GO富集

应用场景：在做单细胞分析时，其数据就是不同细胞和多个维度（feature即基因的表达），在处理单细胞数据时会给这些大量的细胞分一个类(形成几个cluster)，如何实现聚类呢？这个内容涉及请看其它文章……。聚类也是根据其每个细胞的基因表达来实现，涉及到的主要函数有这些处理……。聚类之后我们又能干什么，它们任然只是数据，并没有带来特别的生物学信息，如何从数据到表型/性状的认识呢？从数据分析的角度，可以利用基因与表型的对应关系来推测其表型，比如有些细胞的光合基因比较多，可能分为“光合细胞”？，对于一个cluster来说，那么这个cluster里面的细胞推测就可以按这个思路推测为“光合细胞”。基因与表型的对应关系可以在多个数据库查阅，来实现这个过程。

问题1：有那些数据库可以被用到基因到表型的对应呢？

分享12个基因信息查询相关的数据库 <https://mp.weixin.qq.com/s/9dZXetC_YzQQAYn1NSec0Q>

收藏！分享10个基因功能注释的数据库

<https://mp.weixin.qq.com/s/KTxH-da1hOaO5YEjZHoMqQ>

实现要准备的内容：cluster里面的特征基因，一个cluster总共的feature是很多的，我们要先进行一个简单过滤，通过选取pvalue\_adj小的基因作为检验基因

将目标基因与基因功能目录进行对应，或者说叫匹配/比对。再简单说就是我现在有两个词“单细胞”和“基因富集”，我要去词典里面去查这两个词的意思。欸，那么现在目标基因有了，用于检索的基因库呢？

问题2：如何拿到这个背景基因库/文库，用于检索？

有了这两个东西之后，用什么方法来做基因富集，具体而言就是你要选择哪个package/function来实现基因富集。

问题3：现有的基因富集的method不少，选择哪一种方法呢？

问题4：有了基因富集的结果，怎么来解读这个结果呢，以及这些结果如何更好的可视化？

手把手教学｜如何利用DAVID数据库进行GO、KEGG富集分析<https://mp.weixin.qq.com/s/nb7_mPa8fqjZ2YnQwYT42g>

GO富集分析主要从生物学过程（Biological Process，BP）、细胞组分（Cellular Component，CC）、分子功能（Molecular Function，MF）三个维度揭示各基因集中基因在基因本体（Gene Ontology，GO）上的富集情况。

KEGG（Kyoto Encyclopedia of Genes and Genomes）富集分析则是用于分析基因或蛋白在KEGG通路中的富集情况的一种常见的生物信息学分析方法。其中，KEGG Pathway是用于反映分子相互作用、反应及关系网络的路径图，主要包含代谢（Metabolism）、遗传信息处理（Genetic Information Processing）、环境信息处理（Environmental Information Processing）、细胞过程（Cellular Processes）、有机系统（Organismal Systems）、人类疾病（Human Diseases）、药物开发（Drug Development）七个方面的内容。

富集分析二三事（GO & KEGG）

<https://mp.weixin.qq.com/s/m-RpzMEK5cwvURTJKnCFrQ>

GSEA 基因富集分析一般的差异分析（GO和Pathway）往往侧重于比较两组间的基因表达差异，集中关注少数几个显著上调或下调的基因，对于差异基因检出的阈值，异常的敏感，客户需要给出差异基因的一个明确的定义(阈值)，例如abs(logFC) ≧2.0 & FDR ≦ 0.05，这容易遗漏部分差异表达不显著却有重要生物学意义的基因，忽略一些基因的生物特性、基因调控网络之间的关系及基因功能和意义等有价值的信息。GSEA 不需要指定明确的差异基因阈值，算法会根据实际数据的整体趋势，为研究者们提供了一种合理地解决目前芯片分析瓶颈问题的方法，即使在没有先前经验存在的情况下也能在表达谱整体层次上对数条基因进行分析，从而从数理统计上把表达谱芯片数据与生物学意义很好地衔接起来，使得研究者们能够更轻松、更合理地解读芯片结果。「MSigDB（Molecular Signatures Database）」：分子特征数据库是一组带注释的基因集，可与 GSEA 软件一起使用。这是一组用于 GSEA 软件的带注释基因集，数据库中定义了已知的基因集合：http://software.broadinstitute.org/gsea/msigdb.

GEO芯片注释缺少基因注释信息咋办？不慌！大海哥教你重注释

<https://mp.weixin.qq.com/s/c56zoevOEnKnIAv4ymnnQg>

EggNOG功能注释数据库在线和本地使用

<https://mp.weixin.qq.com/s/1p0oCtgM2HQkXGqfMiHsMQ>

1获得表达矩阵里面的核酸序列；2基于核酸序列拿到注释文件（通用id和蛋白质序列），利用seqmap或者blast与文库进行比对获得；3利用获得的蛋白质序列提交到eggnog上获得GO号和KEGG号；4利唯一的基因id和号的关系构建对应org库用于后续的富集分析；5富集分析准备target\_list和package\_org即可，简单使用clusterprofile的enrichGO

按照这个思路的话，转录本的核酸序列肯定是有的，但是我们需要它的一个通用名称，或者说是一个兼容基因组对应和蛋白组对应的名称

我去看了一下拟兰芥在NCBI上的基因组，广一个基因组就多的超乎想象的多的文件，so 我该如何认识不同文件，找到我所需要的高质量的reference

NCBI使用指南

<https://mp.weixin.qq.com/s/EMez9xO_h5KIqh-fDtP0kQ>

问题太多、太杂，回到关键问题，实现从核苷酸序列到蛋白质序列，到了蛋白质序列之后就可以用eggnog来得到GO和KEGG的号，这样就接上了基因富集的部分，所以现在学习基因的注释annotation部分！

注释的流程实在复杂，brain遭不住，后面还得系统学习基因组装和基因注释的流程