**生物医药合作项目开发**

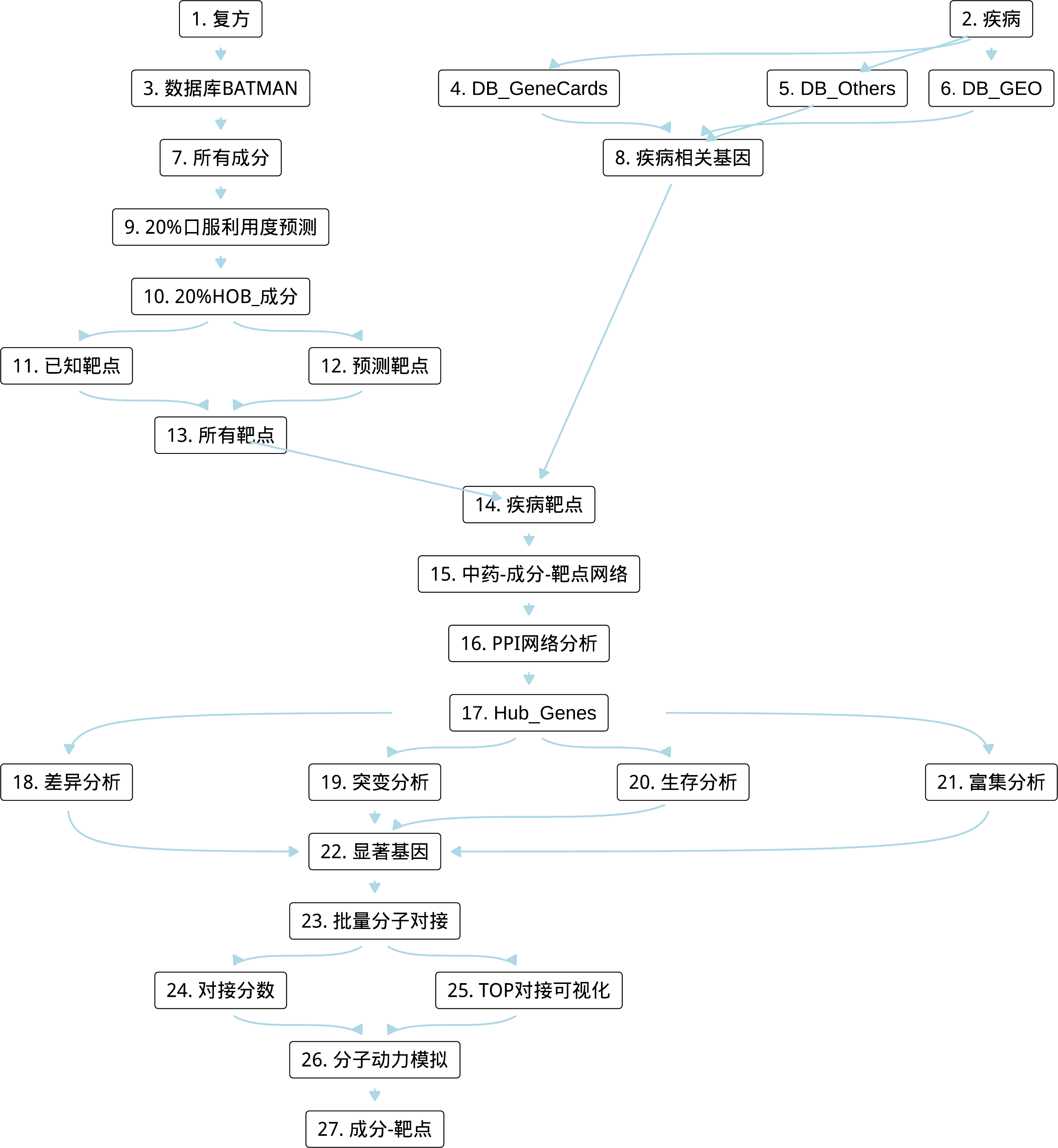
**研究方向： 网络药理学分析+分子动力模拟 ;**

**委托人： 方涛 ;**

**受托人： 杭州铂赛生物科技有限公司 .**

# 1 研究背景

## 1.1 思路



**Fig.** Route

# 2 GEO 检索方法

## 2.1 数据分析平台

在 Linux pop-os x86\_64 (6.9.3-76060903-generic) 上，使用 R version 4.4.2 (2024-10-31) (<https://www.r-project.org/>) 对数据统计分析与整合分析。

## 2.2 GSE 数据搜索 (Dataset: INSOMNIA)

使用 Entrez Direct (EDirect) <https://www.ncbi.nlm.nih.gov/books/NBK3837/> 搜索 GEO 数据库 (esearch -db gds)，查询信息为: ((Insomnia[Description]) AND ((6:300[Number of Samples]) AND (GSE[Entry Type]) AND (Homo sapiens[Organism]))，转化为数据表格。 以正则匹配，滤除 ‘summary’ 或 ‘title’ 中包含 ‘single cell’ 或 ‘scRNA’ 的数据例。仅查询临床数据，因此滤除匹配到关键词 in vitro, cell line, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。 (注：以上仅为查找合适的 GEO 数据所做的数据筛选，与实际分析无关) 。仅获取类型包含 ‘Expression profiling by high throughput sequencing’ 或 ‘Expression profiling by array’ 的数据例。此外，排除 summary 或 title 中匹配到字符集 EX (KO, WT, *WT*, *KO*, wildtype, mutant, knock, deficien, absen, SuperSeries, transgenic, CD[0-9]+) 的数据例。上述得到 2 个 GSE 数据集。以 R 抓取网页 (例如，RCurl::getURL 抓取 )，解析 ‘Overall design’ 和 ‘Samples’ 模块，匹配字符集 EX，排除匹配到的数据例。排除 ‘Overall design’ 中包含 in vitro, cell line, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。为了获得包含对照样本的数据集，(描述字符过少的例外) 以正则匹配 (‘Samples’, ‘Summary’, ‘Overall design’) 包含 normal, control, healthy, ctrl, adjacent, N[0-9]+, C[0-9]+ 的数据例。共 1 个。仅获取包含 ‘protein coding’ 测序的数据集，排除 ‘Samples’ 和 ‘Overall design’ 中包含 siRNA, miRNA, miR, lncRNA 字符的数据例。余下共 1 个。 以 GEOquery 获取 GSE 数据集 (n=1)。

## 2.3 GSE 数据搜索 (Dataset: SD)

使用 Entrez Direct (EDirect) <https://www.ncbi.nlm.nih.gov/books/NBK3837/> 搜索 GEO 数据库 (esearch -db gds)，查询信息为: ((Sleep disorders[Description]) AND ((6:300[Number of Samples]) AND (GSE[Entry Type]) AND (Homo sapiens[Organism]))，转化为数据表格。 以正则匹配，滤除 ‘summary’ 或 ‘title’ 中包含 ‘single cell’ 或 ‘scRNA’ 的数据例。仅查询临床数据，因此滤除匹配到关键词 in vitro, cell line, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。 (注：以上仅为查找合适的 GEO 数据所做的数据筛选，与实际分析无关) 。仅获取类型包含 ‘Expression profiling by high throughput sequencing’ 或 ‘Expression profiling by array’ 的数据例。此外，排除 summary 或 title 中匹配到字符集 EX (KO, WT, *WT*, *KO*, wildtype, mutant, knock, deficien, absen, SuperSeries, transgenic, CD[0-9]+) 的数据例。上述得到 3 个 GSE 数据集。以 R 抓取网页 (例如，RCurl::getURL 抓取 )，解析 ‘Overall design’ 和 ‘Samples’ 模块，匹配字符集 EX，排除匹配到的数据例。排除 ‘Overall design’ 中包含 in vitro, cell line, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。为了获得包含对照样本的数据集，(描述字符过少的例外) 以正则匹配 (‘Samples’, ‘Summary’, ‘Overall design’) 包含 normal, control, healthy, ctrl, adjacent, N[0-9]+, C[0-9]+ 的数据例。共 3 个。仅获取包含 ‘protein coding’ 测序的数据集，排除 ‘Samples’ 和 ‘Overall design’ 中包含 siRNA, miRNA, miR, lncRNA 字符的数据例。余下共 3 个。 以 GEOquery 获取 GSE 数据集 (n=3)。

# 3 参考文献和数据集

## 3.1 GSE 数据搜索 (INSOMNIA)

以 Entrez Direct (EDirect) 搜索 GEO 数据库 (检索条件见方法章节) 。 可用数据，及其组别为：

* **GSE40562**, **Type**: Microarray
  + NPL (n = 1)
  + NT (n = 1)
  + PLFFI (n = 3)
  + TFFI (n = 3)

## 3.2 GSE 数据搜索 (SD)

以 Entrez Direct (EDirect) 搜索 GEO 数据库 (检索条件见方法章节) 。 可用数据，及其组别为：

* **GSE48113**, **Type**: Microarray
  + 1 (n = 149)
  + 4 (n = 138)
* **GSE81761**, **Type**: Microarray
  + No\_Follow\_Up\_Data (n = 16)
  + No\_PTSD (n = 47)
  + PTSD\_Improved (n = 24)
  + PTSD\_Not\_Improved (n = 22)
* **GSE82113**, **Type**: Microarray
  + in\_phase (n = 11)
  + out\_of\_phase (n = 12)