



短串联重复序列介导的基因表达对荷斯坦牛复杂性状的影响

陶欣月, 白云飞, 唐永杰, 倪妙卓, 俞 英*

(中国农业大学动物科学技术学院, 北京 100193)

摘要: 旨在系统鉴定奶牛中与基因表达显著相关的短串联重复序列(expression short tandem repeats, eSTRs), 解析其在调控基因表达及复杂经济性状中的作用, 为奶牛分子育种提供功能变异资源。本研究选用108头中国荷斯坦牛进行尾静脉血样采集, 提取基因组DNA和总RNA, 分别进行30×深度全基因组重测序和转录组测序。因部分样本质量较差, 最终保留105头样本用于分析。对于遗传变异, 利用TRF和HipSTR完成STR群体规模的注释和定量。对于基因表达, 以Trimmomatic进行质量控制, STAR比对到参考基因组以及edegR定量。接着, 以基因转录起始位点±1 Mb范围为窗口, 采用线性模型将基因表达与STR进行关联分析, 鉴定得到cis-eSTR。之后结合加权共表达网络分析(weighted correlation network analysis, WGCNA)及Gene Ontology(GO)富集分析, 探索其在复杂性状中的潜在功能。共鉴定出25 154对显著的cis-eSTR-基因表达关联, cis-eSTR主要分布于内含子区及转录起始位点附近。加权基因共表达网络分析显示, 受cis-eSTR调控的基因被划分为多个功能模块, 部分模块与奶牛的产奶及免疫相关性状显著相关。GO富集分析显示, 这些模块广泛参与免疫应答、脂质代谢、细胞周期、DNA复制与稳定性等生物过程。多个关键基因(如IRF7、BoLA-DRB、CCT2等)在免疫调控与乳腺功能中具有潜在功能意义。cis-eSTR在调控奶牛基因表达和复杂性状方面具有重要作用。本研究所构建的cis-eSTR资源为奶牛健康等复杂性状的遗传机制研究和分子育种实践提供了功能注释信息标记。

关键词: 荷斯坦牛; 顺式表达数量性状位点; 短串联重复序列; 加权基因共表达; 复杂性状

中图分类号: S823.2

文献标志码: A

文章编号: 0366-6964(2026)02-0728-12

The Effect of Short Tandem Repeat Sequence Mediated Gene Expression on Complex Traits of Holstein Cattle

TAO Xinyue, BAI Yunfei, TANG Yongjie, NI Miaozhuo, YU Ying*

(College of Animal Science and Technology, Chinese Agricultural University, Beijing 100193, China)

Abstract: This study aimed to systematically identify expression short tandem repeats (eSTRs) in dairy cattle that are significantly associated with gene expression, in order to explore their regulatory roles in gene expression and complex economic traits, thereby providing functional variant resources for molecular breeding. A total of 108 Chinese Holstein cows were sampled for tail vein blood collection. Genomic DNA and total RNA were extracted and subjected to 30× whole-genome resequencing and transcriptome sequencing, respectively. After quality control, 105 samples were

收稿日期: 2025-07-04

基金项目: “十四五”国家重点研发计划(2021YFD1200903; 2023YFF1000902); 河北省科技计划“揭榜挂帅”项目(23227602Z); 国家自然科学基金面上项目(32472864); 现代农业产业技术体系项目(CARS-36)

作者简介: 陶欣月, 硕士生, 主要从事奶牛抗病育种研究, E-mail: a1076146543@163.com; 白云飞, 硕士生, 主要从事奶牛抗病育种研究, E-mail: byf09202018@163.com; 陶欣月和白云飞为同等贡献作者

* 通信作者: 俞 英, 主要从事动物抗病分子育种及表观遗传调控机理研究, E-mail: yuying@cau.edu.cn

retained for further analysis. STRs were annotated and genotyped at the population scale using TRF and HipSTR. Gene expression levels were quantified after quality filtering with Trimmomatic, alignment using STAR, and normalization with edgeR. Based on a ± 1 Mb window around transcription start sites (TSS), a linear model was used to assess the association between STR dosage and gene expression to identify *cis*-eSTRs. Weighted correlation network analysis (WGCNA) and Gene Ontology (GO) were further applied to explore the functional relevance of *cis*-eSTRs in complex traits. A total of 25 154 significant *cis*-eSTR-gene expression associations were identified. These *cis*-eSTRs were predominantly located in intronic regions and near TSSs. WGCNA revealed that *cis*-eSTR regulated genes clustered into several functional modules, some of which were significantly associated with traits related to milk production and immune. GO enrichment analysis revealed that these modules are broadly involved in biological processes such as immune response, lipid metabolism, cell cycle, DNA replication and stability. Several key genes (e.g., *IRF7*, *BoLA-DRB*, *CCT2*) might play important functional roles in immune regulation and mammary gland function. *cis*-eSTRs play important roles in regulating gene expression and complex traits in dairy cattle. The *cis*-eSTR resource generated in this study provides a valuable foundation for the functional annotation of complex traits and offers potential molecular markers for breeding applications.

Keywords: Holstein cattle; *cis* expression quantitative trait loci; short tandem repeat sequence; WGCNA; complex traits

* **Corresponding author:** YU Ying, E-mail: yuying@cau.edu.cn

表达数量性状位点(expression quantitative trait loci, eQTL)分析是揭示基因调控机制与复杂性状关系的重要手段,近年来已广泛应用于奶牛健康性状与生产性状的功能注释研究中^[1-2]。传统的eQTL研究主要聚焦于单核苷酸多态性(single nucleotide polymorphism, SNP),通过关联基因型与转录水平,挖掘影响表型的关键调控位点^[3]。然而,复杂性状的遗传基础往往涉及多种类型的变异,尤其是短串联重复序列(short tandem repeat, STR)等结构变异。STRs是由1~6 bp核心单元首尾相连、多次重复形成的基因组串联重复序列,在整个基因组中广泛分布且具有高度多态性^[4]。STR由于其高度多态性和在基因调控区的频繁分布,在调控基因表达方面可能发挥独特作用^[5-6]。在人类、猪和鸡等多个物种中,已有研究发现STRs通过介导基因表达对复杂性状和疾病产生重要影响^[7-9]。在牛中,牛GTEx项目^[10]的研究揭示了影响多种组织中基因表达与剪接变异的非编码单核苷酸多态性(SNP)和插入与缺失(insertion-deletion, Indel),但未系统评估STR的作用。近期的研究在牛睾丸组织中鉴定了eSTR,并发现其与精液质量及雄性生育力相关^[11]。然而,关于STR在牛体内多组织表达调控及复杂经济性状中的系统性作用仍缺乏深入

研究。因此,有必要对奶牛eSTR在健康与生产性状调控中的功能进行进一步探索。

本研究以中国荷斯坦牛为研究对象,基于血液样本构建高质量表达谱和STR基因型数据,系统开展*cis*-eSTR识别与功能注释工作。通过基因组定位、WGCNA分析和GO富集,揭示*cis*-eSTR的调控特征及其在免疫、代谢等性状中的潜在作用。研究结果为奶牛复杂性状的调控机制提供新见解,并为分子育种提供潜在的功能变异资源。

