

早产相关基因的挖掘与特征分析

刘玄石, 李巍

国家儿童医学中心, 首都医科大学附属北京儿童医院, 遗传与出生缺陷防治中心; 北京市儿科研究所, 出生缺陷遗传学研究北京市重点实验室; 儿科重大疾病研究教育部重点实验室, 北京 100045

摘要: 早产(preterm birth, PTB)指胎儿在完成 37 周妊娠前出生, 是新生儿死亡的主要原因, 与多种新生儿疾病和成年发生的慢性病相关。据双生子和家系研究报道, 遗传因素约占早产风险的 15%~35%, 然而早产的分子流行病学机制目前尚不明确。本研究通过挖掘文献数据库和疾病数据库中 与早产相关的文献, 并结合两重过滤的方法, 筛选出 355 个与早产相关基因。富集分析发现早产相关基因主要分子功能包括: 受体配体活性、细胞因子受体结合、细胞因子活性和生长因子活性等; 主要通路包括 KEGG 中富集的糖尿病并发症中的 AGE-RAGE 信号通路、Chagas 病和 IL-17 信号通路和 TNF 信号通路等, 以及 Reactome 中富集的多个与免疫相关的通路。早产相关基因与基因组其他基因相比较, 转录本数量有差异($\alpha = 0.1$, $P = 0.06$), 但在 GC 含量和基因长度上没有明显差异。本研究结果提示早产基因大多集中在免疫相关通路, 具备与免疫过程密切相关的分子功能, 为早产的遗传机制研究提供了重要资源。

关键词: 早产; 数据挖掘; 富集分析; 基因特征; 转录本数量

Mining and characterization of preterm birth related genes

Xuanshi Liu, Wei Li

Beijing Key Laboratory for Genetics of Birth Defects, Beijing Pediatric Research Institute; MOE Key Laboratory of Major Diseases in Children; Genetics and Birth Defects Control Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China

Abstract: Preterm birth (PTB) refers to birth before 37 completed gestational weeks. PTB is the leading cause of neonatal deaths and is associated with various neonatal complications and adult-onset chronic diseases. According to twin and family studies, genetic variants account for about 15% to 35% of the incidence of PTB. However, the molecular epidemiology of PTB is still unclear. By mining the PTB-related researches in the literature database and the disease databases, and combining two filtering methods, 355 PTB-related genes were selected. The enrichment analyses of molecular function revealed that the main functions of PTB-related genes include: receptor ligand activity, cytokine receptor binding, cytokine activity, growth factor activity, etc.; the main pathways from KEGG enrichment were the AGE-RAGE signaling pathway in diabetic complications, Chagas disease, and the IL-17 signaling pathway, the TNF

收稿日期: 2019-03-21; 修回日期: 2019-05-08

作者简介: 刘玄石, 博士研究生, 助理研究员, 专业方向: 生物信息学。E-mail: liuxs2017bioinf@163.com

通讯作者: 李巍, 博士, 教授, 博士生导师, 研究方向: 医学生物化学, 医学遗传, 细胞生物学, 产前诊断以及遗传咨询。E-mail: liwei@bch.com.cn

DOI: 10.16288/j.ycz.19-078

网络出版时间: 2019/5/10 15:23:07

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20190510.1522.002.html>

signaling pathway, etc, as well as several immune-related pathways from Reactome enrichment. There were differences in the number of transcripts between PTB-related genes and other genes in the genome ($\alpha = 0.1$, $P = 0.06$), but there was no significant difference in GC content and gene lengths. The results suggest that PTB-related genes are mostly in immune-related pathways, and have molecular functions closely related to immunity. Our work provides an important resource for the study of the genetical mechanisms of PTB.

Keywords: preterm birth; data mining; enrichment analysis; gene features; transcript number

早产是指胎儿在完成 37 周妊娠前出生。2010 年,世界卫生组织等国际组织对全世界 184 个国家的调查发现,新生儿的早产率大致是 5%~18%^[1],中国的早产率大约是 7%,每年约有 120 万早产婴儿,全球排名第二,仅低于印度^[2]。除死亡风险外,早产还可能伴有脑瘫、肺部疾病、听觉和视觉缺陷等风险^[1,2],甚至有研究发现早产与成年后发生的一些慢性疾病相关,如心血管疾病和糖尿病等^[3]。目前,早产的发生机制尚不明确。根据双生子及家系研究的估算,遗传因素对早产风险的影响大约占 15%~35%^[4~6]。早期对早产遗传机制的研究,通常根据早产病理学特点,选择可能相关的基因展开研究。例如,与新生儿出生体重和月经期有关的 *PON2*^[7],参与炎症反应的 *TNF*、*IL10*^[8]和 *TLR2*^[9],与血管生成有关的 *VEGF*^[10,11]等。近年来,采用高通量测序技术对早产遗传因素的研究,发现了大量相关的位点和基因,包括采用全基因组关联分析找到的与自发早产相关的 3 个位点(rs17053026、rs17527054 和 rs3777722)^[12],以及位于 *EBF1*、*EEFSEC* 和 *AGTR2* 基因上的与早产相关的位点^[13];利用全外显子测序发现与早产最显著相关的位点落在 *CRI* 基因外显子上^[14];全基因组、转录组和甲基化数据的结果提示 *RAB31* 和 *RBPJ* 基因与早产相关^[15]等。虽然针对早产遗传因素的研究已经积累了大量数据,然而由于早产的遗传机制相当复杂,现有研究结果也缺乏较好的归纳和整合,如 Database for Preterm Birth (dbPTB)最后一次更新是 2014 年,这使得后续采用生物信息学手段对早产遗传信息的挖掘和早产遗传模型的构建变得困难^[16]。因此,本研究利用生物信息学方法,通过挖掘文献数据库以及疾病基因数据库中报道的早产相关基因信息,整合并分析早产相

关基因的特征,为早产的遗传研究提供重要资源。

1 材料与方法

1.1 数据库和软件

(1)文献数据库:美国国家医学图书馆(PubMed, <https://www.ncbi.nlm.nih.gov/pubmed/>);(2)疾病数据库:人类孟德尔遗传数据库(OMIM, <https://www.omim.org/>, 下载时间:2019 年 1 月 18 日)、人类基因组变异数据库(ClinVar, <https://www.ncbi.nlm.nih.gov/clinvar/>, 下载时间:2019 年 2 月 11 日)以及毒物基因组学数据库(CTD, <http://ctdbase.org/>, 下载时间 2019 年 2 月 6 日);(3)基因特征数据通过 Ensembl 数据库收集(<http://grch37.ensembl.org/biomart/martview/b3df3ce0609b9d96d3347ff1d09e4348>, 数据下载时间:2019 年 3 月 10 日)。基因数据均统一使用人类参考基因组 GRCh37/hg19;(4)统计应用软件 R, 版本号 3.5.1。R 包 ClusterProfiler (版本 3.10.1)用于富集分析^[17];(5)网页版文本挖掘工具 SciMiner (<http://hurlab.med.und.edu/SciMiner/>, 使用时间:2019 年 3 月 10 日)^[18]。

1.2 文献数据库的信息挖掘

2019 年 3 月 8 日,通过计算机检索 PubMed 数据库,采用关键词检索式“preterm birth”AND “gene”,检索年限为建库至 2019 年 3 月。整理出所有文献的 PMID,输入文本挖掘工具 SciMiner。SciMiner 软件通过关键字“preterm birth”,以及软件内置的正则表达规则和基因字典,挖掘文献中与早产相关基因。为避免过度匹配,对 SciMiner 挖掘

结果设置阈值和人工审核的两层过滤方式。首先根据设置的阈值,删除了仅在2篇及以下文献中出现的基因。其次通过人工核查摘要,删除摘要中没有直接提及早产的基因。最后筛选出用于后续分析的基因列表。

1.3 疾病数据库的信息挖掘

通过Shell脚本程序,搜索疾病数据库OMIM, ClinVar和CTD,查找与“preterm birth”或其同义词匹配的记录,提取记录下的基因信息,并合并进文献数据库筛选出的基因列表。

1.4 基因富集分析

采用R软件包ClusterProfiler对筛选出的基因,进行了基因功能(Gene Ontology, GO)和KEGG通路(京都基因与基因组大百科全书数据库, Kyoto Encyclopedia of Genes and Genomes)以及Reactome通路^[19]的富集分析,对结果进行多重检验后,获得显著的功能和通路,以 $FDR < 0.05$ (false discovery rate)作为显著性的阈值。

1.5 基因特征的收集

采用Ensembl的BioMart,收集了20320个基因的长度,转录本数量,GC含量特征(人基因组版本GRCh37.p13/hg19)。根据筛选出的基因列表,采用Shell脚本程序,从BioMart数据中提取了所需基因的特征信息。

2 结果与分析

2.1 早产相关基因数据库挖掘结果

通过计算机检索PubMed数据库获得来源于800种杂志的2264篇相关文献的摘要,并通过PMID和SciMiner软件挖掘出了文献中与早产可能相关的2149个基因。其中,文献数量居前5%的杂志多数是临床专业期刊(附表1)。经过阈值和人工审核的两层过滤,筛选出在1274篇文献里出现的355个基因(附表2)表1列出了在文献数量中排名前5%的基因。

