure 1A), and the Pearson correlation coefficient for each group of replicates also indicated a high repeatability ($|r| \geq 0.8$ for all) (Figure 1B). A total of 32,057 transcripts from 16 samples were fused to construct the co-expression network which was constructed by R packages of WGCNA. Firstly, samples were clustered by *hclust* (h > 5,500), no outlier samples were found in the hierarchical clustering (Figure 1C). In order to build a network with scale-free distribution and preserve the information of DEGs as much as possible, we found the best soft-thresholding powers β (β = 6). The connectivity between genes in the network is relatively high (β = 6, R^2 = 0.813), indicating that the network was scale-free (Figure 1D). Genes were divided into 22 modules (Figure 1E; Supplementary Table S2). In order to study the mechanism of

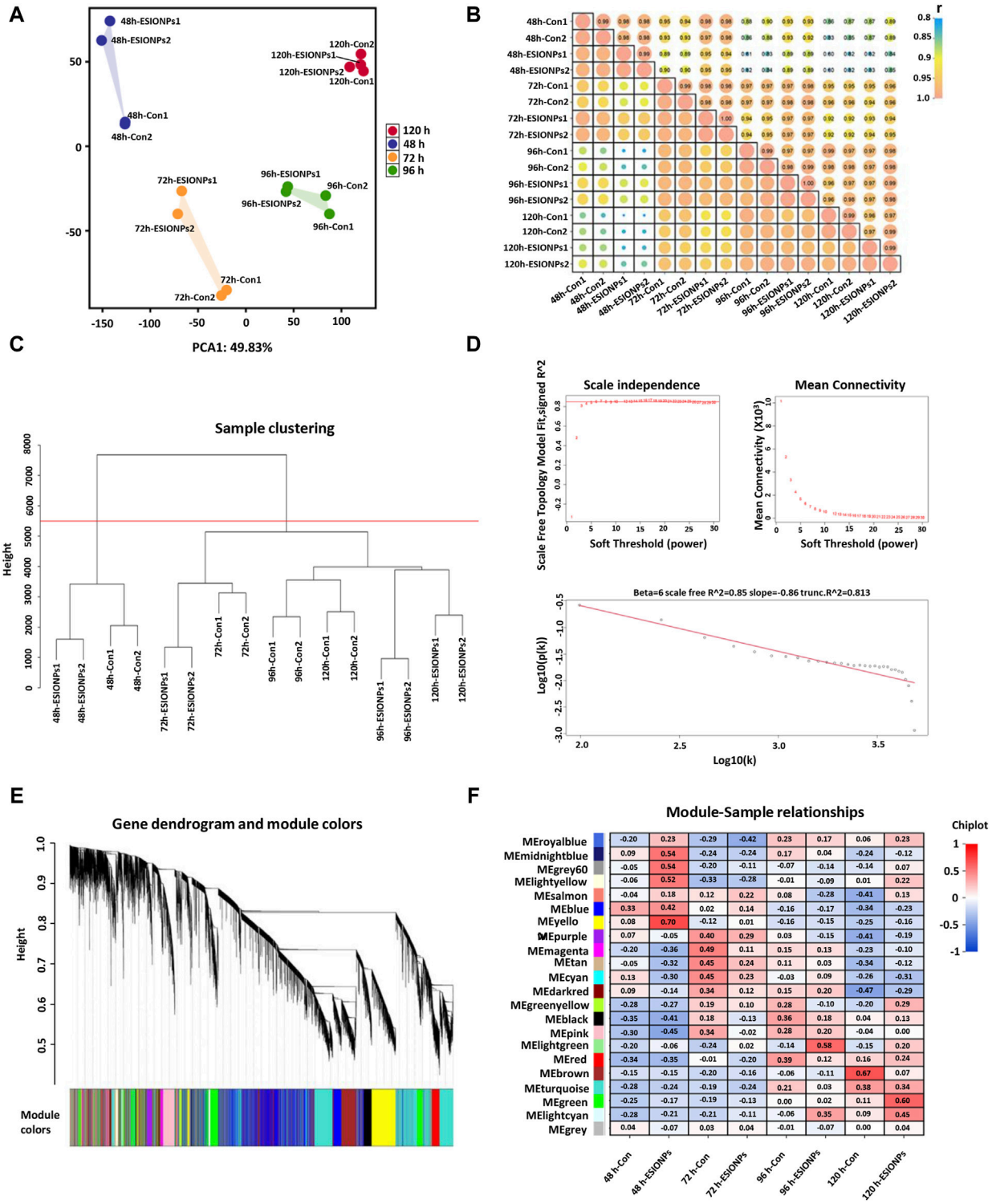


FIGURE 1
Stage-Specific Gene Co-Expression Networks via WGCNA. **(A)** PCA analysis of all RNA sequencing samples. **(B)** Pearson correlation co-efficient analysis of all RNA sequencing samples. **(C)** Hierarchical clustering information of all RNA sequencing samples. **(D)** The determination of soft threshold power, when β = 6, the scale-free network fitting. **(E)** Based on the hierarchical clustering and adjacency dissimilarity, a gene clustering tree diagram of 22 modules was obtained. **(F)** Module-sample relationship, where the horizontal axis represents the samples and the vertical axis represents the modules. The numbers in each grid represent the correlation between the modules and the samples.

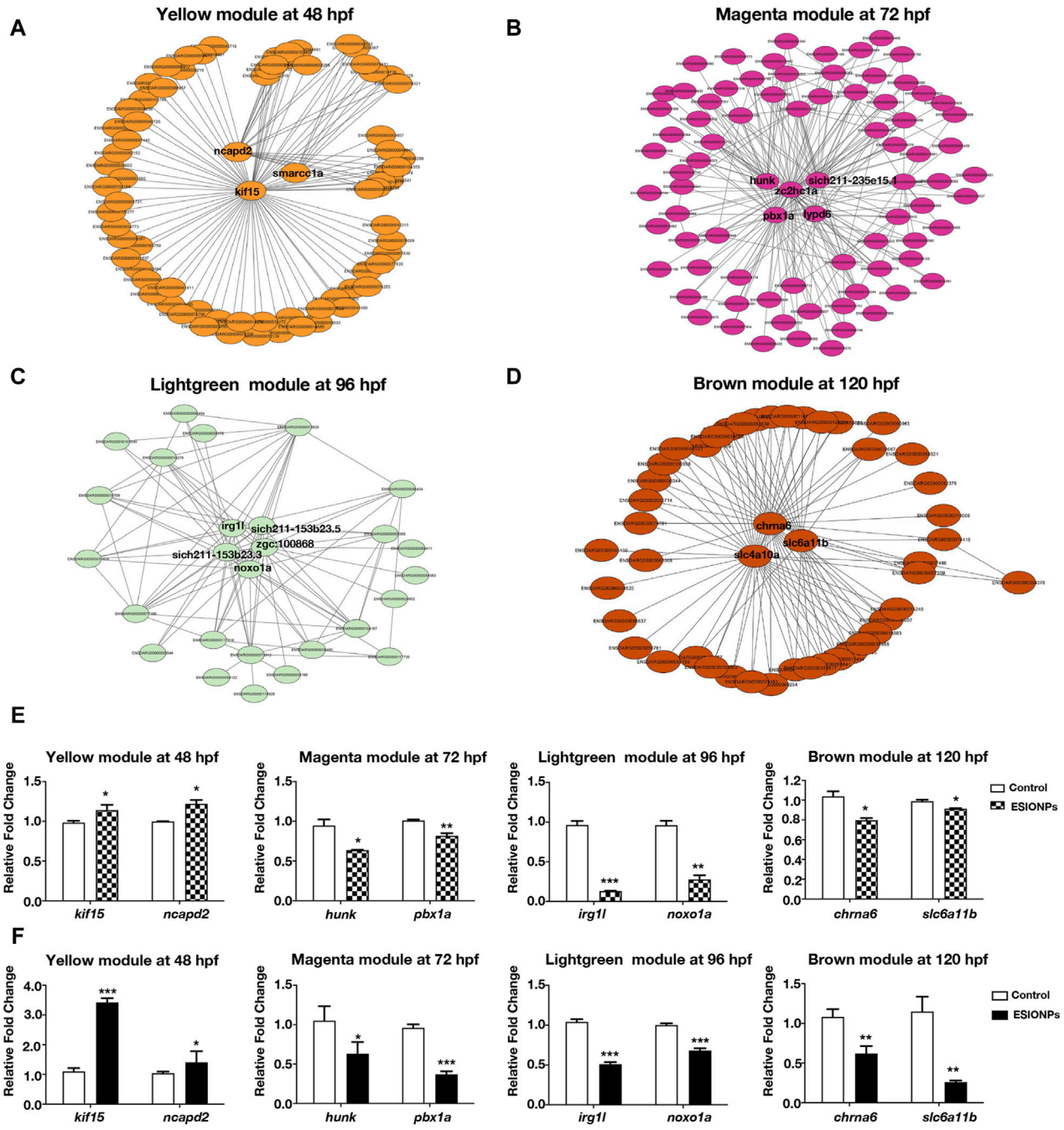


FIGURE 2
Networks of genes in different modules. **(A)** The key hub genes were shown in Yellow Module (*ncapd2*, *smarcc1a*, *kif15*). **(B)** The key hub genes were shown in Magenta Module (*hunk*, *pbx1a*, *zc2hc1a*, *lypd6*, *sich211-235e15.1*). **(C)** The key hub genes were shown in Lightgreen Module (*irg1l*, *sich211-153b23.5*, *noxo1a*, *sich211-153b23.3*, *zgc:100868*). **(D)** The key hub genes were shown in Magenta Module (*chrna6*, *slc4a10a*, *slc6a11b*). **(E,F)** Gene expression of the key hub genes in 4 modules by RNA-seq **(E)** and qRT-PCR **(F)**. Value are normalized to Control group, and represent mean ± SE from three independent experiments, * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001 (Student's t-test).

toxicity of ESIONPs at each development stages, the module with the highest correlation of changes was selected at each stage. The yellow module ($|R^2|$ = 0.62), magenta module ($|R^2|$ = 0.3765), lightgreen module ($|R^2|$ = 0.725), and brown module ($|R^2|$ = 0.6) showed the highest correlation of changes at 48 hpf, 72 hpf, 96 hpf, and 120 hpf, which were used for further research (Figure 1F).

Functional annotations of key modules to each development stages

Genes in key modules were screened out according to the eigengene-based connectivity (kME) values (|kME| > 0.9). There were 61 hub genes in the yellow module, 128 hub genes in the

magenta module, 54 hub genes in the lightgreen module, and 1,157 hub genes in the brown module (Supplementary Table S3).

In the enrichment analysis of the GO pathways, we presented top 10 GO annotations (Supplementary Figures S2A–D; Supplementary Tables S4–S7). The important terms of enrichment in the yellow module (48 hpf) were related to neuron system and muscle development (Supplementary Figure S2A). The terms of related neuron development also mainly were enriched in the magenta module (72 hpf) (Supplementary Figure S2B). The terms of inflammatory were enriched in the lightgreen module (96 hpf, mainly including inflammatory response and chemotaxis of immunocyte (Supplementary Figure S2C). The terms of neuronal signal transmission were enriched in the brown module (120 hpf), including (Supplementary Figure S2D).

In enrichment analysis of KEGG pathways (Supplementary Figures S2E–H; Supplementary Tables S4–S7), yellow module genes (48 hpf) were mainly enriched in embryonic development key signaling pathways, including notch signaling pathway, wnt signaling pathway, hedgehog signaling pathway (Supplementary Figure S2E). The terms of hormone secretion and metabolism were enriched in the magenta module (72 hpf), including pentose phosphate pathway, ubiquitin mediated proteolysis, protein processing in endoplasmic reticulum, GnRH signaling pathway, Endocytosis (Supplementary Figure S2F). Lightgreen module genes (96 hpf) were enriched in metabolism pathways (Biosynthesis of nucleotide sugars, arachidonic acid metabolism, amino sugar and nucleotide sugar metabolism, glycerophospholipid metabolism), necroptosis, C-type lectin receptor signaling pathway, and ferroptosis (Supplementary Figure S2G). Brown module genes (120 hpf) were enriched in pathways related to neuronal signal transmission such as calcium signaling pathway, neuroactive ligand-receptor interaction, cell adhesion molecules, and oxidative phosphorylation (Supplementary Figure S2H).

The enrichment analysis of modules corresponding to each stage of embryonic development by GO and KEGG indicated that ESIONPs might mainly be toxic to the nervous system development, neural conduction, and motor system of zebrafish, and might induce inflammation and ferroptosis in zebrafish embryos.

Hub genes identified in each module by WGCNA

To investigate the mechanism of toxicity of ESIONPs on zebrafish embryos, we next filtered out the hub genes affected by ESIONPs. Networks were constructed to explore relationships among hub genes, which were used as nodes of the scale-free network and had the highest correlation. The top hub genes as the most important nodes in each module were identified and highlighted (Figure 2). There were 3 hub genes in the yellow module: *kif15*, *ncapd2*, *smarcc1a* (Figure 2A), 5 hub genes in the magenta module: *zc2hc1a*, *hunk*, *pbx1a*, *lypd6*, *si:ch211-235e15.1* (Figure 2B), 5 hub genes in the lightgreen module: *irg1l*, *si:ch211-153b23.3*, *si:ch211-153b23.5*, *zgc:100868*, *noxo1a* (Figure 2C), and 3 hub genes in the brown module: *chrna6*, *slc4a10a*, *slc6a11b* (Figure 2D). Although some hub genes (*si:ch211-235e15.1*, *si:ch211-153b23.3*, *si:ch211-153b23.5*, *zgc:100868*) of the magenta and lightgreen modules remain unannotated, the other hub genes of each module could reflect the toxicity of ESIONPs to embryos. For example, genes involved in neuron development (*kif15*, *ncapd2*, *smarcc1a*, *hunk*, *pbx1a*, *lypd6*), neurotransmission (*chrna6*, *slc4a10a*, *slc6a11b*), immune system regulation (*irg1l*), and oxygen stress (*noxo1a*). For each module, two hub genes were selected for validation by qRT-PCR (Figure 2F), which exhibited similar expression trends to the RNA-Seq profiles (Figure 2E). The WGCNA analysis indicated that ESIONPs might have neurotoxicity, which could damage neuron development, nerve conduction, and synaptic transmission. In addition, ESIONPs might also cause ferroptosis in zebrafish embryos.

ESIONPs resulted in neurotoxicity in zebrafish embryos

In *Tg* (*eef1a1l1:EGFP*) transgenic zebrafish embryos, ESIONPs-exposed embryos exhibited significant abnormal development in the nervous system at 72 hpf (Figure 3A), compared with the control embryos. Furthermore, the expression of neuron developmental markers (*pax2a*, *neurog1*, *axin2*) was significantly downregulated in ESIONPs-exposed embryos (Figure 3B). The analysis of the movement track at 120 hpf showed that the movement ability of the ESIONPs-exposed group was significantly weakened (Supplementary Figure S3A), and the expression of the neuromuscular junction and synapse markers (*lrp4*, *musk*, *mpz*) was also downregulated in ESIONPs-exposed embryos (Supplementary Figure S3B). Moreover, the muscle polarization of zebrafish larvae exposed to ESIONPs was significantly reduced (Figure 3C), and the expression of muscle markers (*acta2*, *ttn*, *lpin1*) was also downregulated (Figure 3D). These results showed that ESIONPs not only impaired neuron development, synaptic signal transmission, and neuromuscular junction signal transmission, but also reduced muscle development.

Ferroptosis induced by ESIONPs could lead to oxidative stress, cell apoptosis and inflammatory response in zebrafish embryos

Next, we determined whether ferroptosis occurred by detecting oxidative stress and apoptosis in zebrafish embryos (Dixon et al., 2012). The oxidative stress was significantly increased in the ESIONPs-exposed embryos (Figure 3E), and the expression of oxidative stress markers (*cybb*, *nox1*, *rac2*) was significantly upregulated in ESIONPs-exposed embryos (Figure 3G). The ESIONPs-exposed embryos also displayed significant cell apoptosis (Figure 3F), and the expression of apoptotic markers (*jnk1*, *bcl2a*, *tp53*) was significantly upregulated (Figure 3H). In addition, the expression of inflammatory markers (*il1b*, *il6*, *tnfa*) was also increased significantly (Supplementary Figure S3C). These results revealed that ESIONPs could induce ferroptosis, resulting in oxidative stress, cell apoptosis, and inflammatory response in zebrafish embryos.

Discussion

Iron oxide nanoparticles typically consist of a core of magnetic iron oxide surrounded by a stable coating. Fe_3O_4, Fe_2O_3, and FeO

10.3389/fgene.2024.1402771

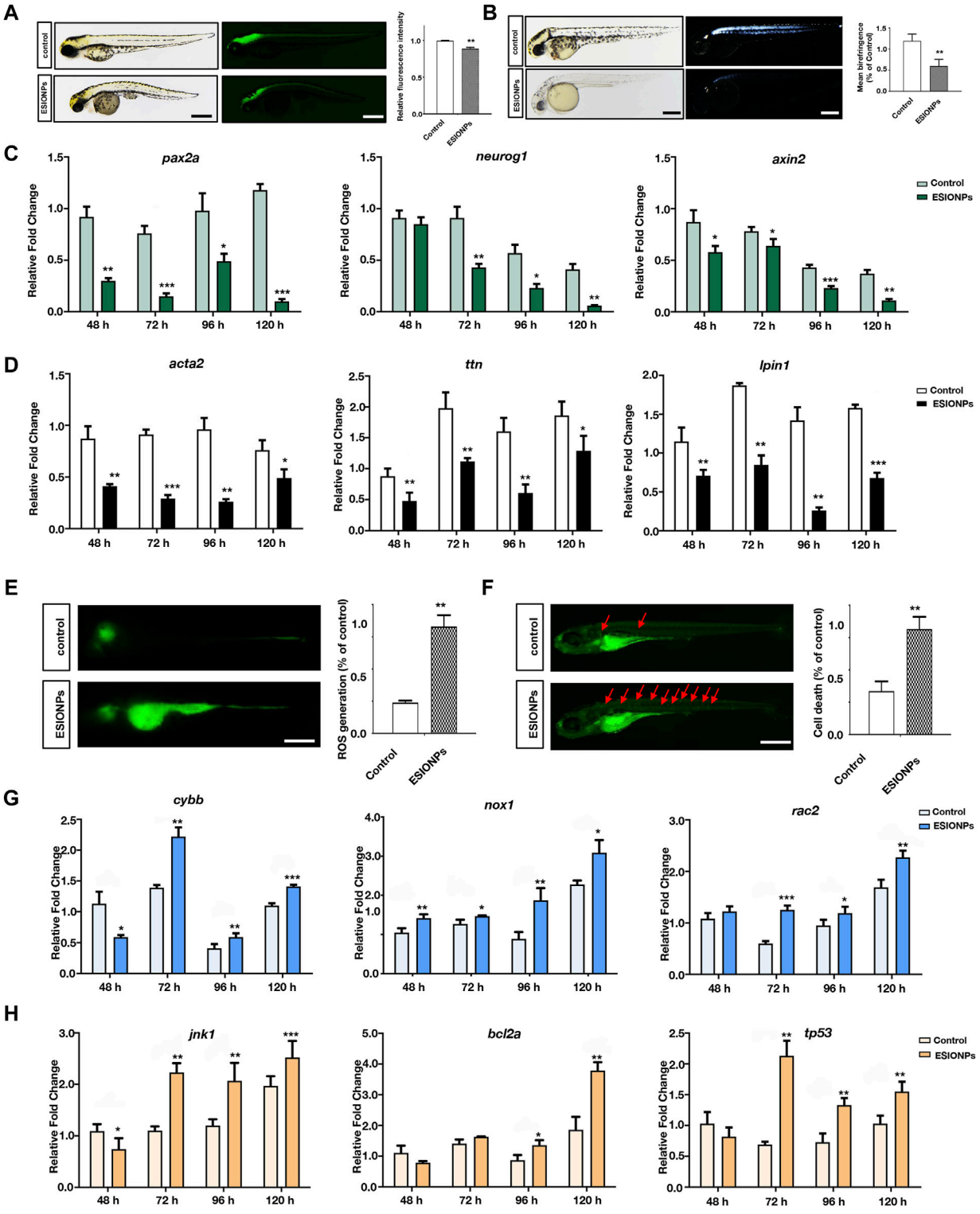


FIGURE 3
The neurotoxicity and Ferroptosis identified in zebrafish embryos. **(A)** Neuron system fluorescence signals of control group and ESIONPs-exposed group in transgenic zebrafish *Tg* (*eef1a1l1:EGFP*). **(B)** Polarized light intensity of control group and ESIONPs-exposed group of zebrafish muscles. **(C)** The expression of neuron-developmental markers (*pax2a*, *neurog1*, and *axin2*) at different developmental stages of zebrafish. **(D)** The expression of muscle markers (*acta2*, *ttn*, *lpin1*) at different developmental stages of zebrafish. **(E)** ROS generation in zebrafish embryos was detected with fluorescent probe DCFH-DA staining. **(F)** The prevalence of apoptosis in zebrafish embryos was detected with fluorescent dye AO staining; red arrows indicate the apoptotic cells. **(G)** The expression of oxidative stress biomarkers (*cybb*, *nox1*, *rac2*) at different developmental stages of zebrafish. **(H)** The expression of apoptosis biomarkers (*jnk1*, *bcl2a*, *tp53*) at different developmental stages of zebrafish. The fluorescence intensity and polarized light intensity was quantified for individual zebrafish using ImageJ analysis. Scale bar, 200 μm. Value are normalized to Control group, and represent mean ± SE from three independent experiments, * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001 (Student's t-test).

nanoparticles containing iron ions of different valences all belong to iron oxide nanoparticles (Laurent et al., 2008). However, the iron ions released from iron oxide nanoparticles are toxic. The iron ions released from iron oxide nanoparticles and can lead to iron accumulation, oxidative stress and protein aggregation in the neural cells (Yarjanli et al., 2017). It is reported that iron overload may decrease the AChE activity in the brains and livers of zebrafish, which are highly susceptible to iron exposure (Sant Anna et al., 2011). Iron overload zebrafish also exhibit dysregulation in metal homeostasis and decreased neurophysiological performance (Hassan and Kwong, 2020). In this study, after exposure to ESIONPs, the neuron development of zebrafish embryos was significantly disrupted, the motor ability mediated by nerve impulses was also reduced, and ferroptosis might be induced. These phenotypes were similar to organ damage caused by iron ions released from iron oxide nanoparticles. Therefore, we assumed that the neurotoxicity of ESIONPs is mainly caused by the release of iron ions into the environment.

The release and accumulation of ESIONPs in the environment significantly endanger water ecosystems, aquatic organisms, and human health. The small size and high surface activity of NPs allow them to persist in aquatic environments, evade conventional water treatments, and accumulate in aquatic organisms, posing risks to the entire ecosystems (Auffan et al., 2009). ESIONPs can disrupt cellular processes, induce oxidative stress, damage structural integrity, and thus affect the health and reproduction of organisms. In humans, ESIONPs might enter the human body through inhalation or skin contact during water-related activities, arising certain risks to respiratory and other organ systems (Mahmoudi et al., 2011). Our results also showed that ESIONPs had significant neurotoxic effects, which might affect neurological function and lead to degenerative changes or behavioral abnormalities. Therefore, it is imperative that future efforts should focus on reducing the toxicity of ESIONPs while preserving their advanced imaging capabilities to improve their biosafety.

Because of the obvious toxicity of iron oxide nanoparticles, which is caused by the release of iron ions, synthetic and coating strategies have been continuously modified to reduce this toxicity. Iron oxide nanoparticles synthesized by Pouteria caimito fruit can significantly reduce cytotoxicity (Veeramani et al., 2022), and Ag also can reduce the toxicity of Fe3O4 in iron oxide nanoparticles synthesizing (Qi et al., 2022). The neurotoxicity of iron oxide nanoparticles in clinical application also can be reduced by Quercetin in conjugated form as supplementation (Bardestani et al., 2021). Recently, an iron nanoparticle (3 nm in diameter) modified with polyethylene glycol-ethoxy-benzyl ligand on the surface ($MnFe_2O_4$-EOB-PEG) was reported that it can substantially reduce the risk of potential neurotoxicity in rabbits, pigs and macaques (Zhang et al., 2023). Thus, employing synthesis methods with low biological toxicity and various coating techniques for iron oxide nanoparticles might be effective ways to reduce their toxicity.

Transcriptome analysis is the main technology used for toxicity investigation (Teeguarden et al., 2014; Zheng et al., 2018; Arsiwala et al., 2022). However, studies on the toxicity of ESIONPs often focus on a single stage, neglecting the dynamics of embryonic development. The toxicity analysis of a single stage results in many potential or critical factors caused by ESIONPs being concealed. To avoid this problem, this study used WGCNA for toxicity analysis at different stages of embryonic development, which reflected the dynamic impact of ESIONPs on the development of zebrafish embryos. Here, key hub genes in yellow, magenta, lightgreen, and brown modules corresponding to the developmental stages of 48 hpf, 72 hpf, 96 hpf, and 120 hpf were identified. Meanwhile, the expression trends of these hub genes were consistent with the neural development, neural signal transformation, and ferroptosis. Hence, dynamic and continuous analysis of the toxicity of nanoparticles during embryonic development could comprehensively identify key hub genes or novel biomarkers.

Iron is considered an important target for neurodegenerative diseases, and ferroptosis is a type of iron dependent cell death (Yan et al., 2021; Li et al., 2023). An increasing number of studies have confirmed that ferroptosis is associated with the pathological changes of neurological diseases such as Alzheimer's disease, Parkinson's disease, and Huntington's disease, mainly manifesting as neuronal cell death, neuronal loss, and synaptic damage (Li et al., 2022). In this study, ESIONPs caused oxidative stress and cell apoptosis in the neuronal system, ultimately leading to movement disorders, similar to human neurological disorders induced by ferroptosis. It suggested that the use of ESIONPs might induce ferroptosis in the human brain, leading to neuronal damage and death, and increasing the probability of neurological disease occurrence and development.

In this study, through WGCNA analysis, we revealed that exposure to ESIONPs could lead to neurodevelopmental abnormalities and ferroptosis. Since ESIONPs can enter the circulatory system and may have an impact on various body organs. Therefore, a comprehensive evaluation of the safe dosage, *in vivo* distribution, and potential toxicity to other organs is still needed to ensure its safety in clinical applications.

Scope statement

With the increasing use of iron oxide nanoparticles as contrast agents in clinical practice, extremely small iron oxide nanoparticles (<5 nm in diameter) (ESIONPs) have been synthesizing and modified for better absorption and imaging. However, the toxicity of IONPs might lead to chronic neurological and motor system diseases, so research on the toxicity of ESIONPs is urgent. Here, we used zebrafish as a model animal to explore the potential toxicity of ESIONPs on embryonic development. By performing RNA-Seq on control and ESIONPs-exposed embryos at 48 hpf, 72 hpf, 96 hpf, and 120 hpf, WGCNA analysis revealed different module corresponding to each embryonic development stage, and key biomarkers were identified in each module. The expression trends of these key biomarkers were further validated by qRT-PCR. Moreover, exposure to ESIONPs might disrupt the neuronal and muscle development of zebrafish, and induced ferroptosis, leading to oxidative stress, cell apoptosis, and inflammatory response in zebrafish larvae. The toxicity study of ESIONPs herein provides certain suggestions for the potential clinical application of ESIONPs.

Data availability statement

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive in National

Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA016266) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

Ethics statement

The animal study was approved by the Experimental Animal Management and Ethics Committee of Northwest University. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

ZL: Data curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. YK: Data curation, Investigation, Project administration, Writing–review and editing. YF: Validation, Writing–review and editing, Visualization. YX: Writing–review and editing, Validation, Resources. BY: Writing–review and editing, Methodology, Software. HZ: Formal Analysis, Writing–review and editing. JT: Conceptualization, Funding acquisition, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China, grant number 32170618.

Acknowledgments

We would like to thank Prof. Haiming Fan for providing nanomaterials.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1402771/full#supplementary-material

References

Arsiwala, T., Vogt, A. S., Barton, A. E., Manolova, V., Funk, F., Flühmann, B., et al. (2022). Kupffer cells and Blood Monocytes Orchestrate the clearance of iron-Carbohydrate nanoparticles from Serum. *Int. J. Mol. Sci.* 23 (5), 2666. doi:10.3390/ijms23052666

Auffan, M., Rose, J., Bottero, J. Y., Lowry, G. V., Jolivet, J. P., and Wiesner, M. R. (2009). Towards a definition of inorganic nanoparticles from an environmental, health and safety perspective. *Nat. Nanotechnol.* 4 (10), 634–641. doi:10.1038/nnano.2009.242

Bardestani, A., Ebrahimpour, S., Esmaeili, A., and Esmaeili, A. (2021). Quercetin attenuates neurotoxicity induced by iron oxide nanoparticles. *J. nanobiotechnology* 19 (1), 327. doi:10.1186/s12951-021-01059-0

Calderón-Garcidueñas, L., González-Maciel, A., Reynoso-Robles, R., Silva-Pereyra, H. G., Torres-Jardón, R., Brito-Aguilar, R., et al. (2022). Environmentally toxic Solid nanoparticles in Noradrenergic and Dopaminergic Nuclei and Cerebellum of Metropolitan Mexico City Children and Young Adults with neural Quadruple Misfolded protein Pathologies and high exposures to Nano Particulate Matter. *Toxics* 10 (4), 164. doi:10.3390/toxics10040164

Cao, Y., Mao, Z., He, Y., Kuang, Y., Liu, M., Zhou, Y., et al. (2020). Extremely small iron oxide nanoparticle-Encapsulated Nanogels as a Glutathione-Responsive T1 contrast agent for Tumor-Targeted magnetic resonance imaging. *ACS Appl. Mater Interfaces* 12 (24), 26973–26981. doi:10.1021/acsami.0c07288

Chemello, G., Randazzo, B., Zarantoniello, M., Fifi, A. P., Aversa, S., Ballarin, C., et al. (2019). Safety assessment of antibiotic administration by magnetic nanoparticles in *in vitro* zebrafish liver and intestine cultures. *Toxicol. Pharmacol. CBP* 224, 108559. doi:10.1016/j.cbpc.2019.108559

Chen, T., Chen, X., Zhang, S., Zhu, J., Tang, B., Wang, A., et al. (2021). The genome sequence archive aamily: toward explosive data growth and diverse data types. *Genom. Proteom. Bioinform.* 19 (4), 578–583. doi:10.1016/j.gpb.2021.08.001

CNCB-NGDC Members and Partners (2024). Database resources of the national genomics data center, China national Center for Bioinformation in 2024. *Nucleic Acids Res.* 52 (D1), D18–D32. doi:10.1093/nar/gkad1078

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 (5), 1060–1072. doi:10.1016/j.cell.2012.03.042

Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic acids Res.* 49 (D1), D325–D334. doi:10.1093/nar/gkaa1113

Groult, H., Carregal-Romero, S., Castejón, D., Azkargorta, M., Miguel-Coello, A. B., Pulagam, K. R., et al. (2021). Heparin length in the coating of extremely small iron oxide nanoparticles regulates *in vivo* theranostic applications. *Nanoscale* 13 (2), 842–861. PMID: 33351869. doi:10.1039/d0nr06378a

Han, J., Tian, Y., Wang, M., Li, Y., Yin, J., Qu, W., et al. (2022). Proteomics unite traditional toxicological assessment methods to evaluate the toxicity of iron oxide nanoparticles. *Front. Pharmacol.* 13, 1011065. doi:10.3389/fphar.2022.1011065

Hassan, A. T., and Kwong, R. W. M. (2020). The neurophysiological effects of iron in early life stages of zebrafish. *Environ. Pollut. Barking, Essex 1987* 267, 115625. doi:10.1016/j.envpol.2020.115625

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic acids Res.* 51 (D1), D587–D592. doi:10.1093/nar/gkac963

Kim, B. H., Lee, N., Kim, H., An, K., Park, Y. I., Choi, Y., et al. (2011). Large-scale synthesis of uniform and extremely small-sized iron oxide nanoparticles for high-resolution T1 magnetic resonance imaging contrast agents. *J. Am. Chem. Soc.* 133 (32), 12624–12631. doi:10.1021/ja203340u

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Laurent, S., Forge, D., Port, M., Roch, A., Robic, C., Vander Elst, L., et al. (2008). Magnetic iron oxide nanoparticles: synthesis, stabilization, vectorization, physicochemical characterizations, and biological applications. *Chem. Rev.* 108 (6), 2064–2110. doi:10.1021/cr068445e

Lee, N., Yoo, D., Ling, D., Cho, M. H., Hyeon, T., and Cheon, J. (2015). Iron oxide based nanoparticles for Multimodal imaging and Magnetoresponsive Therapy. *Chem. Rev.* 115 (19), 10637–10689. doi:10.1021/acs.chemrev.5b00112

Li, J., Jia, B., Cheng, Y., Song, Y., Li, Q., and Luo, C. (2022). Targeting molecular Mediators of ferroptosis and oxidative stress for neurological disorders. *Oxidative Med. Cell. Longev.* 2022, 3999083. doi:10.1155/2022/3999083

Li, X., Wang, X., Huang, B., and Huang, R. (2023). Sennoside A restrains TRAF6 level to modulate ferroptosis, inflammation and cognitive impairment in aging mice with Alzheimer's Disease. *Int. Immunopharmacol.* 120, 110290. doi:10.1016/j.intimp.2023.110290

Lu, S., Lyu, Z., Wang, Z., Kou, Y., Liu, C., Li, S., et al. (2021). Lipin 1 deficiency causes adult-onset myasthenia with motor neuron dysfunction in humans and neuromuscular junction defects in zebrafish. *Theranostics* 11 (6), 2788–2805. doi:10.7150/thno.53330

Mahmoudi, M., Sant, S., Wang, B., Laurent, S., and Sen, T. (2011). Superparamagnetic iron oxide nanoparticles (SPIONs): development, surface modification and applications in chemotherapy. *Adv. drug Deliv. Rev.* 63 (1-2), 24–46. doi:10.1016/j.addr.2010.05.006

Mishra, S. K., Herman, P., Crair, M., Constable, R. T., Walsh, J. J., Akif, A., et al. (2022). Fluorescently-tagged magnetic protein nanoparticles for high-resolution optical and ultra-high field magnetic resonance dual-modal cerebral angiography. *Nanoscale* 14 (47), 17770–17788. doi:10.1039/d2nr04878g

Mohammadinejad, R., Moosavi, M. A., Tavakol, S., Vardar, D. Ö., Hosseini, A., Rahmati, M., et al. (2019). Necrotic, apoptotic and autophagic cell fates triggered by nanoparticles. *Autophagy* 15 (1), 4–33. doi:10.1080/15548627.2018.1509171

Pereira, A. C., Gonçalves, B. B., Brito, R. D. S., Vieira, L. G., Lima, E. C. O., and Rocha, T. L. (2020). Comparative developmental toxicity of iron oxide nanoparticles and ferric chloride to zebrafish Danio rerio after static and semi-static exposure. *Chemosphere* 254, 126792. doi:10.1016/j.chemosphere.2020.126792

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33 (3), 290–295. doi:10.1038/nbt.3122

Pitt, J. A., Kozal, J. S., Jayasundara, N., Massarsky, A., Trevisan, R., Geitner, N., et al. (2018). Uptake, tissue distribution, and toxicity of polystyrene nanoparticles in developing zebrafish Danio rerio. *Aquat. Toxicol. Amst. Neth.* 194, 185–194. doi:10.1016/j.aquatox.2017.11.017

Qi, J., Zhang, J., Jia, H., Guo, X., Yue, Y., Yuan, Y., et al. (2022). Synthesis of silver/Fe3O4@chitosan@polyvinyl alcohol magnetic nanoparticles as an antibacterial agent for accelerating wound healing. *Int. J. Biol. Macromol.* 221, 1404–1414. doi:10.1016/j.ijbiomac.2022.09.030

Qiu, W., Liu, S., Yang, F., Dong, P., Yang, M., Wong, M., et al. (2019). Metabolism disruption analysis of zebrafish larvae in response to BPA and BPA analogs based on RNA-Seq technique. *Ecotoxicol. Environ. Saf.* 174, 181–188. doi:10.1016/j.ecoenv.2019.01.126

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Sant Anna, M. C., Soares, V. M., Seibt, K. J., Ghisleni, G., Rico, E. P., Rosemberg, D. B., et al. (2011). Iron exposure modifies acetylcholinesterase activity in zebrafish Danio rerio tissues: distinct susceptibility of tissues to iron overload. *Fish Physiology and Biochem.* 37 (3), 573–581. doi:10.1007/s10695-010-9459-7

Schütz, C. A., Staedler, D., Crosbie-Staunton, K., Movia, D., Chapuis Bernasconi, C., Kenzaoui, B. H., et al. (2014). Differential stress reaction of human colon cells to oleic-acid-stabilized and unstabilized ultrasmall iron oxide nanoparticles. *Int. J. nanomedicine* 9, 3481–3498. doi:10.2147/IJN.S65082

Shen, Z., Wu, A., and Chen, X. (2017). Iron oxide nanoparticle based contrast agents for magnetic resonance imaging. *Mol. Pharm.* 14 (5), 1352–1364. doi:10.1021/acs.molpharmaceut.6b00839

Teeguarden, J. G., Mikheev, V. B., Minard, K. R., Forsythe, W. C., Wang, W., Sharma, G., et al. (2014). Comparative iron oxide nanoparticle cellular dosimetry and response in mice by the inhalation and liquid cell culture exposure routes. *Part. fibre Toxicol.* 11, 46. doi:10.1186/s12989-014-0046-4

Thirumurthi, N. A., Raghunath, A., Balasubramanian, S., and Perumal, E. (2022). Evaluation of maghemite nanoparticles-induced developmental toxicity and oxidative stress in zebrafish embryos/larvae. *Biol. trace Elem. Res.* 200 (5), 2349–2364. doi:10.1007/s12011-021-02830-y

Tian, J., Shao, J., Liu, C., Hou, H. Y., Chou, C. W., Shboul, M., et al. (2019). Deficiency of lrp4 in zebrafish and human LRP4 mutation induce aberrant activation of Jagged-Notch signaling in fin and limb development. *Cell. Mol. life Sci. CMLS* 76 (1), 163–178. doi:10.1007/s00018-018-2928-3

Veeramani, C., El Newehy, A. S., Alsaif, M. A., and Al-Numair, K. S. (2022). Vitamin A- and C-rich Pouteria camito fruit derived superparamagnetic nanoparticles synthesis, characterization, and their cytotoxicity. *Afr. health Sci.* 22 (1), 673–680. doi:10.4314/ahs.v22i1.78

Wang, Z., Liu, P., Hu, M., Lu, S., Lyu, Z., Kou, Y., et al. (2021). Naoxintong restores ischemia injury and inhibits thrombosis via COX2-VEGF/NFκB signaling. *J. Ethnopharmacol.* 270, 113809. doi:10.1016/j.jep.2021.113809

Wu, J., Ding, T., and Sun, J. (2013). Neurotoxic potential of iron oxide nanoparticles in the rat brain striatum and hippocampus. *Neurotoxicology* 34, 243–253. doi:10.1016/j.neuro.2012.09.006

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov. Camb. Mass.* 2 (3), 100141. doi:10.1016/j.mtbio.2021.100141

Yan, H. F., Zou, T., Tuo, Q. Z., Xu, S., Li, H., Belaidi, A. A., et al. (2021). Ferroptosis: mechanisms and links with diseases. *Signal Transduct. Target. Ther.* 6 (1), 49. doi:10.1038/s41392-020-00428-9

Yang, F., Qiu, W., Li, R., Hu, J., Luo, S., Zhang, T., et al. (2018). Genome-wide identification of the interactions between key genes and pathways provide new insights into the toxicity of bisphenol F and S during early development in zebrafish. *Chemosphere* 213, 559–567. doi:10.1016/j.chemosphere.2018.09.133

Yarjanli, Z., Ghaedi, K., Esmaeili, A., Rahgozar, S., and Zarrabi, A. (2017). Iron oxide nanoparticles may damage to the neural tissue through iron accumulation, oxidative stress, and protein aggregation. *BMC Neurosci.* 18 (1), 51. doi:10.1186/s12868-017-0369-9

Zhang, C., Huang, W., Huang, C., Zhou, C., Tang, Y., Wei, W., et al. (2022). VHPKQHR Peptide modified ultrasmall Paramagnetic iron oxide nanoparticles targeting Rheumatoid Arthritis for T_1-weighted magnetic resonance imaging. *Front. Bioeng. Biotechnol.* 10, 821256. doi:10.3389/fbioe.2022.821256

Zhang, H., Guo, Y., Jiao, J., Qiu, Y., Miao, Y., He, Y., et al. (2023). A hepatocyte-targeting nanoparticle for enhanced hepatobiliary magnetic resonance imaging. *Nat. Biomed. Eng.* 7 (3), 221–235. doi:10.1038/s41551-022-00975-2

Zheng, M., Lu, J., and Zhao, D. (2018). Toxicity and transcriptome sequencing (RNA-seq) Analyses of adult zebrafish in response to exposure Carboxymethyl Cellulose stabilized iron Sulfide nanoparticles. *Sci. Rep.* 8 (1), 8083. doi:10.1038/s41598-018-26499-x

Zhu, H., Wang, Z., Wang, W., Lu, Y., He, Y. W., and Tian, J. (2022). Bacterial *quorum*-Sensing signal DSF inhibits LPS-induced inflammations by Suppressing Toll-like receptor signaling and Preventing Lysosome-mediated apoptosis in zebrafish. *Int. J. Mol. Sci.* 23 (13), 7110. doi:10.3390/ijms23137110

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 08 August 2024
DOI 10.3389/fgene.2024.1434681

OPEN ACCESS

EDITED BY
Maritha J. Kotze,
Tygerberg Hospital, South Africa

REVIEWED BY
Liane S. Canas,
King's College London, United Kingdom
Iván Galván-Femenía,
Institute for Research in Biomedicine, Spain

*CORRESPONDENCE
Yuriy Lvovich Orlov,
orlov@d-health.institute
Olga Yurievna Bushueva,
olga.bushueva@inbox.ru

RECEIVED 18 May 2024
ACCEPTED 29 July 2024
PUBLISHED 08 August 2024

CITATION
Loktionov AV, Kobzeva KA, Karpenko AR, Sergeeva VA, Orlov YL and Bushueva OY (2024) GWAS-significant loci and severe COVID-19: analysis of associations, link with thromboinflammation syndrome, gene-gene, and gene-environmental interactions. *Front. Genet.* 15:1434681. doi: 10.3389/fgene.2024.1434681

GWAS-significant loci and severe COVID-19: analysis of associations, link with thromboinflammation syndrome, gene-gene, and gene-environmental interactions

Alexey Valerevich Loktionov[1,2], Ksenia Andreevna Kobzeva[2], Andrey Romanovich Karpenko[1,2], Vera Alexeevna Sergeeva[1], Yuriy Lvovich Orlov[3]* and Olga Yurievna Bushueva[2,4]*

[1]Department of Anesthesia and Critical Care, Institute of Continuing Education, Kursk State Medical University, Kursk, Russia, [2]Laboratory of Genomic Research, Research Institute for Genetic and Molecular Epidemiology, Kursk State Medical University, Kursk, Russia, [3]Institute of Biodesign and Complex Systems Modeling, Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia, [4]Department of Biology, Medical Genetics and Ecology, Kursk State Medical University, Kursk, Russia

Objective: The aim of this study was to replicate associations of GWAS-significant loci with severe COVID-19 in the population of Central Russia, to investigate associations of the SNPs with thromboinflammation parameters, to analyze gene-gene and gene-environmental interactions.

