卵巢早衰的基因拷贝数变异与卵巢高表达长非编码 RNA 分析

完成人
任路瑶

指导小组成员

张锋 教授
石乐明 教授

目 录

摘 要

卵巢早衰（Premature Ovarian Failure, POF）是指女性在 40 岁以前出现的伴有促性腺激素分泌过多的卵巢功能衰退，可导致闭经甚至不孕不育。POF 的发病率约为 1%，致病因素异质性较强，其中遗传因素被认为是导致 POF 的重要原因之一。本课题首先运用比较基因组杂交芯片（array-based Comparative Genomic Hybridization, aCGH）对 21 例 POF 的散发病例、2 对姐妹进行基因拷贝数变异（Copy Number Variations，CNVs）的筛查。在其中两例散发病患中各发现了一个不同的 CNV 缺失，分别影响到两个长非编码 RNA（long noncoding RNA，lncRNA）ENSG00000233967 和 ENSG00000253671。从 Illumina Human BodyMap2 数据集中发现，这两个 lncRNA 在正常卵巢中是高表达的，其表达水平比在其他组织中高出两倍以上，呈现卵巢特异性高表达的特征，预示其对卵巢功能的重要性。此外，为了排除实验误差、测序误差和人个体之间的差异，本课题进一步对 GTEx 数据集进行了分析，得知这两个 lncRNA 确实在卵巢中表达较高，具有特异性。本课题的结果提示，lncRNA ENSG00000233967 和 ENSG00000253671 在卵巢正常功能中发挥作用，其缺失可能是两例 POF 的重要致病原因，后续工作应该集中研究这两个 lncRNA 的生物学功能。

关键词：卵巢早衰，基因拷贝数变异，RNA-seq，组织特异性表达，长非编码 RNA

Abstract

Premature ovarian failure (POF), causing amenorrhoea due to cessation of ovarian function before the age of 40 years, occurs in 1% of woman. People with POF suffer from anovulation, hypoestrogenism, infertility etc. POF is heterogeneous, with a wide spectrum of causes, but with a significant genetic contribution. In this research, 21 idiopathic cases and 2 pairs of sisters have been studied, employing Agilent array-based Comparative Genomic Hybridization (aCGH) microarray to find rare copy number variations (CNVs). Two CNVs hitting on two different long noncoding RNAs (lncRNA) have been found in two idiopathic patients, which are highly expressed in normal ovary. In the RNA-seq dataset of Illumina Human BodyMap 2 Project, these two lncRNAs are highly specific expressed in ovary. They are expressed more than 2-times higher in ovary, compared with other organs. In order to exclude experimental and sequencing errors, and inter-individual differences, we found that these two lncRNAs were highly expressed in ovary in the GTEx Project dataset. The results of this study show that ENSG00000233967 and ENSG00000253671 are important in maintaining ovarian normal functions, and CNV loss of these two lncRNAs probably causes POF. Future research should focus on their biological functions in ovarian development.

Key Word: premature ovarian failure, copy number variations, RNA-seq, tissue-specific expression, long noncoding RNA

一、 前言

1.1 卵巢早衰

卵巢早衰（Premature Ovarian Failure, POF）是指女性在 40 岁以前出现的卵巢功能衰退，可分为原发性闭经或 4-6 个月经周期无月经来潮的继发性闭经，同时伴有激素水平的变化，血清中卵泡刺激素（Follicle Stimulating Hormone, FSH）升高至大于 40 IU/L，雌二醇（Estradiol, E2）降低[1]。

卵巢早衰的发病率为 1%，多为散发病例，其致病原因异质性较强，机理尚不明确，主要有遗传致病、自身免疫疾病引发、肿瘤放疗或化疗产生的医源性伤害、药物影响、以及病毒感染等 [2] 。

遗传因素是一个重要的致病原因，能解释卵巢早衰病例的 20-25%，主要可以分为以下几类：（1）10-13%是由于染色体异常引起的卵巢早衰，如 X 染色体单体，X 染色体大段缺失，X 染色体三体，X 染色体与常染色体的平衡易位等；（2）一些 X 染色体或常染色体上的单基因突变也可能导致卵巢早衰，但每个研究报道的单基因仅能解释其研究人群的 1-2%；（3）单基因突变的引起的综合症性卵巢早衰，如 FMR1 导致的脆性染色体综合症（Fragile X syndrome），GALT 导致的半乳糖血症（Galactosemia）等；（4）线粒体基因突变，卵母细胞积累了很多线粒体基因的产物，对卵子发生和成熟有重要作用；（5）根据患者和正常人群中基因突变频率的差异筛选的候选基因（Candidate genes） [3] 。

卵巢早衰遗传致病机理主要有减数分裂异常影响卵母细胞分裂，原始卵泡数量减少，卵母细胞闭锁加速，卵泡成熟障碍以及卵泡激素受体功能丧失，对体内激素调节不反应等 [2] 。

目前研究卵巢早衰致病基因的方法主要有以下几种：

（1）在模式动物中敲除某基因，根据突变表型判断其是否可能导致卵巢早衰。例如，研究发现，敲除 Bmp15 的小鼠生育能力下降，排卵速率降低[4]；Ar 突变的小鼠具有与卵巢早衰相似的表型，卵泡发育受到影响[5]；Nobox 纯和缺失的小鼠出生后的卵泡流失较快，卵泡被纤维组织取代，卵子中高表达的基因 Pou5f1 和 Gdf9 表达量下调等[6]。随后这些研究也在卵巢早衰病人中找到这些基因的有害突变，并在大量患者中计算突变频率加以证明。

（2）在卵巢早衰的病例中筛选突变基因。例如，PGRMC1 首先是在一对母女中被发现的，X 染色体和常染色体在 t（X，11）（q24；q13）位置发生平衡易位，影响到了 PGRMC1，使其表达量下降，之后在 67 名患者中又发现了一位携带无义突变的病例，但是该基因还没有后续的功能研究[7]。

（3）运用比较基因组杂交芯片（array-based Comparative Genomic Hybridization, aCGH）、全基因组外显子测序（Whole-Exome Sequencing, WES）和全基因组关联分析（Genome-Wide Association Study, GWAS）等方法筛选候选基因。这类实验是对大样本量进行全基因组基因组拷贝数变异（Copy Number Variations, CNVs）、SNP（Single Nucleotide Polymorphisms）或 indel（insertion-deletion）突变的筛查，根据实验组和正常对照组的突变频率对比寻找候选基因，但其中大多候选基因都没有验证功能[3]。

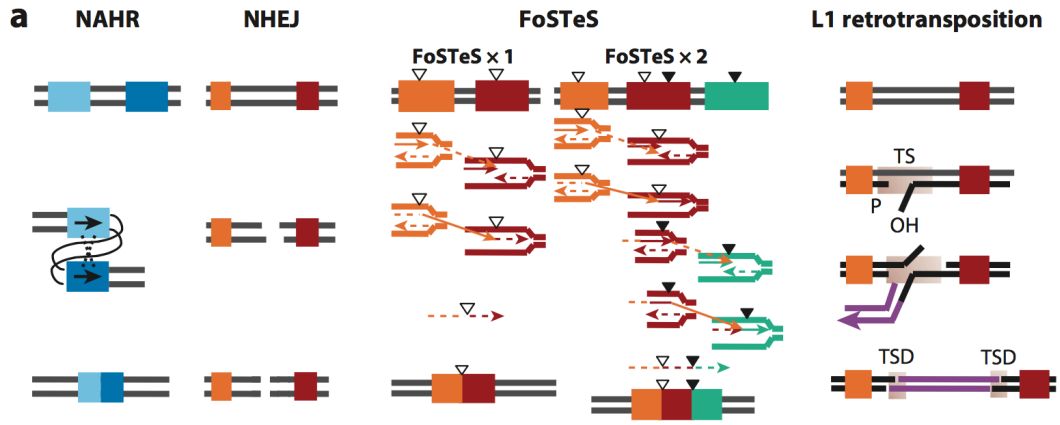
卵巢早衰遗传致病原因的研究不仅帮助了解卵巢的生理功能，而且为遗传咨询和生育保健提供指导，帮助患者提前预测绝经年龄，应用激素替代治疗，冷冻卵子和人工受精等方法防治不孕不育。

1.2 基因组拷贝数变异的研究进展

基因组拷贝数变异是指长度 1kb 以上的 DNA 大片段的重复、缺失、倒位或异位。CNV 的突变率发生率为 10^{-1}～10^{-5}，约为点突变的 1000 以上[8, 9]。

CNV 突变产生的分子机理主要有：（1）重复序列导致非等位基因同源重组（non-allelic homologous recombination, NAHR）；NAHR 是 CNV 产生的重要机制，常发生在有丝分裂和减数分裂过程中，联会期时两段拥有相似序列的基因片段发生重组。基因组上的低拷贝重复序列（low-copy repeats，LCRs）、片段重复（segmental duplications, SDs）和高拷贝重复序列（high-copy repeats）都是 NAHR 产生 CNV 机制中的关键结构性介导因素。染色体间交叉、染色体内（或染色单体间）交叉和染色单体内交叉是 NAHR 产生 CNV 的 3 种方式[10-12]；（2）DNA 修复中的非同源末端连接（nonhomologous end-joining, NHEJ）；NHEJ 是修复 DNA 双链断裂（DNA double-strand breaks, DSBs）的一种重要机制，蛋白质-DNA 复合体直接将断裂的两个 DNA 末端连接起来，不需要同源的 DNA 序列，但修复后往往会在连接末端引入一些碱基的插入。 因为在断裂点附近没有重复序列

等明显的分子结构特征，所以 NHEJ 是形成小的简单的非重复发生的 CNV 的重要机制[13，14]；（3）DNA 复制机制引起的 CNV；在 DNA 复制过程中，重复序列也会诱导基因组的不稳定性，包括复制叉停滞与模板交换（Fork Stalling and Template Switching, FoSTeS）[15]和微同源序列介导的断裂后复制（Microhomology Mediated Break-induced Replication, MMBIR）[16]两种机制，其特点是在 CNV 连接处有微同源序列的复杂基因组重排。



b	NAHR	NHEJ	FoSTeS	Retrotransposition
Structural variation type	dup, del, inv	dup, del	dup, del, inv, complex	ins
Homology flanking breakpoint (before rearrangement)?	Yes (LCR/SD, *Alu*, L1, or pseudogene)	No	No	No
Breakpoint	Inside homology	Addition or deletion of basepairs, or microhomology	Microhomology	No specification
Sequence undergoing SV	Any	Any	Any	Transcribed sequences

图 1.1 CNV 形成的不同分子机理。a. CNV 产生的四种不同机制；b.不同机制产生 CNV 的特点[17]

很多研究报道了 CNV 突变导致的各类出生缺陷、孟德尔式遗传病和一些复杂疾病，如自闭症、智力发育障碍等。CNV 致病的主要机制有：（1）CNV 缺失导致某基因单倍剂量不足（haploinsufficiency），不能编码足够的蛋白维持正常的生理功能，或 CNV 重复导致基因扩增，基因表达水平升高，翻译出多余的蛋白。（2）CNV 不直接影响致病基因，而是影响该基因的转录调控区域，有些 CNV 甚至可以影响到 1Mb 以外的基因。（3）CNV 破坏了基因的结构，导致基因功能丧失。（4）两个基因间存在 CNV 缺失导致基因融合产生功能获得性突变，编码出有害的蛋白。（5）CNV 导致的杂合性缺失（loss of heterozygosity, LOH）使有害的隐形基因发挥功能导致疾病[18]。

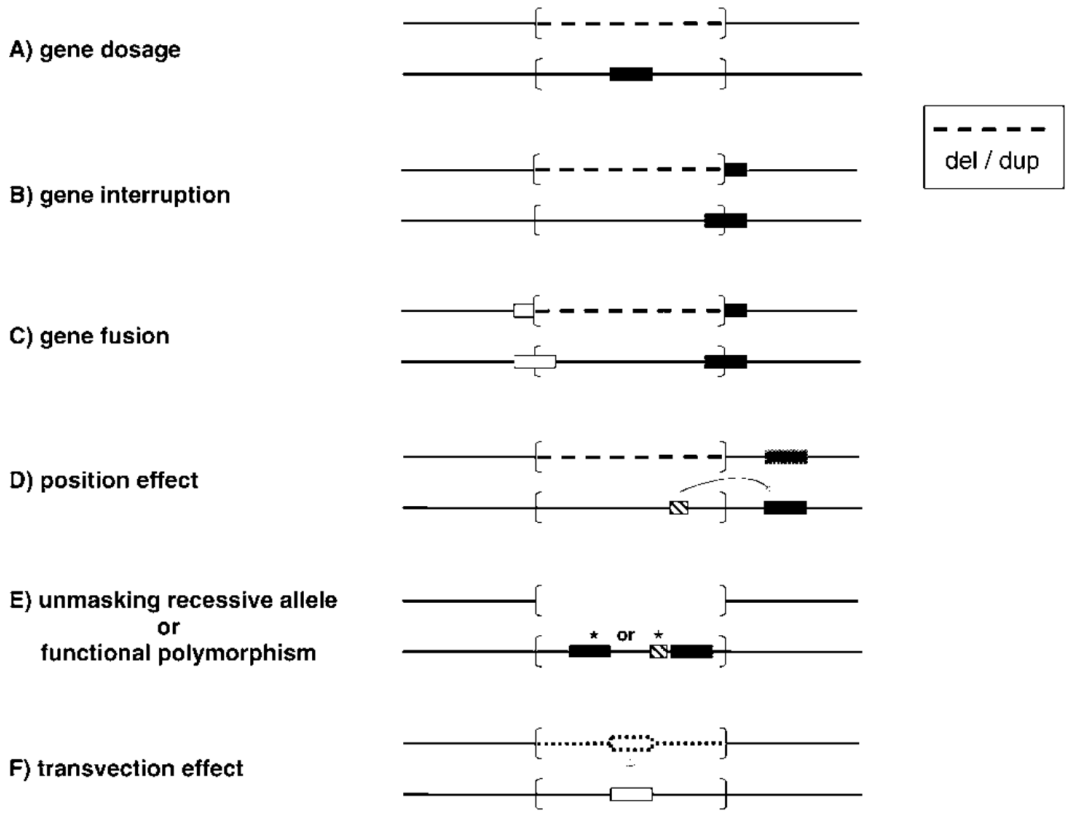


图 1.2 CNV 致病机制 a.基因单倍剂量不足；b.基因被截断；c.基因融合；d.位置效应，基因的调控区受 CNV 突变影响；e.隐形基因暴露；f.基因转位作用[18]

本课题研究 CNV 的技术手段是比较基因组杂交芯片（aCGH）。一百万个长约 60bp 的探针非均匀地分布在全基因组上，探针一般会被设计在没有重复序列的位置上，覆盖密度约为 3kb 一个探针。它的原理是用 Cy5 标记的样本 DNA 和用 Cy3 标记的参考 DNA 混合后同时与芯片上的探针杂交，通过计算荧光信号比值 log2（Cy5/Cy3）得到对应基因片段的拷贝数。通常连续三个探针有稳定信号才被判断为一个 CNV。如果某段基因区域 log2ratio 值等于约 0.6 时表示样本该基因片段比参考样本多一拷贝；如果某段基因区域 log2ratio 值等于-1.0 时表示样本的这段基因片段比参考样本少一拷贝。

1.3 长非编码 RNA 的研究进展

长非编码 RNA（long noncoding RNA, lncRNA）是指长度大于 200nt，5`端加帽，3`端有 poly A 尾，没有开放阅读框（open reading frame, ORF）不编码蛋白质的 RNA。lncRNA 可分为五种类型：（1）正义 lncRNA；（2）反义 lncRNA，即

与另一转录本的一个或一个以上的外显子重合，从同一条 DNA 链上被翻译或另一条不同链翻译；（3）双向 lncRNA，即在某个 lncRNA 开始翻译时，它的不远处的相反链上一个编码蛋白的的基因同时开始表达；（4）内含子上的 lncRNA；（5）两个基因间的 lncRNA[19]。

lncRNA 的平均表达水平较低，约为编码蛋白基因的十分之一，具有组织特异性表达的特点。虽然 lncRNA 的保守性较低，但一些 lncRNA 包含较为保守的基因元件，或是一级结构不保守但二级茎环结构十分保守 [20] 。

lncRNA曾经一度被认为是转录的噪音，是RNA聚合酶转录时的一种副产物，没有生物学功能。但是越来越多的研究发现，lncRNA在染色体修饰、基因转录和转录后调节、细胞分化、器官发育过程中起到重要作用。lncRNA与疾病的产生也有紧密的联系[21]。例如：

（1）lncRNA调控基因表观遗传沉默，INK4b/ARF/INK4a基因座编码了三个肿瘤抑制基因，lncRNA ANRIL由INK4b的反义链编码，与CBX7相互作用，调控了与INK4a的表观遗传沉默。在前列腺癌组织中发现了CBX7和ANRIL表达水平上调，而INK4a表达下降[22]。

（2）lncRNA调控基因剪切，lncRNA MALAT-1与调控前mRNA剪切的SR蛋白相互作用，在敲除MALAT-1的细胞中，SR蛋白不被磷酸化且发生错误定位[23]。在已转移的非小细胞肺癌中MALAT-1是没有转移肿瘤中的三倍，虽然MALAT-1是怎么发挥作用的还尚不可知，但研究者认为它是非小细胞肺癌的转移和预测的很好的标志物[24]。