通过对疾病数据库OMIM、ClinVar和CTD的挖掘,找到1个早产相关基因(*SERPINH1*)。由于该

基因已存在于上述355个基因中,因此最终用于分析的基因数目不变。

GO富集分析发现174种显著的生物学功能($FDR < 0.05$)。根据显著性由高到低排列,前10种生物学功能包括:受体配体活性(receptor ligand activity)、细胞因子受体结合(cytokine receptor binding)、细胞因子活性(cytokine activity)、生长因子活性(growth factor activity)、生长因子结合(growth factor binding)、蛋白酶结合(protease binding)、血红素结合(heme binding)、生长因子受体结合(growth factor receptor binding)、四吡咯结合(tetrapyrrole binding)和脂多糖结合(lipopolysaccharide binding)(图1,附表3)。其中具有受体配体活性功能的基因数量最多,共有61个。

KEGG富集分析发现的显著信号通路达到158个($FDR < 0.05$)。前10条通路根据显著性由高到低排列分别是:糖尿病并发症中的AGE-RAGE信号通路(AGE-RAGE signaling pathway in diabetic complications), Chagas病(美洲锥虫病), IL-17信号通路(IL-17 signaling pathway), TNF信号通路(TNF signaling pathway), PI3K-Akt信号通路(PI3K-Akt signaling pathway), Toll样受体信号通路(Toll-like receptor signaling pathway), 结核(tuberculosis), 炎症性肠病(inflammatory bowel disease (IBD)), 乙型肝炎(hepatitis B)和流体剪切力和动脉粥样硬化(fluid shear stress and atherosclerosis)(图2,附表4)。

Reactome通路富集分析中前10个显著通路分别是:白细胞介素信号(Signaling by Interleukins), 白细胞介素4和白细胞介素-13信号传导(Interleukin-4 and Interleukin-13 signaling), 白细胞介素10信号传导(Interleukin-10 signaling), Toll样受体级联(Toll-like Receptor Cascades), Toll样受体4(TLR4)级联(Toll Like Receptor 4 (TLR4) Cascade), Toll样受体TLR1:TLR2级联(Toll Like Receptor TLR1:TLR2 Cascade), Toll样受体2(TLR2)级联(Toll Like Receptor 2 (TLR2) Cascade), 免疫系统疾病(Diseases of Immune System)与TLR信号级联相关疾病(Diseases associated with the TLR signaling cascade), 质膜上启动的MyD88:MAL (TIRAP)级联(MyD88:MAL (TIRAP) cascade initiated on plasma membrane)(图3,附表5)。

表 1 筛选出的基因列表中排前 5%的早产相关基因

Table 1 Top 5% preterm birth related genes after filtering

基因名称	基因 ID	基因名全称	有基因记录的文献数量
<i>TNF</i>	11892	Tumor necrosis factor (TNF superfamily, member 2)	156
<i>IL6</i>	6018	Interleukin 6 (Interferon, beta 2)	155
<i>IL1B</i>	5992	Interleukin 1 beta	140
<i>IL8</i>	6025	Interleukin 8	85
<i>NFKB1</i>	7794	Nuclear factor of kappa light polypeptide gene Enhancer in B-cells 1 (p105)	68
<i>COL1A1</i>	2197	Collagen type I alpha 1 chain	68
<i>PTGS2</i>	9605	Prostaglandin-endoperoxide synthase 2 (Prostaglandin G/H synthase and cyclooxygenase)	63
<i>TLR4</i>	11850	Toll-like receptor 4	57
<i>VEGFA</i>	12680	Vascular endothelial growth factor A	57
<i>IL10</i>	5962	Interleukin 10	53
<i>MT-RNR2</i>	7471	Mitochondrially encoded 16S RNA	51
<i>INS</i>	6081	Insulin	46
<i>PGR</i>	8910	Progesterone receptor	42
<i>IGF1</i>	5464	Insulin-like growth factor 1 (Somatomedin C)	39
<i>TGFB1</i>	11766	Transforming growth factor beta 1	39
<i>SFTPD</i>	10803	Surfactant, pulmonary-associated protein D	38
<i>MMP9</i>	7176	Matrix metalloproteinase 9 (Gelatinase B, 92kDa gelatinase, 92 kDa type IV collagenase)	36
<i>NR3C1</i>	7978	Nuclear receptor subfamily 3, group C, member 1 (Glucocorticoid receptor)	35
<i>SFTPA2B</i>	23441	Surfactant, pulmonary-associated protein A2B	34
<i>IL1A</i>	5991	Interleukin 1 alpha	33

