Yu et al. *Zool. Res.* 2024, 45(3): 491−505
https://doi.org/10.24272/j.issn.2095-8137.2023.336

ZR | Zoological Research

Article Open Access

Combining genome-wide association study and transcriptome analysis to identify molecular markers and genetic basis of population-asynchronous ovarian development in *Coilia nasus*

Yue Yu[1], Shi-Ming Wan[1,2,*], Cheng-You Huang[3], Shuang-Meng Zhang[1], Ai-Li Sun[1], Jun-Qi Liu[1], Shun-Yao Li[1], Yong-Fu Zhu[4], Shu-Xin Gu[5], Ze-Xia Gao[1,2,*]

[1] *College of Fisheries*, *Hubei Hongshan Laboratory/Key Lab of Freshwater Animal Breeding*, *Ministry of Agriculture and Rural Affairs/Engineering Research Center of Green Development for Conventional Aquatic Biological Industry in the Yangtze River Economic Belt*, *Ministry of Education*, *Huazhong Agricultural University*, *Wuhan*, *Hubei* 430070, *China*

[2] *Laboratory for Marine Biology and Biotechnology*, *Qingdao Marine Science and Technology Center*, *Qingdao*, *Shandong* 266237, *China*

[3] *Wuhan Hengyuxin Technology Breeding Professional Cooperative*, *Wuhan*, *Hubei* 430117, *China*

[4] *Hubei Provincial Aquatic Science Research Institute*, *Wuhan*, *Hubei* 430208, *China*

[5] *Zhenjiang Jiangzhiyuan Fishery Technology Co.*, *Ltd.*, *Zhenjiang*, *Jiangsu* 212213, *China*

ABSTRACT

Coilia nasus, a migratory fish species found in the middle and lower reaches of the Yangtze River and along offshore areas of China, possesses considerable aquacultural and economic potential. However, the species faces challenges due to significant variation in the gonadal development rate among females, resulting in inconsistent ovarian maturation times at the population level, an extended reproductive period, and limitations on fish growth rate due to ovarian prematurity. In the present study, we combined genome-wide association study (GWAS) and comparative transcriptome analysis to investigate the potential single nucleotide polymorphisms (SNPs) and candidate genes associated with population-asynchronous ovarian development in *C. nasus*. Genotyping of the female population based on whole-genome resequencing yielded 2 120 695 high-quality SNPs, 39 of which were suggestively associated with ovarian development. Of note, a significant SNP peak on LG21 containing 30 suggestively associated SNPs was identified, with *cpne5a* determined as the causal gene of the peak. Therefore, single-marker and haplotype association analyses were performed on *cpne5a*, revealing four genetic markers (P<0.05) and seven haplotypes (r^2>0.9) significantly associated with the phenotype. Comparative transcriptome analysis of precociously and normally maturing individuals screened out 29 and 426 overlapping differentially expressed genes in the brain and ovary, respectively, between individuals of different body sizes. Integrating the GWAS and transcriptome analysis results, this study identified genes and pathways related to hypothalamic-pituitary-gonadal axis hormone secretion, extracellular matrix, angiogenesis, and gap junctions involved in population-asynchronous ovarian development. The insights gained from this study provide a basis for a deeper understanding of the molecular mechanisms underlying ovarian development in fish and may facilitate the genetic breeding of *C. nasus* strains exhibiting population-synchronous ovarian development in the future.

Keywords: *Coilia nasus*; GWAS; Transcriptome; Ovarian development; SNP

INTRODUCTION

Oocyte development directly affects fish spawning strategies. Specifically, based on the growth pattern of oocytes, ovarian development can be classified as synchronous, asynchronous, or group-synchronous (Wallace & Selman, 1981). In species with synchronous ovarian development, oocytes are recruited and mature collectively in a single batch, leading to the release of all mature oocytes over a brief period within the reproductive season. Their spawning typically

Received: 01 December 2023; Accepted: 25 December 2023; Online: 26 December 2023

Foundation items: This work was supported by the National Key R & D Program of China (2022YFD2400904), Key R&D Projects in Hubei Province (2022BBA008), Zhenjiang Jinshan Talents, and Wuhan Yangtze River Characteristic Fish Breeding and Domestication Project

*Corresponding authors, E-mail: wansm@mail.hzau.edu.cn; gaozx@mail.hzau.edu.cn

occurs once a year, as observed in coho salmon (*Oncorhynchus kisutch*) (Heins & Brown-Peterson, 2022). Conversely, ovaries showing asynchronous development contain several heterogeneous batches of different sized oocytes, enabling fish to produce newly matured oocytes shortly after the completion of a previous spawning event. Such species often demonstrate iteroparous batch spawning (Murua & Saborido-Rey, 2003). For group-synchronous ovarian development, the ovaries are characterized by the presence of at least two oocyte populations during the reproductive season: a group of small primary growth oocytes, and one or more groups of larger, secondary growth oocytes (Lubzens et al., 2010; Murua & Saborido-Rey, 2003). Fish with this ovarian pattern may engage in either total or batch spawning (Heins & Brown-Peterson, 2022).

The estuarine tapertail anchovy (*Coilia nasus*), an important anadromous fish species, inhabits the Yangtze River in China. Juveniles grow in the brackish waters of the Yangtze River estuary, while adults migrate upstream to the middle and lower reaches of the river in early spring each year (Huang et al., 2022). Spawning events occur gradually during migration as their ovaries mature. Previous studies have shown that *C. nasus* females exhibit group-synchronous ovarian development, with iteroparous spawning occurring once during the annual reproductive migration season (Duan et al., 2016; Zheng et al., 2012). The species displays a highly complex natural population composition, with significant variation in ovarian development rates among females within the same population. This phenomenon, known as the population-asynchronous ovarian development system, greatly extends the spawning period of the reproductive population (Xu et al., 2016), potentially benefiting the distribution of offspring, alleviating resource competition and predation pressure, and mitigating mass mortality due to sudden, extreme weather events. However, *C. nasus* is currently listed as an endangered species, attributed to the destruction of wild resources from water conservancy projects and overfishing, leading to a ban on commercial fishing (Wang et al., 2021). In response, artificial breeding and aquaculture efforts for *C. nasus* have developed rapidly over the past decade (Shi et al., 2015). However, the challenges posed by the population-asynchronous ovarian development and extended spawning period have hindered industrial-scale development of *C. nasus*.

Furthermore, growth retardation associated with premature ovarian development represents another limiting factor in the aquaculture industry. Gonadal maturation occurs at the cost of stored energy, with massive amounts of nutrients mobilized from somatic cells to support germ cell development, potentially compromising muscle nutritional composition (Manor et al., 2012). Fish often experience a decrease in growth rate following the first sexual maturation (Felip et al., 2008; Taranger et al., 2010). As such, premature gonadal development can adversely affect early growth, as energy allocation is early redirected towards gonadal development (Arndt, 2000; Jonsson & Jonsson, 2014). In cultured *C. nasus*, environmental deterioration and long-term captivity have led to severe degradation of genetic quality, evident premature ovarian maturation, and a marked trend towards population miniaturization. Notably, the onset of female sexual maturity has decreased from 2–3 years old to 1–2 years old (Liu et al., 2008). Thus, there is an urgent need to carry out genetic breeding programs targeting premature maturation in *C. nasus* to improve production performance.

Current research on *C. nasus* has primarily focused on natural resource distribution (Ma et al., 2020), migratory behavior (Yin et al., 2020), genetic diversity (Zhang et al., 2023), and stress regulation (Du et al., 2014; Gao et al., 2021), with limited studies investigating population-asynchronous ovarian development (Xu et al., 2016). The development of high-throughput sequencing has significantly contributed to the advancement of genetic breeding in aquatic animals. Concurrently, GWAS has emerged as an efficient tool for the identification of genetic markers and candidate genes associated with economic traits (Palaiokostas & Houston, 2018). The integration of GWAS and RNA-seq has facilitated explorations into various economic traits of importance to aquaculture, such as sex determination (Sun et al., 2023), growth rates (Wang et al., 2023), and hypoxia tolerance (Ding et al., 2023). However, the application of this combined analytical approach to gonadal development-related traits remains unexplored.

In the present study, we integrated GWAS with comparative transcriptome analysis to elucidate the molecular mechanisms underlying population-asynchronous ovarian development in *C. nasus*. Females exhibiting either precocious or normal ovarian maturation were genotyped using whole-genome resequencing and subjected to GWAS. Subsequently, single-marker and haplotype association analyses were performed to identify causal genes related to significant single nucleotide polymorphism (SNP) peaks. Additionally, to address the potential confounding effects of growth on transcriptome analysis of ovarian development, RNA-seq based on extreme growth phenotypic grouping was applied, then combined with GWAS to precisely identify candidate genes and pathways related to ovarian development. This research investigated the molecular mechanisms underlying population-asynchronous ovarian development in *C. nasus*, providing a strong theoretical foundation and genetic markers for future genetic selection of *C. nasus* strains with population-synchronized ovarian development.

MATERIALS AND METHODS

Ethics approval

This study was conducted in accordance with the “Guidelines for Experimental Animals” of the Ministry of Science and Technology (Beijing, China). The protocols for fish handling and sampling were approved by the Institutional Animal Care and Use Ethics Committee of Huazhong Agricultural University (No. HZAUFI-2023-0009). All individuals were anesthetized before sampling, using 100 mg/L MS-222.

Sample collection

All *C. nasus* samples were obtained from a farmed population at Zhenjiang Jiangzhiyuan Fishery Technology Co., Ltd. (Zhenjiang, China), all descendants of wild ancestors from the Yangtze River. A total of 500 two-year-old *C. nasus* fish with a female-to-male ratio of 1:1.5 were subjected to microflow stimulation for simulated ecological artificial reproduction in 2021. After the fertilized eggs hatched, approximately 6 000 offspring from the same batch were reared in a 2 000 m^2 pond. After one year of cultivation, a total of 252 one-year-old *C. nasus* fish were randomly collected in June 2022. Their fins were collected and preserved in 95% ethanol at −20°C for DNA extraction. Gonads from one side were fixed in Bouin’s

solution and transferred to 90% ethanol to determine phenotypic sex, with progressive dehydration, paraffin embedding, and sectioning into 6 μm slices for hematoxylin and eosin (H&E) staining, in accordance with routine histological procedures. Gonads on the other side and brain tissues were collected and rapidly frozen in liquid nitrogen and stored at −80°C for RNA extraction. Subsequently, all female *C. nasus* fish were classified into precociously maturing (PM) and normally maturing (NM) individuals based on the developmental stage of oocytes at the cellular level (Brown-Peterson et al., 2011). Females only containing primary growth oocytes were classified into the NM group, while those containing abundant vitellogenic oocytes were classified into the PM group. Growth-related traits, including body length and weight, were recorded for all fish. Significance analysis of growth traits was completed using independent samples *t*-test in IBM SPSS Statistics v.25.0.