3）lncRNA调控蛋白翻译，lncRNA BACE1-AS是BACE1的反义链上编码的lncRNA，BACE1的活性与大脑的正常发育有关，合成Aβ（amyloid β-peptide）。BACE1-AS上调使得BACE1表达增加，加快淀粉前体蛋白（amyloid precursor protein）处理，从而沉积过多有害的Aβ，这是阿尔茨海默症的一条重要致病通路[25]。

1.4 本论文主要研究内容

本课题首先应用比较基因组杂交芯片对 21 例卵巢早衰的散发病例和 2 对姐妹进行基因拷贝数变异分析，结合 NCBI、UCSC 等生物信息学数据库筛选可能的候选基因，在此 25 例病人中发现两例分别携带两个不同的 CNV 缺失，影响到了两个不同的在卵巢中高表达的 lncRNA。然后通过对被 UCSC 引用的 Cabili 发表的 RNA-seq 表达谱数据的分析得知，这两个 lncRNA 在卵巢中的表达水平高，是其他组织的 2 倍以上，即在卵巢中特异高表达。但由于 Cabili 的数据集没有生物学重复，所以应用样本量更大的 GTEx 数据集对结果进行了验证，同样表明这两个 lncRNA 在卵巢中是较高表达的，具有卵巢特异性。实验流程图如图 1.3 所示。

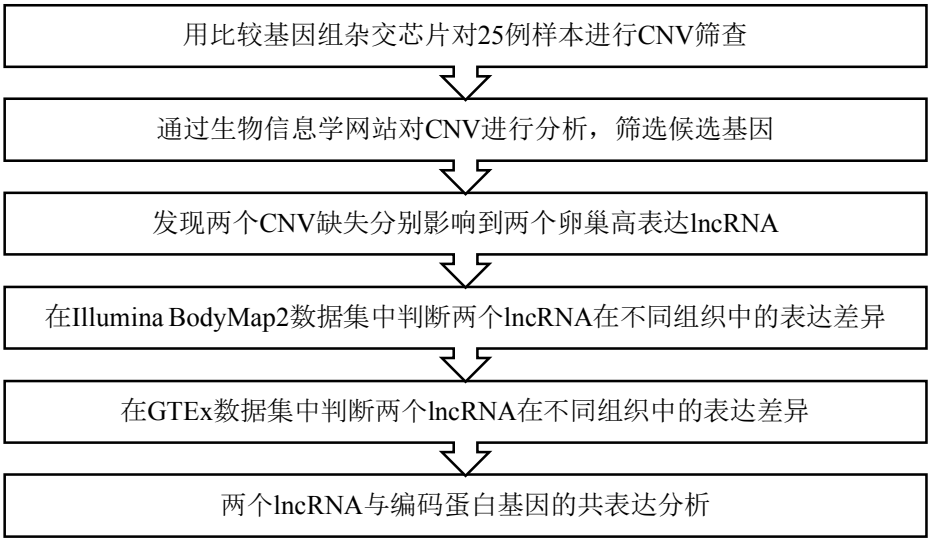


图 1.3 本课题实验流程图。

二、 材料与方法

2.1 实验材料

2.1.1 研究对象

本研究与浙江大学医学院附属妇产科医院及生殖遗传教育部重点实验室合作， 21 例散发和 2 对姐妹病例样本的 DNA 样本及临床信息都由浙江大学医学院提供。研究样本全部符合卵巢早衰临床判断标准：（1）40 岁以前闭经或原发性闭经；（2）FSH 高于 40 UI/L，并且排除了自身免疫疾病、肿瘤放化疗等医源性损害和药物影响等干扰因素。病患血清中 FSH 大于 40IU/L，E2 的波动较大，除少部分患者升高大于 100 pg/ml，大部分患者的 E2 水平都降低，初潮年龄属于正常范围，但闭经年龄提前至 40 岁以前。样本编号与临床信息如表 2.1 和 2.2 所示。

表 2.1 25 例卵巢早衰样本编号与临床信息

样本编号	FSH（UI/L）	E2（pg/ml）	初潮年龄	闭经年龄
1110029	87.5	54.77	14	34
1120082	>40	—	13	18
1130009	63.88	<10.0	15	<40
1130031	63.59	182.7	14	37
1130032	61.03	<18.35	12	35
1130120	75.15	10	15	17
1130152	64.6	126.6	14	23
1130194	74.2	100	13	27
1130196	92.52	249.3	15	17
1130204	69	<10.0	13	23
1140019	123.3	28.9	原发闭经	
1140024	121.96	<10.0	12	29
1140025	78.57	<18.35	17	25
1140027	77.21	79.36	15	29
1140033	53.17	<18.35	17	21

1140036	56.68	101	13	17
1140037	116.47	52	15	24
1140056	43.63	26.06	14	35
1140058	77.1	25.43	14	30
1140064	73.96	<10.0	15	32
1140120	78.46	<10.0	16	33
1120217-1	84.02	40.6	13	39
1120217-2	119.6	55.07	14	39
1120234-1	117.2	25.31	13	16
1120234-2	84.23	48.57	13	17

表 2.2 25 例卵巢早衰病例激素水平、初潮和闭经年龄统计。

FSH（IU/L）	81.54 ± 22.41
E2（pg/ml）	54.61 ± 59.03
初潮年龄(岁)	14.13 ± 1.33
闭经年龄(岁)	26.83 ± 7.59

2.1.2 仪器

Veriti 96 孔梯度 PCR 仪（ABI 公司）

5424 高速离心机（Eppendorf 公司）

芯片杂交炉（SL）

芯片扫描仪器（Agilent 公司）

NanoDrop2000 紫外分光光度仪（Thermo 公司）

Tanon 2500R 凝胶成像仪

Tanon EPS 300 电泳仪

Hoefer HE33 水平电泳槽

漩涡震荡器

电热恒温培养箱

磁力搅拌器

电热恒温水浴锅

DNP-9052BS-III 电热恒温培养箱（Cimo）

2.1.3 材料

一次性手套

2.5 ul, 10 ul, 20 ul, 100 ul, 200 ul, 1000 ul 微量加样器（Eppendorf 公司）

1.5 ml EP 管

八连管

过滤柱（Agilent 公司）

Agilent 公司 SurePrint G3 Human CGH Microarray, 1x1M

Agilent 公司芯片夹、扫描夹

2.1.4 试剂

a) 比较基因组杂交芯片实验所用试剂

限制性酶切体系: 10x buffer C、Acetylated BSA (10 g/μL)、AluI (10 U/μL)、RsaI (10 U/μL)，Promega Female Genome DNA control (208 ng/ul)

荧光标记时的随机引物及标记体系: Random Primer，5x Buffer C、10x dNTP、Cyanine 3-dUTP (1.0 mM)、Cyanine 5-dUTP (1.0 mM)、Exo-Klenow fragment

纯化: 1x TE (pH 8.0)

杂交前的封闭体系: Cot-1 DNA (1.0 mg/mL)、Agilent 10x Blocking Agent、Agilent 2x Hybridization Buffer

洗芯片: Oligo aCGH Wash Buffer 1、Oligo aCGH Wash Buffer 2、Acetonitrile、Stabilization and Drying Solution

b) 聚合酶链式反应（PCR）

ddH_2O, 2.5 mM dNTP Mixture, 10x LA PCR Buffer (Mg^{2+}), 25 mM $MgCl_2$, TaKaRa LA Taq DNA Polymerase (5 U/uL), 上下游引物（由上海睿迪生物科技有限公司合成），测序引物（由上海铂尚生物有限公司合成并测序）

c) 电泳检测

1x TAE, 琼脂糖，TaKaRa 2000 bp DNA Marker, TaKaRa 10000 bp DNA Marker, 6x Loading Buffer（已加 GelRed）

2.2 实验方法

2.2.1 Agilent aCGH 芯片制作

（1）实验前的准备：

a. 样本必须用电泳检测基因组 DNA 是否有降解，若有较为严重的弥散则不能用；

b. 用 NanoDrop 紫外分光光度仪检测 1 ul gDNA 的浓度和品质，A260/A280 比值应该 1.8~2.0，表明没有蛋白质污染。A260/A230>2.0，说明没有其他有机化合物如异硫氰酸胍、酒精、酚以及碳水化合物之类的污染物。

c. 确认样本性别。

（2）酶切：

a. 计算 DNA 体积

C_{DNA}*V_{DNA}=2.0 ug（1.5 ~3.0 ug 均可）

V_{DNA}+V_{H2O}=22.2 ul

b. 按照酶切体系加入到八连管中

酶切体系：（先加水再加 DNA 然后加其他）

试剂	体积(ul)
10 x Buffer C	2.6
BSA	0.2
Alu I	0.5
Rsa I	0.5
DNA＋H_2O	22.2
总体积	26

c. 轻轻混匀后离心，将酶切体系放到 PCR 仪中，酶切条件如下：

37 ℃　　120 min

65 ℃　　20 min

4 ℃　　保存

d. 酶切结束后，取 2 ul 电泳，条带弥散在 200-500 bp，酶切成功则可进入下一步。

（3）荧光标记（避光）：

a. 取出样本和对照瞬时离心；

b. 样本和对照中分别加入 5 ul 随机引物，此时总体系共 29 ul，震荡离心；

c. 在 PCR 仪上按照以下程序进行扩增：

95 ℃　　3 min

冰水浴 0-4 ℃　　5 min

d. 扩增结束后，在冰上避光配置以下荧光标记体系：

试剂	体积(ul)
Nuclease-free Water	2.0
5 x gDNA Buffer	10.0
10 x dNTP	5.0
Cye3-dUTP/Cye5- dUTP	3.0
Klenow	1.0
扩增后 DNA	29
总体积	50

e. 按照以下程序在 PCR 仪上进行荧光标记反应：

37 ℃　　120 min

65 ℃　　10 min

12 ℃　　保存

（4） 标记产物纯化：

a. 取出 PCR 管离心；

b. 取出收集管和纯化柱，标记相应样本编号；

c. 往每只柱子中加入 430 ul 1x TE（pH 8.0），将标记体系加入纯化柱中，室温离心 14000 g，10 分钟，弃滤液；

d. 再往每个柱子中加入 480 ul 1x TE（pH 8.0），室温离心 14000 g，10 分钟，弃滤液；

e. 弃去滤液，室温离心，14000 g，10 分钟，弃滤液；

f. 取新的收集管，分别做好标记，将纯化柱倒置其中，14000 g 离心 2 分钟，收集纯化后的产物，体积约为 20 ul；

g. 加 1 x TE 补至 80.5 ul，指弹混匀后，瞬离；

h. 用锡箔纸将样品包裹好避光保存；

i. 取 1.5ul 用 NanoDrop 检测荧光标记效率。第一，荧光浓度值：根据经验值，一般 Cye5 大于 3 pmol/ul，Cye3 大于 4 pmol/ul。第二，荧光比活性计算：根据公式 Specific Activity = (pmol per μL dye) / (μg per μL genomic DNA)。第三，标记产物产量 DNA 浓度大于 100 ng/ul。

（5） 预杂交：

a. 将实验组和对照组正确对应，混合到一个新的 EP 管中；

b. 向上一步骤产物中加入 362 ul 封闭体系, 轻轻吹匀后瞬离

试剂	体积 (ul)
Mouse Cot-1 DNA	50
10 x Blocking Agent	52
2 x Hybridization Buffer	260
总体积	362

c. 金属浴 95 ℃ 3 分钟，水浴 37 ℃ 30 分钟。

（6） 清洗旧芯片：

a. 洗液配置：800 ml ddH_2O+1.52 ml K_2HPO_4+2.48 ml KH_2PO_4 加入一高一矮两个烧杯。高烧杯放置于加热板上加热；矮烧杯放置于冰桶中。注意：（1）洗液必须现用现配;（2）区分 K_2HPO_4 和 KH_2PO_4;（3）两个烧杯都用保鲜膜封口；

b. 高烧杯加热至 65 ℃（用带着手套的手去触摸烧杯壁，感觉到烫手；约需 20 min）时，取出旧芯片放入装有乙腈的液缸中，300~350 r/min，3 min。3 min 后，取出芯片观察表面是否还有污迹。若有的话，可以放入乙腈缸中再洗 1min。注意：芯片在乙腈缸中清洗的时间不能超过 10 min；

c. 将芯片放到芯片架上，缓慢放入已经加热至 65 ℃的高烧杯中。注意：芯片需放在靠近芯片架边缘的凹槽中，以借助沸腾的起泡冲洗芯片（因为起泡产生自芯片架边缘与烧杯壁接触位置）；

d. 从 65 ℃加热至沸腾至少需要 40 min。从沸腾开始计时，10-15 min；

e. 缓慢取出芯片；缓慢放入至于冰桶的矮烧杯中，计时 2 min；再缓慢取出芯片。此时，可将洗好的芯片放到一个空的芯片盒中，用纸小心轻轻拭去芯片标签上的水滴。旧芯片清洗干净，可待使用。

(7) 芯片杂交：

a. 准备好芯片及干净的 gasket，将 gasket 放入 chamber 中；

b. 取出样品瞬时离心；

c. 用 200 ul 微量加样器将混合液缓缓滴加到 gasket 表面，枪头不能接触芯片，并尽量避免产生气泡；

d. 取出芯片，按照“Agilent kissing Agilent”的方向缓慢盖上芯片，避免样品溢出；

e. 记录芯片序列号及 gasket 的编号，将 chamber 拧紧，并尽量排除橡胶圈间的气泡，将 chamber 放入杂交炉中，并放入一个空的 chamber 以保持平衡。65 ℃杂交 40 小时，转速为 20 rpm, 记录好杂交时间。

(8) 杂交后下芯片以及洗芯片

a. 扫描前准备：提前一天晚上将 buffer2 放入恒温箱预热（乙腈、drying、buffer1 最多用 2 次，用 2 次倒掉清洗缸，并放在恒温箱风干）；

b. 取出扫描夹，开启电脑后开启扫描仪，等机器稳定后再开软件；

c. 将芯片放入无转子的 buffer1 中，分开芯片和垫片；

d. 轻涮后卡入有转子的 buffer1 中，550 r/min，5min；

f. 将 buffer2 缸下垫浮漂，将芯片卡入 buffer2 缸中，550 r/min，1 min；

e. 缓慢卡入乙腈缸中（乙腈缸下需垫纸），出完标签时等上部乙腈风干后再缓慢取出，放入扫描夹中，Agilent 向上，盖上盖子；

f. 设置 start slot 和 end slot，profile：Agilent 3_CGH；

g. 在 out path 中选择存储文件夹（提前在电脑中新建好）；

h. 等待 scan slot 可点击后，将芯片放入扫描仪中，点击 scan slot；

i. 扫描完成后，将芯片卡入 drying 中（方法同乙腈，缸下仍需垫纸），取出保存。

(9) 数据分析：

将芯片扫描结果导入 Agilent Genomic Workbench 进行拷贝数变异分析。

2.2.2 长片段 PCR 验证断点

本研究中使用的 Agilent 1x1 million 芯片包含一百万个探针，两个探针之间的距离约为 3kb，包含连续三个探针以上的缺失或重复被认为是 CNV。包含探针个数太少，或 log2ratio 值不稳定的 CNV 需要用长片段 PCR 和 Sanger 测序排除系统和实验操作带来的假阳性。如果 CNV 是真实存在的，则有长片段 PCR 产物，如果 CNV 是假阳性，则没有产物。

首先将 CNV 的最大范围的左右两个探针的坐标输入 UCSC Genome Browser，查找并获取上游探针的前约 1 kb 和下游探针的后 1 kb 的序列，再利用 Oligo 6 软件在这两个区域分别设计上下游引物。本研究中验证的包含卵巢高表达 lncRNA-2 的 CNV 缺失（Chr8:17674902-17697149）的扩增和测序引物如下：

引物名称	引物序列 5'→3'	扩增片段长度
Lnc2-1-F	AAGGATGAACTCAGCAAACTTGGAAAACTG	最长 7813 bp
Lnc2-1-R	TTGATGGCAGAGGGGTATCTTCTATGTCAT	
Lnc2-2-F	ATTGGGAATAAGGATGAACTCAGCAAACTT	最长 7818 bp
Lnc2-2-R	TGGCAGAGGGGTATCTTCTATGTCATTAGA	

测序引物名称	引物序列 5'→3'
seq-1F	GGATGAACTCAGCAAACTTGGA
seq-1R	GCAGAGGGGTATCTTCTATGTC

PCR 体系：

试剂	体积（ul）
ddH_2O	6.4
10x LA PCR Buffer (Mg^{2+} Plus)	1.0
2.5 mM dNTP Mixture	1.6
上游引物	0.2
下游引物	0.2
DNA	0.5
TaKaRa LA Taq DNA Polymerase (5U/u1)	0.1
总体积	10.0

PCR 扩增条件：

98℃　3 min

30 cycles { 98℃　30 sec
　　60℃　30 sec
　　68℃　7 min

68℃　10 min

12℃　∞

各取 1 ul PCR 产物与 1 ul 6x Loading Buffer 混合，用 1%琼脂糖凝胶进行电泳检测，均有清晰单一且明亮产物条带，背景干净且条带大小与预期相符。将成功扩增的 PCR 产物于 4℃冰箱保存。

将保存于 4℃冰箱的 PCR 扩增产物送于上海铂尚生物公司进行测序，将测序结果用 FinchTV 软件读取序列并用 UCSC 进行 Blat 比对，进行 CNV 验证和断点分析。