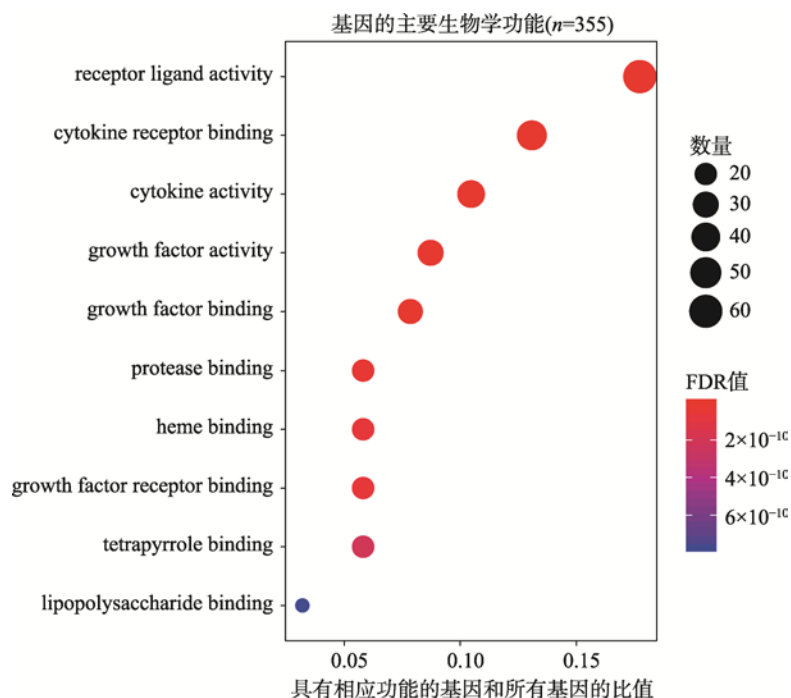


图 1 基因分子功能的 GO 富集

Fig. 1 GO enrichment analysis of molecular function in genes

颜色代表 FDR 值的大小, 由蓝色到红色 FDR 值逐渐变小, 圆点的面积代表基因的数量。

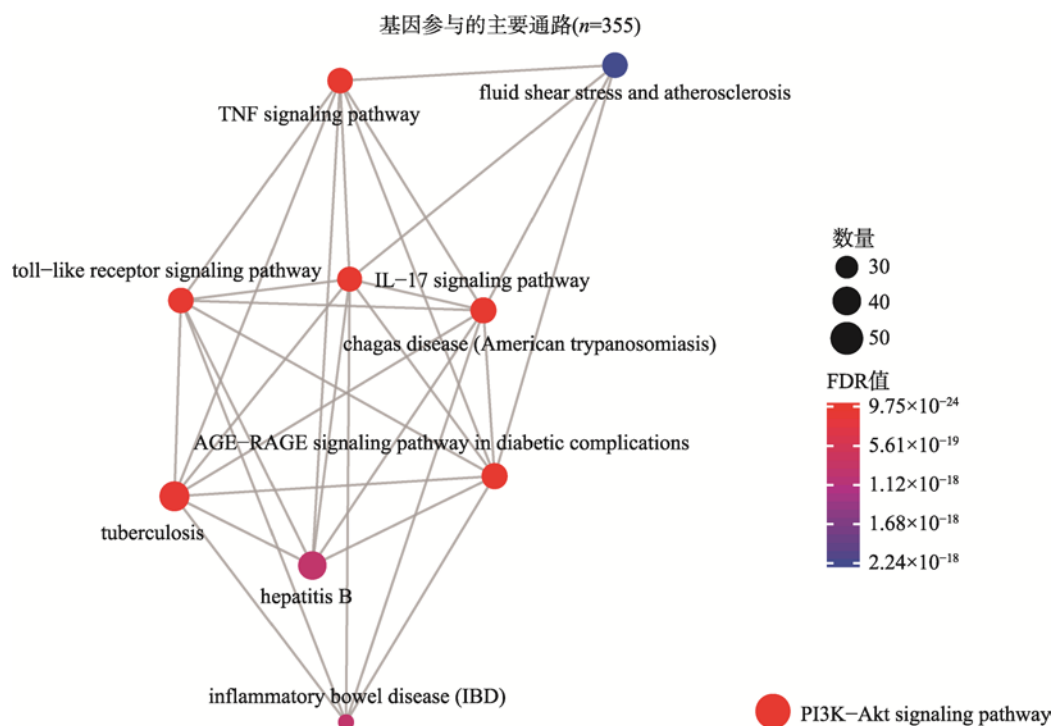


图 2 基因 KEGG 通路的富集结果

Fig. 2 KEGG enrichment analysis of genes

颜色代表 FDR 值的大小, 由蓝色到红色 FDR 值逐渐变小, 圆点的面积代表基因的数量。

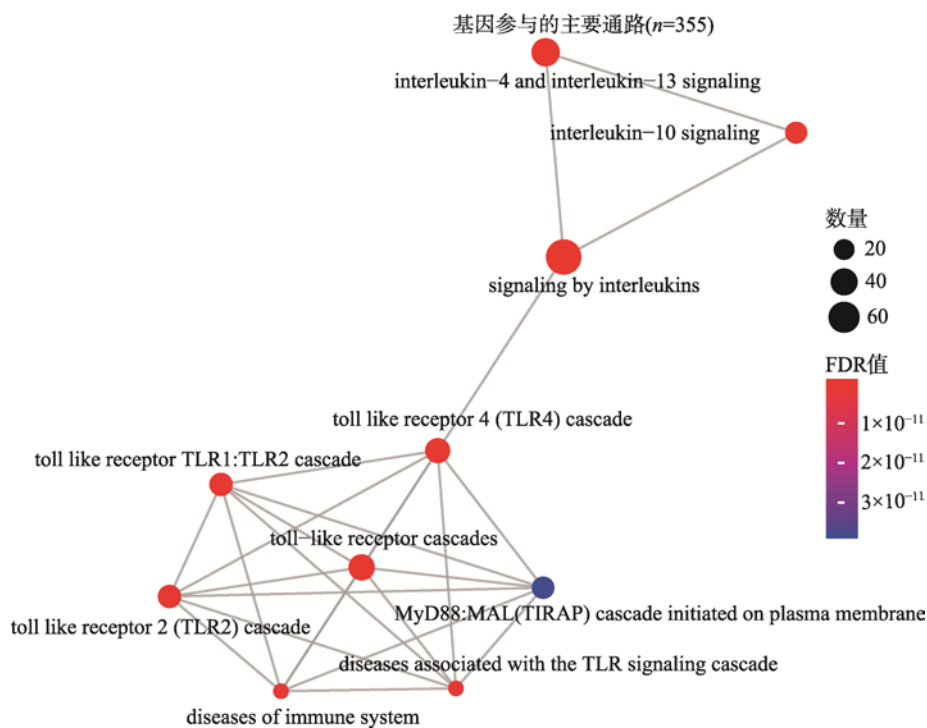


图 3 基因 Reactome 通路的富集

Fig. 3 Reactome enrichment analysis of genes

颜色代表 FDR 值的大小, 由蓝色到红色 FDR 值逐渐变小, 圆点的面积代表基因的数量。

2.2 基因特征的收集与分析结果

对比早产基因的每个基因转录本数量和全基因组每个基因的转录本数量,早产基因的转录本数量平均值(8.2)要高于全基因组基因的转录本数量平均值(7.5) (图 4A)。在显著性水平 $\alpha=0.1$ 的情况下,差异显著(t 检验: $P=0.06$)。针对 GC 含量的比较,早产基因和全基因组基因之间没有明显差异(t 检验: $P=0.70$, $\alpha=0.1$) (图 4B)。

在早产基因长度和全基因组编码蛋白的基因长度的比较中发现,早产基因的平均长度为 63 100 bp,而全基因组基因的长度平均为 61 191 bp (图 5)。在显著性水平 $\alpha=0.1$ 的情况下,差异不显著(t 检验: $P=0.73$)。

3 讨论

早产是新生儿健康研究领域的一个极其重要的研究方向。虽然关于早产发生发展的分子作用机制尚不明确,但是已有大量研究表明早产的发生与遗传有关,并已产生了大量的数据。本研究通过文本挖掘工具挖掘 PubMed 中所检索的 2264 篇早产相关

文献中的基因,结合阈值和人工审核的两层过滤以及疾病数据库记录,最终锁定 355 个早产相关基因。这是目前为止从文献中挖掘的最新的早产相关基因数据集。富集分析表明早产相关基因大多集中在免疫相关通路,基因特征分析发现早产相关基因和全基因组基因对比,GC 含量和基因长度没有差异,而转录本数量有差异。

以往的研究发现,免疫和炎症反应对维持妊娠和决定分娩时间起重要作用^[8,20,21]。其中,由于父源和母源抗原的同时存在,母胎免疫耐受的维持在妊娠期间起重要作用,而这种稳态的破坏,可能会导致早产的发生^[20]。先天免疫细胞通过释放炎症因子来影响妊娠过程和分娩时间,例如巨噬细胞释放的炎症因子可能促进催产素的产生,从而使子宫发生收缩,为分娩做准备^[22]。同时,先天免疫和获得性免疫之间的失衡,也可能导致早产发生^[23]。本研究采用挖掘得到的早产相关基因进行 KEGG 和 Reactome 富集分析,结果发现早产基因大多集中在免疫和炎症反应相关通路,这一点与以往的研究发现相吻合。先天免疫系统反映了对感染的应答作用,包括但不限于巨噬细胞、toll-like 受体、噬中性粒细胞和细胞因子等;获得性免疫系统主要是 T 淋巴细胞和 B 淋

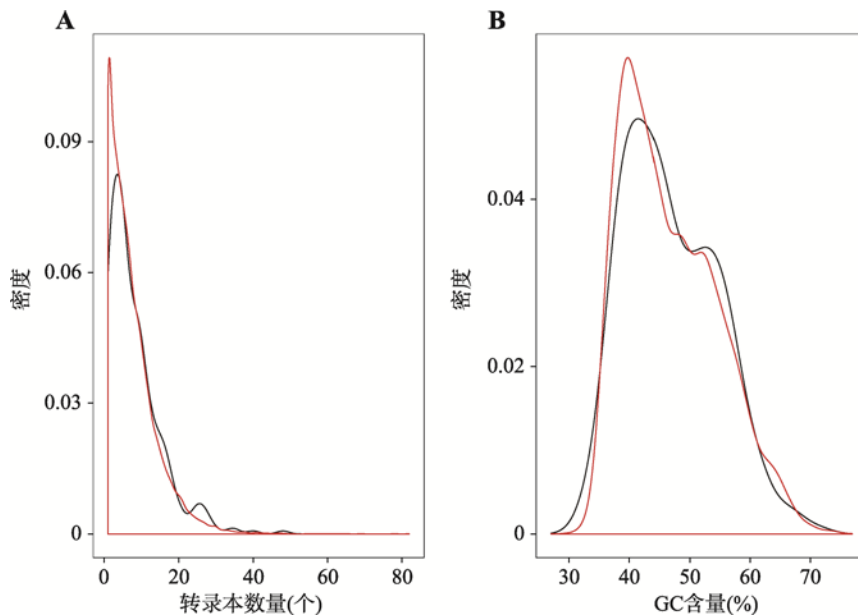


图 4 对比早产基因和全基因组基因的转录本数量以及 GC 含量

Fig. 4 Comparisons between preterm birth related genes and genes in whole genome in terms of transcript numbers and GC contents

A: 转录本数量分布(个); B: GC 含量分布(%). 红色的曲线代表全基因组,黑色的曲线代表早产基因。

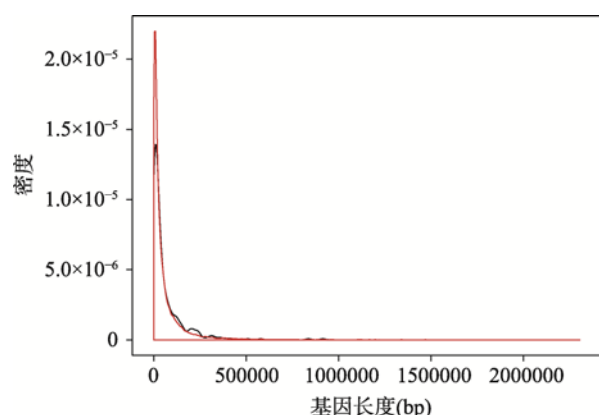


图5 对比早产基因和全基因组编码蛋白基因的长度

Fig. 5 Comparisons between preterm birth related genes and protein coding genes in whole genome in terms of gene lengths

红色的曲线代表全基因组, 黑色的曲线代表早产基因。

巴细胞^[24]。GO 富集分析的结果也体现了早产相关基因具备与免疫过程密切相关的分子功能, 包括受体配体活性、细胞因子受体活性等。本研究找到的前 20 个早产相关基因中, 大多与免疫直接或间接相关。其中研究 *TNF* 基因的文献数目最多, 研究包括胎儿肠膜发育和早产介导炎症^[25]、环境内分泌物与孕期炎症生物标志物^[26]。