1 材料与方法

1.1 样本采集及核酸提取

本研究共选择了108头中国荷斯坦奶牛(87头来源于国家核心育种场,21头来自另一大型奶牛场),均于夏季采集尾静脉血样,使用EDTA抗凝管保存,用于基因组DNA和总RNA提取,并同步采集DHI乳样。样本胎次覆盖1~7胎,其中1胎泌乳牛占比最高(41.9%)。生产性能数据显示,国家核心育种场奶牛的平均日产奶量为(44.05 $\pm$ 6.51) kg,另一大型牛场为(39.95 $\pm$ 10.93) kg,表明所有个体均具有较好的生产水平。DNA提取采用天根生物血液基因组DNA试剂盒(Tiagen Blood Genomic DNA Extraction Kit,天根生物科技有限公司),

RNA 提取使用 TRIzol 试剂,遵循标准 TRIzol 试剂操作流程,包括氯仿分层、异丙醇沉淀和 75% 乙醇洗涤。提取得到的 RNA 样品经浓度测定和完整性评估,确保满足后续高通量测序的质量要求。

1.2 血液 DNA 及 RNA 测序

基因组 DNA 样本在华大基因(BGI)平台进行 30× 深度全基因组重测序,每头牛约产出 90 Gb 数据。RNA 样品在 Illumina NovaSeq6000 平台进行 150 bp 双端测序。对比率较低的样本($n=3$)在后续分析中剔除,最终保留 105 头奶牛样本。

1.3 STR 分型

本研究采用 Tandem Repeat Finder(TRF)^[12]工具对 ARS-UCD1.2 版本^[13]的牛参考基因组进行 STR 的识别和注释,保留重复单元长度为 1~6 bp 的序列,并生成 STR 注释文件。群体 STR 分型首先利用 BWA^[14]和 samtools^[15]工具对测序读段进行比对和处理,生成 BAM 格式文件。随后,使用 HipSTR^[16]软件,并以前述产生的 STR 注释文件为参考,对样本进行 STR 基因分型。HipSTR 输出的 VCF 文件经进一步过滤,仅保留多态性明显且质量得分高(>0.9)的 STR 位点,用于后续分析。

1.4 基因表达定量

首先使用 Trimmomatic^[17]进行质量控制和过滤,随后利用 STAR^[18]软件将其比对到 ARS-UCD1.2 参考基因组。保留在至少 20% 样本中 TPM >0.1 且 counts >6 的基因,最终共获得 14 765 个表达基因。最后,使用 edgeR^[19]对基因的 counts 进行 TMM 转化和定量。

1.5 cis-eSTR 鉴定的线性模型中协变量的选择分析

为了考虑混杂因素对 cis-eSTR 鉴定的影响,识别相关变量并将其作为协变量纳入模型。首先,使用 SNPrelate^[20]软件计算基因型主成分(principal components, PCs),并将前 5 个主成分作为协变量纳入模型。接着,对基因表达的前 10 个主成分(expression principal components, EPCs)与已知的生物学因素(包括:泌乳天数(days in milk, DIM)、胎次(parity)、血常规指标(白细胞数-WBC、中性粒细胞数-Neu、淋巴细胞数-Lym、单核细胞数-Mon、嗜碱性粒细胞数-Bas、嗜酸性粒细胞数-Eos)、奶牛群体改良(dairy herd improvement, DHI)信息、RNA 完整值(RNA integrity number, RIN)及牛场(farm))进行皮尔逊相关分析,具有显著性影响的因素也被纳

入模型作为协变量。此外,使用 PEER^[21]软件(R 包)通过估计 TMM 标准化基因表达数据的残差来识别未知的混杂因素。为确定纳入模型中 PEER 因子的最优数量,本研究遵循参考文献[21]中的方法,通过绘制因子相关性诊断图(基于 ARD 参数),系统评估了不同数量 PEER 因子所能解释的基因表达方差。当新增因子对方差解释度的边际贡献趋近于 0 时,停止增加选择的因子数量。随后,进行皮尔逊相关分析,以评估 PEER 因子是否捕获了先前未考虑的额外混杂效应。

1.6 cis-eSTR 的鉴定

本研究使用线性模型进行 cis-eSTR 分析,以探究转录起始位点(transcription start sites, TSS)上下游 ± 1 Mb 范围内的调控效应。通过关联 STR 剂量与基因表达水平,评估 STR 变异对基因表达的影响。STR 剂量定义为两个等位基因相较于参考基因组序列长度差异的总和。用于 cis-eSTR 分析的统计模型为: $Y=aX+b$ 。其中, Y 代表基因表达水平, a 是效应值, X 表示 STR 剂量, b 为协变量。关联分析采用 FastQTL^[22]软件在置换检验模式(permutation mode)下进行,随后使用 R 包 q value^[23]中实现的 Benjamini-Hochberg 方法进行多重检验校正,以控制错误发现率(false discovery rate, FDR)。只有 FDR <0.05 的 STR-基因对被认为具有显著性。

1.7 cis-eSTR 的注释与功能富集分析

为系统解析 cis-eSTR 在基因组中的分布特征及其潜在功能,本研究采用 Ensembl 开发的变异注释工具 Variant Effect Predictor (VEP)对显著 cis-eSTR 位点进行功能注释,统计其在基因功能区域的分布。

1.8 加权基因共表达网络分析(WGCNA)

为识别 cis-eSTR 调控基因与表型性状之间的潜在共表达关系,本研究采用 WGCNA^[24]方法构建基因共表达网络。通过相似性矩阵和拓扑重叠矩阵识别模块,设定最小模块基因数进行模块划分。将各模块与荷斯坦牛的生产、健康和免疫性状进行相关性分析,并对与性状显著相关的模块基因进行 GO 富集分析,以揭示其潜在功能机制。