Genomic DNA extraction, genotyping, and quality control

Individual DNA from whole-genome resequencing was extracted from fin clips using a Genomic DNA Extraction Kit (Cret-Biotech, China), and DNA concentration was quantified using Qubit v.4.0. The extracted DNA was sent to Wuhan Yingzi Gene Technology Co., Ltd. (China) for library construction using the DNBSEQ-T7 platform (MGI, China). The quality-controlled clean reads were then aligned to the reference genome of *C. nasus* (GenBank: GCA_007927625.1) using Sentieon software for genetic variation detection and genotyping (Xu et al., 2020). Variant calling was performed using GATK v.4.4.0.0 with the following filter criteria: Quality(QUAL)<30.0, QualByDepth(OD)<2.0, RMSMapping Quality(MQ)<40.0, FisherStrand(FS)>60.0, StrandOdds Ratio(SOR)>3.0, MQRankSum<−12.5, and ReadPos RankSum<−8.0 (McKenna et al., 2010). Plink v.1.90 was then used for the final screening, removing: (1) SNP sites with a minor allele frequency (MAF) below 5%; (2) individuals with a variant missing rate greater than 10% (mind>0.1) and SNP sites with a genotyping call rate below 10% (geno<0.1); and (3) SNP sites with a Hardy Weinberg equilibrium *P*-value of the chi-square test less than 1×10^{-5} (hwe<1×10^{-5}) (Purcell et al., 2007).

LD and population structure analyses

PopLDdecay v.3.41 was employed to calculate the LD and correlation coefficient (r^2) between each pair of SNPs and to generate LD decay plots (Zhang et al., 2019). LD decay distance was defined as half of the maximum r^2 value. Prior to genetic analysis, principal component analysis (PCA) was performed using plink v.1.90 based on the filtered genotype data to visualize the population structure of *C. nasus*. PHYLIP v.3.697 (Retief, 1999) was used to construct a phylogenetic tree based on the neighbor-joining (NJ) method.

GWAS

The phenotypic data were defined as a binary trait according to the stage of ovarian development at the histological level, with 77 PM and 24 NM individuals used for association analysis (denoted as 0 and 1, respectively). General linear (GLM) and mixed linear models (MLM) were employed using GEMMA v.0.98.3 (Zhou & Stephens, 2012) for GWAS of the phenotype and genotype data. Principal components (PCs) and a kinship matrix (K) were incorporated into the models to control for population stratification. The GWAS model and run number of PCs were determined based on the quantile-quantile (QQ) plot and genomic inflation factor (λ). Associated λ values close to 1.00 were considered as optimal mode (Ai et al., 2023). The equation used in the present study is as follows:

$$Y = X\beta + G\gamma + Zu + \varepsilon \quad (1)$$

where Y is the observed vector of phenotypes for all samples; X is the design matrix comprising genetic variation and other explanatory variables; β is the vector of regression coefficients; G is the covariate matrix consisting of three or five PCs as fixed effects; γ is the corresponding coefficient vector; Z is the matrix for random effects (kinship matrix of MLM); u is the random effect vector for kinship; and ε is the residual vector. To avoid over-conservativeness of Bonferroni correction, the GWAS threshold was adjusted using the SimpleM package to calculate the effective number of independent tests (MeffG) (Gao et al., 2008; Gao, 2011). A total of 2 120 695 genotype SNP markers were compressed to 752 310 MeffG. The suggestive and significant thresholds for the final genetic analysis were defined as 1/MeffG and 0.05/MeffG, respectively. QQ plots, Manhattan plots, and SNP density plots were generated using the CMplot package in RStudio v.4.2.2 (Yin et al., 2021).

LD haplotype association analysis

LDBlockShow v.1.40 was utilized to generate locuszoom plots and LD heatmaps for the 200 kb upstream and downstream regions of the lead SNP located within significant peaks. This approach aimed to identify genetic SNP markers having strong LD with the lead SNP (Dong et al., 2021). Single-marker association analysis was performed on the SNPs within the coding sequence (CDS) region of the genes with a high density of strong LD genetic markers. The SNPs within the CDS region of the causal gene, showing significant correlation with the phenotype in single-marker association analysis (P<0.05), as well as the suggestively associated SNPs on the causal gene from GWAS analysis, were collectively input into Haploview v.4.2 to identify LD blocks. The LD block regions were estimated using a strict standard of r^2>0.9 (Renner et al., 2009; van Es et al., 2009). Finally, LD haplotype association analysis was performed using SHEsisPlus (Li et al., 2009).

RNA extraction and transcriptome analysis

Due to the presence of exaggerated growth variations within the sample populations of PM and NM, an RNA-seq grouping approach was designed based on body size to minimize the confounding effects of irrelevant variables such as growth. Three extremely large individuals (average 16.53±0.69 cm body length and 14.91±2.54 g body weight in the NM group; 17.17±0.78 cm body length and 18.09±1.39 g body weight in the PM group) and three extremely small individuals (average 11.77±0.95 cm body length and 5.13±0.80 g body weight in the NM group; 10.27±0.41 cm body length and 3.77±0.53 g body weight in PM group) were selected from the NM and PM groups, respectively (Supplementary Table S1). Transcriptomic analysis was performed using three strategies: based on large individuals (three LNM vs. three LPM), small individuals (three SNM vs. three SPM), and all samples (six NM vs. six PM). The final results were determined at the overlapping intersection of the three sets of analyses. The ovaries and brain tissues from these 12 individuals were preserved in liquid nitrogen and sent to Wuhan Yingzi Gene Technology Co., Ltd. (China) for sequencing and construction

of RNA libraries.

RNA sequencing was conducted using the DNBSEQ-T7 sequencing platform (MGI, China) to construct paired-end libraries. The raw data were entered into Trimmomatic v.0.39 for quality control. Adapter contamination, reads with a high proportion of N bases (≥10%), and low-quality sequences (bases with a quality score of Q≤10 comprising more than 50% of the total bases) were adequately screened out. Following data filtering, the clean reads were aligned to the *C. nasus* reference genome (GenBank: GCA_007927625.1). Clean reads are quantified for gene expression using reads per kilobase per million reads (RPKM) as a measurement metric. DESeq2 v.1.40.1 was used to determine the number of differentially expressed genes (DEGs) between PM and NM samples (Love et al., 2014). Genes with $|\log_2$ fold-change$|>1$ and $P<0.05$ were considered significant DEGs (Ding et al., 2023). The DEGs obtained from RNA-seq of large-size individuals, small-size individuals, and all individuals were compared to identify overlapping intersections. These overlapping DEGs represented the most conservative differences in *C. nasus* ovarian development and were input into ClusterProfiler v.4.6.2 in RStudio for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses.

Validation of RNA-seq data by quantitative real-time polymerase chain reaction (qPCR)

To validate the reliability of RNA-seq, 10 genes with high differential expression levels from the overlapping DEGs in the brain and ovary were selected for validation through qPCR, using the primers listed in Supplementary Table S2. qPCR was performed to validate gene expression using SYBR® PreMix DimerEraser™ (TaKaRa, Japan) with the QuantStudio™ 6 Flex qRT-PCR system (Applied Biosystems, USA). Three biological replicates were conducted to confirm the reliability of the results, and *C. nasus* beta-actin was used as a reference for internal standardization. The thermal cycling program of PCR was as follows: activation at 95°C for 10 min, followed by 40 cycles at 95°C for 15 s, 60°C for 30 s, and 72°C for 30 s. The relative expression levels of genes were calculated using the $2^{-\Delta\Delta Ct}$ method (Livak & Schmittgen, 2001). Independent samples *t*-test was performed using IBM SPSS Statistics v.25.0 to determine the significance of gene expression differences between groups, with statistical significance set to $P<0.05$.

Candidate gene identification and functional annotation

Based on the physical location of all significant genetic markers in the annotation file of the *C. nasus* reference genome, bedtools v.2.30.0 (Quinlan & Hall, 2010) was employed to retrieve all genes within a 100 kb upstream and downstream region (200 kb for lead SNP within significant peaks). The coding sequences of these potential candidate genes were then inputted into the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/) for manual identification using the Basic Local Alignment Search Tool (BLAST). All potential candidate genes were thoroughly compared with the RNA-seq analysis results. Only genes classified as DEGs in the RNA-seq analysis for at least one body size category, whether large or small, were selected as the final candidate genes. These genes were also subjected to KEGG enrichment analysis using the above methods described for RNA-seq.

RESULTS

Descriptive statistics of phenotype

At the anatomical level, NM and PM ovaries showed marked morphological differences despite similar body sizes (Figure 1A). The NM ovaries were smaller in size, closely adhered to the upper abdominal wall, and displayed a pale pink or translucent appearance. In contrast, the PM ovaries contained well-defined eggs and nearly filled the entire body cavity.

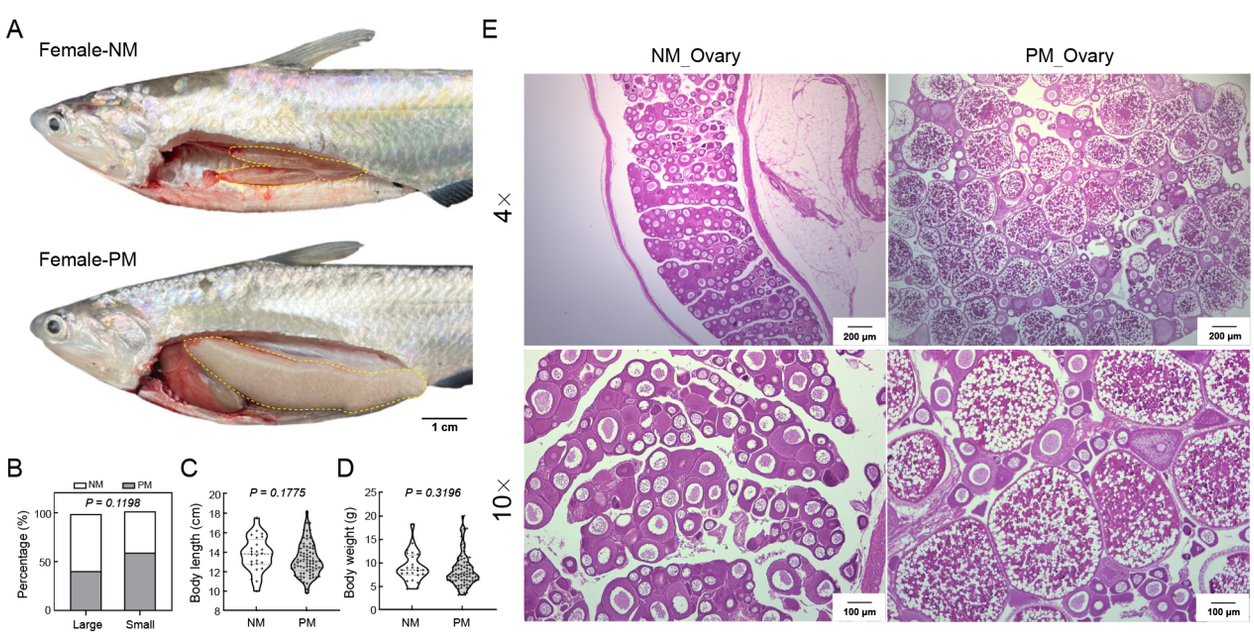

Figure 1 Phenotypic statistics of growth and ovarian development

A: Gonad morphology of NM and PM ovaries at anatomical level. B: Proportions of NM and PM individuals in large body-size and small body-size groups. C, D: Discrete comparison of growth phenotypes between NM and PM individuals. E: Gonadal histology of NM and PM ovaries. NM, normally maturing individual; PM, precociously maturing individual.