据文献报道, 人类基因组可能在疾病中具备一定特征^[27,28], 如慢性阻塞性肺疾病相关的基因转录本复杂度与对照组显著不同^[29], 内源性疾病的基因编码区具有高 GC 含量^[30], 在神经发育和神经退行性疾病中发现基因的长度扮演重要角色^[31], 其中在自闭症可能的候选基因中有许多长基因^[32]。为进一步探索早产相关基因的基因组特征, 本研究对比了早产相关基因与全基因组基因在转录本数量、GC 含量和基因长度上的差异。其中, 转录本数量存在差异。有研究发现, 具有较多转录本数量的基因多为管家基因或必需基因, 在生物学上起重要作用^[33], 然而针对转录本数量较多的早产相关基因, 目前尚无文献报道。这些基因在早产所起的作用, 仍需要进一步研究。GC 含量在本研究中反映的是鸟嘌呤和胞嘧啶在每个基因中所占的比例。本研究并未发现早产相关基因与全基因组基因 GC 含量上存在显著差异。同时, 早产基因在基因长度上与全基因组的所有基因相比, 也无明显差异。

然而, 本研究也有一定的局限性。首先, 在数据库的甄选上, 挖掘文献中早产相关基因时, 也可以考虑包括中文数据库, 例如 CNKI, 可以挖掘更多与中国人早产相关的研究和相关基因。其次, 对基因的特征分析可以引入更多的变量, 如种族信息等。对不同种族的研究, 或许可以找到疾病相关且种族特异的遗传背景^[34]。

综上所述, 本研究结合文本挖掘和两层过滤方法以及疾病数据库记录, 最终锁定 355 个早产相关基因, 是截止到投稿时, 最新的早产相关基因的整合记录。富集分析表明早产相关基因大多集中在免疫相关信号通路, 基因特征分析提示了早产相关基因的转录本数量对比全基因组基因有一定差异。本研究对早产基因的挖掘和整合, 可以为早产的遗传研究提供重要资源和提示相关研究方向。

附录

附表 1~5 见文章电子版 www.chinagene.cn。

参考文献(References):

- [1] Liu L, Oza S, Hogan D, Chu Y, Perin J, Zhu J, Lawn JE, Cousens S, Mathers C, Black RE. Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the sustainable development goals. *Lancet*, 2016, 388(10063): 3027-3035. [DOI]
- [2] Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller AB, Narwal R, Adler A, Vera Garcia C, Rohde S, Say L, Lawn JE. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet*, 2012, 379(9832): 2162-2172. [DOI]
- [3] Sipola-Leppänen M, Vääräsmäki M, Tikanmäki M, Matinlinna HM, Miettola S, Hovi P, Wehkalampi K, Ruokonen A, Sundvall J, Pouta A, Eriksson JG, Järvelin MR, Kajantie E. Cardiometabolic risk factors in young adults who were born preterm. *Am J Epidemiol*, 2015, 181(11): 861-873. [DOI]
- [4] Wu W, Witherspoon DJ, Fraser A, Clark EA, Rogers A, Stoddard GJ, Manuck TA, Chen K, Esplin MS, Smith KR, Varner MW, Jorde LB. The heritability of gestational age in a two-million member cohort: implications for spontaneous

- preterm birth. *Hum Genet*, 2015, 134(7): 803–808. [DOI]
- [5] Kistka ZA, DeFranco EA, Ligthart L, Willemsen G, Plunkett J, Muglia LJ, Boomsma DI. Heritability of parturition timing: an extended twin design analysis. *Am J Obstet Gynecol*, 2008, 199(1): 43.e1–5. [DOI]
- [6] York TP, Eaves LJ, Lichtenstein P, Neale MC, Svensson A, Latendresse S, Långström N, Strauss JF 3rd. Fetal and maternal genes' influence on gestational age in a quantitative genetic analysis of 244,000 Swedish births. *Am J Epidemiol*, 2013, 178(4): 543–550. [DOI]
- [7] Liang HY, Wu BY, Chen DF, Yang F, Hu HY, Chen L, Xu XP. Association of PON2 Gene Polymorphisms in Neonates with Preterm. *Hereditas(Beijing)*, 2002, 24(5): 515–518. 梁红业, 吴白燕, 陈大方, 杨帆, 胡海燕, 陈栋, 徐希平, 新生儿 PON2 基因多态性与早产的关系. 遗传, 2002, 24(5): 515–518. [DOI]
- [8] Annells MF, Hart PH, Mullighan CG, Heatley SL, Robinson JS, Bardy P, McDonald HM. Interleukins-1, -4, -6, -10, tumor necrosis factor, transforming growth factor-beta, FAS, and mannose-binding protein C gene polymorphisms in Australian women: risk of preterm birth. *Am J Obstet Gynecol*, 2004, 191(6): 2056–2067. [DOI]
- [9] Krediet TG, Wiertsema SP, Vossers MJ, Hoeks SB, Fleer A, Ruven HJ, Rijkers GT. Toll-like receptor 2 polymorphism is associated with preterm birth. *Pediatr Res*, 2007, 62(4): 474–476. [DOI]
- [10] Papazoglou D, Galazios G, Koukourakis MI, Kontomanolis EN, Maltezos E. Association of -634G/C and 936C/T polymorphisms of the vascular endothelial growth factor with spontaneous preterm delivery. *Acta Obstet Gyn Scan*, 2004, 83(5): 461–465. [DOI]
- [11] Chen BH, Carmichael SL, Shaw GM, Iovannisci DM, Lammer EJ. Association between 49 infant gene polymorphisms and preterm delivery. *Am J Med Genet A*, 2007, 143A(17): 1990–1906. [DOI]
- [12] Zhang H, Baldwin DA, Bukowski RK, Parry S, Xu Y, Song C, Andrews WW, Saade GR, Esplin MS, Sadovsky Y, Reddy UM, Ileki J, Varner M, Biggio JR Jr. A genome-wide association study of early spontaneous preterm delivery. *Genet Epidemiol*, 2015, 39(3): 217–226. [DOI]
- [13] Zhang GB, Feenstra B, Bacelis J, Liu X, Muglia LM, Juodakis J, Miller DE, Litterman N, Jiang PP, Russell L, Hinds DA, Hu Y, Weirauch MT, Chen X, Chavan AR, Wagner GP, Pavličev M, Nnamani MC, Maziarz J, Karjalainen MK, Rämetsä M, Sengpiel V, Geller F, Boyd HA, Palotie A, Momany A, Bedell B, Ryckman KK, Huusko JM, Forney CR, Kottyan LC, Hallman M, Teramo K, Nohr EA, Davey Smith G, Melbye M, Jacobsson B, Muglia LJ. Genetic associations with gestational duration and spontaneous preterm birth. *New Engl J Med*, 2017, 377(12): 1156–1167. [DOI]
- [14] McElroy JJ, Gutman CE, Shaffer CM, Busch TD, Puttonen H, Teramo K, Murray JC, Hallman M, Muglia LJ. Maternal coding variants in complement receptor 1 and spontaneous idiopathic preterm birth. *Hum Genet*, 2013, 132(8): 935–942. [DOI]
- [15] Knijnenburg TA, Vockley JG, Chambwe N, Gibbs DL, Humphries C, Huddleston KC, Klein E, Kothiyal P, Tasseff R, Dhankani V, Bodian DL, Wong WSW, Glusman G, Mauldin DE, Miller M, Slagel J, Elasedy S, Roach JC, Kramer R, Leinonen K, Linthorst J, Baveja R, Baker R, Solomon BD, Eley G, Iyer RK, Maxwell GL, Bernard B, Shmulevich I, Hood L, Niederhuber JE. Genomic and molecular characterization of preterm birth. *Proc Natl Acad Sci USA*, 2019, 116(12): 5819–5827. [DOI]
- [16] Uzun A, Laliberte A, Parker J, Andrew C, Winterrowd E, Sharma S, Istrail S, Padbury JF. DbPTB: a database for preterm birth. *Database(Oxford)*, 2012, 2012: bar069. [DOI]
- [17] Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*, 2012, 16(5): 284–287. [DOI]
- [18] Hur J, Schuyler AD, States DJ, Feldman EL. SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics*, 2009, 25(6): 838–840. [DOI]
- [19] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. *Nucleic Acids Res*, 2018, 46(D1): D649–D655. [DOI]
- [20] Romero R, Dey SK, Fisher SJ. Preterm labor: one syndrome, many causes. *Science(New York, N.Y.)*, 2014, 345(6198): 760–765. [DOI]
- [21] Macones GA, Parry S, Elkousy M, Clothier B, Ural SH, Strauss JF 3rd. A polymorphism in the promoter region of TNF and bacterial vaginosis: preliminary evidence of gene-environment interaction in the etiology of spontaneous preterm birth. *Am J Obstet Gynecol*, 2004, 190(6): 1509–1519. [DOI]
- [22] Fang X, Wong S, Mitchell BF. Effects of LPS and IL-6 on oxytocin receptor in non-pregnant and pregnant rat uterus. *Am J Reprod Immunol*, 2000, 44(2): 65–72. [DOI]