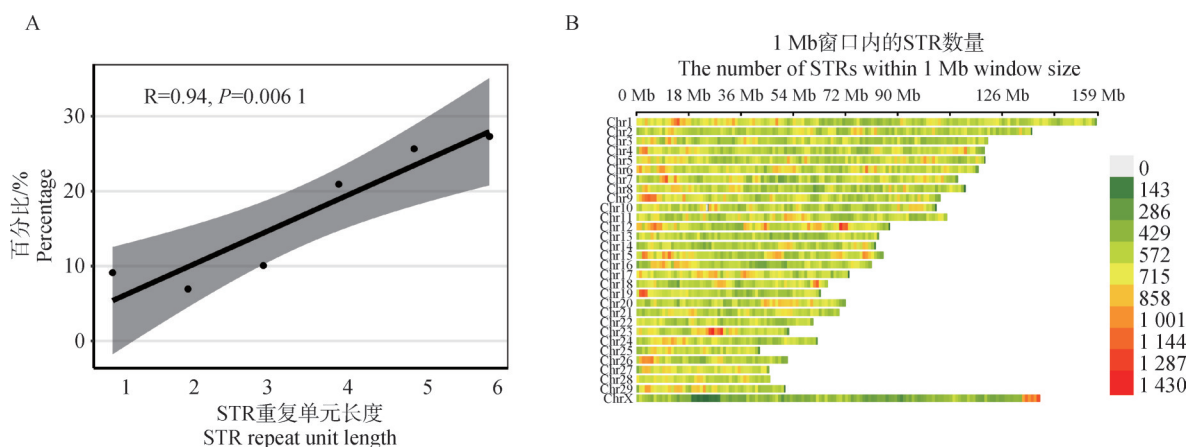
2 结 果

2.1 荷斯坦牛 STR 的全基因组分布与多态性特征

为获取牛基因组中 STR 的分布信息并为后续

分型提供参考文件,本研究基于ARS-UCD1.2参考基因组,使用TRF软件对常染色体和X染色体进行注释,共鉴定出6 882 952个核心重复单元为1~6 bp的STR位点,其中单核苷酸627 731个、二核苷酸477 988个、三核苷酸692 752个、四核苷酸1 439 792个、五核苷酸1 765 386个、六核苷酸1 879 303个。各类型STR比例依次递增,STR重复单元长度与其数量呈显著正相关(图1A)。在此基础上,使用HipSTR软件对105份中国荷斯坦牛

样本进行群体规模的STR基因分型。利用上述方法生成的注释文件,共在常染色体和X染色体上成功推断出6 676 126个位点的基因型,分型成功率为96.99%。随后,使用VCFtools对原始基因型进行质控,去除等位基因数少于两个的低多态性位点,最终保留1 601 869个多态性STR用于后续eSTR分析。本研究获得的荷斯坦牛全基因组STR多态性位点密度分布如图1B所示,1 Mb窗口内最多可包含1 430个多态性STR。



A. 荷斯坦牛参考基因组STR基序长度与数量占比的相关性分析;B. 荷斯坦牛群体的基因组STR位点密度图

A. Correlation analysis between the length and proportion of STR motifs in the reference genome of Holstein cattle; B. Genomic STR locus density map of Holstein cattle population

图1 STR重复基序长度分布趋势与多态性STR位点的基因组分布

Fig. 1 Trend of STR repeat length distribution and genomic distribution of polymorphic STR loci

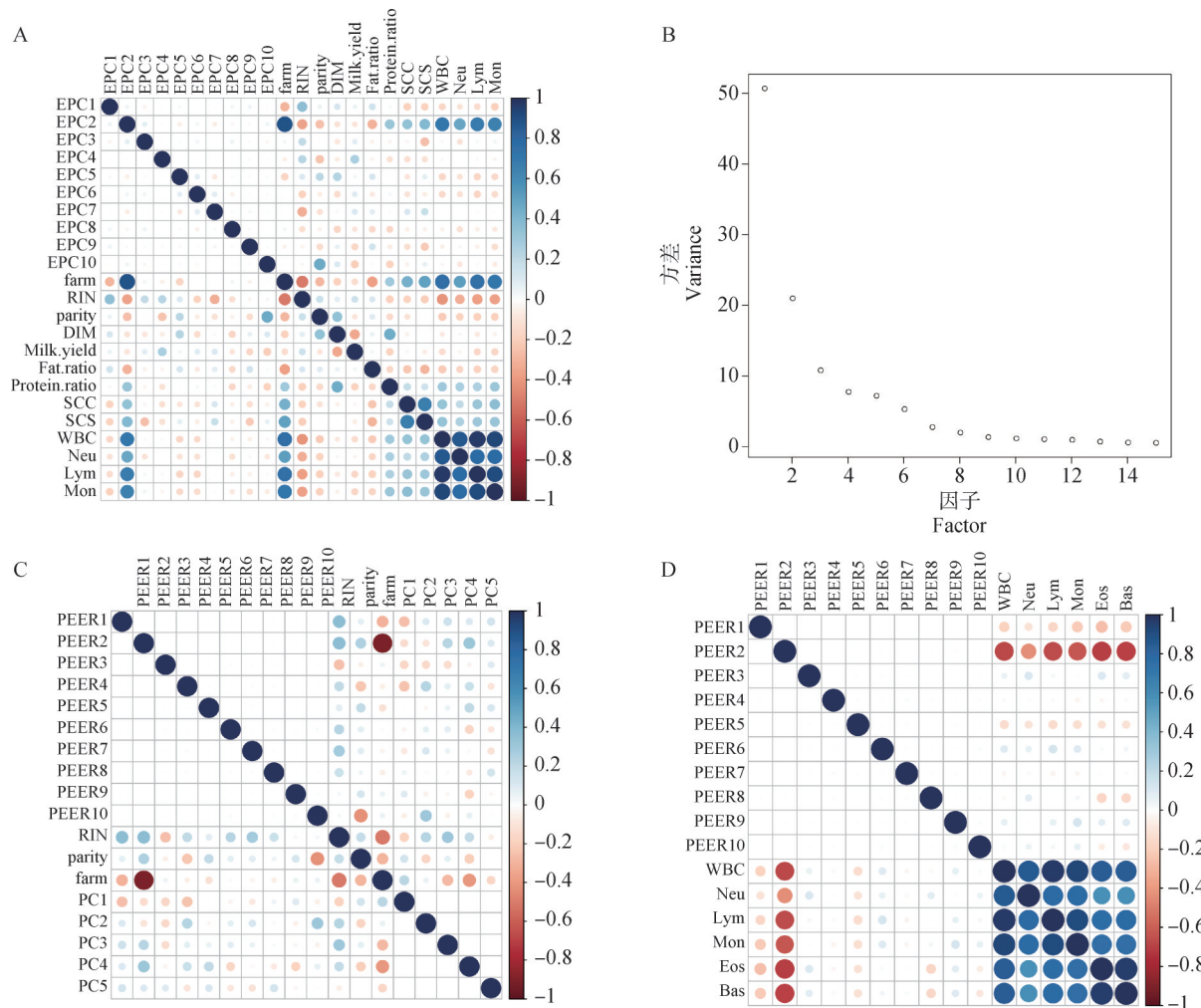
2.2 荷斯坦牛 *cis*-eSTR 鉴定中的协变量

在 *cis*-eSTR 鉴定过程中,基因表达数据会受到生物和技术因素的影响。为确保结果的可靠性,本研究首先评估了已知的生物学因素对基因表达的影响。相关分析结果显示, *farm*、*parity* 和 *RIN* 三者与前10个EPCs的总体相关性较高(图2A),因此被作为协变量纳入模型。虽然血常规指标和DHI信息也与EPCs存在相关性,但由于可能存在与 *farm* 相关的混杂效应,因此未将其列为协变量。此外,还使用了PEER因子来考虑未知的混杂因素。因子相关性诊断图表明,前10个PEER因子可解释基因表达的最大变化。第10个因子之后,解释度几乎不变且趋近于0(图2B)。因此,前10个PEER因子被选为模型的协变量。为了验证PEER因子的可靠性,进一步将其与模型中的已知协变量(如前5个基因型PC、*farm*、*parity* 和 *RIN*)进行相关分析。结果显示,PEER因子与 *farm* 和 *parity* 等已知混杂因素