Based on gonadal histology, a total of 105 *C. nasus* were identified as females among 252 fish. Of these, 77 individuals contained ovaries filled with vitellogenic oocytes and were classified as the PM group, while 24 individuals contained ovaries filled with primary growth oocytes and were classified as the NM group. Four samples were in an intermediate ambiguous stage and were excluded from final genetic association analysis (Figure 1E). Average body lengths and weights were 13.78±1.79 cm and 9.17±3.22 g (mean±standard deviation (SD), *n*=24) in the NM group and 13.23±1.68 cm and 8.39±3.39 g (mean±SD, *n*=77) in the PM group, respectively.

The *C. nasus* growth phenotypes showed substantial population variation, particularly concerning body weight, with a maximum of 20.05 g and a minimum of 3.21 g. The coefficients of variation (CV) for body length and weight in the experimental population reached 12.8% and 39.0%, respectively (Table 1). No significant differences were observed when comparing the dispersion of growth parameters between the two groups (Figure 1C, D). Subsequently, body size was differentiated based on average body length of 101 female individuals (13.37 cm), with those with a body length longer than 13.37 cm classified as large individuals and those with a body length shorter considered as small individuals. Figure 1B shows that NM individuals accounted for the majority in fish in the large-size group, while PM individuals were more predominant in the small-size group, although there was no significant difference at the one-year-old stage (chi-square test, *P*=0.1198).

Genotyping and population structure analyses

In total, 105 *C. nasus* samples were genotyped from whole-genome resequencing, yielding 1 109 Gb of clean reads. The mapping rate of each individual was 93.96%±0.32% (mean±SD, *n*=105), and the sequencing depth ranged from 6× to 13×. After thorough quality control, a total of 2 120 695 high-quality SNPs from 98 individuals were used for the final GWAS. LD decay within the population was found to be remarkably rapid. When the average distance between adjacent SNPs was 11.5 kb, LD decayed to half of the maximum value (r^2=0.269), indicating a high level of genetic diversity within the population (Lu et al., 2011) (Figure 2A). The size of the reference genome used for analysis was about 812.2 Mb. Theoretically, to ensure the accuracy of association analysis, a minimum of 70 626 uniformly distributed genetic markers (812.2 Mb/11.5 kb) is required (Ding et al., 2023). The 2 120 695 SNPs used far exceeded this number and were sufficient to precisely pinpoint significant signals. The SNP density plot shows the number of SNPs per 1 Mb on each linkage group (LG), indicating relatively even distribution on the 24 LGs, with an average marker density of 383 bp/SNP (Figure 2B). The highest number of SNPs was found on LG14 (132 994 SNPs), and the lowest was on LG4 (48 815 SNPs). PCA indicated that the population structure showed slight stratification within the experimental population (Figure 2C). The NJ-based phylogenetic tree showed no obvious branching clusters (Supplementary Figure S1). These results demonstrate no significant genetic differences between the PM and NM groups in genetic information. However, to minimize the impact of population stratification on genetic association results, we incorporated PCs and a kinship matrix to correct the following analysis models.

GWAS and LD haplotype analysis

To assess the most effective model for correcting group stratification, four analytical models were evaluated: GLM with three PCs, GLM with five PCs, MLM with three PCs and a kinship matrix, MLM with five PCs and a kinship matrix. The corresponding λ values of these models were 0.997, 1.104, 0.956, and 0.997, respectively (Figure 2D). The λ values for GLM with three PCs and MLM with five PCs were the same and closest to 1. Thus, GLM with three PCs was selected for subsequent analysis as it exhibited more significant and abundant positive signals. According to the Manhattan plot (Figure 3A), a total of 39 SNPs were identified with suggestive correlations to ovarian development (Table 2). Of these, 32 SNPs were located within the intronic regions of six genes, including fibroblast growth factor 12b (*fgf12b*), WD repeat and FYVE domain containing 3 (*wdfy3*), decorin (*dcn*), two copine V homologous genes (called *cpne5a* and *cpne5b* in this study), and one uncharacterized gene, while seven SNPs were located in the intergenic regions. Notably, a significant SNP peak was detected on LG21, comprising 30 suggestively associated SNPs. The Locuszoom plot, displaying the linkage relationship between the lead SNP of a significant peak (rs.21-35691138) and all genetic markers 200 kb upstream and downstream (Figure 3B), indicated that the SNPs with a high LD relationship to rs.21-35691138 were all located around two genes, *cpne5a* and *cpne5b* (D'>0.8). The SNPs located in the CDS region of these two genes were selected for single-marker association analysis. Four out of the seven SNPs located in the CDS region of the *cpne5a* gene showed significant associations with ovarian development (Table 3), whereas no SNPs were identified in the CDS region of *cpne5b* (Table 2). Thus, *cpne5a* was identified as the causal gene for the significant SNP peak.

All significant markers located on the *cpne5a* gene, including suggestively associated SNPs from GWAS analysis and significant SNPs from single-marker analysis, were collected and subjected to LD haplotype analysis. Seven significant haplotypes were identified (Table 4; Figure 3C), including Hap.Ov1 (rs.21-3562278, rs.21-35622426), Hap.Ov2 (rs.21-35646100, rs.21-35646108), Hap.Ov3 (rs.21-35652249, rs.21-35656228, rs.21-35658231, rs.21-35659819, rs.21-35664850, rs.21-35664851), Hap.Ov4 (rs.21-35680954, rs.21-35682653, rs.21-35684130, rs.21-35686438), Hap.Ov5 (rs.21-35691138, rs.21-35697505, rs.21-35697513, rs.21-35701504), Hap.Ov6 (rs.21-35707111, rs.21-35707919, rs.21-

Table 1 Growth phenotype statistics for *Coilia nasus*

Group	Body length (cm)				Body eight (g)			
	Max	Min	mean±SD	CV (%)	Max	Min	mean±SD	CV (%)
NM	17.50	10.00	13.78±1.79	13.0	18.30	4.48	9.17±3.22	35.1
PM	18.20	9.80	13.24±1.68	12.7	20.05	3.21	8.39±3.39	40.4
Total	18.20	9.80	13.37±1.71	12.8	20.05	3.21	8.58±3.35	39.0

NM, normally maturing individual; PM, precociously maturing individual; Min, minimum; Max, maximum; SD, standard deviation; CV, coefficients of variation.

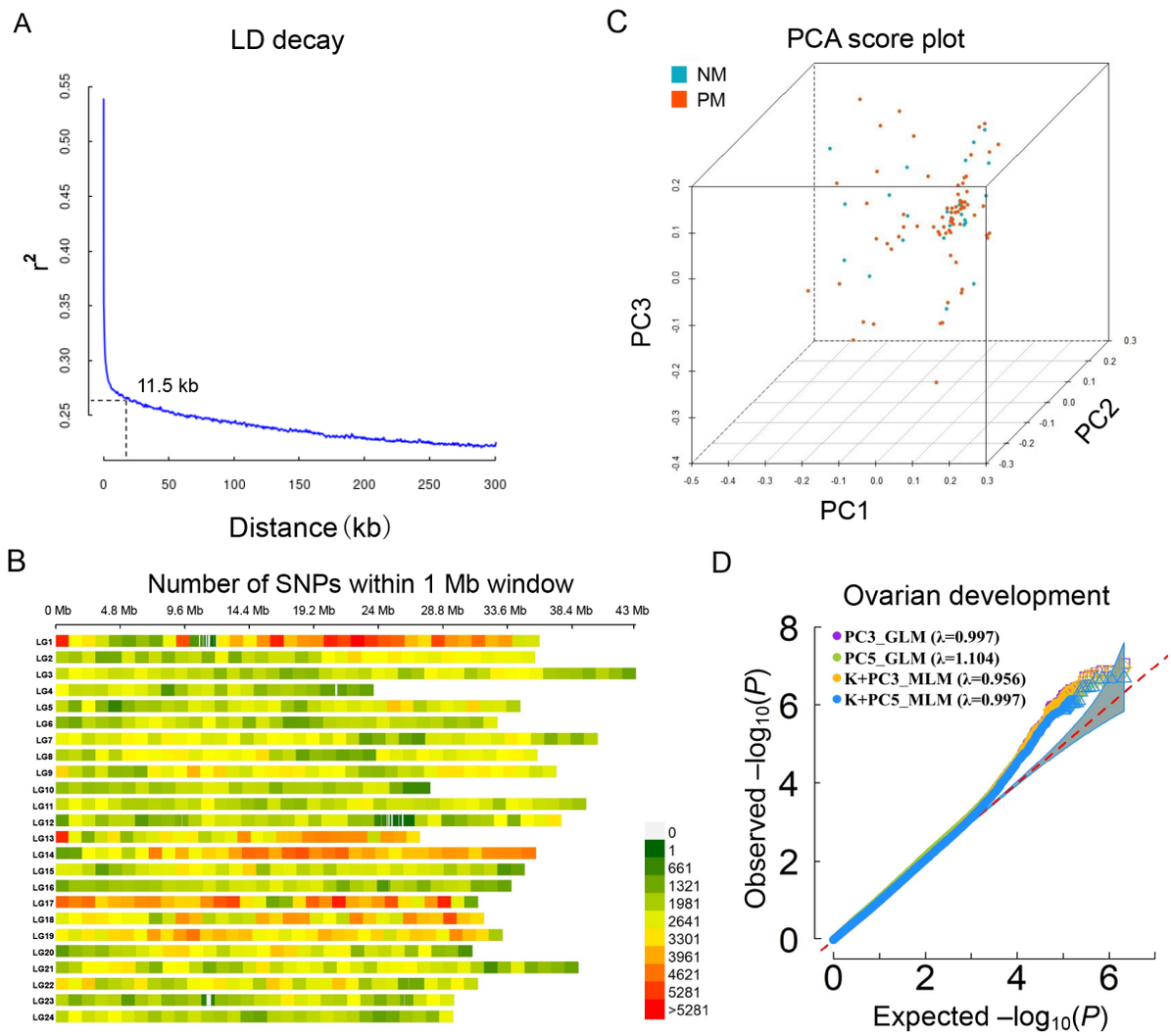

Figure 2 Genetic background investigation and model selection

A: LD decay plots of experimental individuals. X-axis represents physical location. Y-axis represents LD value (r^2). B: Identification of high-quality SNPs and their distribution across 24 linkage groups (LGs) of *C. nasus*. Gradient colors from green to red indicate increasing SNP density within 1 Mb interval. C: Population structure of *Coilia nasus* revealed by PCA. Red and blue dots are individuals in precocious maturing (PM) and normally maturing (NM) groups, respectively. D: Q-Q plot for ovarian development. The four analysis models included: (1) GLM with three PCs, (2) GLM with five PCs, (3) MLM with three PCs and a kinship matrix, and (4) MLM with five PCs and a kinship matrix. PC, principal component. K, kinship matrix.

35708107, rs.21-35708121), and Hap.Ov7 (rs.21-35708583, rs.21-35708584, rs.21-35708649). The distribution of the LD haplotypes differed significantly between the NM and PM groups (Figure 3D). Notably, LD haplotype association analysis demonstrated that the seven haplotypes were all significantly correlated with ovarian development and were thus identified as key candidate markers to guide future breeding programs (Table 4).