Materials and Methods: DNA samples from 798 unrelated Caucasian subjects from Central Russia (199 hospitalized COVID-19 patients and 599 controls with a mild or asymptomatic course of COVID-19) were genotyped using probe-based polymerase chain reaction for 10 GWAS-significant SNPs: rs143334143 *CCHCR1*, rs111837807 *CCHCR1*, rs17078346 *SLC6A20-LLZTFL1*, rs17713054 *SLC6A20-LLZTFL1*, rs7949972 *ELF5*, rs61882275 *ELF5*, rs12585036 *ATP11A*, rs67579710 *THBS3, THBS3-AS1*, rs12610495 *DPP9*, rs9636867 *IFNAR2*.

Results: SNP rs17713054 *SLC6A20-LZTFL1* was associated with increased risk of severe COVID-19 in the entire group (risk allele A, OR = 1.78, 95% CI = 1.22–2.6, $p = 0.003$), obese individuals (OR = 2.31, 95% CI = 1.52–3.5, $p = 0.0002$, ($p_{bonf} = 0.0004$)), patients with low fruit and vegetable intake (OR = 1.72, 95% CI = 1.15–2.58, $p = 0.01$, ($p_{bonf} = 0.02$)), low physical activity (OR = 1.93, 95% CI = 1.26–2.94, $p = 0.0035$, ($p_{bonf} = 0.007$)), and nonsmokers (OR = 1.65, 95% CI = 1.11–2.46, $p = 0.02$). This SNP correlated with increased BMI ($p = 0.006$) and worsened thrombodynamic parameters (maximum optical density of the formed clot, D ($p = 0.02$), delayed appearance of spontaneous clots, Tsp ($p = 0.02$), clot size 30 min after coagulation activation, CS ($p = 0.036$)). SNP rs17078346 *SLC6A20-LZTFL1* was linked with increased BMI ($p = 0.01$) and severe COVID-19 in obese individuals (risk allele C, OR = 1.72, 95% CI = 1.15–2.58, $p = 0.01$, ($p_{bonf} = 0.02$)). SNP rs12610495 *DPP9* was associated with increased BMI ($p = 0.01$), severe COVID-19 in obese patients (risk allele G, OR = 1.48, 95% CI = 1.09–2.01, $p = 0.01$, ($p_{bonf} = 0.02$)), and worsened thrombodynamic parameters (time to the start of clot growth, Tlag ($p = 0.01$)). For rs7949972 *ELF5*, a protective effect against severe COVID-19 was observed in

non-obese patients (effect allele T, OR = 0.67, 95% CI = 0.47–0.95, p = 0.02, (p_{bonf} = 0.04)), improving thrombodynamic parameters (CS (p = 0.02), stationary spatial clot growth rates, Vst (p = 0.02)). Finally, rs12585036 *ATP11A* exhibited a protective effect against severe COVID-19 in males (protective allele A, OR = 0.51, 95% CI = 0.32–0.83, p = 0.004). SNPs rs67579710 *THBS3*, *THBS3-AS1*, rs17713054 *SLC6A20-LZTFL1*, rs7949972 *ELF5*, rs9636867 *IFNAR2*—were involved in two or more of the most significant G×G interactions ($p_{perm} \leq 0.01$). The pairwise combination rs67579710 *THBS3*, *THBS3-AS1* × rs17713054 *SLC6A20-LZTFL1* was a priority in determining susceptibility to severe COVID-19 (it was included in four of the top five most significant SNP-SNP interaction models).

Conclusion: Overall, this study represents a comprehensive molecular-genetic and bioinformatics analysis of the involvement of GWAS-significant loci in the molecular mechanisms of severe COVID-19, gene-gene and gene-environmental interactions, and provides evidence of their relationship with thromboinflammation parameters in patients hospitalized in intensive care units.

KEYWORDS

chronic diseases, genotyping, COVID-19, GWAS, thromboinflammation syndrome, rs17713054, rs17078346, rs12610495

1 Introduction

The emergence of coronavirus disease 2019 (COVID-19) at the close of 2019 brought forth an array of symptoms and outcomes stemming from Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Globally, the case fatality rate of COVID-19 ranges from 1% to 17%. Various factors, like the size of the population tested, demographic characteristics, ethnicity, the effectiveness of healthcare systems, and virus variants can affect the rates of mortality from COVID-19 (https://coronavirus.jhu.edu/data/mortality (accessed 18 March 2024)). However, the most common cause of death from COVID-19 is severe disease, manifested by immune dysregulation and the onset of a cytokine storm (CS) (Kim et al., 2021), characterized by a rapid surge in proinflammatory cytokines and other markers of inflammation. This hyperinflammation leads to coagulopathies, oxidative stress, organ failure, and ultimately, mortality (Silva et al., 2023). Hypercoagulation and micro-clot formation are critical factors in the molecular pathogenesis of COVID-19, contributing significantly to its complications and adverse outcomes (Pretorius et al., 2020). Furthermore, we are becoming increasingly aware of COVID-19 long-term consequences on various organ systems, including the pulmonary, cardiovascular, hematologic, renal, central nervous system, gastrointestinal, and psychosocial manifestations (Joshee et al., 2022; Ma et al., 2022). This growing comprehension underscores the imperative to delve deeper into the understanding of COVID-19.

Understanding why some individuals experience asymptomatic or mild courses while others face intensive care unit (ICU) admissions with severe organ failure and mortality remains a critical challenge and the subject of much research worldwide (Carvalho et al., 2023; Collins et al., 2023).

To date, it is known that lifestyle factors such as fruit and vegetable consumption and physical activity significantly influence the severity of COVID-19 (Yedjou et al., 2021; Tadbir Vajargah et al., 2022; Tavakol et al., 2023). However, host genetic factors play no less a significant role, as evidenced by findings from molecular-genetic studies. Genes such as *SLC6A20*, *LZTFL1*, *IFNAR2*, *DPP9*, *CCHCR1*, *ELF5*, *ATP11A* and *THBS3* have been identified as potentially contributing to severe COVID-19 and hospitalization in genome-wide association studies (Severe Covid-19 GWAS Group et al., 2020; Lee et al., 2021; Horowitz et al., 2022; Kousathanas et al., 2022; Pairo-Castineira et al., 2021). Many of the genetic variants identified by GWAS have been replicated in different populations around the world, demonstrating their high predictive value for the risk of severe COVID-19 (Rescenko et al., 2021; Garg et al., 2024).

Despite the wealth of genetic data, there is a significant lack of research worldwide on the relationship between genetic variants and the severity of thromboinflammatory syndrome in COVID-19 patients, as well as intergenic interactions, interactions between genetic variants and environmental factors that could either mitigate or exacerbate the impact of genetic variants on the severity of the disease.

Therefore, the aim of this pilot study was to i) investigate the association between common single nucleotide polymorphisms identified by GWAS and the risk of severe COVID-19 in a Russian population; ii) investigate the most significant gene-gene interactions associated with severe COVID-19; iii) evaluate the joint influence of polymorphisms and environmental risk factors on disease susceptibility; and iv) find out how COVID-19 GWAS loci influence the features of the clinical manifestations of the disease, including thrombodynamic parameters.

2 Materials and methods

2.1 Study design

The study's fundamental structure, along with the materials and tools employed, are outlined in Figure 1.

2.2 Study participants

The study included 798 unrelated individuals from Central Russia, comprising 199 hospitalized COVID-19 patients and 599 patients of the control group. The Ethical Review Committee

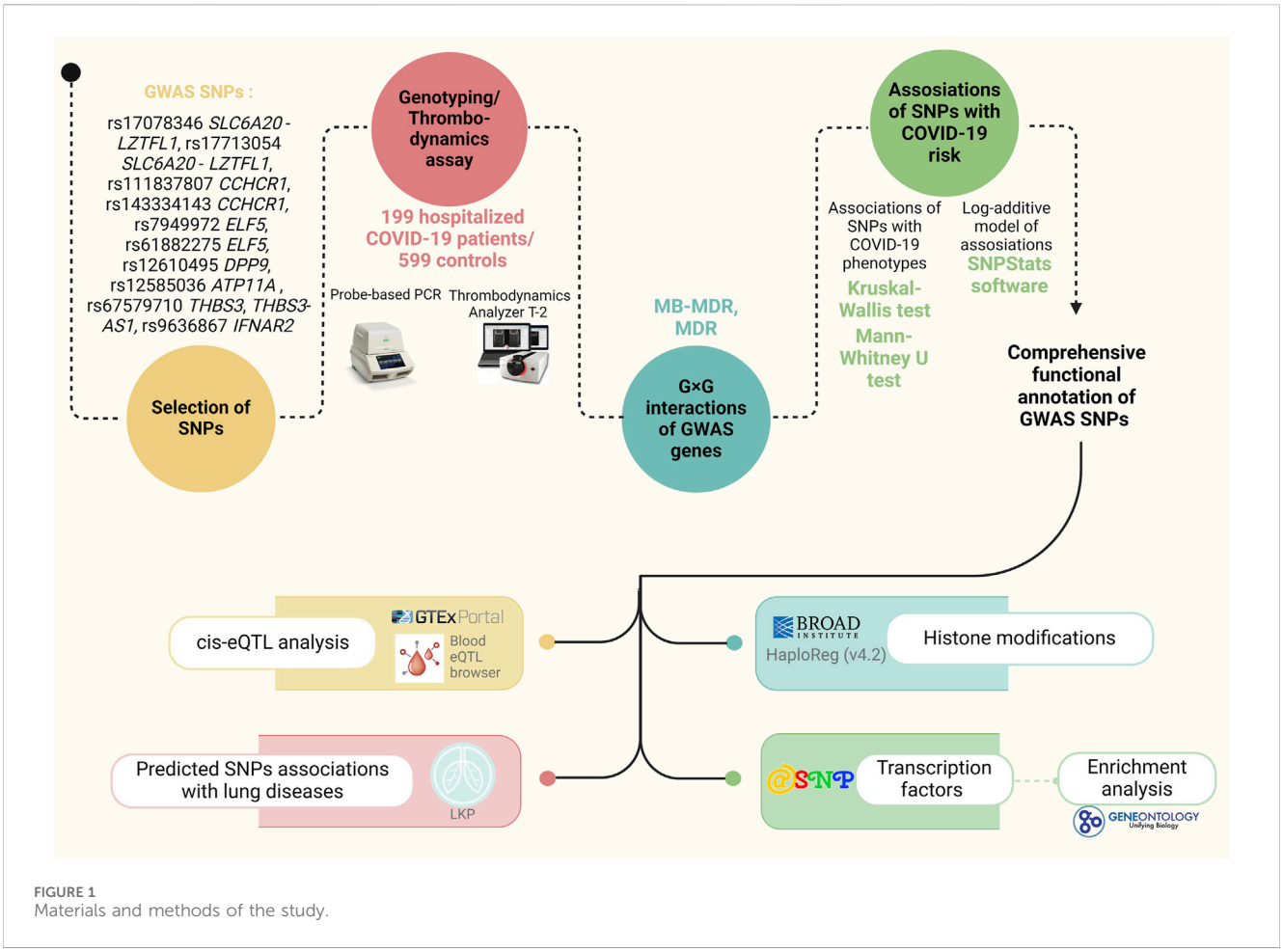


FIGURE 1
Materials and methods of the study.

of Kursk State Medical University approved the study protocol (protocol №1 from 11 January 2022), and all participants provided written informed consent. The patients were enrolled in the study during the COVID-19 pandemic from 2020 to 2022 at the intensive care units (ICU) of Kursk Regional Hospital №6 and Kursk Regional Tuberculosis Dispensary. All patients had a PCR-confirmed diagnosis of COVID-19. The control group consisted of healthy volunteers from Biobank of Research Institute for Genetic and Molecular Epidemiology (Bushueva O. et al., 2015; OYu et al., 2015) who had mild or asymptomatic COVID-19 and did not need ICU admission (Bushueva, 2020; Kobzeva et al., 2022; Belykh et al., 2023). Supplementary Table S1 provides the baseline and clinical characteristics of the study cohort.

In accordance with WHO guidelines (Amine et al., 2003), low fruit and vegetable consumption was defined as consuming less than 400 g per day. Adequate consumption of fresh vegetables and fruits was defined as consuming 400 g or more, equivalent to 3-4 servings per day, excluding starchy tubers like potatoes. Insufficient physical activity was characterized by engaging in less than 180 min per week of moderate to vigorous physical activities. This encompassed various forms of exercise, including leisure activities such as walking and running as well as fitness club exercises like treadmill running, aerobics, or resistance training. Obesity is assessed using the Body Mass Index (BMI), a measurement based on a person's height and weight. A BMI of 30 or higher is generally considered indicative of obesity.

2.3 Selection of genes and polymorphisms

For this study, we selected SNPs from the largest GWAS meta-analysis of severe COVID-19 (top 20 SNPs with *p*-level of significance of $\leq 1 \times 10^{-20}$) (Pairo-Castineira et al., 2023). Then, SNPs with a minor allele frequency <0.05 were excluded from the analysis, as well as loci for which was unable to design probes for TaqMan-based-PCR (low CG composition, presence of GC clamps, runs of identical nucleotides). In total, 10 SNPs were included in the genotyping: rs143334143 *CCHCR1* (chr6:31153649 (GRCh38)), rs111837807 *CCHCR1* (chr6:31153455 (GRCh38)), rs17078346 *SLC6A20-LZTFL1*(chr3:45804256 (GRCh38)), rs17713054 *SLC6A20-LZTFL1* (chr3:45818159 (GRCh38)), rs7949972 *ELF5* (chr11:34480495 (GRCh38)), rs61882275 *ELF5* (chr11:34482745 (GRCh38)), rs12585036 *ATP11A* (chr13:112881427 (GRCh38)), rs67579710 *THBS3*, *THBS3-AS1* (chr1: 155203736 (GRCh38)), rs12610495 *DPP9* (chr19:4717660 (GRCh38)), rs9636867 *IFNAR2* (chr21:33639 (GRCh38)).

2.4 Genetic analysis

The Laboratory of Genomic Research at the Research Institute for Genetic and Molecular Epidemiology of Kursk State Medical University (Kursk, Russia) performed genotyping. Up to 5 mL of venous blood from each participant was collected from a cubital

vein, put into EDTA-coated tubes, and kept at −20 C until it was processed. Defrosted blood samples were used to extract genomic DNA using the standard methods of phenol/chloroform extraction and ethanol precipitation. The purity, quality, and concentration of the extracted DNA samples were assessed using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, United States).

Genotyping of the SNPs was performed using allele-specific probe-based polymerase chain reaction (PCR) according to the protocols designed in the Laboratory of Genomic Research at the Research Institute for Genetic and Molecular Epidemiology of Kursk State Medical University. The Primer3 software was used for primer design (Koressaar and Remm, 2007). A real-time PCR procedure was performed in a 25 μL reaction solution containing 1.5 units of Hot Start Taq DNA polymerase (Biolabmix, Novosibirsk, Russia), approximately 10 ng of DNA, and the following concentrations of reagents: 0.25 μM of each primer; 0.1 μM of each probe; 250 μM of each dNTP; 3 mM $MgCl_2$ for rs7949972, 3.5 mM $MgCl_2$ for rs61882275, 2 mM $MgCl_2$ for rs12610495, and 2.5 mM $MgCl_2$ for the remining SNPs; 1xPCR buffer (67 mM Tris-HCl, pH 8.8, 16.6 mM $(NH_4)_2SO_4$, 0.01% Tween-20). The PCR procedure comprised an initial denaturation for 10 min at 95°C, followed by 39 cycles of 92 °C for 30 s and 57 °C, 59 °C, 60 °C, 61 °C, 62 °C, 63 °C, 65 °C, 66 °C for 1 min (for rs12610495 *DPP9*, rs17078346 *SLC6A20-LZTFL1*, rs17713054 *SLC6A20-LZTFL1*, rs111837807 *CCHCR1*, rs9636867 *IFNAR2*, rs143334143 *CCHCR1* and rs7949972 *ELF5*, rs12585036 *ATP11A* and rs61882275 *ELF5*, rs67579710 *THBS3*, *THBS3-AS1*, respectively). 10% of the DNA samples were genotyped twice, blinded to the case-control status, in order to assure quality control. Over 99% of the data were concordant. Due to the Hardy-Weinberg equilibrium deviation in the control group for SNP rs12610495 *DPP9*, all locus samples underwent re-genotyping. The results were entirely consistent (100%) with the initial genotypes.

2.5 Thrombodynamics analysis

The analysis utilized venous blood samples obtained from the peripheral veins of patients upon admission to the ICU, prior to the initiation of drug therapy or any other manipulations. Blood collection involved vacuum tubes containing sodium citrate 3.2%, with a maximum interval of 45 min between collection and centrifugation.

To isolate platelet-free plasma for the thrombodynamics test, a "soft" double centrifugation method was used: samples underwent initial centrifugation at 1,600 *g* for 15 min, followed by an additional 20 min at 1,600 g. Platelet-free plasma (120 μL) was used for the test within 3 h.

The thrombodynamics test was performed using the laboratory diagnostic system "Thrombodynamics Recorder TD-2". Blood plasma was introduced into specialized cuvettes, into which an "activator-insert" containing lipids and tissue factor protein was added. This factor initiated the clotting process, simulating damage to the blood vessel wall. Coagulation is initiated on the surface of an activator fixed in space and extends into a thin layer of non-stirred plasma. The growth of the fibrin clot was recorded by the device in sequential photography mode with a digital camera using the dark field method for 30 min.

Based on the obtained images, the Thrombodynamics Recorder TD-2 software calculated the quantitative parameters of the spatial dynamics of fibrin clot growth and spontaneous thrombus formation, including: time to the start of clot growth (Tlag), initial Vi) and stationary (Vst) spatial clot growth rates (the slopes of the clot size curve vs time for the segments of 2–6 min and 15–25 min from the clot growth start for Vi and V, respectively), the clot size at 30 min after coagulation activation (CS), the maximum optical density of the formed clot (D), characterizing its quality, and the time of appearance of spontaneous clots in the sample (Tsp). This latter characteristic has substantial clinical value because spontaneous clots (i.e., those that do not grow from the activator surface) may only be observed in cases of serious hypercoagulable states.

2.6 Statistical and bioinformatic analysis

The STATISTICA software (v13.3, United States) was utilized for statistical processing. The normality of the distribution for quantitative data was assessed using the Shapiro-Wilk's test. Given that the majority of quantitative parameters exhibited deviations from normal distribution, they were presented as the median (Me) along with the first and third quartiles [Q1 and Q3]. The Kruskal–Wallis test was used to compare quantitative variables among three independent groups. Following that, groups were contrasted pairwise using the Mann–Whitney test. To compare quantitative variables among two independent groups, the Mann-Whitney test was also performed. For categorical variables, differences in statistical significance were evaluated using Pearson's chi-squared test with Yates's correction for continuity.

The compliance of genotype distributions with Hardy-Weinberg equilibrium was evaluated using Fisher's exact test. The study groups' genotype frequencies and their associations with disease risk were analyzed using the SNPStats software (https://www.snpstats.net/start.htm (accessed on 18 February 2024)). The additive model was considered for the genotype association analysis. Associations within the entire group of COVID-19 patients/controls were adjusted for age and gender. Given the potentially significant modifying influence of environmental risk factors on the association of genetic markers with disease (Bushueva et al., 2016; Polonikov et al., 2017), associations were analyzed based on the presence or absence of the risk factor. When information about the environmental risk factor was unavailable in the control group (for fruit/vegetable intake, physical activity levels), the patient group was compared to the overall control group. In such cases, the Bonferroni correction was applied to account for multiple comparisons.

The MB-MDR analysis tested two-, three-, and four-level genotype combinations (G×G) and genotype-environment combinations with the including of smoking as an environmental risk factor (G×E). Smoking was analyzed as an environmental risk factor in the analysis of G×E interactions (due to the high pathogenetic significance of this environmental factor in the development of severe COVID-19, as well as the lack of data about other environmental factors like physical activity levels and levels of fruit and vegetable intake in control group). Since SNPs located in the same genes are in linkage disequilibrium, and linkage

groups included no more than two SNPs, one of the SNPs was included in the MB-MDR analysis. For each model, the empirical *p*-value (p_{perm}) was estimated using a permutation test. Permutation testing was employed to improve the validity of the results obtained (Calle et al., 2010). Because the default call to MB-MDR is designed to simultaneously test all possible interactions of a given order, we used 1,000 permutations to obtain accurate p-values. Models with $p_{perm} < 0.01$ were considered as statistically significant. All calculations were adjusted for gender and age. Statistical analysis was carried out using the R software environment. Models (on average 3-4 models of each level) with the highest Wald statistics and the lowest *p*-level of significance were included in the study. Additionally, using the MB-MDR method, individual combinations of genotypes associated with the studied phenotypes were established ($p < 0.05$). Calculations were performed in the MB-MDR program for the R software environment (Version 3.6.3) (Ivanova, 2024).

Additionally, the most significant G×G and G×E models were analyzed using the MDR method (the analysis included genes that appeared in the 2 or more best models of 2-, 3- and 4-locus G×G models in the analysis of intergenic interactions/smoking and genes included in 2 or more best models of 2-, 3- and 4-locus G×E models in the analysis of gene-environment interactions with the including of smoking as an environmental risk factor). The analysis was implemented in the MDR program (v.3.0.2) (http://sourceforge.net/projects/mdr (accessed on 25 February 2024)). The MDR method was used to assess the mechanisms of interactions (synergy, antagonism, additive interactions (independent effects)) and the strength of interactions (the contribution of individual genes/environmental factors as the purpose of the study, to the entropy of a trait and the contribution of interactions, calculated as a percentage). The results of the MDR analysis were visualized as a graph.

We conducted a mediation analysis using the "statsmodels" package for Python to assess whether rs17713054, identified as a genetic risk factor in overall group in our study, influences SARS-CoV-2 directly or indirectly through other clinical conditions such as essential hypertension (EH), coronary artery disease (CAD), cerebrovascular accident (CVA) in anamnesis, chronic obstructive pulmonary disease (COPD), and diabetes mellitus type 2 (T2D).

The functional effects of SNPs were examined using bioinformatics resources, the methodologies and functionalities of which were comprehensively described in our prior research (Kobzeva et al., 2023; Shilenok et al., 2023; Stetskaya et al., 2024):

- The bioinformatic tool GTExportal (http://www.gtexportal.org/ (accessed on 28 February 2024)) was used to analyze the link of SNPs with expression quantitative trait loci (eQTLs) in lungs, whole blood, blood vessels, and adipose tissue (Consortium, 2020).
- For additional examination of binding SNPs to expression quantitative trait loci (eQTL) in peripheral blood, the eQTLGen resource available at https://www.eqtlgen.org/ (accessed on 28 February 2024) was employed (Võsa et al., 2018).
- HaploReg (v4.2), a bioinformatics tool available at https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php (accessed on 28 February 2024), was utilized to assess the associations between GWAS SNPs and specific histone modifications marking promoters and enhancers. These modifications included acetylation of lysine residues at positions 27 and 9 of the histone H3 protein, as well as mono-methylation at position 4 (H3K4me1) and tri-methylation at position 4 (H3K4me3) of the histone H3 protein. Additionally, the tool was applied to investigate the positioning of SNPs in DNase hypersensitive regions (Ward and Kellis, 2012).
- The atSNP Function Prediction online tool (http://atsnp.biostat.wisc.edu/search (accessed on 29 February 2024)) was used to evaluate the impact of SNPs on the gene affinity to transcription factors (TFs) depending on the carriage of the reference/alternative alleles (Shin et al., 2019). TFs were included based on the degree of influence of SNPs on the interaction of TFs with DNA calculated on the basis of a positional weight matrix.
- Using the Gene Ontology online tool (http://geneontology.org/ (accessed on 29 February 2024)), it was feasible to analyze the joint involvement of TFs linked to the reference/SNP alleles in overrepresented biological processes directly related to the pathogenesis of severe COVID-19 (Consortium, 2019). Biological functions controlled by transcription factors associated with SNPs were used as functional groups.
- The Lung Disease Knowledge Portal (LKP) (https://cd.hugeamp.org/ (accessed on 29 February 2024)), which combines and analyzes the results of genetic associations of the largest consortiums for the study of lung diseases, was used for bioinformatics analysis of associations of SNPs with COVID-19 and intermediate phenotypes (such as FEV1, FEV1 to FVC ratio, etc.).

3 Results

3.1 Genetic correlates between GWAS-significant loci and the risk of severe COVID-19

The genotype frequencies of SNPs within the study cohorts are detailed in Supplementary Table S2. Because associations of genetic markers with disease can lead to deviations from equilibrium, we relied on the results of Hardy-Weinberg equilibrium analysis in the control group. Within the control group, all studied SNPs exhibited genotype frequencies consistent with Hardy-Weinberg equilibrium ($p > 0.05$), except for rs12610495 *DPP9* (Supplementary Table S2). However, due to the fact that repeated genotyping of rs12610495 showed 100% reproducibility of the primary results, this SNP was included in the statistical analysis.

The analysis of the entire group (Table 1) revealed an association between rs17713054 *SLC6A20-LZTFL1* and the increased risk of severe COVID-19 course, regardless of sex and age: risk allele A, OR = 1.78, 95% CI = 1.22–2.6, $p = 0.003$. Sex-stratified analysis (Supplementary Table S3) showed that rs17713054 *SLC6A20-LZTFL1* elevates the risk of severe COVID-19 both in males (OR = 1.91, 95% CI = 1.12–3.26, $p = 0.02$) and females (OR = 1.63, 95% CI = 1.03–2.58, $p = 0.04$); additionally, we found that

TABLE 1 Results of the analysis of associations between GWAS SNPs and severe COVID-19 risk in the entire group.

Genetic variant	Effect allele	Other allele	N	OR [95% CI][1]	p^2
rs143334143 *CCHCR1*	A	G	752	1.07 [0.72–1.59]	0.74
rs111837807 *CCHCR1*	C	T	751	0.98 [0.64–1.50]	0.94
rs17713054 *SLC6A20-LZTFL1*	A	G	753	**1.78 [1.22–2.60]**	**0.003**
rs17078346 *SLC6A20-LZTFL1*	C	A	754	1.41 [0.99–2.02]	0.059
rs12585036 *ATP11A*	T	C	749	0.87 [0.65–1.18]	0.37
rs12610495 *DPP9*	G	A	749	1.03 [0.79–1.34]	0.82
rs7949972 *ELF5*	T	C	743	0.92 [0.71–1.21]	0.56
rs61882275 *ELF5*	A	G	751	1.17 [0.91–1.51]	0.21
rs67579710 *THBS3, THBS3-AS1*	A	G	749	0.65 [0.41–1.05]	0.072
rs9636867 *IFNAR2*	G	A	751	0.85 [0.65–1.11]	0.24

All calculations were performed relative to the minor alleles (Effect allele) with adjustment for sex, age; 1 - odds ratio and 95% confidence interval; 2– *p*-value; statistically significant differences are marked in bold.

rs12585036 *ATP11A* lowers the risk of severe COVID-19 in males (protective allele T; OR = 0.51, 95% CI = 0.32–0.83, $p = 0.004$).

Mediation analysis revealed that the indirect effect of rs17713054 through T2D, CVA, and EH was insignificant, accounting for 0%, 2.79%, and 5.48% respectively, in the sequential analysis of these conditions. Adding these variables to the logistic regression model did not render the influence of rs17713054 on SARS-CoV-2 statistically insignificant.

Conversely, adding COPD or CAD to the model rendered the influence of rs17713054 on SARS-CoV-2 insignificant (with the model including COPD showing a weaker effect and the significance level remaining within the statistical trend, $p < 0.1$). Mediation analysis for these variables showed that the contribution of rs17713054 to SARS-CoV-2, mediated through COPD and CAD, was 18.57% and 71.54% respectively. This suggests that the influence of rs17713054 on SARS-CoV-2 is likely mediated through other clinical conditions, primarily through CAD (Supplementary Table S4).

3.2 Gene-gene interactions associated with severe COVID-19

Using the MB-MDR method, five most significant models of intergenic interactions associated with the severe course of COVID-19 were established: one two-locus model, three three-locus and one four-locus models ($p_{perm} \leq 0.001$) (Table 2). In total, the best models of G×G interactions included eight polymorphic loci, four of which—rs67579710 *THBS3*, *THBS3-AS1*, rs17713054 *SLC6A20-LZTFL1*, rs7949972 *ELF5*, rs9636867 *IFNAR2*—were involved in 2 or more of the most significant G×G interactions. We analyzed the interactions of these genetic variants using the MDR method (Figure 2).

The MDR method, firstly, showed that the genetic variants included in the best G×G models are characterized by antagonism/additive (independent) effects. Secondly, the mono-effects of SNPs are comparable to the effects of gene-gene interactions in terms of their contribution to the entropy of COVID-19, with the exception of rs17713054, which showed the most prominent mono-effect (1.15%). Thirdly, combinations of genotypes of GWAS-significant SNPs associated with severe COVID-19 are listed in Supplementary Table S4. The combinations with the most pronounced associations with severe COVID-19 are as follows: rs67579710 *THBS3, THBS3-AS1* G/G×rs17713054 *SLC6A20-LZTFL1* A/G (Beta = 0.15378, $p = 0.0001$); rs67579710 *THBS3, THBS3-AS1* G/G×rs17713054 *SLC6A20-LZTFL1* A/G×rs143334143 *CCHCR1* G/G (Beta = 0.16359, $p = 0.0002$ rs7949972 *ELF5* T/C×rs67579710 *THBS3, THBS3-AS1* G/G×rs12610495 *DPP9* G/A) (Beta = 0.11149, $p = 0.01$); rs9636867 *IFNAR2* G/G×rs67579710 *THBS3, THBS3-AS1* G/G×rs17713054 *SLC6A20-LZTFL1* A/G (Beta = 0.215831, $p = 0.0003$); rs7949972 *ELF5* T/C×rs9636867 *IFNAR2* G/G×rs67579710 *THBS3, THBS3-AS1* G/G×rs17713054 *SLC6A20-LZTFL1* A/G (Beta = 0.278009, $p = 0.002$) (Supplementary Table S5).

3.3 Environmental-associated correlates of GWAS SNPs

GWAS SNPs were assessed for their potential contribution to COVID-19 severity in combination with environmental risk factors

TABLE 2 Gene-gene interactions associated with severe COVID-19 (MB-MDR modeling).

Gene-gene interaction models	NH	beta H	WH	NL	beta L	WL	Wmax	p_{perm}
The best two-locus models of intergenic interactions (for models with Pmin. < 0.001, 1,000 permutations)								
rs67579710 *THBS3*, *THBS3-AS1* × rs17713054 *SLC6A20-LZTFL1*	1	0.1538	15.21	1	−0.05222	2.886	15.21	0.002
The best three-locus models of intergenic interactions (for models with Pmin. < 1×10^{-4}, 1,000 permutations)								
rs67579710 *THBS3*, *THBS3-AS1* × rs17713054 *SLC6A20-LZTFL1*× rs143334143 *CCHCR1*	2	0.1673	17.60	1	−0.05878	4.001	17.60	0.005
rs7949972 *ELF5* × rs67579710 *THBS3*, *THBS3-AS1* × rs12610495 *DPP9*	3	0.1370	19.41	2	−0.08620	6.019	19.41	0.008
rs9636867 *IFNAR2* × rs67579710 *THBS3*, *THBS3-AS1* × rs17713054 *SLC6A20-LZTFL1*	2	0.2242	18.84	1	−0.05784	3.391	18.84	0.018
The best four-locus models of gene-gene interactions (for models with Pmin. < 1×10^{-5}, 1,000 permutations)								
rs7949972 *ELF5* × rs9636867 *IFNAR2* × rs67579710 *THBS3*, *THBS3-AS1* × rs17713054 *SLC6A20-LZTFL1*	4	0.1990	24.34	2	−0.11414	7.205	24.34	0.046

Note: NH, is the number of interacting high-risk genotypes; beta H—regression coefficient for high-risk interactions identified at the second stage of analysis; WH, Wald statistics for high-risk interactions; NL, number of interacting low-risk genotypes; beta L—regression coefficient for low-risk interactions identified at the second stage of analysis; WL, Wald statistics for low-risk interactions; p_{perm}—permutational significance levels for models (all models are adjusted for gender and age); Loci included in 2 or more best G×G models are indicated in bold.

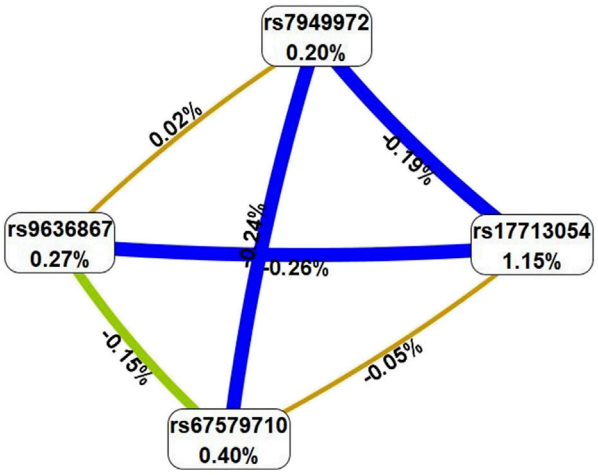


FIGURE 2
Graph reflecting the structure and strength of the most significant G×G interactions of GWAS-significant loci associated with severe COVID-19. (Note: the color of the lines reflects the nature of the interaction: red and orange lines mean pronounced and moderate synergism, brown means additive effect of genes (independent effects); % reflects the strength and direction of the phenotypic effect of gene interaction (% entropy)).

such as smoking, fresh fruit and vegetable consumption, and physical activity level (Supplementary Table S5). SNP rs17713054 *SLC6A20-LZTFL1* was associated with an increased risk of severe COVID-19 risk among nonsmokers (risk allele A; OR = 1.65, 95% CI = 1.11–2.46, $p = 0.02$), patients with low fruit and vegetable intake (OR = 1.72, 95% CI = 1.15–2.58, $p = 0.01$, $p_{bonf} = 0.02$), and patients with low levels of physical activity (OR = 1.93, 95% CI = 1.26–2.94, $p = 0.0035$, $p_{bonf} = 0.007$) (Supplementary Table S6).

Using the MB-MDR approach, the eight most significant models of gene-environment interactions associated with severe COVID-19 were identified: two two-level model, two three-order models, and four four-level models ($p_{perm} \leq 0.01$) (Table 3). In total, the best G×E models included smoking in interaction with seven loci, five of which—rs7949972 *ELF5*, rs17713054 *SLC6A20-LZTFL1*, rs9636867 *IFNAR2*, rs12585036_ATP11A, rs12610495_*DPP9*—were involved in two or more of the most significant G×E interactions. In the next step, we analyzed the interactions between these genetic variants and smoking using the multivariate dimensionality reduction (MDR) method (Figure 3).

Firstly, MDR revealed that smoking as an environmental risk factor has the least mono-effect (0.04% contribution to the entropy of severe COVID-19). Secondly, the mono-effects of SNPs/smoking (0.04%–1.15%) are comparable to the effects of gene-environment interactions (0.01%–0.53%). Thirdly, rs17713054 has the maximum mono-effect among the SNPs involved in the most significant gene-environment interactions. (1.15% contribution to entropy). Fourthly, smoking is characterized by multidirectional effects in interaction with SNPs included in the best G×E models: pronounced synergism in interaction with rs12610495, moderate synergism in interaction with rs9636867, additive (independent) effects in interaction with rs17713054, rs12585036, rs7949972. Fifth, the interactions between the genetic variants included in the most significant G×E models are antagonistic/independent (additive effects), with the exception of the interactions between rs7949972 and rs12610495, which exhibit pronounced synergism in interaction with each other. Sixthly, the list of models of gene-environment interactions between GWAS SNPs' genotypes and smoking is presented in Supplementary Table S6. The following gene-smoking interactions show the strongest correlation with severe COVID-19: non-smokers × rs17713054 *SLC6A20-LZTFL1* G/G (Beta = 0.06466, $p = 0.031$); smokers × rs9636867 *IFNAR2* A/A (Beta = 0.179684, $p = 0.049$); non-smokers ×rs67579710 *THBS3, THBS3-AS1* G/G×rs17713054 *SLC6A20-LZTFL1* A/G (Beta = 0.162513, $p = 6.77 \times 10^{-5}$); non-smokers ×rs7949972 *ELF5* T/C×rs9636867 *IFNAR2* G/G×rs17713054 *SLC6A20-LZTFL1* A/G (Beta = 0.3137824, $p = 0.002$); smokers ×rs9636867 *IFNAR2*

TABLE 3 Gene-environmental interactions, associated with severe COVID-19 (MB-MDR modeling).

Gene-gene interaction models	NH	beta H	WH	NL	beta L	WL	Wmax	p_{perm}
The best two-order models of gene-smoking interactions (for G×E models with Pmin. < 0.005, 1,000 permutations)								
SMOKE × rs17713054 *SLC6A20-LZTFL1*	2	0.10911	8.737	1	−0.06466	4.696	8.737	0.02
SMOKE × rs9636867 *IFNAR2*	2	0.11407	7.307	1	−0.08059	3.742	7.307	0.048
The best three-order models of gene-smoking interactions (for G×E models with Pmin. < 0.005, 1,000 permutations)								
SMOKE × rs67579710 *THBS3*, *THBS3-AS1* × rs17713054 *SLC6A20-LZTFL1*	2	0.17901	14.892	2	−0.06870	5.292	14.892	0.009
SMOKE × rs9636867 *IFNAR2* × rs12585036 *ATP11A*	2	0.19410	13.116	0	NA	NA	13.116	0.045
The best four-order models of gene-smoking interactions (for G×E models with Pmin. < 1×10^{-5}, 1,000 permutations)								
SMOKE × rs7949972 *ELF5* × rs9636867 *IFNAR2* × rs17713054 *SLC6A20-LZTFL1*	4	0.3663	25.92	2	−0.12671	8.166	25.92	0.018
SMOKE × rs9636867 *IFNAR2* × rs12585036 *ATP11A* × rs17713054 *SLC6A20-LZTFL1*	6	0.2709	27.16	1	−0.12411	4.163	27.16	0.024
SMOKE × rs9636867 *IFNAR2* × rs12610495 *DPP9* × rs17713054 *SLC6A20-LZTFL1*	5	0.2858	25.01	2	−0.13819	8.247	25.01	0.039
SMOKE × rs7949972 *ELF5* × rs12610495 *DPP9* × rs12585036 *ATP11A*	7	0.2118	27.69	2	−0.14928	7.334	27.69	0.045

Note: NH, is the number of high-risk interactions; beta H—regression coefficient for high-risk interactions identified at the second stage of analysis; WH, Wald statistics for high-risk interactions; NL, number of interacting low-risk interactions; beta L—regression coefficient for low-risk interactions identified at the second stage of analysis; WL, Wald statistics for low-risk interactions; p_{perm}—permutational significance levels for models (all models are adjusted for gender, age); Loci included in 2 or more best G× E models are indicated in bold.