2.2.3 组织特异高表达长非编码 RNA 的筛选

Cabili 等[20]在 2011 年发表了关于 20 种人体器官组织和细胞的长非编码 RNA 的研究。他们利用了 Illumina Human Body Map 2 Project 的 16 种人体组织和 Rinn 实验室重新测序的 8 种人体组织和细胞的 RNA-seq 数据，其中包括脂肪组织、肾上腺、大脑、胸腺、结肠、心脏、肾脏、肝脏、肺、淋巴结、卵巢、前列腺、骨骼肌、白血细胞、睾丸、甲状腺、人肺成纤维细胞、皮肤成纤维细胞、HeLa 细胞、胎盘（人肺成纤维细胞、大脑、肺、睾丸分别有两组测序数据）。本课题从 UCSC 网站（http://hgdownload.soe.ucsc.edu/downloads.html）下载了该研究发表的共计 8195 个 lncRNA 转录本的 RNA-seq 表达数据（已经统一在 FPKM 值水平上），用于卵巢组织特异高表达长非编码 RNA 的筛选。

Cabili等的研究中，每个人体组织样本没有生物学重复。因此，本课题采用表达差异变化倍数（Fold Change，FC）来判断组织特异性表达lncRNA，阈值设为FC值为大于等于2，即一个lncRNA在卵巢中的表达水平高于或者等于在其他所有19种组织的2倍以上。

因为Illumina Human BodyMap2 Project和Rinn实验室的RNA-seq数据没有生物学重复，也没有考虑到人个体之间的lncRNA表达差异[26]，所以本课题从GTEx

基因表达数据（http://gtexportal.org/home/，为最新版本V6，已经统一在RPKM水平上）再次筛选卵巢特异高表达的lncRNA，并比较了GTEx和Cabili文章的两组数据的异同，对筛选结果的可靠性进行了验证。GTEx Project中有8555个人体组织样本，其中包括脂肪组织、肾上腺、膀胱、血液、血管、骨髓、大脑、胸腺、子宫颈、结肠、食道、输卵管、心脏、肾脏、肝脏、肺、肌肉、神经、卵巢、胰腺、垂体、前列腺、唾液腺、皮肤、小肠、脾脏、胃、睾丸、甲状腺、子宫、阴道等53种人体组织，共计195747个转录本。因为GTEx数据中每种组织有多个生物学重复，所以筛选卵巢特异高表达的lncRNA的方法是用t检验判断一个基因在卵巢与其他组织表达是否有差异，当同时满足Bonferroni矫正后的P值≤0.05和FC≥2时，认为该基因是在卵巢中特异表达的[27]。

本课题中认为当一个lncRNA的FPKM或RPKM值大于等于1时说明该lncRNA是有表达的，并且其表达量足够在功能实验中检测到[28]。

三、研究结果

3.1 卵巢早衰的基因拷贝数变异分析

3.1.1 样本DNA质量检测

（1）用NanoDrop 2000紫外分光光度仪检测样本DNA浓度和纯度，如表3.1所示：

表3.1 25例卵巢早衰DNA样本浓度和纯度

样本编号	Conc. (ng/ul)	260/280	260/230
1110029	106.9	1.83	1.78
1120082	181.4	1.86	2.00
1130009	174.8	1.85	1.98
1130031	173.6	1.83	1.89
1130032	138.1	1.85	1.89
1130120	212.0	1.84	1.97
1130152	112.4	1.84	1.74
1130194	162.1	2.07	3.04
1130196	114.3	1.85	1.71
1130204	96.8	1.86	1.72
1140019	102.5	1.93	1.86
1140024	103.0	1.83	1.55
1140025	205.5	1.86	2.02
1140027	231.8	1.86	2.12
1140033	395.4	1.86	2.21
1140036	397.0	1.85	2.24
1140037	144.2	1.83	1.50
1140056	382.9	1.87	2.05
1140058	175.5	1.84	1.91
1140064	207.9	1.86	2.08
1140120	285.0	1.83	1.46
1120217-1	197.8	1.81	1.80
1120217-2	180.9	1.81	1.84
1120234-1	254.0	1.79	1.34
1120234-2	264.8	1.83	2.07

分析：所有样本DNA浓度都在100ng/ul左右或以上，浓度达到芯片杂交试验的要求。A260/A280比值在1.8~2.0，说明蛋白质污染较少。A260/A230值在1.34～2.24，样本的A260/A230值>2.0说明没有其他有机化合物如异硫氰酸胍、酒精、酚以及碳水化合物之类的细胞污染物，有些样本的A260/A230值较低，可能原因是抽提DNA时乙醇洗涤后没有挥发完全。

（2）用1%琼脂糖凝胶电泳进行检测，条带清晰明亮，无弥散无拖尾现象，显示DNA样本完好无降解，可用于芯片杂交试验。本课题中的DNA样本都无降解现象，图3.1示例质量较好没有被降解的DNA的电泳图。

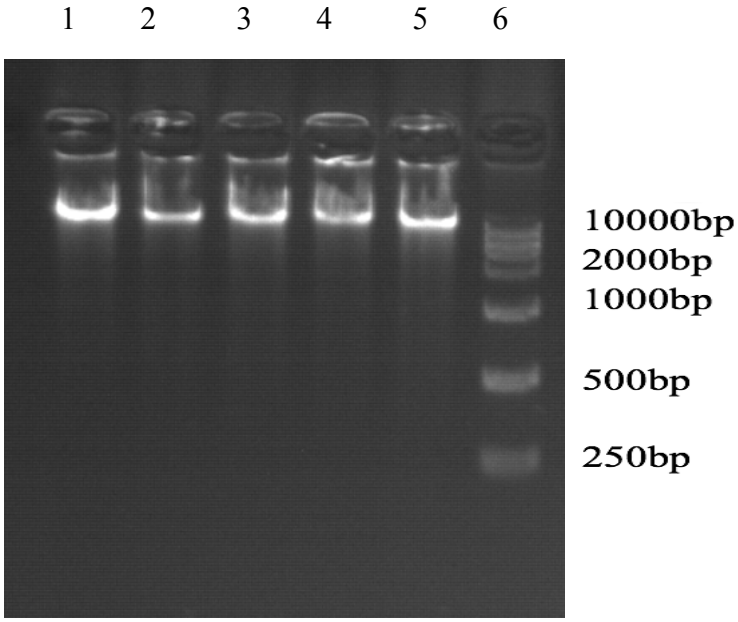


图3.1 电泳检测DNA是否降解。泳道1，样本1120082；泳道2，样本1130009；泳道3，样本1130031；泳道4，样本1130032；泳道5，样本1130194；泳道6，10000 bp DNA Maker。

3.1.2 基因拷贝数变异的分析

本课题首先应用比较基因组杂交芯片对25例卵巢早衰病患样本进行基因拷贝数变异分析，然后将检测到的CNV区域输入UCSC Genome Browser（http://genome.ucsc.edu）通用DGV数据库（Database of Genomic Variants, http://dgv.tcag.ca/dgv/app/home）判断该CNV区域重复或缺失是否罕见，即判断该区域的变化是突变还是多态性。如果变化区域罕见，则判断是否影响到基因的编码区或转录调控的区域，再通过OMIM（http://omim.org）、GeneCards（http://www.genecards.org）等数据库以及文献查找了解该基因的功能是否可能与卵巢早衰的致病相关，也可利用MGI（http://informatics.jax.org）网站查询该基因的杂合或纯和突变小鼠是否存在生殖障碍或不孕不育。最终通过以上步骤确定卵巢早衰的候选基因。

本研究中以连续3个探针的实验组和对照组的荧光信号比值log2ratio值同时升高（log2ratio为0.6是指该段基因区域重复一拷贝）或降低的基因组（log2ratio为-1.0是指该段基因区域缺失一拷贝）。芯片检测到样本1130120的核型是46,X,del(X)(q23)，可能是因为X染色体大片段缺失导致的卵巢早衰。其余的24例病人中一共检测到540个CNV，其中有213个CNV重复和327个CNV缺失，平均每个人的基因组上包含21.6个CNV，变化区域大小为4.1 Kb至3.0 Mb。罕见的CNV一共有123个，其中有69个CNV影响到编码蛋白基因，总共涉及到102个基因。分析这些基因的功能后，发现30个可能与卵巢早衰发病有关的候选基因（表3.2），它们的功能主要包括参与纺锤体形成、控制细胞周期、DNA损伤修复、卵泡发育、激素分泌和X染色体沉默，可能影响减数分裂和卵子发生等过程。

表3.2 用比较基因组杂交芯片筛选到的CNV影响到的编码基因

相关功能	CNV影响的基因
纺锤体形成	*MAP7D3*、*MOB3B*、*PCM1*、*FAM82A1*
X染色体沉默	*FTX*
细胞周期	*TTC28*、*SHCBP1*、*DESI2*、*ANAPC5*、*PTPRD*、*AHCTF1*、*ARGHAP10*、*PPP1R12A*
DNA损伤修复	*NSMCE2*、*FANCA*、*FANCC*、*BTBD2*、*CHAF1A*、*JMY*、*DDX1*、*XRCC2*、*HORMAD2*
卵泡发育	*FGF13*、*OVGP1*、*AQP7*
激素分泌	*HSD17B12*、*PDGFRA*、*TYRO3*、*PGRMC1*、*NCOR2*

本研究发现1130009号样本在6号染色体上携带了一个约为170 Kb的CNV缺失，包含了一个完整的lncRNA的基因座，在后文中称为lncRNA-1（图3.2和图3.3）；1140056号样本在8号染色体上携带了一个约为14.8 Kb的CNV缺失，包含了另一个lncRNA基因座的三分之一，在后文中称为lncRNA-2（图3.4和图3.5）。通过UCSC Genome Browser可知，相比于其他组织，这两个lncRNA在卵巢中表达较高，它们在人与灵长类动物之间比较保守，在人与常用的模式动物小鼠中是不保守的。

这两个lncRNA在正常卵巢中特异性高表达说明它们可能在卵巢中发挥组织特异性的功能[29]。在上述两个病人中缺少一拷贝，可能会导致lncRNA表达量的下降，因为单倍剂量不足从而导致卵巢或生育相关疾病的发生。

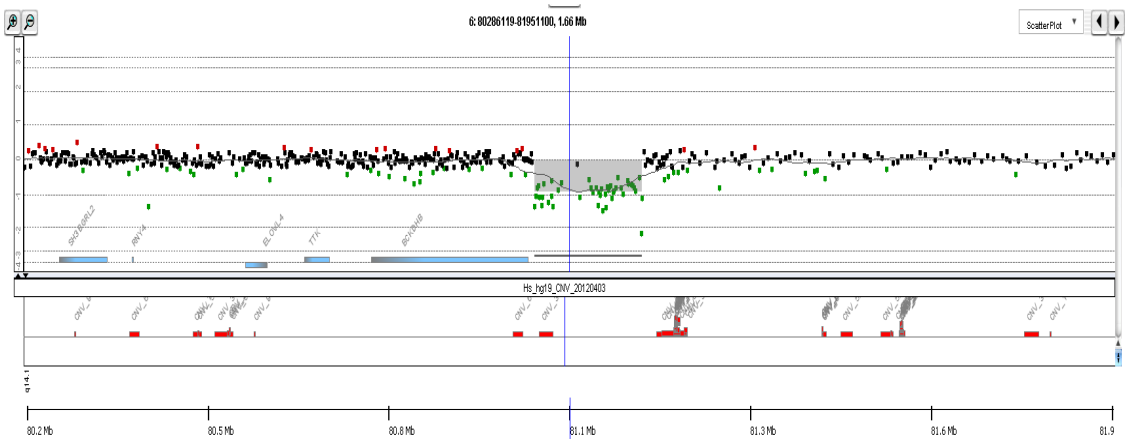
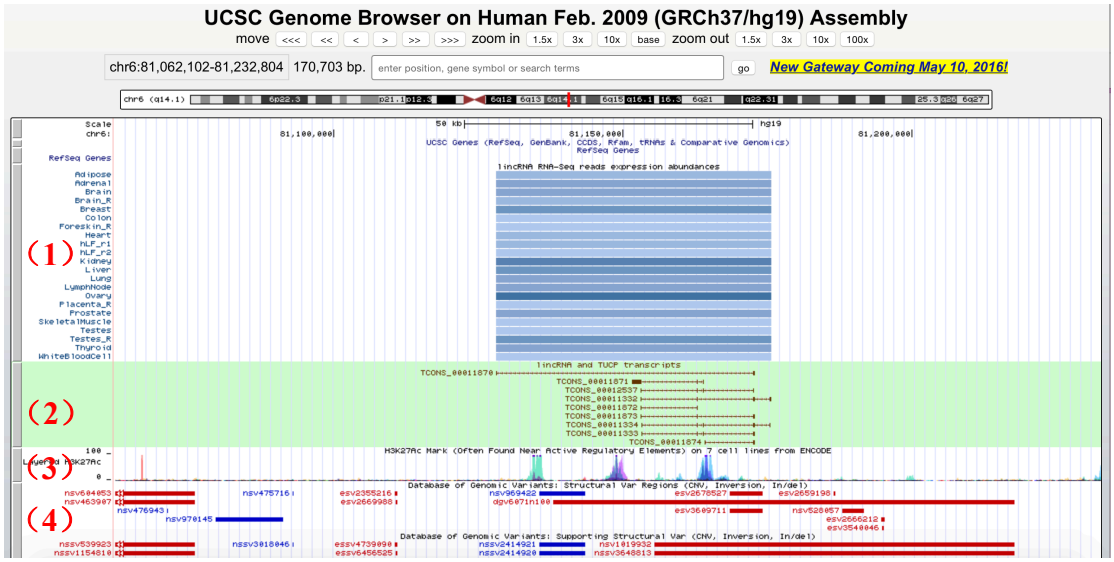


图3.2 lncRNA-1在样本1130009中存在CNV缺失片段。位置为Chr6:81062102-81232804，长度约为170 Kb，探针的点为绿色，log2ratio值为-1，表示该区域缺失一拷贝。



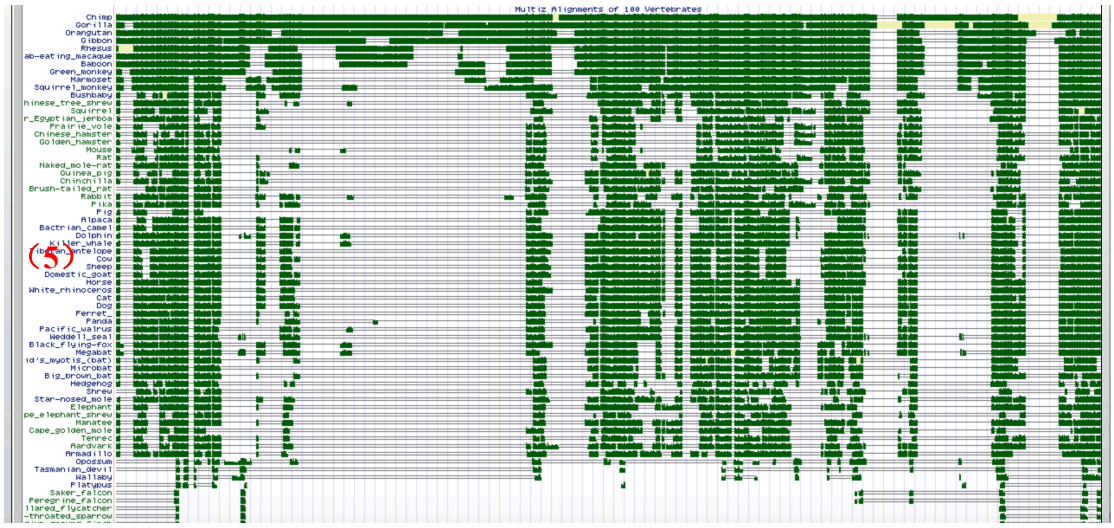


图3.3 样本1130009的CNV缺失区域在UCSC Genome Browser中的截图。（1）蓝色的横线表示lncRNA-1基因座在人体不同器官组织的表达水平，颜色越深代表表达越高，可知lncRNA-1在卵巢中表达最高；（2）该lncRNA基因座的9种可能的剪切体；（3）H3K27ac的不同颜色信号峰表示该位点转录活跃；（4）红色横条表示缺失，蓝色表示扩增，尚未报道与本研究中发现的1130009的CNV缺失相同范围的CNV，说明该CNV可能是罕见的；（5）绿色的横条表示人与某物种的某段区域是保守的，该lncRNA在人与灵长类动物之间较为保守。

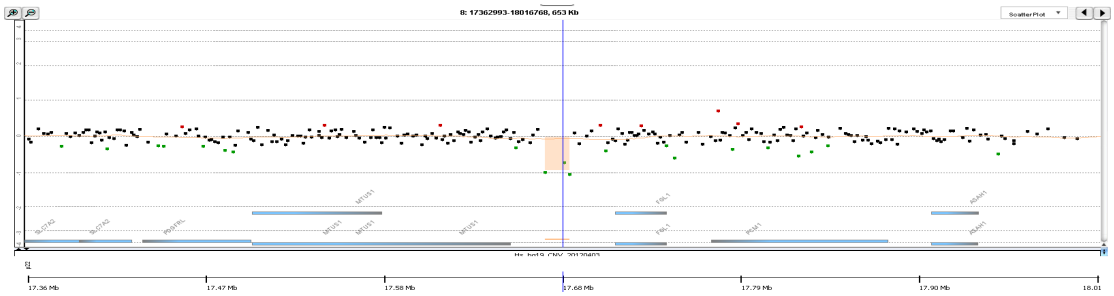


图3.4 lncRNA-2在样本1140056中存在CNV缺失片段。位置为Chr8:17674902-17697149，长度约为14.8 Kb，探针的点为绿色，log2ratio值为-1，表示该区域缺失一拷贝。