Annotation of potential candidate genes regulating ovarian development following GWAS

Considering the rapid LD decay rate, we extended the candidate gene annotation region to 100 kb upstream and downstream of each significant SNP, while the annotation region of the lead SNP of significant peak was extended to 200 kb upstream and downstream. Ultimately, 47 potential candidate genes were functionally annotated, with potential involvement in various biological processes in ovarian maturation, including cell proliferation, meiosis, apoptosis, autophagy, and hypothalamic-pituitary-gonadal (HPG) axis hormone secretion (Supplementary Table S3). These genes were used for subsequent cross-analysis with the transcriptome results.

Analysis of DEGs

An extreme body size-based RNA-seq grouping was designed to minimize the confounding effects of growth on transcriptomic analysis. Six individuals with extreme body size (three large and three small) were selected in the NM and PM groups, respectively. Supplementary Table S1 shows the growth phenotypes of each individual. Notably, substantial differences in growth were observed, especially in body weight, with large-size individuals reaching 4–5 times of small-size individuals.

Transcriptomic analysis based on ovarian development was performed on the 12 individuals (six NM vs. six PM), and 235 (62 up-regulated and 173 down-regulated) and 1 444 (463 up-regulated and 981 down-regulated) DEGs were obtained in the brain and ovary, respectively (Figure 4A, C). Subsequently, transcriptomic analysis was performed on large individuals (three LNM vs. three LPN) and small individuals (three SNM vs. three SPN), yielding 198 and 1 235 DEGs in

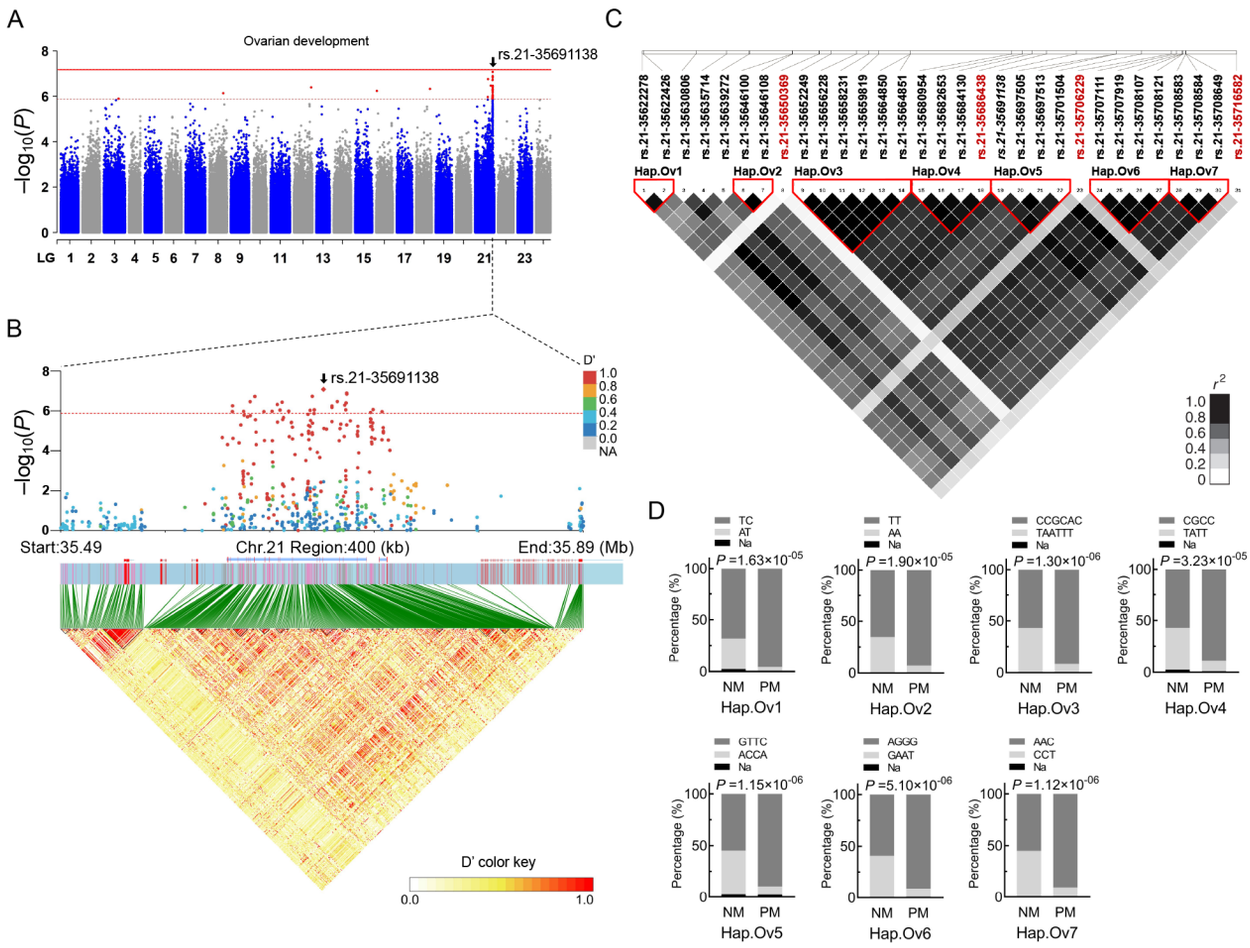

Figure 3 GWAS of ovarian development and LD association analysis of *cpne5a* in *Coilia nasus*

A: Manhattan plot for GWAS of ovarian development. X-axis represents LG numbers. Y-axis represents $-\log_{10}(P)$ for each SNP. Red dotted line indicates suggestive threshold, red line indicates significant threshold. B: Locuszoom plot of significant SNP peak. Dot color indicates LD strength with lead SNP, rs.21-35691138. C: LD plot of *cpne5a* gene. Black font indicates suggestively associated SNPs in GWAS analysis, red font indicates significantly associated SNPs on CDS region in single marker association analysis. Intensity of black square increases with degree of linkage and r^2 value. D: Proportion of seven haplotypes in NM and PM individuals. NM, normally maturing individual; PM, precociously maturing individual.

the brain and ovary of large individuals, and 798 and 1 245 DEGs in the brain and ovary of small individuals, respectively (Figure 4B, D). Overlapping the RNA-seq results of different body sizes and comparing them with the initial RNA-seq results of all individuals resulted in the identification of 29 (three up-regulated and 26 down-regulated) and 426 (49 up-regulated and 377 down-regulated) highly conserved DEGs in the brain and ovary, respectively (Figure 4B, D; Supplementary Tables S4, S5). These genes, potentially representing the phenomenon of population-asynchronous ovarian development in *C. nasus*, were then subjected to functional enrichment analysis. Among the DEGs exhibiting the most significant change in expression levels in the ovary and brain, several are posited to play key roles in ovarian development and follicle maturation, including two gap junction proteins (*gja11*, *gja13.2*), retinol dehydrogenase 10b (*rdh10b*), cytochrome P450 family 19 subfamily A polypeptide 1a (*cyp19a1a*), phospholipid transfer protein (*pltp*), solute carrier family 6 member 11a (*slc6a11a*), sizzled (*szl*), LIM homeobox 8 (*lhx8*), and natriuretic peptide B (*nppb*).

A total of 455 DEGs were significantly enriched in 330 GO terms. Among the top 20 significant terms (Figure 5A), “sprouting angiogenesis”, “extracellular matrix (ECM)”, “angiogenesis”, “blood vessel morphogenesis”, and “ovarian follicle development” may be directly related to population-asynchronous ovarian development. In addition, among the 15 significantly enriched KEGG pathways, several were associated with ovarian development, including “ECM-receptor interaction”, “GnRH signaling pathway”, and “steroid biosynthesis” (Figure 5B).

qPCR validation results

Based on the overlapping RNA-seq results, five genes were selected for qPCR validation among the DEGs identified in the brain and ovary tissues, respectively. The qPCR results showed that the NM individuals exhibited significant down-regulation in *foxg1*, *dlx2a*, *lhx8*, *nppb*, *gja11*, *gja13.2*, *rdh11b*, and *cyp19a1a* compared to the PM individuals, while *unc13a* and *tmem59l* were significantly up-regulated (Figure 6). These results are consistent with the expression patterns observed from RNA-seq. Additionally, the differential fold changes between NM and PM individuals for qPCR and RNA-seq results were $\log_2$ normalized and subjected to correlation analysis. The results suggested an extremely significant correlation between them (R^2=0.884, P=0.001), indicating high reliability of the transcriptomic data.

Comprehensive GWAS and RNA-seq analyses

All potential candidate genes predicted through GWAS were

Table 2 SNPs significantly associated with ovarian development in *Coilia nasus*

SNP ID	LG	Position	Allele	MAF (%)	Beta	*P*-value	Location
rs.3-29957384	3	29957384	T/C	46.9	0.30	1.27e–06	Intronic (*fgf12b*)
rs.8-28521866	8	28521866	A/G	8.4	0.58	7.32e–07	Intronic (*wdfy3*)
rs.12-35213305	12	35213305	T/A	13.3	0.45	4.05e–07	Intergenic
rs.16-1482918	16	1482918	G/A	8.6	0.47	5.84e–07	Intronic (*dcn*)
rs.18-28290664	18	28290664	G/T	11.1	0.48	4.72e–07	Intergenic
rs.21-25913866	21	25913866	T/A	13.2	0.45	1.08e–06	Intergenic
rs.21-26311407	21	26311407	C/T	9.2	0.53	1.74e–07	Intergenic
rs.21-32018434	21	32018434	A/G	8.5	0.54	3.38e–07	Intronic (*uncharacterized*)
rs.21-35133061	21	35133061	G/C	7.1	0.59	1.26e–06	Intergenic
rs.21-35622278	21	35622278	A/T	11	0.44	1.08e–06	Intronic (*cpne5a*)
rs.21-35622426	21	35622426	T/C	10.3	0.46	5.62e–07	Intronic (*cpne5a*)
rs.21-35630806	21	35630806	A/G	18	0.37	1.06e–06	Intronic (*cpne5a*)
rs.21-35635714	21	35635714	A/T	23.2	0.34	3.49e–07	Intronic (*cpne5a*)
rs.21-35639272	21	35639272	C/T	17.2	0.44	1.92e–07	Intronic (*cpne5a*)
rs.21-35646100	21	35646100	A/T	14.1	0.42	1.09e–06	Intronic (*cpne5a*)
rs.21-35646108	21	35646108	A/T	14.1	0.42	1.09e–06	Intronic (*cpne5a*)
rs.21-35652249	21	35652249	T/C	18	0.37	1.05e–06	Intronic (*cpne5a*)
rs.21-35656228	21	35656228	A/C	18.2	0.38	4.82e–07	Intronic (*cpne5a*)
rs.21-35658231	21	35658231	A/G	17	0.38	9.22e–07	Intronic (*cpne5a*)
rs.21-35659819	21	35659819	T/C	17.7	0.39	3.70e–07	Intronic (*cpne5a*)
rs.21-35664850	21	35664850	T/A	18	0.37	1.04e–06	Intronic (*cpne5a*)
rs.21-35664851	21	35664851	T/C	18	0.37	1.04e–06	Intronic (*cpne5a*)
rs.21-35680954	21	35680954	T/C	17.3	0.36	9.50e–07	Intronic (*cpne5a*)
rs.21-35682653	21	35682653	A/G	18.5	0.38	1.89e–07	Intronic (*cpne5a*)
rs.21-35684130	21	35684130	T/C	17.6	0.37	8.96e–07	Intronic (*cpne5a*)
rs.21-35691138	21	35691138	A/G	15.6	0.41	8.47e–08	Intronic (*cpne5a*)
rs.21-35697505	21	35697505	C/T	17	0.38	6.81e–07	Intronic (*cpne5a*)
rs.21-35697513	21	35697513	C/T	17.9	0.39	5.71e–07	Intronic (*cpne5a*)
rs.21-35701504	21	35701504	A/C	16.3	0.40	3.45e–07	Intronic (*cpne5a*)
rs.21-35707111	21	35707111	G/A	16.5	0.40	5.18e–07	Intronic (*cpne5a*)
rs.21-35707919	21	35707919	A/G	17	0.39	8.97e–07	Intronic (*cpne5a*)
rs.21-35708107	21	35708107	A/G	16.3	0.39	4.06e–07	Intronic (*cpne5a*)
rs.21-35708121	21	35708121	T/G	16.3	0.39	4.05e–07	Intronic (*cpne5a*)
rs.21-35708583	21	35708583	C/A	18.5	0.37	1.30e–07	Intronic (*cpne5a*)
rs.21-35708584	21	35708584	C/A	18.5	0.37	1.30e–07	Intronic (*cpne5a*)
rs.21-35708649	21	35708649	T/C	17.8	0.37	1.46e–07	Intronic (*cpne5a*)
rs.21-35726606	21	35726606	G/A	19.7	0.36	1.21e–06	Intergenic
rs.21-35728574	21	35728574	T/C	20.7	0.37	8.67e–07	Intergenic
rs.21-35735318	21	35735318	C/A	21.8	0.32	1.10e–06	Intronic (*cpne5b*)