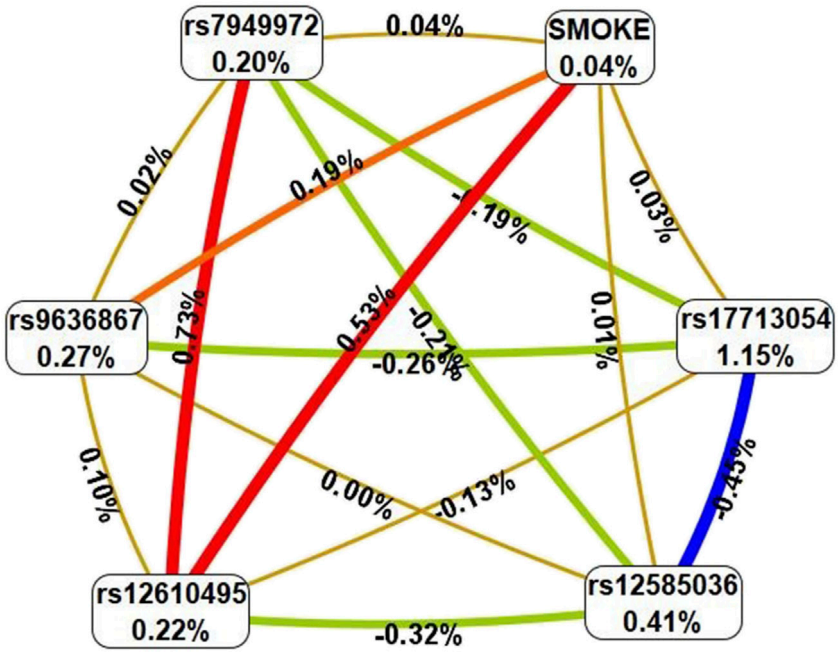


FIGURE 3
Graph reflecting the structure and power of the most significant G×E interactions of GWAS loci associated with severe COVID-19. (Note: The color of the lines reflects the nature of the interaction: red means strong synergism, brown means additive (independent) effects, and % reflects the strength and direction of the phenotypic effect of gene-environmental interaction (% of entropy)).

A/A×rs12585036 *ATP11A* C/C×rs17713054 *SLC6A20-LZTFL1* G/G (Beta = 0.648395, p = 0.0003); smokers ×rs9636867 *IFNAR2* A/A×rs12610495 *DPP9* G/A×rs17713054 *SLC6A20-LZTFL1* G/G (Beta = 0.485735, p = 0.003); non-smokers ×rs7949972 *ELF5* C/C ×rs12610495 *DPP9* A/A×rs12585036 *ATP11A* C/T (Beta = 0.157332, p = 0.01) (Supplementary Table S7).

TABLE 4 Results of the analysis of associations between GWAS SNPs and severe COVID-19 in obese and non-obese patients.

Genetic variant	Effect allele	Other allele	N	OR [95% CI][1]	p^2 (p_{bonf})	N	OR [95% CI][1]	p^2 (p_{bonf})
			BMI <30			BMI ≥30		
rs143334143 *CCHCR1*	A	G	657	0.82 [0.48–1.38]	0.44 (0.88)	658	1.12 [0.70–1.80]	0.63 (1.26)
rs111837807 *CCHCR1*	C	T	656	0.65 [0.36–1.18]	0.14 (0.28)	657	1.09 [0.67–1.78]	0.73 (1.46)
rs17713054 *SLC6A20-LZTFL1*	A	G	657	1.14 [0.69–1.88]	0.61 (1.22)	659	**2.31 [1.52–3.50]**	**0.0002 (0.0004)**
rs17078346 *SLC6A20-LZTFL1*	C	A	657	1.02 [0.64–1.63]	0.93 (1.86)	660	**1.72 [1.15–2.58]**	**0.01 (0.02)**
rs12585036 *ATP11A*	T	C	654	0.80 [0.55–1.16]	0.23 (0.46)	656	0.85 [0.60–1.22]	0.38 (0.76)
rs12610495 *DPP9*	G	A	655	0.85 [0.61–1.19]	0.34 (0.68)	656	**1.48 [1.09–2.01]**	**0.01 (0.02)**
rs7949972 *ELF5*	T	C	647	**0.67 [0.47–0.95]**	**0.02 (0.04)**	650	1.13 [0.82–1.55]	0.46 (0.92)
rs61882275 *ELF5*	A	G	656	0.91 [0.66–1.26]	0.56 (1.12)	657	**1.40 [1.02–1.91]**	**0.036** (0.072)
rs67579710 *THBS3, THBS3-AS1*	A	G	653	0.79 [0.44–1.42]	0.42 (0.84)	657	0.75 [0.42–1.35]	0.33 (0.66)
rs9636867 *IFNAR2*	G	A	655	0.82 [0.59–1.16]	0.26 (0.52)	657	1.02 [0.74–1.42]	0.91 (1.82)

All calculations were performed relative to the minor alleles (Effect allele); 1 - odds ratio and 95% confidence interval; 2– *p*-value; statistically significant differences are marked in bold.

3.4 Obesity-depended associations of GWAS SNPs with severe COVID-19

Considering the potential impact of BMI, particularly obesity, on the severity of COVID-19, we carried out an analysis of associations of GWAS SNPs with severe COVID-19 in groups of patients stratified by BMI. Among patients with a BMI less than 30 (non-obese patients), the rs7949972 *ELF5* variant was associated with a reduced risk of severe COVID-19 (protective allele T, OR = 0.67, 95% CI = 0.47–0.95, $p = 0.02$, $p_{bonf} = 0.04$) (Table 4). However, in patients with obesity (BMI ≥30), increased risk of severe COVID-19 was observed for the rs17713054 *SLC6A20-LZTFL1* (risk allele A, OR = 2.31, 95% CI = 1.52–3.5, $p = 0.0002$, $p_{bonf} = 0.0004$), rs12610495 *DPP9* (risk allele G, OR = 1.48, 95% CI = 1.09–2.01, $p = 0.01$, $p_{bonf} = 0.02$), and rs17078346 *SLC6A20-LZTFL1* (risk allele C, OR = 1.72, 95% CI = 1.15–2.58, $p = 0.01$, $p_{bonf} = 0.02$) (Table 4).

3.5 Relationship between GWAS- significant loci and the clinical characteristics of severe COVID-19 patients

The results of the associations between GWAS SNPs and clinical characteristics of severe COVID-19 patients are presented in Figure 4 and Supplementary Table S8.

Upon the analysis of clinical characteristics among hospitalized COVID-19 patients, it was observed that rs17713054 *SLC6A20-LZTFL1* ($p = 0.006$), rs12610495 *DPP9* ($p = 0.01$), and rs17078346 *SLC6A20-LZTFL1* ($p = 0.01$) were found to be linked with increased BMI (Supplementary Table S7; Figures 4A–C). Additionally, rs12610495 *DPP9* correlated with a reduction in the duration of oxygen therapy (Figure 4D). The maximum optical density of the formed clot (D) was associated with rs17713054 *SLC6A20-LZTFL1* ($p = 0.02$) (Figure 4E). SNP rs12585036 *ATP11A* ($p = 0.006$) increased the count of platelets (Figure 4F). Meanwhile, SNP rs7949972 *ELF5* ($p = 0.02$) reduced in stationary spatial clot growth rates (Vst, μm/minutes) (Figure 4G), rs12610495 *DPP9* ($p = 0.01$) increased the time to the start of clot growth (Tlag, minutes) (Figure 4H), while rs17713054 *SLC6A20-LZTFL1* ($p = 0.036$) and rs7949972 *ELF5* ($p = 0.02$) decreased the clot size at 30 min post-coagulation activation (CS, μm) (Figure 4I, J respectively). Notably, the time of appearance of spontaneous clots (Tsp) was extended in the overall patient group with rs17713054 *SLC6A20-LZTFL1* ($p = 0.0036$) (Figure 4K). Given the strong correlation between rs17713054 *SLC6A20-LZTFL1*, rs12610495 *DPP9*, and rs17078346 *SLC6A20-LZTFL1* with BMI, we conducted a comparison of clinical characteristics between two patient groups based on BMI status. In patients with a BMI ≥30, SNP rs17713054 *SLC6A20-LZTFL1* (Figure 4L) was associated with an elevation in Tsp ($p = 0.02$), while among patients without obesity (BMI <30) rs61882275 *ELF5* ($p = 0.003$) was found to increase Tsp (Figure 4M).

3.6 Functional annotation of severe COVID-19-related SNPs

3.6.1 QTL-effects

The results of the cis-eQTL analysis (Table 5) shed light on the impact of specific genetic variants on gene expression. According to

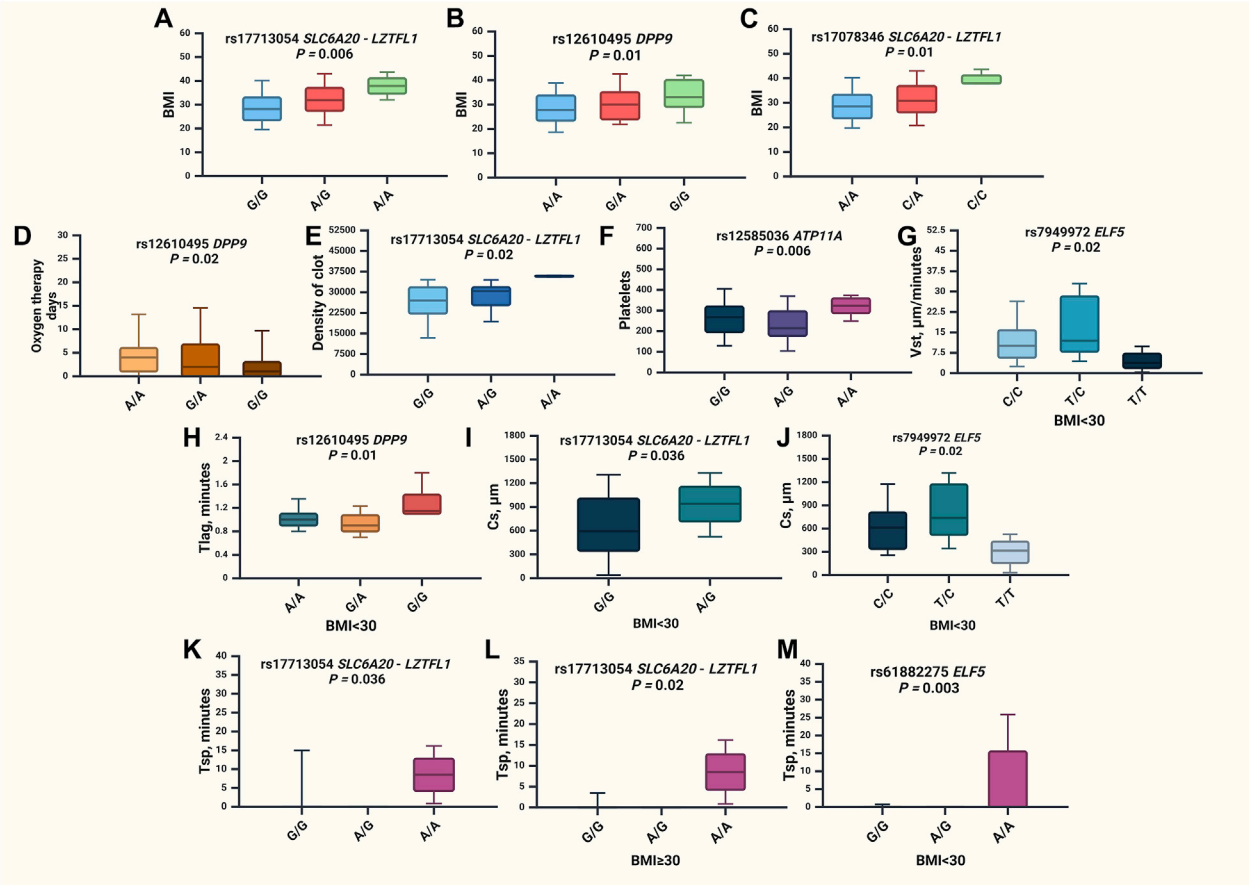


FIGURE 4
Associations of GWAS loci and clinical characteristics of severe COVID-19 patients. **(A)** BMI values for rs17713054 *SLC6A20-LZTFL1* in the entire group (p = 0.006), **(B)** BMI values for rs12610495 *DPP9* in the entire group (p = 0.01), **(C)** BMI values for rs17078346 *SLC6A20-LZTFL1* in the entire group (p = 0.01), **(D)** oxygen therapy days for rs12610495 *DPP9* in the entire group (p = 0.02), **(E)** maximum optical density of the formed clot **(D)** values for rs17713054 *SLC6A20-LZTFL1* in the entire group (p = 0.02), **(F)** platelets count for rs12585036 *ATP11A* in the entire group (p = 0.006), **(G)** stationary spatial clot growth rates (Vst, μm/minutes) values for rs7949972 *ELF5* in the group of patients with BMI <30 (p = 0.02), **(H)** time to the start of clot growth (Tlag, minutes) for rs12610495 *DPP9* in the group of patients with BMI <30 (p = 0.01), **(I)** clot size at 30 min post-coagulation activation (CS, μm) values for rs17713054 *SLC6A20-LZTFL1* in the group of patients with BMI <30 (p = 0.036), **(J)** clot size at 30 min post-coagulation activation (CS, μm) values for rs7949972 *ELF5* in the group of patients with BMI <30 (p = 0.02), **(K)** time of appearance of spontaneous clots (Tsp) values for rs17713054 *SLC6A20-LZTFL1* in the entire group of patients (p = 0.036), **(L)**—Tsp values for rs17713054 *SLC6A20-LZTFL1* in the group of patients with BMI<30 (p = 0.02), **(M)**—Tsp values for rs61882275 *ELF5* in the group of patients with BMI <30 (p = 0.003).

the eQTLGen Browser, rs17713054 *SLC6A20-LZTFL1* and rs17078346 *SLC6A20-LZTFL1* were associated with a decrease in the expression of *FLT1P1*, *CCR3*, *CCR1*, *SACM1L*, *CCR5*, *CCR2*, *RP11-24F11.2*, and *CXCR6*, while these two SNPs were linked to an increase in the expression of CCR9 in blood. Moreover, data from the GTEx Portal indicated that rs17713054 *SLC6A20-LZTFL1* was associated with reduced expression levels of *CXCR6* in tibial artery and adipose tissues (subcutaneous), alongside an elevation in the expression of *LZTFL1* in adipose tissue (subcutaneous).

Additionally, rs12585036 *ATP11A* was correlated with decreased expression levels of *ATP11A* in the blood and aorta, as well as *RP11-88E10.5* in coronary arteries. rs12610495 DPP9 showed associations with reduced expression of *DPP9* in blood, lung, and arteries (tibial artery and aorta), while it was linked to an increase in the expression levels of *TNFAIP8L1* in blood. Notably, rs7949972 *ELF5* demonstrated a decrease in expression levels of *CAT* in whole blood and artery (tibial), while *ABTB2* expression was reduced solely in whole blood by the influence of this SNP. Furthermore, *ELF5* expression was found to be decreased in the lungs, indicating the effects of rs7949972.

3.6.2 Histone modifications

Using the bioinformatics tool HaploReg (v4.2), we analyzed histone modifications associated with SNPs identified in our study as linked to an increased risk of severe COVID-19 (Table 6).

SNP rs17713054 *SLC6A20-LZTFL1* is situated in a DNA-binding region associated with histone H3 monomethylation at the fourth lysine residue (H3K4me1) in lung, aorta, and adipose tissue. Moreover, this SNP has further influence on H3K27ac, which marks enhancers, particularly in lung tissues and the aorta.

Similarly, rs12610495 *DPP9* is located in a DNA-binding region associated with H3K4me1 in both the lungs and blood. In lung tissue, it also binds to H3K4me3. Additionally, the impact of these histone modifications is further enhanced by the presence of H3K27ac.

TABLE 5 Association of SNPs with cis-eQTL-Mediated Expression Profiles of GWAS Genes.

SNP	Effect allele	eQTLGen Browser data			GTEx Portal data			
		Gene expressed	Z-score	*p*-value	Gene expressed	*p*-value	Effect (NES)	Tissue
rs17713054 *SLC6A20-LZTFL1*	A	*CXCR6*	↓(−13.9294)	4.20×10^{-44}	*CXCR6*	1.7×10^{-7}	↓(-0.42)	Artery - Tibial
		FLT1P1	↓(−15.1094)	1.40×10^{-51}				
		CCR3	↓(−14.3393)	1.24×10^{-46}				
		CCR9	↑(5.1055)	3.30×10^{-7}	*CXCR6*	6.7×10^{-5}	↓(-0.30)	Adipose - Subcutaneous
		CCR1	↓(−7.4173)	1.19×10^{-13}				
		SACM1L	↓(−5.7667)	8.08×10^{-9}	*LZTFL1*	1.0×10^{-4}	↑(0.21)	Adipose - Subcutaneous
		CCR5	↓(−5.206)	1.93×10^{-7}				
		CCR2	↓(−4.9694)	6.71×10^{-7}	*CCR9*	1.6×10^{-4}	↑(0.33)	Whole Blood
		RP11-24F11.2	↓(−4.6342)	3.58×10^{-6}				
rs17078346 *SLC6A20-LZTFL1*	C	*CCR3*	↓(−13.0025)	1.18×10^{-38}	-			
		FLT1P1	↓(−12.847)	8.94×10^{-38}				
		CXCR6	↓(−11.6247)	3.08×10^{-31}				
		CCR1	↓(−6.5185)	7.10×10^{-11}				
		CCR5	↓(−5.5666)	2.59×10^{-8}				
		SACM1L	↓(−5.4113)	6.25×10^{-8}				
		CCR2	↓(−5.3267)	9.99×10^{-8}				
		CCR9	↑(5.0282)	4.95×10^{-7}				
		RP11-24F11.2	↓(−4.5892)	4.44×10^{-6}				
rs12585036 *ATP11A*	T	*ATP11A*	↓(−7.9284)	2.22×10^{-15}	*ATP11A*	7.3×10^{-8}	↓(−0.18)	Artery - Aorta
					RP11-88E10.5	2.2×10^{-6}	↓(−0.34)	Artery - Coronary
rs12610495 *DPP9*	G	*DPP9*	↓(−14.4364)	3.05×10^{-47}	*DPP9*	4.50×10^{-9}	↓(−0.18)	Lung
		TNFAIP8L1	↑(7.6938)	1.43×10^{-14}	*DPP9*	8.90×10^{-8}	↓(−0.15)	Artery - Tibial
					DPP9	4.1×10^{-6}	↓(−0.17)	Artery - Aorta
rs7949972 *ELF5*	T	*CAT*	↓(−56.0274)	3.27×10^{-310}	*ELF5*	2.50×10^{-15}	↓(−0.23)	Lung
					CAT	3.20×10^{-14}	↓(−0.25)	Whole Blood
		ABTB2	↓(-15.164)	6.12×10^{-52}	*CAT*	4.3×10^{-6}	↓(−0.15)	Artery - Tibial
					ABTB2	0.00007	↓(−0.17)	Whole Blood

Finally, rs7949972 *ELF5* falls within a region of DNA binding to H3K4me1 exclusively in lung tissue.

3.6.3 Analysis of transcription factors

The risk allele A of rs17713054 *SLC6A20-LZTFL1* is associated with the generation of DNA binding sites for 48 transcription factors (TFs) (Supplementary Table S9). These TFs are involved in four overrepresented biological processes: integrated stress response signaling (GO:0140467; FDR = 1.48×10^{-12}), positive regulation by host of viral transcription (GO:0043923; FDR = 4.68×10^{-2}), fat cell differentiation (GO:0045444; FDR = 3.45×10^{-4}), transforming growth factor beta receptor signaling pathway (GO:0007179; FDR = 4.94×10^{-2}). The protective allele G of rs17713054 *SLC6A20-LZTFL1* creates binding sites for 24 TFs, jointly involved in response to hypoxia (GO:0001666; FDR = 2.8×10^{-2}).

The protective allele T rs12585036 *ATP11A* generates DNA binding sites for 104 TFs (Supplementary Table S10) involved in response to (GO:1990785; FDR = 8.08×10^{-3}), response to testosterone (GO:0033574; FDR = 4.09×10^{-11}), androgen receptor signaling pathway (GO:0030521; FDR = 8.49×10^{-3}), canonical Wnt signaling pathway (GO:0060070; FDR = 3.54×10^{-3}).

As for the risk allele C rs17078346 *SLC6A20-LZTFL1,* it creates DNA binding regions for 31 TFs (Supplementary Table S11), that are involved in three overrepresented biological processes: epithelial

TABLE 6 The impact of GWAS SNPs on histone tags in various tissues.

SNP (Ref/Alt allele)	Tissues Marks	Lung	Vessels—aorta	Blood	Adipose tissue
rs17713054 (G/A) *SLC6A20-LZTFL1*	H3K4me1	Enh	Enh	-	Enh
	H3K4me3	-	-	-	-
	H3K27ac	Enh	Enh	-	-
rs12610495 (A/G) *DPP9*	H3K4me1	Enh	-	Enh	-
	H3K4me3	Pro	-	-	-
	H3K27ac	Enh	-	-	-
rs7949972 (C/T) *ELF5*	H3K4me1	Enh	-	-	-

H3K4me1—mono-methylation at the fourth lysine residue of the histone H3 protein; H3K4me3—tri-methylation at the fourth lysine residue of the histone H3 protein; H3K9ac—the acetylation at the ninth lysine residues of the histone H3 protein; H3K27ac—acetylation of the lysine residues at N-terminal position 27 of the histone H3 protein; effect alleles are marked in bold. Enh—histone modification in the enhancer region; Pro—histone modification at the promoter region.

tube branching involved in lung morphogenesis (GO:0060441; FDR = 7.59×10^{-4}), Notch signaling pathway (GO:0007219; FDR = 1.29×10^{-3}).

Protective allele A rs12610495 *DPP9* is associated with the generation of DNA binding sites for 39 TFs (Supplementary Table S12). These TFs jointly participate in positive regulation of regulation of cytokine production (GO:0001817; FDR = 0.0475).

Finally, risk allele C rs7949972 *ELF5* creates DNA binding sites for 32 TFs (Supplementary Table S13), that are jointly involved in the following overrepresented biological processes: positive regulation of CD8-positive, alpha-beta T cell differentiation (GO: 0043378; FDR = 0.00247), negative regulation of CD4-positive, alpha-beta T cell differentiation (GO:0043371; FDR = 0.0301), defense response to virus (GO:0051607; FDR = 0.00177), positive regulation of interferon-alpha production (GO:0032727; FDR = 0.0413), positive regulation of interferon-beta production (GO: 0032728; 0.00251).

3.6.4 Bioinformatic analysis of the associations of GWAS SNPs with COVID-19-related phenotypes

According to the bioinformatic resource Lung Disease Knowledge Portal, the GWAS SNPs rs17713054, rs12585036, rs17078436, rs12610495 are linked to the higher risk of hospitalization of COVID-19 patients and to the severe respiratory confirmed COVID-19. Additionally, rs12585036 is associated with a reduction in lung capacity parameters such as forced vital capacity (FVC), forced expired volume in 1 s (FEV1), FEV1 to FVC ratio, peak expiratory flow. Conversely, rs7949972 is associated with a lower risk of hospitalization in COVID-19 patients while increasing the lung capacity parameters (Table 7).

4 Discussion

In the present study, we replicated associations of the rs17713054 *SLC6A20-LZTFL1*, rs17078346 *SLC6A20-LZTFL1*, rs12610495 *DPP9* and rs7949972 *ELF5* with severe COVID-19 within the Caucasian population of Central Russia. For the first time in the world, we assessed the impact of COVID-19 GWAS loci on a wide range of clinical manifestations of the disease, primarily on thrombodynamic parameters, identified the most significant intergenic interactions, and also assessed how environmental risk factors and obesity modify associations of GWAS loci with the risk of severe COVID-19; conducted a comprehensive functional annotation of severe COVID-19-associated SNPs to analyse their involvement in the molecular mechanisms of the disease.

Figure 5 summarizes the principal molecular mechanisms underlying the involvement of GWAS SNPs to severe COVID-19.

First of all, we identified that both studied polymorphic variants located in the *SLC6A20-LZTFL1* region are associated with COVID-19: rs17713054 *SLC6A20-LZTFL1* (risk allele A) increases the risk of severe COVID-19 regardless of sex and age; however, this risk can be modified by smoking status, intake of fresh fruit and vegetables, and higher levels of physical activity. Moreover, rs17713054 (risk allele A) was found to be associated with an increase in body mass index and worsening thrombodynamic parameters, including an increase in the maximum optical density of the formed clot (D), delayed appearance of spontaneous clots (Tsp), and larger clot size 30 min after coagulation activation (CS). It is noteworthy that rs17713054 showed an association with severe COVID-19 in a large number of replication studies conducted around the world (Roberts et al., 2020; Downes et al., 2021; Roozbehani et al., 2023; Udomsinprasert et al., 2023). However, the possible influence of rs17713054 on both the development of COVID-19 and the development of coronary artery disease is a topic of active discussion in the literature (Wang et al., 2023). According to our mediation analysis, the contribution of rs17713054 to SARS-CoV-2 susceptibility may be mediated through comorbid disease in severe COVID-19 patients, to a lesser extent by chronic obstructive pulmonary disease, and to a greater extent by coronary artery disease.

SNP rs17078346 *SLC6A20-LZTFL1* (risk allele C) also was associated with the increased the risk of severe COVID-19 in our study, exclusively in obese patients. Possible molecular mechanisms of the involvement of these genetic variants in the risk of developing severe COVID-19 may be associated with their regulation of the *LZTFL1* gene (Leucine Zipper Transcription Factor Like 1), which regulates protein trafficking to the ciliary membrane, the violation of which may play an important role in weakened airway viral clearance in a patient with COVID-19 (Robinot et al., 2021).

TABLE 7 Results of aggregated bioinformatic analyzes of associations between GWAS SNPs and the risk of severe COVID-19 course.

No	SNP	Phenotype	*p*-value	Beta (OR)	Sample size
1	rs17713054 *SLC6A20-LZTFL1* (G/A)	Very severe respiratory confirmed COVID-19 vs. population	1.15×10^{-80}	$_{OR}$▲1.8111	7,252
2		Hospitalized COVID-19 vs. population	1.09×10^{-51}	$_{OR}$▲1.8134	908,494
3		Hospitalized vs. non-hospitalized COVID-19	2.04×10^{-28}	$_{OR}$▲1.3555	10,216
4		COVID-19 vs. population	4.80×10^{-26}	$_{OR}$▲1.3121	1,299,010
5		Very severe respiratory confirmed vs. non-hospitalized COVID-19	1.43×10^{-5}	$_{OR}$▲2.8766	957
6		COVID-19 vs. no COVID-19	3.12×10^{-5}	$_{OR}$▲1.1324	127,879
7	rs12585036 *ATP11A* (C/T)	Idiopathic pulmonary fibrosis	2.36×10^{-16}	$_{OR}$▼0.9994	57,913
8		FEV1 to FVC ratio	1.97×10^{-6}	$_{Beta}$▼-0.0141	793,368
9		Very severe respiratory confirmed COVID-19 vs. population	8.12×10^{-6}	$_{OR}$▲1.1025	7,376
10		Forced expired volume in 1 s (FEV1)	8.26×10^{-6}	$_{Beta}$▼-0.0128	793,442
11		Peak expiratory flow	5.68×10^{-4}	$_{Beta}$▼-0.0105	690,530
12		Hospitalized vs. non-hospitalized COVID-19	0.002	$_{OR}$▲1.0604	10,013
13		Hospitalized COVID-19 vs. population	0.004	$_{OR}$▲1.0604	908,494
14		Forced vital capacity (FVC)	0.025	$_{Beta}$▼-0.0063	792,938
15		Airway wall area in COPD	0.029	$_{Beta}$▼-0.0334	12,031
16	rs17078346 *SLC6A20-LZTFL1* (A/C)	Very severe respiratory confirmed COVID-19 vs. population	2.96×10^{-39}	$_{OR}$▲1.5011	5,855
17		Hospitalized COVID-19 vs. population	1.08×10^{-18}	$_{OR}$▲1.4711	898,438
18		Hospitalized vs. non-hospitalized COVID-19	1.01×10^{-16}	$_{OR}$▲1.2208	10,256
19		COVID-19 vs. population	3.57×10^{-9}	$_{OR}$▲1.1637	1,288,650
20		COVID-19 vs. no COVID-19	8.72×10^{-5}	$_{OR}$▲1.1040	127,879
21		Very severe respiratory confirmed vs. non-hospitalized COVID-19	2.42×10^{-4}	$_{OR}$▲2.2105	957
22	rs12610495 *DPP9* (A/G)	Very severe respiratory confirmed COVID-19 vs. population	1.64×10^{-15}	$_{OR}$▲1.2015	5,642
23		Idiopathic pulmonary fibrosis	4.11×10^{-15}	$_{OR}$▲1.0003	58,925
24		Hospitalized COVID-19 vs. population	4.84×10^{-8}	$_{OR}$▲1.1914	895,822
25		Hospitalized vs. non-hospitalized COVID-19	1.73×10^{-5}	$_{OR}$▲1.0769	9,939
26		COVID-19 vs. population	1.64×10^{-4}	$_{OR}$▲1.0704	1,274,140
27		COVID-19 vs. no COVID-19	0.0025	$_{OR}$▲1.0603	101,592
28	rs7949972 *ELF5* (C/T)	Very severe respiratory confirmed COVID-19 vs. population	6.47×10^{-7}	$_{OR}$▼0.9079	7,225
29		FEV1 to FVC ratio	3.9×10^{-6}	$_{Beta}$▼-0.0112	808,254
30		Hospitalized vs. non-hospitalized COVID-19	7.01×10^{-5}	$_{OR}$▼0.9392	10,256
31		Hospitalized COVID-19 vs. population	0.0021	$_{OR}$▼0.9254	905,878
32		Forced vital capacity (FVC)	0.0034	$_{Beta}$▲0.0071	807,822
33		COVID-19 vs. population	0.022	$_{OR}$▼0.9642	1,289,590
34		Emphysema in COPD (percentage low attenuation area −950 HU)	0.024	$_{Beta}$▲0.0375	12,031
35		Emphysema in COPD (15th percentile HU)	0.048	$_{Beta}$▼-0.0245	12,031

data obtained using the bioinformatic resource Lung Disease Knowledge Portal https://lung.hugeamp.org/
Effect alleles are marked in bold.

Moreover, *LZTFL1* regulates the transition of epithelial cells to mesenchymal cells (https://www.genecards.org/cgi-bin/carddisp.pl?gene=LZTFL1), thereby participating in the regulation of the viral response pathway associated with epithelial-mesenchymal transition (Downes et al., 2021), an important regulator of the innate immune response.

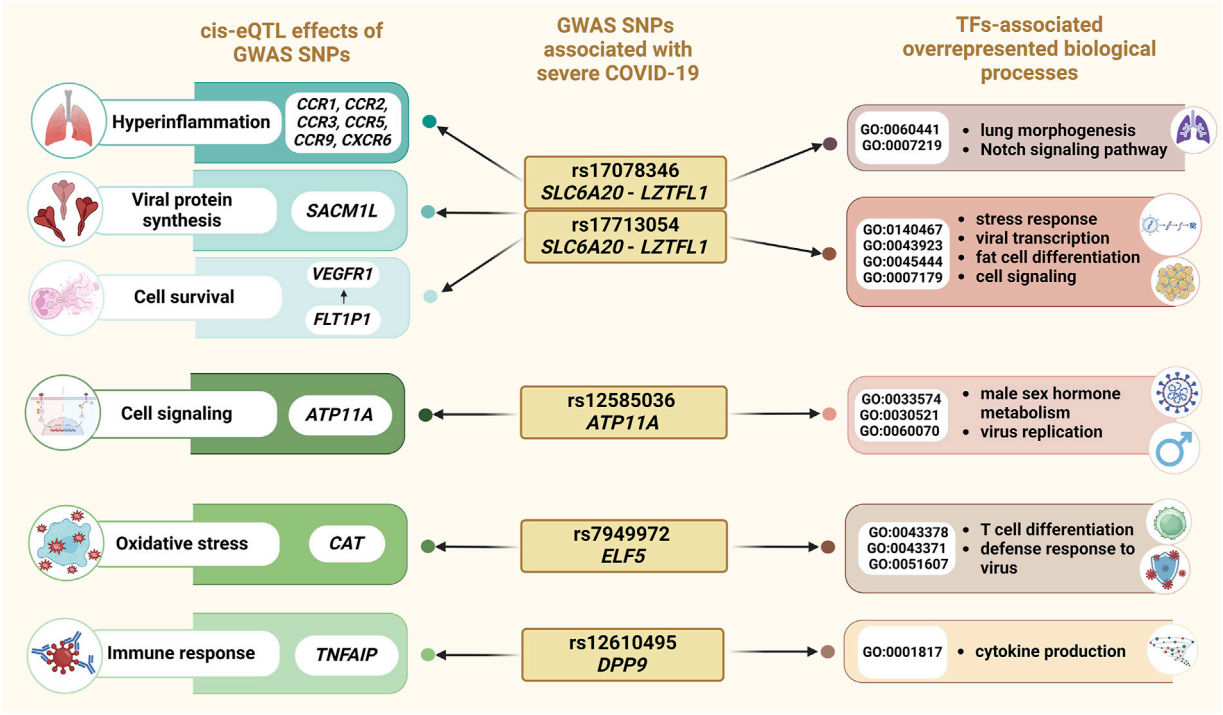


FIGURE 5
Overview of the results of an integrated bioinformatics investigation of severe COVID-19-associated SNPs.

Our bioinformatic analysis revealed that allele A rs17713054 *SLC6A20-LZTFL1* and allele C rs17078346 *SLC6A20-LZTFL1* influence the expression of other genes through cis-eQTL-effects: these SNPs are associated with a decrease in the expression of *FLT1P1* in blood, potentially resulting in dysregulation of vascular endothelial growth factor receptor 1 (VEGFR1) expression (Ye et al., 2015). Numerous studies have demonstrated a correlation between elevated VEGFR1 levels and COVID-19 severity, as well as the ICU admission of COVID-19 patients (Krock et al., 2011; Ackermann et al., 2020; Nagashima et al., 2020; Pine et al., 2020; Miggiolaro et al., 2023; Pius-Sadowska et al., 2023). In addition, we noted eQTL effects of rs17713054 and rs17078346 on the expression levels of chemokine receptors (*CCR1, CCR2, CCR3, CCR5, CCR9,* and *CXCR6*). Previous studies have implicated these chemokine receptors in virus infections and COVID-19 pathogenesis, suggesting their role in lung infiltration by monocytes and macrophages during viral infection, contributing to the hyperinflammation observed in severe COVID-19 cases (Coperchini et al., 2020; Khalil et al., 2021; Mahmoodi et al., 2024). Among other genes with altered expression levels caused by rs17713054 and rs17078346 is *SACM1L,* which was previously identified as a putative causal gene for COVID-19 severity (Wu et al., 2021). SACM1L mediates lipid transfer between closely opposed ER and endosomal membranes with several other lipid transfer proteins (Reinisch and Prinz, 2021). It was found that SACM1L concentrated at the viral factories in infected cells, contrasting its typical distribution in uninfected cells, where it is primarily found in the ER and Golgi apparatus (García-Dorival et al., 2023) (Figure 5).

TFs binding to the risk allele A rs17713054 are associated with positive regulation by host of viral transcription (GO:0043923), integrated stress response signaling (GO:0140467), the transforming growth factor beta receptor signaling pathway (GO:0007179), and fat cell differentiation (GO:0045444), while also resulting in a loss of function in response to hypoxia (GO:0001666). These findings provide insights into the association of rs17713054 with severe COVID-19 and obesity, a known risk factor for severe COVID-19 progression. Risk allele C rs17078346 *SLC6A20-LZTFL1* affects DNA binding to TFs jointly involved in epithelial tube branching involved in lung morphogenesis (GO:0060441), and the Notch signaling pathway (GO:0007219) (Figure 5). These findings suggest its potential role in COVID-19 severity by regulating immune response, and apoptosis.

The correlation between rs17713054 *SLC6A20-LZTFL1* and obesity is supported by previous research indicating that *LZTFL1* may regulate leptin signaling, and participate in the LepRb signaling pathway in the hypothalamus, which controls energy homeostasis (Wei et al., 2018). Elevated levels of circulating leptin are generally attributed to the development of leptin resistance (Zieba et al., 2020), a hallmark of obesity, which is already recognized as a risk factor for severe COVID-19 (Rebello et al., 2020; Maurya et al., 2021). Notably, Lztfl1 knockout mice exhibit hyperphagia, leptin resistance, and obesity (Tomlinson, 2024). Moreover, polyphenolic compounds found in fruits and vegetables, along with regular exercise, have been shown to enhance sensitivity to leptin (Aragonès et al., 2016; Fedewa et al., 2018). Based on these findings, we hypothesize that individuals carrying the allele A of rs17713054 *SLC6A20-LZTFL1*, who consume higher levels of fruit and vegetables and engage in more physical activity, may experience reduced inflammation by lowering serum leptin levels, potentially leading to a less severe course of COVID-19. Additionally, the manifestation of the risk

effects of rs17713054 *SLC6A20-LZTFL1* in patients with low levels of physical activity may be explained by the significant suppression of *Slc6A20* expression observed in mouse models following exercise (Walz et al., 2021). Considering that *SLC6A20* expression is positively associated with infiltrating neutrophils and immune-related signatures (Acar, 2023), the downregulation of this gene through exercise may further contribute to the mitigation of COVID-19 severity. The presence of the rs17713054 *SLC6A20-LZTFL1* association in patients with low consumption of fresh vegetables and fruits—one of the main environmental risk factors for oxidative stress—may be associated with the effect of reactive oxygen species on the expression level of the *SLC6A20* and *LZTFL1* genes. In particular, it was found that hydrogen peroxide, along with plant extracts, may affect the expression of *SLC6A20* mRNA and *LZTFL1* mRNA (Briedé et al., 2010; Tomé-Carneiro et al., 2013).

The association of rs17713054 *SLC6A20-LZTFL*1 with severe COVID-19 in non-smoking individuals can be explained; on the one hand, smoking itself is a known risk factor for severe COVID-19 due to its upregulation of *ACE-2* expression in the lungs, the host receptor for SARS-CoV-2, making smokers more susceptible to the disease (Reddy et al., 2021). This increased susceptibility to COVID-19 in smokers may exceed the effect of rs17713054, leading to the observed association specifically in non-smokers. On the other hand, previous research has shown that smoking affects the expression of genes located near rs17713054, the level of *SLC6A20* mRNA, and the decreased expression of *LZTFL1* (Xiong et al., 2021). Another study showed that benzopyrene, one of the main components of cigarette smoke, increases methylation of the *LZTFL1* gene promoter and exon *SLC6A20* (Jiang et al., 2017) and also reduces the expression of *SLC6A20* mRNA (Qiu et al., 2011; Kreuzer et al., 2020). Considering that increased methylation is a significant regulatory mechanism for decreased gene expression, this finding can be interpreted as further evidence that smoking influences the decreased expression of *LZTFL1* and *SLC6A20*.

Furthermore, our study showed that rs12610495 *DPP9* (risk allele G) is associated with a higher risk of severe COVID-19 in patients with obesity and also affects BMI in patients with severe COVID-19. Additionally, a significant association was found between rs12610495 and thrombodynamic parameters, in particular with prolongation of the time to the start of clot growth (Tlag). Several studies have already pointed to rs12610495 *DPP9* as a risk polymorphic variant for severe COVID-19 (Degenhardt et al., 2022; Horowitz et al., 2022; Thibord et al., 2022; Pairo-Castineira et al., 2023). *DPP9* plays a diverse role in immune regulation: it participates in the activation of inflammasomes (Okondo et al., 2018), its inhibition induces pro-caspase-1-dependent monocyte and macrophage pyroptosis (Okondo et al., 2017). Knockdown of *Dpp9* significantly impairs preadipocytes differentiation (Han et al., 2015), supporting our findings that rs12610495 *DPP9* associates with BMI. This SNP has a high regulatory potential in lung tissue, being marked by the enhancer tags H3K4me1 and H3K27ac as well as by the promoter tag H3K4me3. The risk allele G rs12610495 *DPP9* disrupts the regulation of cytokine production (GO:0001817), potentially leading to dysregulated production of proinflammatory cytokines. This dysregulation may cause excessive immune cell infiltration in pulmonary tissues, leading to tissue damage (Nagashima et al., 2020). In blood, rs12610495 *DPP9* alters the expression through cis-eQTL effects of *TNFAIP8L1*, a member of the TNFAIP family, which plays a modulating role in immune response (Li et al., 2018; Hua et al., 2021). Additionally, Pahl et al. reported that *TNFAIP8L1* levels were significantly downregulated in monocytes from COVID-19 patients compared to healthy controls (Pahl et al., 2022) (Figure 5).

