转录组分析报告

1 数据简介

小鼠胎盘组织的 RNA-seq 数据，敲了x基因，含2个时间点 (E9.5， E11.5)。

E9.5（E10）：6个（3个KO，3个WT）

E11.5（E12）：4个（2个KO，2个WT）

前期数据已经经过SOAOnuke过滤软件对原始数据进行质控，接下来的分析则使用处理过后的cleandata。其中原始数据所测的为双端150bp，因此使用STAR软件构建相应的参考基因组。

Sample
0001_D10-5-wt
013-02-D10-WT
013-10-D10-WT
0001_D10-8-ko
013-4-D10-KO
013-7-D10-KO
0003_D12-5-ko
0003_D12-6-wt
0003_D12-7-wt
0005_D12_2-ko

2 定量分析

比对软件使用的为STAR，基因定量为subread中featurecounts函数。

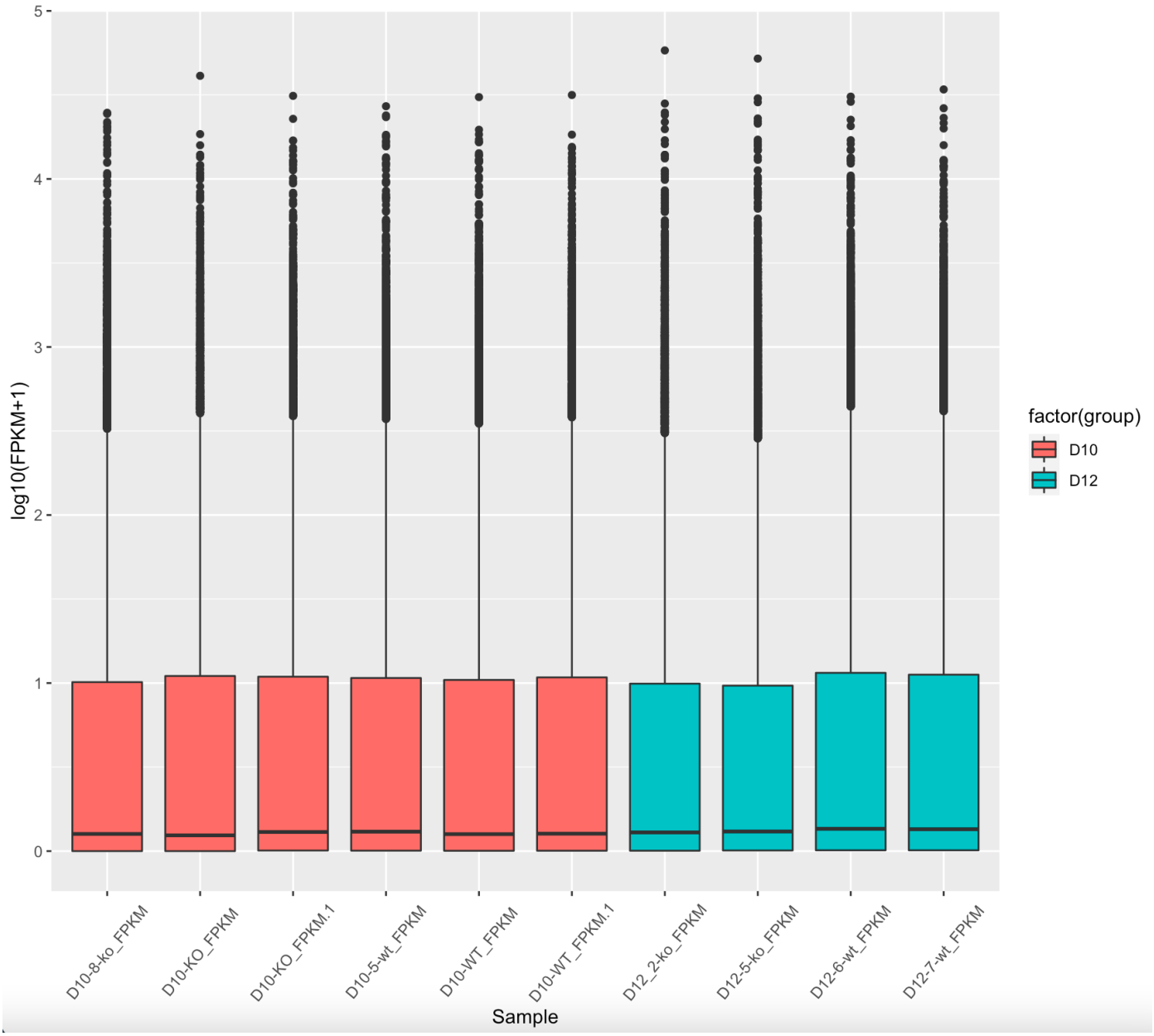
all_feature.txt代表所有样本比对得到的counts文件。

- Genename：基因名
- sample：各样本定量所得的原始read count值
- chr：基因所在的染色体名称
- start：基因所在染色体的起始位置
- end：基因所在染色体的终止位置
- strand：基因所在染色体的正负链信息
- length：基因长度，基因起始到终止所有exon非重叠区域的总和

Genename	D10-8-ko	D10-KO	D10-KO	D10-5-wt	D10-WT	D10-WT	D12_2-ko	D12-5-ko	D12-6-wt	D12-7-wt	Chr	Start	End	Strand
Gm26206	0	0	0	0	0	0	0	0	0	0	NC_000067.(	3102016	3102125	+
Xkr4	3	14	12	2	4	4	10	5	4	8	NC_000067.(	3199733;321	3207317;321	-;-;-;-;-;-;-;-;-;-
Rp1	1	0	3	1	1	0	0	1	1	1	NC_000067.(	4115673;411	4115948;411	-;-
Sox17	201	245	222	205	180	240	319	269	250	210	NC_000067.(	4490928;449	4492668;449	-;-
Gm22307	1	0	0	0	0	0	1	0	0	1	NC_000067.(	4529017	4529123	+
Gm37323	0	0	1	0	0	1	0	0	2	2	NC_000067.(	4547146;458	4547852;458	-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-
Gm39586	0	0	0	0	0	1	0	0	0	1	NC_000067.(	4595057;459	4595422;459	+;+;+
Gm46082	0	1	0	1	0	0	0	0	0	0	NC_000067.(	4602909;460	4604042;460	-;-;-
Gm29874	1	0	3	1	1	1	1	2	1	3	NC_000067.(	4608905;460	4609002;460	+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+