有一定的相关性,这表明它们可以部分解释这些影响。此外,PEER因子还显示与其他未作为协变量的变量(如血液参数)存在相关性(图2C、D)。这凸显了PEER识别隐藏混杂因素的能力,进一步提高了本研究中荷斯坦牛 *cis*-eSTR 识别结果的可靠性。

2.3 荷斯坦牛 *cis*-eSTR 的鉴定及调控特征

本研究使用FastQTL软件识别转录起始位点周围±1 Mb窗口内的 *cis*-eSTRs,同时加入已知和推断的协变量(5个基因型PC、*farm*、*parity*、*RIN* 和10个PEER因子),以考虑潜在的混杂因素。在STR剂量和归一化基因表达之间进行关联分析,确定了25 154个 *cis*-eSTR-基因关联($FDR < 0.05$)和3 082个与 *cis*-eSTR关联的基因。例如,16号染色体上的 *bta-mir-29b-2* 基因与其附近区域的多个 *cis*-eSTR显著相关(表1),表明其在转录后基因调控中的作用。此外,研究发现23号染色体上的 *cis*-



A. 基因表达前 10 个 PCs (EPCs) 与已知生物学因素的相关性; B. 每个 PEER 因子对基因表达变异的贡献度; C. PEER 因子与 RIN、parity、farm 及基因型主成分 (PCs) 的相关性; D. PEER 因子与血液参数的相关性

A. The correlation between the top 10 PCs of gene expression (EPCs) and known biological factors; B. The contribution of each PEER factor to gene expression variation; C. The correlation between PEER factors and RIN, parity, farm, and genotype principal components (PCs); D. The correlation between PEER factors and blood parameters

图 2 协变量相关性分析

Fig. 2 Covariate correlation analysis

eSTR 23_27826583_AG 与附近 11 个基因的表达相关 (表 2), 是本研究中关联基因数量最多的 *cis*-eSTR。

为了从基因组学角度了解荷斯坦牛 *cis*-eSTR 的调控特征, 进一步分析了已鉴定的 *cis*-eSTR 与其关联基因的转录起始位点之间的距离。结果发现, 大多数 *cis*-eSTR 都集中在其关联基因的 TSS 附近。随着与 TSS 物理距离的增加, *cis*-eSTR 的数量逐渐减少, 与下游区域相比, 位于基因上游的 *cis*-eSTR 略多 (图 3A)。为了描述 *cis*-eSTR 的基因组分布和功能特征, 使用了广泛采用的变异效应预测器

(VEP) 工具进行功能注释。分析结果显示, 大部分 *cis*-eSTR 分布在基因间区 (intergenic) 和内含子区 (intron), 而较少分布在编码序列 (coding DNA sequence, CDS) 和非翻译区 (untranslated regions, UTR) (图 3B)。

2.4 荷斯坦牛 *cis*-eSTR 调控基因的共表达网络及其功能特征分析

基于本研究鉴定的 3 082 个 *cis*-eSTR 调控基因, 构建了加权基因共表达网络, 并分析其与荷斯坦牛生产及健康性状的潜在关联。按照 WGCNA 标准流程, 选择软阈值为 4 ($R^2 > 0.85$) 构建无标度

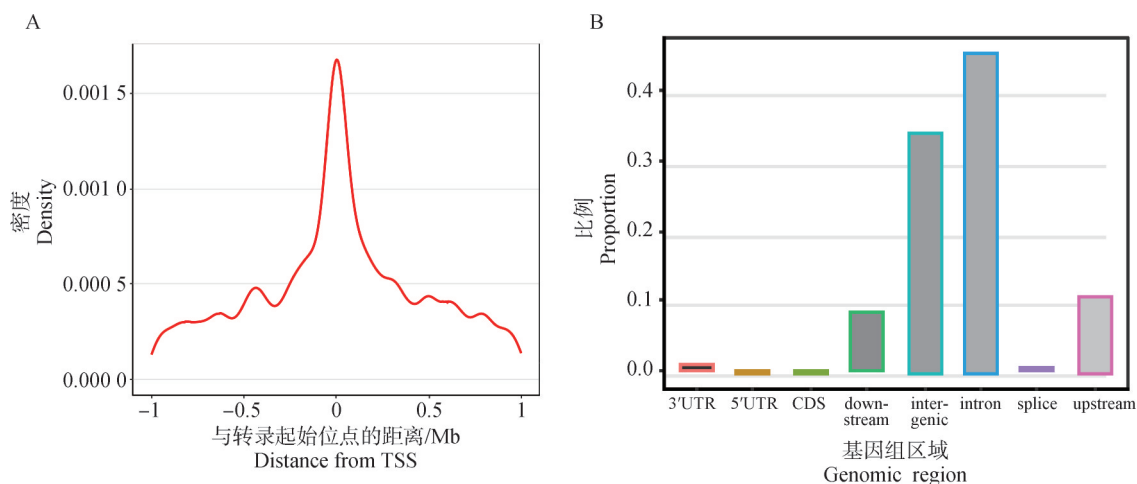
表1 前10个最显著的 *cis*-eSTR 及其关联的基因Table 1 The top 10 most significant *cis*-eSTRs and their associated genes

顺式表达数量性状短串联重复序列 <i>cis</i> -eSTR	关联基因 Associated gene	距离/bp Distance	显著性 <i>P</i> 值 <i>P</i> value	效应值 Effect size
10_86879661_GTT	ENSBTAG00000011985	-1 957	2.43E-43	0.12
9_87281756_ATTT	ENSBTAG00000036061	73 951	3.18E-32	0.30
9_86877091_A	ENSBTAG00000036061	-330 714	1.46E-30	1.18
16_75522657_A	<i>bta-mir-29b-2</i>	-137 713	2.06E-29	1.11
16_75522038_AGC	<i>bta-mir-29b-2</i>	-138 332	1.05E-28	0.37
16_21773538_AGGGGG	<i>TGFB2</i>	-34 465	2.87E-27	-0.52
23_39522637_T	<i>KIF13A</i>	-3 465	4.11E-27	-0.35
16_75522510_A	<i>bta-mir-29b-2</i>	-137 860	4.12E-27	0.61
18_57557982_TGCC	ENSBTAG00000019227	-83 311	4.48E-27	-1.32
16_75518951_T	<i>bta-mir-29b-2</i>	-141 419	9.57E-27	1.12

表2 与 23_27826583_AG 显著关联的基因

Table 2 Genes significantly associated with 23_27826583_AG

关联基因 Associated gene	距离/bp Distance	显著性 <i>P</i> 值 <i>P</i> value	效应值 Effect size
ENSBTAG00000001476	-70 319	2.20E-18	-0.53
ENSBTAG00000012208	-44 623	2.84E-17	-0.49
<i>MIC1</i>	-86 615	1.07E-13	0.44
ENSBTAG00000053433	-134 699	1.77E-13	0.46
<i>MIC2</i>	-22 420	3.63E-09	-0.39
ENSBTAG00000051047	-117 674	2.76E-08	-0.33
<i>BoLA-NC1</i>	-726 506	3.49E-08	-0.37
ENSBTAG00000005146	-701 400	7.34E-08	0.34
<i>VAR2</i>	-332 504	2.37E-07	0.21
ENSBTAG00000001477	-665 948	1.15E-05	0.26
<i>CCHCR1</i>	-166 171	1.84E-05	0.21