We determined that rs7949972 *ELF5* (effect allele T) had a protective effect only in COVID-19 patients with a BMI <30. In this subgroup, we observed that the protective allele T reduces the clot size at 30 min after coagulation activation (CS) and stationary spatial clot growth rates (Vst). ELF5, a member of the erythroblast transformation-specific (Ets) transcription factor family, has been extensively studied in breast cancer contexts (Chakrabarti et al., 2012; Kalyuga et al., 2012). However, recent research has highlighted its role in COVID-19, revealing that key host factors for SARS-CoV-2 (*Ace2* and *Tmprss4*) are upregulated in *Elf5*-overexpressing AT2 cells (Pietzner et al., 2022). *ELF5,* through cis-eQTL effects, also regulates the expression of *CAT*, an antioxidant enzyme, in whole blood and in the tibial artery. Levels of catalase, along with other markers of oxidative stress, were found to be elevated in COVID-19 patients (Martín-Fernández et al., 2021). Oxidative stress may not only pose a risk for severe COVID-19 but also contribute to the development of atherosclerosis (Sorokin et al., 2015; Sorokin et al., 2016)and atherosclerosis-associated cardiovascular diseases (Vialykh et al., 2012; Bushueva OY. et al., 2015; Bushueva et al., 2021), exacerbating patient prognosis (Hessami et al., 2021). Upon analyzing the impact of the risk allele C rs7949972 *ELF5* on TFs binding sites, we hypothesize that this allele may result in a more severe COVID-19 course. This could be due to its positive regulation of CD8-positive, alpha-beta T cell differentiation (GO:0043378), and negative regulation of CD4-positive, alpha-beta T cell differentiation (GO: 0043371), as well as its involvement in the defense response to viruses (GO:0051607) (Figure 5). These processes may contribute to excessive inflammation and worsen the course of COVID-19. Additionally, data from the Lung Knowledge Portal indicates that protective allele T rs7949972 correlates with an increase in parameters such as forced vital capacity (FVC), forced expired volume in 1 s (FEV1), FEV1 to FVC ratio, and peak expiratory flow.

Finally, allele T rs12585036 *ATP11A* exhibited a protective effect against severe COVID-19, but exclusively in men. The ATPase phospholipid transporting 11A (*ATP11A*) gene encodes a membrane ATPase responsible for translocating phosphatidylserine (PtdSer) (Segawa et al., 2021). Phagocytosis associated with PtdSer translocation could serve as an early event linked to viral infections (Takizawa et al., 1993; Banki et al., 1998). Moreover, PtdSer has been implicated as a potential mechanism or participant in inflammation and coagulation abnormalities in COVID-19 patients (Argañaraz et al., 2020; Wang et al., 2022). We hypothesize that the protective effect of the T allele of rs12585036 *ATP11A* regarding the risk of severe COVID-19 specifically in men is due to the fact that female sex hormones, in particular estradiol, lead to increased expression of *ATP11A* (Logan et al., 2010; Vydra et al., 2019). Considering the fact that the protective T allele is associated with a decrease in *ATP11A* expression through cis-eQTL effects, it can be assumed that the influence of female sex hormones can neutralize this effect by

increasing the level of *ATP11A*. Moreover, bioinformatics analysis revealed that the protective T allele rs12585036 creates DNA binding sites for TFs involved in overrepresented biological processes related to male sex hormone metabolism (response to testosterone (GO: 0033574) and androgen receptor signaling pathway (GO:0030521)) and regulation of canonical Wnt signaling pathway (GO:0060070; FDR = 3.54×10^{-3}), which has been shown to inhibit the replication of SARS-CoV-2 *in vitro*, and reduce viral load, inflammation and clinical symptoms in a mouse model of COVID-19 (Xu et al., 2024). This finding suggests a potential explanation for why SNP rs12585036 *ATP11A* protects against COVID-19 in men.

5 Study limitations

Firstly, our study was limited in its scope, as we were unable to investigate other genes implicated in the progression of severe COVID-19. Secondly, we lacked data on the vaccination status of the control group, as well as laboratory parameters, including venous blood for thrombodynamics testing, which could only be obtained during hospitalization. This limitation prevented us from conducting a formal comparative analysis of laboratory parameters, incl. thrombodynamic parameters between control group patients and patients with severe COVID-19. Additionally, the effectiveness of different types of vaccines remains controversial, adding further complexity to the analysis of data and the role of vaccination in protecting against severe COVID-19. Thirdly, essential environmental factors such as vegetable intake and physical activity levels were not available for the control group, preventing their inclusion in the MB-MDR analysis of gene-environmental interactions.

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.

Ethics statement

The studies involving humans were approved by The Ethical Review Committee of Kursk State Medical University. The studies were conducted in accordance with the local legislation and institutional requirements. The 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

AL: Formal Analysis, Writing–original draft. KK: Formal Analysis, Writing–original draft. AK: Formal Analysis, Investigation, Writing–original draft. VS: Investigation, Methodology, Validation, Visualization, Writing–original draft. YO: Investigation, Writing–review and editing. OB: Conceptualization, Data curation, Funding acquisition, Methodology, Resources, Supervision, Validation, Visualization, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by Kursk State Medical University, Kursk, Russia.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1434681/full#supplementary-material

References

Acar, A. (2023). Pan-cancer analysis of the COVID-19 causal gene SLC6A20. *ACS Omega* 8, 13153–13161. doi:10.1021/acsomega.3c00407

Ackermann, M., Verleden, S. E., Kuehnel, M., Haverich, A., Welte, T., Laenger, F., et al. (2020). Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in Covid-19. *N. Engl. J. Med.* 383, 120–128. doi:10.1056/NEJMoa2015432

Amine, E. K., Baba, N. H., Belhadj, M., Deurenberg-Yap, M., Djazayery, A., Forrestre, T., et al. (2003). Diet, nutrition and the prevention of chronic diseases. *World Health Organ. Tech. Rep. Ser.* doi:10.1093/ajcn/60.4.644a

Aragonès, G., Ardid-Ruiz, A., Ibars, M., Suárez, M., and Bladé, C. (2016). Modulation of leptin resistance by food compounds. *Mol. Nutr. and Food Res.* 60, 1789–1803. doi:10.1002/mnfr.201500964

Argañaraz, G. A., Palmeira, J. da F., and Argañaraz, E. R. (2020). Phosphatidylserine inside out: a possible underlying mechanism in the inflammation and coagulation abnormalities of COVID-19. *Cell. Commun. Signal* 18, 190. doi:10.1186/s12964-020-00687-7

Banki, K., Hutter, E., Gonchoroff, N. J., and Perl, A. (1998). Molecular ordering in HIV-induced apoptosis. Oxidative stress, activation of caspases, and cell survival are regulated by transaldolase. *J. Biol. Chem.* 273, 11944–11953. doi:10.1074/jbc.273.19.11944

Belykh, A. E., Soldatov, V. O., Stetskaya, T. A., Kobzeva, K. A., Soldatova, M. O., Polonikov, A. V., et al. (2023). Polymorphism of SERF2, the gene encoding a heat-resistant obscure (Hero) protein with chaperone activity, is a novel link in ischemic stroke. *IBRO Neurosci. Rep.* 14, 453–461. doi:10.1016/j.ibneur.2023.05.004

Briedé, J. J., van Delft, J. M., de Kok, T. M., van Herwijnen, M. H., Maas, L. M., Gottschalk, R. W., et al. (2010). Global gene expression analysis reveals differences in cellular responses to hydroxyl-and superoxide anion radical–induced oxidative stress in caco-2 cells. *Toxicol. Sci.* 114, 193–203. doi:10.1093/toxsci/kfp309

Bushueva, O. (2020). Single nucleotide polymorphisms in genes encoding xenobiotic metabolizing enzymes are associated with predisposition to arterial hypertension. *Res. Results Biomed.* 6, 447–456. doi:10.18413/2658-6533-2020-6-4-0-1

Bushueva, O., Barysheva, E., Markov, A., Belykh, A., Koroleva, I., Churkin, E., et al. (2021). DNA Hypomethylation of the MPO gene in peripheral blood Leukocytes is associated with cerebral stroke in the Acute phase. *J. Mol. Neurosci.* 71, 1914–1932. doi:10.1007/s12031-021-01840-8

Bushueva, O., Solodilova, M., Ivanov, V., and Polonikov, A. (2015a). Gender-specific protective effect of the −463G>A polymorphism of myeloperoxidase gene against the risk of essential hypertension in Russians. *J. Am. Soc. Hypertens.* 9, 902–906. doi:10.1016/j.jash.2015.08.006

Bushueva, O. Y., Ivanov, V. P., Ryzhaeva, V. N., Ponomarenko, I. V., Churnosov, M. I., and Polonikov, A. V. (2016). Association of the -844G>A polymorphism in the catalase gene with the increased risk of essential hypertension in smokers. *Ter. Arkh* 88, 50–54. doi:10.17116/terarkh201688950-54

Bushueva, O. Y., Stetskaya, T. A., Polonikov, A. V., and Ivanov, V. P. (2015b). The relationship between polymorphism 640A>G of the CYBA gene with the risk of ischemic stroke in the population of the Central Russia. *Zh Nevrol. Psikhiatr Im. S S Korsakova* 115, 38–41. doi:10.17116/jnevro20151159238-41

Calle, M. L., Urrea, V., Malats, N., and Van Steen, K. (2010). mbmdr: an R package for exploring gene–gene interactions associated with binary or quantitative traits. *Bioinformatics* 26, 2198–2199. doi:10.1093/bioinformatics/btq352

Carvalho, C. R. R., Lamas, C. A., Chate, R. C., Salge, J. M., Sawamura, M. V. Y., Albuquerque, A. L. P. de, et al. (2023). Long-term respiratory follow-up of ICU hospitalized COVID-19 patients: prospective cohort study. *PLoS ONE* 18, e0280567. doi:10.1371/journal.pone.0280567

Chakrabarti, R., Hwang, J., Andres Blanco, M., Wei, Y., Lukačišin, M., Romano, R.-A., et al. (2012). Elf5 inhibits the epithelial–mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2. *Nat. Cell. Biol.* 14, 1212–1222. doi:10.1038/ncb2607

Collins, R., Vallières, F., and McDermott, G. (2023). The experiences of post-ICU COVID-19 survivors: an existential perspective using interpretative phenomenological analysis. *Qual. Health Res.* 33, 589–600. doi:10.1177/10497323231164556

Consortium, G. O. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi:10.1093/nar/gky1055

Consortium, G. T.Ex (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi:10.1126/science.aaz1776

Coperchini, F., Chiovato, L., Croce, L., Magri, F., and Rotondi, M. (2020). The cytokine storm in COVID-19: an overview of the involvement of the chemokine/chemokine-receptor system. *Cytokine and Growth Factor Rev.* 53, 25–32. doi:10.1016/j.cytogfr.2020.05.003

Degenhardt, F., Ellinghaus, D., Juzenas, S., Lerga-Jaso, J., Wendorff, M., Maya-Miles, D., et al. (2022). Detailed stratified GWAS analysis for severe COVID-19 in four European populations. *Hum. Mol. Genet.* 31, 3945–3966. doi:10.1093/hmg/ddac158

Downes, D. J., Cross, A. R., Hua, P., Roberts, N., Schwessinger, R., Cutler, A. J., et al. (2021). Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. *Nat. Genet.* 53, 1606–1615. doi:10.1038/s41588-021-00955-3

Fedewa, M. V., Hathaway, E. D., Ward-Ritacco, C. L., Williams, T. D., and Dobbs, W. C. (2018). The effect of chronic exercise training on leptin: a systematic review and meta-analysis of randomized controlled trials. *Sports Med.* 48, 1437–1450. doi:10.1007/s40279-018-0897-1

García-Dorival, I., Cuesta-Geijo, M. Á., Galindo, I., del Puerto, A., Barrado-Gil, L., Urquiza, J., et al. (2023). Elucidation of the cellular interactome of african swine fever virus fusion proteins and identification of potential therapeutic targets. *Viruses* 15, 1098. doi:10.3390/v15051098

Garg, E., Arguello-Pascualli, P., Vishnyakova, O., Halevy, A. R., Yoo, S., Brooks, J. D., et al. (2024). Canadian COVID-19 host genetics cohort replicates known severity associations. *PLoS Genet.* 20, e1011192. doi:10.1371/journal.pgen.1011192

Han, R., Wang, X., Bachovchin, W., Zukowska, Z., and Osborn, J. W. (2015). Inhibition of dipeptidyl peptidase 8/9 impairs preadipocyte differentiation. *Sci. Rep.* 5, 12348. doi:10.1038/srep12348

Hessami, A., Shamshirian, A., Heydari, K., Pourali, F., Alizadeh-Navaei, R., Moosazadeh, M., et al. (2021). Cardiovascular diseases burden in COVID-19: systematic review and meta-analysis. *Am. J. Emerg. Med.* 46, 382–391. doi:10.1016/j.ajem.2020.10.022

Horowitz, J. E., Kosmicki, J. A., Damask, A., Sharma, D., Roberts, G. H., Justice, A. E., et al. (2022). Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. *Nat. Genet.* 54, 382–392. doi:10.1038/s41588-021-01006-7

Hua, J., Zhuang, G., and Qi, Z. (2021). Current research status of TNFAIP8 in tumours and other inflammatory conditions (Review). *Int. J. Oncol.* 59, 46. doi:10.3892/ijo.2021.5226

Ivanova, T. A. (2024). Sex-specific features of interlocus interactions determining susceptibility to hypertension. *Res. Results Biomed.* 10, 53–68. doi:10.18413/2658-6533-2024-10-1-0-3

Jiang, C.-L., He, S.-W., Zhang, Y.-D., Duan, H.-X., Huang, T., Huang, Y.-C., et al. (2017). Air pollution and DNA methylation alterations in lung cancer: a systematic and comparative study. *Oncotarget* 8, 1369–1391. doi:10.18632/oncotarget.13622

Joshee, S., Vatti, N., and Chang, C. (2022). Long-term effects of COVID-19. *Mayo Clin. Proc.* 97, 579–599. doi:10.1016/j.mayocp.2021.12.017

Kalyuga, M., Gallego-Ortega, D., Lee, H. J., Roden, D. L., Cowley, M. J., Caldon, C. E., et al. (2012). ELF5 suppresses estrogen sensitivity and underpins the acquisition of antiestrogen resistance in luminal breast cancer. *PLOS Biol.* 10, e1001461. doi:10.1371/journal.pbio.1001461

Khalil, B. A., Elemam, N. M., and Maghazachi, A. A. (2021). Chemokines and chemokine receptors during COVID-19 infection. *Comput. Struct. Biotechnol. J.* 19, 976–988. doi:10.1016/j.csbj.2021.01.034

Kim, J. S., Lee, J. Y., Yang, J. W., Lee, K. H., Effenberger, M., Szpirt, W., et al. (2021). Immunopathogenesis and treatment of cytokine storm in COVID-19. *Theranostics* 11, 316–329. doi:10.7150/thno.49713

Kobzeva, K. A., Shilenok, I. V., Belykh, A. E., Gurtovoy, D. E., Bobyleva, L. A., Krapiva, A. B., et al. (2022). C9orf16 (BBLN) gene, encoding a member of Hero proteins, is a novel marker in ischemic stroke risk. *Res. Results Biomed.* 8, 278–292. doi:10.18413/2658-6533-2022-8-3-0-2

Kobzeva, K. A., Soldatova, M. O., Stetskaya, T. A., Soldatov, V. O., Deykin, A. V., Freidin, M. B., et al. (2023). Association between HSPA8 gene variants and ischemic stroke: a pilot study providing additional evidence for the role of heat shock proteins in disease pathogenesis. *Genes* 14, 1171. doi:10.3390/genes14061171

Koressaar, T., and Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23, 1289–1291. doi:10.1093/bioinformatics/btm091

Kousathanas, A., Pairo-Castineira, E., Rawlik, K., Stuckey, A., Odhams, C. A., Walker, S., et al. (2022). Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* 607, 97–103. doi:10.1038/s41586-022-04576-6

Kreuzer, K., Böhmert, L., Alhalabi, D., Buhrke, T., Lampen, A., and Braeuning, A. (2020). Identification of a transcriptomic signature of food-relevant genotoxins in human HepaRG hepatocarcinoma cells. *Food Chem. Toxicol.* 140, 111297. doi:10.1016/j.fct.2020.111297

Krock, B. L., Skuli, N., and Simon, M. C. (2011). Hypoxia-induced angiogenesis: good and evil. *Genes and Cancer* 2, 1117–1133. doi:10.1177/1947601911423654

Lee, J.-W., Lee, I.-H., Sato, T., Kong, S. W., and Iimura, T. (2021). Genetic variation analyses indicate conserved SARS-CoV-2–host interaction and varied genetic adaptation in immune response factors in modern human evolution. *Dev. Growth and Differ.* 63, 219–227. doi:10.1111/dgd.12717

Li, T., Wang, W., Gong, S., Sun, H., Zhang, H., Yang, A.-G., et al. (2018). Genome-wide analysis reveals TNFAIP8L2 as an immune checkpoint regulator of inflammation and metabolism. *Mol. Immunol.* 99, 154–162. doi:10.1016/j.molimm.2018.05.007

Logan, P. C., Ponnampalam, A. P., Rahnama, F., Lobie, P. E., and Mitchell, M. D. (2010). The effect of DNA methylation inhibitor 5-Aza-2′-deoxycytidine on human endometrial stromal cells. *Hum. Reprod.* 25, 2859–2869. doi:10.1093/humrep/deq238

Ma, Y., Deng, J., Liu, Q., Du, M., Liu, M., and Liu, J. (2022). Long-term consequences of COVID-19 at 6 Months and above: a systematic review and meta-analysis. *Int. J. Environ. Res. Public Health* 19, 6865. doi:10.3390/ijerph19116865

Mahmoodi, M., Mohammadi Henjeroei, F., Hassanshahi, G., and Nosratabadi, R. (2024). Do chemokine/chemokine receptor axes play paramount parts in trafficking and oriented locomotion of monocytes/macrophages toward the lungs of COVID-19 infected patients? A systematic review. *Cytokine* 175, 156497. doi:10.1016/j.cyto.2023.156497

Martín-Fernández, M., Aller, R., Heredia-Rodríguez, M., Gómez-Sánchez, E., Martínez-Paz, P., Gonzalo-Benito, H., et al. (2021). Lipid peroxidation as a hallmark of severity in COVID-19 patients. *Redox Biol.* 48, 102181. doi:10.1016/j.redox.2021.102181

Maurya, R., Sebastian, P., Namdeo, M., Devender, M., and Gertler, A. (2021). COVID-19 severity in obesity: leptin and inflammatory cytokine interplay in the link between high morbidity and mortality. *Front. Immunol.* 12, 649359. doi:10.3389/fimmu.2021.649359

Miggiolaro, AFRS, da Silva, F. P. G., Wiedmer, D. B., Godoy, T. M., Borges, N. H., Piper, G. W., et al. (2023). COVID-19 and pulmonary angiogenesis: the possible role of hypoxia and hyperinflammation in the overexpression of proteins involved in alveolar vascular dysfunction. *Viruses* 15, 706. doi:10.3390/v15030706

Nagashima, S., Mendes, M. C., Camargo Martins, A. P., Borges, N. H., Godoy, T. M., Miggiolaro, A. F. R. dos S., et al. (2020). Endothelial dysfunction and thrombosis in patients with COVID-19—brief report. *Arteriosclerosis, Thrombosis, Vasc. Biol.* 40, 2404–2407. doi:10.1161/ATVBAHA.120.314860

Okondo, M. C., Johnson, D. C., Sridharan, R., Go, E. B., Chui, A. J., Wang, M. S., et al. (2017). DPP8 and DPP9 inhibition induces pro-caspase-1-dependent monocyte and macrophage pyroptosis. *Nat. Chem. Biol.* 13, 46–53. doi:10.1038/nchembio.2229

Okondo, M. C., Rao, S. D., Taabazuing, C. Y., Chui, A. J., Poplawski, S. E., Johnson, D. C., et al. (2018). Inhibition of dpp8/9 activates the Nlrp1b inflammasome. *Cell. Chem. Biol.* 25, 262–267.e5. doi:10.1016/j.chembiol.2017.12.013

Oyu, B., Bulgakova, I. V., Ivanov, V. P., and Polonikov, A. V. (2015). Association of flavin monooxygenase gene E158K polymorphism with chronic heart disease risk. *Bull. Exp. Biol. Med.* 159, 776–778. doi:10.1007/s10517-015-3073-8

Pahl, M. C., Le Coz, C., Su, C., Sharma, P., Thomas, R. M., Pippin, J. A., et al. (2022). Implicating effector genes at COVID-19 GWAS loci using promoter-focused Capture-C in disease-relevant immune cell types. *Genome Biol.* 23, 125. doi:10.1186/s13059-022-02691-1

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., et al. (2021). Genetic mechanisms of critical illness in COVID-19. *Nature* 591, 92–98. doi:10.1038/s41586-020-03065-y

Pairo-Castineira, E., Rawlik, K., Bretherick, A. D., Qi, T., Wu, Y., Nassiri, I., et al. (2023). GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. *Nature* 617, 764–768. doi:10.1038/s41586-023-06034-3

Pietzner, M., Chua, R. L., Wheeler, E., Jechow, K., Willett, J. D. S., Radbruch, H., et al. (2022). ELF5 is a potential respiratory epithelial cell-specific risk gene for severe COVID-19. *Nat. Commun.* 13, 4484. doi:10.1038/s41467-022-31999-6

Pine, A. B., Meizlish, M. L., Goshua, G., Chang, C.-H., Zhang, H., Bishai, J., et al. (2020). Circulating markers of angiogenesis and endotheliopathy in COVID-19. *Pulm. Circ.* 10, 2045894020966547. doi:10.1177/2045894020966547

Pius-Sadowska, E., Kulig, P., Niedźwiedź, A., Baumert, B., Łuczkowska, K., Rogińska, D., et al. (2023). VEGFR and DPP-IV as markers of severe COVID-19 and predictors of ICU admission. *Int. J. Mol. Sci.* 24, 17003. doi:10.3390/ijms242317003

Polonikov, A. V., Samgina, T. A., Nazarenko, P. M., Bushueva, O. Y., and Ivanov, V. P. (2017). Alcohol consumption and cigarette smoking are important modifiers of the association between Acute pancreatitis and the PRSS1-PRSS2 locus in men. *Pancreas* 46, 230–236. doi:10.1097/MPA.0000000000000729

Pretorius, E., Venter, C., Laubscher, G. J., Lourens, P. J., Steenkamp, J., and Kell, D. B. (2020). Prevalence of readily detected amyloid blood clots in 'unclotted' Type 2 Diabetes Mellitus and COVID-19 plasma: a preliminary report. *Cardiovasc Diabetol.* 19, 193. doi:10.1186/s12933-020-01165-7

Qiu, C., Cheng, S., Xia, Y., Peng, B., Tang, Q., and Tu, B. (2011). Effects of subchronic benzo (a) pyrene exposure on neurotransmitter receptor gene expression in the rat hippocampus related with spatial learning and memory change. *Toxicology* 289, 83–90. doi:10.1016/j.tox.2011.07.012

Rebello, C. J., Kirwan, J. P., and Greenway, F. L. (2020). Obesity, the most common comorbidity in SARS-CoV-2: is leptin the link? *Int. J. Obes.* 44, 1810–1817. doi:10.1038/s41366-020-0640-5

Reddy, R. K., Charles, W. N., Sklavounos, A., Dutt, A., Seed, P. T., and Khajuria, A. (2021). The effect of smoking on COVID-19 severity: a systematic review and meta-analysis. *J. Med. Virology* 93, 1045–1056. doi:10.1002/jmv.26389

Reinisch, K. M., and Prinz, W. A. (2021). Mechanisms of nonvesicular lipid transport. *J. Cell. Biol.* 220, e202012058. doi:10.1083/jcb.202012058

Rescenko, R., Peculis, R., Briviba, M., Ansone, L., Terentjeva, A., Litvina, H. D., et al. (2021). Replication of LZTFL1 gene region as a susceptibility locus for COVID-19 in Latvian population. *Virol. Sin.* 36, 1241–1244. doi:10.1007/s12250-021-00448-x

Roberts, G. H., Park, D. S., Coignet, M. V., McCurdy, S. R., Knight, S. C., Partha, R., et al. 2020. *AncestryDNA COVID-19 host genetic study identifies three novel loci*

Robinot, R., Hubert, M., de Melo, G. D., Lazarini, F., Bruel, T., Smith, N., et al. (2021). SARS-CoV-2 infection induces the dedifferentiation of multiciliated cells and impairs mucociliary clearance. *Nat. Commun.* 12, 4354. doi:10.1038/s41467-021-24521-x

Roozbehani, M., Keyvani, H., Razizadeh, M., Yousefi, P., Gholami, A., Tabibzadeh, A., et al. (2023). LZTFL1 rs17713054 polymorphism as an indicator allele for COVID-19 severity. *Mol. Genet. Microbiol. Virol.* 38, 124–128. doi:10.3103/S0891416823020088

Segawa, K., Kikuchi, A., Noji, T., Sugiura, Y., Hiraga, K., Suzuki, C., et al. (2021). A sublethal ATP11A mutation associated with neurological deterioration causes aberrant phosphatidylcholine flipping in plasma membranes. *J. Clin. Invest.* 131, e148005. doi:10.1172/JCI148005

Severe Covid-19 GWAS Group, Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., et al. (2020). Genomewide association study of severe Covid-19 with respiratory failure. *N. Engl. J. Med.* 383, 1522–1534. doi:10.1056/NEJMoa2020283

Shilenok, I., Kobzeva, K., Stetskaya, T., Freidin, M., Soldatova, M., Deykin, A., et al. (2023). SERPINE1 mRNA binding protein 1 is associated with ischemic stroke risk: a comprehensive molecular–genetic and bioinformatics analysis of SERBP1 SNPs. *Int. J. Mol. Sci.* 24, 8716. doi:10.3390/ijms24108716

Shin, S., Hudson, R., Harrison, C., Craven, M., and Keleş, S. (2019). atSNP Search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding. *Bioinformatics* 35, 2657–2659. doi:10.1093/bioinformatics/bty1010

Silva, M. J. A., Ribeiro, L. R., Gouveia, M. I. M., Marcelino, B. dos R., Santos, C. S. dos, Lima, K. V. B., et al. (2023). Hyperinflammatory response in COVID-19: a systematic review. *Viruses* 15, 553. doi:10.3390/v15020553

Sorokin, A., Kotani, K., Bushueva, O., Taniguchi, N., and Lazarenko, V. (2015). The cardio-ankle vascular index and ankle-brachial index in young russians. *J. Atheroscler. Thromb.* 22, 211–218. doi:10.5551/jat.26104

Sorokin, A. V., Kotani, K., Bushueva, O. Y., and Polonikov, A. V. (2016). Antioxidant-related gene polymorphisms associated with the cardio-ankle vascular index in young Russians. *Cardiol. Young* 26, 677–682. doi:10.1017/S104795111500102X

Stetskaya, T. A., Kobzeva, K. A., Zaytsev, S. M., Shilenok, I. V., Komkova, G. V., Goryainova, N. V., et al. (2024). HSPD1 gene polymorphism is associated with an increased risk of ischemic stroke in smokers. *Res. Results Biomed.* 10, 175–186. doi:10.18413/2658-6533-2024-10-2-0-1

Tadbir Vajargah, K., Zargarzadeh, N., Ebrahimzadeh, A., Mousavi, S. M., Mobasheran, P., Mokhtari, P., et al. (2022). Association of fruits, vegetables, and fiber intake with COVID-19 severity and symptoms in hospitalized patients: a cross-sectional study. *Front. Nutr.* 9, 934568. doi:10.3389/fnut.2022.934568

Takizawa, T., Matsukawa, S., Higuchi, Y., Nakamura, S., Nakanishi, Y., and Fukuda, R. (1993). Induction of programmed cell death (apoptosis) by influenza virus infection in tissue culture cells. *J. Gen. Virol.* 74 (11), 2347–2355. doi:10.1099/0022-1317-74-11-2347

Tavakol, Z., Ghannadi, S., Tabesh, M. R., Halabchi, F., Noormohammadpour, P., Akbarpour, S., et al. (2023). Relationship between physical activity, healthy lifestyle and COVID-19 disease severity; a cross-sectional study. *J. Public Health (Berl)* 31, 267–275. doi:10.1007/s10389-020-01468-9

Thibord, F., Chan, M. V., Chen, M.-H., and Johnson, A. D. (2022). A year of COVID-19 GWAS results from the GRASP portal reveals potential genetic risk factors. *HGG Adv.* 3, 100095. doi:10.1016/j.xhgg.2022.100095

Tomé-Carneiro, J., Larrosa, M., Yáñez-Gascón, M. J., Dávalos, A., Gil-Zamorano, J., Gonzálvez, M., et al. (2013). One-year supplementation with a grape extract containing resveratrol modulates inflammatory-related microRNAs and cytokines expression in peripheral blood mononuclear cells of type 2 diabetes and hypertensive patients with coronary artery disease. *Pharmacol. Res.* 72, 69–82. doi:10.1016/j.phrs.2013.03.011

Tomlinson, J. W. (2024). Bardet-Biedl syndrome: a focus on genetics, mechanisms and metabolic dysfunction. *Obes. Metabolism* 26, 13–24. doi:10.1111/dom.15480

Udomsinprasert, W., Nontawong, N., Saengsiwaritt, W., Panthan, B., Jiaranai, P., Thongchompoo, N., et al. (2023). Host genetic polymorphisms involved in long-term symptoms of COVID-19. *Emerg. Microbes and Infect.* 12, 2239952. doi:10.1080/22221751.2023.2239952

Vialykh, E. K., Solidolova, M. A., Bushueva, O. I., Bulgakova, I. V., and Polonikov, A. V. (2012). Catalase gene polymorphism is associated with increased risk of cerebral stroke in hypertensive patients. *Zh Nevrol. Psikhiatr Im. S S Korsakova* 112, 3–7.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *BioRxiv*, 447367. doi:10.1038/s41588-021-00913-z

Vydra, N., Janus, P., Toma-Jonik, A., Stokowy, T., Mrowiec, K., Korfanty, J., et al. (2019). 17β-Estradiol activates HSF1 via MAPK signaling in erα-positive breast cancer cells. *Cancers* 11, 1533. doi:10.3390/cancers11101533

Walz, C., Brenmoehl, J., Trakooljul, N., Noce, A., Caffier, C., Ohde, D., et al. (2021). Control of protein and energy metabolism in the pituitary gland in response to three-week running training in adult male mice. *Cells* 10, 736. doi:10.3390/cells10040736

Wang, J., Yu, C., Zhuang, J., Qi, W., Jiang, J., Liu, X., et al. (2022). The role of phosphatidylserine on the membrane in immunity and blood coagulation. *Biomark. Res.* 10, 4. doi:10.1186/s40364-021-00346-0

Wang, S., Peng, H., Chen, F., Liu, C., Zheng, Q., Wang, M., et al. (2023). Identification of genetic loci jointly influencing COVID-19 and coronary heart diseases. *Hum. Genomics* 17, 101. doi:10.1186/s40246-023-00547-8

Ward, L. D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi:10.1093/nar/gkr917

Wei, Q., Gu, Y.-F., Zhang, Q.-J., Yu, H., Peng, Y., Williams, K. W., et al. (2018). Lztfl1/BBS17 controls energy homeostasis by regulating the leptin signaling in the hypothalamic neurons. *J. Mol. Cell. Biol.* 10, 402–410. doi:10.1093/jmcb/mjy022

Wu, L., Zhu, J., Liu, D., Sun, Y., and Wu, C. (2021). An integrative multiomics analysis identifies putative causal genes for COVID-19 severity. *Genet. Med.* 23, 2076–2086. doi:10.1038/s41436-021-01243-5

Xiong, R., Wu, Y., Wu, Q., Muskhelishvili, L., Davis, K., Tripathi, P., et al. (2021). Integration of transcriptome analysis with pathophysiological endpoints to evaluate cigarette smoke toxicity in an *in vitro* human airway tissue model. *Archives Toxicol.* 95, 1739–1761. doi:10.1007/s00204-021-03008-0

Xu, Z., Elaish, M., Wong, C. P., Hassan, B. B., Lopez-Orozco, J., Felix-Lopez, A., et al. (2024). The Wnt/β-catenin pathway is important for replication of SARS-CoV-2 and other pathogenic RNA viruses. *Npj Viruses* 2, 6–15. doi:10.1038/s44298-024-00018-4

Ye, X., Fan, F., Bhattacharya, R., Bellister, S., Boulbes, D. R., Wang, R., et al. (2015). VEGFR-1 pseudogene expression and regulatory function in human colorectal cancer cells. *Mol. Cancer Res.* 13, 1274–1282. doi:10.1158/1541-7786.MCR-15-0061

Yedjou, C. G., Alo, R. A., Liu, J., Enow, J., Ngnepiepa, P., Long, R., et al. (2021). Chemo-preventive effect of vegetables and fruits consumption on the COVID-19 pandemic. *J. Nutr. Food Sci.* 4, 029.

Zieba, D. A., Biernat, W., and Barć, J. (2020). Roles of leptin and resistin in metabolism, reproduction, and leptin resistance. *Domest. Anim. Endocrinol.* 73, 106472. doi:10.1016/j.domaniend.2020.106472

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 01 November 2024
DOI 10.3389/fgene.2024.1445033

OPEN ACCESS

EDITED BY
Yuriy L. Orlov,
I.M.Sechenov First Moscow State Medical University, Russia

REVIEWED BY
Rocío Salceda,
National Autonomous University of Mexico, Mexico
Shuyin Bao,
Inner Mongolia University for Nationalities, China

*CORRESPONDENCE
Xiaoyu Liu,
lxy2002sk@ntu.edu.cn
Xuchu Duan,
dxd2002sk@ntu.edu.cn
Yuqing Chen,
yuqingchen0312@163.com

RECEIVED 06 June 2024
ACCEPTED 18 October 2024
PUBLISHED 01 November 2024

CITATION
Yao L, Xu J, Zhang X, Tang Z, Chen Y, Liu X and Duan X (2024) Bioinformatical analysis and experimental validation of endoplasmic reticulum stress-related biomarker genes in type 2 diabetes mellitus.
Front. Genet. 15:1445033.
doi: 10.3389/fgene.2024.1445033

Bioinformatical analysis and experimental validation of endoplasmic reticulum stress-related biomarker genes in type 2 diabetes mellitus

Lili Yao[1], Jie Xu[1], Xu Zhang[2], Zhuqi Tang[1], Yuqing Chen[1]*, Xiaoyu Liu[1]* and Xuchu Duan[1]*

[1]Key Laboratory of Neuroregeneration of Jiangsu and Ministry of Education, Nantong Laboratory of Development and Diseases, Department of Endocrine, Department of Pharmacy, School of Life Science, Co-innovation Center of Neuroregeneration, Medical School, Affiliated Hospital of Nantong University, Nantong University, Nantong, China, [2]Clinical Medical Research Center, Wuxi No. 2 People's Hospital, Jiangnan University Medical Center, Wuxi, China

Introduction: Endoplasmic reticulum stress (ERS) is a prominent etiological factor in the pathogenesis of diabetes. Nevertheless, the mechanisms through which ERS contributes to the development of diabetes remain elusive.

Methods: Transcriptional expression profiles from the Gene Expression Omnibus (GEO) datasets were analyzed and compared to obtain the differentially expressed genes (DEGs) in T2DM. Following the intersection with ERS associated genes, the ERS related T2DM DEGs were identified. Receiver operating characteristic (ROC) and Least Absolute Shrinkage and Selection Operator (LASSO) analysis were performed to screen out the ERS related biomarker genes and validate their diagnostic values. Gene expression level was detected by qPCR and Elisa assays in diabetic mice and patient serum samples.

Results: By analyzing the transcriptional expression profiles of the GEO datasets, 49 T2DM-related DEGs were screened out in diabetic islets. RTN1, CLGN, PCSK1, IAPP, ILF2, IMPA1, CCDC47, and PTGES3 were identified as ERS-related DEGs in T2DM, which were revealed to be involved in protein folding, membrane composition, and metabolism regulation. ROC and LASSO analysis further screened out CLGN, ILF2, and IMPA1 as biomarker genes with high value and reliability for diagnostic purposes. These three genes were then demonstrated to be targeted by the transcription factors and miRNAs, including CEBPA, CEBPB, miR-197-5p, miR-6133, and others. Among these miRNAs, the expression of miR-197-5p, miR-320c, miR-1296-3P and miR-6133 was down-regulated, while that of miR-4462, miR-4476-5P and miR-7851-3P was up-regulated in diabetic samples. Small molecular drugs, including D002994, D001564, and others, were predicted to target these genes potentially. qPCR and Elisa analysis both validated the same expression alteration trend of the ERS-related biomarker genes in diabetic mice and T2DM patients.

Discussion: These findings will offer innovative perspectives for clinical diagnosis and treatment strategies for T2DM.

KEYWORDS

type 2 diabetes mellitus, biomarker gene, endoplasmic reticulum stress, bioinformatical analysis, experimental validation, diagnosis

Introduction

Type 2 diabetes mellitus (T2DM) is presently recognized as the third most severe chronic ailment globally, following cancer and cardiovascular disease, posing a significant threat to human wellbeing (Chen et al., 2012). The pathogenesis of T2DM encompasses intricate mechanisms, including immune dysfunction (Geerlings and Hoepelman, 1999), oxidative stress (Darenskaya et al., 2021), mitochondrial dysfunction (Wada and Nakatsuka, 2016), glucose toxicity (Robertson and Harmon, 2006) and endoplasmic reticulum stress (ERS) (Lee and Lee, 2022). The complexities of diabetic complications, such as diabetic neuropathy and nephropathy, pose challenges in the development of efficacious medications. However, due to the limited understanding of the pathogenesis, clinically, there are no reliable biomarkers for early detection of diabetes, and treatment continues to be challenging.

The endoplasmic reticulum (ER), one of the most extensive organelles within cells, possesses a vast membrane structure and serves as the site for initial protein synthesis and folding. The disruption of internal environmental stability in the endoplasmic reticulum, caused by the stimulation of factors such as ion storage and lipid synthesis, can lead to protein misfolding, referred to as ERS (Cnop et al., 2017). Excessive ERS can impede cellular function, resulting in aberrant protein synthesis and degradation, cellular oxidation, apoptosis, inflammatory responses, and the onset of various diseases (Zhang et al., 2022). ERS plays a pivotal role in the pathogenesis and progression of numerous disorders, including diabetes, Alzheimer's disease, Parkinson's disease, and other related conditions (Ghemrawi and Khair, 2020).

Previous studies have been conducted to explore the interplay between ER stress and diabetes and its complications. A prior study has elucidated the involvement of the IP3R1-GRP75-VDAC1 complex in mediating ER stress and mitochondrial oxidative stress and its significant role in atrial remodeling in diabetes (Yuan et al., 2022). Additionally, it has been reported that in diabetic mice, ER stress and autophagy play a regulatory role in neuronal survival and death, with the ER stress pathway potentially contributing to diabetes-induced neurotoxicity and cognitive impairment (Kong et al., 2018). These findings suggested the vital roles of ERS in the occurrence and progression of diabetes and its complications.

T2DM is primarily caused by decreased pancreatic β cells and insulin secretion (Eizirik et al., 2020; Sun et al., 2023). ERS has been considered a critical contributing factor to unfolded protein response (UPR) and the dysfunction of β cells, which is essential factor T2DM pathogenesis (Yong et al., 2021; Sak et al., 2024; Zhang et al., 2023). To explore the critical ERS-related biomarker genes in T2DM, we investigated the association between ERS-related genes and the differentially expressed genes (DEGs) in islets of T2DM patients through bioinformatical analysis. The GEO dataset GSE25724 was employed for Gene Set Enrichment Analysis (GSEA) and for identifying DEGs in T2DM. Following the intersection with the ERS-related genes, T2DM-associated ERS-DEGs were obtained. Three critical biomarker genes were further screened and validated by Receiver Operating Characteristic (ROC) and Least Absolute Shrinkage and Selection Operator (LASSO) analysis. Ultimately, the relative expression levels of the critical biomarker genes were determined using qPCR on constructed T2DM mice and T2DM patients. The findings will help to further understand the pathogenesis and provide novel insights into the clinical diagnosis and treatments of T2DM.