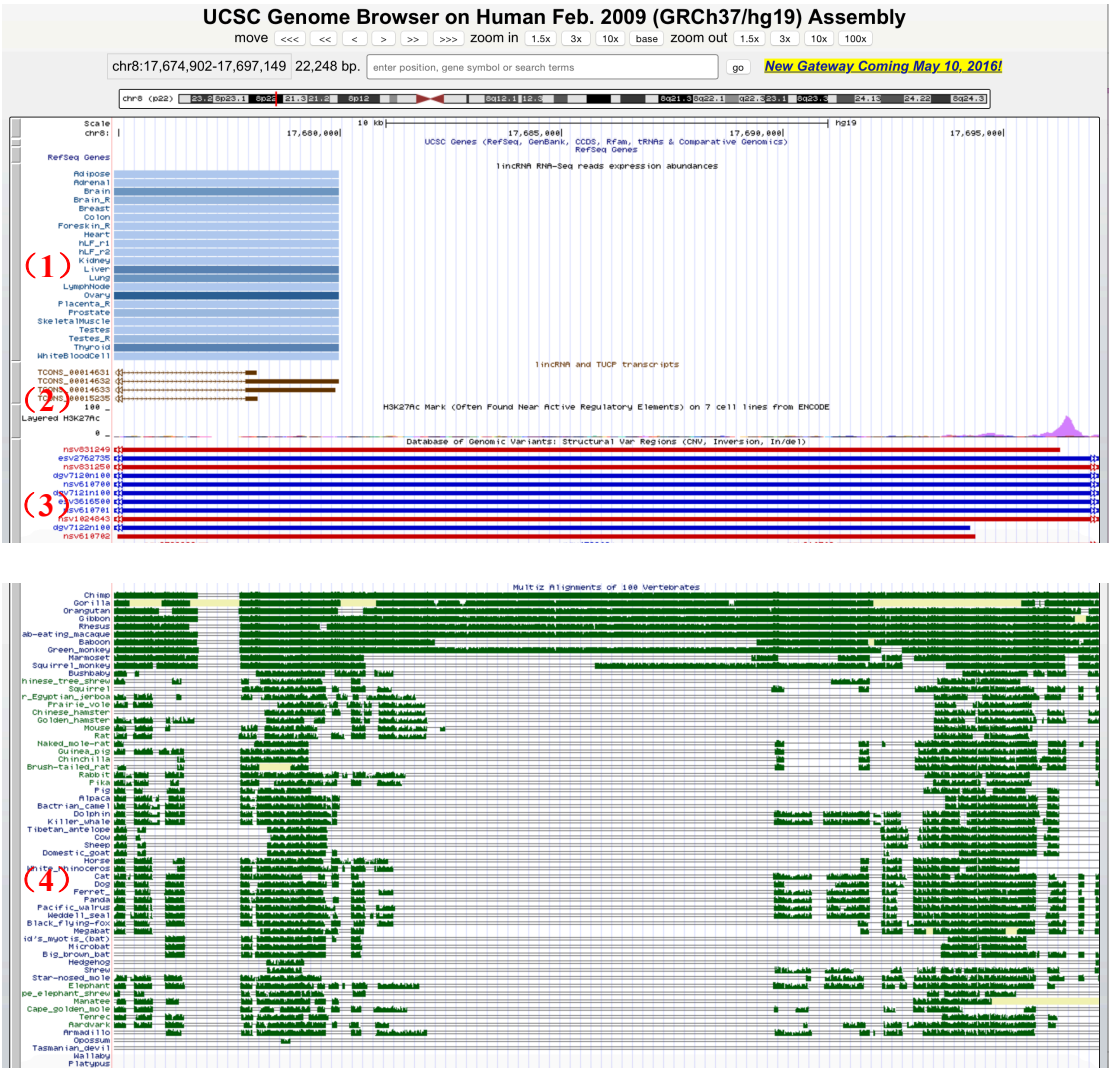


图3.5 样本1140056的CNV缺失区域在UCSC Genome Browser中的截图。（1）蓝色的横线表示lncRNA-2基因座在人体不同器官组织的表达水平，颜色越深代表表达越高，可知lncRNA-2在卵巢中表达最高；（2）该lncRNA基因座的4种不同剪切体；（3）红色横条表示缺失，蓝色表示扩增，尚未报道与本研究中发现的1140056的CNV缺失相同范围的CNV，已报道的大小相近的CNV缺失频率低于1%，说明该CNV是罕见的。（4）绿色的横条表示人与某物种的某段区域是保守的，该lncRNA在人与灵长类动物之间较为保守。

3.1.3 基因拷贝数变异的验证

样本1130009的CNV缺失区域约为170 Kb，范围较大，且缺失位点众多探针的log2ratio值都约为-1。这样的大片段缺失是可信的。

但样本1140056的CNV缺失片段只包含三个探针，为了排除实验误差可能带来的假阳性，本课题用长片段PCR进行验证。分别在最大缺失范围的两个log2ratio

值接近0的显示为黑色点的探针的上游和下游设计引物。如果该CNV缺失是真的，则会有长度最长为接近8 Kb的PCR产物，否则没有任何产物（图3.6）。

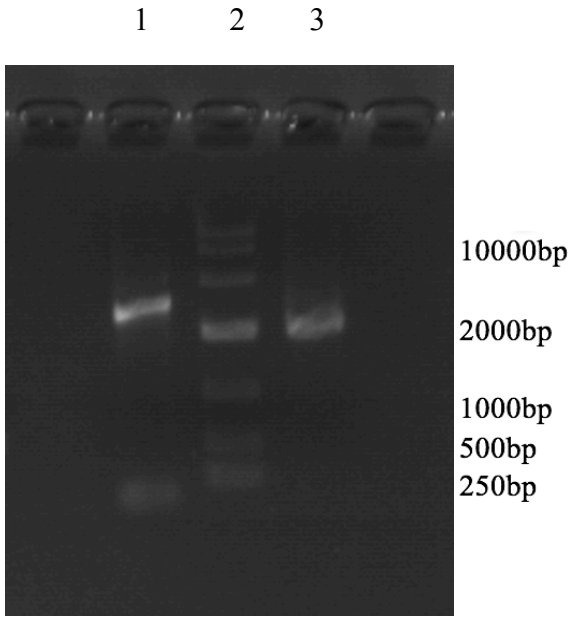


图3.6 长片段PCR验证样本1140056的CNV缺失。泳道1，lnc2-1-F和lnc2-1-R引物的PCR产物，有长度为2 kb至4 kb的产物和引物二聚体；泳道2，10000 bp DNA Maker；泳道3，lnc2-2-F和lnc2-2-R引物的PCR产物，有长度为2 kb至4 kb的产物。

通过电泳结果可知设计的两对引物都有长度接近的PCR产物，说明样本1140056的CNV缺失是真实的。对PCR产物进行测序，返回的测序结果通过UCSC中hg19数据库进行Blat比对，可知CNV断点在chr8:17676655-17697030（图3.7）。

```
GGCATGGCAT TGATTTACCG ATCACAGAGA AGACTTTTTA AAGCAATTTA  17839662
TCTAGAGATG CTACCTTCAA AACGGGTTAC ATTTTAGATT TATAAAGTTT  17839612
TGAAGTTTTG CTTACGTTCT CTAGAGCCTG GtGACTTTTT TTATTTTTTT  17839562
AGGTGGAGTT TCACTCTTGT TGTCCAGGCT GGAGTGCAGT GGCGTGATCT  17839512
TGGCTCACtg caacctccgc ctccctggtt caagcgattc ttgtgcctca  17839462
gcttcccgag aagctgggat tacaggcgcc caccaccacg cctggctaat  17839412
ttttgtattt ttagtaaaga cggggtttca ccatgttggc caggctggtc  17839362

… …

tatataacat atgtaatata cattatgtac attatataac atatgtaata  17819312
tacattatgt acattatata acatatgtaa tatatatatg tgtatataaa  17819262
aaacatacct ggaaaaagca atataaatgt ctgggtctta ctcagtagcc  17819212
cagactagag tgcagtggtg tgatcatggc tcacTGCAGC CTCAACCTCA  17819162
GTCGATCCTC CTGCCTCAGC CTCCCACGTA GCTGAGACtA CAGGCTTGTG  17819112
CCACCACACC TGGCTAATTT TGTTTGTatt ttttgcagag acagggtttc  17819062
```

图3.7 将测序结果通过UCSC网站的Blat进行断点分析，选择Human Genome及Feb. 2009 (GRCh37/hg19) Assembly进行BLAT所得结果，确定CNV缺失范围是chr8:17676655-17697030。

3.2 卵巢特异表达lncRNA的筛选

3.2.1 应用Illumina Human BodyMap2数据集筛选卵巢特异表达lncRNA

在UCSC Genome Browser中提示，样本1130009和样本1140056携带的CNV缺失所影响到的两个lncRNA在卵巢中均为高表达。UCSC数据库引用了2011年Cabili等发表的不同的人体组织和细胞系中的lncRNA的研究，分别包括Illumina Body Map 2 Project的16种人体组织和Rinn实验室的8种人体组织和细胞系，总共20种不同组织的8195个lncRNA基因转录本的RNA-seq表达谱数据。为了进一步确定它们确实是在卵巢中特异高表达的，我下载了Cabili等发表的lncRNA的RNA-seq表达谱数据，进行了卵巢组织特异性转录本的表达分析。因为该数据集中这些样本没有生物学重复，所以只用表达差异倍数变化FC≥2来筛选卵巢特异表达的lncRNA，筛选后得到364个lncRNA转录本。当某个lncRNA的FPKM值≥1时，认为它是有表达的，再次筛选后得到316个lncRNA转录本（附表1），来自于86个lncRNA基因座，这些转录本即是卵巢特异高表达的lncRNA转录本（图3.8）。

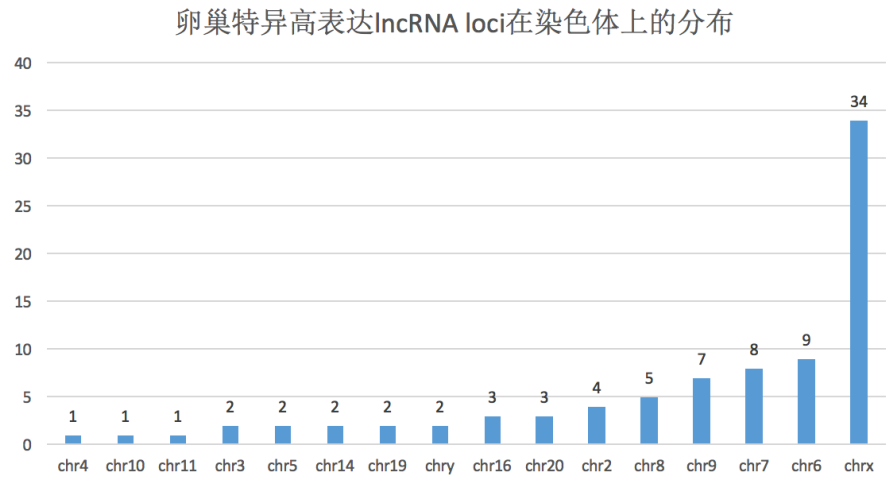


图3.8 卵巢特异高表达lncRNA在染色体上的分布，在X染色体上最多，1、12、13、15、17、18、21和22号染色体上没有分布，Y染色体上有分布可能是实验误差或序列比对时出现问题。

本研究中用比较基因组杂交芯片发现的两个lncRNA基因座lncRNA-1和lncRNA-2被包含在筛选出来的316个卵巢组织特异性高表达的lncRNA中，它们在上皮组织、人肺成纤维细胞和骨骼肌中都没有表达。倍数变化的计算用log2（FPKM＋0.01）进行变换，两种组织中的表达水平log2（FPKM＋0.01）值相差1则表示表达差异为2倍。这两个lncRNA基因座在卵巢中的表达量远远高于其他组织，是其他组织的4倍或4倍以上（图3.9）。

一个基因的不同转录本是通过包含剪切位点（splice site）的读段（reads）发现和确认的。Cabili等认为，由于当时的测序技术和计算方法的局限，当一个基因的所有读段不包含剪切位点时就不能分辨出这个基因的不同转录本，从而影响计算得到不同转录本的表达量的准确性，所以他们没有估算同一个基因的不同转录本的表达水平，而取一个lncRNA基因座的最高表达水平（maximal expression level）[20]。

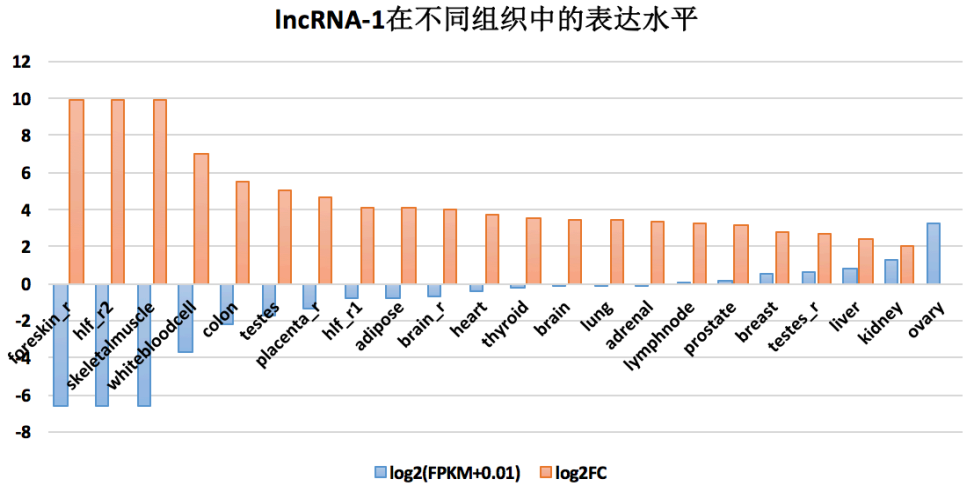


图3.9 1130009号样本CNV缺失包含的lncRNA-1在不同组织中的表达情况。蓝色代表绝对表达水平，橙色代表lncRNA-1在卵巢中的表达水平高于其他组织中表达水平的倍数（log2FC）。它在上皮组织、人肺成纤维细胞和肌肉组织中是没有表达的，而在卵巢中的log2（FPKM+0.01）值为9.8，高于其他组织的4倍或4倍以上。

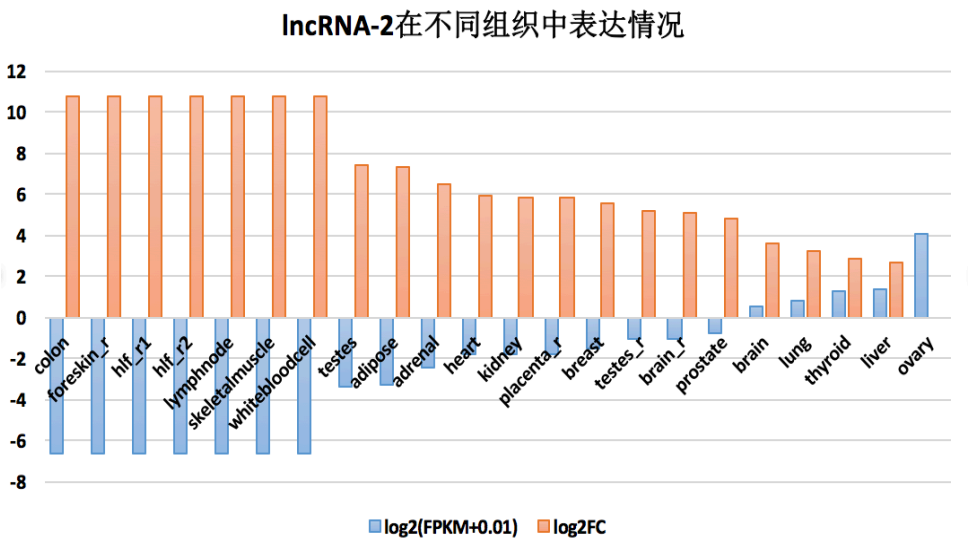


图3.10 1140056号样本包含部分CNV缺失的lncRNA在不同组织中的表达情况。蓝色代表绝对表达水平，橙色代表lncRNA-1在卵巢中的表达水平高于其他组织中表达水平的倍数（log2FC）。它在结肠、上皮组织、人肺成纤维细胞、淋巴结和肌肉组织中是没有表达的，而在卵巢中的log2（FPKM+0.01）值为17.08，表达量是其他组织的4倍以上。

3.2.2 应用GTEx数据集筛选卵巢特异表达lncRNA

Cabili分析的数据，人体组织只有一个或两个样本，没有生物学重复，样本来源是美籍非裔或高加索人，而lncRNA在人的个体中是有一定表达差异性的[26]。为了排除实验误差、个体和人群的差异，进一步验证芯片检测到lncRNA-1和lncRNA-2是否在卵巢中特异高表达，本研究应用了更为全面的GTEx数据集进行[30]分析。

2016年更新的GTEx V6版本包含了来自544个捐献者的53种不同组织，样本量达到8555个（图3.11），其中包括了1.0%的亚洲人、13.7%的美籍非裔和84.3%的白人。GTEx数据集将一个器官组织更为精细地分为不同的部分，如将大脑分为杏仁核、前扣带皮质、尾状核、小脑、额叶皮质、海马、下丘脑、伏核、壳核、脊髓、黑质区等不同区域。GTEx数据集中共有97个卵巢样本，但没有被细分为不同区域进行研究，依然被作为一个整体，所以只能在组织器官水平上分析卵巢的表达谱。该数据集详细地包含了一个基因座下不同转录本的表达水平RPKM值。