Mrpl15	3301	2995	2891	2665	2896	2460	1863	2246	3019	2555	NC_000067.(	4773198;477	4776801;477	-;-
Lypla1	2604	3213	3258	2521	3495	3560	2189	2080	2796	3264	NC_000067.(	4807823;480	4807982;480	+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+
Tcea1	3100	4632	4595	2942	4132	3950	2365	2349	2272	3164	NC_000067.(	4857694;485	4857976;485	+;+
Rgs20	2	2	0	2	1	1	4	2	2	2	NC_000067.(	4909573;490	4910662;491	-;-
Gm16041	1	0	0	0	5	2	2	1	1	0	NC_000067.(	4970857;497	4971128;497	+;+;+
Atp6v1h	2265	3396	3716	2241	4124	4826	1964	2680	2319	2317	NC_000067.(	5082978;508	5083278;508	+;+
Oprk1	0	3	3	0	0	0	3	1	1	0	NC_000067.(	5588450;558	5588663;558	+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+
Npbwr1	0	0	0	1	1	0	0	0	0	0	NC_000067.(	5913707	5917398	-
Gm46206	22	2	2	16	3	10	82	50	54	38	NC_000067.(	6163888;616	6164938;616	-;-;-
4732440D04	62	82	88	46	37	54	77	79	63	71	NC_000067.(	6209866;621	6215400;621	-;-
Rb1cc1	1702	2411	2469	1740	2977	2975	2121	1892	1836	2392	NC_000067.(	6214634;621	6214957;621	+;+
Alkal1	0	0	0	0	0	0	0	0	0	0	NC_000067.(	6359331;638	6359514;638	+;+;+;+;+
St18	1	0	0	0	0	0	0	1	0	0	NC_000067.(	6487231;648	6487405;648	+;+
Pcmtd1	456	552	578	425	569	611	655	648	523	732	NC_000067.(	7087077;708	7089216;708	+;+
Gm23274	0	0	0	0	0	0	0	0	0	0	NC_000067.(	7265803	7265932	+
Gm26901	25	10	5	13	5	4	6	5	21	5	NC_000067.(	7314637;738	7314932;738	-;-;-
Gm39589	0	0	0	0	0	0	0	0	0	1	NC_000067.(	8108629;810	8108911;810	-;-;-;-