Materials and methods

Data collection

The dataset GSE25724, which contains the transcriptional expression profiles of normal and diabetic tissue samples, was obtained from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) and used for the analysis of DEGs in T2DM. The expression profile analysis was conducted on the GPL96 platform, and the sequencing was performed using Affymetrix Human Genome U133A Array technology. The subjects included seven healthy human islet tissue samples and six islet tissue samples from patients with type II diabetes. GSE118139 and GSE20966, which contain two control islet tissue samples and two diabetic samples, ten control islet samples, and ten diabetic samples, respectively, were used for LASSO analysis to validate and select the biomarker genes. GSE15932, GSE15653, GSE166467, GSE55650, and GSE20966 were used to validate the expression levels of the biomarker genes. GSE15932 contains peripheral blood samples of eight patients with T2DM and eight healthy controls. The GSE15653 dataset consisted of 13 obese (9 with T2DM) and five control subjects from human surgical liver biopsies. GSE166467 comprised the mRNA expression data for both proliferating myoblasts and differentiated myotubes from 13 T2DM patients and 13 controls. GSE55650 consisted of the muscle biopsies from 6 T2DM patients and six controls.

The human genes related to ERS were collected by combining the genes from the GeneCards database (GeneCards; https://www.genecards.org/) (970 protein genes) and a list from the literature (26 protein genes) (Shen et al., 2022). After the intersection, 973 ERS-related protein genes were obtained and used for the subsequent analysis.

Screening of differentially expressed ERS-related genes

Based on the expression data provided by the dataset GSE25724, DEGs between normal and diabetic islet samples were analyzed. Firstly, batch effects were excluded by principal component analysis (PCA). The "limma" package of R software was used to identify the DEGs in diabetes. With the absolute value | log2 FC |>2, the genes with significant expression differences were selected after calibration for $p < 0.05$. The heatmap and volcano plots of the DEGs were generated using the "heatmap" and "ggplot2″ packages of R software. The intersection of the ERS-related genes and the DEGs from GSE25724 is considered the ERS-related DEGs in diabetes.

Protein-protein interaction networks analysis

The Protein-Protein Interaction (PPI) network analysis was conducted using the Search Tool (STRING) database. A

TABLE 1 The qPCR primers of the ERS-related biomarker genes.

Genes	Forward (5′-3′)	Reverse (5′-3′)
CLGN-mouse	AGTGGTAATGTCTGAGCAA	AGGAGTTTGTAGTGATGTTTG
CLGN-human	TATATGACCCACATTTACCTAGT	AACCCATTATCCTTGTATTCAAA
IMPA1-mouse	AGAATTGGAATCGGACAGA	CAAGTTTAGATCAGTGGATAGC
IMPA1-human	TTGCCTGTAATCTTTCCAAC	TCTAAGAAGTCCTGTTACTCAA
ILF2-mouse	TGGCTTCTATAACCTCAGTAG	GCTTTCACCCCACATTTAG
ILF2-human	GTAGGGCTCTTGGTCTTT	AGGTTCCAGGAGTTTGTC

TABLE 2 The specific qPCR primers of the miRNAs predicted to target the ERS-related biomarker genes.

Genes	Forward (5′-3′)
miR-197–5p	CGCGGGTAGAGAGGGCAGT
miR-6133	CGCGTGAGGGAGGAGGT
miR-7851–3p	CGGAGTGGGGCTTCGACC
miR-1296–3p	CGGAGTGGGGCTTCGACC
miR-320c	CGCGAAAAGCTGGGTTGA
miR-4776–5p	CGGTGGACCAGGATGGCA
miR-4462	GTGACACGGAGGGTGGCT
MiR-16–5p (reference)	CGCGTAGCAGCACGTAAATA

composite score of >0.4 was considered statistically significant. The analysis results were visualized using Cytoscape software (version 3.8.1). The Spearman correlation of candidate genes was analyzed using R software's "coreplot" package.

Functional enrichment analysis

GO and KEGG pathway enrichment analysis was conducted using the "GO plot" package of R software to predict the potential molecular functions of the biomarker genes. The GO settings include molecular function (MF), biological process (BP), and cellular composition (CC). GSEA analysis was conducted using the Xiantao Academic Analysis Platform (https://www.xiantaozi.com) to find the pathways to enrich the DEGs.

Validation of the ERS-related DEGs

Receiver operating characteristic (ROC) curves were performed using the "pROC" package to evaluate the reliability of the ERS-related DEGs. Subsequently, the genes selected by ROC were further screened by Least Absolute Shrinkage and Selection Operator (LASSO) Cox analysis using the datasets GSE25724, GSE118139, and GSE20966 to give out the critical biomarker genes. The biomarker genes for predicting were then analyzed in the validation sets, including GSE118139, GSE15932, GSE15653, GSE166467, GSE20966, and GSE55650.

Prediction of transcription factors and microRNAs

The transcription factors regulating the biomarker genes were predicted by the "hTFtarget" database (http://bioinfo.life.hust.edu.cn/hTFtarget#!/). The gene transcription factor network diagram was generated by "Cytoscape." MiRNAs targeting the biomarker genes were predicted in miRWalk (http://mirwalk.umm.uni-heidelberg.de/) and miRTarBase (https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php), by setting the conditions for "number_of_pairings>15, binding_region_length>20 and longest_consecutive_pairings>10".

Drug analysis of the biomarker genes

Potential drugs with CAS numbers interacting with the biomarker genes were predicted by the CTD online tool (https://ctdbase.org/). The drugs were screened using the website's scores. The top 20 drugs for each biomarker gene were collected.

Construction of diabetic mouse model

Forty 6-week-old wild-type B6 male mice (weighing about 18–20 g) were randomly divided into the control group and the model group. The model group was fed a high-fat diet (45% fat content) for 6 weeks and treated by intraperitoneal injection of 30 mg/kg streptozotocin once. The type 2 diabetic mice were constructed when the Fasting blood glucose was higher than 7.8 mmol/L, and the random blood glucose was higher than 16.7 mmol/L. All animal protocols were approved by the Committee of Nantong University (SYXK (SU) 2017-0046) and the Administration Committee of Experimental Animals, Jiangsu Province, China.

Human samples collection

For qPCR detection of the biomarker genes, eight blood samples were collected from the Department of Endocrinology, Affiliated Hospital of Nantong University, containing 4 T2DM samples and four non-related control samples. For qPCR detection of the miRNAs and Elisa detection of the biomarker genes, another 12 blood samples (6 T2DM and six non-related control samples)

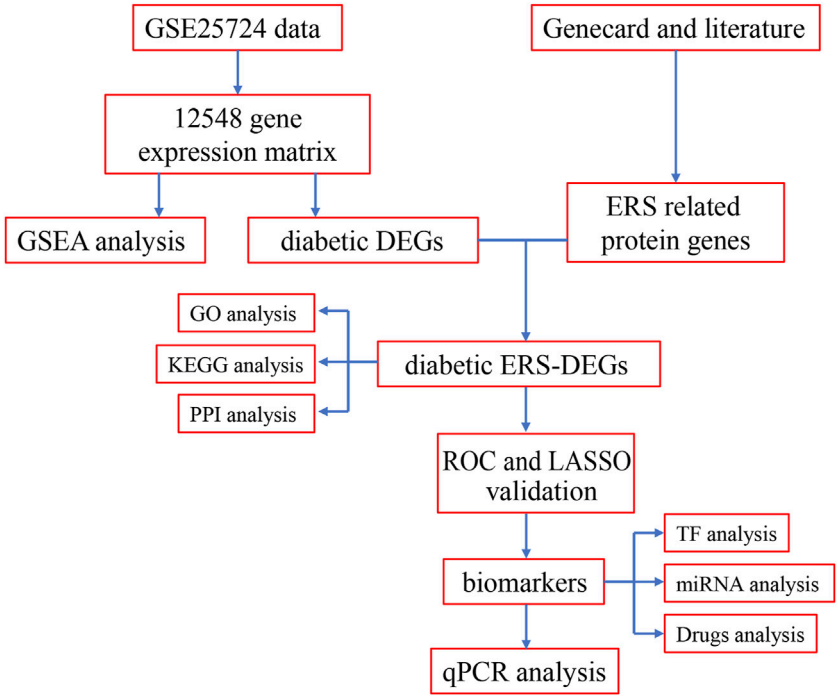


FIGURE 1
Flow chart of methodologies applied in the study.

were collected from the same hospital department. The case information of all the samples is listed in (Supplementary Table. 1, 2). After overnight fasting, 5 mL venous blood samples were collected and centrifuged at 3,000 *g* for 10 min, followed by serum separation and storage at −80°C. The study and experiments were approved by the Administration Committee of Nantong University Affiliated Hospital (2018-K016), Jiangsu Province, China. All volunteers involved in this study provided written consent for publication.

RNA extraction and qPCR analysis

For detecting the expression of the critical ERS-related DEGs, the total RNA of diabetic and normal islet tissues of mice and human serum plasma samples was isolated with Trizol (Invitrogen) and stored at −80°C. The cDNA was synthesized using the Transcriptor First Strand cDNA Synthesis Kit (Roche) according to the manufacturer's instructions and stored at −20°C. The qPCR reaction was performed in triplicates using the FastStart Universal SYBR Green Master Mix (Roche Applied Science) on a real-time PCR detection system (StepOneTM Real-Time PCR Systems). The primers of the critical ERS-related DEGs and reference gene (18S) for qPCR were designed by Beacon Designer eight and the sequences are listed below (Table 1):

For miRNA expression analysis, all small RNAs were extracted from serum plasma by using mirVana™ miRNA isolation kit (ThermoFisher). The reverse transcription and qPCR reaction were performed using miRNA first Strand cDNA Synthesis Kit (by stem-loop) and miRNA Universal SYBR qPCR Master Mix (Vazyme), respectively. While the reverse universal primer is provided by the Universal SYBR qPCR Master Mix, the specific forward primers for miRNAs' qPCR detection were designed according to the manufacturer's instruction and listed below (Table 2):

Statistical analysis

R software (version 3.6.2) was utilized for statistical analysis. Student's t-test was used to evaluate the significance of variance. A *p*-value of <0.05 was considered statistically significant. The overview of the workflow is shown in Figure 1.

Results

Identification of ERS-related DEGs in diabetes

From the dataset, GSE25724 yielded 12,548 genes found to be diabetes-related protein genes. In order to investigate the pathways associated with diabetes, Gene Set Enrichment Analysis (GSEA) was conducted on these 12,548 genes. The findings revealed a significant enrichment of the "Unfolded Protein Response" (UPR) pathway, which demonstrated a close correlation with T2DM, thereby suggesting the crucial involvement of ERS in the development of T2DM (Figure 2). Principal Component Analysis (PCA) was employed to validate the reproducibility and reliability of the data obtained from GSE25724 (Figure 3A). Subsequently, the

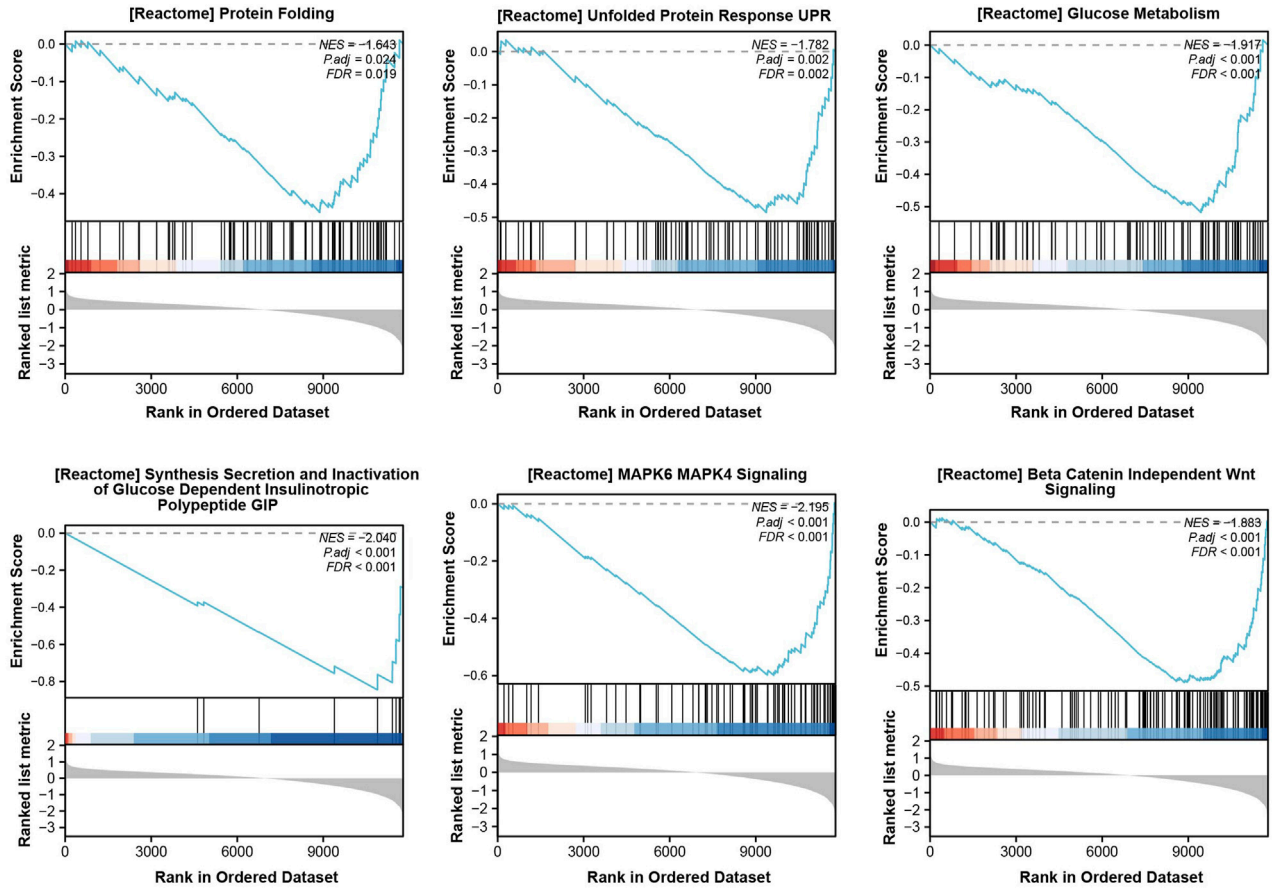


FIGURE 2
GSEA enrichment analysis for T2DM-related pathways.

gene expression profiles of normal and diabetic islet tissues in GSE25724 were analyzed for selecting the DEGs related to ERS, with the screening criteria of |logFC| >2 and $p < 0.05$. As a result, 49 protein genes were identified as T2DM-associated DEGs, which were visualized in a volcano plot (Figure 3B). After intersection with the 973 ERS related genes, 8 ERS-related DEGs (RTN1, CLGN, PCSK1, IAPP, ILF2, IMPA1, CCDC47, and PTGES3) were screened out as T2DM related ERS-DEGs, which were indicated to be downregulated in diabetic samples (Figures 3C–E).

Functional enrichment and PPI network analysis

In order to gain further insight into the functions and pathways associated with the eight genes selected above, GO and KEGG enrichment analyses were conducted. The findings revealed that these genes are primarily involved in protein folding, the ubiquitin-dependent ERAD pathway, and cellular carbohydrate biosynthetic metabolism. The protein products of the genes are primarily localized on the rough endoplasmic reticulum, the intrinsic component of the endoplasmic reticulum membrane, and the integral component of the endoplasmic reticulum membrane. These proteins are involved in protein folding chaperones, unfolded protein binding, and RNA-directed DNA polymerase activity (Figures 4A, B).

PPI network analysis was conducted using the online tools “GeneMANIA” and “Cytoscape” software to investigate the interaction between the eight candidate genes and other protein genes. The results revealed strong associations between the genes. For example, CAPRIN and PCSK1 were co-expressed, and RELA was co-expressed with CCDC47 and PTGES3 (Figure 4C).

Screening and validation of the biomarker genes for T2DM

To evaluate the predictive and diagnostic value of the ERS-related genes in T2DM, the Receiver Operating Characteristic (ROC) curves were employed to assess the diagnostic efficacy of the eight genes selected above. The findings revealed that six genes (CCDC47, CLGN, ILF 2, IMPA 1, PTGES3, RTN 1) exhibited an area under the curves (AUC) exceeding 0.9, indicating a substantial diagnostic value (Figure 5A).

To further observe and evaluate the correlation of these six genes with T2DM, LASSO regression analysis was executed based on the gene expression databases GSE25724, GSE118139, and GSE20966, wherein CLGN, ILF2 and IMPA1 were identified as

10.3389/fgene.2024.1445033

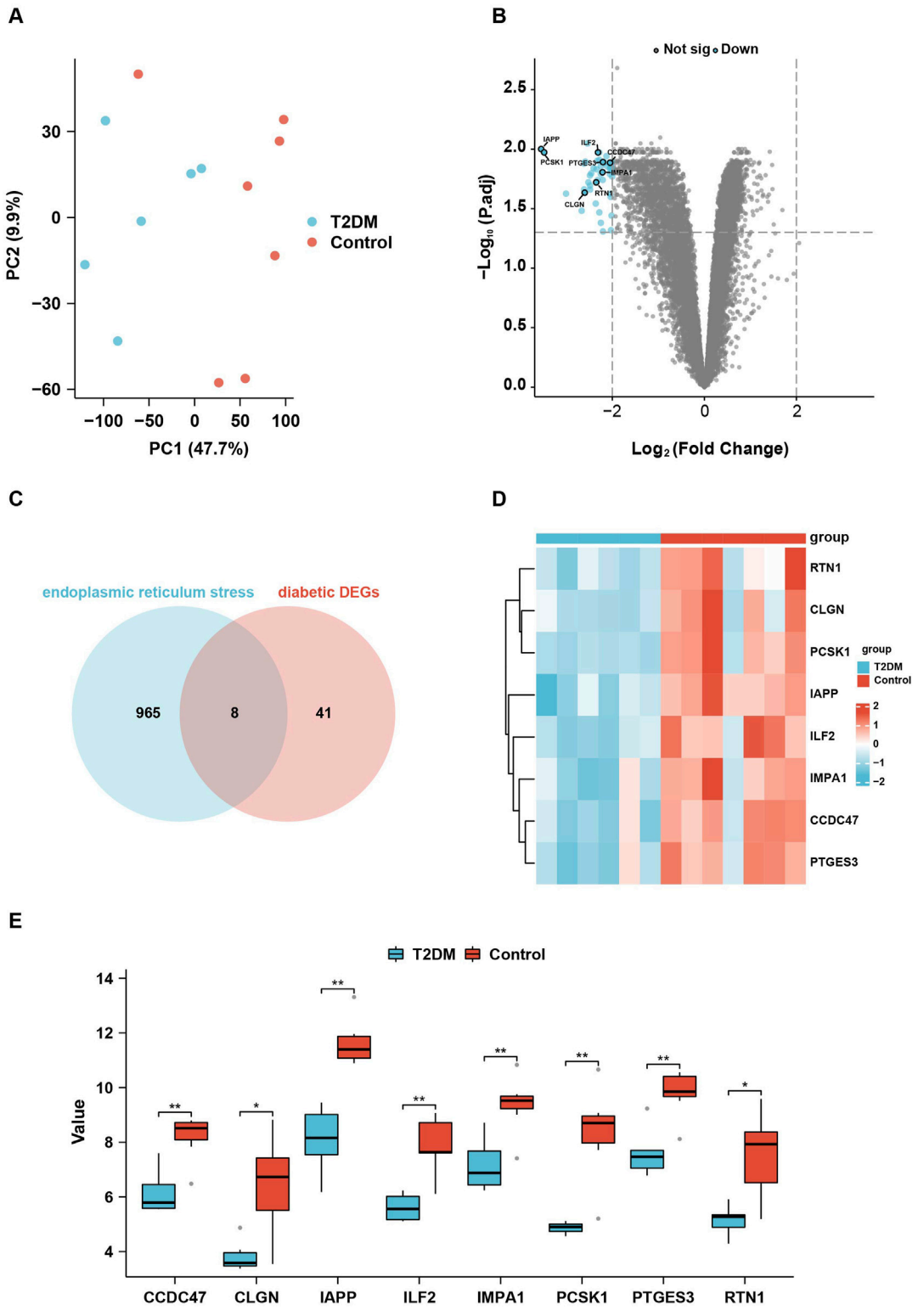


FIGURE 3
Identification of the ERS-related DEGs in T2DM. **(A)**, PCA analysis of the gene expression data from GSE25724. **(B)**, the volcano plot showing the DEGs in T2DM. **(C)**, the Venn diagram showing the intersection between the DEGs in T2DM and the ERS-related genes. **(D, E)**, the heatmap and box plots showing the eight relative expression levels of ERS-related DEGs in T2DM.

10.3389/fgene.2024.1445033

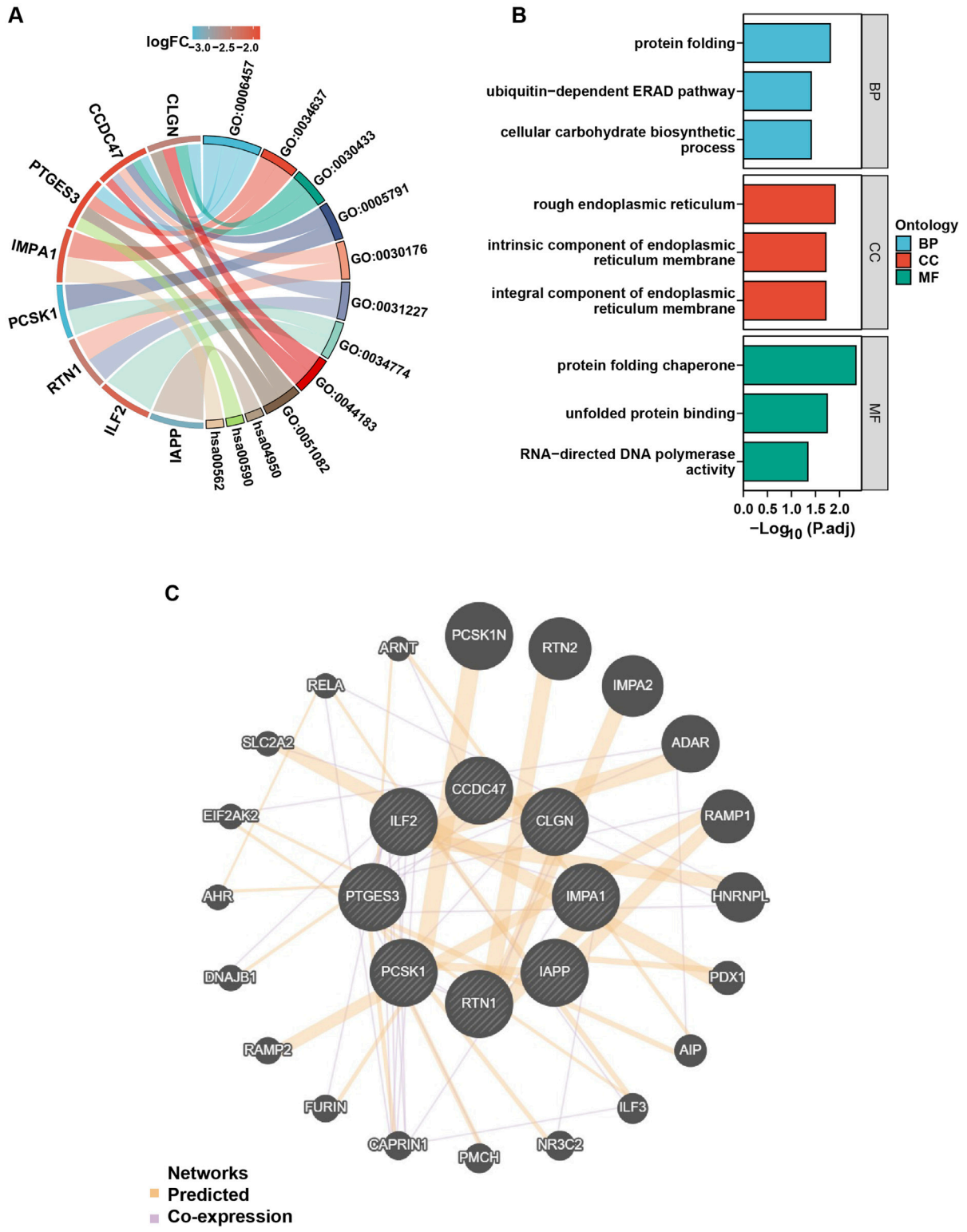


FIGURE 4
GO, KEGG, and PPI analysis of the ERS-related DEGs in T2DM. **(A)**, the chord diagram of GO analysis on the ERS-related DEGs. GO: 0006457 = protein folding, GO: 0034637 = cellular carbohydrate biosynthetic process, GO: 0030433 = ubiquitin-dependent ERAD pathway, GO: 0005791 = rough endoplasmic reticulum, GO: 0030176 = integral component of endoplasmic reticulum membrane, GO: 0031227 = intrinsic component of endoplasmic reticulum membrane, GO: 0034774 = secretory granule lumen, GO: 0044183 = protein folding chaperone, GO: 0051082: unfolded protein binding, hsa00562 = inositol phosphate metabolism, hsa00590 = arachidonic acid metabolism, hsa04950 = maturity onset diabetes of the young, p. adjust value < 0.05. **(B)**, GO enrichment map with ERS-related DEGs, p. adjust value < 0.05. **(C)**, PPI network analysis of the ERS-related DEGs.

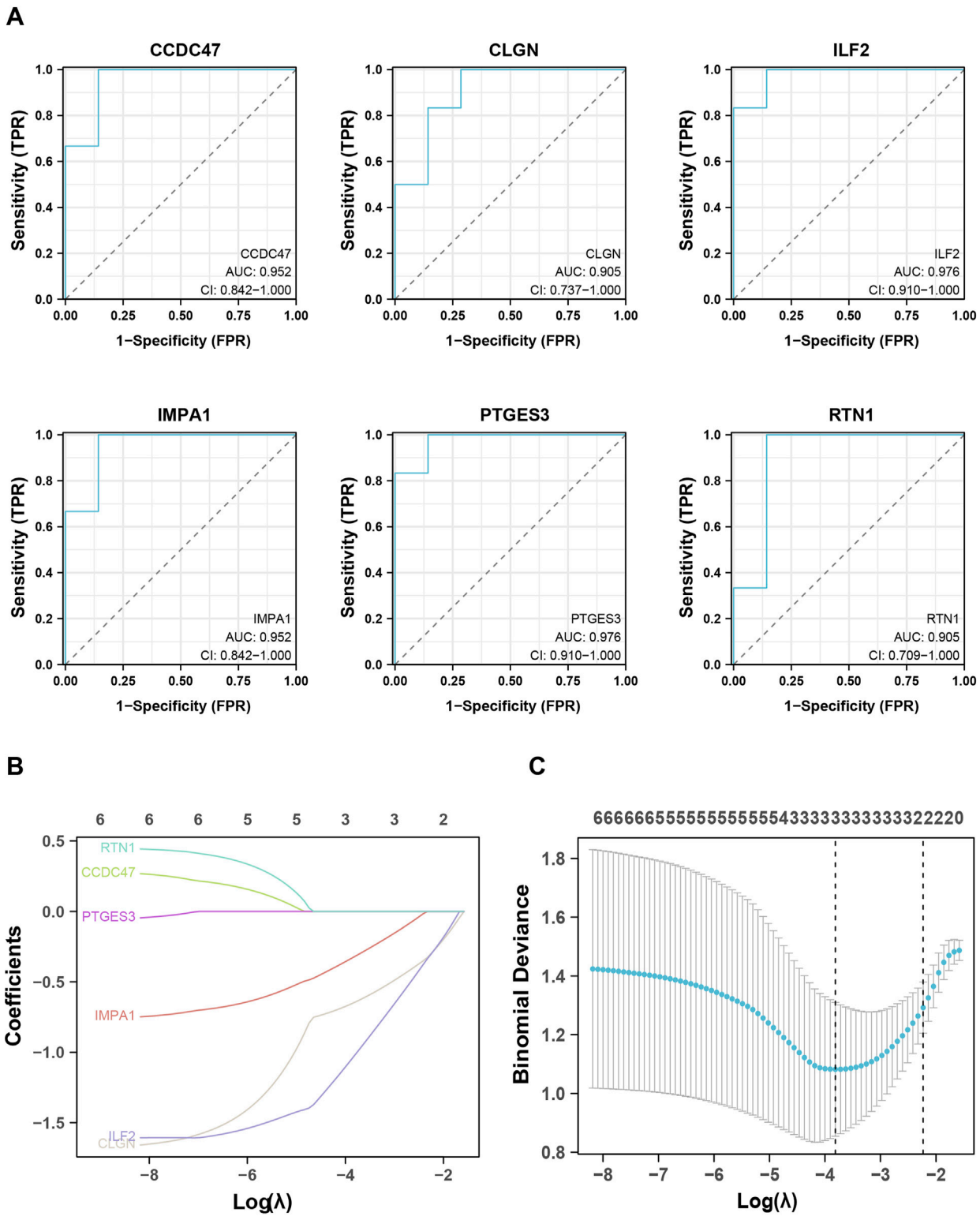


FIGURE 5
Identification and validation of the biomarker genes for diagnosis. **(A)**, ROC analysis shows the six critical genes with AUC values higher than 0.9. **(B, C)**, LASSO analysis screened out the three biomarker genes.

critical biomarker genes for T2DM and therefore used for subsequent analysis (Figures 5B, C). In addition, the expression levels of these three genes were examined in other validation datasets. As a result, IMPA1 was found to be significantly downregulated in the diabetes group in databases GSE15932, GSE15653, and GSE166467. CLGN and ILF2 were

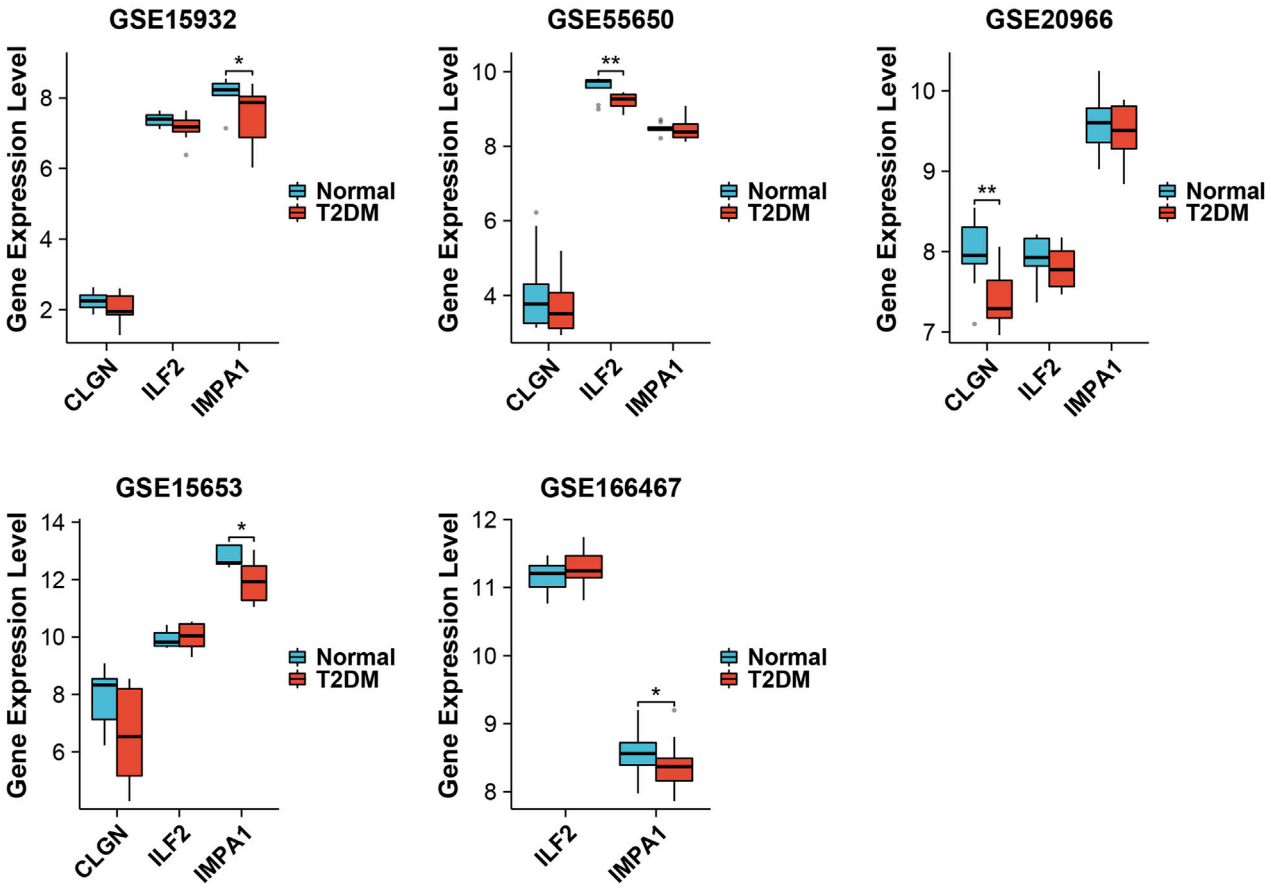


FIGURE 6
Validation of the biomarker genes in other datasets. "*" represents $p < 0.05$, "**" represents $p < 0.01$.

significantly downregulated in the diabetic group in GSE20966 and GSE55650, respectively (Figure 6).

Transcription factor analysis

To further elucidate the upstream regulators of the biomarker genes, a transcription factor network analysis was performed to investigate the transcription factors regulating the biomarker genes with significant diagnostic potential. The findings revealed that multiple transcription factors potentially regulate most of these genes. For instance, CLGN was predicted to be targeted by FOXA1, FOXA2, CEBPA, CEBPB, and others. E2F1, CDK9, MAZ, KLF1, and others can target IMPA1. ILF2 might be targeted by MAX, KLF5, KLF4, JUND, and so on (Figure 7). The same transcription factors, such as HDAC1, HDAC2, CEBPA, and CEBPB could also regulate different biomarker genes.

miRNA analysis

To understand the potential roles of miRNAs involved in regulating these three biomarker genes, miRWalk and miRTarBase were utilized for microRNA prediction. Numerous miRNAs were predicted to be potential upstream regulators of the biomarker genes. Among them, hsa-miR-197–5p, hsa-miR-6133, hsa-miR-7851–3p, hsa-miR-1296–3p, hsa-miR-320c, hsa-miR-4776–5p and hsa-miR-4462 were predicted to simultaneously target two of the biomarker genes (Figure 8A). We then performed qPCR analysis to detect the relative expression levels of these miRNAs in serum samples, which were revealed to be all significantly changed in diabetic patients. Among them, the expression of miR-197–5p, miR-320c, miR-1296–3P and miR-6133 was downregulated, while that of miR-4462, miR-4476–5P and miR-7851–3P was upregulated in diabetic samples (Figure 8B).

Drug identification and selection

To identify personalized medicines for diabetes, the CTD website was utilized to predict small molecular drugs. Based on the website's scores, the top 20 drugs for each gene were selected and presented herein. Interestingly, D019813 (1, 2-Dimethylhydrazine) was predicted to target all three biomarker genes. At the same time, several other drugs, such as D002994 (Clofibrate), D001564 (Benzo(a)pyrene), C016403 (2, 4-dinitrotoluene), D016604 (Aflatoxin B1), D000082 (acetaminophen), D003471(Cuprizone), C006780 (bisphenol A), D016572 (Cyclosporine), D019327 (Copper Sulfate) and D003300 (Copper), exhibited the potential to target two

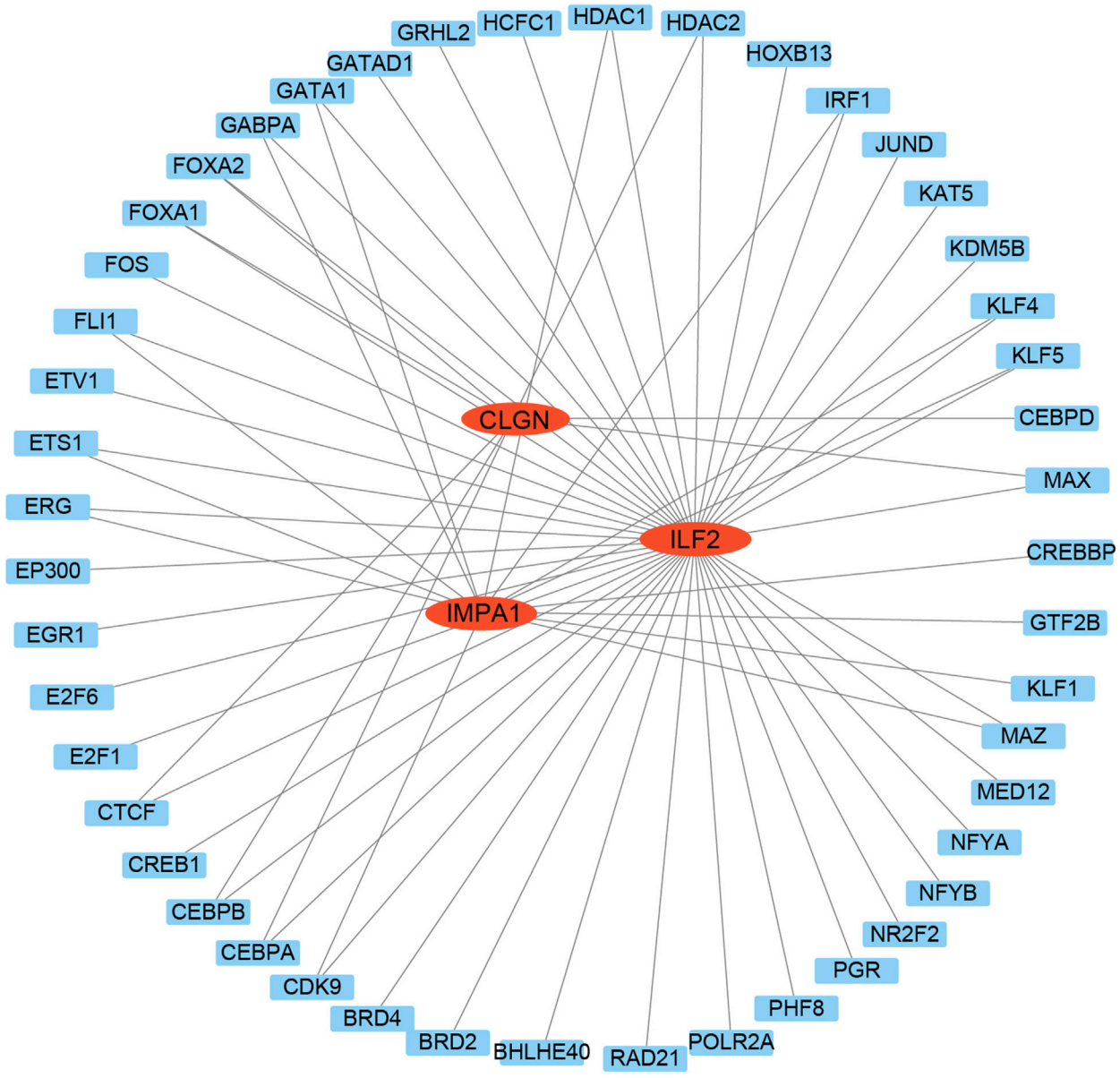


FIGURE 7
Transcription factors analysis of the biomarker genes. The TFs were used as nodes in an interconnected regulatory network. Red ellipses represent the biomarker genes, and blue rectangles represent the TFs.

biomarker genes simultaneously. The result suggests that these drugs may serve as effective multi-target medications for T2DM (Figure 9).

Experimental validation of the expression of the ERS-related biomarker genes

In order to provide additional evidence for the differential expression of the ERS-related biomarker genes in diabetes, we conducted qPCR analysis on the islet tissues of healthy and diabetic mice. The results demonstrated a significant decrease in the transcriptional expression of all three genes in the diabetic samples compared to the control samples (Figure 10A). In addition, to determine whether these biomarker genes can be used for clinical detection, we also performed the qPCR and Elisa analysis on human serum samples of T2DM patients and non-related individuals. The results revealed the same changing trend of gene expression in diabetes, although the alteration of the serum protein level of ILF2 was not significant (Figures 10A, B). The experimental findings aligned with the results obtained from the bioinformatical analysis.