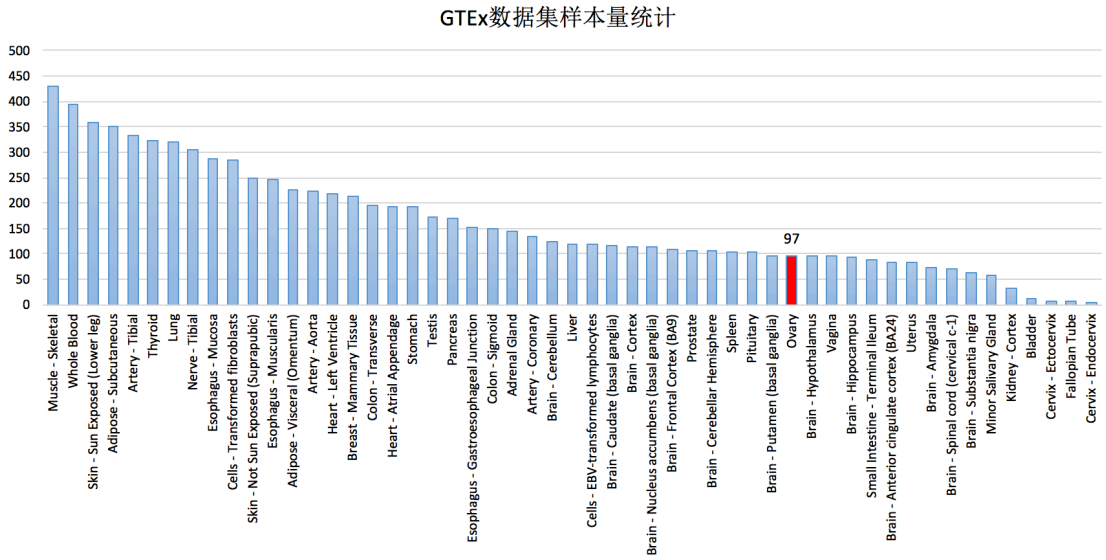


图3.11 GTEx数据集样本量统计，53种不同组织共8555个样本，来自于544个捐献者。

因为GTEx数据集有多个生物学样本重复，所以判断卵巢组织特异性高表达的转录本的方法为，用t检验确定某个转录本在卵巢和其他不同组织中的表达水平是否有显著差异，当结果满足Bonferroni矫正后的P值≤0.05，同时在卵巢的表达水平是其他组织的两倍数以上时（log2标准化后的值相差1则表达倍数差异为2倍），则认为该转录本在卵巢中特异高表达。由于筛选得到115个转录本（附表

2），其中4个lncRNA的转录本，共涉及到92个不同的基因，其中包含了CNV芯片筛查的3个候选基因，分别是*MOB3B*、*PDGFRA*和*TYRO3*。

lncRNA-1在Ensembl中被注释的基因ID是ENSG00000233967，它有三个转录本，分别是ENST00000427048、ENST0000452402和ENST00000443221（图3.12）。首先用log2（RPKM＋0.01）进行变换，再用卵巢标准化后的值减去其他组织的log2（RPKM＋0.01）值，大于1则表示倍数变化为2倍（图3.13和3.14）。转录本ENST00000427048的表达水平是其他所有组织的2倍以上，相对较高；转录本ENST0000452402的表达水平也高于其他所有组织，但与唾液腺、垂体、神经组织的表达差异没有达到2倍以上，但它的绝对表达水平是三个转录本中最高的；转录本ENST0000443221在肝脏中的表达是最高的，其次是在卵巢中。这三个转录本在全血、骨骼肌和左心室等组织表达最低。通过t检验发现，这三个转录本在卵巢和输卵管、唾液腺、外子宫颈与内子宫颈等组织的表达差异是不显著的，考虑到输卵管、子宫颈和卵巢都是属于生殖系统，它们之间的表达差异可能与其他组织相比会小一些。

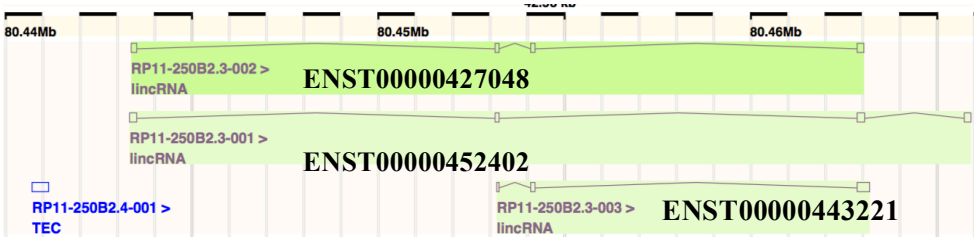


图3.12 lncRNA-1（ENSG00000233967）的基因模型。它的三个转录本为：ENST00000427048（Chr 6: 80,443,390-80,463,053），ENST00000452402（Chr6: 80,443,344-80,465,927），ENST00000443221（Chr6: 80,453,199-80,463,207）。

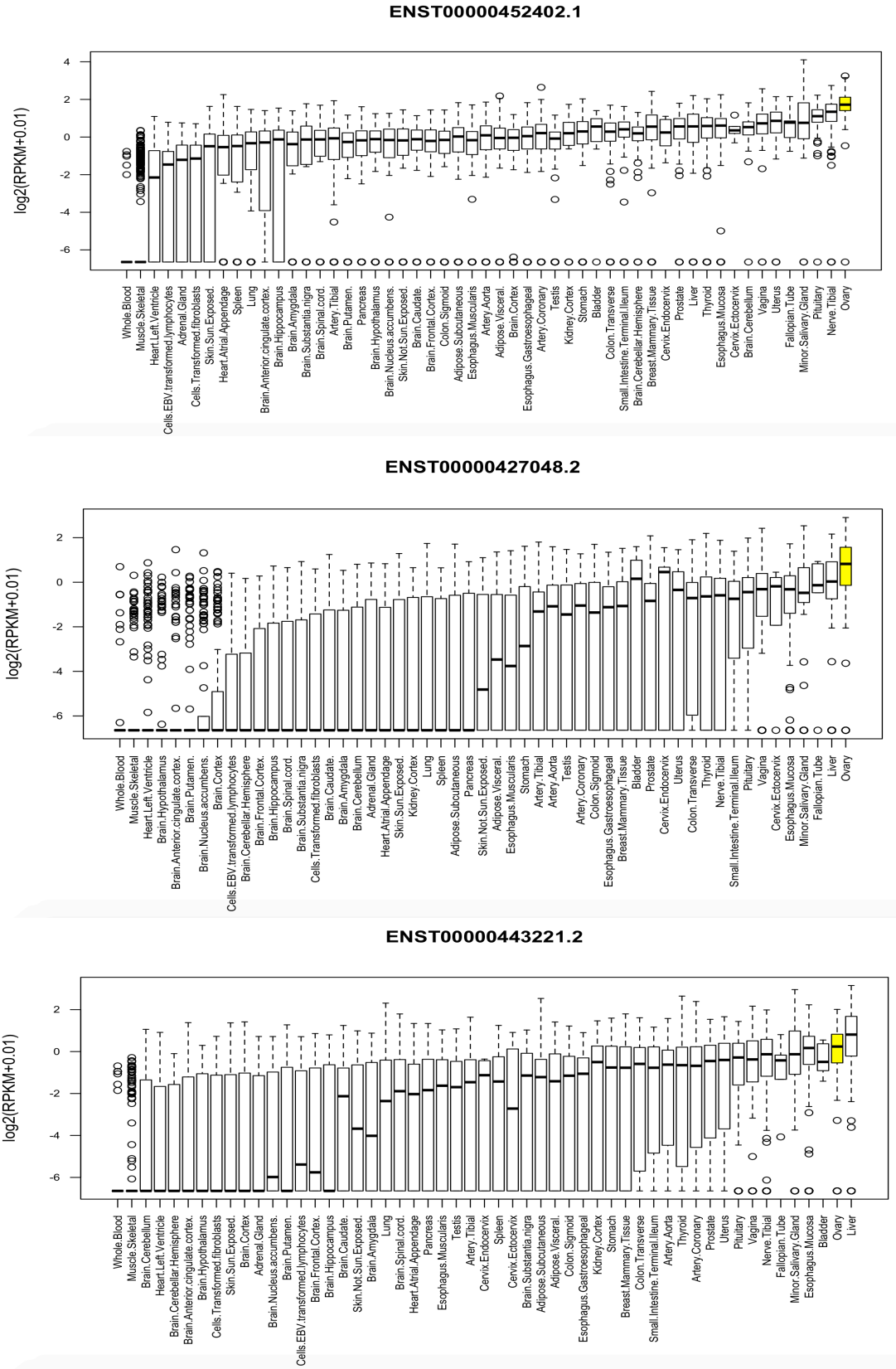


图3.13 lncRNA-1（ENSG00000233967）的三个转录本ENST00000452402、ENST00000427048、ENST00000443221的log2(RPKM+0.01)表达水平的箱线图，以表达水平的平均值排序。三个转录本的表达水平与其他组织相比都较高，其中ENST00000452402的绝对表达水平最高。每个转录本在相同组织中表达的差异性也是很大的。

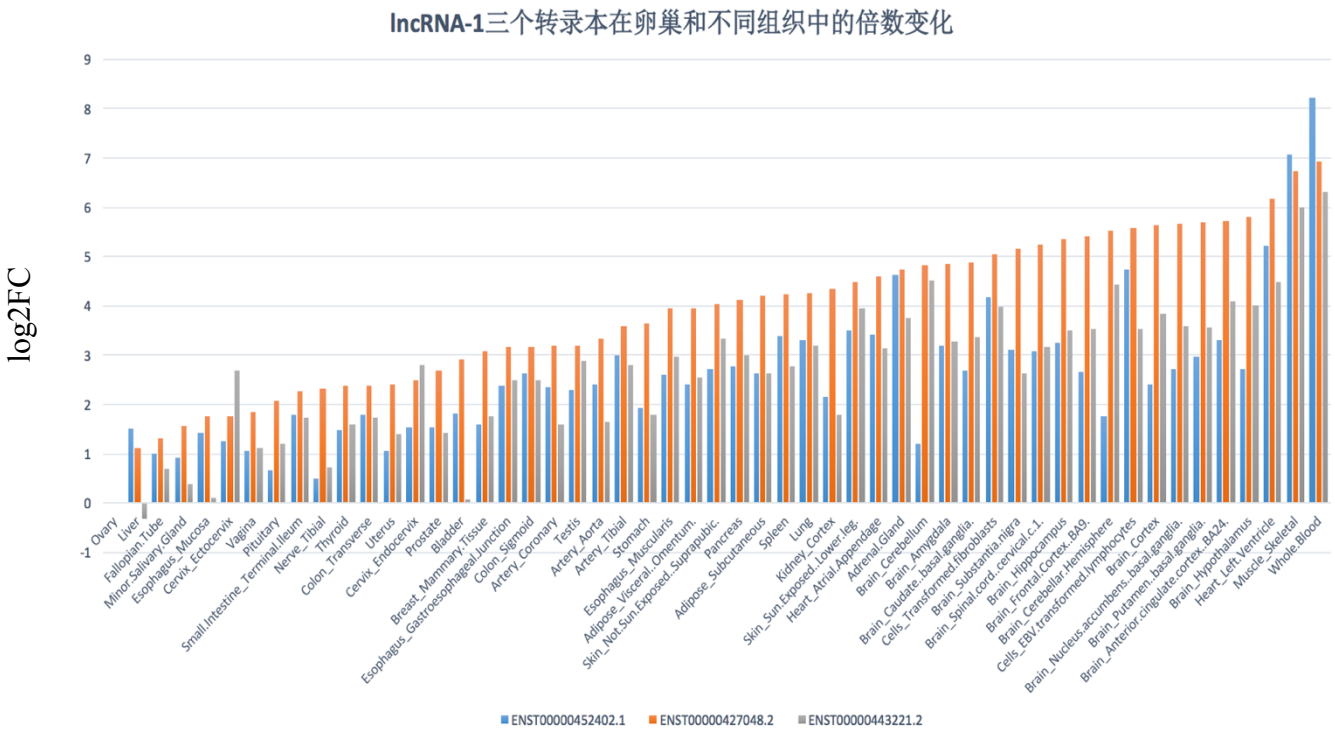


图3.14 lncRNA-1的三个转录本在卵巢和其他不同组织中表达差异的倍数变化，统一在log2（RPKM＋0.01）的水平上，值相差1则表示表达差异为2倍

lncRNA-2在Ensembl中被注释的基因ID是ENSG00000253671，它有两个转录本，分别是ENST00000520646和ENST00000520156（图3.15、3.16和3.17）。计算方法与上述相同，结果表明转录本ENST00000520646在肾上腺、唾液腺、胫动脉和输卵管的表达量高于卵巢，在肾上腺中的表达量是卵巢的4倍。转录本ENST00000520156在肾上腺、唾液腺、胰腺和前列腺的表达水平高于卵巢。通过t检验发现，这两个转录本在胰腺、前列腺、子宫颈、唾液腺和输卵管中的表达水平与其在卵巢中的表达水平不存在显著差异。

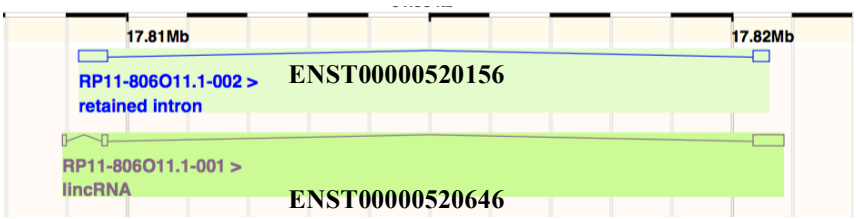


图3.15 lncRNA-2（ENSG00000253671）的基因模型，它的两个转录本为：ENST00000520646（Chromosome 8: 17,808,941-17,820,868），ENST00000520156（Chromosome 8: 17,809,211-17,820,625）

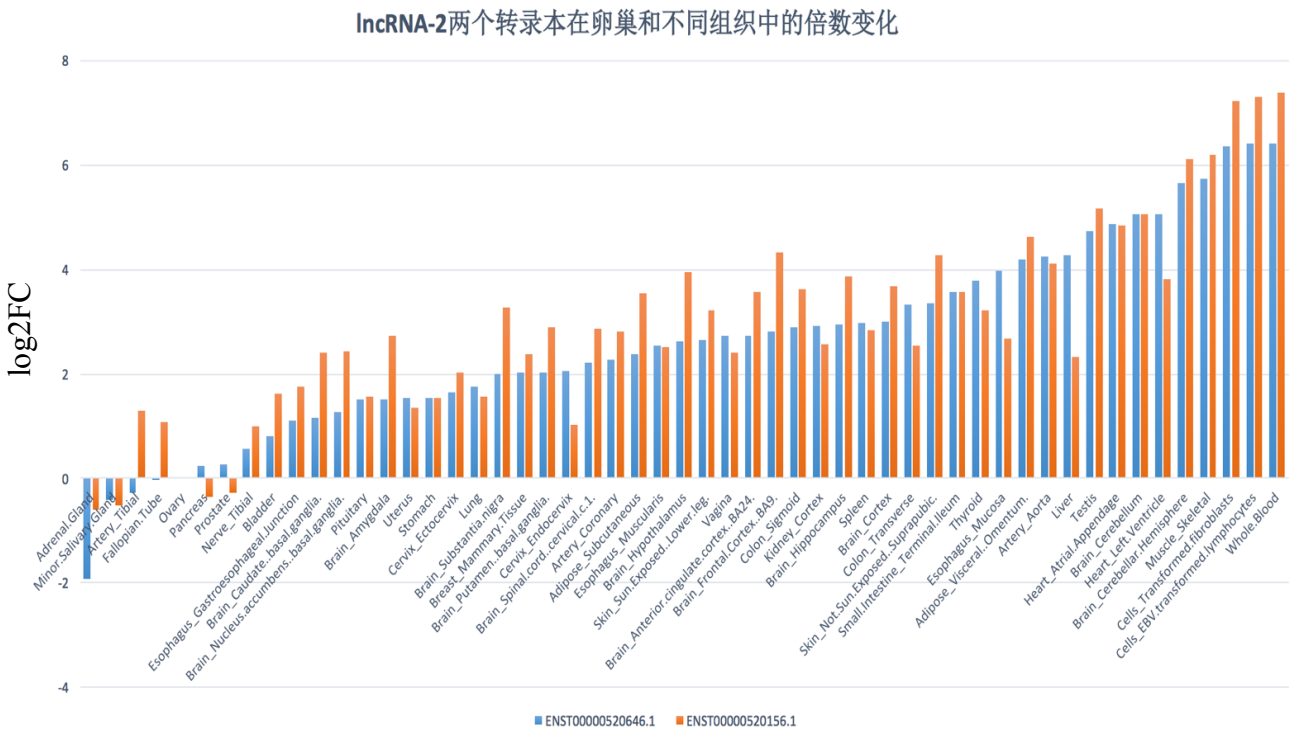
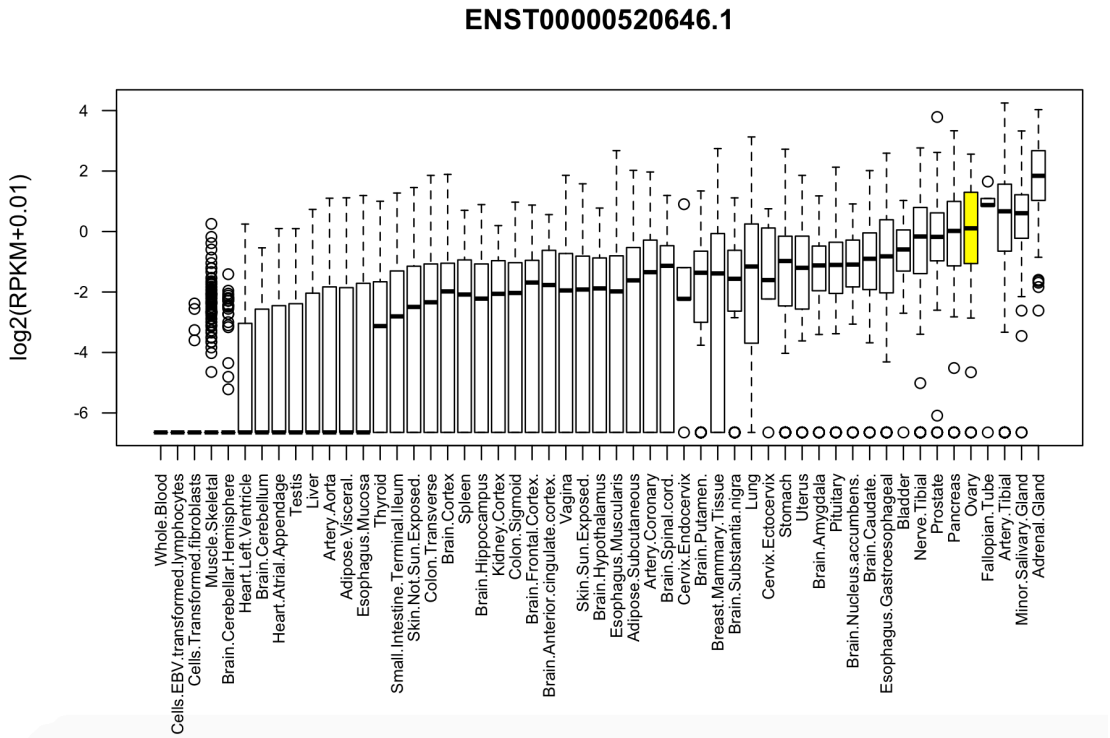