- [23] Gomez-Lopez N, StLouis D, Lehr MA, Sanchez-Rodriguez EN, Arenas-Hernandez M. Immune cells in term and preterm labor. *Cell Mol Immunol*, 2014, 11(6): 571–581. [DOI]
- [24] Melville JM, Moss TJ. The immune consequences of preterm birth. *Front Neurosci-Switz*, 2013, 7: 79. [DOI]
- [25] Schreurs R, Baumdick ME, Sagebiel AF, Kaufmann M, Mokry M, Klarenbeek PL, Schaltenberg N, Steinert FL, van Rijn JM, Drewniak A, The SML, Bakx R, Derikx JPM, de Vries N, Corpeleijn WE, Pals ST, Gagliani N, Friese MA, Middendorp S, Nieuwenhuis EES, Reinshagen K, Geijtenbeek TBH, van Goudoever JB, Bunders MJ. Human fetal TNF- α -Cytokine-Producing CD4⁺ effector memory T cells promote intestinal development and mediate inflammation early in life. *Immunity*, 2019, 50(2): 462–476.e8. [DOI]
- [26] Ferguson KK, Cantonwine DE, Rivera-González LO, Loch-Carusio R, Mukherjee B, Anzalota Del Toro LV, Jiménez-Vélez B, Calafat AM, Ye X, Alshawabkeh AN, Cordero JF, Meeker JD. Urinary phthalate metabolite associations with biomarkers of inflammation and oxidative stress across pregnancy in Puerto Rico. *Environ Sci Technol*, 2014, 48(12): 7018–7025. [DOI]
- [27] Collins A. The genomic and functional characteristics of disease genes. *Brief Bioinform*, 2014 16(1): 16–23. [DOI]
- [28] Pengelly RJ, Vergara-Lope A, Alyousfi D, Jabalameli MR, Collins A. Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. *Brief Bioinform*, 2019, 20(1): 267–273. [DOI]
- [29] Lackey L, McArthur E, Laederach A. Increased transcript complexity in genes associated with chronic obstructive pulmonary disease. *PLoS One*, 2015, 10(10): e0140885. [DOI]
- [30] Peng Z, Uversky VN, Kurgan L. Genes encoding intrinsic disorder in Eukaryota have high GC content. *Intrinsically Disord Proteins*, 2016, 4(1): e1262225. [DOI]
- [31] Zylka MJ, Simon JM, Philpot BD. Gene length matters in neurons. *Neuron*, 2015, 86(2): 353–355. [DOI]
- [32] King IF, Yandava CN, Mabb AM, Hsiao JS, Huang HS, Pearson BL, Calabrese JM, Starmer J, Parker JS, Magnuson T, Chamberlain SJ, Philpot BD, Zylka MJ. Topoisomerases facilitate transcription of long genes linked to autism. *Nature*, 2013, 501(7465): 58–62. [DOI]
- [33] Ryu JY, Kim HU, Lee SY. Human genes with a greater number of transcript variants tend to show biological features of housekeeping and essential genes. *Mol Biosyst*, 2015, 11(10): 2798–2807. [DOI]
- [34] Rappoport N, Toung J, Hadley D, Wong RJ, Fujioka K, Reuter J, Abbott CW, Oh S, Hu D, Eng C, Huntsman S, Bodian DL, Niederhuber JE, Hong X, Zhang G, Sikora-Wohfeld W, Gignoux CR, Wang H, Oehlert J, Jelliffe-Pawlowski LL, Gould JB, Darmstadt GL, Wang X, Bustamante CD, Snyder MP, Ziv E, Patsopoulos NA, Muglia LJ, Burchard E, Shaw GM, O'Brodovich HM, Stevenson DK, Butte AJ, Sirota M. A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth. *Sci Rep*, 2018, 8(1): 226. [DOI]

(责任编辑: 方向东)

附录

附表 1 PubMed 检索结果中居前 5% 的杂志

Supplementary Table 1 Top 5% Journals from PubMed

排序	期刊	用于过滤的文献数量	排序	期刊	用于过滤的文献数量
1	<i>PLoS One</i>	109	11	<i>Mol Hum Reprod</i>	25
2	<i>Pediatr Res</i>	75	12	<i>Sci Rep</i>	22
3	<i>Am J Obstet Gynecol</i>	71	13	<i>Endocrinology</i>	22
4	<i>J Matern Fetal Neonatal Med</i>	36	14	<i>Hum Mol Genet</i>	20
5	<i>Reprod Sci</i>	35	15	<i>Am J Physiol Lung Cell Mol Physiol</i>	20
6	<i>Placenta</i>	30	16	<i>J Perinatol</i>	19
7	<i>Am J Reprod Immunol</i>	29	17	<i>Neonatology</i>	18
8	<i>Am J Med Genet A</i>	29	18	<i>Am J Pathol</i>	18
9	<i>Proc Natl Acad Sci U S A</i>	28	19	<i>J Reprod Immunol</i>	17
10	<i>Biol Reprod</i>	26	20	<i>Pediatrics</i>	16

附表 2 早产相关基因

Supplementary Table 2 Preterm birth related genes

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>TNF</i>	11892	tumor necrosis factor (TNF superfamily, member 2)	156
<i>IL6</i>	6018	interleukin 6 (interferon, beta 2)	155
<i>IL1B</i>	5992	interleukin 1, beta	140
<i>IL8</i>	6025	interleukin 8	85
<i>NFKB1</i>	7794	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)	68
<i>COL1A1</i>	2197	collagen, type I, alpha 1	68
<i>PTGS2</i>	9605	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	63
<i>TLR4</i>	11850	toll-like receptor 4	57
<i>VEGFA</i>	12680	vascular endothelial growth factor A	57
<i>IL10</i>	5962	interleukin 10	53
<i>MT-RNR2</i>	7471	mitochondrially encoded 16S RNA	51
<i>INS</i>	6081	insulin	46
<i>PGR</i>	8910	progesterone receptor	42
<i>IGF1</i>	5464	insulin-like growth factor 1 (somatomedin C)	39
<i>TGFB1</i>	11766	transforming growth factor, beta 1	39
<i>SFTPD</i>	10803	surfactant, pulmonary-associated protein D	38
<i>MMP9</i>	7176	matrix metalloproteinase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	36
<i>NR3C1</i>	7978	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	35
<i>SFTPA2B</i>	23441	surfactant, pulmonary-associated protein A2B	34
<i>IL1A</i>	5991	interleukin 1, alpha	33
<i>SFTPA1</i>	10798	surfactant, pulmonary-associated protein A1	32

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>CCL2</i>	10618	chemokine (C-C motif) ligand 2	29
<i>F2</i>	3535	coagulation factor II (thrombin)	26
<i>IL4</i>	6014	interleukin 4	25
<i>TLR2</i>	11848	toll-like receptor 2	25
<i>SFTPB</i>	10801	surfactant, pulmonary-associated protein B	24
<i>IFNG</i>	5438	interferon, gamma	24
<i>IL1RN</i>	6000	interleukin 1 receptor antagonist	24
<i>MBL2</i>	6922	mannose-binding lectin (protein C) 2, soluble (opsonic defect)	23
<i>SFTPC</i>	10802	surfactant, pulmonary-associated protein C	23
<i>OXTR</i>	8529	oxytocin receptor	23
<i>MTHFR</i>	7436	5,10-methylenetetrahydrofolate reductase (NADPH)	23
<i>MAPK1</i>	6871	mitogen-activated protein kinase 1	23
<i>NOS2A</i>	7873	nitric oxide synthase 2A (inducible, hepatocytes)	22
<i>ACE</i>	2707	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	21
<i>REN</i>	9958	renin	21
<i>CRH</i>	2355	corticotropin releasing hormone	20
<i>ALB</i>	399	albumin	20
<i>CYP1A1</i>	2595	cytochrome P450, family 1, subfamily A, polypeptide 1	18
<i>MMP1</i>	7155	matrix metalloproteinase 1 (interstitial collagenase)	18
<i>GSTT1</i>	4641	glutathione S-transferase theta 1	17
<i>GJA1</i>	4274	gap junction protein, alpha 1, 43kDa	17
<i>CD14</i>	1628	CD14 molecule	17
<i>CASP3</i>	1504	caspase 3, apoptosis-related cysteine peptidase	17
<i>APOE</i>	613	apolipoprotein E	16
<i>NOS3</i>	7876	nitric oxide synthase 3 (endothelial cell)	16
<i>F5</i>	3542	coagulation factor V (proaccelerin, labile factor)	16
<i>JUN</i>	6204	jun oncogene	15
<i>IGF2</i>	5466	insulin-like growth factor 2 (somatomedin A)	15
<i>LEP</i>	6553	leptin	15
<i>BCL2</i>	990	B-cell CLL/lymphoma 2	15
<i>GAPDH</i>	4141	glyceraldehyde-3-phosphate dehydrogenase	15
<i>FOS</i>	3796	v-fos FBJ murine osteosarcoma viral oncogene homolog	15
<i>MMP2</i>	7166	matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	15
<i>SERPINH1</i>	1546	serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1)	14
<i>FLT1</i>	3763	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	14
<i>NFKBIA</i>	7797	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	14
<i>GSTM1</i>	4632	glutathione S-transferase M1	13
<i>SERPINE1</i>	8583	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	13