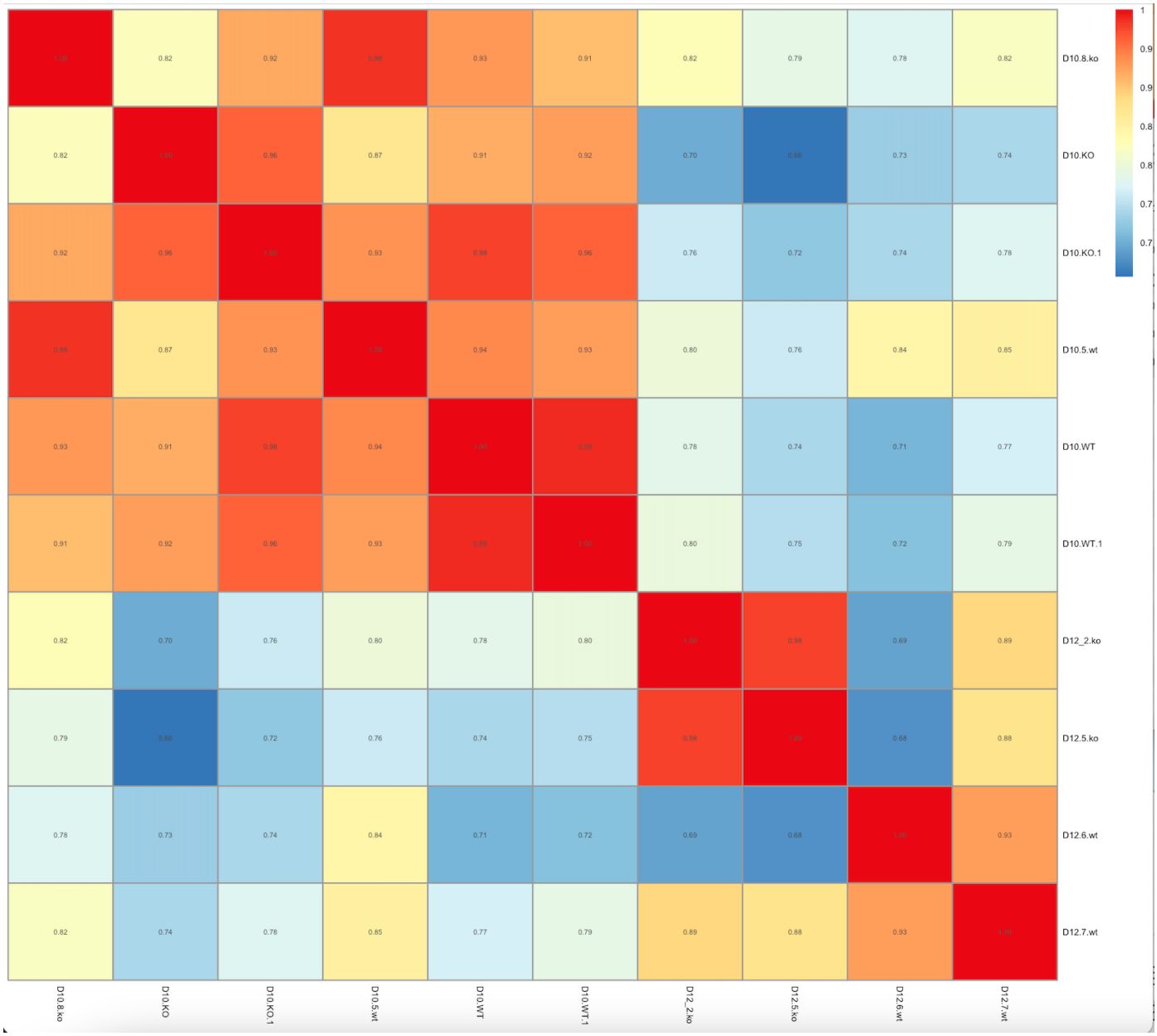
2.1 基因表达分布

由于测序深度和基因长度的影响，RNA-seq的基因表达值一般不用read count来表示，而是用FPKM，FPKM先后对测序深度和基因长度进行了校正[8]。我们计算各样本所有基因的表达值(FPKM)后，通过盒形图展示不同样本基因表达水平的分布情况，如下图所示。图中横坐标为样品名称，纵坐标为log2(FPKM+1)，每个区域的盒形图对五个统计量(至上而下分别为最大值，上四分位数，中位数，下四分位数和最小值)，见结果文件：boxplot.pdf。



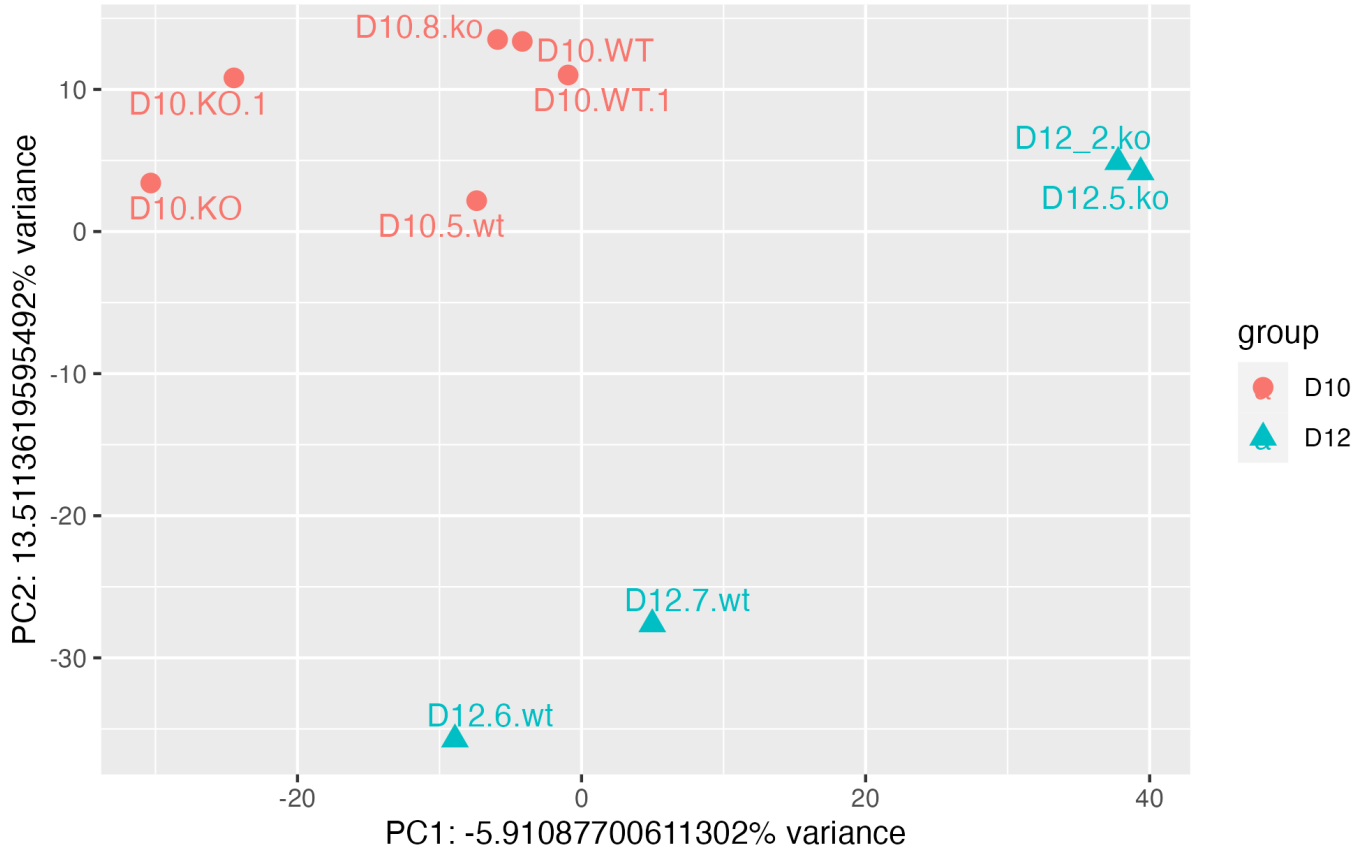
2.2 样本间相关性

生物学重复通常是任何生物学实验所必须的，目前主流期刊也基本要求生物学重复。生物学重复主要有两个用途：一个是证明所涉及的生物学实验操作不是偶然，而是可重复的。另一个是为了确保后续的差异基因分析得到更可靠的结果。样品间基因表达水平相关性是检验实验可靠性和样本选择是否合理的重要指标。相关系数越接近1，表明样品之间表达模式的相似度越高。Encode计划建议皮尔逊相关系数的平方(R2)大于0.92(理想的取样和实验条件下)。具体的项目操作中，我们要求生物学重复样品间R2至少要大于0.8，否则需要对样品做出合适的解释，或者重新进行实验。根据各样本所有基因的FPKM值计算组内及组间样本的相关性系数，绘制成热图，可直观显示组间样本差异及组内样本重复情况。样本间相关性系数越高，其表达模式越为接近，样本相关性热图如下图所示，见结果文件： gene_cor.pdf 。



2.3 主成分分析

主成分分析（PCA）也常用来评估组间差异及组内样本重复情况，PCA采用线性代数的计算方法，对数以万计的基因变量进行降维及主成分提取。我们对所有样本的基因表达值（FPKM）进行PCA分析，如下图所示。理想条件下，PCA图中，组间样本应该分散，组内样本应该聚在一起，见结果文件：PCA.png。



3 差异分析

基因表达定量完成后，需要对其表达数据进行统计学分析，筛选样本在不同状态下表达水平显著差异的基因。差异分析主要分为三个步骤。

- 首先对原始的readcount进行标准化（normalization），主要是对测序深度的校正。
- 然后统计学模型进行假设检验概率（pvalue）的计算
- 最后进行多重假设检验校正，得到FDR值（错误发现率，padj是其常见形式)。

针对不同的实验情况，我们选用合适的软件进行基因表达差异显著性分析，具体如下表所示。

表3.5 表达差异分析所用软件及差异基因筛选标准

类型	软件	标准化方法	pvalue计算模型	FDR计算方法	差异基因筛选标准
有生物学重复	DESeq2(Anders et al, 2014)	DESeq	负二项分布	BH	|log2(FoldChange)| >= 1 & padj <= 0.05
无生物学重复	edgeR(Robinson et al, 2010)	TMM	负二项分布	BH	|log2(FoldChange)| >= 1 & padj <= 0.05

为了找到更多的基因，将阈值调整为|log2(FoldChange)| >= 0

实验设计：

1. D10：处理组KO与对照组WT做差异分析
2. D12：处理组KO与对照组WT做差异分析

- **gene_id**：基因编号
- **basemean**：表示所有样本经过归一化系数矫正的read counts（counts/sizeFactor）的均值。
- **log2FoldChange**：处理组与对照组基因表达水平的比值，再经过差异分析软件收缩模型处理，最后以2为底取对数
- **pvalue**：显著性检验的值
- **padj**:校正后的p值

文件结果：

1. D10组：

   D10_DEGs.csv 为padj<0.05的差异基因

   upDEGsD10.csv为padj<0.05的上调差异基因

   downDEGsD10.csv为padj<0.05的下调差异基因

2. D12组：

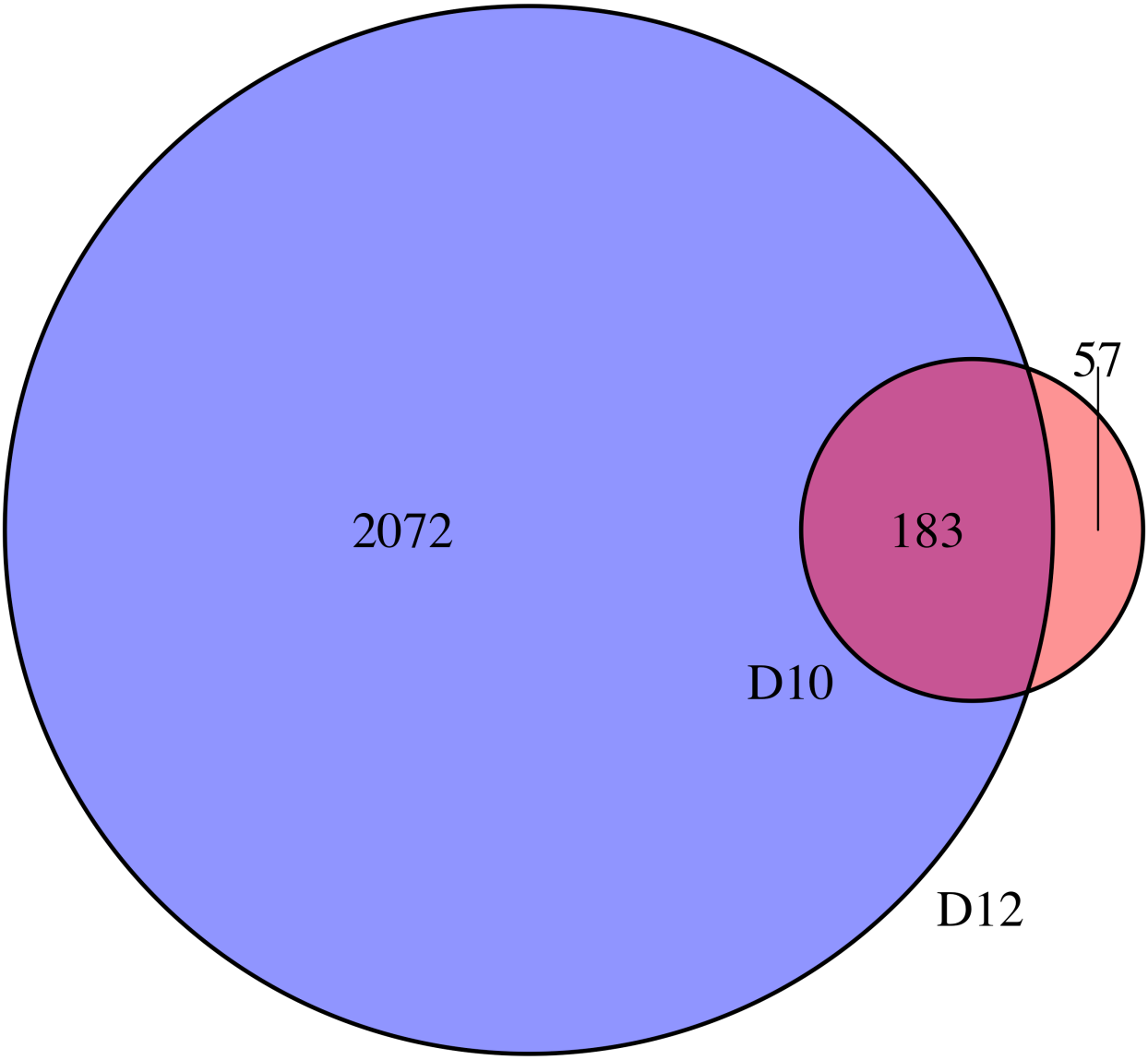
   D12_DEGs.csv 为padj<0.05的差异基因