Discussion

T2DM is a metabolic syndrome characterized by insulin resistance, relative insulin deficiency, and impaired glucose

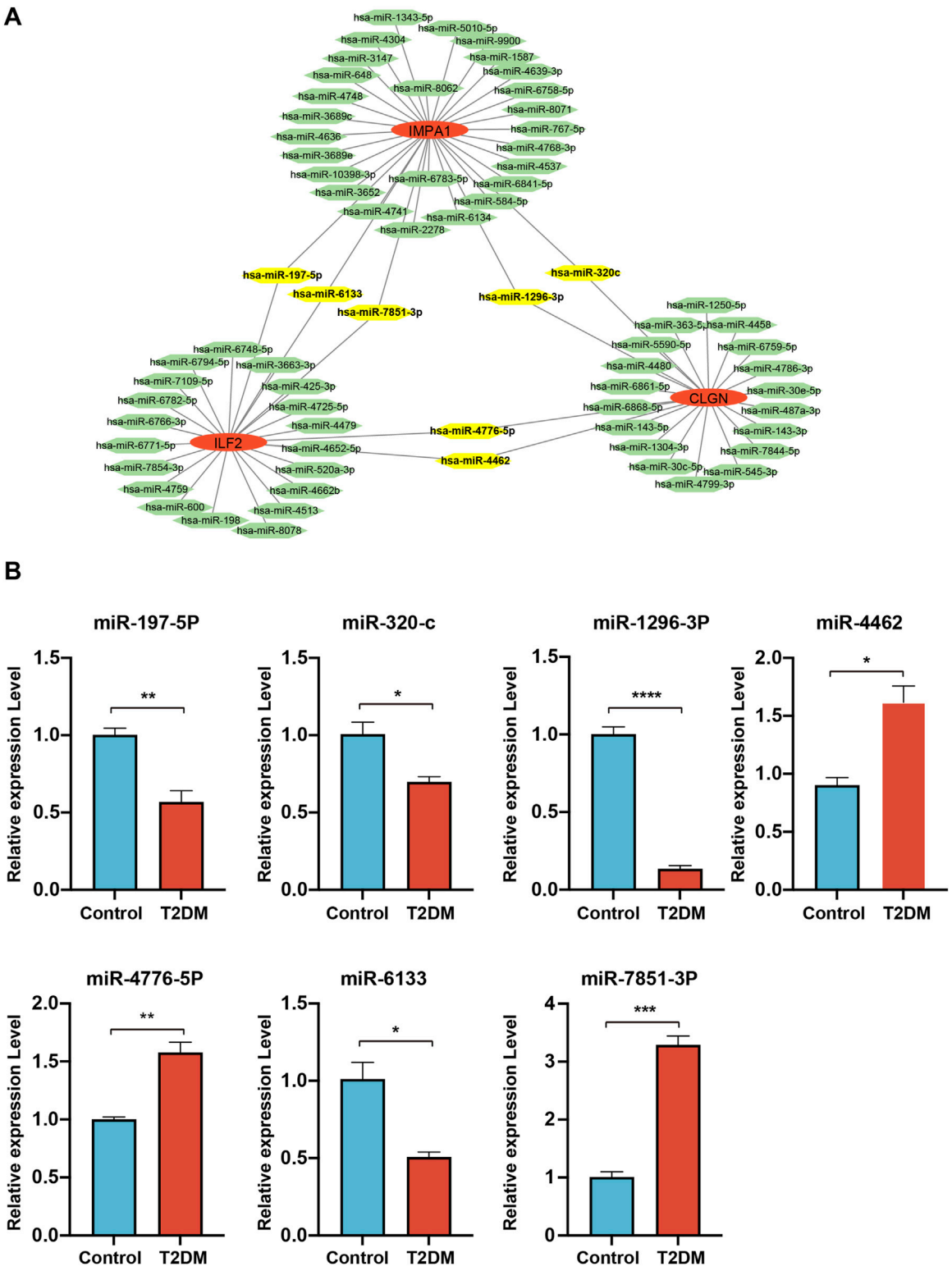


FIGURE 8
miRNA analysis of the biomarker genes. **(A)**, the miRNAs are predicted to target the biomarker genes. MiRNAs are represented by green hexagons. **(B)**, qPCR detection of the expression levels of the miRNAs simultaneously targeting two of the biomarker genes.

tolerance (Artasensi et al., 2020). The occurrence of ERS is implicated in regulating diverse physiological processes (Jayasooriya et al., 2018), such as inflammation (Zhang et al., 2020), tumor development (Cubillos-Ruiz et al., 2017), anti-viral response (Ong et al., 2018), and lipid metabolism (Zhao et al., 2020). ERS induced by elevated levels of glucose, fat, and cytokine

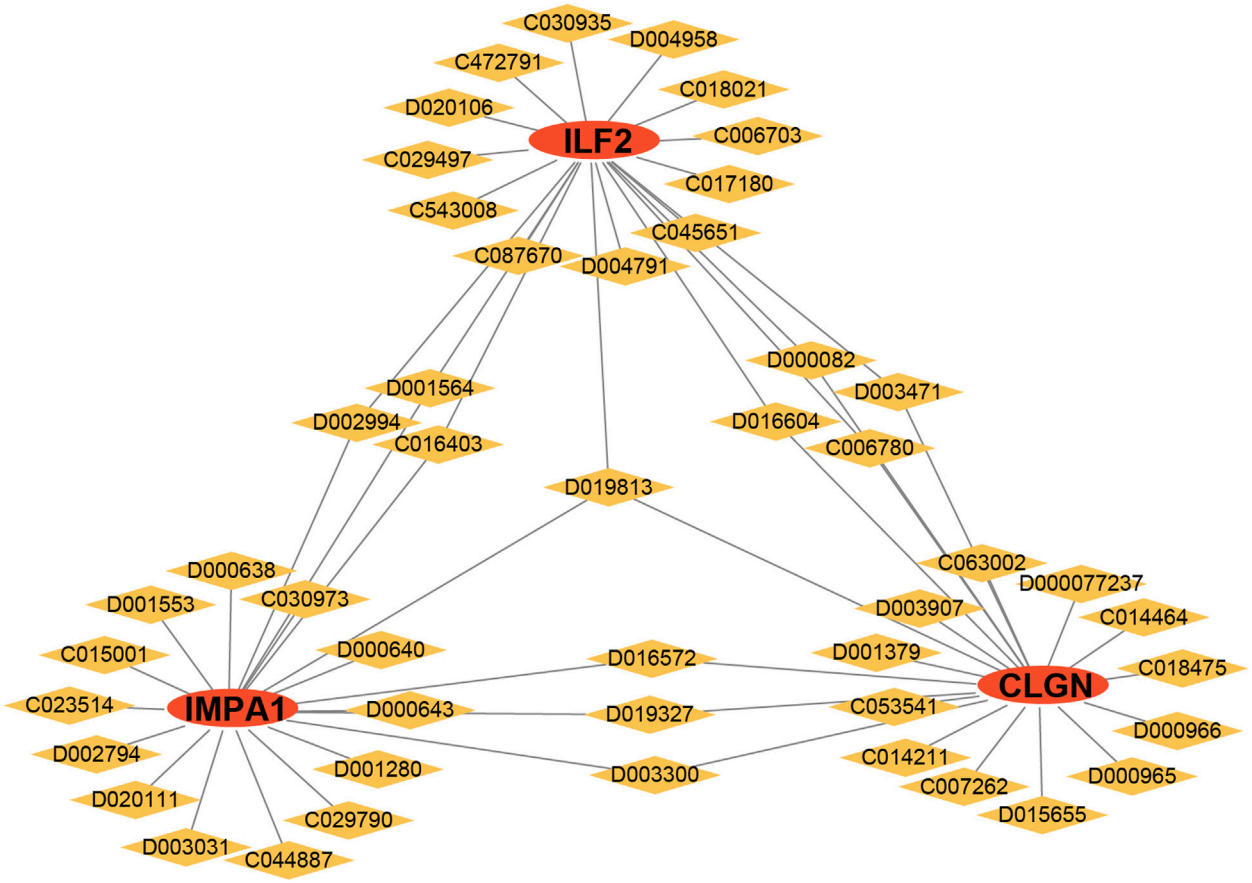


FIGURE 9
Drug analysis of the biomarker genes. Red ellipses represent the biomarker genes, and yellow diamonds represent the drug IDs.

stimulation has been observed to contribute to insulin resistance and the deterioration of islet β-cell function in T2DM (Vallée et al., 2020; Yilmaz, 2017; Deng et al., 2022; Chen et al., 2018). Although numerous studies have demonstrated the association between ERS and T2DM, few studies have been done to explore the utility of the ERS-related biomarker genes for T2DM diagnosis.

In this study, T2DM-associated genes and ERS-associated genes were extracted from the GEO and GeneCards databases, respectively. GSEA analysis of the T2DM-associated genes revealed the enrichment of the diabetic proteins in the pathway of UPR. UPR is crucial in coordinating protein synthesis, folding, and degradation to maintain protein stability, which is vital for cell survival and activity. Prior research has indicated that sustained activation of the UPR plays a role in mitigating ERS-induced disruptions in glucose regulation, chronic inflammation, and the advancement of T2DM (Herrema et al., 2022; Keestra-Gounder et al., 2016; Ma et al., 2014). After the intersection of the two groups of genes, we identified eight ERS-DEGs associated with T2DM, which were all downregulated in the patient samples. These proteins are functionally involved in various biological processes, including cell carbohydrate synthesis, protein folding, inositol phosphatase metabolism, etc. Dysfunctions in the metabolic response to inositol phosphatase have been linked to insulin resistance and the development of long-term microvascular complications in individuals with diabetes (Croze and Soulage, 2013). The concurrent administration of phosphoinositol and inositol has been shown to safeguard hepatocyte integrity and enhance its antioxidant capacity in individuals with T2DM (Foster et al., 2017).

Utilizing ROC and LASSO analysis on the eight ERS-DEGs, three critical genes were subsequently screened out as valuable predictors of the disease, including CLGN, ILF2, and IMPA1. CLGN, known as ER chaperone calmegin, is a highly expressed ER-associated gene in aldosterone-producing adenomas, but its role in diabetes remains unclear (Itcho et al., 2020). ILF2, known as nuclear factor 45 (NF45), is crucial in regulating RNA stability and inflammatory response (Yin et al., 2022). The formation of a complex between ILF2 and S6K protein influences insulin levels and consequently contributes to the progression of metabolic diseases (Das et al., 2018; Pavan et al., 2016). IMPA1 is an enzyme responsible for inositol synthesis; deficiency in IMPA1 leads to a decline in inositol and mitochondrial fission, ultimately leading to the development of diabetes and mitochondrial diseases (Hsu et al., 2021). Recently, a case-control study demonstrated that in gestational diabetes, higher maternal glycemia is associated with decreased protein and mRNA expression levels of IMPA1 (Pillai et al., 2021). In addition, these three genes were further validated in other separated datasets comprising the data from different organs or tissues, including liver, muscle, and peripheral blood of T2DM patients, exhibiting a consistent alteration in expression, thereby suggesting their potential diagnostic value for clinical use.

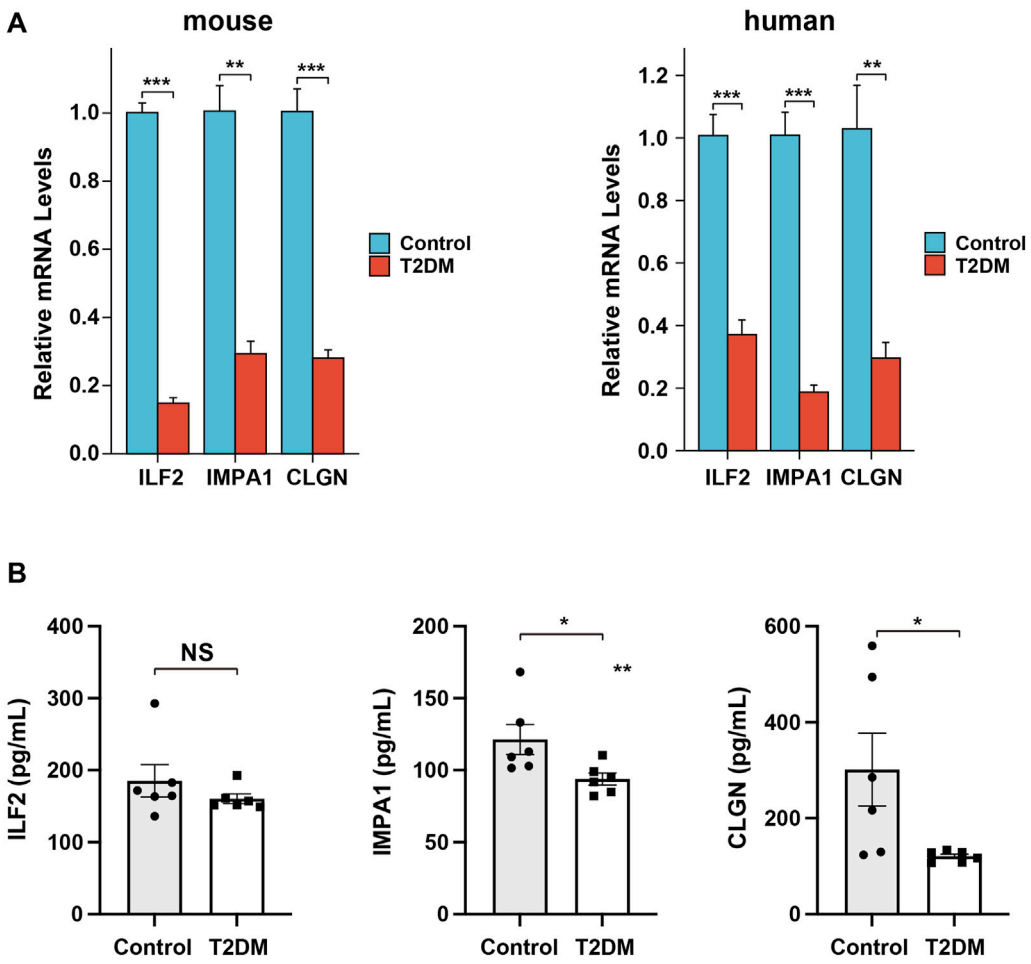


FIGURE 10
Expression analysis of the biomarker genes in mice and patient serum samples. **(A)**, qPCR analysis of the relative expression levels of the biomarker genes in diabetic mice and patient serum samples. **(B)**, Elisa analysis of the serum protein levels of the biomarker genes in diabetic and non-related patient samples.

Our transcription factor analysis identified several crucial transcription factors closely associated with diabetes. Among them, CEBPA and CEBPB are adipogenic transcription factors that not only impact adipogenesis but also influence the occurrence of diabetes (Oh et al., 2013; Wicks et al., 2015). HDAC1 and HDAC2, known as histone deacetylases, are widely recognized for their essential role in regulating DNA structure and gene activity, and they also contribute to the development of diabetic complications (Hou et al., 2018; Zheng et al., 2022; Draney et al., 2018; Li et al., 2022). Additionally, FOXA1 has been demonstrated to play significant roles in maintaining glucose homeostasis and promoting α-cell differentiation (Heddad Masson et al., 2014). In the context of miRNA analysis, it is noteworthy that a few miRNAs exhibit potential targeting capabilities towards two biomarker genes. Our experimental analysis further revealed significant expression alteration of these miRNAs in patient serum samples, suggesting their potential value in diagnosis for T2DM. Among these miRNAs, miR-6133 and miR-320c have been reported to be downregulated in urinary exosomes of T2DM patients in the previous study (Delić et al., 2016), consistent with our results.

Recently, several drugs or biomolecules have been shown to have potential in treating diabetes and its complications via modulating ERS, such as Ghrelin, Rosuvastatin, Selenium Nanodots (SENDs) and Astragalus polysaccharide (Li et al., 2024; Zhao et al., 2023; Huang et al., 2023; Chen et al., 2023). Our study also predicted the drugs having the potential to target the three biomarker genes. Among the top 20 drugs targeting over two genes, cyclosporine has been used for treating T2DM and related complications for a long time (Mahon et al., 1993; Wang et al., 2020). Copper and copper sulfate have been shown to potentially ameliorate diabetes and its complications in animal models (Sitasawad et al., 2001; Sakurai, 2012). However, bisphenol A is a risk factor associated with the occurrence and development of T2DM (Provvisiero et al., 2016). In addition, 1,2-Dimethylhydrazine, Benzo(a)pyrene, and Aflatoxin B1 have been reported to be inducers for cancers including diabetic colon cancer, lung cancer, and hepatocellular carcinoma, suggesting their potentially detrimental effects in application (Terai et al., 2006;

Kasala et al., 2015; Cao et al., 2022). The results of the drug analysis indicate that the targeted drugs, which were screened based on the critical diabetic ERS-related DEGs, exhibit significant therapeutic relevance for diabetes treatment. However, it is important to note that certain drugs also present substantial risks, particularly in terms of side effects that may induce carcinogenesis. Therefore, comprehensive evaluation and rigorous testing are imperative during the drug development process.

To further confirm the expression of the biomarker genes, we conducted qPCR experiments using constructed diabetic mice and human serum samples. The results demonstrated a significant decrease in their expression in both diabetic islets and human serum, aligning with the computational analysis. Moreover, the results suggested that these biomarker genes have immense potential in fundamental research and clinical application in T2DM pathogenesis and diagnosis. Comparing with the recent studies which also conducted the bioinformatical analysis of ERS-related biomarkers in T2DM and diabetes nephropathy (Su et al., 2023; Liang et al., 2023), our research has provided more evidence with animal models and clinical samples for the biomarker genes.

There are some limitations of this study. For example, the limited sample size of the T2DM database and the lack of clinical validation. In addition, the mechanism of the critical biomarker genes in regulating T2DM remains unclear. We will continue to delve deeper into the subsequent research about our findings.

Conclusion

Using bioinformatical methods, the identification and screening of ERS-related genes in diabetes were conducted. Subsequently, three critical biomarker genes were identified and validated through bioinformatical analysis and experimental detection, establishing their utility as biomarkers for T2DM diagnosis. These findings contribute to a deeper comprehension of the interplay between ERS and the onset and progression of diabetes while also offering potential targets for future diagnostic and therapeutic interventions.

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 studies involving humans were approved by the Administration Committee of Nantong University Affiliated Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. The animal study was approved by Committee of Nantong University. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

LY: Validation, Writing–original draft. JX: Funding acquisition, Validation, Writing–original draft. XZ: Funding acquisition, Investigation, Writing–original draft. ZT: Writing–review and editing, Funding acquisition, Resource. YC: Funding acquisition, Methodology, Writing–original draft. XL: Conceptualization, Writing–review and editing. XD: Conceptualization, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was grants from Jiangsu Provincial Research Hospital (YJXYY202204-YSB38 received by Zhuqi Tang), supported by the Training Project for College Students of Nantong University (202310304126Y received by Jie Xu), the Project of Wuxi Health Committee (Q202229 received by Xu Zhang), Projects of Nantong Health Committee and Jiangsu Key Laboratory of New Drug Research and Clinical Pharmacy (QN2023005 and KFKT-2316 received by Yuqing Chen), and the Large Instruments Open Foundation of Nantong University (KFJN2449 received by Xiaoyu Liu).

Acknowledgments

During the preparation of this work the authors used the web tool on "HOME for Researchers" (https://www.home-for-researchers.com/static/index.html#/) to polish the language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1445033/full#supplementary-material

References

Artasensi, A., Pedretti, A., Vistoli, G., and Fumagalli, L. (2020). Type 2 diabetes mellitus: a review of multi-target drugs. *Mol. Basel, Switz.* 25 (8), 1987. doi:10.3390/molecules25081987

Cao, W., Yu, P., Yang, K., and Cao, D. (2022). Aflatoxin B1: metabolism, toxicology, and its involvement in oxidative stress and cancer development. *Toxicol. Mech. Methods* 32 (6), 395–419. doi:10.1080/15376516.2021.2021339

Chen, G., Wu, Y., Wang, T., Liang, J., Lin, W., Li, L., et al. (2012). Association between serum endogenous secretory receptor for advanced glycation end products and risk of type 2 diabetes mellitus with combined depression in the Chinese population. *Diabetes Technol. and Ther.* 14 (10), 936–942. doi:10.1089/dia.2012.0072

Chen, X., Shen, W. B., Yang, P., Dong, D., Sun, W., and Yang, P. (2018). High glucose inhibits neural stem cell differentiation through oxidative stress and endoplasmic reticulum stress. *Stem cells Dev.* 27 (11), 745–755. doi:10.1089/scd.2017.0203

Chen, X., Shi, C., He, M., Xiong, S., and Xia, X. (2023). Endoplasmic reticulum stress: molecular mechanism and therapeutic targets. *Signal Transduct. Target Ther.* 8 (1), 352. doi:10.1038/s41392-023-01570-w

Cnop, M., Toivonen, S., Igoillo-Esteve, M., and Salpea, P. (2017). Endoplasmic reticulum stress and eIF2α phosphorylation: the Achilles heel of pancreatic β cells. *Mol. Metab.* 6 (9), 1024–1039. doi:10.1016/j.molmet.2017.06.001

Croze, M. L., and Soulage, C. O. (2013). Potential role and therapeutic interests of myo-inositol in metabolic diseases. *Biochimie* 95 (10), 1811–1827. doi:10.1016/j.biochi.2013.05.011

Cubillos-Ruiz, J. R., Bettigole, S. E., and Glimcher, L. H. (2017). Tumorigenic and immunosuppressive effects of endoplasmic reticulum stress in cancer. *Cell* 168 (4), 692–706. doi:10.1016/j.cell.2016.12.004

Darenskaya, M. A., Kolesnikova, L. I., and Kolesnikov, S. I. (2021). Oxidative stress: pathogenetic role in diabetes mellitus and its complications and therapeutic approaches to correction. *Bull. Exp. Biol. Med.* 171 (2), 179–189. doi:10.1007/s10517-021-05191-7

Das, S., Reddy, M. A., Senapati, P., Stapleton, K., Lanting, L., Wang, M., et al. (2018). Diabetes mellitus-induced long noncoding RNA Dnm3os regulates macrophage functions and inflammation via nuclear mechanisms. *Arteriosclerosis, thrombosis, Vasc. Biol.* 38 (8), 1806–1820. doi:10.1161/ATVBAHA.117.310663

Delić, D., Eisele, C., Schmid, R., Baum, P., Wiech, F., Gerl, M., et al. (2016). Urinary exosomal miRNA signature in type II diabetic nephropathy patients. *PloS one* 11 (3), e0150154. doi:10.1371/journal.pone.0150154

Deng, J., Zheng, C., Hua, Z., Ci, H., Wang, G., and Chen, L. (2022). Diosmin mitigates high glucose-induced endoplasmic reticulum stress through PI3K/AKT pathway in HK-2 cells. *BMC complementary Med. Ther.* 22 (1), 116. doi:10.1186/s12906-022-03597-y

Draney, C., Austin, M. C., Leifer, A. H., Smith, C. J., Kener, K. B., Aitken, T. J., et al. (2018). HDAC1 overexpression enhances β-cell proliferation by down-regulating Cdkn1b/p27. *Biochem. J.* 475 (24), 3997–4010. doi:10.1042/BCJ20180465

Eizirik, D. L., Pasquali, L., and Cnop, M. (2020). Pancreatic β-cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.* 16 (7), 349–362. doi:10.1038/s41574-020-0355-7

Foster, S. R., Dilworth, L. L., Thompson, R. K., Alexander-Lindo, R. L., and Omoruyi, F. O. (2017). Effects of combined inositol hexakisphosphate and inositol supplement on antioxidant activity and metabolic enzymes in the liver of streptozotocin-induced type 2 diabetic rats. *Chemico-biological Interact.* 275, 108–115. doi:10.1016/j.cbi.2017.07.024

Geerlings, S. E., and Hoepelman, A. I. (1999). Immune dysfunction in patients with diabetes mellitus (DM). *FEMS Immunol. Med. Microbiol.* 26 (3-4), 259–265. doi:10.1111/j.1574-695X.1999.tb01397.x

Ghemrawi, R., and Khair, M. (2020). Endoplasmic reticulum stress and unfolded protein response in neurodegenerative diseases. *Int. J. Mol. Sci.* 21 (17), 6127. doi:10.3390/ijms21176127

Heddad Masson, M., Poisson, C., Guérardel, A., Mamin, A., Philippe, J., and Gosmain, Y. (2014). Foxa1 and Foxa2 regulate α-cell differentiation, glucagon biosynthesis, and secretion. *Endocrinology* 155 (10), 3781–3792. doi:10.1210/en.2013-1843

Herrema, H., Guan, D., Choi, J. W., Feng, X., Salazar Hernandez, M. A., Faruk, F., et al. (2022). FKBP11 rewires UPR signaling to promote glucose homeostasis in type 2 diabetes and obesity. *Cell metab.* 34 (7), 1004–1022.e8. doi:10.1016/j.cmet.2022.06.007

Hou, Q., Hu, K., Liu, X., Quan, J., and Liu, Z. (2018). HADC regulates the diabetic vascular endothelial dysfunction by targetting MnSOD. *Biosci. Rep.* 38 (5). doi:10.1042/BSR20181042

Hsu, C. C., Zhang, X., Wang, G., Zhang, W., Cai, Z., Pan, B. S., et al. (2021). Inositol serves as a natural inhibitor of mitochondrial fission by directly targeting AMPK. *Mol. cell* 81 (18), 3803–3819.e7. doi:10.1016/j.molcel.2021.08.025

Huang, Q., Liu, Z., Yang, Y., Yang, Y., Huang, T., Hong, Y., et al. (2023). Selenium Nanodots (SENDs) as antioxidants and antioxidant-prodrugs to rescue islet β cells in type 2 diabetes mellitus by restoring mitophagy and alleviating endoplasmic reticulum stress. *Adv. Sci. (Weinh)* 10 (19), e2300880. doi:10.1002/advs.202300880

Itcho, K., Oki, K., Gomez-Sanchez, C. E., Gomez-Sanchez, E. P., Ohno, H., Kobuke, K., et al. (2020). Endoplasmic reticulum chaperone calmegin is upregulated in aldosterone-producing adenoma and associates with aldosterone production. *Hypertens. Dallas, Tex* 75 (2), 492–499. doi:10.1161/HYPERTENSIONAHA.119.14062

Jayasooriya, R., Dilshara, M. G., Karunarathne, W., Molagoda, I. M. N., Choi, Y. H., and Kim, G. Y. (2018). Camptothecin enhances c-Myc-mediated endoplasmic reticulum stress and leads to autophagy by activating Ca(2+)-mediated AMPK. *Food Chem. Toxicol.* 121, 648–656. doi:10.1016/j.fct.2018.09.057

Kasala, E. R., Bodduluru, L. N., Barua, C. C., Sriram, C. S., and Gogoi, R. (2015). Benzo(a)pyrene induced lung cancer: role of dietary phytochemicals in chemoprevention. *Pharmacol. Rep.* 67 (5), 996–1009. doi:10.1016/j.pharep.2015.03.004

Keestra-Gounder, A. M., Byndloss, M. X., Seyffert, N., Young, B. M., Chávez-Arroyo, A., Tsai, A. Y., et al. (2016). NOD1 and NOD2 signalling links ER stress with inflammation. *Nature* 532 (7599), 394–397. doi:10.1038/nature17631

Kong, F. J., Ma, L. L., Guo, J. J., Xu, L. H., Li, Y., and Qu, S. (2018). Endoplasmic reticulum stress/autophagy pathway is involved in diabetes-induced neuronal apoptosis and cognitive decline in mice. *Clin. Sci. Lond. Engl. 1979* 132 (1), 111–125. doi:10.1042/CS20171432

Lee, J. H., and Lee, J. (2022). New understandings from the biophysical study of the structure, dynamics, and function of nucleic acids 2.0. *Int. J. Mol. Sci.* 23 (9), 15822. doi:10.3390/ijms232415822

Li, T., Yu, X., Zhu, X., Wen, Y., Zhu, M., Cai, W., et al. (2022). Vaccarin alleviates endothelial inflammatory injury in diabetes by mediating miR-570-3p/HDAC1 pathway. *Front. Pharmacol.* 13, 956247. doi:10.3389/fphar.2022.956247

Li, X., Ji, Q., Zhong, C., Wu, C., Wu, J., Yuan, C., et al. (2024). Ghrelin regulates the endoplasmic reticulum stress signalling pathway in gestational diabetes mellitus. *Biochem. Biophys. Res. Commun.* 709, 149844. doi:10.1016/j.bbrc.2024.149844

Liang, B., Chen, S. W., Li, Y. Y., Zhang, S. X., and Zhang, Y. (2023). Comprehensive analysis of endoplasmic reticulum stress-related mechanisms in type 2 diabetes mellitus. *World J. Diabetes* 14 (6), 820–845. doi:10.4239/wjd.v14.i6.820

Ma, J. H., Wang, J. J., and Zhang, S. X. (2014). The unfolded protein response and diabetic retinopathy. *J. diabetes Res.* 2014, 160140. doi:10.1155/2014/160140

Mahon, J. L., Dupre, J., and Stiller, C. R. (1993). Lessons learned from use of cyclosporine for insulin-dependent diabetes mellitus. The case for immunotherapy for insulin-dependent diabetics having residual insulin secretion. *Ann. N. Y. Acad. Sci.* 696, 351–363. doi:10.1111/j.1749-6632.1993.tb17171.x

Oh, Y. S., Lee, Y. J., Kang, Y., Han, J., Lim, O. K., and Jun, H. S. (2013). Exendin-4 inhibits glucolipotoxic ER stress in pancreatic β cells via regulation of SREBP1c and C/EBPβ transcription factors. *J. Endocrinol.* 216 (3), 343–352. doi:10.1530/JOE-12-0311

Ong, H. K., Soo, B. P. C., Chu, K. L., and Chao, S. H. (2018). XBP-1, a cellular target for the development of novel anti-viral strategies. *Curr. protein and peptide Sci.* 19 (2), 145–154. doi:10.2174/1389203718666170911144812

Pavan, I. C., Yokoo, S., Granato, D. C., Meneguello, L., Carnielli, C. M., Tavares, M. R., et al. (2016). Different interactomes for p70-S6K1 and p54-S6K2 revealed by proteomic analysis. *Proteomics* 16 (20), 2650–2666. doi:10.1002/pmic.201500249

Pillai, R. A., Islam, M. O., Selvam, P., Sharma, N., Chu, A. H. Y., Watkins, O. C., et al. (2021). Placental inositol reduced in gestational diabetes as glucose alters inositol transporters and IMPA1 enzyme expression. *J. Clin. Endocrinol. Metab.* 106 (2), e875–e890. doi:10.1210/clinem/dgaa814

Provvisiero, D. P., Pivonello, C., Muscogiuri, G., Negri, M., de Angelis, C., Simeoli, C., et al. (2016). Influence of bisphenol A on type 2 diabetes mellitus. *Int. J. Environ. Res. Public Health* 13 (10), 989. doi:10.3390/ijerph13100989

Robertson, R. P., and Harmon, J. S. (2006). Diabetes, glucose toxicity, and oxidative stress: a case of double jeopardy for the pancreatic islet beta cell. *Free Radic. Biol. and Med.* 41 (2), 177–184. doi:10.1016/j.freeradbiomed.2005.04.030

Sak, F., Sengul, F., and Vatansev, H. (2024). The role of endoplasmic reticulum stress in metabolic diseases. *Metab. Syndr. Relat. Disord.* 22, 487–493. doi:10.1089/met.2024.0013

Sakurai, H. (2012). Copper compounds ameliorate cardiovasclur dysfunction and diabetes in animals. *Yakugaku Zasshi* 132 (3), 285–291. doi:10.1248/yakushi.132.285

Shen, Y., Cao, Y., Zhou, L., Wu, J., and Mao, M. (2022). Construction of an endoplasmic reticulum stress-related gene model for predicting prognosis and immune features in kidney renal clear cell carcinoma. *Front. Mol. Biosci.* 9, 928006. doi:10.3389/fmolb.2022.928006

Sitasawad, S., Deshpande, M., Katdare, M., Tirth, S., and Parab, P. (2001). Beneficial effect of supplementation with copper sulfate on STZ-diabetic mice (IDDM). *Diabetes Res. Clin. Pract.* 52 (2), 77–84. doi:10.1016/s0168-8227(00)00249-7

Su, J., Peng, J., Wang, L., Xie, H., Zhou, Y., Chen, H., et al. (2023). Identification of endoplasmic reticulum stress-related biomarkers of diabetes nephropathy based on bioinformatics and machine learning. *Front. Endocrinol.* 14, 1206154. doi:10.3389/fendo.2023.1206154

Sun, Y., Guo, L.-q., Wang, D.-g., Xing, Y.-j., Bai, Y.-p., Zhang, T., et al. (2023). Metformin alleviates glucolipotoxicity-induced pancreatic β cell ferroptosis through regulation of the GPX4/ACSL4 axis. *Eur. J. Pharmacol.* 956, 175967. doi:10.1016/j.ejphar.2023.175967

Terai, K., Sakamoto, K., Goto, M., Matsuda, M., Kasamaki, S., Shinmura, K., et al. (2006). Greater development of 1,2-dimethylhydrazine-induced colon cancer in a rat model of type 2 diabetes mellitus. *J. Int. Med. Res.* 34 (4), 385–389. doi:10.1177/147323000603400407

Vallée, D., Blanc, M., Lebeaupin, C., and Bailly-Maitre, B. (2020). Endoplasmic reticulum stress response and pathogenesis of non-alcoholic steatohepatitis. *Med. Sci. M/S* 36 (2), 119–129. doi:10.1051/medsci/2020008

Wada, J., and Nakatsuka, A. (2016). Mitochondrial dynamics and mitochondrial dysfunction in diabetes. *Acta medica Okayama* 70 (3), 151–158. doi:10.18926/AMO/54413

Wang, P., Chen, F., and Zhang, X. (2020). Cyclosporine-a attenuates retinal inflammation by inhibiting HMGB-1 formation in rats with type 2 diabetes mellitus. *BMC Pharmacol. Toxicol.* 21 (1), 9. doi:10.1186/s40360-020-0387-6

Wicks, K., Torbica, T., Umehara, T., Amin, S., Bobola, N., and Mace, K. A. (2015). Diabetes inhibits gr-1+ myeloid cell maturation via cebpa deregulation. *Diabetes* 64 (12), 4184–4197. doi:10.2337/db14-1895

Yilmaz, E. (2017). Endoplasmic reticulum stress and obesity. *Adv. Exp. Med. Biol.* 960, 261–276. doi:10.1007/978-3-319-48382-5_11

Yin, X., Yang, Z., Zhu, M., Chen, C., Huang, S., Li, X., et al. (2022). ILF2 contributes to hyperproliferation of keratinocytes and skin inflammation in a KLHDC7B-DT-Dependent manner in psoriasis. *Front. Genet.* 13, 890624. doi:10.3389/fgene.2022.890624

Yong, J., Johnson, J. D., Arvan, P., Han, J., and Kaufman, R. J. (2021). Therapeutic opportunities for pancreatic β-cell ER stress in diabetes mellitus. *Nat. Rev. Endocrinol.* 17 (8), 455–467. doi:10.1038/s41574-021-00510-4

Yuan, M., Gong, M., He, J., Xie, B., Zhang, Z., Meng, L., et al. (2022). IP3R1/GRP75/VDAC1 complex mediates endoplasmic reticulum stress-mitochondrial oxidative stress in diabetic atrial remodeling. *Redox Biol.* 52, 102289. doi:10.1016/j.redox.2022.102289

Zhang, J., Guo, J., Yang, N., Huang, Y., Hu, T., and Rao, C. (2022). Endoplasmic reticulum stress-mediated cell death in liver injury. *Cell death and Dis.* 13 (12), 1051. doi:10.1038/s41419-022-05444-x

Zhang, R., Bian, C., Gao, J., and Ren, H. (2023). Endoplasmic reticulum stress in diabetic kidney disease: adaptation and apoptosis after three UPR pathways. *Apoptosis* 28 (7-8), 977–996. doi:10.1007/s10495-023-01858-w

Zhang, Y., Chen, W., and Wang, Y. (2020). STING is an essential regulator of heart inflammation and fibrosis in mice with pathological cardiac hypertrophy via endoplasmic reticulum (ER) stress. *Biomed. and Pharmacother. = Biomedecine and Pharmacother.* 125, 110022. doi:10.1016/j.biopha.2020.110022

Zhao, T., Wu, K., Hogstrand, C., Xu, Y. H., Chen, G. H., Wei, C. C., et al. (2020). Lipophagy mediated carbohydrate-induced changes of lipid metabolism via oxidative stress, endoplasmic reticulum (ER) stress and ChREBP/PPARγ pathways. *Cell. Mol. life Sci. CMLS* 77 (10), 1987–2003. doi:10.1007/s00018-019-03263-6

Zhao, Z., Wang, X., Lu, M., and Gao, Y. (2023). Rosuvastatin improves endothelial dysfunction in diabetes by normalizing endoplasmic reticulum stress via calpain-1 inhibition. *Curr. Pharm. Des.* 29 (32), 2579–2590. doi:10.2174/0113816128250494231016065438

Zheng, Z., Zhang, S., Chen, J., Zou, M., Yang, Y., Lu, W., et al. (2022). The HDAC2/SP1/miR-205 feedback loop contributes to tubular epithelial cell extracellular matrix production in diabetic kidney disease. *Clin. Sci. Lond. Engl. 1979* 136 (3), 223–238. doi:10.1042/CS20210470

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 05 March 2025
DOI 10.3389/fgene.2025.1560172

OPEN ACCESS

EDITED BY
Yuriy L. Orlov,
I.M.Sechenov First Moscow State Medical University, Russia

REVIEWED BY
Zhu Yaodong,
Ministry of Health, Malaysia
Shiyan Wang,
Huaiyin Institute of Technology, China

*CORRESPONDENCE
Bo Han,
bhan@shsmu.edu.cn
Jun Lu,
lujun512@yahoo.com
Qin Shi,
shiqin0506@163.com

†These authors have contributed equally to this work

RECEIVED 15 January 2025
ACCEPTED 10 February 2025
PUBLISHED 05 March 2025

CITATION
Wang J, Zhu L, Li Y, Ding M, Wang X, Xiong B, Chen H, Chang L, Chen W, Han B, Lu J and Shi Q (2025) Multi-omics analysis reveals Jianpi formula-derived bioactive peptide-YG-22 potentially inhibited colorectal cancer via regulating epigenetic reprogram and signal pathway regulation.
Front. Genet. 16:1560172.
doi: 10.3389/fgene.2025.1560172

Multi-omics analysis reveals Jianpi formula-derived bioactive peptide-YG-22 potentially inhibited colorectal cancer via regulating epigenetic reprogram and signal pathway regulation

Jun Wang[1,2†], Lijuan Zhu[3†], Yuanyuan Li[4†], Mingming Ding[1], Xiyu Wang[1], Bo Xiong[5], Hongyu Chen[1], Lisheng Chang[1], Wenli Chen[1], Bo Han[6*], Jun Lu[7*] and Qin Shi[1*]

[1]Department of Oncology, Baoshan District Hospital of Integrated Traditional Chinese and Western Medicine of Shanghai, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2]Department of General Surgery, Baoshan District Hospital of Integrated Traditional Chinese and Western Medicine of Shanghai, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Department of Anorectal, Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [4]Department of Neurology and Institute of Neurology, Ruijin Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China, [5]Department of Clinical Pharmacy, Baoshan District Hospital of Integrated Traditional Chinese and Western Medicine of Shanghai, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [6]Key Laboratory for Translational Research and Innovative Therapeutics of Gastrointestinal Oncology, Hongqiao International Institute of Medicine, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [7]Shanghai Institute of Thoracic Oncology, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Introduction: Colorectal cancer (CRC) is a prevalent malignancy worldwide, often treated with chemotherapy despite its limitations, including adverse effects and resistance. The traditional Chinese medicine (TCM) Jianpi formula has been demonstrated to improve efficacy of chemotherapy, however the underlying mechanisms still need to be explored. In this study, we aim to screen bioactive peptides derived from the blood of CRC patients through peptidomics and explore the molecular mechanisms of the candidate peptides in the inhibition of CRC using multi-omics analysis.

Methods: In this study, we recruited 10 patients with CRC who had received either adjuvant chemotherapy or adjuvant chemotherapy combined with the traditional Chinese medicine Jianpi formula after surgery. We collected plasma samples at 2 cycles of adjuvant therapy and performed peptidomic analysis on these samples. The differentially bioactive peptides were screened using a model of HCT116 cells *in vitro*. To investigate the molecular mechanism underlying YG-22's inhibition of the colorectal cancer cell line HCT116, we performed a multi-omics analysis, including transcriptome, metabolome, chromatin accessibility, H3K4Me3 histone methylation, and NF-κB binding site analyses.

Results: Differential peptides were identified in plasma samples from patients treated with adjuvant chemotherapy combined with the Jianpi formula. Among these peptides, YG-22 exhibited the strongest cytotoxic effect on HCT116 cells, reducing cell viability in a dose- and time-dependent manner. Transcriptome analysis highlighted that YG-22 treatment in CRC modulates key pathways

associated with lysosome-mediated degradation and apoptosis. Metabolomic profiling further indicated disruptions in tumor-supportive metabolic pathways. Chromatin accessibility and histone modification analyses suggested that YG-22 induces epigenetic reprogramming. Additionally, treatment with YG-22 resulted in significant changes in NF-κB binding and pathway activation.

Conclusions: This study demonstrates that combining chemotherapy with TCM Jianpi formula enriches the molecular landscape and generates bioactive peptides with strong antitumor activity. Furthermore, this study also lays the foundation for further development of peptide-based therapies and highlights the value of combining traditional and modern therapeutic strategies for CRC management.