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>IL6R</i>	6019	interleukin 6 receptor	13
<i>TP53</i>	11998	tumor protein p53	13
<i>TWIST1</i>	12428	twist homolog 1 (acrocephalosyndactyly 3; Saethre-Chotzen syndrome) (<i>Drosophila</i>)	13
<i>SOD1</i>	11179	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))	13
<i>IL2</i>	6001	interleukin 2	13
<i>CD4</i>	1678	CD4 molecule	13
<i>AGT</i>	333	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	12
<i>PPARG</i>	9236	peroxisome proliferator-activated receptor gamma	12
<i>CAT</i>	1516	catalase	12
<i>CYP2B6</i>	2615	cytochrome P450, family 2, subfamily B, polypeptide 6	12
<i>PGF</i>	8893	placental growth factor	11
<i>S100A9</i>	10499	S100 calcium binding protein A9	11
<i>GSTA1</i>	4626	glutathione S-transferase A1	11
<i>KDR</i>	6307	kinase insert domain receptor (a type III receptor tyrosine kinase)	11
<i>PRKCA</i>	9393	protein kinase C, alpha	11
<i>STAT1</i>	11362	signal transducer and activator of transcription 1, 91kDa	10
<i>MBP</i>	6925	myelin basic protein	10
<i>IL13</i>	5973	interleukin 13	10
<i>EDN1</i>	3176	endothelin 1	10
<i>LTA</i>	6709	lymphotoxin alpha (TNF superfamily, member 1)	10
<i>TFAP2A</i>	11742	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)	10
<i>TLR5</i>	11851	toll-like receptor 5	10
<i>TBXAS1</i>	11609	thromboxane A synthase 1 (platelet, cytochrome P450, family 5, subfamily A)	10
<i>IL1R1</i>	5993	interleukin 1 receptor, type I	10
<i>CYP3A5</i>	2638	cytochrome P450, family 3, subfamily A, polypeptide 5	9
<i>IGFBP1</i>	5469	insulin-like growth factor binding protein 1	9
<i>EGR1</i>	3238	early growth response 1	9
<i>FAS</i>	11920	Fas (TNF receptor superfamily, member 6)	9
<i>ADRB2</i>	286	adrenergic, beta-2-, receptor, surface	9
<i>MMP8</i>	7175	matrix metalloproteinase 8 (neutrophil collagenase)	9
<i>PTGER4</i>	9596	prostaglandin E receptor 4 (subtype EP4)	8
<i>CSF3</i>	2438	colony stimulating factor 3 (granulocyte)	8
<i>TNFRSF1A</i>	11916	tumor necrosis factor receptor superfamily, member 1A	8
<i>NES</i>	7756	nestin	8
<i>FGFR3</i>	3690	fibroblast growth factor receptor 3 (achondroplasia, thanatophoric dwarfism)	8
<i>MAPK3</i>	6877	mitogen-activated protein kinase 3	8
<i>COX8A</i>	2294	cytochrome c oxidase subunit 8A (ubiquitous)	8
<i>ESR2</i>	3468	estrogen receptor 2 (ER beta)	8
<i>PPARA</i>	9232	peroxisome proliferator-activated receptor alpha	8
<i>FSHR</i>	3969	follicle stimulating hormone receptor	8

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>MAPK14</i>	6876	mitogen-activated protein kinase 14	8
<i>RPS27A</i>	10417	ribosomal protein S27a	8
<i>H19</i>	4713	H19, imprinted maternally expressed transcript	8
<i>SP1</i>	11205	Sp1 transcription factor	7
<i>SOD2</i>	11180	superoxide dismutase 2, mitochondrial	7
<i>MT-CO2</i>	7421	mitochondrially encoded cytochrome c oxidase II	7
<i>IL17A</i>	5981	interleukin 17A	7
<i>IGFBP3</i>	5472	insulin-like growth factor binding protein 3	7
<i>PTGER3</i>	9595	prostaglandin E receptor 3 (subtype EP3)	7
<i>IRF6</i>	6121	interferon regulatory factor 6	7
<i>MYD88</i>	7562	myeloid differentiation primary response gene (88)	7
<i>PLAT</i>	9051	plasminogen activator, tissue	7
<i>ICAM1</i>	5344	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor	7
<i>MAPK8</i>	6881	mitogen-activated protein kinase 8	7
<i>MMP3</i>	7173	matrix metalloproteinase 3 (stromelysin 1, progelatinase)	7
<i>HSD11B2</i>	5209	hydroxysteroid (11-beta) dehydrogenase 2	7
<i>CD8A</i>	1706	CD8a molecule	7
<i>SLC6A3</i>	11049	solute carrier family 6 (neurotransmitter transporter, dopamine), member 3	7
<i>SLC12A1</i>	10910	solute carrier family 12 (sodium/potassium/chloride transporters), member 1	6
<i>NFE2L2</i>	7782	nuclear factor (erythroid-derived 2)-like 2	6
<i>HPGD</i>	5154	hydroxyprostaglandin dehydrogenase 15-(NAD)	6
<i>PTGER2</i>	9594	prostaglandin E receptor 2 (subtype EP2), 53kDa	6
<i>ABCA3</i>	33	ATP-binding cassette, sub-family A (ABC1), member 3	6
<i>SMN1</i>	11117	survival of motor neuron 1, telomeric	6
<i>CXCL10</i>	10637	chemokine (C-X-C motif) ligand 10	6
<i>DEFB1</i>	2766	defensin, beta 1	6
<i>LPAL2</i>	21210	lipoprotein, Lp(a)-like 2	6
<i>NOS1</i>	7872	nitric oxide synthase 1 (neuronal)	6
<i>FGFR1</i>	3688	fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)	6
<i>CASP1</i>	1499	caspase 1, apoptosis-related cysteine peptidase (interleukin 1, beta, convertase)	6
<i>AR</i>	644	androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)	6
<i>ATM</i>	795	ataxia telangiectasia mutated	6
<i>ZMPSTE24</i>	12877	zinc metalloproteinase (STE24 homolog, <i>S. cerevisiae</i>)	6
<i>CXCL1</i>	4602	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	6
<i>NDP</i>	7678	Norrie disease (pseudoglioma)	6
<i>TLR3</i>	11849	toll-like receptor 3	6
<i>FLG</i>	3748	filaggrin	5
<i>SLC6A4</i>	11050	solute carrier family 6 (neurotransmitter transporter, serotonin), member 4	5
<i>RUNX2</i>	10472	runt-related transcription factor 2	5

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>ABCB1</i>	40	ATP-binding cassette, sub-family B (MDR/TAP), member 1	5
<i>NR3C2</i>	7979	nuclear receptor subfamily 3, group C, member 2	5
<i>EGFR</i>	3236	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	5
<i>LOR</i>	6663	loricrin	5
<i>HDAC9</i>	14065	histone deacetylase 9	5
<i>TNFRSF1B</i>	11917	tumor necrosis factor receptor superfamily, member 1B	5
<i>EPO</i>	3415	erythropoietin	5
<i>NOD2</i>	5331	nucleotide-binding oligomerization domain containing 2	5
<i>LEPR</i>	6554	leptin receptor	5
<i>CTNNB1</i>	2514	catenin (cadherin-associated protein), beta 1, 88kDa	5
<i>THBS1</i>	11785	thrombospondin 1	5
<i>TNFAIP3</i>	11896	tumor necrosis factor, alpha-induced protein 3	5
<i>S100A6</i>	10496	S100 calcium binding protein A6	5
<i>TGFB2</i>	11768	transforming growth factor, beta 2	5
<i>IL5</i>	6016	interleukin 5 (colony-stimulating factor, eosinophil)	5
<i>SLC2A4</i>	11009	solute carrier family 2 (facilitated glucose transporter), member 4	5
<i>ACPP</i>	125	acid phosphatase, prostate	5
<i>TCEAL1</i>	11616	transcription elongation factor A (SII)-like 1	5
<i>COL1A2</i>	2198	collagen, type I, alpha 2	5
<i>CTGF</i>	2500	connective tissue growth factor	5
<i>F2R</i>	3537	coagulation factor II (thrombin) receptor	5
<i>CD163</i>	1631	CD163 molecule	5
<i>JAG1</i>	6188	jagged 1 (Alagille syndrome)	5
<i>IL12A</i>	5969	interleukin 12A (natural killer cell stimulatory factor 1, cytotoxic lymphocyte maturation factor 1, p35)	5
<i>TIRAP</i>	17192	toll-interleukin 1 receptor (TIR) domain containing adaptor protein	5
<i>FOXP3</i>	6106	forkhead box P3	5
<i>MEST</i>	7028	mesoderm specific transcript homolog (mouse)	5
<i>CFH</i>	4883	complement factor H	5
<i>IRAK1</i>	6112	interleukin-1 receptor-associated kinase 1	5
<i>PRKAR2A</i>	9391	protein kinase, cAMP-dependent, regulatory, type II, alpha	5
<i>TIMP2</i>	11821	TIMP metalloproteinase inhibitor 2	5
<i>CDKN1C</i>	1786	cyclin-dependent kinase inhibitor 1C (p57, Kip2)	4
<i>GORASP1</i>	16769	golgi reassembly stacking protein 1, 65kDa	4
<i>HLA-G</i>	4964	major histocompatibility complex, class I, G	4
<i>PON1</i>	9204	paraoxonase 1	4
<i>RAF1</i>	9829	v-raf-1 murine leukemia viral oncogene homolog 1	4
<i>PTPN11</i>	9644	protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1)	4
<i>LCN2</i>	6526	lipocalin 2	4
<i>CALCA</i>	1437	calcitonin-related polypeptide alpha	4

