+----------+
| 转录组项目信息分析结果说明文档 |
+----------+

目录结构及文件说明：

****

*

BMK_3_geneExpression/

|-- BMK_1_randcheck

| |-- Sample1.randcheck.png #样品 1 测序 Reads 在转录本上的位置分布图

| |-- Total.randcheck.png #各样品测序 Reads 在转录本上的位置分布整合图

|-- BMK_2_insertSize

| |-- Sample1.insertSize.png #样品 1 插入片段长度模拟分布图

|-- BMK_3_saturation

| |-- Sample1.express.gene_tag.png #样品 1 测序数据饱和度模拟图

| |-- Total.gene_tag.png #各样品测序数据饱和度模拟整合图

|-- BMK_4_density

| |-- all.fpkm_density.png #各样品 FPKM 密度分布对比图

| |-- all.fpkm_box.png #各样品 FPKM 箱线图

| |-- Sample1.fpkm_density.png #样品 1 的 FPKM 密度分布图

|-- BMK_5_correlation

| |-- Sample1_vs_Sample2.cor.png #样品 1 和样品 2 的表达量相关性散点图

| |-- correlation.txt #两两样品的表达量相关性（皮尔逊相关系数的平方）统计表

| |-- cor.cluster.png #两两样品的表达量相关性热图

|-- All_geneExpression.xls #样品 1 基因表达量分析结果文件

|-- readme.pdf #目录结果说明

##########

文件：All_gene_expression.xls

描述：所有基因表达量分析结果文件

字段解释：

ID：Unigene ID

Sample_count：该样品的 Unigene 的 reads count 值

Sample_FPKM：该样品的 Unigene 的 FPKM 值

##########

文件：correlation.txt

描述：生物学重复相关性统计表，描述两个样品间的生物学重复相关性

字段解释：

第一行：Sample：百迈客对样品的统一编号，其余表示样品名称

第一列：样品名称，每一个值都代表两个样品的相关性。

##########

##
图片：Sample1.insertSize.png
描述：样品插入片段长度
意义：插入片段长度的离散程度能直接反映出文库制备过程中切胶或磁珠纯化的效果
X:双端 Reads 在 Unigene 库中比对起止点之间的距离，范围为 0 到 800bp
Y:纵坐标为比对起止点之间不同距离的双端 Reads 或插入片段数量
##
图片：Sample1.randcheck.png
描述：mRNA 片段化随机性检验,RNA 片段化后的插入片段大小选择，可以理解为从 mRNA 序列中独立随机地抽取子序列，如果样本量（mRNA 数目）越大、打断方式和时间控制得越合适，那么目的 RNA 每个部分被抽取到的可能性就越接近，即 mRNA 片段化随机性越高，mRNA 上覆盖的 Reads 越均匀。
意义：由于参考的 mRNA 长度不同，作图时将每个 mRNA 按照长度划分成 100 个区间，进而统计每一区间内的 Mapped Reads 数目及所占的比例，图中反映的是所有 mRNA 各个区间内的 Mapped Reads 比例的汇总。
X:mRNA 位置
Y:对应位置区间内 Reads 在 Mapped Reads 中所占百分比
##
图片：Sample1.express.gene_tag.png
描述：样品的 Mapped Reads 对检测到的基因数目的饱和度，充足的有效数据是信息分析准确的必要条件。
意义：转录组测序检测到的基因数目与测序数据量成正相关性，即测序数据量越大，检测到的基因数目越多。但一个物种的基因数目是有限的，而且基因转录具有时间特异性和空间特异性，所以随着测序量的增加，检测到的基因数目会趋于饱和。为了评估数据是否充足，需要查看随着测序数据量的增加，新检测到的基因是否越来越少或没有，即检测到的基因数目是否趋于饱和。通过将 Mapped Reads 等量地分成 100 份，逐渐增加数据查看检测到的基因数量来绘制饱和度曲线。
X:Reads 数目
Y:检测到的基因数量
##
图片：Sample1_vs_Sample2.cor.png
描述：两样品的基因相关性散点图
意义：基因表达量散点图中每个点代表一个基因，点越偏离对角线，说明对应基因在两个样品间的表达水平差异越大。另外，偏离对角线的点越多，说明两样品表达量的相关性越低，表达量差异越大；反之亦然。
X:该基因在第一个样品中表达量加 1（FPKM+1）的对数值
Y:该基因在第二个样品中表达量加 1（FPKM+1）的对数值
##
图片：sample_cluster.png
描述：同一条件的每一对生物学重复样品的基因表达量做相关性图,图中不同的列代表不同的样品，不同的行代表不同的基因。颜色代表了基因在样品中的表达量 FPKM 以 2 为底的对数值。
意义：颜色从红到绿，相关性逐渐增大。相关性相近的聚类到一起。
##

图片：Sample1.fpkm_density.png
描述：样品基因表达量总体分布
意义：利用转录组数据检测基因表达具有较高的灵敏度。通常情况下，能够测序到的蛋白质编码基因表达水平 FPKM 值横跨 10^-2 到 10^4 六个数量级。
X：图中不同颜色的曲线代表不同的样品，曲线上点的横坐标表示对应样品 FPKM 的对数值
Y：表示概率密度。
##
图片：all.fpkm_box.png
描述：各样品的 FPKM 分布箱线图
意义：从箱线图中不仅可以查看单个样品基因表达水平分布的离散程度，还可以直观的比较不同样品的整体基因表达丰度。
X：不同的样品
Y：纵坐标表示样品表达量 FPKM 的对数值
##
文件：Sample1.Mapped.stat.xls
描述：样品测序数据与 Unigene 或转录本序列的比对统计表
字段解释：
Total Reads：测序 Reads 总数目；
Mapped Reads：比对到 Unigene 上的测序 Reads 总数目，及其在测序 Reads 总数目中所占的百分比；
Uniq mapped Reads：比对到 Unigene 唯一位置上的测序 Reads 总数目，及其在测序 Reads 总数目中所占的百分比；
Multi mapped Reads：比对到多个 Unigene 或一个 Unigene 多个位置上的测序 Reads 总数目，及其在测序 Reads 总数目中所占的百分比。
##