KEYWORDS

colorectal cancer, bioactive peptide, YG-22, Jianpi formula, multi-omics

Introduction

Colorectal cancer (CRC) ranks as the third most prevalent malignancy globally and is a leading cause of cancer-related deaths (Xi and Xu, 2021; Morgan et al., 2023; Siegel et al., 2023). Its significant incidence and mortality rates pose a major burden on healthcare systems worldwide (Xi and Xu, 2021; Alsakarneh et al., 2024; Klimeck et al., 2023). While advancements in surgery, chemotherapy, and radiotherapy have markedly improved patient outcomes, these conventional therapies are often accompanied by severe limitations (Morris et al., 2023; Feria and Times, 2024). High recurrence rates, drug resistance, and debilitating side effects—such as myelosuppression and gastrointestinal toxicity—frequently compromise their effectiveness (Adebayo et al., 2023; Al Bitar et al., 2023). Compounding these challenges is the alarming rise in CRC cases among younger populations, further underscoring the urgent need for novel therapeutic approaches that are both efficacious and less toxic (Constantinou and Constantinou, 2024).

Chemotherapy, a cornerstone of CRC treatment, continues to play a critical role in disease management (Morris et al., 2023). However, its efficacy remains limited, and its adverse effects significantly impact patients' quality of life. In recent years, complementary and alternative medicine, particularly traditional Chinese medicine (TCM), has garnered attention as an adjunctive strategy in CRC therapy (Jiang et al., 2023; Wu et al., 2024; Lin et al., 2023; Chen et al., 2018; Chen et al., 2019; McCulloch et al., 2016). The Jianpi formula, a TCM approach rooted in holistic principles, has shown potential in addressing some of the limitations associated with chemotherapy (Zhou et al., 2019). Evidence suggests that the Jianpi formula exhibits antitumor properties, including the inhibition of tumor proliferation, induction of apoptosis, and modulation of the tumor microenvironment (He et al., 2025; Peng et al., 2018). Furthermore, it has been reported to mitigate chemotherapy-induced complications, such as neutropenia, while enhancing patients' immune responses and overall quality of life (Zhou et al., 2019).

Despite these promising outcomes, the precise molecular mechanisms underlying the synergistic effects of chemotherapy and the Jianpi formula remain largely unexplored. This gap in knowledge has hindered the broader clinical adoption of this integrative therapeutic approach. Recent advancements in high-throughput technologies, such as peptidomics (Wang et al., 2012) and multi-omics analyses (Zhao et al., 2024), provide an opportunity to investigate these mechanisms in greater depth. Bioactive peptides, which are short amino acid sequences with regulatory functions, have been identified as critical players in various biological processes, including cancer progression and treatment response (Quintal-Bojórquez and Segura-Campos, 2021; Cui et al., 2019; Zhang et al., 2023). By screening bioactive peptides derived from the blood of CRC patients, researchers can uncover key molecular pathways influenced by these peptides. However, limited research has explored whether combining the Jianpi formula with chemotherapy alters the peptide profile in the peripheral blood of CRC patients and whether these peptides contribute to enhancing chemotherapy efficacy.

In this study, we aim to identify and characterize bioactive peptides in CRC patients' blood using peptidomics, and to explore their biological functions in inhibiting CRC. By integrating multi-omics approaches, including transcriptomics, proteomics, and metabolomics, we seek to elucidate the molecular mechanisms through which candidate peptides exert their effects.

Materials and methods

Patient enrollment, grouping, treatment, and blood sample collection

This study enrolled 10 colorectal cancer (CRC) patients, divided into two treatment groups: chemotherapy alone (n = 5) and chemotherapy combined with Jianpi formula (n = 5). Chemotherapy was performed as standard postoperative adjuvant therapy. The composition of Jianpi formula including Huangqi (*Astragalus membranaceus*, 30 g), Dangshen (*Codonopsis pilosula*,12 g), Chao-baizhu (*Atractylodes macrocephala*, 12 g), Fuling (*Poria cocos,* 12 g), Zhi-gancao (*Glycyrrhiza uralensis Fisch*, 6 g), Banxia (*Pinellia ternate*, 6 g), Tianlong (*Gekko japonicus*, 6 g), Hongteng (*Caulis sargentodoxae*, 30 g), Tengligen (*Actinidia arguta*, 30 g), Sheng-muli (*Ostreae Concha*, 30 g). The formula was decocted into a solution and taken orally twice a day. The extract from the Jianpi formula has been shown to induce apoptosis in HCT116 cells, prevent HCT116 cell invasion, and inhibit HCT116 cell viability *in vitro* (data not shown). Patients' clinical characteristics, including gender, age (54–79 years), and histopathological subtypes such as ulcerative adenocarcinoma, moderate-differentiated adenocarcinoma, poor-differentiated

TABLE 1 Clinical information of the colorectal cancer patients who received chemotherapy or chemotherapy plus Jianpi formula.

Groups	Patients	Gender	Age	Types of pathology	Time of blood collection
Chemotherapy	Patient 1	Female	79	Ulcerative adenocarcinoma	2 cycles of therapy
	Patient 2	Male	74	Moderate-differentiated adenocarcinoma	2 cycles of therapy
	Patient 3	Female	54	Poor-differentiated adenocarcinoma	2 cycles of therapy
	Patient 4	Male	75	Moderate-differentiated adenocarcinoma	2 cycles of therapy
	Patient 5	Female	68	Moderate-differentiated adenocarcinoma	2 cycles of therapy
Chemotherapy + Jianpi formula	Patient 6	Female	71	Tubular adenocarcinoma	2 cycles of therapy
	Patient 7	Male	74	Ulcerative adenocarcinoma	2 cycles of therapy
	Patient 8	Female	67	Poor-differentiated adenocarcinoma	2 cycles of therapy
	Patient 9	Female	60	Poor-differentiated adenocarcinoma	2 cycles of therapy
	Patient 10	Male	71	Poor-differentiated adenocarcinoma	2 cycles of therapy

adenocarcinoma, and tubular adenocarcinoma, are listed in Table 1. Blood samples were collected after two cycles of therapy and were processed to isolate plasma, then stored at −80°C for peptidome analysis.

Polypeptide extraction

Polypeptides were extracted following a rigorous multi-step protocol. The protein samples extracted from plasma were lysed using a buffer containing 8 M urea and a 1× protein inhibitor cocktail (Roche Ltd., Basel, Switzerland). Mechanical disruption was performed through three intervals of 400 s each, followed by incubation on ice for 30 min. High-speed centrifugation at 15,000 rpm for 15 min at 4°C was used to collect the supernatant. Filtration with 3 kDa ultrafiltration spin columns (Millipore, Billerica) removed high-molecular-weight proteins, retaining peptides in the 0–3 kDa range. For peptides in the 3–10 kDa range, enzymatic hydrolysis was conducted by drying the samples, redissolving them in 100 μL of 100 mM TEAB, and incubating overnight with trypsin (Promega, Madison, WI) at 37°C. Peptides were desalted using C18 Zip Tips (MonoSpin C18, GL), dried under vacuum, and stored at −80°C for mass spectrometry analysis.

Nano-HPLC-MS/MS analysis and bioinformatic analysis

Re-dissolved peptides were analyzed using a Thermo Scientific™ Orbitrap Fusion Lumos mass spectrometer coupled to an EASY-nanoLC 1,200 system. A 3 μL sample was loaded onto a 25 cm analytical column (75 μm inner diameter, 1.9 μm resin, Dr. Maisch) and separated using a 130-min gradient. Buffer B (80% acetonitrile with 0.1% formic acid) was increased from 4% to 50% over 120 min, followed by an increase to 95% for the final 9 min. The column flow rate was maintained at 250 nL/min, and the temperature was set at 55°C. The mass spectrometer operated in data-dependent acquisition (DDA) mode, alternating between MS and MS/MS scans. Full-scan spectra were acquired at a resolution of 120,000 with an m/z range of 350–1,500, an AGC target of 8×10^5, and a maximum injection time of 50 ms. Precursor ions were fragmented using higher-energy collision dissociation (HCD) with normalized collision energies of 25, 30, and 35. MS/MS spectra were acquired at a resolution of 30,000 with an AGC target of 1×10^5 and a maximum injection time of 54 ms. A dynamic exclusion window of 30 s was applied to prevent repeated ion selection, ensuring comprehensive peptide profiling. Numbers of peptides unique peptides, and proteins were identified. Then, the distribution of proteins based on the number of peptides was calculated. Lastly, protein sequence coverage distribution, KEGG pathway enrichment, pearson correlation, and heatmap of differential peptides were analyzed.

Candidate peptides screening and cell viability evaluation

After comparing the differential expressed peptides between two treatment groups: chemotherapy alone (n = 5) and chemotherapy combined with Jianpi formula (n = 5). Six candidate peptides were finally screened out, including YP-16 (YGRKKRRQRRR-GPSVP), YP-17 (YGRKKRRQRRR-OLTSGP), YG-22 (YGRKKRRQRRR-DGSPGKDGVRG), YM-22 (YGRKKRRQRRR-LGEAFDGARDM), YP-23 (YGRKKRRQRRR-MEPLGRQLTSGP), and YD-28 (YGRKKRRQRRR-EDPQGDAAOKTDTSHHD). And then, the candidate peptides were synthesized based on their amino acid sequences to evaluate their effects on HCT116 cell viability. Briefly, a total of 1,500 cells per well were seeded in 96-well plates and incubated overnight in a culture medium under appropriate conditions. Following incubation, the cells were treated with indicated peptides (5 mg/mL) at the desired concentrations for 24 h. Cell viability was then assessed using the CCK-8 assay (Dojindo, Japan), following the manufacturer's instructions. After adding the CCK-8 reagent and incubating for the recommended time, absorbance was measured at 450 nm using a spectrophotometric plate reader (Bio-Tek, United States). Furthermore, HCT116 cells were treated with varying concentrations of YG-22 (0, 2, 4, 6, 8, and 10 mg/mL) for either 24 or 48 h to determine the IC50 values. After the incubation with

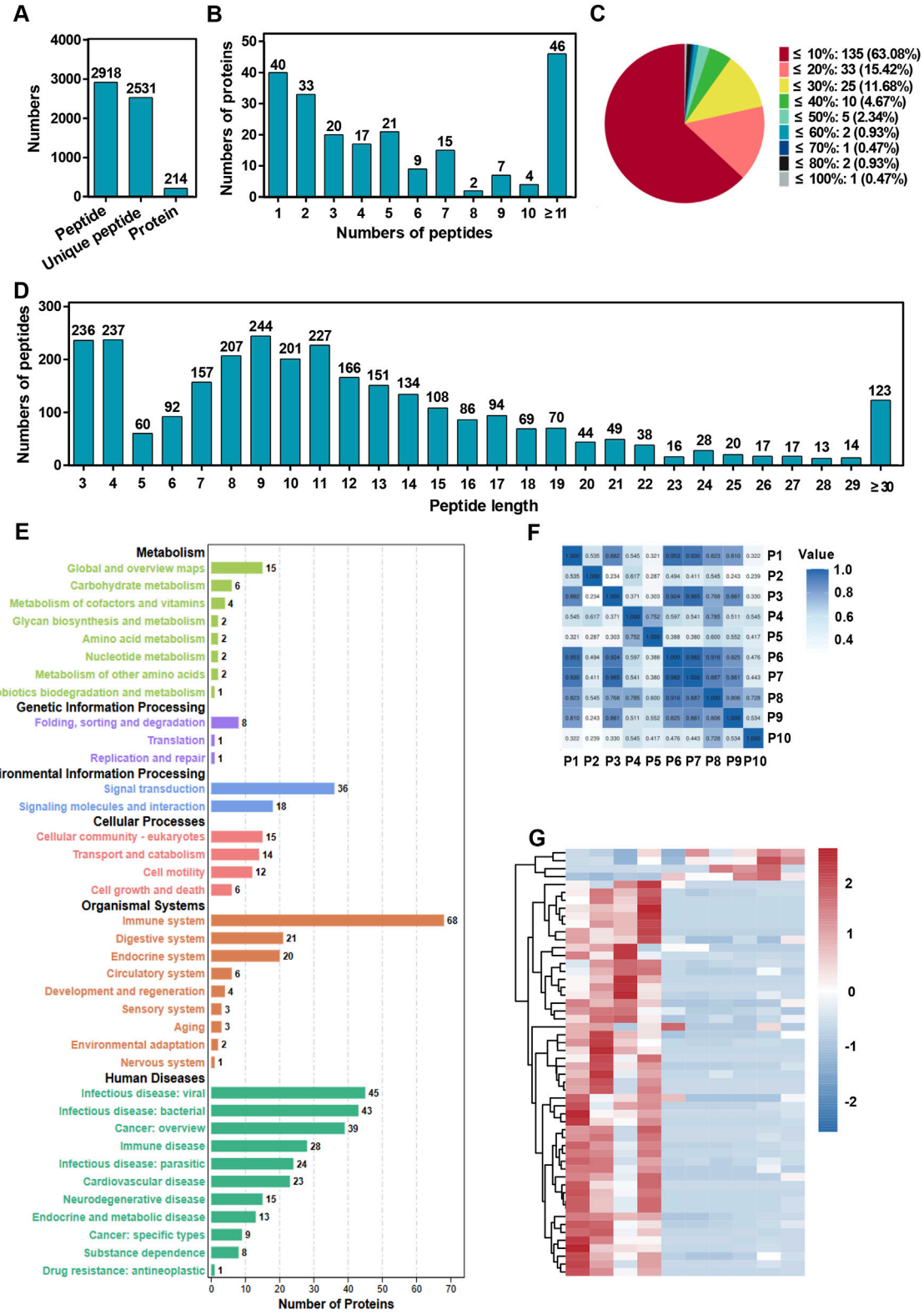


FIGURE 1
Peptidome analysis of plasma samples from colorectal cancer patients treated with chemotherapy or chemotherapy combined with the Jianpi formula. **(A)** Total number of identified peptides, unique peptides, and proteins. The bar chart summarizes the overall counts of peptides and proteins detected in the plasma samples. **(B)** Distribution of proteins based on the number of peptides identified. The histogram shows the frequency of proteins identified with varying numbers of peptides. **(C)** Protein sequence coverage distribution across the identified peptides. The pie chart illustrates the proportion of proteins categorized by sequence coverage percentages. **(D)** Protein sequence coverage distribution across the identified peptides. The pie chart illustrates the proportion of proteins categorized by sequence coverage percentages. **(E)** KEGG pathway analysis of proteins corresponding to the

(*Continued*)

FIGURE 1 (Continued)
identified peptides. The bar chart categorizes proteins into pathways related to metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases. The numbers on the bars represent the count of proteins associated with each category. **(F)** Pearson correlation heatmap comparing plasma sample data from colorectal cancer patients in the chemotherapy group and the chemotherapy plus Jianpi formula group. The heatmap illustrates the correlation coefficients between samples. **(G)** Heatmap of differential peptides comparing plasma sample data from colorectal cancer patients in the chemotherapy group and the chemotherapy plus Jianpi formula group. The heatmap illustrates the foldchange between samples.

YG-22 (IC50 concentration) for 48 h, the HCT116 cells were collected for subsequent analyses, including RNA sequencing (RNA-seq) to evaluate gene expression profiles, liquid chromatography-mass spectrometry (LC-MS) for metabolite analysis, assay for transposase-accessible chromatin using sequencing (ATAC-seq) to assess chromatin accessibility, and chromatin immunoprecipitation sequencing (ChIP-seq) to study the H3K4Me3 profiling and NF-κB protein-DNA interactions.

Transcriptome analysis

For transcriptome analysis, treated and control HCT116 cells (1×10^5) are collected, and total RNA is extracted using RNA isolation kit (Qiagen, Germany) following the manufacturer's instructions. Total RNA samples were prepared with an initial concentration of at least 20 ng/μL and a total quantity of at least 2 μg, ensuring an A260/A280 ratio between 1.9 and 2.1 for quality control. mRNA was isolated using oligo-dT beads to capture polyA-tailed transcripts, followed by thermal fragmentation into 200–300 bp fragments. Reverse transcription was performed using a strand synthesis master mix to generate cDNA. Library preparation involved end-repair, A-tailing, and ligation of sequencing adapters, followed by PCR amplification and size selection for fragments of 300–400 bp, including adapter sequences. The prepared libraries underwent high-throughput sequencing on the Illumina NovaSeq 6,000 platform, producing comprehensive transcriptomic data for downstream bioinformatics analysis. Differential gene expression analysis is performed using DESeq2 or edgeR, identifying significantly up- or downregulated genes based on fold changes and adjusted *P*-values. Functional enrichment analysis, including GO and KEGG pathway analysis, is conducted to interpret biological implications, and results are visualized with heatmaps, volcano plots, and pathway diagrams.

Metabolomics analysis

For metabolomics analysis, treated and control HCT116 cells (1×10^5) are collected, washed with cold PBS, and lysed using 80% methanol to extract metabolites. The lysates are centrifuged at 12,000–15,000 × g at 4°C, and the supernatants are stored at −80°C. Metabolite profiling is performed using liquid chromatography-mass spectrometry (LC-MS) with a reverse-phase LC column and gradient elution, followed by high-resolution mass spectrometry in positive and/or negative ion modes. Data preprocessing involves peak detection, alignment, normalization, and filtering using software such as XCMS or Compound Discoverer. Metabolites are identified by matching m/z values, retention times, and fragmentation patterns to databases like HMDB or METLIN, with MS/MS used for structural confirmation. Statistical analyses, including PCA, PLS-DA, and univariate tests, are conducted to identify significantly altered metabolites, with pathway mapping performed using tools like MetaboAnalyst or KEGG Mapper. The results are visualized with heatmaps, volcano plots, and pathway diagrams, providing insights into metabolic changes induced by treatment.

Chromatin accessibility analysis

For chromatin accessibility analysis using ATAC-seq, the process begins by lysing live HCT116 cells (5×10^5) to isolate nuclei using a lysis buffer containing RSB, NP-40, Tween-20, and digitonin, followed by centrifugation to remove supernatant. The isolated nuclei are treated with a transposase mix containing Tn5 transposase at 37°C for 30 min to fragment accessible chromatin regions and insert sequencing adapters. The DNA is then purified using MinElute kits (Qiagen, Germany) and subjected to PCR amplification to optimize library quality, with purified libraries dissolved in 10 mM Tris buffer. Following library preparation, high-throughput sequencing is conducted using the Illumina NovaSeq 6,000 platform. Data processing involves quality control with FastQC, adapter trimming using Trimmomatic, genome alignment with Bowtie2, and peak calling with MACS2. Peaks are annotated using ChIPseeker and enriched motifs identified using HOMER. Differential analysis of peaks between samples is performed using DiffBind, and functional enrichment analysis is carried out to link chromatin accessibility changes to biological pathways. Results are visualized through heatmaps, volcano plots, and motif enrichment diagrams, providing insights into chromatin structure and regulatory dynamics.

ChIP-seq analysis for H3K4Me3 profiling and NF-κB protein-DNA interactions

ChIP-seq analysis for H3K4Me3 profiling and NF-κB protein-DNA interactions was conducted using a combination of chromatin immunoprecipitation and high-throughput sequencing. HCT116 cells (5×10^6) were fixed with 1% formaldehyde to crosslink proteins and DNA, followed by quenching with glycine and lysis to isolate chromatin. The chromatin was sonicated to fragment DNA, and immunoprecipitation was performed using specific antibodies for H3K4Me3 and NF-κB, coupled with Protein A + G magnetic beads. The immunoprecipitated complexes were reverse crosslinked, and DNA was purified. High-throughput sequencing libraries were prepared through end repair, A-tailing, adapter ligation, and PCR amplification, targeting fragment sizes of 100–300 bp. Sequencing was performed on the Illumina NovaSeq 6,000 platform.

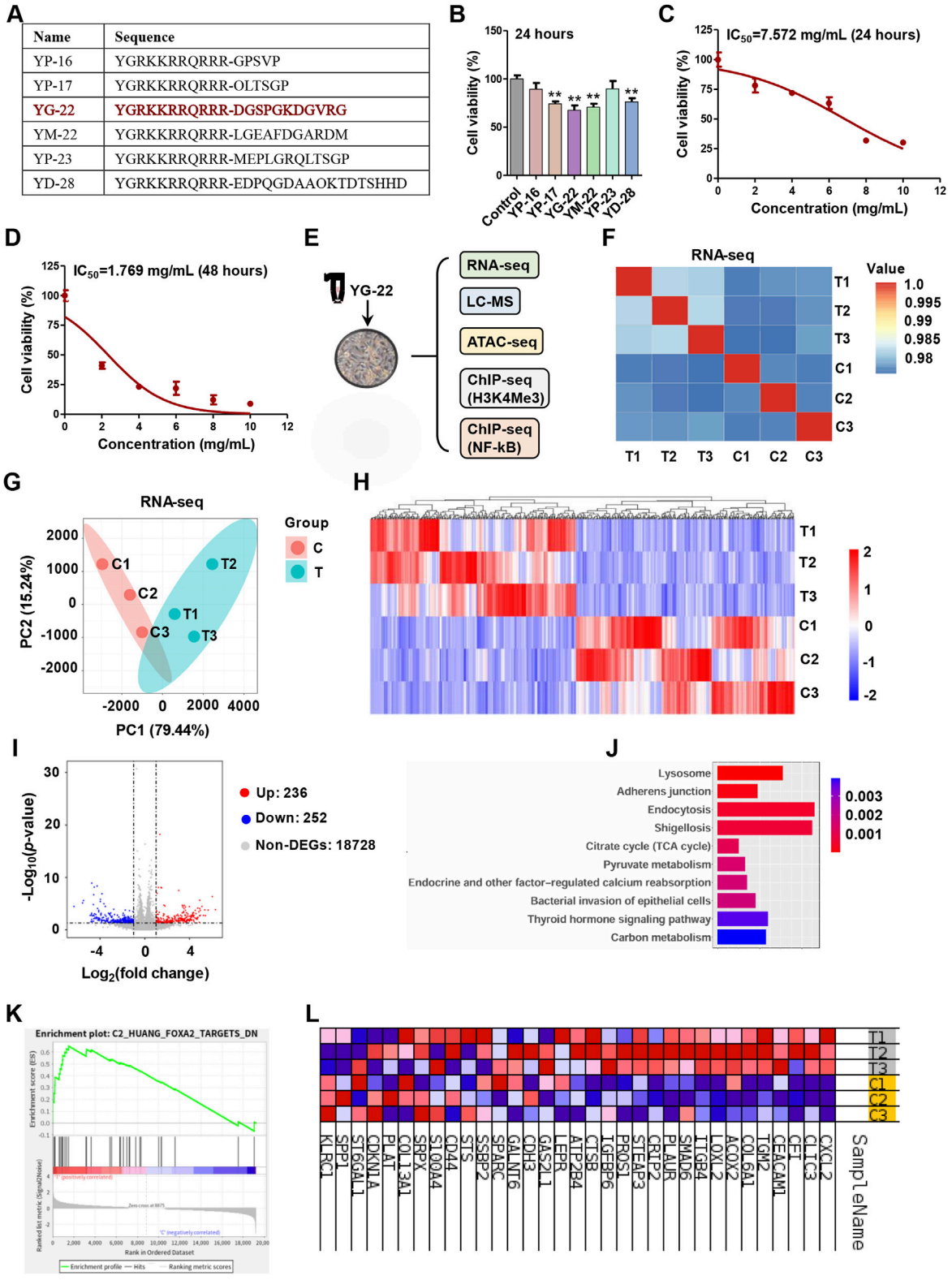


FIGURE 2
Evaluation of candidate peptide effects on HCT116 cell viability and transcriptome analysis of YG-22 treatment. **(A)** Amino acid sequences of the candidate differential peptides. The table lists the sequences for peptides YP-16, YP-17, YG-22, YM-22, YP-23, and YD-28. **(B)** Assessment of cell viability after treatment with candidate peptides (5 mg/mL) for 24 h. Data are presented as mean ± SEM (n = 3). $^{**}P < 0.01$ indicates statistically significant differences compared to the control. **(C)** Dose-response curve showing the viability of HCT116 cells treated with YG-22 at various concentrations (2, 4, 6, 8, and 10 mg/mL) for 24 h. IC50 for 24 h = 7.572 mg/mL. Data are presented as mean ± SD (n = 3). **(D)** Dose-response curve showing the viability of HCT116 cells treated with YG-22 at various concentrations for 48 h. IC50 for 48 h = 1.769 mg/mL. Data are presented as mean ± SD (n = 3). **(E)** Schematic
(*Continued*)

FIGURE 2 (Continued)
of multi-omics analysis performed on HCT116 cells treated with YG-22 (1.769 mg/mL) for 48 h. The analysis includes RNA-seq, LC-MS, ATAC-seq, and ChIP-seq targeting H3K4Me3 and NF-κB. **(F)** Pearson correlation heatmap comparing RNA-seq data from control (C1-C3) and YG-22-treated (T1-T3) samples. The heatmap illustrates high correlation within groups. **(G)** Principal Component Analysis (PCA) of RNA-seq data showing clear clustering between control and treated groups based on gene expression profiles. **(H)** Heatmap of differentially expressed genes (DEGs) between control and YG-22-treated groups. Red indicates upregulated genes, while blue indicates downregulated genes. **(I)** Heatmap of differentially expressed genes (DEGs) between control and YG-22-treated groups. Red indicates upregulated genes, while blue indicates downregulated genes. **(J)** KEGG pathway enrichment analysis of DEGs. Bar plots show significantly enriched pathways, with *P*-values represented by bar color intensity. **(K)** Gene Set Enrichment Analysis (GSEA) highlighting enriched pathways affected by YG-22 treatment. The plot shows pathway enrichment scores and ranks. **(L)** Heatmap of enriched differentially expressed genes involved in key pathways. Red and blue indicate higher and lower expression levels, respectively, across samples.

Bioinformatics analysis included quality control using FastQC, adapter trimming with fastp, and alignment of clean reads to the human reference genome (hg38) using Bowtie2. Peak calling was conducted with MACS2 to identify binding sites and histone modifications, followed by annotation using ChIPseeker. Motif analysis, using MEME and HOMER, identified enriched motifs within peak regions, particularly for NF-κB binding. Functional enrichment analysis was carried out with KEGG pathways to link identified peaks to pathways. Visualization tools, such as heatmaps and Circos plots, were used to highlight binding site distribution and signal enrichment across the genome, providing insights into chromatin modifications and transcription factor interactions under experimental conditions.

Statistical analysis

There were at least three biological replicates, excluding ATAC-seq and ChIP-seq analysis, for each group. Cell viability evaluation data were reported as means ± SEM. Student's t-test (two-tailed) or one-way ANOVA with Bonferroni's multiple comparison test were used. *P*-values of <0.05, or 0.01, or 0.001 were deemed significant.

Results

Clinical information of colorectal cancer patients enrolled in this study

The baseline clinical characteristics of colorectal cancer patients enrolled in this study are presented in Table 1. Patients were divided into two groups based on their treatment: chemotherapy alone (n = 5) and chemotherapy combined with Jianpi formula (n = 5). In the chemotherapy group, patients included a 79-year-old female with ulcerative adenocarcinoma, a 74-year-old male with moderate-differentiated adenocarcinoma, a 54-year-old female with poor-differentiated adenocarcinoma, a 75-year-old male with moderate-differentiated adenocarcinoma, and a 68-year-old female with moderate-differentiated adenocarcinoma. Similarly, in the chemotherapy plus Jianpi formula group, patients included a 71-year-old female with tubular adenocarcinoma, a 74-year-old male with ulcerative adenocarcinoma, a 67-year-old female with poor-differentiated adenocarcinoma, a 60-year-old female with poor-differentiated adenocarcinoma, and a 71-year-old male with poor-differentiated adenocarcinoma. Blood samples were collected from all patients following 2 cycles of therapy, ensuring consistent timing for peptidome analysis.

Peptidome and lnc-peptidome analysis in colorectal cancer patients

The peptidome analysis conducted on plasma samples from colorectal cancer patients who had received conventional chemotherapy alone and combination chemotherapy highlighted significant differences in the protein and peptide profiles between peptidome and long non coding-peptidome (lnc-peptidome). A total of 2,918 peptides, 2,531 unique peptides and 214 proteins were identified in peptidome, while the lnc-peptidome exhibited a remarkable increase, identifying 5,115 peptides, 4,908 unique peptides and 383 proteins (Figure 1A; Supplementary Figure 1A). The distribution of these proteins based on their identified peptides showed that the majority of proteins were linked to a limited number of peptides; however, a considerable proportion difference between peptides and lnc-peptides (Figure 1B; Supplementary Figure 1B). Additionally, sequence coverage analysis revealed that while many proteins enriched in <20% coverage in peptides, while with a higher proportion showing greater than 30% coverage in lnc-peptides (Figure 1C; Supplementary Figure 1C). Notably, peptide length distribution illustrated a significant difference, with the peptides presenting a higher frequency of peptide lengths between 7 and 12 amino acids compared to a more uniform distribution in the lnc-peptides (Figure 1D; Supplementary Figure 1D).

The KEGG pathway analysis for peptides and lnc-peptides demonstrated similar enrichment patterns, with additional pathways linked to development and cellular response represented predominantly in the lnc-peptidome (Figure 1E; Supplementary Figure 1E). The heatmaps further illustrated varying expression levels of differentially expressed peptides and lnc-peptides between the treatment groups, highlighting distinct clustering patterns that suggest a functional divergence in Jianpi formula treatments (Figures 1F, G; Supplementary Figures 1F, G). These comprehensive analyses indicate that the incorporation of the Jianpi formula markedly enriches the molecular landscape in plasma samples, underscoring its potential role in enhancing therapeutic efficacy in colorectal cancer.

Evaluation of candidate peptide effects on HCT116 cell viability and transcriptome analysis of YG-22 treatment

Among these differential peptides induced by Jianpi formula, we further screened 6 candidate peptides for cytotoxicity evaluation (Figure 2A). The analysis of candidate peptides revealed significant effects on HCT116 cell viability. The peptides, including YP-16, YP-17, YG-22, YM-22, YP-23, and YD-28, were tested at a

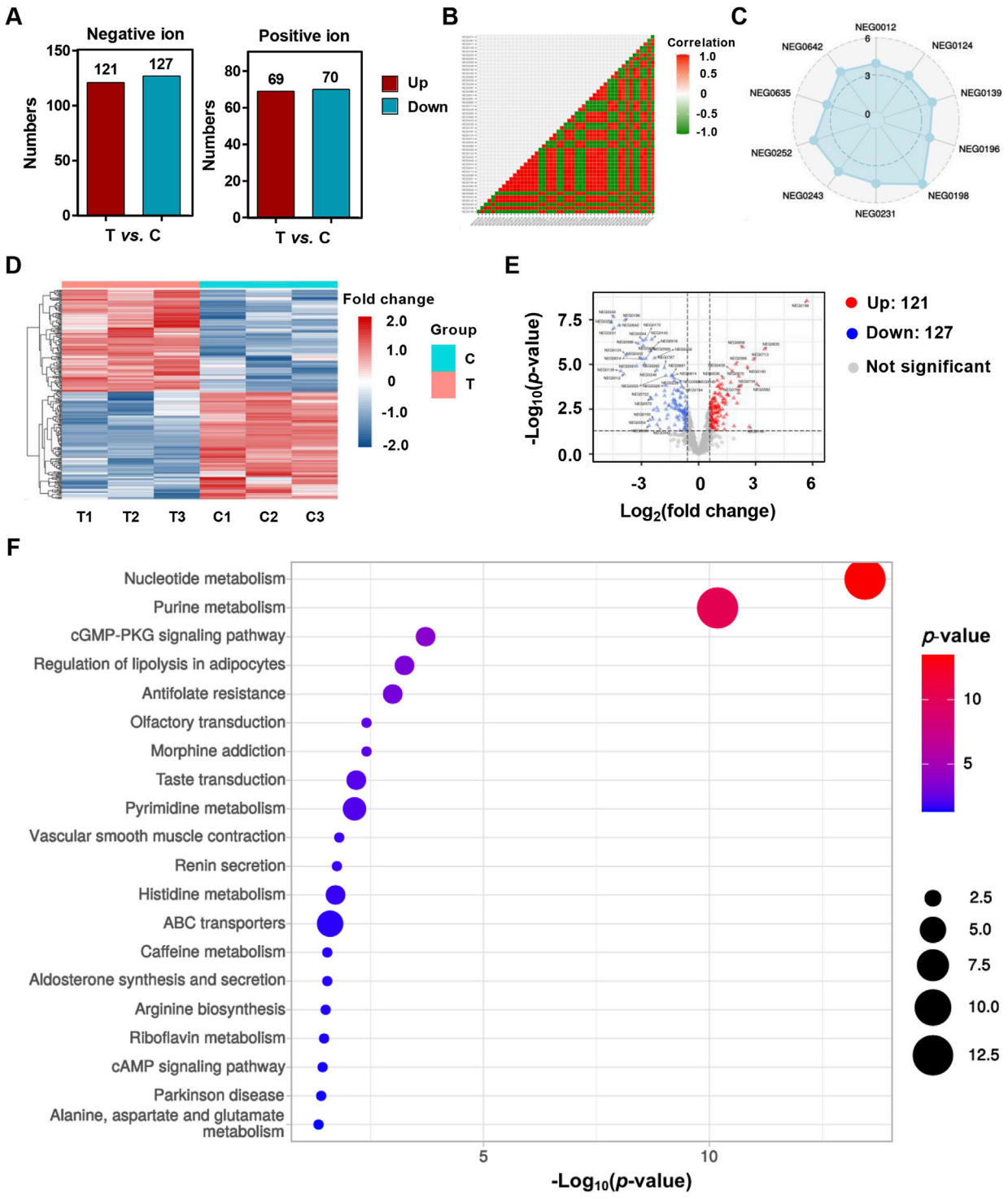


FIGURE 3
Metabolomics analysis of HCT116 cells treated with YG-22. **(A)** Total number of upregulated and downregulated metabolites identified in negative ion (NEG) and positive ion (POS) models. The bar chart shows the counts of significantly altered metabolites between treated (T) and control (C) groups. **(B)** Correlation matrix of metabolites detected in control and YG-22-treated samples based on the NEG mode. The heatmap displays correlation coefficients between samples, with red and green indicating positive and negative correlations, respectively. **(C)** Radar plot visualizing the distribution of specific metabolites detected in the NEG mode. The plot highlights the relative abundance of key metabolites across samples. **(D)** Heatmap of differentially expressed metabolites identified through NEG mode analysis. The heatmap shows hierarchical clustering and fold change in metabolite expression levels between control (C1-C3) and YG-22-treated (T1-T3) samples. **(E)** Volcano plot highlighting significant differential metabolites between control and YG-22-treated samples in NEG mode. Red dots indicate upregulated metabolites, blue dots represent downregulated metabolites, and gray dots denote non-significant changes. **(F)** KEGG pathway analysis of differentially expressed metabolites, categorizing them into metabolic pathways and biological processes. The dot plot represents pathway enrichment, with dot size indicating the number of metabolites involved and color denoting the *P*-value significance.

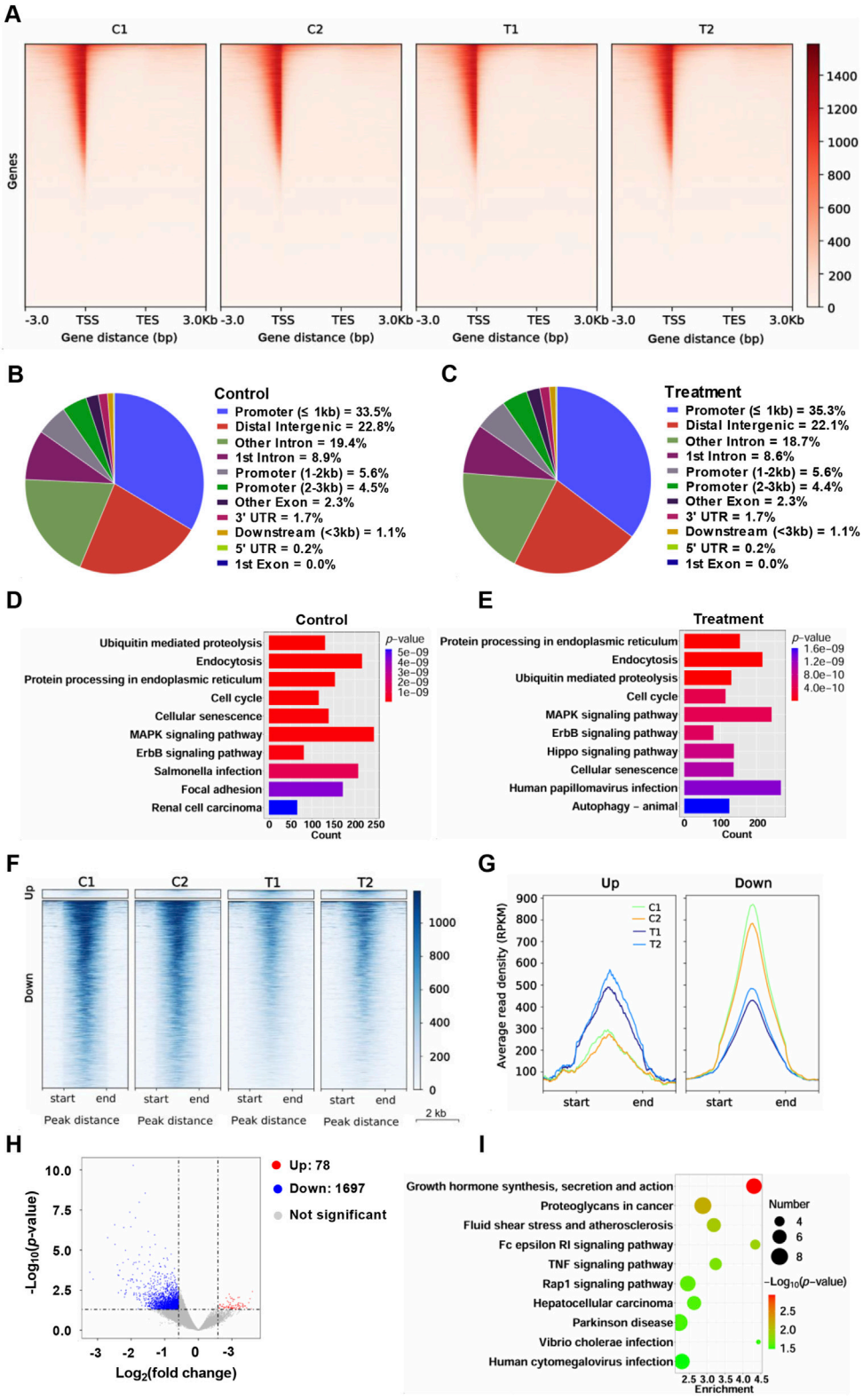


FIGURE 4
Chromatin accessibility analysis of HCT116 cells treated with YG-22. **(A)** Gene body coverage analysis across different samples. The heatmaps represent the chromatin accessibility signals distributed along gene bodies [from transcription start site (TSS) to transcription end site (TES)]. **(B)** Distribution of chromatin accessibility peaks across genomic regions in control samples. The pie chart highlights the proportion of peaks associated with various genomic features, including promoters, introns, and intergenic regions. **(C)** Distribution of chromatin accessibility peaks across genomic regions in YG-22-treated samples. The pie chart shows the peak distribution percentages across different genomic regions. **(D)** KEGG pathway analysis of promoter-associated chromatin accessibility peaks in control samples. The bar chart shows enriched pathways, with the length of the bars reflecting the

(*Continued*)

FIGURE 4 (Continued)
number of peaks associated with each pathway and the color indicating the *P*-value significance. **(E)** KEGG pathway analysis of promoter-associated chromatin accessibility peaks in YG-22-treated samples. The enriched pathways and their significance are represented as in **(D)**. **(F)** Heatmap showing differential chromatin accessibility peaks between control and YG-22-treated samples. Clustering highlights distinct patterns of chromatin accessibility for upregulated and downregulated peaks. **(G)** Average read density (RPKM) of upregulated and downregulated peaks in control and YG-22-treated samples. Line plots show changes in chromatin accessibility signal density for each condition. **(H)** Volcano plot illustrating significant differential chromatin accessibility peaks between control and YG-22-treated samples. Red dots indicate upregulated peaks, blue dots represent downregulated peaks, and gray dots correspond to non-significant peaks. **(I)** KEGG pathway enrichment analysis for differentially accessible chromatin regions. The bubble plot represents enriched pathways, where bubble size reflects the number of peaks associated with each pathway and color indicates statistical significance (*P*-value).

concentration of 5 mg/mL for 24 h. Results indicated that YP-17, YG-22, YM-22 and YD-28 notably reduced cell viability, achieving statistical significance compared to the control ($P < 0.01$) (Figure 2B). Considering that YG-22 derived from collagen type I alpha 1 protein showed the best inhibitory effect on CRC cells, we then evaluated IC50 at 24 h and 48 h respectively. The dose-response curves displayed a clear relationship between YG-22 concentration and cell viability, with the 24-h IC50 determined to be 7.572 mg/mL (Figure 2C) and a reduced IC50 of 1.769 mg/mL observed at 48 h (Figure 2D). These findings suggest that YG-22 exerts a dose- and time-dependent cytotoxic effect on HCT116 cells.