A. eSTR 在目标基因转录起始位点 (TSS) ± 1 Mb 范围内的分布; B. eSTR 在不同基因组功能区域的分布

A. The distribution of eSTR within the range of ± 1 Mb at the transcription start site (TSS) of the target gene; B. Distribution of eSTR in different genomic functional regions

图3 荷斯坦牛 *cis*-eSTR 的分布和功能特征Fig. 3 Distribution and functional characteristics of *cis*-eSTR in Holstein cattle

网络,最终划分出12个表达模块。各模块在产奶量、乳成分(脂肪率、蛋白率、乳糖率)以及白细胞、淋巴细胞等血液指标间呈现显著相关性(图4A)。

其中,模块1与模块3均与荷斯坦牛血细胞类指标(Lym、WBC、Neu、Mon、Eos、Bas)呈正相关,提示其在免疫或造血调控中的潜在作用。GO富集分析显示,模块1中的基因富集于免疫系统过程(immune system process)、MHC II类蛋白复合物(MHC class II protein complex)等相关通路,关键基因如*BoLA-DQB*、*IRF7*、*HP*等在抗原呈递与免疫调节中具有重要功能(图4B)。其中*IRF7*的表达水平与位于染色体29上的一条A重复序列(29_51075001_A)显著相关,呈现等位基因剂量效应(图4C),提示该位点可能通过调控*IRF7*表达影响个体免疫能力。此外,模块1还在氧化还原过程(oxidation-reduction process)及氧化还原酶活性(oxidoreductase activity)等代谢通路中显著富集,提示其同时参与免疫与代谢调控。模块3富集于IL-12受体结合(interleukin-12 receptor binding)等免疫相关通路,同时显著关联于生长与衰老调控过程,包括生长调节(regulation of growth)、端粒酶活性调控(positive regulation of telomerase activity)等(图4D)。该模块中关键基因如*TERT*与*CCT2*,可能在免疫及代谢调控中发挥桥梁作用,其中*CCT2*表达与染色体5上的一条(A)_n重复序列显著相关(图4E)。

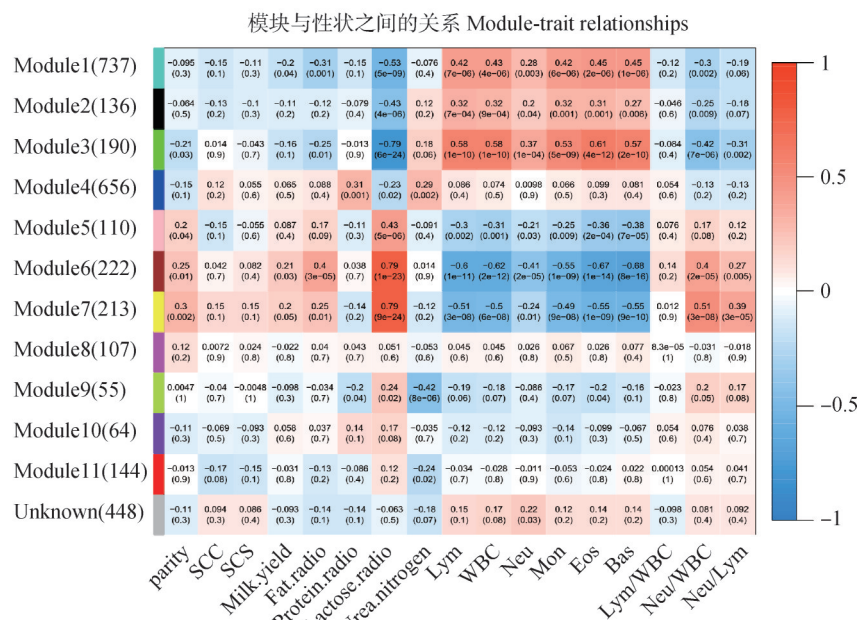
相较之下,模块6和模块7与荷斯坦牛血细胞参数呈强负相关,表明它们在某些生理状态下具有调节免疫抑制或免疫调节的作用。GO富集分析显示,这两个模块的基因都富集在免疫相关通路中,包括免疫反应的调节(regulation of immune response)、急性炎症反应(acute inflammatory response)和自噬(autophagy)等,表明它们在微调免疫功能中发挥作用。此外,*TIPIN*、*SMARCA1*和*DONSON*等关键基因也富集于DNA复制相关过程,如复制叉加工(replication fork processing)、复制叉保护(replication fork protection)等。*DONSON*和*TIPIN*的表达与位于1_2021547_T和10_13178513_A的*cis*-eSTR显著相关,且这些*cis*-eSTR对基因表达有明显的等位基因效应(图4F、G)。

3 讨论

3.1 大规模STR多态性位点的检测与调控热点识别

本研究基于105头中国荷斯坦牛的重测序数据共鉴定出6 676 126个STR位点,覆盖常染色体和X染色体,其中约160万个为多态性位点。相比之下,Xu等^[25]基于5头荷斯坦牛的深度重测序数据(30~50×)共鉴定出约6万个多态性STR位点。相较于该研究,本研究使用更大的样本群体,大幅度扩展了多态性STR的检测范围。结果表明,增加样本量可显著提高STR遗传变异的检测能力,为后续开展与性状相关的功能研究提供了丰富的资源基础。

