**癌症早诊标记物设计及其演化机理认知**

癌症是一种遗传变异相关的疾病。在很多情况下，患者在确诊的时候就已经处于疾病的中晚期了，错过了治疗的最佳时机。因此，如何开展面向癌症早期的诊断标记物，一直以来都是临床医学面临的主要难题之一。

目前的癌症诊断及病理分期判定，主要依赖肿瘤的大小、肿瘤是否转移、患者的生化指标、临床病理图像分析等手段，这些方法非常依赖检测的精准度和医生的临床经验。而根据生物学中心法则可知，基因是决定表观性状变化的根本因素。患者的肿瘤状态变化、生化指标变化、或病理图像的异常，往往会晚于基因数据的异常。因此，通过对基因数据的分析，有可能更早的发现癌症，为疾病的早诊提供了新的思路，也能够为更深入的了解癌症发生发展的机理提供更深的见解。

基因表达数据就是这样一种常见的基因相关数据。它反映的是基因转录产物（如信使RNA，也叫mRNA）在细胞中的表达丰度。这些数据可以用于分析不同的条件下哪些基因的表达发生了改变、基因的活动是如何受外界条件的变化而变化、以及基因之间有何共表达关系。这些数据与特诊对于临床诊断标记物设计、癌症发生发展的机制认知等有着重要的作用。

请仅根据题目所提供的数据，尝试回答下列问题：

1. 在癌症的早期与晚期，分别是哪些基因的表达发生了较为显著的变化，这些基因之间的关系如何？
2. 请设计或分析，哪些基因或基因之间关系所构成的特征能够被用来作为诊断的标记物，以此判定或分析患者当前的病理状态。请给出具体的量化指标（或特征集合），并评价所设计的诊断标记物的识别准确性；更具体的，给出早期癌症诊断标记物；
3. 不同的患者之间存在着极强的异质性，如何在分析过程中尽可能的消除异质性的影响？

名词解释：

**基因表达数据**：基因转录产物在细胞中的表达丰度，用以描述当前基因的活跃状态；

**生物标记物（或诊断标记物）：**一种用于判定患者病理状态的特征，可以是某个或某些基因表达值，也可以是某些基因表达关系的某种量化值。通常，生物标记物希望规模越小越好，这样可以节省检测成本；使用生物标志物作为输入特征，可以输出或判定该患者的病理分期或病理状态；

**异质性：**数据之间的差异程度的一种描述。不同的患者，即使他们处于同一种病理阶段，他们的基因表达数据、生化指标、医学影像数据也存在着极大的差异性，会给数据分析带来较大的挑战；

数据文件说明：

1. 附件1：Normal\_exp.csv: 正常人的基因表达数据；
2. 附件2：EarlyStage\_exp.csv：早期患者的基因表达数据；
3. 附件3：LaterStage\_exp.csv：晚期患者的基因表达数据；

上述文件中，第一行为表头，给出各个患者的样本ID（第一列除外）；第一列为基因名称，第二列至最后一列，为各个患者的基因表达值；

参考文献：

[1] <https://ww2.mathworks.cn/help/bioinfo/ug/gene-expression-profile-analysis.html?requestedDomain=cn>

[2] Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). https://doi.org/10.1186/1471-2105-9-559