To further understand the molecular mechanisms impacted by YG-22 treatment, a comprehensive multi-omics analysis was conducted, including transcriptome, metabolomics, chromatin accessibility profiling, H3K4me3 profiling and NF-κB binding profiling (Figure 2E). The transcriptome results generated a Pearson correlation heatmap, indicating a relative high degree of correlation between control and YG-22-treated groups (Figure 2F). Further Principal Component Analysis (PCA) revealed distinct clustering between the treated and control groups based on gene expression profiles (Figure 2G). Differentially expressed genes (DEGs) were identified, with illustrating 236 upregulated and 252 downregulated genes in response to YG-22 treatment (Figures 2H, I). KEGG pathway enrichment analysis highlighted significant pathways impacted by YG-22, including lysosome, adherens junction, and endocytosis pathways (Figure 2J). Gene Set Enrichment Analysis (GSEA) further supported these findings, identifying enriched pathways critical for cellular responses to YG-22 treatment (Figure 2K). A heatmap summarizing the enriched DEGs indicated distinct expression patterns across samples, providing insights into the key biological processes modulated by YG-22 (Figure 2L). These results underscore the potential therapeutic mechanisms of YG-22 in colorectal cancer treatment.

Metabolomics analysis of HCT116 cells treated with YG-22

The metabolomics analysis of HCT116 cells treated with YG-22 highlighted significant alterations in metabolite profiles, showcasing distinct differences between the control and YG-22-treated groups. In the negative ion (NEG) mode, a total of 121 metabolites were upregulated, while 127 were downregulated, indicating notable metabolic shifts due to YG-22 treatment (Figure 3A). Correlation analyses demonstrated strong relationships among some metabolites, with a radar plot emphasizing the relative abundance of key metabolites across samples (Figure 3B). Heatmap representations illustrated hierarchical clustering of differentially expressed metabolites, reinforcing the distinctions between control and treated (Figure 3D). The volcano plot further elucidated these differences, with red dots denoting upregulated and blue dots marking downregulated metabolites, providing a clear visual of significant changes (Figure 3E). KEGG pathway analysis categorized these metabolites into various metabolic pathways, revealing important involvement in processes such as nucleotide and purine metabolism (Figure 3F).

In addition, the positive ion (POS) mode analysis complemented the findings from the NEG mode, providing further insights into the metabolic changes associated with YG-22 treatment. Similar correlation patterns were observed, reinforcing the relationships among metabolites in the control and YG-22-treated groups (Supplementary Figure 2A). The radar plots for selected metabolites in POS mode illustrated variations in abundance (Supplementary Figure 2B), while heatmap and volcano plot analyses confirmed the differential expression of metabolites, with 69 upregulated and 70 downregulated components visually represented significantly (Supplementary Figures 2C, D). The KEGG pathway analysis for this mode also highlighted a range of metabolic pathways impacted by YG-22 treatment, including ABC transporters and cAMP signaling pathways (Supplementary Figure 2E). Collectively, these analyses underscore the extensive metabolic reprogramming induced by YG-22 in HCT116 cells, highlighting critical biological processes that may contribute to its therapeutic efficacy in colorectal cancer.

Chromatin accessibility analysis of HCT116 cells treated with YG-22

To further observe the effect of YG-22 on chromatin status of HCT116 cells, we performed chromatin accessibility analysis. Chromatin accessibility analysis of HCT116 cells treated with YG-22 revealed significant differences in the chromatin landscape between control and YG-22-treated samples. Gene body coverage analysis, represented in heatmaps, indicated varied chromatin accessibility signals enriched around the transcription start site (TSS) (Figure 4A). The distribution of chromatin accessibility peaks was assessed through pie charts, highlighting the proportion of peaks associated with genomic features. In control samples, the analysis showed peaks primarily located within promoter, distal intergenic and intron (Figure 4B), while YG-22-treated samples displayed a slightly altered distribution of peaks across various regions (Figure 4C). KEGG pathway analysis

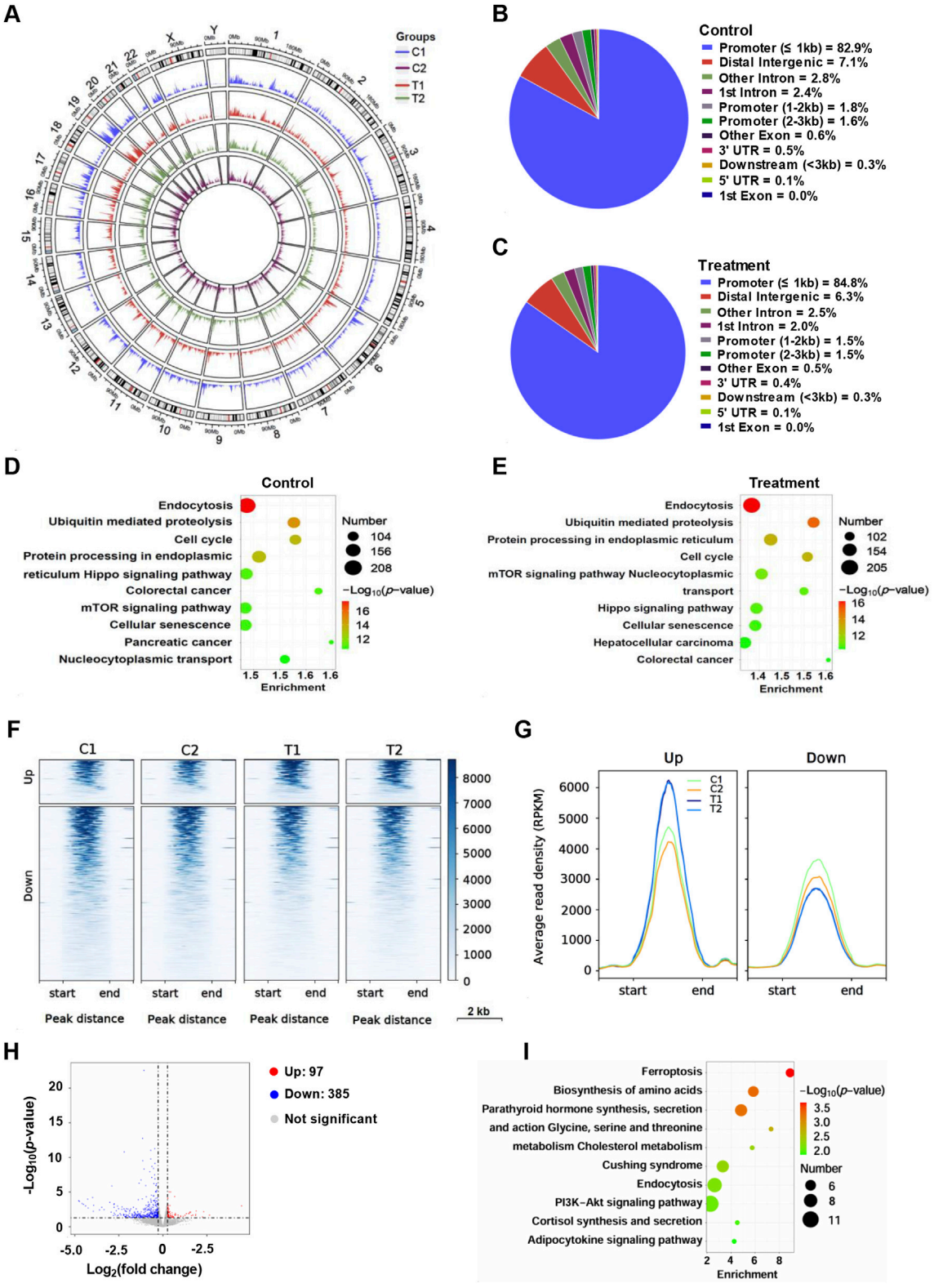


FIGURE 5
H3K4me3 profiling analysis of HCT116 cells treated with YG-22. **(A)** Circular plot visualizing the genome-wide distribution of H3K4me3 peaks across different samples (C1, C2, T1, T2). The plot highlights peak density within various genomic regions for each condition. **(B)** Distribution of H3K4me3 peaks across different chromatin regions in control samples. The pie chart represents the proportion of peaks located in promoters, introns, intergenic regions, and other genomic features. **(C)** Distribution of H3K4me3 peaks across different chromatin regions in YG-22-treated samples. The pie chart shows the genomic localization of peaks after treatment, indicating changes in distribution patterns. **(D)** KEGG pathway analysis of promoter-associated H3K4me3 peaks in control samples. The bubble plot highlights enriched pathways, with bubble size representing the number of peaks and color

(*Continued*)

FIGURE 5 (Continued)
indicating the *P*-value. **(E)** KEGG pathway analysis of promoter-associated H3K4me3 peaks in YG-22-treated samples. Enrichment analysis reveals pathways significantly associated with treatment-related changes in H3K4me3 marks. **(F)** Heatmap showing differential H3K4me3 peaks between control and YG-22-treated samples. The clustering patterns illustrate upregulated and downregulated H3K4me3 peaks across genomic regions. **(G)** Average read density (RPKM) of upregulated and downregulated H3K4me3 peaks in control and YG-22-treated samples. Line plots display chromatin signal intensities near the differential peaks. **(H)** Average read density (RPKM) of upregulated and downregulated H3K4me3 peaks in control and YG-22-treated samples. Line plots display chromatin signal intensities near the differential peaks. **(I)** KEGG pathway enrichment analysis of differentially accessible promoter-associated H3K4me3 peaks. Bubble size reflects the number of peaks associated with each pathway, and color represents the statistical significance (*P*-value).

demonstrated that numerous pathways were enriched in both groups, with differences in the number of accessible chromatin peaks associated with specific pathways, such as the ubiquitin-mediated proteolysis pathway and focal adhesion pathway (Figures 4D, E).

Differential chromatin accessibility between control and YG-22-treated samples was further elucidated using heatmap representations, which highlighted distinct patterns of upregulated and downregulated peaks (Figure 4F). The average read density (RPKM) analysis, indicated in line plots, illustrated changes in chromatin accessibility levels for both upregulated and downregulated peaks across conditions (Figure 4G). A volcano plot provided a visual summary of significant differential chromatin peaks, with 78 upregulated peaks and 1,697 downregulated peaks (Figure 4H). The KEGG pathway enrichment analysis for differentially accessible chromatin regions revealed that the pathways, including growth hormone synthesis, secretion and action pathway, proteoglycans in cancer pathway, were enriched significantly (Figure 4I). Supplementary analyses further clarified these findings, showcasing gene body coverage, Pearson correlation comparisons, and circular plots visualizing peak distributions across the genome (Supplementary Figure 3). This collective data underscores the extensive reconfiguration of chromatin accessibility induced by YG-22 in HCT116 cells, shedding light on potential regulatory mechanisms involved in its therapeutic effects.

H3K4me3 profiling analysis of HCT116 cells treated with YG-22

Furthermore, we performed ChIP-seq to analyze the H3K4Me3 profiling after YG-22 treatment upon HCT116 cells. The genome-wide distribution of H3K4me3 peaks was illustrated using a circular plot, showing peak density across various genomic regions for both control and treatment samples (Figure 5A). In control samples, 86.3% of H3K4me3 peaks were located in promoter regions, while 5.2% were found in introns and 7.1% in intergenic regions (Figure 5B). Upon treatment with YG-22, there was a slight change, with promoter-associated peaks increasing to 87.8%, intronic peaks decreasing to 4.5%, and intergenic peaks decreasing to 6.3% (Figure 5C). KEGG pathway analysis highlighted significant pathways associated with these peaks, revealing that pathways related to endocytosis and ubiquitin mediated proteolysis were enriched both in control and treatment (Figures 5D, E). Differential analysis indicated distinct clustering of upregulated and downregulated H3K4me3 modifications (Figure 5F). The average read density (RPKM) analysis, indicated in line plots, illustrated changes in H3K4me3 modification levels for both upregulated and downregulated peaks across conditions (Figure 5G). Volcano plot provided a visual summary of significant differential chromatin peaks, with 97 upregulated peaks and 385 downregulated peaks (Figure 5H). The KEGG pathway enrichment analysis for differentially H3K4me3 modification revealed that the pathways, including ferroptosis, biosynthesis of amino acids, parathyroid hormone synthesis, secretion, were enriched significantly (Figure 5I).

Supplementary analyses provided further insights into the consistency and significance of the H3K4me3 modifications. A heatmap analysis of ChIP quality confirmed the reliability of the ChIP-seq data (Supplementary Figures 4A, B). The average read density (RPKM) line plots indicated an increase in signal intensity at treatment samples (Supplementary Figure 4C). Additionally, combined circular peak distribution plots showed similar results as Figure 5A (Supplementary Figure 4D). TSS analyses further emphasized these findings, revealing that the signal density surrounding TSS regions (Supplementary Figure 4E). Collectively, these results underscore the pivotal role of H3K4me3 in modulating gene expression and illustrate the therapeutic potential of YG-22 in colorectal cancer through these alterations in chromatin modification.

ChIP-seq analysis of NF-κB binding in HCT116 cells treated with YG-22

Finally, ChIP-seq analysis was conducted to investigate the genome-wide distribution. After treated with YG-22 in HCT116 cells for 48 h, enrichment of NF-κB binding peaks was collected. A circular plot revealed the NF-κB binding distribution across chromosomes, highlighting changes between control and treated groups (Figure 6A). In the control group, NF-κB peaks were predominantly located in distal intergenic regions (28.9%) and promoters (39.0%) (Figure 6B), while treatment with YG-22 led to a shift, with 38.4% of peaks observed in distal intergenic regions and a reduced percentage in promoter regions (Figure 6C). KEGG pathway analysis demonstrated significant enrichment of pathways such as “NF-kappa B signaling,” “TNF signaling,” and “Necroptosis” in the control group, with treatment further enhancing pathways like “MAPK signaling” and “C-type lectin receptor signaling” (Figures 6D, E). Heatmap analysis displayed clustering patterns of upregulated and downregulated peaks, reflecting differential NF-κB binding between the two groups (Figure 6F). The average read density (RPKM) plots illustrated the chromatin signal intensities near these peaks, highlighting

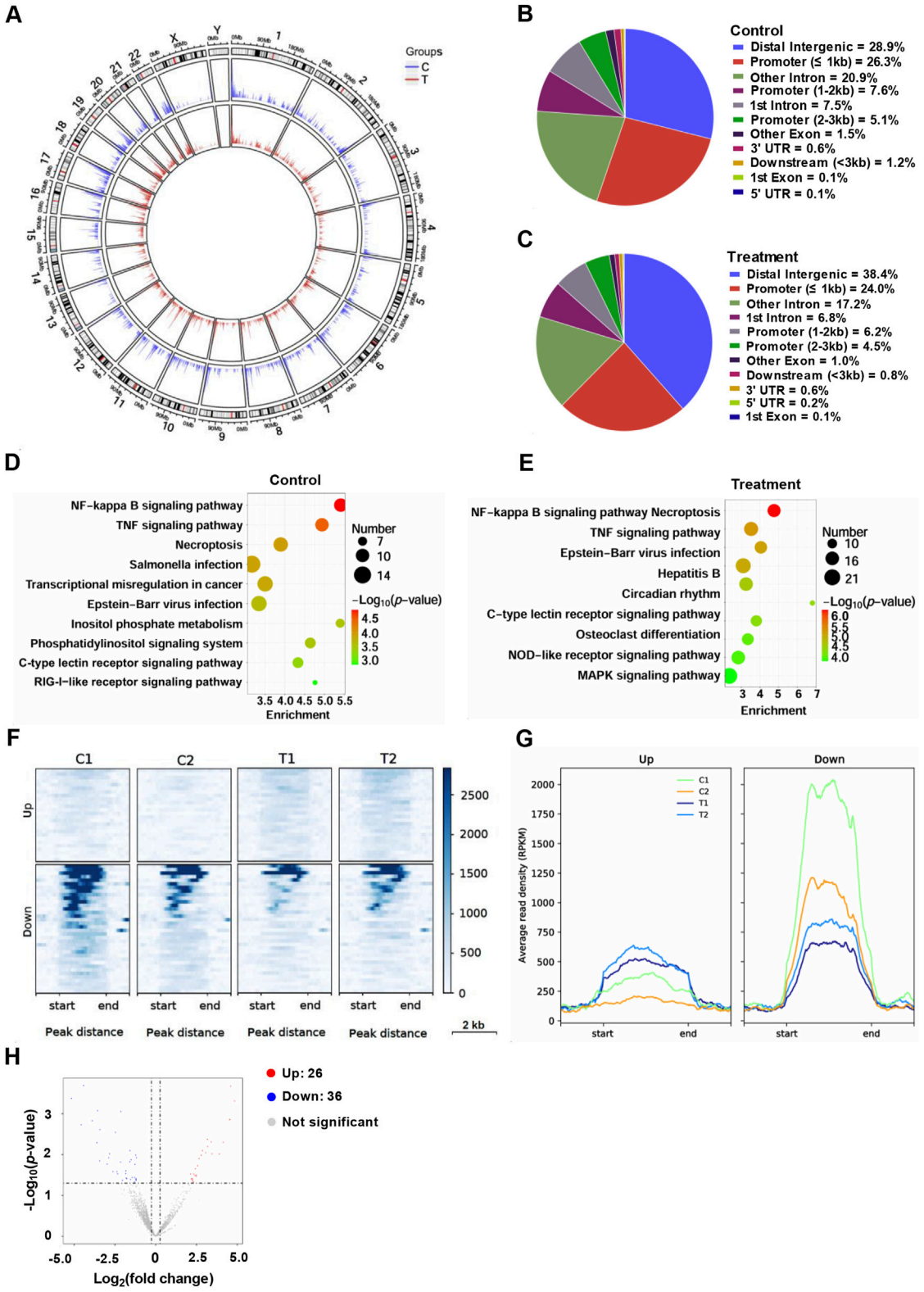


FIGURE 6
ChIP-seq analysis of NF-κB profiling in HCT116 cells treated with YG-22. **(A)** Circular plot visualizing genome-wide distribution of NF-κB peaks across different groups (control and treatment). The plot highlights the density of NF-κB binding across chromosomes. **(B)** Distribution of NF-κB peaks in various chromatin regions in control samples. The pie chart illustrates the percentage of peaks in promoters, introns, intergenic regions, and other genomic features. **(C)** Distribution of NF-κB peaks in various chromatin regions in YG-22-treated samples. The pie chart represents the proportion of peaks localized to distinct genomic regions after treatment. **(D)** KEGG pathway analysis of promoter-associated NF-κB peaks in control samples. The bubble plot shows enriched pathways, with bubble size indicating the number of peaks and color denoting statistical significance (*P*-value). **(E)** KEGG
(*Continued*)

FIGURE 6 (Continued)
pathway analysis of promoter-associated NF-κB peaks in YG-22-treated samples. Pathway enrichment is represented similarly to panel D, highlighting pathways associated with treatment-induced changes. **(F)** Heatmap analysis showing differential NF-κB peaks between control and YG-22-treated samples. The heatmap emphasizes the clustering patterns of upregulated and downregulated peaks. **(G)** Average read density (RPKM) of upregulated and downregulated NF-κB peaks in control and YG-22-treated samples. Line plots display chromatin signal intensities near the differential peaks. **(H)** Volcano plot highlighting significant differential NF-κB peaks between control and treated samples. Red dots represent upregulated peaks, blue dots indicate downregulated peaks, and gray dots correspond to non-significant changes.

significant differences between control and treated samples (Figure 6G). Finally, the volcano plot identified 26 upregulated and 36 downregulated NF-κB peaks, providing a comprehensive view of the impact of YG-22 on NF-κB activity (Figure 6H). These findings indicate that YG-22 treatment induces significant changes in NF-κB binding and pathway activation, emphasizing its potential as a modulator of NF-κB signaling.

Collectively, these data examined the peptide alterations of chemotherapy combined with the Jianpi formula in colorectal cancer patients, revealing significant changes in the peptidome and lnc-peptidome profiles, and screening the candidate bioactive peptide YG-22. Furthermore, YG-22 treatment in HCT116 cells demonstrated dose-dependent cytotoxicity, altered gene expression, metabolic reprogramming, chromatin accessibility, and significant modifications in histone and NF-κB binding, highlighting its potential as a therapeutic agent in colorectal cancer.

Discussion

This study aimed to identify differential bioactive peptides in patients treated with chemotherapy alone and those receiving chemotherapy combined with traditional TCM-Jianpi formula. By screening these peptides, we sought to investigate their potential therapeutic effects, using *in vitro* model to evaluate cytotoxicity in HCT116 cells. The most effective peptide was then subjected to multi-omics analyses to explore its underlying mechanisms of action.

Human plasma is a vital resource in clinical and biological research, serving as a reservoir for proteins secreted by various organs. It provides insights into a patient's physiological and pathological states and may contain biomarkers for disease detection and treatment response (Geyer et al., 2017; Xu et al., 2019). Plasma peptidomes has been used for screening novelty bioactive peptides or biomarker (Xu et al., 2019; Taguchi et al., 2021; Lu et al., 2022). In the present study, the results revealed a significant enhancement in the diversity and complexity of the plasma peptidome in patients treated with the combination therapy compared to chemotherapy alone. This suggests that the Jianpi formula not only complements chemotherapy but may also contribute to regulating key molecular pathways involved in tumor suppression, such as apoptosis and immune modulation. Among the candidate peptides identified, YG-22 exhibited the most potent cytotoxic effect, with clear dose- and time-dependent reductions in HCT116 cell viability. These findings underscore the therapeutic potential of differential peptides derived from the combined treatment approach.

Multi-omics technologies, which include transcriptomics (gene expression analysis), metabolomics (metabolic profiling), and epigenomics (chromatin and histone modification analysis), provide a comprehensive understanding of drug mechanisms (Zhao et al., 2024; Lou et al., 2022). The transcriptome primarily focuses on alterations in gene expression resulting from drug intervention, whereas metabolomics examines the impact of drug intervention on metabolites (Cui and Paules, 2010; Wilmes et al., 2013; Astarita et al., 2023; Lu et al., 2019). Chromatin accessibility analysis primarily examines the impact of drug action on chromatin accessibility and its potential influence on gene expression (Zhang et al., 2020), whereas the H3K4Me3 profiling analysis focuses on histone modifications in regions associated with active or inactive gene expression (Igolkina et al., 2019; Karlić et al., 2010). NF-κB binding can be analyzed using ChIP-seq by identifying the specific DNA regions where NF-κB transcription factors interact with the genome, providing insights into its regulatory roles in gene expression under various conditions (Mulero et al., 2019). Here, our multi-omics analysis provided valuable insights into the mechanisms of action of YG-22 in inhibiting CRC. Transcriptomic data revealed distinct gene expression changes, with enrichment in pathways related to lysosome-mediated degradation, cell adhesion, and apoptosis—processes pivotal in tumor progression and metastasis. Metabolomic profiling highlighted significant metabolic reprogramming in YG-22-treated cells, including disruptions in pathways essential for cell survival and proliferation. Furthermore, epigenomic analyses demonstrated notable alterations in chromatin accessibility and histone modifications, suggesting that YG-22 induces epigenetic reprogramming, which may enhance its antitumor effects.

These findings collectively suggest that combining TCM-Jianpi formula with chemotherapy may augment therapeutic outcomes by enriching the molecular and cellular responses to treatment. The Jianpi formula appears to amplify the generation of bioactive peptides, such as YG-22, which exhibit strong antitumor activity and target multiple pathways critical for cancer progression. This integrated approach provides a promising avenue for improving treatment efficacy. However, while this study offers valuable insights, the precise mechanisms by which the Jianpi formula enhances peptide generation and therapeutic efficacy require further investigation. Additionally, the clinical applicability and safety of the identified peptides require validation in larger clinical trials. Future studies should focus on exploring the molecular interactions of YG-22, validating its efficacy in animal models, and assessing the broader clinical implications of TCM-based combination therapies for CRC.

In conclusion, this study demonstrates the potential of TCM-Jianpi formula-derived bioactive peptides to improve colorectal cancer treatment outcomes. The identification and functional characterization of differential bioactive peptides (such as YG-22)

provide a foundation for developing innovative, peptide-based therapeutic strategies.

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 NCBI database (accession numbers: OMIX008933, GSE289006, GSE289007 and GSE289008).

Ethics statement

The studies involving humans were approved by the Ethics Committee Baoshan District Hospital of Integrated Traditional Chinese and Western Medicine of Shanghai, Shanghai University of Traditional Chinese Medicine (Approval Number: 202310). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JW: Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing–original draft, Data curation. LZ: Data curation, Formal Analysis, Investigation, Writing–original draft, Supervision, Visualization. YL: Data curation, Formal Analysis, Investigation, Writing–original draft, Methodology. MD: Investigation, Writing–review and editing. XW: Investigation, Writing–review and editing. BX: Investigation, Writing–review and editing. HC: Investigation, Writing–review and editing. LC: Investigation, Writing–review and editing. WC: Investigation, Writing–review and editing, Resources. BH: Investigation, Resources, Writing–review and editing, Conceptualization, Formal Analysis, Validation. JL: Conceptualization, Formal Analysis, Investigation, Writing–review and editing, Visualization, Writing–original draft. QS: Conceptualization, Formal Analysis, Investigation, Visualization, Writing–original draft, Writing–review and editing, Funding acquisition, Methodology, Project administration, Resources, Software, Validation.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the foundation of National Natural Science Foundation of China (Project No. 82104944); the 2023 Outstanding Young Expert Talent Training Award from the Shanghai Baoshan District Chinese and Western Medicine Combined Hospital (Project No. 2023BY002); and the project leader of the 3rd Shanghai Baoshan District Famous Traditional Chinese Medicine Inheritance Studio (Project No. BSMZYGZS-2024-12).

Acknowledgments

The authors thank the patients for their participation in the study, and the investigators for collecting and analyzing the sequencing data.

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2025.1560172/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Lnc-peptidome analysis of plasma samples from colorectal cancer patients treated with chemotherapy or chemotherapy combined with the Jianpi formula. **(A)** Total number of identified peptides, unique peptides, and proteins. The bar chart presents the overall counts of peptides and proteins detected in the plasma samples. **(B)** Distribution of proteins based on the number of peptides identified. The histogram illustrates the frequency of proteins with varying numbers of peptides detected. **(C)** Protein sequence coverage distribution across the identified peptides. The pie chart shows the proportion of proteins categorized by sequence coverage percentages. **(D)** Distribution of peptides by length. The bar chart highlights the number of peptides grouped according to their length in amino acids. **(E)** KEGG pathway analysis of proteins corresponding to the identified lnc-peptides. The bar chart categorizes these proteins into pathways related to metabolism, genetic and environmental information processing, cellular processes, organismal systems, and human diseases. The numbers indicate the count of proteins associated with each category. **(F)** Pearson correlation heatmap comparing plasma sample data from colorectal cancer patients in the chemotherapy group and the chemotherapy plus Jianpi formula group. The heatmap displays correlation coefficients between samples. **(G)** Heatmap of differentially expressed peptides between the two treatment groups. The heatmap emphasizes clustering patterns and relative expression levels of peptides, with red and blue indicating upregulated and downregulated peptides, respectively.

SUPPLEMENTARY FIGURE 2
Metabolomics analysis of HCT116 cells treated with YG-22 based on positive ion (POS) analysis. **(A)** Correlation analysis of metabolites detected in control and YG-22-treated samples based on POS mode. The heatmap shows pairwise correlation coefficients, with red indicating positive correlations and

green indicating negative correlations. **(B)** Radar plot displaying the distribution of selected metabolites detected in POS mode. The plot highlights the relative abundance of key metabolites across samples. **(C)** Heatmap of differentially expressed metabolites identified through POS mode analysis. Hierarchical clustering shows differences in metabolite expression levels between control (C1- C3) and treated (T1-T3) groups. **(D)** Volcano plot illustrating significant differential metabolites between control and YG-22-treated samples in POS mode. Red dots indicate upregulated metabolites, blue dots indicate downregulated metabolites, and gray dots represent metabolites with non-significant changes. **(E)** KEGG pathway analysis of differentially expressed metabolites, categorized by metabolic pathways and biological processes. The bubble plot represents pathway enrichment, where dot size reflects the number of involved metabolites and color indicates the level of statistical significance (*P*-value).

SUPPLEMENTARY FIGURE 3
Differences in chromatin accessibility across different samples. **(A)** Average read density (RPKM) analysis for gene body coverage across different samples (C1, C2, T1, and T2). Line plots show the chromatin accessibility signal along gene bodies, from transcription start site (TSS) to transcription end site (TES). **(B)** Pearson correlation analysis comparing chromatin accessibility between control (C1 and C2) and treatment (T1 and T2) samples. Scatter plots illustrate pairwise correlations, with Pearson correlation coefficients indicated for each comparison. **(C)** TSS analysis of chromatin accessibility across different samples. Combined line plots and heatmaps depict the density of chromatin accessibility signals near the TSS regions, emphasizing differences between control and treated samples. **(D)** Circular plot visualizing chromatin accessibility peak distribution across the genome for all samples (C1, C2, T1, and T2). The plot highlights genomic regions with significant chromatin accessibility changes. **(E)** Heatmap analysis of chromatin accessibility peaks across different samples. The upper panel shows the average read density (RPKM) near peaks, while the lower panel illustrates heatmaps of peak signals, emphasizing differential chromatin accessibility clustering patterns.

SUPPLEMENTARY FIGURE 4
Differences in H3K4Me3 profiling across different samples. **(A)** Heatmap analysis of ChIP quality for both immunoprecipitated (IP) and input samples. The heatmap displays pairwise correlation coefficients among samples, emphasizing the quality and consistency of ChIP-seq data. **(B)** Pearson correlation analysis comparing H3K4Me3 profiling between control and treatment samples for both IP and input datasets. The scatter plots represent pairwise correlations, with Pearson correlation coefficients noted. **(C)** Average read density (RPKM) for IP and input samples. Line plots depict the signal intensity distribution near peak centers, comparing control and treated samples. **(D)** Circular plot visualizing the genome-wide distribution of H3K4Me3 peaks across control and treatment groups. The plot highlights differences in peak density between groups. **(E)** Transcription start site (TSS) analysis for different samples. Combined line plots and heatmaps show H3K4Me3 signal density near TSS regions, illustrating differences in chromatin accessibility between control and treated samples for both IP and input datasets.

References

Adebayo, A. S., Agbaje, K., Adesina, S. K., and Olajubutu, O. (2023). Colorectal cancer: disease process, current treatment options, and future perspectives. *Pharmaceutics* 15, 2620. doi:10.3390/pharmaceutics15112620

Al Bitar, S., El-Sabban, M., Doughan, S., and Abou-Kheir, W. (2023). Molecular mechanisms targeting drug-resistance and metastasis in colorectal cancer: updates and beyond. *World J. Gastroenterology* 29, 1395–1426. doi:10.3748/wjg.v29.i9.1395

Alsakarneh, S., Jaber, F., Beran, A., Aldiabat, M., Abboud, Y., Hassan, N., et al. (2024). The national burden of colorectal cancer in the United States from 1990 to 2019. *Cancers* 16, 205. doi:10.3390/cancers16010205

Astarita, G., Kelly, R. S., and Lasky-Su, J. (2023). Metabolomics and lipidomics strategies in modern drug discovery and development. *Drug Discov. Today* 28, 103751. doi:10.1016/j.drudis.2023.103751

Chen, D., Zhao, J., and Cong, W. (2018). Chinese herbal medicines facilitate the control of chemotherapy-induced side effects in colorectal cancer: progress and perspective. *Front. Pharmacol.* 9, 1442. doi:10.3389/fphar.2018.01442

Chen, P., Ni, W., Xie, T., and Sui, X. (2019). Meta-analysis of 5-fluorouracil-based chemotherapy combined with traditional Chinese medicines for colorectal cancer treatment. *Integr. Cancer Ther.* 18, 1534735419828824. doi:10.1177/1534735419828824

Constantinou, V., and Constantinou, C. (2024). Focusing on colorectal cancer in young adults. *Mol. Clin. Oncol.* 20, 1–10.

Cui, H., Han, W., Zhang, J., Zhang, Z., and Su, X. (2019). Advances in the regulatory effects of bioactive peptides on metabolic signaling pathways in tumor cells. *J. Cancer* 10, 2425–2433. doi:10.7150/jca.31359

Cui, Y., and Paules, R. S. (2010). Use of transcriptomics in understanding mechanisms of drug-induced toxicity. *Pharmacogenomics* 11, 573–585. doi:10.2217/pgs.10.37

Feria, A., and Times, M. (2024). Effectiveness of standard treatment for stage 4 colorectal cancer: traditional management with surgery, radiation, and chemotherapy. *Clin. Colon Rectal Surg.* 37, 062–065. doi:10.1055/s-0043-1761420

Geyer, P. E., Holdt, L. M., Teupser, D., and Mann, M. (2017). Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* 13, 942. doi:10.15252/msb.20156297

He, S., Hao, L., Chen, Y., Gong, B., and Xu, X. (2025). Chinese herbal Jianpi Jiedu formula suppressed colorectal cancer growth *in vitro* and *in vivo* via modulating hypoxia-inducible factor 1 alpha-mediated fibroblasts activation. *J. Ethnopharmacol.* 337, 118753. doi:10.1016/j.jep.2024.118753

Igolkina, A. A., Zinkevich, A., Karandasheva, K. O., Popov, A. A., Selifanova, M. V., Nikolaeva, D., et al. (2019). H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 histone tags suggest distinct regulatory evolution of open and condensed chromatin landmarks. *Cells* 8, 1034. doi:10.3390/cells8091034

Jiang, H. Z., Jiang, Y. L., Yang, B., Long, F. X., Yang, Z., and Tang, D. X. (2023). Traditional Chinese medicines and capecitabine-based chemotherapy for colorectal cancer treatment: a meta-analysis. *Cancer Med.* 12, 236–255. doi:10.1002/cam4.4896

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci.* 107, 2926–2931. doi:10.1073/pnas.0909344107

Klimeck, L., Heisser, T., Hoffmeister, M., and Brenner, H. (2023). Colorectal cancer: a health and economic problem. *Best Pract. and Res. Clin. Gastroenterology* 66, 101839. doi:10.1016/j.bpg.2023.101839

Lin, X., Yang, X., Yang, Y., Zhang, H., and Huang, X. (2023). Research progress of traditional Chinese medicine as sensitizer in reversing chemoresistance of colorectal cancer. *Front. Oncol.* 13, 1132141. doi:10.3389/fonc.2023.1132141

Lou, Y., Shi, Q., Zhang, Y., Qi, Y., Zhang, W., Wang, H., et al. (2022). Multi-omics signatures identification for LUAD prognosis prediction model based on the integrative analysis of immune and hypoxia signals. *Front. Cell Dev. Biol.* 10, 840466. doi:10.3389/fcell.2022.840466

Lu, J., Shi, Q., Zhang, L., Wu, J., Lou, Y., Qian, J., et al. (2019). Integrated transcriptome analysis reveals KLK5 and L1CAM predict response to anlotinib in NSCLC at 3rd line. *Front. Oncol.* 9, 886. doi:10.3389/fonc.2019.00886

Lu, J., Zhang, W., Yu, K., Zhang, L., Lou, Y., Gu, P., et al. (2022). Screening anlotinib responders via blood-based proteomics in non-small cell lung cancer. *FASEB J.* 36, e22465. doi:10.1096/fj.202101658R

McCulloch, M., Ly, H., Broffman, M., See, C., Clemons, J., and Chang, R. (2016). Chinese herbal medicine and fluorouracil-based chemotherapy for colorectal cancer: a quality-adjusted meta-analysis of randomized controlled trials. *Integr. Cancer Ther.* 15, 285–307. doi:10.1177/1534735416638738

Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C., Laversanne, M., et al. (2023). Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut* 72, 338–344. doi:10.1136/gutjnl-2022-327736

Morris, V. K., Kennedy, E. B., Baxter, N. N., Benson III, A. B., Cercek, A., Cho, M., et al. (2023). Treatment of metastatic colorectal cancer: ASCO guideline. *J. Clin. Oncol.* 41, 678–700. doi:10.1200/JCO.22.01690

Mulero, M. C., Wang, V.Y.-F., Huxford, T., and Ghosh, G. (2019). Genome reading by the NF-κB transcription factors. *Nucleic Acids Res.* 47, 9967–9989. doi:10.1093/nar/gkz739

Peng, W., Zhang, S., Zhang, Z., Xu, P., Mao, D., Huang, S., et al. (2018). Jianpi Jiedu decoction, a traditional Chinese medicine formula, inhibits tumorigenesis, metastasis, and angiogenesis through the mTOR/HIF-1α/VEGF pathway. *J. Ethnopharmacol.* 224, 140–148. doi:10.1016/j.jep.2018.05.039

Quintal-Bojórquez, N., and Segura-Campos, M. R. (2021). Bioactive peptides as therapeutic adjuvants for cancer. *Nutr. Cancer* 73, 1309–1321. doi:10.1080/01635581.2020.1813316

Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., and Jemal, A. (2023). Colorectal cancer statistics, 2023. *CA A Cancer J. Clin.* 73, 233–254. doi:10.3322/caac.21772

Taguchi, T., Kodera, Y., Oba, K., Saito, T., Nakagawa, Y., Kawashima, Y., et al. (2021). Suprabasin-derived bioactive peptides identified by plasma peptidomics. *Sci. Rep.* 11, 1047. doi:10.1038/s41598-020-79353-4

Wang, C., Lu, Y., Chen, Z., Liu, X., Lin, H., Zhao, H., et al. (2012). Serum proteomic, peptidomic and metabolomic profiles in myasthenia gravis patients during treatment with Qiangji Jianli Fang. *Chin. Med.* 7, 16–18. doi:10.1186/1749-8546-7-16

Wilmes, A., Limonciel, A., Aschauer, L., Moenks, K., Bielow, C., Leonard, M. O., et al. (2013). Application of integrated transcriptomic, proteomic and metabolomic profiling for the delineation of mechanisms of drug induced cell stress. *J. proteomics* 79, 180–194. doi:10.1016/j.jprot.2012.11.022

Wu, Z., Fu, X., Jing, H., Huang, W., Li, X., Xiao, C., et al. (2024). Herbal medicine for the prevention of chemotherapy-induced nausea and vomiting in patients with advanced colorectal cancer: a prospective randomized controlled trial. *J. Ethnopharmacol.* 325, 117853. doi:10.1016/j.jep.2024.117853

Xi, Y., and Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Transl. Oncol.* 14, 101174. doi:10.1016/j.tranon.2021.101174

Xu, M., Deng, J., Xu, K., Zhu, T., Han, L., Yan, Y., et al. (2019). In-depth serum proteomics reveals biomarkers of psoriasis severity and response to traditional Chinese medicine. *Theranostics* 9, 2475–2488. doi:10.7150/thno.31144

Zhang, L., Lu, J., Liu, R., Hu, M., Zhao, Y., Tan, S., et al. (2020). Chromatin accessibility analysis reveals that TFAP2A promotes angiogenesis in acquired resistance to anlotinib in lung cancer cells. *Acta Pharmacol. Sin.* 41, 1357–1365. doi:10.1038/s41401-020-0421-7

Zhang, Y., Wang, C., Zhang, W., and Li, X. (2023). Bioactive peptides for anticancer therapies. *Biomater. Transl.* 4, 5–17. doi:10.12336/biomatertransl.2023.01.003

Zhao, M., Che, Y., Gao, Y., and Zhang, X. (2024). Application of multi-omics in the study of traditional Chinese medicine. *Front. Pharmacol.* 15, 1431862. doi:10.3389/fphar.2024.1431862

Zhou, W.-J., Wei, B., Cai, F.-F., Yang, M.-D., Chen, X.-L., Chen, Q.-L., et al. (2019). Therapeutic effect of Jianpi decoction combined with chemotherapy on postoperative treatment of colorectal cancer: a systematic review and meta-analysis. *World J. Traditional Chin. Med.* 5, 228–235. doi:10.4103/wjtcm.wjtcm_25_19

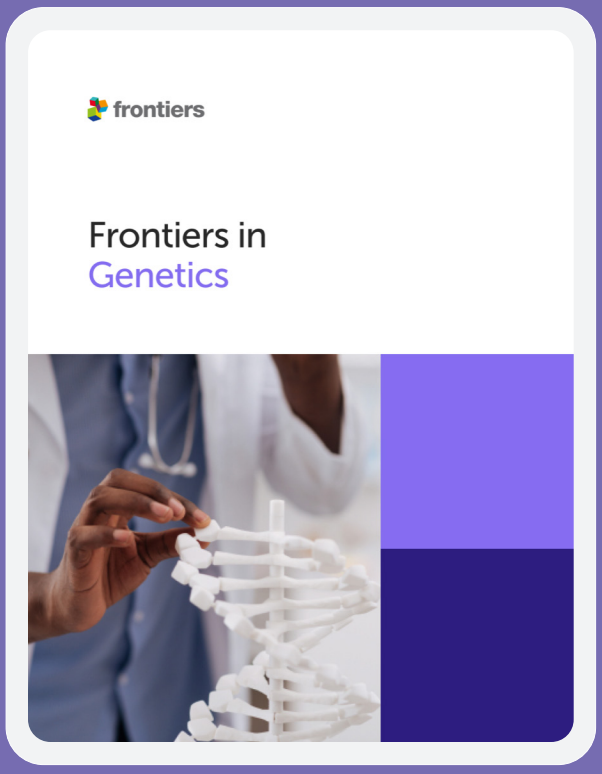
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact