小鼠与牛在精子形成过程中同源差异基因比对分析

Differentially expressed homologous genes between mouse and cattle during the spermiogenes

G01组员：李　欣　岳静伟
潘建飞　孙丹丹

目录

1.本组组员研究课题简介

潘建飞：

题 目：利用CRISPR/Cas9技术构建的UCP1定点敲入猪腹股沟脂肪组织转录组分析

背 景: 1.哺乳动物主要存在两种脂肪：WHA 和BAT。

2.BAT通过UCP1来消耗能量来维持体温和抵抗肥胖，但是家猪缺乏UCP1基因，导致新生仔猪不耐寒，给生产带来了巨大的困难。

内 容: 利用CRISPR/Cas9构建了UCP1基因定点敲入猪，实现了UCP1基因在猪白色脂肪组织的特异表达。

结 果: UCP1猪在急性冷刺激情况下的体温维持能力显著优于野生型猪。生长性能测定结果显示UCP1基因敲入猪脂肪率显著降低，背膘厚度显著降低，瘦肉率显著上升。

目 的: UCP1作用的分子机制尚不清楚，通过猪腹股沟脂肪组织转录组分析来阐明！

1.本组组员研究课题简介

孙丹丹：

题 目：叶酸缺乏影响骨骼肌发育的分子机制

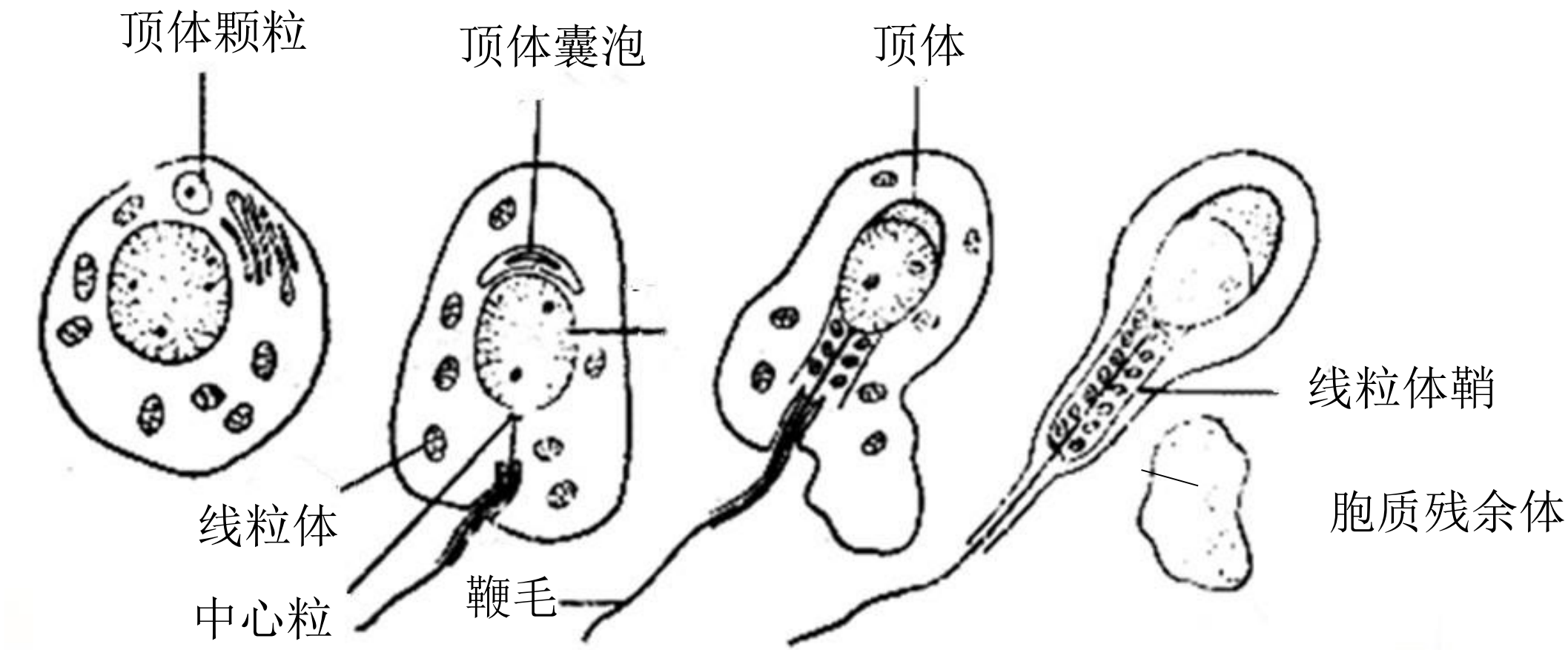
背 景: 叶酸是人体必需的微量元素，体内不能合成，当摄入量不能满足需要时，引起很多健康问题。叶酸缺乏影响骨骼肌发育的分子机制尚不清楚。

目 的: 叶酸缺乏对骨骼肌肌细胞生成和增殖抑制的分子机制。

岳静伟：

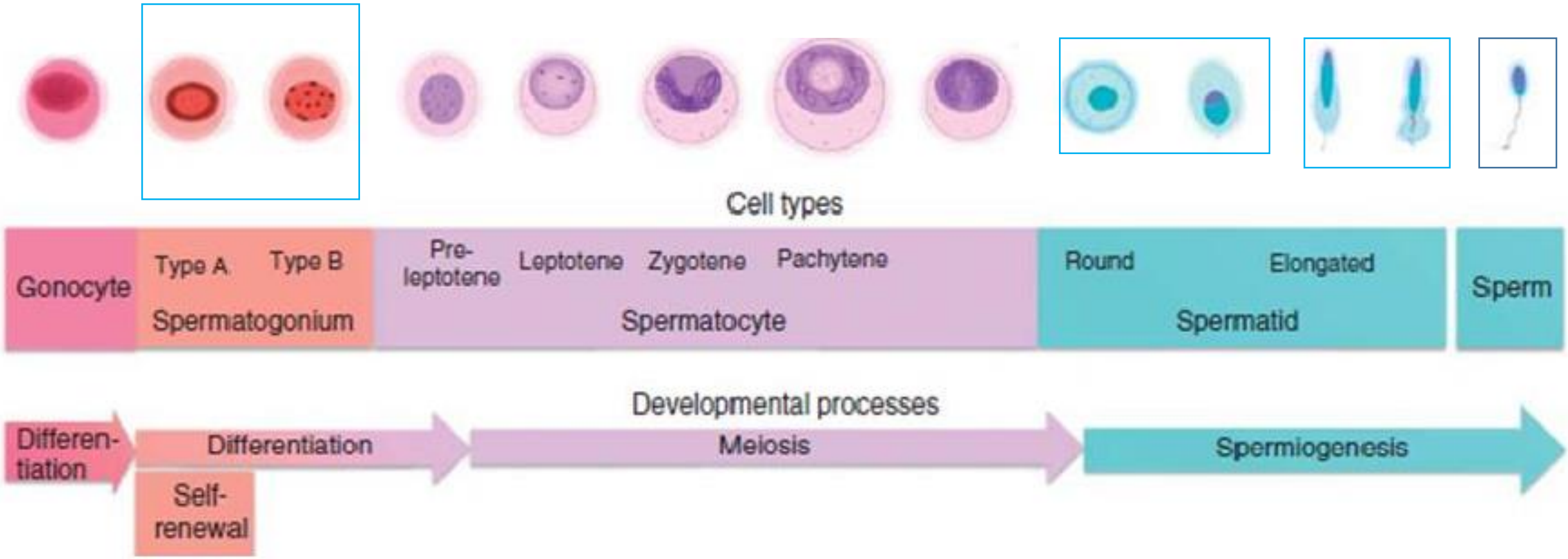
课题尚未确定

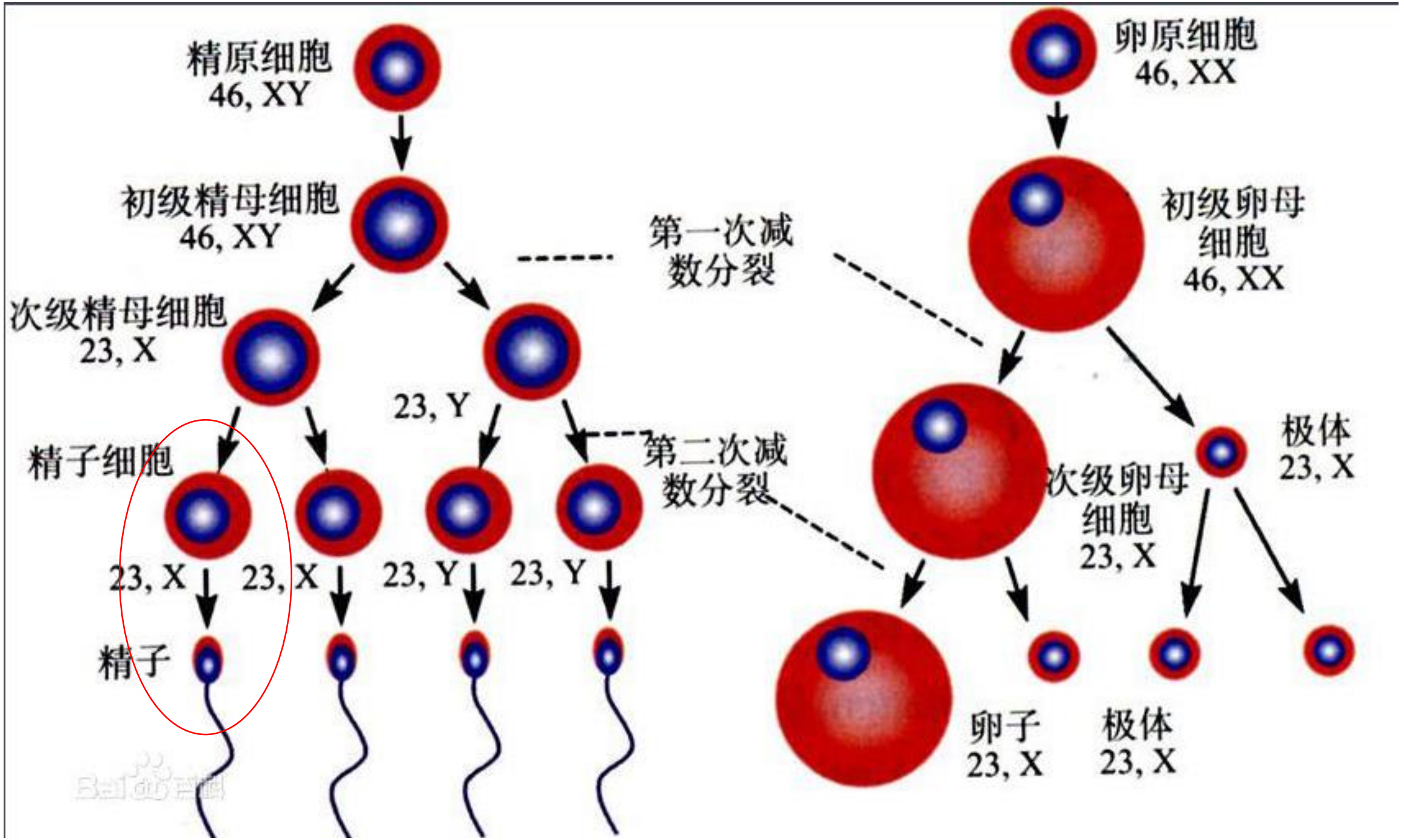
2、研究背景



精子形成过程

2、研究背景


Cell types
Gonocyte
Type A
Type B
Spermatogonium
Pre-leptotene
Leptotene
Zygotene
Pachytene
Spermatocyte
Round
Elongated
Spermatid
Sperm
Developmental processes
Differen-tiation
Differentiation
Self-renewal
Meiosis
Spermiogenesis


精原细胞
46, XY
初级精母细胞
46, XY
次级精母细胞
23, X
23, Y
精子细胞
23, X
23, X
23, Y
23, Y
精子
第一次减数分裂
第二次减数分裂
卵原细胞
46, XX
初级卵母细胞
46, XX
次级卵母细胞
23, X
极体
23, X
卵子
23, X
极体
23, X

3、研究目的

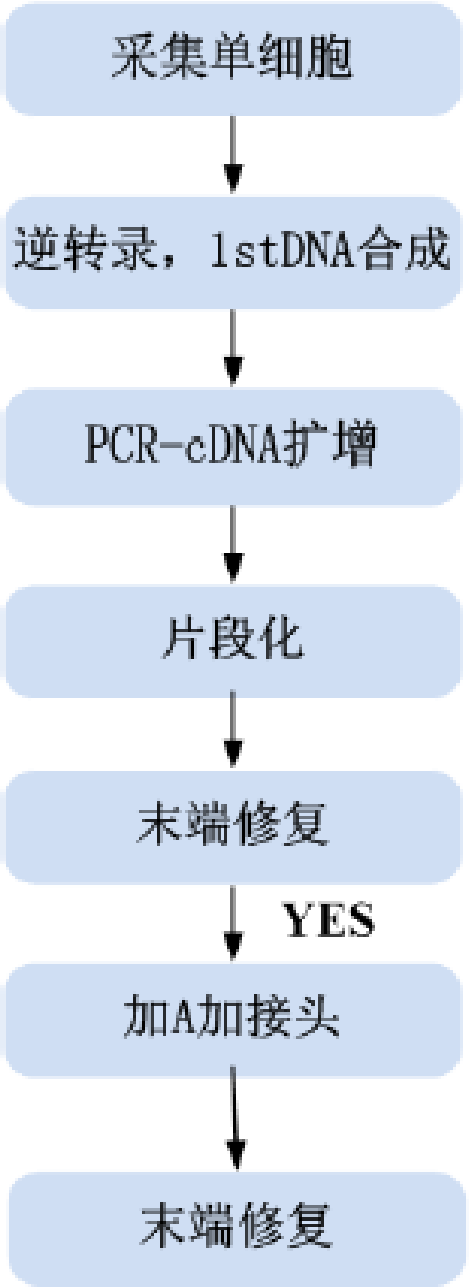
1、寻找小鼠和牛的精子变形过程中四个阶段的同源基因

2、寻找牛精子变形过程中同源基因中四个阶段的差异基因

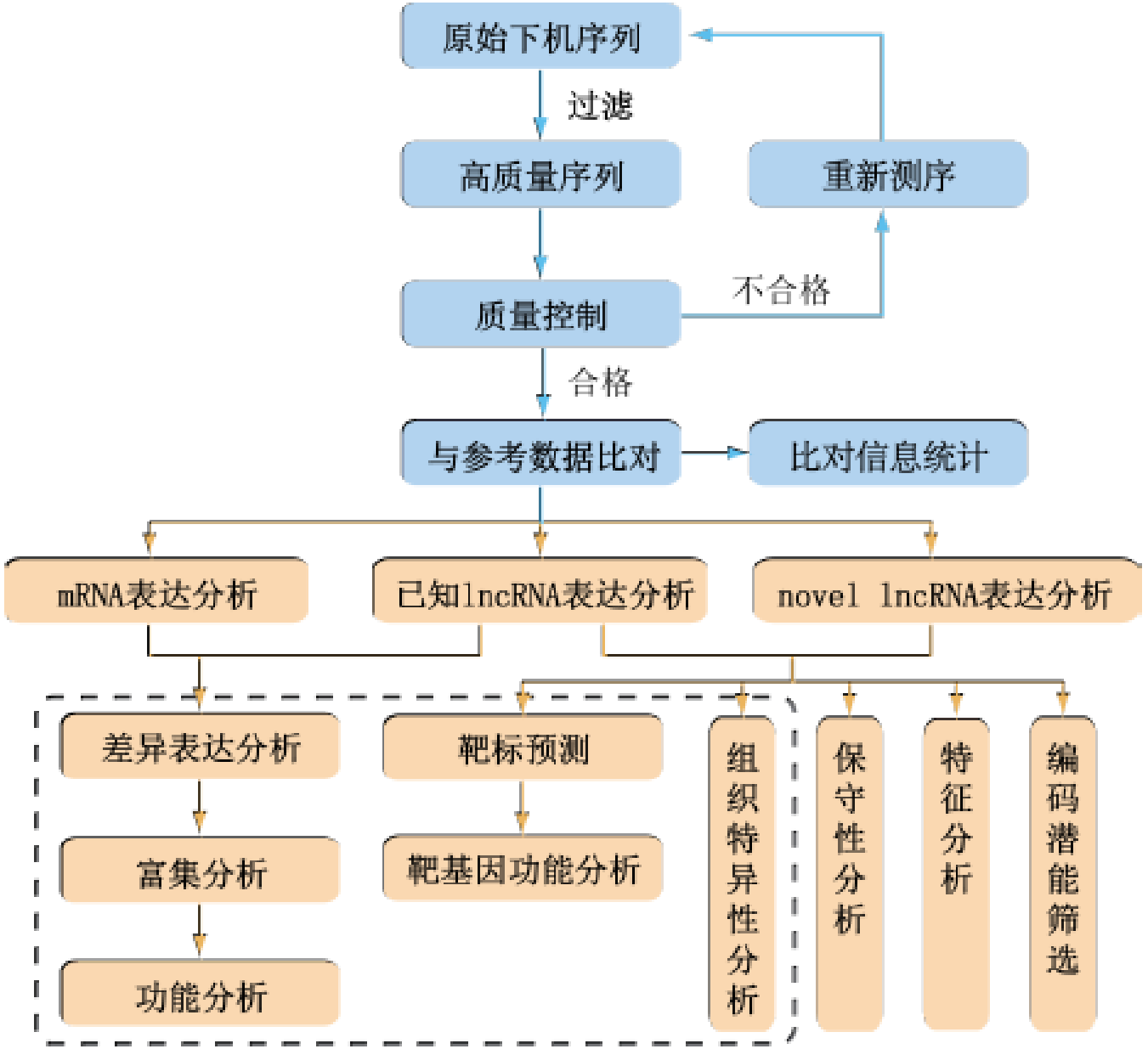
3、寻找小鼠精子变形过程中同源基因中四个阶段的差异基因

4 材料和方法

- 实验材料：
- 方法：单细胞转录组测序
- 测序平台：Illumina HiSeq4000测序平台，



技术路线流程图


原始下机序列
过滤
高质量序列
重新测序
质量控制
不合格
合格
与参考数据比对
比对信息统计
mRNA表达分析
已知lncRNA表达分析
novel lncRNA表达分析
差异表达分析
富集分析
功能分析
靶标预测
靶基因功能分析
组织特异性分析
保守性分析
特征分析
编码潜能筛选

5 已取得的实验进展

- 5.1 小鼠数据的收集
- Ying Z, Chong T, Tian Y, et al. MicroRNAs control mRNA fate by compartmentalization based on 3′ UTR length in male germ cells[J]. Genome Biology, 2017, 18(1):105.
- SRR3395024，SRR3395025，SRR3395026，SRR3395033，SRR3395034，SRR3395035，SRR3395030，SRR3395031，SRR3395032，SRR3395039，SRR3395040，SRR3395041

5 已取得的实验进展

- 5.1 小鼠数据收集
- *K, Singh A, Nguyen T, et al. TSPAN8 Expression Distinguishes Spermatogonial Stem Cells in the Prepubertal Mouse Testis1[J]. Biology of Reproduction, 2016, 95(6):117.*
- SRR3662178，SRR3662179，SRR3662181
- *Li X, Ao J, Wu J. Systematic identification and comparison of expressed profiles of lncRNAs and circRNAs with associated co-expression and ceRNA networks in mouse germline stem cells:[J]. Oncotarget, 2017, 8(16):26573-26590.*
- SRR4413830，SRR4413831，SRR4413832

SRA号	测序方法	文库类型	样品类型	小鼠品系
SRR4423201 SRR4423202 SRR4423203 SRR4423204	Illumina HiSeq 4000	Strand-specific	Mature Sperm	C57BL/6J
SRR3395024 SRR3395025 SRR3395026	Illumina Hi-Seq 2000	TruSeq Stranded Total RNA Library	elongating polysome	C57BL/6J
SRR3395033 SRR3395034 SRR3395035			elongating rnp	
SRR3395030 SRR3395031 SRR3395032			round polysome	
SRR3395039 SRR3395040 SRR3395041			round rnp	
SRR3662178 SRR3662179 SRR3662181	Illumina HiSeq 2500	Smart-seq	SSCs	C57BL/6J
SRR4413830 SRR4413831 SRR4413832	Illumina HiSeq 2000	First--strand cDNA	SSCs	C57BL/6J dx4-Cre
	Illumina HiSeq 2000 or 2500	strand-specific RNA sequencing	SSCs	C57BL/6

5.2 转录组数据与基因组比对

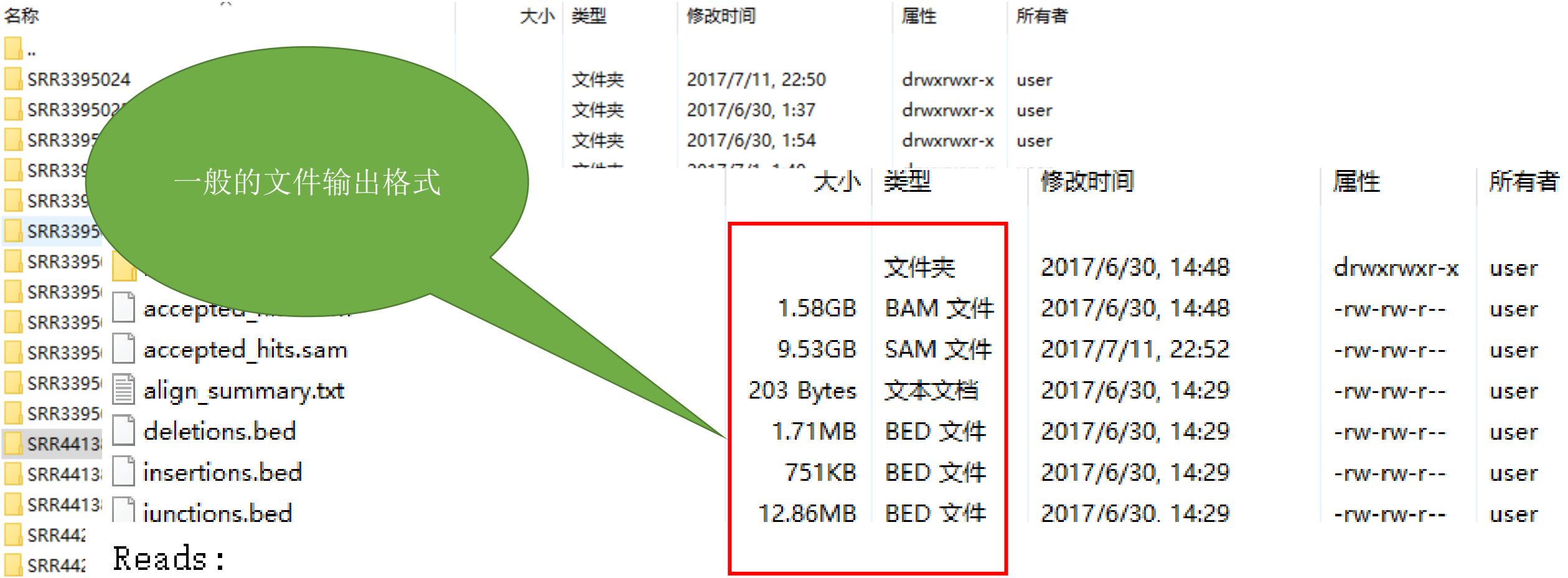
Using the SRA Toolkit to convert .sra files into other formats

- 3.2.1 使用**sratoolkit软件**将查找到的转录组原始数据sra文件转换成fastq数据文件；

- 3.2.2 将得到的fastq数据文件中的转录组数据与基因组比对：（1）使用bowtie2-build根据基因组数据生成基因组索引文件；（2）利用Tophat将得到的fastq数据与基因组索引文件和基因组注释文件进行比对，生成基因组比对bam文件并进行sort；

将文件进行排序并输出

可以看到每条reads在参考基因组的位置，这条reads是在哪一条染色体，又是在这条染色体的哪个位置就可以一目了然。




```
Reads:
          Input     :  39631708
           Mapped   :  36838255 (93.0% of input)
            of these:   2342870 ( 6.4%) have multiple alignments (92591
have >20)
93.0% overall read mapping rate.
```

5.3、估算基因表达量（FPKM）

- 使用cufflink输入基因注释文件和sort之后的基因组比对文件估算基因表达量（FPKM）；
- FPKM: expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced (每百万测序碱基中每千个转录子测序碱基中所包含的测序片断数)
- 在随机抽样的情况下，序列较长的基因被抽到的概率本来就会比序列短的基因高，如此一来，序列长的基因永远会被认为表达量较高，而错估基因真正的表达量。在测序深度不同的情况下，测序深度更深的样品中，比对到每个基因的Read数量更多。
- 为排除因基因的长度、测序深度等因素造成的干扰，FPKM等方法就应运而生了。

5.3、估算基因表达量（FPKM）

- Cufflinks是加利福尼亚大学伯克利分校数学和计算机生物实验室，由LiorPachter领导的StevenSalzberg’s团队，和马里兰大学生物信息和计算机生物中心的Steven Salzberg小组，以及加州理工学院的Barbara Wold实验室联合作用的结果。

- Cufflinks 利用Tophat比对的结果（alignments）来组装转录本，估计这些转录本的丰度，并且检测样本间的差异表达及可变剪接。这个软件其实是个套装，包括四个部分分别命名为：cufflinks、cuffcompare、cuffmerge及cuffdiff。

5.4 Cufflinks的一般步骤：

利用tophat/bowtie比对结果（bam格式）及参考基因组构建转录本，最终的转录本是以gtf格式保存的。

评估转录本构建情况，此外，根据构建的转录本与已知ENSEMBL数据库中的转录本的相对位置定义了一系列分类，例如内含子区域、反义、基因间区域转录本等等近10种分类。

将多个转录本集合合并成一套转录本集合

衡量两个或多个样本间差异表达的基因；还能衡量差异可变剪接体。

tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
ENSMUSG00000064842	-	-	ENSMUSG00000064842	Gm26206	-	1:3102015-3102125	-	-	0	0	0	OK
ENSMUSG00000102348	-	-	ENSMUSG00000102348	Gm10568	-	1:3680154-3681788	-	-	0	0	0	OK
ENSMUSG00000102592	-	-	ENSMUSG00000102592	Gm38385	-	1:3752009-3754360	-	-	0	0	0	OK
ENSMUSG00000102693	-	-	ENSMUSG00000102693	4933401J01Rik	-	1:3073252-3074322	-	-	0.0440964	0	0.12944	OK
ENSMUSG00000088333	-	-	ENSMUSG00000088333	Gm27396	-	1:3783875-3783933	-	-	0	0	0	OK
ENSMUSG00000102343	-	-	ENSMUSG00000102343	Gm37381	-	1:3905738-3986215	-	-	0.47239	0	0.94478	OK
ENSMUSG00000102269	-	-	ENSMUSG00000102269	Gm7357	-	1:4522904-4526737	-	-	0	0	0	OK
ENSMUSG00000096126	-	-	ENSMUSG00000096126	Gm22307	-	1:4529016-4529123	-	-	0	0	0	OK
ENSMUSG00000103003	-	-	ENSMUSG00000103003	Gm38076	-	1:4534836-4535286	-	-	0	0	0	OK
ENSMUSG00000104328	-	-	ENSMUSG00000104328	Gm37323	-	1:4583128-4586252	-	-	0.220565	0.0998921	0.332974	OK

3.4、差异表达分析

Cuffmerge将各个Cufflinks生成的transcripts.gtf文件融合称为一个更加全面的transcripts注释结果文件merged.gtf。以利于用Cuffdiff来分析基因差异表达。

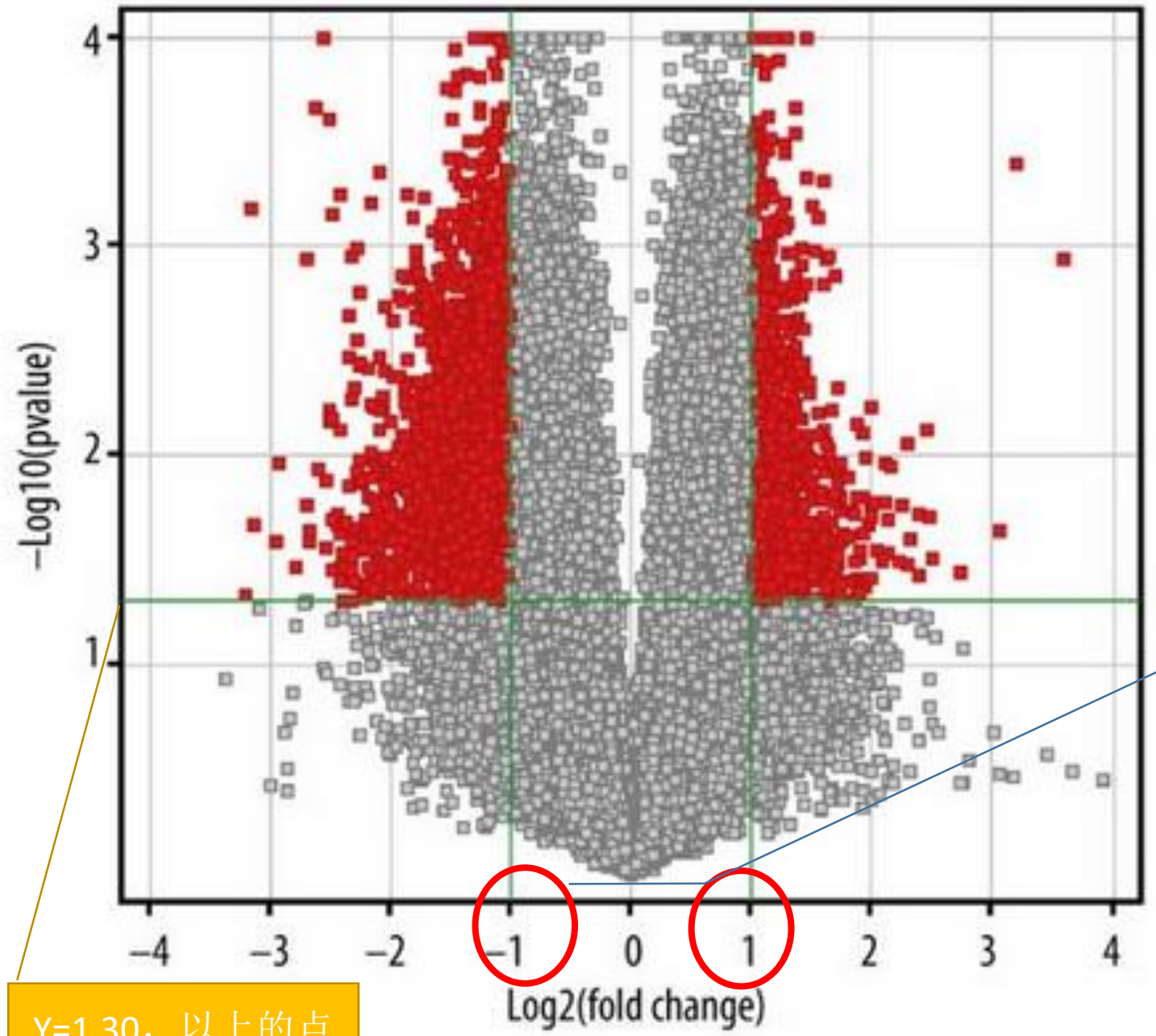
- 5.1 使用cuffmerge将基因表达量gtf文件转换成merged.gtf文件；
- 5.2 使用cuffdiff，输入各样本的merged.gtf文件、样本信息及转录组信息得到差异分析结果；
- 5.3 根据差异基因筛选条件（一般认为fdr≤0.05 ， fold change≥2的为差异基因）提取差异表达基因。

用于寻找转录子表达的显著性差异。

```
test_id     gene_id     gene  locus sample_1    sample_2    statusvalue_
1     value_2     log2(fold_change) test_stat   p_value     q_value
      significant
XLOC_000001 XLOC_000001 4933401J01Rik1:3073252-3074322 E     M     NOTEST
      0.124143    0     -inf  0     1     1     no
XLOC_000002 XLOC_000002 Gm26206     1:3102015-3102125 E     M     NOTEST
      0     0     0     0     1     1
XLOC_000003 XLOC_000003 Gm18956     1:3205900-3671498 E     M     NOTEST
      0     0     0     0     1     1     no
XLOC_000004 XLOC_000004 Gm19921:3205900-3671498 E     M     NOTESTO
      0     0     0     1     1     no
XLOC_000005 XLOC_000005 Gm73411:3205900-3671498 E     M     NOTESTO
      0     0     0     1     1     no
XLOC_000006 XLOC_000006 Gm10568     1:3680154-3681788 E     M     NOTEST
      0     0     0     0     1     1     no
XLOC_000007 XLOC_000007 Gm38385     1:3752009-3754360 E     M     NOT
      0     0     0     0     1     1     no
XLOC_000008 XLOC_000008 Gm37587     1:4490930-4499558 E     M
      2.36671     0     -inf  -nan  0.01745     0.031416    yes
XLOC_000009 XLOC_000009 Gm73571:4522904-4526737 E     M     NOTES
      0.0960108   inf   0     1     1     no
XLOC_000010 XLOC_000010 Gm22307     1:4529016-4529123 E     M
      0     0     0     0     1     1     no
XLOC_000011 XLOC_000011 Gm73691:4610470-4611406 E     M     NOTEST
      0.449787    0.242816    -0.889377   0     1     1     no
```

p-value或q-value小于0.05代表具有显著性差异

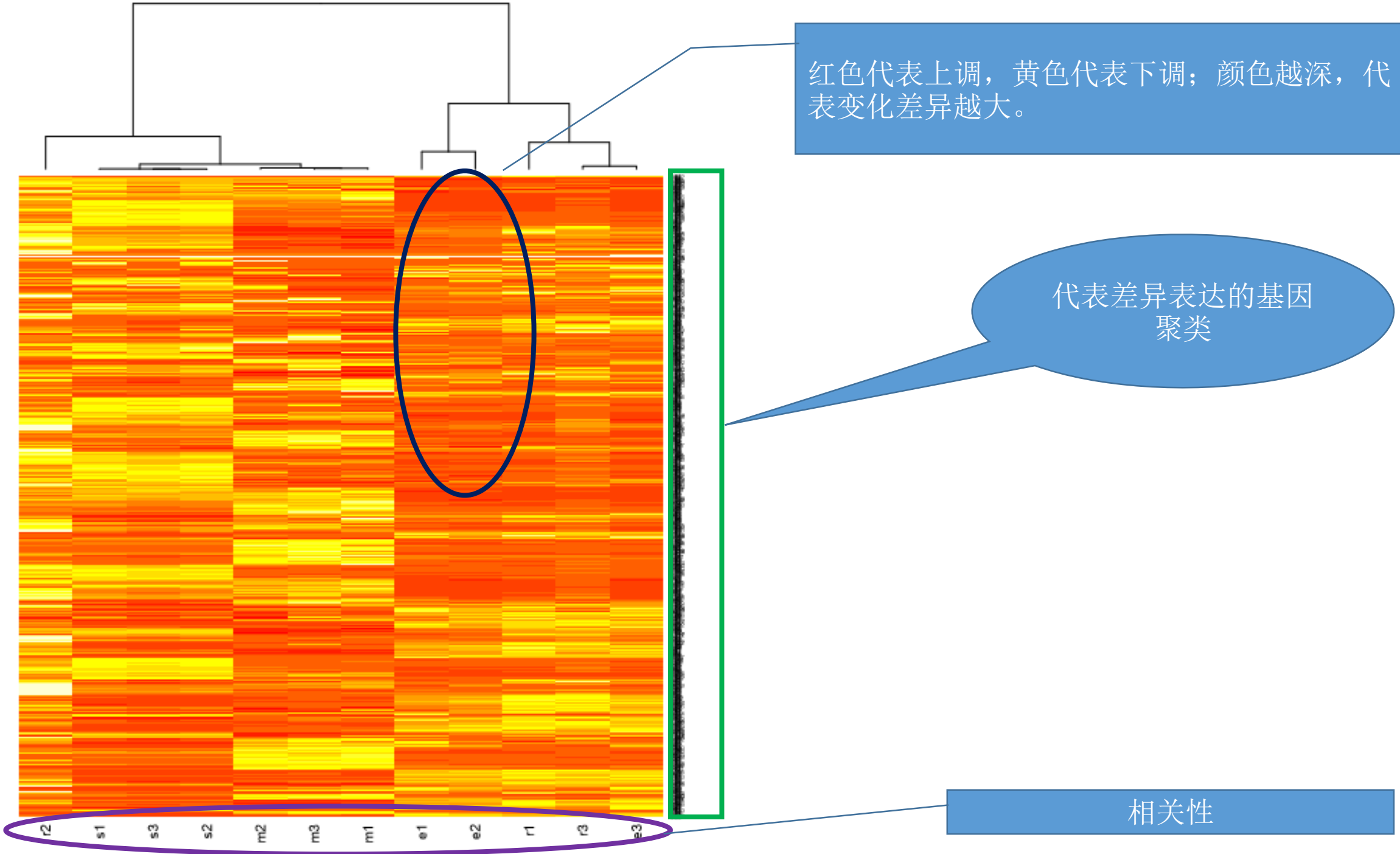
表示实验组比上对照组的差异表达倍数，一般表达相差2倍以上是有意义的，放宽要求1.5倍或者1.2倍也可以接受。



解读：火山图可反映总体基因的表达情况，横坐标代表log2（Fold Change）,纵坐标表示-log10（P值），每个点代表一个基因，颜色用以区分基因是否差异表达，图中红色的点代表差异表达基因，灰白色的点代表没有差异表达的基因。

X=-1左侧的点是下调2倍以上的基因；X=1右侧点是代表上调2倍以上的基因。

Y=1.30，以上的点表示显著



红色代表上调，黄色代表下调；颜色越深，代表变化差异越大。

代表差异表达的基因聚类

相关性

6 下阶段工作任务

1. 通过同源基因对将小鼠和牛精子形成各阶段的转录组进行相似性分析；

2. 将小鼠和牛的差异基因与同源基因对进行比对，找出两个物种之间共同表达的同源差异基因；

3. 使用STEM分别对两个物种在精子形成四个阶段共同表达的同源差异基因进行变化趋势分析，再对两个物种间变化趋势一致的基因进行重叠性分析，找到变化趋势相同的同源差异基因；

6 下阶段工作任务

4. 使用R语言将小鼠与牛之间共同表达的同源差异基因分别与小鼠和牛的GO注释、KEGG通路进行比对分析，得到这些基因在这两个物种上的功能和参与的通路；

5. 通过查看变化趋势相同的同源差异基因的功能及通路筛选出具有调控精子形态变化的同源差异表达基因。

Thank You!