SNP ID, combination of chromosome number and physical location of SNP on reference genome; LG, linkage group; Allele, alternative/reference; MAF, minor allele frequency; Beta, effect size of each SNP; SNP ID is shown as LG-position.

compared with the DEGs identified from RNA-seq analysis, with any gene that matched with the RNA-seq findings for at least one body size designated as a final candidate gene. Among the 47 potential candidate genes, 12 exhibited differential expression in the ovaries or brains of NM and PM individuals, with five showing expression differences only in large-size or small-size individuals (Table 5; Figure 7A). Four KEGG pathways were significantly enriched in these final candidate genes, among which the gap junction and TGF-beta signaling pathways are implicated in ovarian maturation (Figure 7B).

DISCUSSION

Despite its high commercial value, *C. nasus* faces challenges in aquaculture development due to its population-asynchronous ovarian development. This condition, widespread in both natural and cultured populations, results in dispersed spawning times among females and longer reproductive periods across the population, greatly hindering large-scale aquacultural development. Furthermore, genetic deterioration and long-term captivity have led to an earlier age of sexual maturity in *C. nasus*, which not only exacerbates the degree of ovarian asynchrony within populations but also affects early growth (Liu et al., 2008). The intricate link between gonadal development and growth is well-documented. Teleost fish rely on endogenous nutrition to meet the increasing energy demands during gonadal maturation and spawning activities, and dynamically regulate the balance between gonadal development and growth by mobilizing ketone body stores (Manor et al., 2012; Medford & Mackay, 1978; Tveranger, 1985). Sexual maturation in certain fish is associated with reduced food intake (Rowe & Thorpe, 1990;

Table 3 Single-marker association analysis of SNPs in CDS region of *cpne5a* gene

SNP ID	RAF	ALT	Chi2	*P* value	OR (95% CI)
rs.21-35619045	A	G	0.016	0.897	0.946 (0.412–2.173)
rs.21-35639111	C	T	1.354	0.244	0.416 (0.091–1.9)
rs.21-35650369	C	T	4.021	**0.044**	0.298 (0.086–1.033)
rs.21-35686438	C	T	24.567	**7.18E-07**	6.333 (2.892–13.867)
rs.21-35706229	C	T	7.96	**0.004**	2.586 (1.321–5.059)
rs.21-35716582	T	C	6.654	**0.009**	2.367 (1.219–4.596)
rs.21-35716611	G	C	2.382	0.122	0.563 (0.27–1.173)

Bold font indicates significantly associated SNP. SNP ID, combination of chromosome number and physical location of SNP on reference genome; RAF, reference allele; ALT, alternative allele; Chi2, Chi-square value; OR, odds ratio; CI, confidence interval.

Table 4 Haplotype association analysis of loci on *cpne5a* gene

Hap ID	Haplotype	Number (frequency)		Chi2	*P*-value	OR (95% CI)
		PM	NM			
Hap.Ov1	TC	30 (0.681)	130 (0.955)	15.519	2.09e–04	0.23 (0.107–0.495)
	AT	13 (0.295)	6 (0.044)	21.958	2.37e–05	8.79 (3.12–24.762)
Hap.Ov2	TT	30 (0.652)	128 (0.927)	13.342	5.99e–04	0.26 (0.122–0.551)
	AA	16 (0.347)	10 (0.072)	22.25	1.52e–05	6.9 (2.865–16.613)
Hap.Ov3	CCGCAC	25 (0.568)	117 (0.914)	13.208	6.57e–04	0.287 (0.144–0.574)
	TAATTT	19 (0.431)	11 (0.085)	28.901	7.82e–07	8.159 (3.509–18.972)
Hap.Ov4	CGCC	24 (0.571)	105 (0.889)	7.067	0.013	0.409 (0.21–0.798)
	TATT	17 (0.404)	12 (0.101)	21.44	1.90e–05	6.215 (2.694–14.333)
Hap.Ov5	GTTC	23 (0.547)	120 (0.895)	20.206	2.25e–05	0.214 (0.106–0.432)
	ACCA	18 (0.428)	11 (0.082)	25.992	2.79e–06	7.472 (3.201–17.444)
Hap.Ov6	AGGG	26 (0.59)	109 (0.908)	6.417	0.018	0.422 (0.215–0.83)
	GAAT	18 (0.409)	10 (0.083)	27.976	1.27e–06	8.28 (3.475–19.724)
Hap.Ov7	AAC	22 (0.55)	123 (0.904)	26.16	1.13e–06	0.171 (0.084–0.35)
	CCT	18 (0.45)	12 (0.088)	24.153	5.79e–06	6.8 (2.963–15.603)

NM, normally maturing individual; PM, precociously maturing individual; Chi2, Chi-square value; OR, odds ratio; CI, confidence interval.

Simpson et al., 1996), slower growth rates (Lundqvist, 1980; Saunders et al., 1982; Tveranger, 1985), decreased muscle protein content (Aussanasuwannakul et al., 2011; Salem et al., 2006), and depleted energy reserves (Arndt, 2000). Despite these findings, our study revealed no direct correlation between ovarian maturity and growth in *C. nasus*, as evidenced by the lack of significant differences in growth phenotype dispersion between the PM and NM groups. However, the observed distribution of NM and PM individuals across the different size populations suggested a potential interaction between sexual maturity and growth in *C. nasus*. We speculate that the growth differences between NM and PM individuals may become more pronounced as females approach sexual maturity. These observations underscore the pressing need for genetic breeding aimed at ovarian development to mitigate the prolonged population spawning period and possible growth inhibition in *C. nasus*.

In the present study, we conducted the first GWAS on population-asynchronous ovarian development in aquatic animals. In total, 39 suggestively associated SNPs were identified, although none exceeded the threshold for significant association, possibly due to the limited sample size (Wu et al., 2022). Notably, a significant SNP peak containing 30 suggestively associated SNPs was identified on LG21. Analysis using Locuszoom plot identified *cpne5a* as the possible causal gene of this significant peak, as most of the suggestively associated SNPs and SNPs with strong LD linkage relationships to the lead SNP were located around *cpne5a*. The *cpne5* gene (orthologous to *cpne5a*) belongs to the copine gene family and encodes C2 structural domain-containing, Ca^{2+}-dependent, phospholipid-binding proteins, which guide ligands through biofilms according to the binding characteristics of ligands to phospholipids (Damer et al., 2007). This gene is also reported to play a crucial role in the central nervous system (Ding et al., 2008, 2017), and is associated with alcohol dependence and obesity in Caucasians (Wang et al., 2015). Although there is no current research demonstrating the direct function of *cpne5a* in ovarian development, recent GWAS analysis of pre- and post-pubertal heifers detected a significant correlation between the *cpne5* gene and phenotypic traits, with RNA-seq analysis showing differential expression of *cpne5* in the endometrium (Cánovas et al., 2014). Subsequent studies have also identified several SNPs on *cpne5* associated conservatively with heifer puberty across five cattle breeds (Dias et al., 2017). These findings suggest that *cpne5a* may play a role in female sexual maturation and oocyte development. Therefore, haplotype association analysis based on *cpne5a* was performed, identifying seven significant haplotypes associated with ovarian development in *C. nasus*, thus providing valuable markers for future genetic breeding efforts.

Considering the observed variation in growth phenotypes among *C. nasus* in both NM and PM individuals, an extreme body size-based grouping approach was adopted for transcriptomic analysis of population-asynchronous ovarian development in *C. nasus*, aimed at minimizing the confounding effect of growth on analytical outcomes. Comparison analysis of potential candidate genes located within the annotation intervals of suggestively associated SNPs and DEGs from transcriptomic analyses led to the