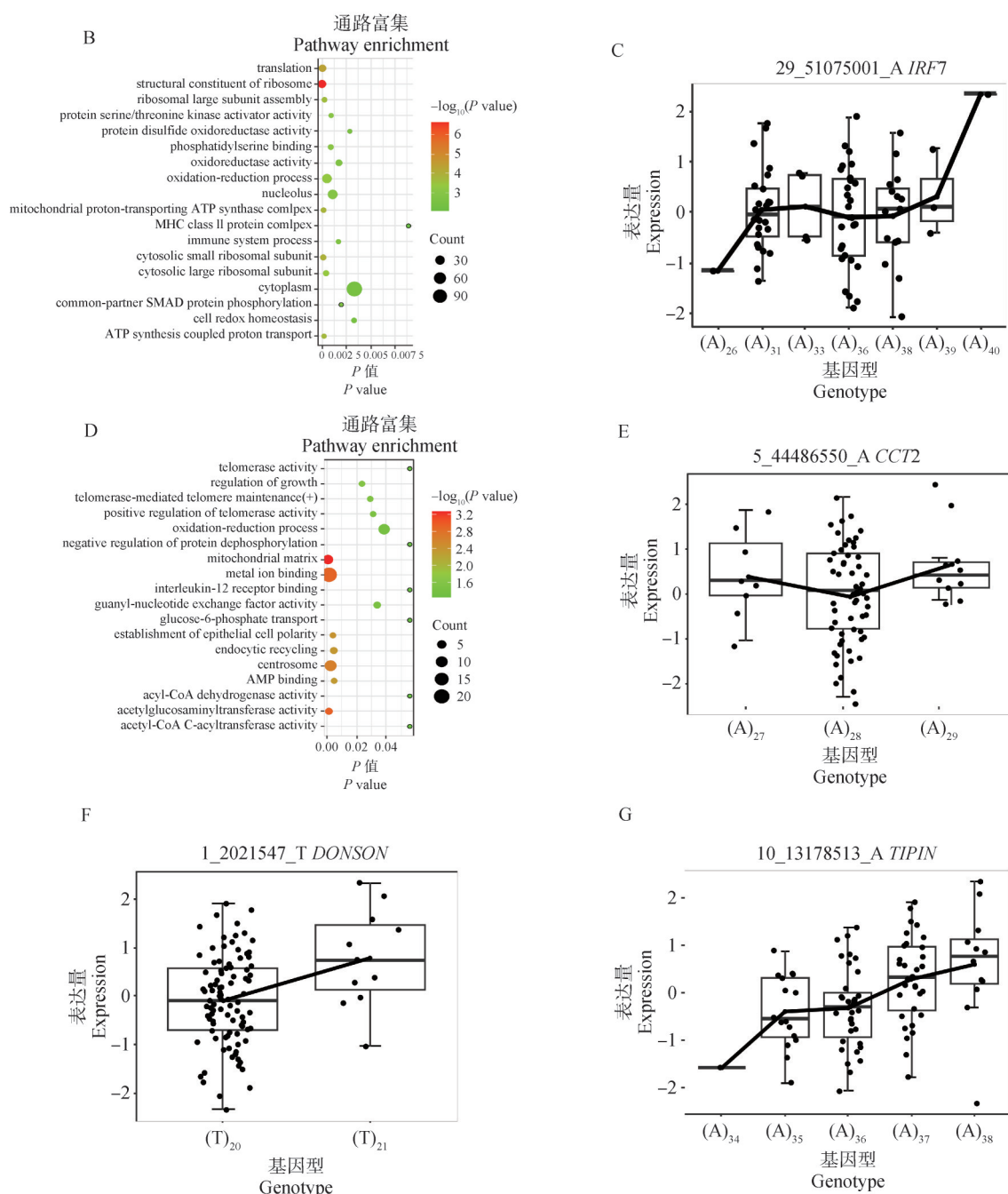
A



(图4 Fig.4)

(转下页 Carried forward)

(续图 4 Continued)



A. 基因共表达模块与荷斯坦牛表型的关联; B. 模块 1 基因的 GO 富集分析; C. 29_51075001_A 对 *IRF7* 表达的等位基因效应; D. 模块 3 基因的 GO 富集分析; E. 5_44486550_A 对 *CCT2* 表达的等位基因效应; F. 1_2021547_T 对 *DONSON* 表达的等位基因效应; G. 10_13178513_A 对 *TIPIN* 表达的等位基因效应

A. Association between gene co-expression module and Holstein cattle phenotype; B. GO enrichment analysis of module 1 genes; C. Allelic effects of 29_51075001_A on *IRF7* expression; D. GO enrichment analysis of module 3 genes; E. Allelic effects of 5_44486550_A on *CCT2* expression; F. Allelic effects of 1_2021547_T on *DONSON* expression; G. Allelic effects of 10_13178513_A on *TIPIN* expression

图 4 WGCNA 基因共表达模块及模块基因功能

Fig. 4 WGCNA gene co-expression modules and modules genes function

此外,在本研究所鉴定的显著的 *cis*-eSTR 中,前 10 个最显著的 *cis*-eSTR 有 5 个位于 BTA16 染色体,呈现出明显的染色体富集倾向,提示其可能是 *cis*-eSTR 的“热点区域”。此前,BTA16 在多个研究中已被证实具有与产奶、乳房炎、生殖疾病等性状显著关联的位点^[26-28],本研究中其富集 *cis*-eSTR 的现象为进一步挖掘功能调控变异提供了候选区域依据。同时,本研究中位于 23 号染色体的 23_27826583_AG 关联了最多的目标基因,提示该位点可能具有广谱型的调控能力。

3.2 荷斯坦牛 *cis*-eSTR 的基因组分布特征

对荷斯坦牛 *cis*-eSTR 的基因组位置进行注释发现,其主要分布在目标基因转录起始位点(TSS)附近,且随着与 TSS 距离的增加,*cis*-eSTR 的数量逐渐减少。这一趋势与人类和猪的研究^[8,29]一致,说明 *cis*-eSTR 作为顺式调控元件的功能可能依赖其与 TSS 的距离。此外,大多数 *cis*-eSTR 位于内含子和基因间区,这一分布模式也与其他牛相关研究的结果一致^[11]。已有研究表明,内含子不仅是剪接元件,还可作为调控单元参与转录调控。Dwyer 等^[30]的研究指出,位于启动子(promoter)或终止子(terminator)附近的内含子往往具有更强的转录增强潜力,而这类区域恰恰是 STRs 的富集位置之一。因此,这些内含子区的 *cis*-eSTR 可能通过其重复单元的扩张或收缩,直接干扰潜在的内含子增强子的活性,或者影响剪接因子结合,从而在转录或转录后水平调控基因表达。此外,越来越多的研究表明,基因间区并非“垃圾 DNA”,而是基因组中重要的调控区域。研究表明,H3K4me3 可标记基因间的活跃调控元件,从而放大其转录活性^[31]。通过全基因组关联分析(genome-wide association study, GWAS)定位于基因间区的遗传变异被发现可通过染色质空间互作与靶基因启动子建立联系,参与复杂性状的调控网络^[32]。结合这些研究结果可以推测,分布在基因间区的 *cis*-eSTR 可能通过影响局部染色质修饰或空间互作结构,介导调控元件与靶基因之间的沟通,从而实现对基因表达的精细调控。这一机制为解释本研究中一些关键的发现提供了新的视角:与多达 11 个基因表达显著关联的“热点”*cis*-eSTR 23_27826583_AG(表 2),其本身可能位于一个关键的调控热点或染色质环的锚定点上。该 STR 的等位基因变异可能通过微调三维基因组架构同时协调多个邻近基因的转录活动。

3.3 *cis*-eSTR 共调控基因的免疫功能富集与奶牛性状关联

本研究通过加权基因共表达网络分析(WGCNA)对 3 082 个受 *cis*-eSTR 调控的基因进行了系统性功能挖掘,识别出多个与奶牛生产和健康性状密切相关的共表达模块。从整体上看,这些模块显著富集于免疫、代谢、细胞生长与分化等关键生物过程,提示 STR 变异可能通过多条功能通路介导复杂性状的表型。