   upDEGsD12.csv为padj<0.05的上调差异基因

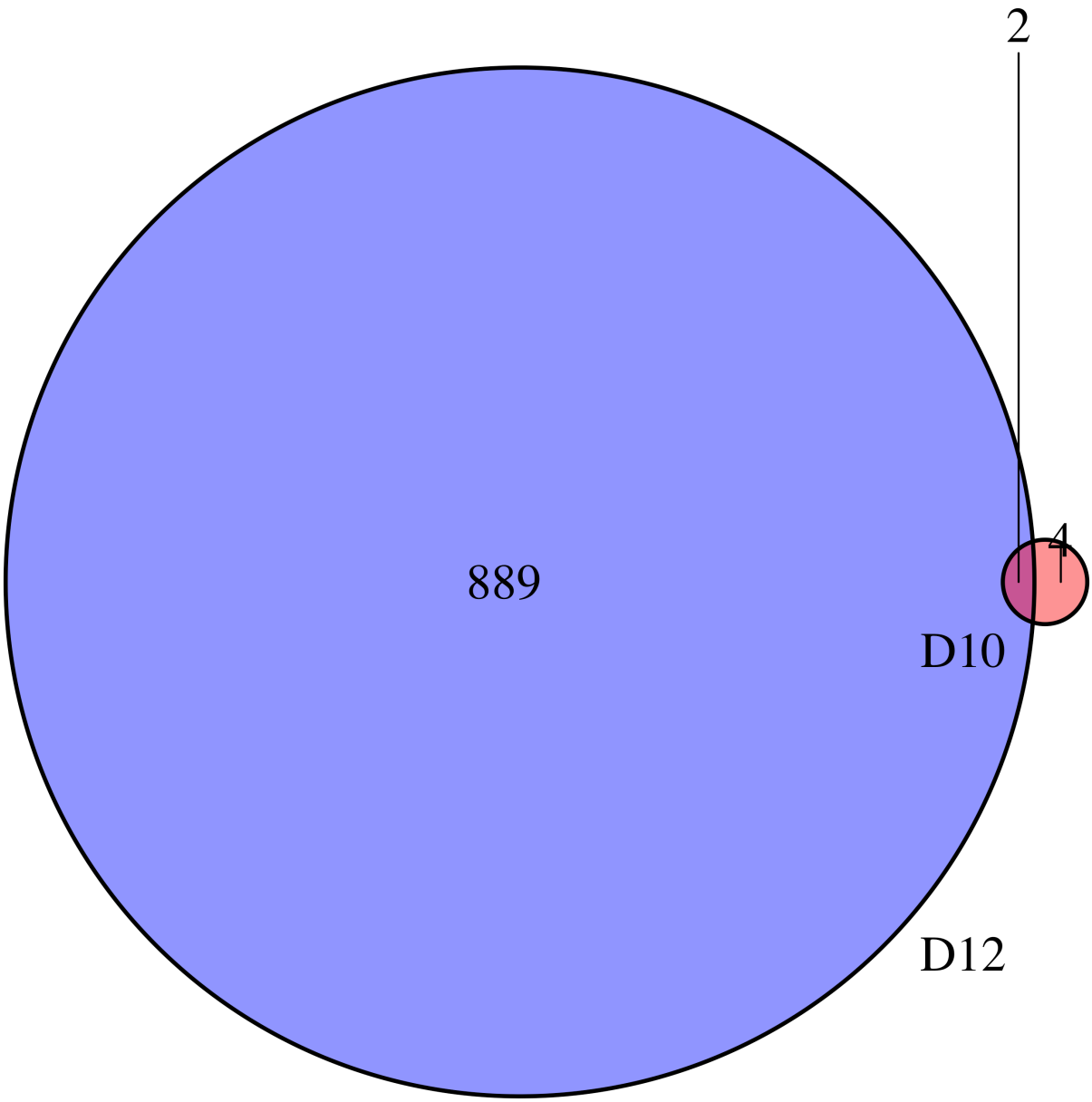
   downDEGsD12.csv为padj<0.05的下调差异基因

Venn图

all_D10_D12_KOvs.WT_0.05_venn



up_D10_D12_KOvs.WT_0.05_venn



down_D10_D12_KOvs.WT_0.05_venn