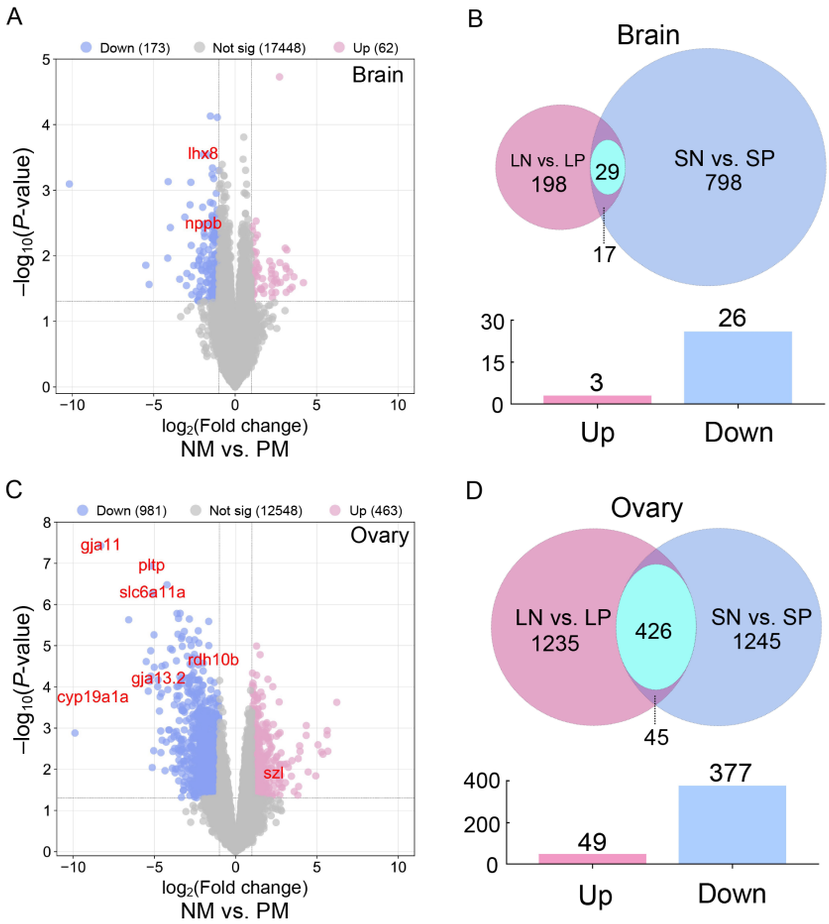

Figure 4 Transcriptomic analysis based on ovarian development in *C. nasus*

A, C: Volcano plots of DEGs in brain and ovary based on RNA-seq of all individuals. Pink and blue dots indicate significantly up-regulated and down-regulated genes, respectively ($|\log_2 FC|\geq 1$ and $P<0.05$). Gray dots indicate genes not significantly differentially expressed between NM and PM groups. B, D: Venn diagrams of DEGs obtained from RNA-seq based on large and small body-size groups. Overlapping intersection is compared with the DEGs from RNA-seq based on all individuals, with resulting blue region representing final conserved DEGs. Bar chart shows number of up-regulated and down-regulated DEGs. LN, large body size normally maturing individual; LP, large body size normally maturing individual; SN, small body size normally maturing individual; SP, small body size precociously maturing individual.

identification of 12 final candidate genes. The process of ovarian development is complex and strictly regulated. By integrating the findings from both GWAS and transcriptomic analyses, this study delineated the molecular mechanisms underlying population-asynchronous ovarian development in *C. nasus*, including HPG axis hormone secretion, ECM interactions, angiogenesis, and gap junction functioning.

Dynamic balance and timely secretion of HPG axis hormones are essential for normal function and fertility of the female reproductive system. Three important regulatory factors in the HPG axis, including *cyp19a1a*, follicle stimulating hormone receptor (*fshr*), and luteinizing hormone/choriogonadotropin receptor (*lhcgr*), were identified based on transcriptomic analysis (Martyniuk et al., 2013; Ramos-Júdez et al., 2022) and showed significant enrichment in “ovarian follicle development” (GO) and “steroid biosynthesis pathway” (KEGG). Knockdown of sulfiredoxin 1 (*srxn1*), identified as an overlapping candidate gene in our study, would reduce the sensitivity and initiating activity of the glycoprotein hormone subunit mRNAs of *Lhb* and *Fshb* in response to gonadotropin-releasing hormone (GnRH) stimulation (Kim et al., 2019). Notably, these genes exhibited lower expression in NM individuals compared to PM individuals, indicating enhanced hormonal interactions within the HPG axis during the vitellogenesis phase of ovarian development compared to the primary growth phase. The ECM constitutes a complex mixture of various molecules that influence cellular behaviors during ovarian development, including migration, division, differentiation, cell death, and cell anchorage (Hay, 1991; Irving-Rodgers & Rodgers, 2005). Based on GO and KEGG functional enrichment analyses, terms related to the ECM, such as “Extracellular matrix” and “External encapsulating structure”, as well as the “ECM-receptor interaction” pathway, were significantly enriched, suggesting the importance of ECM in ovarian follicular development in *C. nasus*. Our final candidate genes included two typical ECM proteins, *dcn* and lumican (*lum*), which are members of the small leucine-rich proteoglycan (SLRP) gene family. Previous studies have identified several SLRPs that interact with fibrillar collagen, a major component of the ovarian follicle layer in teleosts (Nagahama & Yamashita, 2008), and modulate its assembly dynamics (Irving-Rodgers & Rodgers, 2005). Both *dcn* and *lum* are sensitive to human chorionic gonadotropin (hGC) stimulation (Kedem et al., 2022; Peng et al., 2015). Overexpression of *dcn* has been shown to inhibit the proliferation of ovarian cells in Chinese hamsters

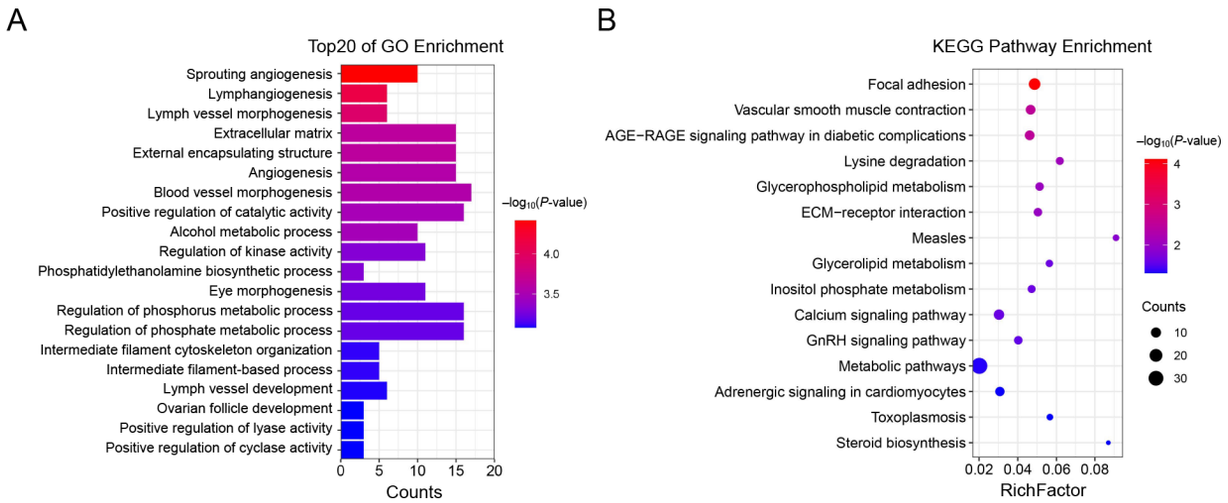

Figure 5 Functional enrichment analysis of overlapping DEGs

A: Top 20 significantly enriched GO terms in 455 conserved DEGs. B: Top 15 significantly enriched KEGG pathways in 455 conserved DEGs.

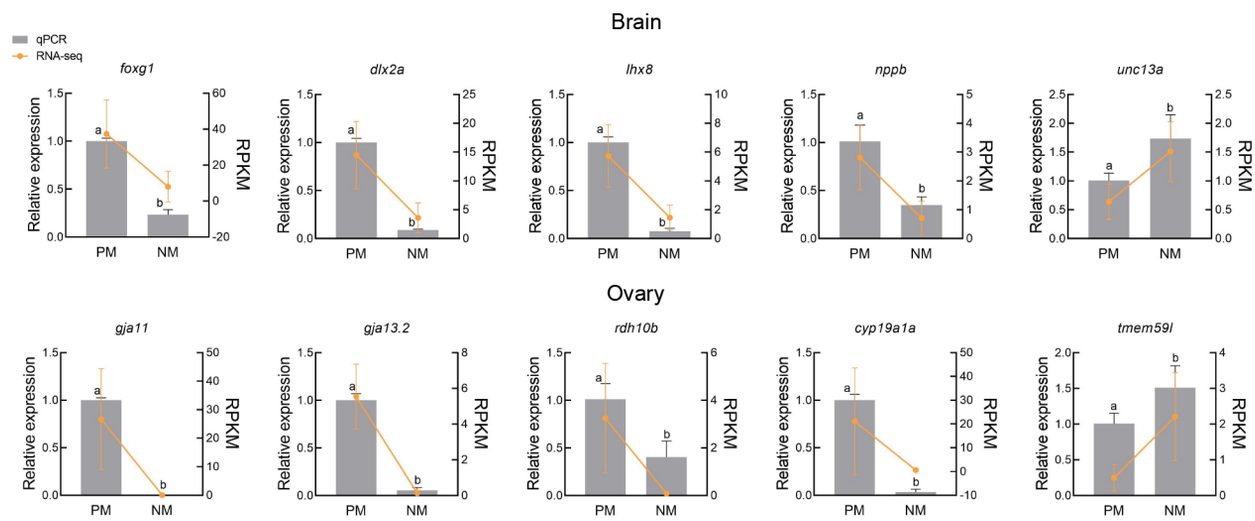

Figure 6 qPCR validation of 10 overlapping DEGs from RNA-seq

Gray column represents relative expression, corresponding to the y-axis on the left. Orange line represents actual expression level measured in RNA-seq, corresponding to the y-axis on the right. Values are mean±SD of the measured values. Significance of qPCR results between PM and NM individuals is indicated using letters. NM, normally maturing individual; PM, precociously maturing individual.

Table 5 Final candidate genes associated with ovarian development in *Coilia nasus*

SNP ID	Candidate gene	LG	Start	End	Log_2FC	*P* value	Description	Tissue	Group
rs.3-29957384	*il1rap*	3	29873857	29878206	1.164334	0.0344064	Interleukin-1 receptor accessory protein	Ovary	All
	il1rap	3	29908272	29936257	1.2734838	0.0378355	Interleukin-1 receptor accessory protein	Ovary	All
rs.8-28721866	*wdfy3*	8	28549826	28582793	1.1265539	0.031702	WD repeat and FYVE domain containing 3	Brain	Small
rs.12-35213305	*tmem198b*	12	35233409	35236894	1.3908257	0.047486	Transmembrane protein 198b	Ovary	All
	tuba1a	12	35118270	35138574	1.1303117	0.0428253	Tubulin alpha chain	Ovary	Small
rs.16-1482918	*lum*	16	1518399	1519729	-3.011686	6.10E-05	Lumican	Ovary	All
	dcn	16	1482217	1503407	-1.613559	0.0050068	Decorin	Ovary	Small
	epyc	16	1547695	1563608	-1.794637	0.0142083	Epiphycan	Brain	All
rs.21-26311407	*bsna*	21	26210330	26260627	-1.250373	0.0033617	Bassoon presynaptic cytomatrix protein a	Ovary	Large
rs.21-32088434	*srxn1*	21	32131172	32132570	-2.010689	0.0004854	Sulfiredoxin-1	Ovary	Large
	uncharacterized gene	21	32010067	32044683	1.6457006	0.00017	Uncharacterized protein	Ovary	All
rs.21-35133061	*adhfe1*	21	35222491	35245287	1.5221476	0.0156807	Alcohol dehydrogenase iron containing 1	Ovary	All

SNP ID, combination of chromosome number and physical location of SNP on reference genome; LG, linkage group; Start and End, physical location of start and end of gene in the reference genome. Group, RNA-seq based on large body size, small body size, and all individuals.