部分共表达模块(模块 1 和 3)与血液中白细胞、淋巴细胞等指标呈正相关,模块内的多个基因富集于抗原提呈、炎症反应、细胞因子受体结合等免疫调节路径。本研究鉴定出多个受 STR 直接调控的关键免疫调节因子。其中,I 型干扰素通路的核心转录因子 *IRF7* 的表达受到一个位于 29 号染色体的 A 重复序列(29_51075001_A)的显著调控。鉴于 *IRF7* 在抗病毒免疫中的核心作用^[33-34]及其表达水平与牛病毒性腹泻病毒(bovine viral diarrhea virus, BVDV)感染结局的已知关联^[35],该 eSTR 为解析奶牛抗病力的遗传基础提供了一个全新的候选调控位点。另一个受 STR 调控的关键免疫基因 *BoLA-DRB*,是牛主要组织相容性复合体中的一个重要组成部分,已知与多种奶牛疾病表型相关联^[36-37],这为阐释该区域经典疾病关联位点的潜在表达调控机制提供了新的证据。同时这些模块与乳脂率、乳糖率等乳成分性状显著相关,模块内基因富集于脂质代谢、氧化还原过程和能量代谢通路,表明 *cis*-eSTR 同时调控产奶性能。*TERT*、*CCT2* 等基因所在模块富集于端粒酶活性调控、细胞周期及衰老相关通路,与细胞寿命调控相关。*CCT2* 在多项研究中被发现参与乳腺疾病过程^[38-39],其表达受 STR 变异调控,在奶牛乳腺相关性状中同样具有研究价值。

此外,部分共表达模块(模块 6 和 7)与中性粒细胞和淋巴细胞的比例等指标呈负相关,富集于免疫反应调节、自噬及 DNA 复制相关功能,涉及 *TIPIN*、*SMARCA1*、*DONSON* 等与 DNA 稳定性维持相关的基因^[40-42]。*TIPIN* 通过调控 TIMELESS 的多聚 ADP 核糖化修饰(poly(ADP-ribosylation)),有助于保护停滞的复制叉并限制 DNA 复制应激,从而维持基因组稳定性^[43]。*SMARCA1* 不仅参与复制应激下的 DNA 修复,还被证实是先天免疫通路的调控因子,其通过调节 *PD-L1* 表达影响免疫逃逸,

提示其可能在感染与肿瘤相关免疫反应中兼具双重调控功能^[44]。DONSON在哺乳动物细胞周期中通过促进CMG(Cdc45-MCM-GINS)解旋酶复合体的组装参与DNA复制起始,维持复制准确性与细胞周期进程^[45]。这些与DNA复制应激和免疫调节密切相关的基因,表达均受到特定*cis*-eSTR的调控,提示STR变异可能通过影响基因组稳定性维护机制,间接调节个体的免疫状态和健康表现。

3.4 局限性与未来展望

本研究基于高深度重测序和系统的表达关联分析,揭示了*cis*-eSTR在奶牛健康与生产性状中的潜在调控作用。在考虑样本量合理性时,本研究参考了Chiang等^[46]在《Nature Genetics》上的开创性工作,该研究表明了100~200的样本量中可有效检测到遗传变异对基因表达的影响。本研究纳入的105头荷斯坦奶牛,在严格控制遗传背景与环境因素后,其样本规模符合探索性eSTR研究的要求。当然,我们承认统计功效对于中低频变异及效应量较小的位点仍可能不足。因此,后续研究可通过扩大样本规模或整合多群体数据,以进一步提升关联分析的灵敏度与稳健性。其次,本研究采用血液组织进行转录组测定,未考虑不同组织或发育阶段可能呈现出差异性的调控模式。未来可结合乳腺、肝脏、卵巢等关键组织的转录数据进行多组织比较分析,从而揭示eSTR在不同生理系统中的时空调控特征。此外,结合表观组学技术和功能验证试验,将有助于深入解析STR变异的具体分子机制,推动其在分子育种中的功能应用。

4 结 论

这项研究表明,STR对牛的基因表达有重要影响,并构建了首个中国荷斯坦牛血液组织*cis*-eSTR数据库。该数据库包括25 154个STR与附近基因表达水平之间的高置信度关联,为探索STR对牛复杂性状的影响提供了宝贵的资源。结果表明,*cis*-eSTRs主要位于内含子区和基因间区,通常集中在目标基因的转录起始位点附近。加权基因共表达网络分析和与荷斯坦牛表型的关联分析发现了与免疫、代谢、生长调节等有关的重要基因模块,强调了*cis*-eSTR在调控荷斯坦牛生产和健康性状中的重要性。

这一发现强调了将STR纳入调控研究的重要性,因为它们独特的突变和功能特性为了解复杂的

基因表达机制提供了宝贵的见解。受*cis*-eSTR影响的基因范围广,凸显了它们作为基因调控网络中关键角色的潜力,以及它们在了解复杂性状遗传基础方面的价值。

参考文献(References):

- [1] TANG Y, ZHANG J, LI W, et al. Identification and characterization of whole blood gene expression and splicing quantitative trait loci during early to mid-lactation of dairy cattle[J]. *BMC Genomics*, 2024, 25(1):445.
- [2] XIANG R, FANG L, LIU S, et al. Gene expression and RNA splicing explain large proportions of the heritability for complex traits in cattle[J]. *Cell Genom*, 2023, 3(10):100385.
- [3] CROTEAU-CHONKA D C, ROGERS A J, RAJ T, et al. Expression quantitative trait loci information improves predictive modeling of disease relevance of non-coding genetic variation[J]. *PLoS One*, 2015, 10(10):e140758.
- [4] MARGOLIASH J, FUCHS S, LI Y, et al. Polymorphic short tandem repeats make widespread contributions to blood and serum traits[J]. *Cell Genom*, 2023, 3(12):100458.
- [5] SHI Y, NIU Y, ZHANG P, et al. Characterization of genome-wide STR variation in 6487 human genomes[J]. *Nat Commun*, 2023, 14(1):2092.
- [6] GYMREK M, WILLEMS T, GUILMATRE A, et al. Abundant contribution of short tandem repeats to gene expression variation in humans[J]. *Nat Genet*, 2016, 48(1):22-29.
- [7] FOTSING S F, MARGOLIASH J, WANG C, et al. The impact of short tandem repeat variation on gene expression[J]. *Nat Genet*, 2019, 51(11):1652-1659.
- [8] WU Z, GONG H, ZHOU Z, et al. Mapping short tandem repeats for liver gene expression traits helps prioritize potential causal variants for complex traits in pigs[J]. *J Anim Sci Biotechnol*, 2022, 13(1):8.
- [9] ZHOU R, DE KONING D J, MCCORMACK H, et al. Short tandem repeats and methylation in the promoter region affect expression of cystathionine beta-synthase gene in the laying hen[J]. *Gene*, 2019, 710:367-374.
- [10] LIU S, GAO Y, CANELA-XANDRI O, et al. A multi-tissue atlas of regulatory variants in cattle[J]. *Nat Genet*, 2022, 54(9):1438-1447.