图3.16 lncRNA-2的两个转录本在卵巢和其他不同组织中表达差异的倍数变化，统一在log2（RPKM＋0.01）的水平上，值相差1则表示表达差异为2倍。



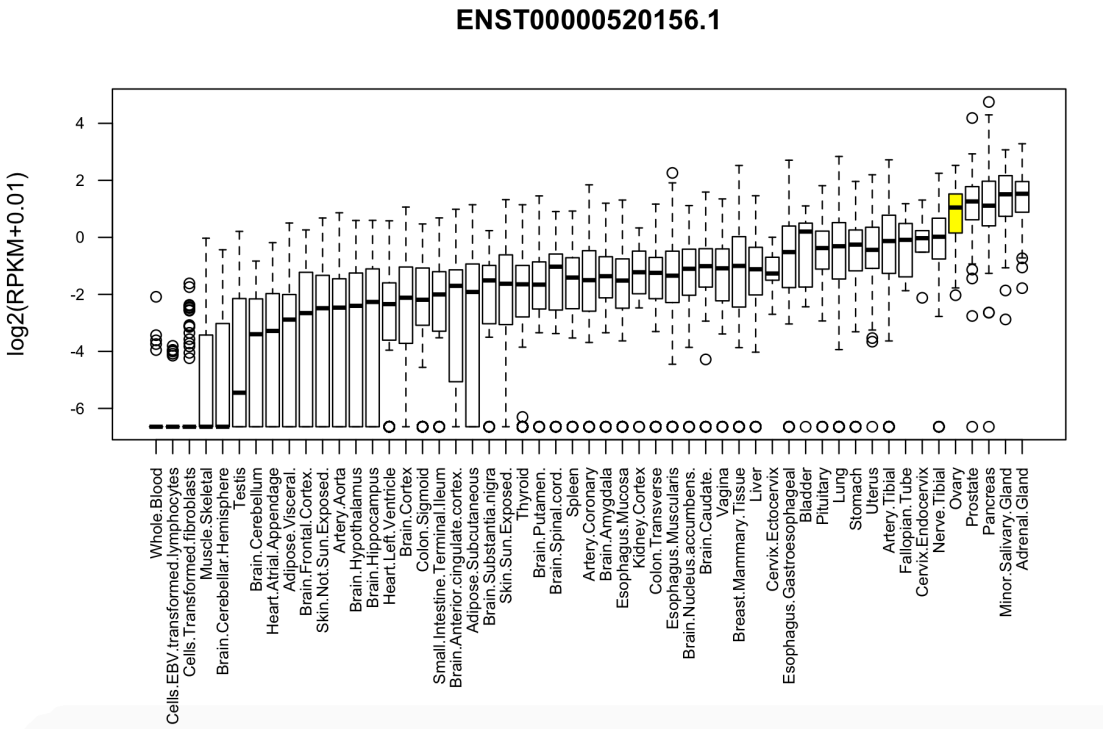


图 3.17 lncRNA-2（ENSG00000253671）的两个转录本ENST00000520646和ENST00000520156的log2（RPKM＋0.01）表达水平的箱线图，以表达水平的平均值排序。虽然两个转录本的绝对表达水平，即RPKM值在卵巢中不是最高的，但是也高于大多数其他组织。每个转录本在相同组织中表达的差异性也是很大的。

GTEx数据集中共有195746个转录本，其中有11609个lncRNA转录本，将这些lncRNA的转录本在卵巢中表达水平的平均值从高到低排序。lncRNA-1的转录本ENST00000452402的RPKM平均值为3.57，排在285位，转录本ENST00000427048的RPKM平均值为2.02，排在443位，转录本ENST00000443221的RPKM平均值为1.32，排在619位。lncRNA-2的转录本ENST000000520646的RPKM平均值为1.63，排在541位，转录本ENST00000520156是retained intron，不是lncRNA。

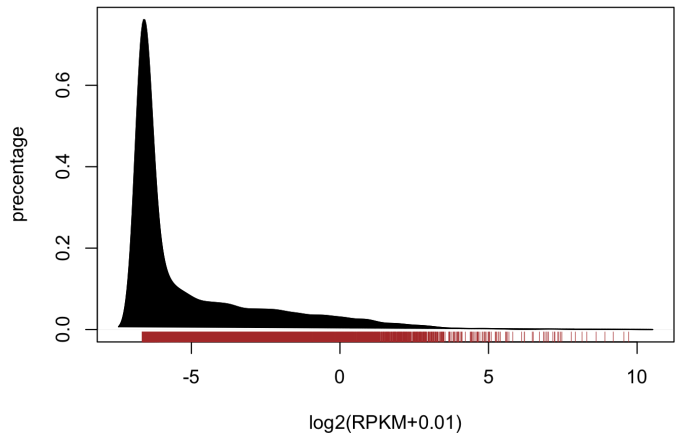


图3.18　GTEx数据集中卵巢中lncRNA表达水平的分布。

通过GTEx数据验证，lncRNA-1的ENST00000427048和ENST00000452402两个转录本在卵巢中的表达量最高，lncRNA-1的转录本ENST00000443221和lncRNA-2的ENST000000520646和ENST00000520156两个转录本虽然不是在卵巢中表达量最高，但与大部分其他组织相比表达量是较高的。结果提示，lncRNA-1很可能与患者的致病机理有关，值得进一步研究其生物学功能。

目前对lncRNA-1和lncRNA-2的研究很少，为了探究它们的生物学功能，我对它们的五个转录本与编码蛋白的转录本进行了共表达分析，通过对共表达的编码蛋白基因的KEGG生物学通路和GO（Gene Ontology）生物学功能注释，可提示这两个lncRNA参与哪些生物学过程中。lncRNA与编码蛋白转录本的共表达分为两种情况：（1）lncRNA与编码蛋白的转录本在不同的组织中的表达水平呈正相关变化，即lncRNA在某种组织中表达水平较高时，编码蛋白转录本也在这种组织中表达水平较高（图3.19）；（2）lncRNA与编码蛋白的转录本在不同组织中的表达水平呈负相关变化，即lncRNA在某种组织中表达水平较高时，编码蛋白转录本却在这种组织中表达水平较低（图3.20）。本研究计算了lncRNA-1和lncRNA-2与所有编码蛋白转录本的相关系数，选取了与两个lncRNA呈正相关或负相关的前100个编码蛋白转录本进行了KEGG通路分析和GO生物学功能分析（表3.3-3.5）。通过分析发现，与lncRNA-1共表达的编码蛋白转录本主要与胰岛素代谢、磷酸肌醇代谢和mRNA的转录调控有关；与lncRNA-2共表达的编码蛋白转录本主要与DNA断裂修复和核苷酸合成有关。

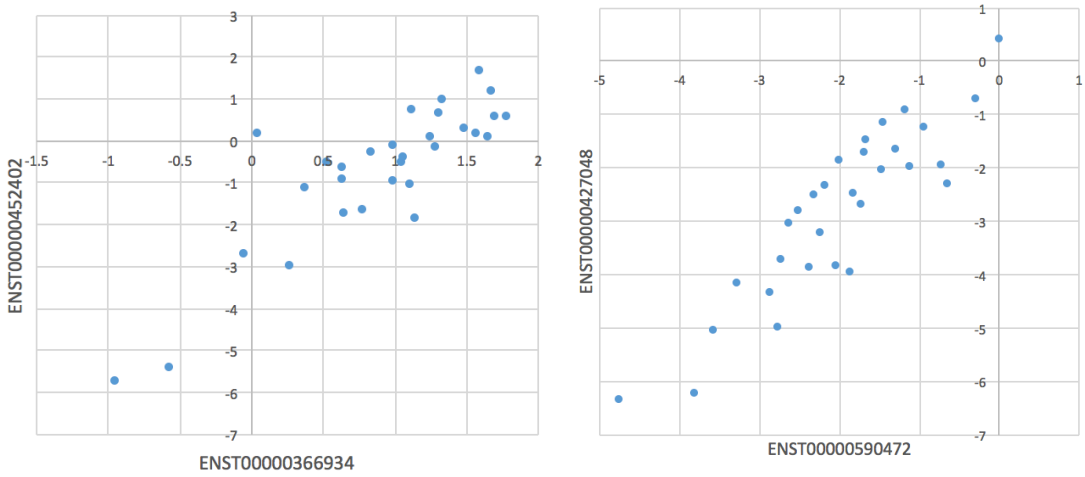


图 3.19 lncRNA 与编码蛋白转录本在不同组织中的表达水平呈正相关。lncRNA-1 的 ENST00000452402 转录本与编码蛋白基因转录本 ENST00000366934 的相关系数为 0.856；lncRNA-1 的 ENST00000427048 转录本与编码蛋白基因转录本 ENST00000590472 的相关系数为 0.90。

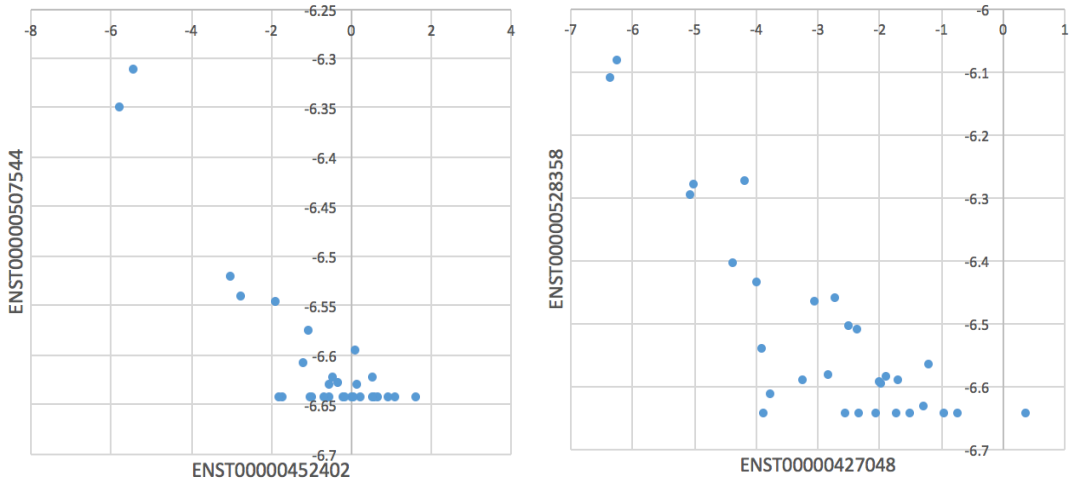


图 3.20 lncRNA 与编码蛋白转录本在不同组织中的表达水平呈负相关。lncRNA-1 的 ENST00000452402 转录本与编码蛋白基因转录本 ENST00000507544 的相关系数为-0.88；lncRNA-1 的 ENST00000427048 转录本与编码蛋白基因转录本 ENST00000528358 的相关系数为-0.84。

表 3.3 与 lncRNA-1 和 lncRNA-2 在不同组织中表达水平相关的编码蛋白转录本的 KEGG 通路分析

lncRNA 转录本名称	正相关	负相关
ENST00000452402	Glycosylphosphatidylinositol(GPI)-anchor	Purine Metabolism
ENST00000443221	Adherens junction	NA
ENST00000427048	NA	T cell receptor signaling pathway
ENST00000520646	Pyrimidine metabolism	NA
ENST00000520156	NA	NA

表 3.4 与 lncRNA-2 在不同组织中表达水平相关的编码蛋白转录本的 GO 功能分析

lncRNA 转录本名称	正相关	负相关
ENST00000520646	pyrimidine nucleotide metabolic process	DNA damage response, signal transduction resulting in induction of apoptosis
	vesicle-mediated transport	induction of apoptosis by intracellular signals
	nucleotide biosynthetic process	response to radiation
	hexose metabolic process	DNA damage reponse, signal transduction
	nucleobase, nucleoside, nucleotide and nucleic acid biosynthetic process	protein localization
	nucleobase, nucleoside and nucleotide biosynthetic process	phosphate metabolic process
	positive regulation of binding	phosphorus metabolic process
	monosaccharide metabolic process	protein transport
	histone modification	establishment of protein localization
	covalent chromatin modification	response to light stimulus
ENST00000520156	positive regulation of binding	NA
	nucleobase, nucleoside and nucleotide biosynthetic process	
	nucleobase, nucleoside, nucleotide and nucleic acid biosynthetic process	
	regulation of Ras protein signal transduction	
	regulation of Rho protein signal transduction	
	tRNA modification	
	regulation of small GTPase mediated signal transduction	

表 3.5 与 lncRNA-1 在不同组织中表达水平相关的编码蛋白转录本的 GO 功能分析

lncRNA 转录本名称	正相关	负相关
ENST00000452402	GPI anchor biosynthetic process	RNA localization
	GPI ancchor metabolic process	RNA splicing
	phosphoinositide biosynthetic process	chromatin assembly or disassembly
	protein amino acid lipidation	
	lipoprotein biosynthetic process	mRNA processing
	regulation of proton transport	
	glycerophospholipid biosynthetic process	pyrimidine nucleoside metabolic process
	ribonuleoprotein complex biogenesis	
	phosphoinositide metabolic process	mRNA metabolic process
	ncRNA processing	
ENST0000044322221	reponse to copper ion	RNA splicing
		mRNA processing
		mRNA metabolic process
	acute inflammatory response	chromatin assembly or disassembly
		cofactor transport
		organelle fusion
	regulation of insulin receptor signaling pathway	vitamin transport
		RNA processing
		RNA splicing, via transesterification reactions
		nuclear mRNA splicing, via spliceosome
ENST00000427048	regulation of insulin receptor signaling pathway	cofactor transport
	rRNA processing	
	rRNA metabolic process	
	negative regulation of cell size	vitamin transport
	negative regulation of insulin receptor signaling pathway	
	ribosome biogenesis	

四、讨 论

本课题首先应用比较基因组杂交芯片对卵巢早衰病患进行基因拷贝数变异分析，在25例病人中发现两例分别携带两个不同的CNV缺失，影响到了两个在卵巢中高表达的lncRNA loci（ENSG00000233967和ENSG00000253671）。通过对多组织的RNA-seq表达谱数据的分析得知这两个lncRNA在正常卵巢中的表达水平是其他组织的2倍以上，呈现明显的卵巢高表达特异性。应用GTEx数据集对结果进行验证，发现ENSG00000233967的ENST00000427048和ENST0000045240两个转录本在卵巢中的表达量最高，且比其他所有组织高两倍以上。ENSG00000233967 的转录本 ENST00000443221 和 ENSG00000253671 的ENST000000520646和ENST00000520156两个转录本虽然不是在卵巢中表达量最高，但与其他大部分组织相比是表达量较高的，所以这两个lncRNA在卵巢中是较高表达的。

从基因拷贝数变异分析可知，lncRNA-1（ENSG00000233967）缺失完整的一拷贝，可能会导致表达量下调，因单倍剂量不足导致卵巢早衰，lncRNA-2（ENSG00000253671）缺少一个拷贝的三分之一，可能会导致表达降低或者编码出错误、无功能的蛋白影响卵巢的正常功能。

通过lncRNA与编码蛋白的共表达分析了这两个lncRNA可能参与的生物学通路和过程，发现与lncRNA-1共表达的编码蛋白转录本主要与胰岛素代谢、磷酸肌醇代谢和mRNA的转录调控有关；与lncRNA-2共表达的编码蛋白转录本主要与DNA断裂修复和核苷酸合成有关。