(Yamaguchi & Ruoslahti, 1988) and induce apoptosis and growth arrest of granulosa cells in goats (Peng et al., 2016). In addition, *lum* is considered a potential molecular marker for bovine oocyte maturation (Bunel et al., 2015; Mamo et al., 2011). During teleost vitellogenesis, the thecal cell layer of the ovary is filled with an abundance of capillaries (Nagahama &

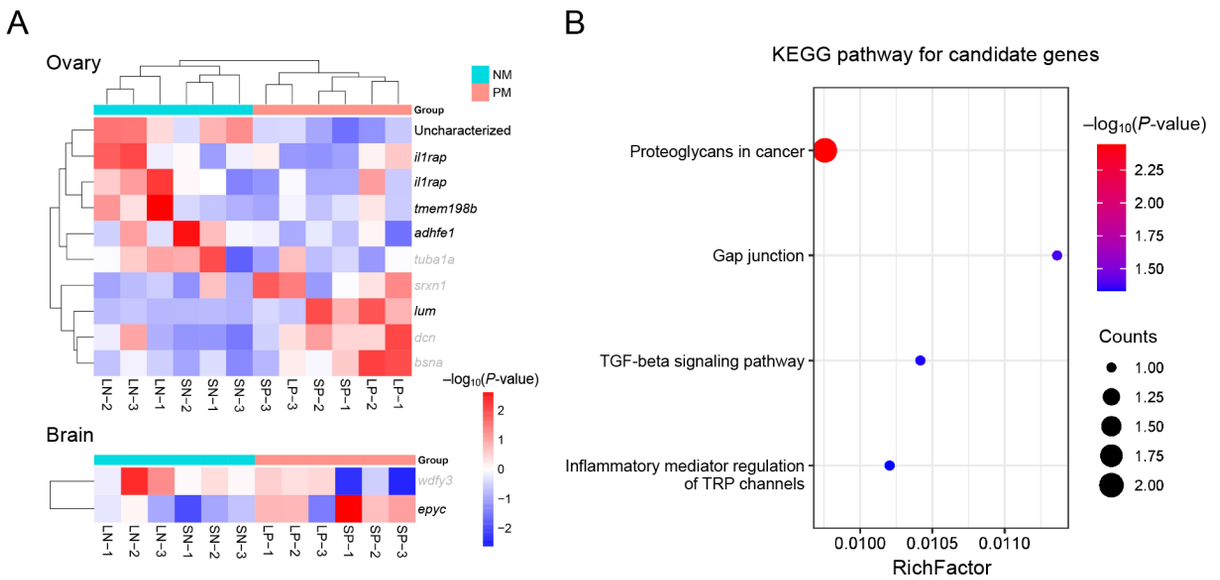

Figure 7 Final candidate genes and functional enrichment analysis

A: Heatmap analysis of final candidate genes in ovary and brain of NM and PM groups. Genes in red are up-regulated, genes in blue are down-regulated. Black font indicates differential expression in all body shapes, gray font indicates differential expression in only one body type. B: Enriched KEGG pathways in final candidate genes. LN, large body size normally maturing individual; LP, large body size normally maturing individual; SN, small body size normally maturing individual; SP, small body size precociously maturing individual.

Yamashita, 2008), facilitating local vascular dynamics associated with endocrine functions (Hazzard & Stouffer, 2000). The significantly enriched GO terms identified based on RNA-seq functional enrichment analysis included several terms related to angiogenesis, namely “sprouting angiogenesis”, “angiogenesis”, and “blood vessel morphogenesis”. The GWAS results identified a suggestively associated SNP on LG3 located within the intronic region of *fgf12b*, a known angiogenesis regulator (Phan et al., 2021; Woo et al., 2022). Notably, ECM proteins *dcn* and *lum* have also been reported as regulatory factors in angiogenesis (Järveläinen et al., 2006; Nikitovic et al., 2014). Gap junction proteins exert a wide range of functions in ovarian processes, including folliculogenesis, meiosis arrest, apoptosis, and steroidogenesis (Gershon et al., 2008). Following RNA-seq analysis of ovarian tissue, two gap junction proteins, gap junction protein alpha 11 (*gja11*) and gap junction protein alpha 13.2 (*gja13.2*), were among the top five DEGs identified. The *gja11* gene is widely expressed in granulosa cells surrounding stage II oocytes in zebrafish (Liu et al., 2022), while the expression of *gja13.2*, which responds to hCG stimulation in the ovaries, is consistent with oocyte competence in Atlantic croaker (Chang et al., 2000). In our results, *gja11* and *gja13.2* were highly expressed in PM individuals, highlighting the importance of gap junctions in the process of vitellogenesis in *C. nasus*.

Ovarian follicular development is a complex and comprehensive regulatory process. In addition to the above, other factors may also affect ovarian development in *C. nasus* via their intrinsic functions and potential pathways. The transmembrane protein 198b (*tmem198b*) and alcohol dehydrogenase iron containing 1 (*adhfe1*) were identified as overlapping genes in the GWAS and RNA-seq results. The *tmem198b* gene can regulate the positive regulation of the Wnt signal pathway by specifically activating low-density lipoprotein receptor-related protein 6 (Liang et al., 2011). The Wnt signaling pathway plays an important role in ovarian folliculogenesis and steroidogenesis (Gifford, 2015; Lapointe & Boerboom, 2011). The *adhfe1* gene, which can affect cell proliferation through self-hypermethylation, was highly expressed in NM individuals, suggesting a potential role in promoting germ cell replication during the early stages of ovarian growth. Several candidate genes, such as *tuba1a* and *dcn*, can induce apoptosis and may also affect the iteration of oocytes through this pathway.

CONCLUSIONS

In this study, we integrated GWAS and transcriptome analysis to identify SNPs regulating population-asynchronous ovarian development in *C. nasus* and to clarify the underlying molecular mechanisms. Based on GWAS, we identified 39 suggestively associated SNPs and a significant peak on LG21, with the *cpne5a* gene identified as a causal factor. From this gene, seven significant haplotype blocks related to the phenotype were obtained. RNA-seq analysis based on extreme body-size grouping revealed 455 conserved DEGs representing population-asynchronous ovarian development in *C. nasus*. Combining the GWAS and RNA-seq results, we identified several genes and biological processes, including HPG axis hormone secretion, ECM, angiogenesis, and gap junction functioning, involved in population-asynchronous ovarian development. These insights provide a theoretical basis for understanding the molecular mechanisms underlying ovarian development and follicular maturation in fish, offering valuable guidance for the genetic selection of *C. nasus* strains optimized for population-synchronized ovarian development.

DATA AVAILABILITY

The *C. nasus* RNA-seq and whole-genome resequencing data generated in this study were submitted to National Center for Biotechnology Information (NCBI SRA; https://www.ncbi.nlm.nih.gov/sra; accession number PRJNA1073042, PRJNA1070325); Genome Sequence Archive (GSA; https://ngdc.cncb.ac.cn/gsa/; accession number PRJCA022839, PRJCA022890) and Science Data Bank (SDB; https://www.scidb.cn/en; Data DOI: 10.57760/sciencedb.15311).

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Y.Y., S.W., and Z.G. conceived and designed the study; S.Z., A.L., J.L., S.L., and S.G. collected the samples and extracted the DNA and RNA; Y.Y. analyzed the sequencing results and performed GWAS analysis; Y.Y. prepared the data, conducted data analysis, designed the charts and tables, and wrote the manuscript. S.W., C.H., Y.Z., and Z.G. provided revisions to the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

We would like to thank the National Center for Biotechnology Information and European Bioinformatics Institute for technical support in data service.

REFERENCES

Ai CH, Zhu ZX, Huang DD, et al. 2023. Identification of SNPs and candidate genes associated with early growth in orange-spotted grouper (*Epinephelus coioides*) by a genome-wide association study. Aquaculture, 565: 739129.

Arndt SKA. 2000. Influence of sexual maturity on feeding, growth and energy stores of wild Atlantic salmon parr. Journal of Fish Biology, 57(3): 589–596.

Aussanasuwannakul A, Kenney PB, Weber GM, et al. 2011. Effect of sexual maturation on growth, fillet composition, and texture of female rainbow trout (*Oncorhynchus mykiss*) on a high nutritional plane. Aquaculture, 317(1-4): 79–88.

Brown-Peterson NJ, Wyanski DM, Saborido-Rey F, et al. 2011. A standardized terminology for describing reproductive development in fishes. Marine and Coastal Fisheries, 3(1): 52–70.

Bunel A, Jorssen EP, Merckx E, et al. 2015. Individual bovine *in vitro* embryo production and cumulus cell transcriptomic analysis to distinguish cumulus-oocyte complexes with high or low developmental potential. Theriogenology, 83(2): 228–237.

Cánovas A, Reverter A, DeAtley KL, et al. 2014. Multi-tissue omics analyses reveal molecular regulatory networks for puberty in composite beef cattle. PLoS One, 9(7): e102551.

Chang XT, Patiño R, Yoshizaki G, et al. 2000. Hormonal regulation and cellular distribution of connexin 32.2 and connexin 32.7 RNAs in the ovary of Atlantic croaker. General and Comparative Endocrinology, 120(2): 146–156.

Damer CK, Bayeva M, Kim PS, et al. 2007. Copine A is required for cytokinesis, contractile vacuole function, and development in *Dictyostelium*. *Eukaryotic Cell*, **6**(3): 430–442.

Dias MM, Cánovas A, Mantilla-Rojas C, et al. 2017. SNP detection using RNA-sequences of candidate genes associated with puberty in cattle. Genetics and Molecular Research, 16(1): gmr16019522.

Ding J, Gao ZY, Wang JY, et al. 2023. Genome-wide association and transcriptome analysis provide the SNPs and molecular insights into the hypoxia tolerance in large yellow croaker (*Larimichthys crocea*). Aquaculture, 573: 739547.

Ding XF, Jin YB, Wu Y, et al. 2008. Localization and cellular distribution of *CPNE5* in embryonic mouse brain. Brain Research, 1224: 20–28.

Ding XF, Wang HJ, Qian L, et al. 2017. Cpne5 is involved in regulating rodent anxiety level. CNS Neuroscience & Therapeutics, 23(3): 266–268.

Dong SS, He WM, Ji JJ, et al. 2021. LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Briefings in Bioinformatics, 22(4): bbaa227.

Du FK, Xu GC, Nie ZJ, et al. 2014. Transcriptome analysis gene expression in the liver of *Coilia nasus* during the stress response. BMC Genomics, 15(1): 558.

Duan JR, Fang DA, Zhang MY, et al. 2016. Changes of gonadotropin-releasing hormone receptor 2 during the anadromous spawning migration in *Coilia nasus*. *BMC Developmental Biology*, **16**(1): 42.

Felip A, Zanuy S, Muriach B, et al. 2008. Reduction of sexual maturation in male *Dicentrarchus labrax* by continuous light both before and during gametogenesis. Aquaculture, 275(1-4): 347–355.

Gao J, Xu GC, Xu P. 2021. Gills full-length transcriptomic analysis of osmoregulatory adaptive responses to salinity stress in *Coilia nasus*. *Ecotoxicology and Environmental Safety*, **226**: 112848.

Gao XY. 2011. Multiple testing corrections for imputed SNPs. Genetic Epidemiology, 35(3): 154–158.

Gao XY, Starmer J, Martin ER. 2008. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic Epidemiology, 32(4): 361–369.

Gershon E, Plaks V, Dekel N. 2008. Gap junctions in the ovary: expression, localization and function. Molecular and Cellular Endocrinology, 282(1-2): 18–25.