- [11] BHATI M, MAPEL X M, LLORET-VILLAS A, et al. Structural variants and short tandem repeats impact gene expression and splicing in bovine testis tissue[J]. *Genetics*, 2023, 225(3):iyad161.
- [12] BENSON G. Tandem repeats finder: a program to analyze DNA sequences[J]. *Nucleic Acids Res*, 1999, 27(2):573-580.
- [13] ROSEN B D, BICKHART D M, SCHNABEL R D, et al. De novo assembly of the cattle reference genome with single-molecule sequencing [J]. *Gigascience*, 2020, 9(3):giaa021.
- [14] LI H, DURBIN R. Fast and accurate short read alignment with Burrows-Wheeler transform [J]. *Bioinformatics*, 2009, 25(14):1754-1760.
- [15] LI H, HANDSAKER B, WYSOKER A, et al. The Sequence Alignment/Map format and SAMtools[J]. *Bioinformatics*, 2009, 25(16):2078-2079.
- [16] WILLEMS T, ZIELINSKI D, YUAN J, et al. Genome-wide profiling of heritable and de novo STR variations[J]. *Nat Methods*, 2017, 14(6):590-592.
- [17] BOLGER A M, LOHSE M, USADEL B. Trimmomatic: a flexible trimmer for Illumina sequence data[J]. *Bioinformatics*, 2014, 30(15):2114-2120.
- [18] DOBIN A, DAVIS C A, SCHLESINGER F, et al. STAR: ultrafast universal RNA-seq aligner [J]. *Bioinformatics*, 2013, 29(1):15-21.
- [19] ROBINSON M D, MCCARTHY D J, SMYTH G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data[J]. *Bioinformatics*, 2010, 26(1):139-140.
- [20] ZHENG X, LEVINE D, SHEN J, et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data [J]. *Bioinformatics*, 2012, 28(24):3326-3328.
- [21] STEGLE O, PARTS L, PIIPARI M, et al. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses [J]. *Nat Protoc*, 2012, 7(3):500-507.
- [22] ONGEN H, BUIL A, BROWN A A, et al. Fast and efficient QTL mapper for thousands of molecular phenotypes [J]. *Bioinformatics*, 2016, 32(10):1479-1485.
- [23] STOREY J D, BASS A J, DABNEY A, et al. qvalue: Q-value estimation for false discovery rate control [CP/OL]. 2023. <https://bioconductor.org/packages/qvalue/>. DOI:10.18129/B9.bioc.qvalue.
- [24] LANGFELDER P, HORVATH S. WGCNA: An R package for weighted correlation network analysis[J]. *BMC Bioinformatics*, 2008, 9:559.
- [25] XU L, HAASL R J, SUN J, et al. Systematic profiling of short tandem repeats in the cattle genome [J]. *Genome Biol Evol*, 2017, 9(1):20-31.
- [26] PENG P, LIU Y, ZHENG W, et al. Identification of genetic effects of ACADVL and IRF6 genes with milk production traits of Holstein cattle in China [J]. *Genes (Basel)*, 2022, 13(12):2393.
- [27] SAHANA G, GULDBRANDTSEN B, THOMSEN B, et al. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle[J]. *J Dairy Sci*, 2014, 97(11):7258-7275.
- [28] GUARINI A R, LOURENCO D A L, BRITO L F, et al. Genetics and genomics of reproductive disorders in Canadian Holstein cattle[J]. *J Dairy Sci*, 2019, 102(2):1341-1353.
- [29] MARTIN-TRUJILLO A, GARG P, PATEL N, et al. Genome-wide evaluation of the effect of short tandem repeat variation on local DNA methylation[J]. *Genome Res*, 2023, 33(2):184-196.
- [30] DWYER K, AGARWAL N, GEGA A, et al. Proximity to the promoter and terminator regions regulates the transcription enhancement potential of an intron[J]. *Front Mol Biosci*, 2021, 8:712639.
- [31] YU H, ZHANG Y, LIAO Z, et al. H3K4me3 amplifies transcription at intergenic active regulatory elements [J]. *Genes Dev*, 2025, 39(21-22):1290-1298.
- [32] SCHIERDING W, ANTONY J, CUTFIELD W S, et al. Intergenic GWAS SNPs are key components of the spatial and regulatory network for human growth [J]. *Hum Mol Genet*, 2016, 25(15):3372-3382.
- [33] NEGISHI H, TANIGUCHI T, YANAI H. The interferon (IFN) class of cytokines and the IFN regulatory factor (IRF) transcription factor family[J]. *Cold Spring Harb Perspect Biol*, 2018, 10(11):a028423.
- [34] MA W, HUANG G, WANG Z, et al. IRF7: role and regulation in immunity and autoimmunity [J]. *Front Immunol*, 2023, 14:1236923.
- [35] MALDONADO N, FREDERICKSEN F, ESPINEIRA C, et al. BVDV-1 induces interferon-beta gene expression through a pathway involving IRF1, IRF7, and NF-kappaB activation [J]. *Mol*

- Immunol*, 2020, 128:33-40.
- [36] YOSHIDA T, FURUTA H, KONDO Y, et al. Association of BoLA-DRB3 alleles with mastitis resistance and susceptibility in Japanese Holstein cows [J]. *Anim Sci J*, 2012, 83(5):359-366.
- [37] LE T D, NGUYEN V S, LO C, et al. Association between BoLA-DRB3 polymorphism and bovine leukemia virus proviral load in Vietnamese Holstein Friesian cattle[J]. *HLA*, 2022, 99(2):105-112.
- [38] ALAGUNDAGID B, GHATE S D, RAJENDRA V, et al. Exploring breast cancer exosomes for novel biomarkers of potential diagnostic and prognostic importance[J]. *3 Biotech*, 2023, 13(1):7.
- [39] SHOWALTER A E, MARTINI A C, NIERENBERG D, et al. Investigating Chaperonin-Containing TCP-1 subunit 2 as an essential component of the chaperonin complex for tumorigenesis [J]. *Sci Rep*, 2020, 10(1):798.
- [40] PROROK P, WOLF E, CARDOSO M C. Timeless-Tipin interactions with MCM and RPA mediate DNA replication stress response [J]. *Front Cell Dev Biol*, 2024, 12:1346534.
- [41] POOLE L A, CORTEZ D. Functions of SMARCAL1, ZRANB3, and HLTf in maintaining genome stability [J]. *Crit Rev Biochem Mol Biol*, 2017, 52(6):696-714.
- [42] KINGSLEY G, SKAGIA A, PASSARETTI P, et al. DONSON facilitates Cdc45 and GINS chromatin association and is essential for DNA replication initiation [J]. *Nucleic Acids Res*, 2023, 51(18):9748-9763.
- [43] RAGEUL J, LO N, PHI A L, et al. Poly (ADP-ribosyl) ation of TIMELESS limits DNA replication stress and promotes stalled fork protection [J]. *Cell Rep*, 2024, 43(3):113845.
- [44] LEUZZI G, VASCIABEO A, TAGLIALATELA A, et al. SMARCAL1 is a dual regulator of innate immune signaling and PD-L1 expression that promotes tumor immune evasion [J]. *Cell*, 2024, 187(4):861-881.
- [45] EVRIN C, ALVAREZ V, AINSWORTH J, et al. DONSON is required for CMG helicase assembly in the mammalian cell cycle [J]. *EMBO Rep*, 2023, 24(11):e57677.
- [46] CHIANG C, SCOTT A J, DAVIS J R, et al. The impact of structural variation on human gene expression [J]. *Nat Genet*, 2017, 49(5):692-699.

(编辑 郭云雁)