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>KCNH2</i>	6251	potassium voltage-gated channel, subfamily H (eag-related), member 2	4
<i>TIMP1</i>	11820	TIMP metalloproteinase inhibitor 1	4
<i>GPX1</i>	4553	glutathione peroxidase 1	4
<i>SERPINB2</i>	8584	serpin peptidase inhibitor, clade B (ovalbumin), member 2	4
<i>NLRP3</i>	16400	NLR family, pyrin domain containing 3	4
<i>MIF</i>	7097	macrophage migration inhibitory factor (glycosylation-inhibiting factor)	4
<i>IL1R2</i>	5994	interleukin 1 receptor, type II	4
<i>ERAL1</i>	3424	Era G-protein-like 1 (E. coli)	4
<i>IFNA1</i>	5417	interferon, alpha 1	4
<i>PLAGL1</i>	9046	pleiomorphic adenoma gene-like 1	4
<i>CYP27B1</i>	2606	cytochrome P450, family 27, subfamily B, polypeptide 1	4
<i>ZEB1</i>	11642	zinc finger E-box binding homeobox 1	4
<i>CXCL12</i>	10672	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)	4
<i>LBP</i>	6517	lipopolysaccharide binding protein	4
<i>WNT4</i>	12783	wingless-type MMTV integration site family, member 4	4
<i>IL4R</i>	6015	interleukin 4 receptor	4
<i>INSR</i>	6091	insulin receptor	4
<i>MAPK10</i>	6872	mitogen-activated protein kinase 10	4
<i>DES</i>	2770	desmin	4
<i>PHEX</i>	8918	phosphate regulating endopeptidase homolog, X-linked (hypophosphatemia, vitamin D resistant rickets)	4
<i>PTPRC</i>	9666	protein tyrosine phosphatase, receptor type, C	4
<i>SLC26A4</i>	8818	solute carrier family 26, member 4	4
<i>TEK</i>	11724	TEK tyrosine kinase, endothelial (venous malformations, multiple cutaneous and mucosal)	4
<i>TLR6</i>	16711	toll-like receptor 6	4
<i>TSHB</i>	12372	thyroid stimulating hormone, beta	4
<i>CCL3</i>	10627	chemokine (C-C motif) ligand 3	4
<i>CYP17A1</i>	2593	cytochrome P450, family 17, subfamily A, polypeptide 1	4
<i>CYP19A1</i>	2594	cytochrome P450, family 19, subfamily A, polypeptide 1	4
<i>FSHB</i>	3964	follicle stimulating hormone, beta polypeptide	4
<i>IL10RA</i>	5964	interleukin 10 receptor, alpha	4
<i>VIM</i>	12692	vimentin	4
<i>ADAMTS2</i>	218	ADAM metalloproteinase with thrombospondin type 1 motif, 2	4
<i>ADAMTS4</i>	220	ADAM metalloproteinase with thrombospondin type 1 motif, 4	4
<i>ATP2A3</i>	813	ATPase, Ca ⁺⁺ transporting, ubiquitous	4
<i>CHRNA9</i>	14079	cholinergic receptor, nicotinic, alpha 9	4
<i>COL2A1</i>	2200	collagen, type II, alpha 1	4
<i>COL5A1</i>	2209	collagen, type V, alpha 1	4
<i>HBG2</i>	4832	hemoglobin, gamma G	4
<i>NOX5</i>	14874	NADPH oxidase, EF-hand calcium binding domain 5	4

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>RELA</i>	9955	v-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3, p65 (avian)	4
<i>TF</i>	11740	transferrin	4
<i>TLR10</i>	15634	toll-like receptor 10	4
<i>PLCB1</i>	15917	phospholipase C, beta 1 (phosphoinositide-specific)	4
<i>MASP2</i>	6902	mannan-binding lectin serine peptidase 2	3
<i>CYP3A4</i>	2637	cytochrome P450, family 3, subfamily A, polypeptide 4	3
<i>GHRL</i>	18129	ghrelin/obestatin preprohormone	3
<i>GJB2</i>	4284	gap junction protein, beta 2, 26kDa	3
<i>BGN</i>	1044	biglycan	3
<i>GHR</i>	4263	growth hormone receptor	3
<i>NEU1</i>	7758	sialidase 1 (lysosomal sialidase)	3
<i>PSEN1</i>	9508	presenilin 1 (Alzheimer disease 3)	3
<i>SMAD7</i>	6773	SMAD family member 7	3
<i>CAMP</i>	1472	cathelicidin antimicrobial peptide	3
<i>DEFB4</i>	2767	defensin, beta 4	3
<i>IGF1R</i>	5465	insulin-like growth factor 1 receptor	3
<i>CAP1</i>	20040	CAP, adenylate cyclase-associated protein 1 (yeast)	3
<i>GDF9</i>	4224	growth differentiation factor 9	3
<i>PHOX2B</i>	9143	paired-like homeobox 2b	3
<i>CCL8</i>	10635	chemokine (C-C motif) ligand 8	3
<i>KCNB1</i>	6231	potassium voltage-gated channel, Shab-related subfamily, member 1	3
<i>SLC27A4</i>	10998	solute carrier family 27 (fatty acid transporter), member 4	3
<i>HMGB1</i>	4983	high-mobility group box 1	3
<i>FASLG</i>	11936	Fas ligand (TNF superfamily, member 6)	3
<i>FZD4</i>	4042	frizzled homolog 4 (Drosophila)	3
<i>TXN</i>	12435	thioredoxin	3
<i>MAP2</i>	6839	microtubule-associated protein 2	3
<i>NGF</i>	7808	nerve growth factor (beta polypeptide)	3
<i>PROK1</i>	18454	prokineticin 1	3
<i>COMT</i>	2228	catechol-O-methyltransferase	3
<i>FOXO1</i>	3819	forkhead box O1	3
<i>FOXO3</i>	3821	forkhead box O3	3
<i>HGF</i>	4893	hepatocyte growth factor (hepapoietin A; scatter factor)	3
<i>KISS1</i>	6341	KISS-1 metastasis-suppressor	3
<i>TYMS</i>	12441	thymidylate synthetase	3
<i>RELB</i>	9956	v-rel reticuloendotheliosis viral oncogene homolog B, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3 (avian)	3
<i>UGT1A1</i>	12530	UDP glucuronosyltransferase 1 family, polypeptide A1	3
<i>CDKN2A</i>	1787	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	3
<i>CYP21A2</i>	2600	cytochrome P450, family 21, subfamily A, polypeptide 2	3

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>GATA6</i>	4174	GATA binding protein 6	3
<i>ITGB4</i>	6158	integrin, beta 4	3
<i>S100A8</i>	10498	S100 calcium binding protein A8	3
<i>SOD3</i>	11181	superoxide dismutase 3, extracellular	3
<i>CD68</i>	1693	CD68 molecule	3
<i>MIRN210</i>	31587	microRNA 210	3
<i>PPT1</i>	9325	palmitoyl-protein thioesterase 1 (ceroid-lipofuscinosis, neuronal 1, infantile)	3
<i>ZEB2</i>	14881	zinc finger E-box binding homeobox 2	3
<i>ABCA1</i>	29	ATP-binding cassette, sub-family A (ABC1), member 1	3
<i>CREBBP</i>	2348	CREB binding protein (Rubinstein-Taybi syndrome)	3
<i>P2RX7</i>	8537	purinergic receptor P2X, ligand-gated ion channel, 7	3
<i>UCP2</i>	12518	uncoupling protein 2 (mitochondrial, proton carrier)	3
<i>CACNA1G</i>	1394	calcium channel, voltage-dependent, T type, alpha 1G subunit	3
<i>MARK2</i>	3332	MAP/microtubule affinity-regulating kinase 2	3
<i>SLC27A1</i>	10995	solute carrier family 27 (fatty acid transporter), member 1	3
<i>TFRC</i>	11763	transferrin receptor (p90, CD71)	3
<i>TICAM1</i>	18348	toll-like receptor adaptor molecule 1	3
<i>FADS2</i>	3575	fatty acid desaturase 2	3
<i>FGF7</i>	3685	fibroblast growth factor 7 (keratinocyte growth factor)	3
<i>OPRM1</i>	8156	opioid receptor, mu 1	3
<i>SPAG8</i>	14105	sperm associated antigen 8	3
<i>CRHR1</i>	2357	corticotropin releasing hormone receptor 1	3
<i>DICER1</i>	17098	dicer 1, ribonuclease type III	3
<i>FTO</i>	24678	fat mass and obesity associated	3
<i>MMP12</i>	7158	matrix metalloproteinase 12 (macrophage elastase)	3
<i>CAV1</i>	1527	caveolin 1, caveolae protein, 22kDa	3
<i>ECE1</i>	3146	endothelin converting enzyme 1	3
<i>FBN1</i>	3603	fibrillin 1	3
<i>FST</i>	3971	folistatin	3
<i>IL9</i>	6029	interleukin 9	3
<i>MT-RNR1</i>	7470	mitochondrially encoded 12S RNA	3
<i>SCNN1A</i>	10599	sodium channel, nonvoltage-gated 1 alpha	3
<i>SDHC</i>	10682	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa	3
<i>WT1</i>	12796	Wilms tumor 1	3
<i>ATG16L1</i>	21498	ATG16 autophagy related 16-like 1 (<i>S. cerevisiae</i>)	3
<i>CASP8</i>	1509	caspase 8, apoptosis-related cysteine peptidase	3
<i>FURIN</i>	8568	furin (paired basic amino acid cleaving enzyme)	3
<i>FUT1</i>	4012	fucosyltransferase 1 (galactoside 2-alpha-L-fucosyltransferase, H blood group)	3
<i>HBA2</i>	4824	hemoglobin, alpha 2	3
<i>IFNB1</i>	5434	interferon, beta 1, fibroblast	3