对这两个卵巢高表达的lncRNA的相关研究极少。两者与其他物种相比保守性较低，只在灵长类动物之间较保守，无法简单地用小鼠模型进行基因敲低或敲除研究它们的功能。lncRNA的表达具有组织特异性，而卵巢中有不同的细胞种类，有卵子、颗粒细胞、膜细胞、卵丘细胞、透明带、黄体等。为了研究这两个lncRNA在生殖发育中发挥什么作用，下一步研究计划是获得卵巢中不同细胞和常用研究卵巢功能的细胞系如KGN细胞的RNA-seq表达谱，找出前面发现的两个lncRNA在哪种细胞中表达量最高，为后续的功能实验找到合适的细胞实验材料。

参考文献

1. Conway GS (2000) Premature ovarian failure. *Br Med Bull* 56,643–649.
2. Goswami, D., & Conway, G. S. (2005). Premature ovarian failure. *Human Reproduction Update*, *11*(4), 391–410.
3. Qin, Y., Jiao, X., Simpson, J. L., & Chen, Z.-J. (2015). Genetics of primary ovarian insufficiency: new developments and opportunities. *Human Reproduction Update*, *21*(6), 787–808.
4. Yan C et al. Synergistic roles of bone morphogenetic protein 15 and growth differentiation factor 9 in ovarian function. *Mol Endocrinol* 2001;6:854 – 866.
5. Shiina H et al. Premature ovarian failure in androgen receptor-deficient mice. *Proc Natl Acad Sci USA* 2006;1:224–229.
6. Rajkovic A et al. NOBOX deficiency disrupts early folliculogenesis and oocyte-specific gene expression. *Science* 2004; 5687:1157 – 1159.
7. Mansouri MR et al. Alterations in the expression, structure and function of progesterone receptor membrane component-1 (PGRMC1) in premature ovarian failure. *Hum Mol Genet* 2008;23:3776–3783.
8. Stankiewica P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*, 2010, 61, 437-455
9. Kondrashov AS. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian disease. *Hum Mutat*, 2003, 21(1):12-27
10. Flores M et al. 2007. Recurrent DNA inversion rearrangements in the human genome. *Proc. Natl. Acad. Sci. USA* 104:6099–106
11. Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* 18:74–82
12. Stankiewicz P et al. 2003. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *Am. J. Hum. Genet.* 72:1101–16
13. LieberMR, LuH, GuJ, SchwarzK. 2008. Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate nonhomologous DNA end joining: relevance to cancer, aging, and the immune system. *Cell Res.* 18:125–33
14. LieberMR,MaY,PannickeU,SchwarzK.2003.Mechanismandregulationofhumann onhomologous DNA end-joining. *Nat. Rev. Mol. Cell Biol.* 4:712–20
15. Zhang F, Carvalho CM, Lupski JR. 2009. Complex human chromosomal and genomic rearrangements. *Trends Genet.* In press, doi: 10.1016/j.tig.2009.05.005
16. Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* 5:e1000327
17. Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, *10*(1), 451–481.
18. Lupski, J.R. and P. Stankiewicz. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet*, 2005, 1(6): e49.

19. Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell*, *136*(4), 629–641.
20. Cabili, M. et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915–1927.
21. Wapinski, O., & Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends in Cell Biology*, *21*(6), 354–361.
22. Yap, K.L. (2010) Molecular Interplay of the non coding RNA ANRIL and methylated histone H3 Lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674
23. Tripathi, V. et al. (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* 39, 925–938
24. Ji, P. et al. (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041
25. Faghihi, M.A. et al. (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med*. 14, 723–730
26. Kornienko et al. (2016). Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biology*, 1–23.
27. Guo, L. et al. (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology*, 24(9), 1162–1169.
28. Yu Y, Fuscoe JC, Zhao C, et al. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nature communications.* 2014, 5: 3230.
29. Lage K et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proceedings of the National Academy of Sciences of the United States of America. 2008, 105(52): 20870-5.
30. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015, 348(6235): 648- 60.

附表 1　Illumina BodyMap2 数据集中卵巢特异高表达 lncRNA

Chrom	Chrom_start	Chrom_end	Name	FPKM
chr10	67330424	67332771	TCONS_00018220	2.8223
chr11	3533504	3566749	TCONS_00019829	1.6265
chr14	24161842	24168143	TCONS_00022430	3.4895
chr14	24161842	24168143	TCONS_00022429	3.4895
chr14	24164069	24168143	TCONS_00022431	3.4895
chr14	104755160	104756411	TCONS_l2_00008273	2.3102
chr14	104755293	104755944	TCONS_l2_00008274	2.3102
chr16	65166561	65173706	TCONS_00024436	1.5813
chr16	65171565	65172724	TCONS_00024861	1.5813
chr16	65175262	65209216	TCONS_00024696	2.2585
chr16	65175530	65210664	TCONS_00025033	2.2585
chr16	65175530	65210664	TCONS_00025032	2.2585
chr16	68756327	68761392	TCONS_00024446	3.9991
chr19	22674955	22676132	TCONS_00026954	1.5967
chr19	50472300	50474212	TCONS_l2_00012612	4.884
chr19	50473579	50474548	TCONS_l2_00013333	4.884
chr2	10702420	10706471	TCONS_00002818	5.1698
chr2	101802574	101807916	TCONS_00004353	2.0236
chr2	105950390	105952068	TCONS_00004365	5.4495
chr2	105950391	105953445	TCONS_00004366	5.4495
chr2	130958856	130962255	TCONS_l2_00013983	5.6482
chr20	46653531	46702929	TCONS_00028194	5.9315
chr20	46653997	46691515	TCONS_00028195	5.9315
chr20	46683232	46692077	TCONS_00028559	5.9315
chr20	46692275	46700463	TCONS_00028560	5.9315
chr20	46698146	46700081	TCONS_00028196	5.9315
chr20	60520107	60523627	TCONS_00028253	1.3968
chr3	129930775	129963708	TCONS_l2_00018856	7.2553
chr3	129931088	129937145	TCONS_l2_00018857	7.2553
chr3	129931104	129931606	TCONS_l2_00018858	7.2553
chr3	129931662	129992648	TCONS_l2_00019883	7.2553
chr3	129931719	129969881	TCONS_l2_00019884	7.2553
chr3	129964290	129967295	TCONS_l2_00019885	7.2553
chr3	129964592	129992631	TCONS_l2_00018859	7.2553
chr3	129969980	129982803	TCONS_l2_00018860	7.2553
chr3	129983284	129986284	TCONS_l2_00019886	7.2553
chr3	184433379	184477461	TCONS_00006332	7.2809
chr3	184433592	184456543	TCONS_00007053	7.3054
chr3	184433592	184445372	TCONS_00007052	7.3054
chr3	184433598	184457421	TCONS_00006333	7.3054

chr3	184433642	184446675	TCONS_00006334	7.3054
chr3	184433831	184456715	TCONS_00006335	7.3054
chr3	184445192	184456749	TCONS_00006336	7.3054
chr3	184451846	184456867	TCONS_00006337	7.3054
chr3	184455325	184456747	TCONS_00007054	7.3054
chr4	11742037	11771223	TCONS_00008436	2.2296
chr4	11742571	11771092	TCONS_00007716	2.2296
chr4	11744478	11748653	TCONS_00009028	2.2296
chr5	36372265	36397435	TCONS_00010302	1.054
chr5	38821814	38845844	TCONS_l2_00022858	8.1015
chr5	38825834	38845924	TCONS_l2_00022859	8.1015
chr5	38843569	38845913	TCONS_l2_00022860	8.1015
chr6	20042902	20045304	TCONS_l2_00024643	9.2572
chr6	32360319	32361247	TCONS_00012469	9.4032
chr6	81128030	81172878	TCONS_00011870	9.7311
chr6	81151576	81163845	TCONS_00011871	9.8096
chr6	81153001	81172632	TCONS_00012537	9.8096
chr6	81153060	81175644	TCONS_00011332	9.8096
chr6	81153072	81162975	TCONS_00011872	9.8161
chr6	81153114	81172913	TCONS_00011873	9.8161
chr6	81153172	81175572	TCONS_00011334	9.8161
chr6	81153172	81172924	TCONS_00011333	9.8161
chr6	81164224	81172890	TCONS_00011874	9.8161
chr6	86367931	86373867	TCONS_l2_00024837	9.8161
chr6	149912890	149913799	TCONS_00011593	9.8161
chr6	166253091	166253578	TCONS_00012358	1.5307
chr6	166315364	166324507	TCONS_00012825	9.8161
chr6	166315364	166324507	TCONS_00012824	9.8161
chr6	166722952	166731019	TCONS_l2_00024517	9.8161
chr6	166723013	166731019	TCONS_l2_00024518	9.8161
chr6	166723264	166731401	TCONS_l2_00024519	9.8161
chr6	169818906	169830893	TCONS_00012370	9.8161
chr6	169818924	169846389	TCONS_00012371	9.8233
chr6	169819100	169825251	TCONS_00011248	9.829
chr6	169821175	169828669	TCONS_00012831	9.829
chr6	169825998	169830792	TCONS_00012372	9.829
chr6	169826471	169828669	TCONS_00012832	9.829
chr6	169827132	169830752	TCONS_00012373	9.829
chr6	169827784	169829155	TCONS_00012833	9.829
chr6	169828814	169830843	TCONS_00012834	9.829
chr7	2761465	2764726	TCONS_00014248	9.8906
chr7	39803197	39807261	TCONS_00013783	10.276
chr7	63539166	63545179	TCONS_l2_00027127	10.791

chr7	63539607	63545689	TCONS_l2_00025931	10.791
chr7	63539998	63546016	TCONS_l2_00025932	10.791
chr7	65112776	65183632	TCONS_l2_00027132	10.791
chr7	65121845	65174629	TCONS_l2_00025967	10.791
chr7	65121856	65154426	TCONS_l2_00025968	10.791
chr7	65121886	65150708	TCONS_l2_00025969	10.791
chr7	65121932	65167794	TCONS_l2_00025970	10.791
chr7	65124599	65183669	TCONS_l2_00025971	10.791
chr7	65156536	65160173	TCONS_l2_00025972	10.791
chr7	65162818	65173519	TCONS_l2_00025973	10.791
chr7	65171484	65173506	TCONS_l2_00025974	10.791
chr7	65180339	65183782	TCONS_l2_00025975	10.791
chr7	79085480	79096779	TCONS_00013201	14.118
chr7	79088786	79093984	TCONS_00014342	14.118
chr7	128166277	128170972	TCONS_00013585	14.198
chr7	130033936	130035446	TCONS_00013589	14.613
chr7	150040314	150054583	TCONS_00013626	2.2771
chr7	150040397	150045221	TCONS_00013627	2.2771
chr7	150040715	150055006	TCONS_00013628	2.2771
chr8	17659187	17678125	TCONS_00014631	17.076
chr8	17661217	17679981	TCONS_00014632	17.076
chr8	17665280	17679892	TCONS_00014633	17.115
chr8	17666370	17678134	TCONS_00015235	17.115
chr8	28915362	28922437	TCONS_00015261	17.115
chr8	28915898	28922445	TCONS_00015262	17.115
chr8	30593692	30594132	TCONS_00014964	17.115
chr8	49293271	49297763	TCONS_l2_00027748	17.115
chr8	49293332	49296676	TCONS_l2_00027749	17.115
chr8	49293568	49297798	TCONS_l2_00027750	17.115
chr8	49293755	49297737	TCONS_l2_00027751	17.115
chr8	49293765	49294938	TCONS_l2_00027752	17.115
chr8	50080641	50105812	TCONS_00014696	1.2737
chr8	50080641	50105812	TCONS_00014695	1.2737
chr9	6645888	6670845	TCONS_l2_00028611	17.606
chr9	6645888	6669881	TCONS_l2_00028610	17.606
chr9	6668943	6670724	TCONS_l2_00028612	17.606
chr9	37509146	37510296	TCONS_00015668	18.275
chr9	68743530	68769869	TCONS_l2_00029973	1.3863
chr9	89562975	89616947	TCONS_00015577	22.73
chr9	89563075	89566324	TCONS_00016041	22.73
chr9	89563557	89613018	TCONS_00016042	23.247
chr9	89563565	89618374	TCONS_00016043	23.299
chr9	89563619	89611321	TCONS_00016044	23.299

chr9	89563703	89565889	TCONS_00016045	23.437
chr9	89612094	89613812	TCONS_00016046	24.095
chr9	130872812	130880957	TCONS_00016871	25.302
chr9	130873449	130880972	TCONS_00015886	25.302
chr9	130873451	130874237	TCONS_00016872	25.302
chr9	130873736	130880116	TCONS_00016873	25.302
chr9	130875159	130877568	TCONS_00016480	25.302
chr9	136125806	136126767	TCONS_00016890	26.972
chr9	139141824	139157976	TCONS_00016506	26.972
chr9	139144132	139148691	TCONS_00016895	26.972
chr9	139144132	139148691	TCONS_00016894	26.972
chr9	139144280	139147807	TCONS_00016507	26.972
chr9	139147883	139166993	TCONS_00016508	26.972
chr9	139150396	139166917	TCONS_00016898	26.972
chr9	139150396	139166917	TCONS_00016896	26.972
chr9	139150396	139166917	TCONS_00016897	26.972
chr9	139150537	139166849	TCONS_00016509	26.972
chr9	139150740	139159464	TCONS_00016510	26.972
chr9	139151964	139157976	TCONS_00016899	26.972
chr9	139158250	139166849	TCONS_00016901	26.972
chr9	139158250	139166849	TCONS_00016900	26.972
chr9	139159243	139166900	TCONS_00016511	26.972
chr9	139159823	139166857	TCONS_00016902	26.972
chr9	139166405	139166839	TCONS_00016512	26.972
chrX	281384	282052	TCONS_00016915	26.972
chrX	281384	282052	TCONS_00016916	26.972
chrX	281388	284630	TCONS_l2_00030415	26.972
chrX	281390	284839	TCONS_l2_00030416	26.972
chrX	281395	284747	TCONS_l2_00030417	27.189
chrX	281724	282586	TCONS_00016957	27.203
chrX	282382	285848	TCONS_l2_00030418	28.477
chrX	2527305	2575270	TCONS_l2_00030653	29.191
chrX	2527305	2556369	TCONS_l2_00030652	29.191
chrX	2527380	2556679	TCONS_l2_00030120	30.943
chrX	2527386	2544686	TCONS_l2_00030121	30.943
chrX	2527388	2534212	TCONS_l2_00030654	30.943
chrX	2527401	2556367	TCONS_l2_00030122	30.943
chrX	2527432	2575270	TCONS_l2_00030655	30.943
chrX	2527515	2544744	TCONS_l2_00030124	32.688
chrX	2527515	2537621	TCONS_l2_00030123	32.688
chrX	2529038	2576883	TCONS_l2_00030125	34.724
chrX	2530195	2576587	TCONS_l2_00030126	34.724
chrX	2530208	2535063	TCONS_l2_00030127	34.724

chrX	2533992	2556735	TCONS_l2_00030656	35.321
chrX	2536278	2556288	TCONS_l2_00030128	35.913
chrX	2536689	2556357	TCONS_l2_00030129	35.913
chrX	2536691	2576947	TCONS_l2_00030130	35.913
chrX	2536762	2540943	TCONS_l2_00030131	35.913
chrX	2536787	2556657	TCONS_l2_00030132	35.913
chrX	2537467	2556225	TCONS_l2_00030133	35.913
chrX	2541205	2544744	TCONS_l2_00030134	36.991
chrX	2541375	2556469	TCONS_l2_00030135	36.991
chrX	3820106	3823822	TCONS_l2_00030431	37.078
chrX	3820865	3838787	TCONS_l2_00030777	37.078
chrX	3823287	3838276	TCONS_l2_00030432	37.078
chrX	3823393	3849842	TCONS_l2_00030433	37.078
chrX	3849696	3855883	TCONS_l2_00030434	38.593
chrX	18884024	18884823	TCONS_l2_00030158	41.9
chrX	26704159	26706126	TCONS_l2_00030166	41.988
chrX	26704217	26705932	TCONS_l2_00030167	41.988
chrX	40122130	40140673	TCONS_00017417	41.988
chrX	40122218	40126475	TCONS_00016978	42.8996
chrX	46404927	46407668	TCONS_00016919	42.8996
chrX	46458742	46461839	TCONS_00017504	42.8996
chrX	47657378	47670374	TCONS_l2_00030193	42.8996
chrX	48306414	48307208	TCONS_l2_00030196	42.8996
chrX	48306415	48307284	TCONS_l2_00030197	42.8996
chrX	49155628	49157930	TCONS_00017170	42.8996
chrX	63264304	63265448	TCONS_l2_00030233	48.784
chrX	65014185	65015782	TCONS_l2_00030240	48.784
chrX	65014704	65015586	TCONS_l2_00030241	48.827
chrX	65041568	65041930	TCONS_l2_00030513	52.47
chrX	70534869	70535653	TCONS_l2_00030521	56.229
chrX	73045949	73047819	TCONS_00017432	64.072
chrX	73164158	73290211	TCONS_l2_00030713	69.624
chrX	73164158	73220326	TCONS_l2_00030712	72.767
chrX	73164171	73183329	TCONS_l2_00030250	76.418
chrX	73164176	73234961	TCONS_l2_00030253	77.246
chrX	73164176	73228621	TCONS_l2_00030252	77.246
chrX	73164176	73169688	TCONS_l2_00030251	77.246
chrX	73164182	73164601	TCONS_l2_00030254	77.246
chrX	73164201	73167265	TCONS_l2_00030714	77.246
chrX	73168807	73169393	TCONS_l2_00030715	77.246
chrX	73207972	73230782	TCONS_l2_00030255	77.246
chrX	73216741	73224554	TCONS_l2_00030256	77.246
chrX	73218520	73224686	TCONS_l2_00030257	77.246