Gifford JAH. 2015. The role of WNT signaling in adult ovarian folliculogenesis. Reproduction, 150(4): R137–R148.

Hay ED. 1991. Cell Biology of Extracellular Matrix. 2nd ed. New York: Springer Science & Business Media.

Hazzard TM, Stouffer RL. 2000. Angiogenesis in ovarian follicular and luteal development. Best Practice & Research Clinical Obstetrics & Gynaecology, 14(6): 883–900.

Heins DC, Brown-Peterson NJ. 2022. The reproductive biology of small fishes and the clutch concept: combining macroscopic and histological approaches. Aquaculture, Fish and Fisheries, 2(4): 253–264.

Huang YF, Rypel AL, Murphy BR, et al. 2022. Declined fitness in larvae born from long-distance migrants of anadromous *Coilia nasus* in the lower reaches of the Yangtze River, China. Zoological Research, 43(3): 404–408.

Irving-Rodgers HF, Rodgers RJ. 2005. Extracellular matrix in ovarian follicular development and disease. Cell and Tissue Research, 322(1): 89–98.

Järveläinen H, Puolakkainen P, Pakkanen S, et al. 2006. A role for decorin in cutaneous wound healing and angiogenesis. *Wound Repair and Regeneration*, 2006, **14**(4): 443–452.

Jonsson B, Jonsson N. 2014. Early environment influences later performance in fishes. Journal of Fish Biology, 85(2): 151–188.

Kedem A, Ulanenko-Shenkar K, Yung Y, et al. 2022. The involvement of Lumican in human ovulatory processes. Reproductive Sciences, 29(2): 366–373.

Kim T, Li DM, Terasaka T, et al. 2019. SRXN1 is necessary for resolution of GnRH-induced oxidative stress and induction of gonadotropin gene expression. Endocrinology, 160(11): 2543–2555.

Lapointe E, Boerboom D. 2011. WNT signaling and the regulation of ovarian steroidogenesis. Frontiers in Bioscience-Scholar, 3(1): 276–285.

Li ZQ, Zhang Z, He ZD, et al. 2009. A partition-ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHEsis (http://analysis.bio-x.cn). *Cell Research*, **19**(4): 519–523.

Liang J, Fu Y, Cruciat CM, et al. 2011. Transmembrane protein 198 promotes LRP6 phosphorylation and Wnt signaling activation. Molecular and Cellular Biology, 31(13): 2577–2590.

Liu Y, Wu Z, Hu M. 2008. Advances on tapertail anchovy *Coilia ectenes* in China. Fisheries Science, 27(4): 205–209. (in Chinese)

Liu YL, Kossack ME, McFaul ME, et al. 2022. Single-cell transcriptome reveals insights into the development and function of the zebrafish ovary. eLife, 11: e76014.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data

using real-time quantitative PCR and the 2-ΔΔC_T method. Methods, 25(4): 402–408.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12): 550.

Lu YL, Shah T, Hao ZF, et al. 2011. Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. PLoS One, 6(9): e24861.

Lubzens E, Young G, Bobe J, et al. 2010. Oogenesis in teleosts: how fish eggs are formed. General and Comparative Endocrinology, 165(3): 367–389.

Lundqvist H. 1980. Influence of photoperiod on growth in Baltic salmon parr (*Salmo salar* L.) with special reference to the effect of precocious sexual maturation. Canadian Journal of Zoology, 58(5): 940–944.

Ma J, Li B, Zhao J, et al. 2020. Environmental influences on the spatio-temporal distribution of *Coilia nasus* in the Yangtze River estuary. Journal of Applied Ichthyology, 36(3): 315–325.

Mamo S, Carter F, Lonergan P, et al. 2011. Sequential analysis of global gene expression profiles in immature and *in vitro* matured bovine oocytes: potential molecular markers of oocyte maturation. BMC Genomics, 12: 151.

Manor ML, Weber GM, Salem M, et al. 2012. Effect of sexual maturation and triploidy on chemical composition and fatty acid content of energy stores in female rainbow trout, *Oncorhynchus mykiss. Aquaculture*, **364–365**: 312–321.

Martyniuk CJ, Prucha MS, Doperalski NJ, et al. 2013. Gene expression networks underlying ovarian development in wild largemouth bass (*Micropterus salmoides*). PLoS One, 8(3): e59093.

McKenna A, Hanna M, Banks E, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9): 1297–1303.

Medford BA, Mackay WC. 1978. Protein and lipid content of gonads, liver, and muscle of northern pike (*Esox lucius*) in relation to gonad growth. Journal of the Fisheries Research Board of Canada, 35(2): 213–219.

Murua H, Saborido-Rey F. 2003. Female reproductive strategies of marine fish species of the North Atlantic. Journal of Northwest Atlantic Fishery Science, 33: 23–31.

Nagahama Y, Yamashita M. 2008. Regulation of oocyte maturation in fish. *Development*, *Growth & Differentiation*, **50**(S1): S195–S219.

Nikitovic D, Papoutsidakis A, Karamanos NK, et al. 2014. Lumican affects tumor cell functions, tumor–ECM interactions, angiogenesis and inflammatory response. Matrix Biology, 35: 206–214.

Palaiokostas C, Houston RD. 2018. Genome-wide approaches to understanding and improving complex traits in aquaculture species. CABI Reviews, 12(55): 1–10.

Peng JY, Gao KX, Xin HY, et al. 2016. Molecular cloning, expression analysis, and function of decorin in goat ovarian granulosa cells. Domestic Animal Endocrinology, 57: 108–116.

Peng JY, Xin HY, Han P, et al. 2015. Expression and regulative function of tissue inhibitor of metalloproteinase 3 in the goat ovary and its role in cultured granulosa cells. Molecular and Cellular Endocrinology, 412: 104–115.

Phan P, Saikia BB, Sonnaila S, et al. 2021. The saga of endocrine FGFs. Cells, 10(9): 2418.

Purcell S, Neale B, Todd-Brown K, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3): 559–575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6): 841–842.

Ramos-Júdez S, Danis T, Angelova N, et al. 2022. Transcriptome analysis of flathead grey mullet (*Mugil cephalus*) ovarian development induced by recombinant gonadotropin hormones. Frontiers in Physiology, 13: 1033445.

Renner O, Harsch S, Schaeffeler E, et al. 2009. Mutation screening of apical sodium-dependent bile acid transporter (*SLC10A2*): novel haplotype block including six newly identified variants linked to reduced expression. Human Genetics, 125(4): 381–391.

Retief JD. 1999. Phylogenetic analysis using PHYLIP. *In*: Misener S, Krawetz SA. Bioinformatics Methods and Protocols. Totowa: Humana Press, 243–258.

Rowe DK, Thorpe JE. 1990. Differences in growth between maturing and non-maturing male Atlantic salmon, *Salmo salar* L. , parr. *Journal of Fish Biology*, **36**(5): 643–658.

Salem M, Kenney PB, Rexroad CE, et al. 2006. Molecular characterization of muscle atrophy and proteolysis associated with spawning in rainbow trout. Comparative Biochemistry and Physiology Part D:Genomics and Proteomics, 1(2): 227–237.

Saunders RL, Henderson EB, Glebe BD. 1982. Precocious sexual maturation and smoltification in male Atlantic salmon (*Salmo salar*). Aquaculture, 28(1-2): 211–229.

Shi YH, Zhang GY, Zhang HM, et al. 2015. Complete artificial propagation and embryonic development of *Coilia nasus. Journal of Shanghai Ocean University*, **24**(1): 36–43. (in Chinese)

Simpson AL, Metcalfe NB, Huntingford FA, et al. 1996. Pronounced seasonal differences in appetite of Atlantic salmon parr, *Salmo salar*: effects of nutritional state and life-history strategy. Functional Ecology, 10(6): 760–767.

Sun DL, Wen HS, Qi X, et al. 2023. Comparative study of candidate sex determination regions in snakeheads (*Channa argus* and *C. maculata*) and development of novel sex markers. Aquaculture, 575: 739771.

Taranger GL, Carrillo M, Schulz RW, et al. 2010. Control of puberty in farmed fish. General and Comparative Endocrinology, 165(3): 483–515.

Tveranger B. 1985. Variation in growth rate, liver weight and body composition at first sexual maturity in rainbow trout. Aquaculture, 49(2): 89–99.

van Es MA, Veldink JH, Saris CGJ, et al. 2009. Genome-wide association study identifies 19p13.3 (*UNC13A*) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. Nature Genetics, 41(10): 1083–1087.

Wallace RA, Selman K. 1981. Cellular and dynamic aspects of oocyte growth in teleosts. American Zoologist, 21(2): 325–343.

Wang KS, Zuo LJ, Pan Y, et al. 2015. Genetic variants in the *CPNE5* gene are associated with alcohol dependence and obesity in Caucasian populations. Journal of Psychiatric Research, 71: 1–7.

Wang MY, Xu GC, Tang YK, et al. 2021. Investigation of the molecular mechanisms of antioxidant damage and immune response downregulation in liver of *Coilia nasus* under starvation stress. Frontiers in Endocrinology, 12: 622315.

Wang T, Wu X, Song LL, et al. 2023. Identification of candidate growth-related SNPs and genes using GWAS and transcriptome analyses in leopard coral grouper (*Plectropomus leopardus*). Aquaculture, 574: 739677.

Woo J, Suh W, Sung JH. 2022. Hair growth regulation by Fibroblast Growth Factor 12 (FGF12). International Journal of Molecular Sciences, 23(16): 9467.

Wu T, Liu ZP, Mak TSH, et al. 2022. Polygenic power calculator: statistical power and polygenic prediction accuracy of genome-wide association studies of complex traits. Frontiers in Genetics, 13: 989639.

Xu GC, Bian C, Nie ZJ, et al. 2020. Genome and population sequencing of a chromosome-level genome assembly of the Chinese tapertail anchovy (*Coilia nasus*) provides novel insights into migratory adaptation. GigaScience, 9(1): giz157.

Xu GC, Du FK, Li Y, et al. 2016. Integrated application of transcriptomics and metabolomics yields insights into population-asynchronous ovary development in *Coilia nasus. Scientific Reports*, **6**(1): 31835.

Yamaguchi Y, Ruoslahti E. 1988. Expression of human proteoglycan in Chinese hamster ovary cells inhibits cell proliferation. Nature, 336(6196): 244–246.

Yin DH, Lin DQ, Ying CP, et al. 2020. Metabolic mechanisms of *Coilia nasus* in the natural food intake state during migration. Genomics, 112(5): 3294–3305.

Yin LL, Zhang HH, Tang ZS, et al. 2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics*, *Proteomics & Bioinformatics*, **19**(4): 619–628.

Zhang C, Dong SS, Xu JY, et al. 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics, 35(10): 1786–1788.

Zhang SM, Xu ZS, Luo LF, et al. 2023. Genetic diversity and population structure of *Coilia nasus* revealed by 2b-RAD sequencing. Biology, 12(4): 600.

Zheng F, Guo HY, Tang WQ, et al. 2012. Age structure and growth characteristics of anadromous populations of *Coilia nasus* in the Yangtze River. Chinese Journal of Zoology, 47(5): 24–31. (in Chinese)

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics, 44(7): 821–824.