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>KRT5</i>	6442	keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types)	3
<i>OPRD1</i>	8153	opioid receptor, delta 1	3
<i>OXT</i>	8528	oxytocin, prepro- (neurophysin I)	3
<i>PDGFRA</i>	8803	platelet-derived growth factor receptor, alpha polypeptide	3
<i>SERPINC1</i>	775	serpin peptidase inhibitor, clade C (antithrombin), member 1	3
<i>SLC2A1</i>	11005	solute carrier family 2 (facilitated glucose transporter), member 1	3
<i>CSH1</i>	2440	chorionic somatomammotropin hormone 1 (placental lactogen)	3
<i>ETS1</i>	3488	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	3
<i>HDAC1</i>	4852	histone deacetylase 1	3
<i>ITGA6</i>	6142	integrin, alpha 6	3
<i>MET</i>	7029	met proto-oncogene (hepatocyte growth factor receptor)	3
<i>TRAF1</i>	12031	TNF receptor-associated factor 1	3
<i>CD1D</i>	1637	CD1d molecule	3
<i>CYP2E1</i>	2631	cytochrome P450, family 2, subfamily E, polypeptide 1	3
<i>DLL4</i>	2910	delta-like 4 (Drosophila)	3
<i>FAM129B</i>	25282	family with sequence similarity 129, member B	3
<i>IRS1</i>	6125	insulin receptor substrate 1	3
<i>LTF</i>	6720	lactotransferrin	3
<i>MTRR</i>	7473	5-methyltetrahydrofolate-homocysteine methyltransferase reductase	3
<i>PLOD1</i>	9081	procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 1	3
<i>VIP</i>	12693	vasoactive intestinal peptide	3
<i>ABCA12</i>	14637	ATP-binding cassette, sub-family A (ABC1), member 12	3
<i>ABCC2</i>	53	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	3
<i>ABO</i>	79	ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase)	3
<i>ADCY10</i>	21285	adenylate cyclase 10 (soluble)	3
<i>APLP2</i>	598	amyloid beta (A4) precursor-like protein 2	3
<i>CCND1</i>	1582	cyclin D1	3
<i>CD69</i>	1694	CD69 molecule	3
<i>COL5A2</i>	2210	collagen, type V, alpha 2	3
<i>DLL1</i>	2908	delta-like 1 (Drosophila)	3
<i>EPAS1</i>	3374	endothelial PAS domain protein 1	3
<i>HAS2</i>	4819	hyaluronan synthase 2	3
<i>IL12RB1</i>	5971	interleukin 12 receptor, beta 1	3
<i>MDK</i>	6972	midkine (neurite growth-promoting factor 2)	3
<i>PLG</i>	9071	plasminogen	3
<i>PRKAA2</i>	9377	protein kinase, AMP-activated, alpha 2 catalytic subunit	3
<i>S100A10</i>	10487	S100 calcium binding protein A10	3
<i>SIRT1</i>	14929	sirtuin (silent mating type information regulation 2 homolog) 1 (<i>S. cerevisiae</i>)	3
<i>TPO</i>	12015	thyroid peroxidase	3

续附表

基因名称	基因 ID	基因全称	有基因记录的文献数
<i>VCAM1</i>	12663	vascular cell adhesion molecule 1	3
<i>AIF1</i>	352	allograft inflammatory factor 1	3
<i>BAX</i>	959	BCL2-associated X protein	3
<i>CCND2</i>	1583	cyclin D2	3
<i>CD27</i>	11922	CD27 molecule	3
<i>CDH17</i>	1756	cadherin 17, LI cadherin (liver-intestine)	3
<i>COL4A3</i>	2204	collagen, type IV, alpha 3 (Goodpasture antigen)	3
<i>CREB1</i>	2345	cAMP responsive element binding protein 1	3
<i>CXCL9</i>	7098	chemokine (C-X-C motif) ligand 9	3
<i>CYCS</i>	19986	cytochrome c, somatic	3
<i>HSD11B1</i>	5208	hydroxysteroid (11-beta) dehydrogenase 1	3
<i>MMP13</i>	7159	matrix metalloproteinase 13 (collagenase 3)	3
<i>MSMB</i>	7372	microseminoprotein, beta-	3
<i>NCAM1</i>	7656	neural cell adhesion molecule 1	3
<i>NCOA1</i>	7668	nuclear receptor coactivator 1	3
<i>NEFL</i>	7739	neurofilament, light polypeptide 68kDa	3
<i>NTRK2</i>	8032	neurotrophic tyrosine kinase, receptor, type 2	3
<i>PARP1</i>	270	poly (ADP-ribose) polymerase family, member 1	3
<i>PYCARD</i>	16608	PYD and CARD domain containing	3
<i>RARA</i>	9864	retinoic acid receptor, alpha	3
<i>RXRA</i>	10477	retinoid X receptor, alpha	3

附表 3 早产相关基因分子功能的 GO 富集

Supplementary Table 3 GO enrichment result of preterm related genes in molecular functions

分子功能	基因数量	P 值	FDR
receptor ligand activity	61	2.34*10 ⁻³²	1.03*10 ⁻²⁹
cytokine receptor binding	45	3.51*10 ⁻²⁸	7.77*10 ⁻²⁶
cytokine activity	36	4.29*10 ⁻²³	6.32*10 ⁻²¹
growth factor activity	30	1.23*10 ⁻²⁰	1.36*10 ⁻¹⁸
growth factor binding	27	1.03*10 ⁻¹⁹	9.14*10 ⁻¹⁸
protease binding	20	4.11*10 ⁻¹³	3.03*10 ⁻¹¹
heme binding	20	1.02*10 ⁻¹²	6.44*10 ⁻¹¹
growth factor receptor binding	20	1.18*10 ⁻¹²	6.52*10 ⁻¹¹
tetrapyrrole binding	20	4.17*10 ⁻¹²	2.05*10 ⁻¹⁰
lipopolysaccharide binding	11	1.75*10 ⁻¹¹	7.75*10 ⁻¹⁰

附表 4 早产相关基因 KEGG 通路富集

Supplementary Table 4 KEGG enrichment result of preterm related genes

通路	基因数量	<i>P</i> 值	FDR
AGE-RAGE signaling pathway in diabetic complications	35	1.02×10^{-25}	9.75×10^{-24}
Chagas disease (American trypanosomiasis)	35	3.26×10^{-25}	1.56×10^{-23}
IL-17 signaling pathway	33	1.63×10^{-24}	5.19×10^{-23}
TNF signaling pathway	34	5.68×10^{-23}	1.36×10^{-21}
PI3K-Akt signaling pathway	57	2.78×10^{-22}	5.33×10^{-21}
Toll-like receptor signaling pathway	32	1.33×10^{-21}	2.13×10^{-20}
Tuberculosis	40	4.46×10^{-21}	6.10×10^{-20}
Inflammatory bowel disease (IBD)	25	8.31×10^{-20}	9.07×10^{-19}
Hepatitis B	37	8.53×10^{-20}	9.07×10^{-19}
Fluid shear stress and atherosclerosis	34	2.34×10^{-19}	2.24×10^{-18}

附表 5 早产相关基因 Reactome 通路富集

Supplementary Table 5 Reactome enrichment result of preterm related genes

通路	基因数量	<i>P</i> 值	FDR
Signaling by Interleukins	78	2.15×10^{-37}	1.47×10^{-34}
Interleukin-4 and Interleukin-13 signaling	40	2.20×10^{-33}	7.54×10^{-31}
Interleukin-10 signaling	20	1.26×10^{-18}	2.87×10^{-16}
Toll-like Receptor Cascades	32	3.57×10^{-18}	6.12×10^{-16}
Toll Like Receptor 4 (TLR4) Cascade	28	1.53×10^{-16}	2.09×10^{-14}
Toll Like Receptor TLR1:TLR2 Cascade	23	1.18×10^{-14}	1.15×10^{-12}
Toll Like Receptor 2 (TLR2) Cascade	23	1.18×10^{-14}	1.15×10^{-12}
Diseases of Immune System	13	2.88×10^{-14}	2.19×10^{-12}
Diseases associated with the TLR signaling cascade	13	2.88×10^{-14}	2.19×10^{-12}
MyD88:MAL(TIRAP) cascade initiated on plasma membrane	34	6.15×10^{-13}	3.83×10^{-11}