chrX	73218605	73225452	TCONS_l2_00030258	77.246
chrX	73289115	73290875	TCONS_l2_00030259	77.246
chrX	73327061	73327716	TCONS_l2_00030260	77.246
chrX	73465728	73495795	TCONS_l2_00030535	77.246
chrX	73492775	73504672	TCONS_l2_00030536	77.246
chrX	73501539	73504697	TCONS_l2_00030833	79.242
chrX	74957498	74966844	TCONS_l2_00030538	79.242
chrX	74959073	74966889	TCONS_l2_00030539	83.172
chrX	89294188	89295476	TCONS_l2_00030550	94.358
chrX	100699053	100787518	TCONS_l2_00030276	99.043
chrX	100699053	100702830	TCONS_l2_00030275	99.043
chrX	100699083	100750930	TCONS_l2_00030277	99.043
chrX	100700985	100754221	TCONS_l2_00030278	99.043
chrX	100740362	100789419	TCONS_l2_00030279	99.043
chrX	100740407	100788446	TCONS_l2_00030724	107.17
chrX	100740407	100788446	TCONS_l2_00030726	118.42
chrX	100740407	100771939	TCONS_l2_00030723	118.42
chrX	100740407	100754031	TCONS_l2_00030722	131.46
chrX	100740436	100760429	TCONS_l2_00030280	131.46
chrX	100740457	100753468	TCONS_l2_00030281	134.47
chrX	100740509	100745181	TCONS_l2_00030282	134.47
chrX	100740882	100754070	TCONS_l2_00030283	134.86
chrX	100740907	100750790	TCONS_l2_00030284	134.86
chrX	100741009	100750790	TCONS_l2_00030285	135.3
chrX	100743428	100753473	TCONS_l2_00030286	135.3
chrX	100743430	100788398	TCONS_l2_00030727	160.14
chrX	100748512	100750788	TCONS_l2_00030287	160.14
chrX	100754056	100764358	TCONS_l2_00030288	160.14
chrX	100759567	100764657	TCONS_00017353	160.14
chrX	100764585	100790666	TCONS_l2_00030289	160.14
chrX	100766011	100787230	TCONS_l2_00030290	160.14
chrX	100766024	100786981	TCONS_l2_00030291	160.14
chrX	100779082	100786916	TCONS_l2_00030292	160.14
chrX	101892937	101894686	TCONS_00017436	160.14
chrX	102024088	102139094	TCONS_l2_00030729	160.14
chrX	102024106	102140334	TCONS_l2_00030730	160.14
chrX	102024108	102094892	TCONS_l2_00030731	160.14
chrX	102024145	102082072	TCONS_l2_00030732	160.14
chrX	102024216	102082052	TCONS_l2_00030298	160.14
chrX	102025332	102082072	TCONS_l2_00030299	160.14
chrX	102071918	102140897	TCONS_l2_00030300	160.14
chrX	102071950	102098298	TCONS_l2_00030301	160.14
chrX	102082048	102089741	TCONS_l2_00030302	160.14

chrX	102085070	102160695	TCONS_l2_00030303	160.14
chrX	102094617	102155728	TCONS_l2_00030304	160.14
chrX	102094769	102155720	TCONS_l2_00030305	160.14
chrX	102094835	102121751	TCONS_l2_00030733	160.14
chrX	102119992	102140218	TCONS_l2_00030306	160.14
chrX	102139070	102152670	TCONS_l2_00030307	160.14
chrX	102139341	102161086	TCONS_l2_00030734	160.14
chrX	102152508	102160619	TCONS_l2_00030735	160.14
chrX	102155692	102170334	TCONS_l2_00030308	160.14
chrX	102156535	102171898	TCONS_l2_00030309	160.14
chrX	102156590	102161770	TCONS_l2_00030310	160.14
chrX	103230501	103232785	TCONS_l2_00030737	160.14
chrX	103230847	103232621	TCONS_l2_00030317	160.14
chrX	103231001	103232047	TCONS_l2_00030318	160.14
chrX	103315075	103316952	TCONS_l2_00030568	163.08
chrX	103315110	103316578	TCONS_l2_00030569	166.41
chrX	103315218	103317502	TCONS_l2_00030839	166.41
chrX	103316416	103317369	TCONS_l2_00030570	166.41
chrX	114937327	114938501	TCONS_l2_00030333	167.884
chrX	119055223	119056803	TCONS_00017085	167.884
chrX	134555991	134561940	TCONS_l2_00030365	167.884
chrX	134556607	134559915	TCONS_l2_00030366	167.884
chrX	135886489	135930734	TCONS_l2_00030613	167.884
chrX	135888045	135897375	TCONS_l2_00030614	167.884
chrX	135889556	135930749	TCONS_l2_00030615	201.33
chrX	135892211	135930743	TCONS_l2_00030616	201.33
chrX	135897099	135904363	TCONS_l2_00030617	219.94
chrX	135991553	136075814	TCONS_l2_00030751	219.94
chrX	135991633	136103777	TCONS_l2_00030752	219.94
chrX	135991643	136104192	TCONS_l2_00030370	219.94
chrX	135996357	136104006	TCONS_l2_00030371	219.94
chrX	136007490	136104280	TCONS_l2_00030372	219.94
chrX	136075645	136103789	TCONS_l2_00030753	219.94
chrX	136075712	136103777	TCONS_l2_00030754	219.94
chrX	136076751	136103735	TCONS_l2_00030373	219.94
chrX	149107956	149109050	TCONS_l2_00030757	219.94
chrX	149107963	149112969	TCONS_l2_00030386	219.94
chrX	149108361	149132635	TCONS_l2_00030387	267.32
chrX	149108373	149185018	TCONS_l2_00030758	267.32
chrX	149109277	149129133	TCONS_l2_00030388	287.26
chrX	149112082	149113822	TCONS_l2_00030389	311.95
chrX	149113092	149129134	TCONS_l2_00030390	311.95
chrX	149113850	149121272	TCONS_l2_00030759	377.52

chrX	149114191	149129256	TCONS_l2_00030391	399.98
chrX	149114193	149115156	TCONS_l2_00030760	399.98
chrX	149114223	149129463	TCONS_l2_00030392	399.98
chrX	149114563	149115435	TCONS_l2_00030761	409.43
chrX	149115588	149121272	TCONS_l2_00030762	416.58
chrX	149116247	149132489	TCONS_l2_00030393	454.72
chrX	149116252	149187439	TCONS_l2_00030394	454.72
chrX	149129099	149131017	TCONS_l2_00030763	454.72
chrX	149183783	149185303	TCONS_l2_00030395	454.72
chrX	149184058	149185102	TCONS_l2_00030396	560.88
chrX	149184136	149184896	TCONS_l2_00030397	830.51
chrX	149184312	149185875	TCONS_l2_00030764	830.51
chrX	154578071	154579221	TCONS_l2_00030412	1297.4
chrX	154578071	154579111	TCONS_l2_00030769	1415.6
chrY	10035279	10036679	TCONS_l2_00030893	1924.1
chrY	13309474	13370674	TCONS_l2_00030931	4419.5
chrY	13317319	13370650	TCONS_l2_00030932	4419.5
chrY	13362084	13370619	TCONS_l2_00030933	4419.5

附表 2 GTEx 数据集中卵巢特异高表达转录本

Ensembl Gene ID	Ensembl Transcript ID	Chr	Start (bp)	End (bp)	Gene
ENSG00000134201	ENST00000256593	1	109712255	109718266	GSTM5
ENSG00000143125	ENST00000271331	1	110451200	110457354	PROK1
ENSG00000143502	ENST00000343846	1	223220819	223364059	SUSD4
ENSG00000143355	ENST00000367390	1	197912505	197935478	LHX9
ENSG00000142937	ENST00000372209	1	44775573	44778779	RPS8
ENSG00000213244	ENST00000401004	1	121118195	121118610	HIST2H3DP1
ENSG00000223345	ENST00000412169	1	121116862	121117242	HIST2H2BA
ENSG00000237749	ENST00000423216	1	37556247	37556499	RP3-423B22.5
ENSG00000234004	ENST00000427600	1	211173488	211173849	RP11-543B16.1
ENSG00000223345	ENST00000430394	1	121108210	121117257	HIST2H2BA
ENSG00000235363	ENST00000432323	1	205351247	205351471	SNRPGP10
ENSG00000121310	ENST00000474789	1	52913720	52921721	ECHDC2
ENSG00000158864	ENST00000478866	1	161202376	161210474	NDUFS2
ENSG00000270380	ENST00000481350	1	110456505	110457354	RP11-470L19.5
ENSG00000162873	ENST00000539253	1	205336065	205357090	KLHDC8A
ENSG00000162873	ENST00000606529	1	205344196	205356836	KLHDC8A
ENSG00000198075	ENST00000272452	2	108378012	108388057	SULT1C4
ENSG00000124006	ENST00000289656	2	219561747	219571464	OBSL1
ENSG00000115641	ENST00000344213	2	105360812	105399224	FHL2

ENSG00000074047	ENST00000361492	2	120797291	120992652	GLI2
ENSG00000115641	ENST00000393353	2	105360826	105399051	FHL2
ENSG00000115641	ENST00000408995	2	105360861	105399050	FHL2
ENSG00000198075	ENST00000409309	2	108378185	108387640	SULT1C4
ENSG00000071082	ENST00000409733	2	101002293	101006419	RPL31
ENSG00000115641	ENST00000409807	2	105360857	105396957	FHL2
ENSG00000238273	ENST00000415627	2	105374093	105376083	AC012360.6
ENSG00000114923	ENST00000425141	2	219627627	219641728	SLC4A3
ENSG00000138395	ENST00000450471	2	201806579	201895550	CDK15
ENSG00000124006	ENST00000456147	2	219555763	219559431	OBSL1
ENSG00000238273	ENST00000457290	2	105363038	105378839	AC012360.6
ENSG00000124006	ENST00000491370	2	219570458	219571859	OBSL1
ENSG00000198075	ENST00000494122	2	108377911	108382922	SULT1C4
ENSG00000269068	ENST00000597192	2	219559083	219559626	RP11-256I23.2
ENSG00000114115	ENST00000232219	3	139517434	139539829	RBP1
ENSG00000152580	ENST00000282466	3	151433494	151458709	IGSF10
ENSG00000183770	ENST00000330315	3	138944224	138947140	FOXL2
ENSG00000206262	ENST00000383165	3	138947234	138953451	FOXL2NB
ENSG00000172986	ENST00000389617	3	72888073	72976926	GXYLT2
ENSG00000174748	ENST00000413699	3	23917131	23920820	RPL15
ENSG00000158258	ENST00000458420	3	139935185	140577397	CLSTN2
ENSG00000242068	ENST00000481241	3	180772750	180773914	RP11-259P15.4
ENSG00000114115	ENST00000483943	3	139524407	139539719	RBP1
ENSG00000145075	ENST00000485055	3	180707589	180870178	CCDC39
ENSG00000013297	ENST00000486429	3	170418875	170423742	CLDN11
ENSG00000114115	ENST00000492918	3	139526405	139539829	RBP1
ENSG00000145075	ENST00000495817	3	180708326	180871005	CCDC39
ENSG00000206262	ENST00000498709	3	138947612	138950828	FOXL2NB
ENSG00000228252	ENST00000504443	3	130212823	130273186	COL6A4P2
ENSG00000002587	ENST00000002596	4	11393150	11429765	HS3ST1
ENSG00000134853	ENST00000257290	4	54229097	54298247	PDGFRA
ENSG00000109625	ENST00000315782	4	8592756	8619749	CPZ
ENSG00000226950	ENST00000425653	4	52712781	52714137	DANCR
ENSG00000231160	ENST00000436901	4	38626408	38664809	KLF3-AS1
ENSG00000231160	ENST00000440181	4	38612701	38664628	KLF3-AS1
ENSG00000109625	ENST00000506287	4	8592708	8601498	CPZ
ENSG00000002587	ENST00000514690	4	11399758	11428597	HS3ST1
ENSG00000109625	ENST00000515606	4	8592726	8619751	CPZ
ENSG00000272620	ENST00000608442	4	7754090	7778928	AFAP1-AS1
ENSG00000164574	ENST00000297107	5	154190730	154420984	GALNT10
ENSG00000171444	ENST00000302475	5	113022099	113294949	MCC
ENSG00000170085	ENST00000481515	5	176296123	176322815	SIMC1

ENSG00000122203	ENST00000508023	5	176359520	176361764	KIAA1191
ENSG00000250579	ENST00000512155	5	5132780	5140054	CTD-2297D10.2
ENSG00000152348	ENST00000514253	5	82164445	82276857	ATG10
ENSG00000248587	ENST00000514532	5	37869907	37874894	GDNF-AS1
ENSG00000128606	ENST00000339431	7	102912991	102944949	LRRC17
ENSG00000122574	ENST00000409123	7	29806486	29906175	WIPF3
ENSG00000128606	ENST00000485478	7	102936184	102944670	LRRC17
ENSG00000253369	ENST00000521558	8	53395211	53395946	RP11-1081M5.1
ENSG00000255394	ENST00000525043	8	11761256	11763223	C8orf49
ENSG00000136574	ENST00000526021	8	11756493	11758765	GATA4
ENSG00000120162	ENST00000262244	9	27325209	27529781	MOB3B
ENSG00000147852	ENST00000382099	9	2622100	2654480	VLDLR
ENSG00000147852	ENST00000382100	9	2621834	2660053	VLDLR
ENSG00000107736	ENST00000224721	10	71396934	71815947	CDH23
ENSG00000107736	ENST00000398792	10	71712604	71742036	CDH23
ENSG00000167992	ENST00000301770	11	61258286	61295091	VWCE
ENSG00000187151	ENST00000334289	11	101890674	101916522	ANGPTL5
ENSG00000167992	ENST00000398808	11	61271165	61273319	VWCE
ENSG00000110700	ENST00000525828	11	17074393	17077668	RPS13
ENSG00000212789	ENST00000526214	11	18261982	18263091	ST13P5
ENSG00000154134	ENST00000532472	11	124876818	124877741	ROBO3
ENSG00000135486	ENST00000340913	12	54280755	54285214	HNRNPA1
ENSG00000187109	ENST00000548044	12	76049589	76084666	NAP1L1
ENSG00000167535	ENST00000552812	12	48819778	48821489	CACNB3
ENSG00000165521	ENST00000554922	14	88612431	88792752	EML5
ENSG00000092445	ENST00000263798	15	41559034	41583586	TYRO3
ENSG00000189136	ENST00000339094	15	84527196	84570795	UBE2Q2P1
ENSG00000244056	ENST00000459938	15	84394072	84394336	RN7SL417P
ENSG00000247809	ENST00000560800	15	96127369	96327021	NR2F2-AS1
ENSG00000005187	ENST00000501740	16	20674510	20755608	ACSM3
ENSG00000260279	ENST00000563087	16	89297508	89298317	AC137932.5
ENSG00000134419	ENST00000565420	16	18782969	18790260	RPS15A
ENSG00000140937	ENST00000569783	16	65118801	65122057	CDH11
ENSG00000125691	ENST00000394332	17	38847865	38853722	RPL23
ENSG00000125691	ENST00000470646	17	38850422	38853424	RPL23
ENSG00000175061	ENST00000477249	17	16439057	16441966	LRRC75A-AS1
ENSG00000172809	ENST00000584577	17	74203693	74209878	RPL38
ENSG00000172000	ENST00000307635	19	2867335	2883445	ZNF556
ENSG00000224864	ENST00000497576	19	20125044	20125484	CTC-260E6.2
ENSG00000267636	ENST00000591132	19	31421997	31427302	AC008992.1
ENSG00000105695	ENST00000593348	19	35309806	35312009	MAG
ENSG00000268119	ENST00000594557	19	21452040	21463884	CTD-2561J22.5

ENSG00000196268	ENST00000599517	19	21411940	21414599	ZNF493
ENSG00000161643	ENST00000602139	19	49969673	49975814	SIGLEC16
ENSG00000101440	ENST00000374954	20	34260365	34269344	ASIP
ENSG00000133519	ENST00000433168	22	23390606	23402726	ZDHHC8P1
ENSG00000188677	ENST00000477795	22	44086684	44100089	PARVB
ENSG00000147119	ENST00000276055	X	46573784	46598408	CHST7
ENSG00000196440	ENST00000354842	X	101485475	101533459	ARMCX4
ENSG00000171004	ENST00000370836	X	132626016	132961395	HS6ST2
ENSG00000157502	ENST00000372552	X	106201736	106208955	MUM1L1
ENSG00000004848	ENST00000379044	X	25003694	25015948	ARX
ENSG00000146938	ENST00000381093	X	5890032	6227203	NLGN4X
ENSG00000147145	ENST00000435339	X	78747728	78757094	LPAR4

致谢

首先十分感谢张锋老师一年多在科研上给予我的悉心指导和帮助，他对生命科学研究的热情和严谨的治学态度深深地影响了我。当我刚进实验室时什么都还不懂，张老师耐心地引导我如何去全面和多角度地思考和分析科学问题，慢慢培养起科学的思维方式。对于科研的任何想法和探索，张老师总是大力支持，鼓励我勇敢地探索。

感谢石乐明老师对我的生物信息学分析思路的指导。感谢实验室的所有师兄师姐，一点一滴地从头带我做实验，耐心地回答我的每一个问题；感谢张雪师姐和我一起讨论课题，感谢陈璐师姐在毕业设计的生物信息学方面的指导。

大学四年是人生中非常重要的时光，感谢复旦大学为我提供了好的平台和机会，让我大胆追逐理想。感谢四年中有幸遇到的所有老师和同学。