# 关于基因富集和marker基因

一些前置知识的学习，理解gene ontology

比较宽泛的全流程GO是什么的讲解<https://mp.weixin.qq.com/s/PxUzjiipfKt4r-LGSNYrDA>

关于GO数据的架构和生成讲的很好<https://mp.weixin.qq.com/s/vURgQakDlSfm5BtE1sf4xQ>

如何制作基因富集的背景库<https://mp.weixin.qq.com/s/J5kOFPAFAH2CZFOKS89OLQ>

教你如何解读富集后的结果<https://mp.weixin.qq.com/s/FCvyoFkBkpGZ94INk3o-og>

相关数据库

eggNOG-mapper <http://eggnog-mapper.embl.de/>

Gene Ontology 官网　<http://geneontology.org/>

EBI 的 GOA 数据库　<https://www.ebi.ac.uk/GOA/>

EBI 对 UniProtKB 参考Proteome <https://www.uniprot.org/proteomes> 进行注释，使用 InterProScan 进行功能注释，并关联GO,

提供的三种信息（BP，biological process；CC，cellular component；MF，molecular function）

Biological Process

生物学过程描述了基因和蛋白质在生物体内参与的各种生物学活动和事件。它涵盖了从分子水平到细胞、组织和整个生物体层面的过程，例如代谢、细胞分裂、发育、免疫反应、信号传导等。生物学过程的目的是描述和理解基因和蛋白质在生物学活动中的功能和相互作用。

Cellular Component

细胞组分描述了细胞内各种生物化学结构和细胞器的定位和组成。它涵盖了细胞膜、细胞核、线粒体、内质网、高尔基体、溶酶体等细胞内部分，以及细胞骨架、细胞间连接等细胞外部分。细胞组分的目的是揭示基因和蛋白质在细胞内的定位和功能，以及它们在细胞结构和组织中的作用。

Molecular Function

分子功能描述了基因和蛋白质分子在细胞中所具有的具体功能和活性。它涵盖了许多不同类型的活性，如催化反应的酶活性、结合特定分子的配体结合活性、信号转导的受体活性等。分子功能的目的是理解基因和蛋白质分子的功能特征，以及它们在生物体内参与的生物学过程和细胞组分中的作用。

背景文件即库怎么获得，自建库就更难了，前面我们做的org.CT的package的意义和内部建构又是什么？

一、什么是GO，GO的意义是什么？

1)什么是gene ontology？

GO的分类体系是一个层次化的结构，其中每个分类都具有唯一的标识符和定义。这种结构使得基因和蛋白质的功能可以通过精确而标准化的术语进行描述，从而促进了跨物种的功能比较和知识共享。GO的应用广泛涉及生物信息学、基因功能注释、基因表达数据分析等领域。它为研究人员提供了一个统一的语言和框架，用于理解和解释基因和蛋白质的功能及其在生物系统中的作用。研究人员可以利用GO进行基因功能富集分析，帮助识别在特定生物学过程中显著富集的基因集合。此外，GO还为基因表达数据的解释和可视化提供了有力的工具。  通过这些基因本体分类（BP、CC、MF）广泛应用于生物信息学研究和基因功能注释中，很好的帮助了研究人员更好地理解基因和蛋白质的功能及其在生物系统中的作用。因此，目前大量的文章中使用了GO分析来注释自己所要关注的基因，以及基因与基因之间的功能。

My viewpoint of the go enrich is a key from abstract number to specific phenotype/bio-meaning

2)GO的数据架构？

GO 定义的术语有着直接非循环式（directed acyclic graphs (DAGs)的特点，而并非是传统的等级制定义方式（随着代数增加，下一级比上一级更为具体）。举个例子来说，生物学途径中有一个定义是己糖合成，它的上 一级为己糖代谢和单糖合成。当某个基因被注解为“己糖合成活性”后，它自动地获得了己糖代谢和单糖合成地注解。因为在GO中，每个术语必须遵循“真途径 “法则，即如果下一代的术语可以用于描述此基因产物，其上一代术语也可以适用。

3)GO的注释（Annotation）[基因——基因产物（关系问题）]

那么，GO中的术语如何和相对应的基因产物相联系的呢？

使用GO的定义方法，对它们所包含的基因产物进行注解，并且提供支持这种注解的参考和证据。每个基因或基因产物都会有一个列表，列出与之相关的GO术语。每个数据库都会给出这些基因产物和GO术语的联系数据库，并且也可以在GO的ftp站点上和WEB方式查询到。而且，GO联合会提供了简化的本体论术语（GO slim），这样，可以在更高级的层面上研究基因组的功能。比如，粗略地估计哪一部分的基因组与信号传导、代谢合成或复制有关。GO对基因和蛋白的注释阐明了基因产物和用于定义他们的GO术语之间的关系。基因产物指一个基因编码的RNA或蛋白产物。因为一个基因可能编码多个具有很不相同性质的产物，所以GO推荐的注释是针对基因产物的而不是基因的。一个基因是和所有适用于它的术语联系在一起的。一个基因产物可以被一种本体论定义的多种分支或多种水平注释。注释需要反映在正常情况下此基因产物的功能，生物途径，定位等，而并不包括其在突变或病理状 态下的情况。GO联合会的各个数据库成员采用手动或自动的方式生成注释，这两种方式共有的原理是：一.所有的注释都需要有来源,可以是文字、另一个数据库或是计算机分析结果；二.注释必须提供支持这种基因产物和GO术语之间联系的证据。

二、如何获得GO的数据

文件格式的多样，从Gene Ontology上面获取的三种文件格式

Gene Association File (GAF)：这是GO的主要数据格式，它是一个扁平的文本文件，包含了基因和蛋白质与GO术语的关联信息。GAF文件每日更新，提供了最新的注释数据，包括基因和蛋白质的功能分类、描述以及与GO术语的关联。OWL/XML：这是GO的本体文件格式，它采用OWL（Web Ontology Language）的标准，以XML形式存储。OWL文件每月更新，提供了GO的完整本体结构和定义，包括术语的层次关系、属性和关联等信息。

MySQL数据库：GO数据还可以以MySQL数据库的形式进行下载和使用。这个数据库每月更新，提供了GO的完整数据集，方便用户进行更复杂和灵活的数据查询和分析。

用户可以从GO的FTP站点上获取这些数据文件，其中包括GAF文件、OWL/XML文件和MySQL数据库的备份文件。对于查找与特定GO术语相关的基因或基因产物，可以使用GAF文件中的注释数据进行查询。GAF文件提供了基因和蛋白质的标识符以及与之关联的GO术语的编号。通过使用这些编号，可以在不同的数据库中查找与之对应的基因或基因产物的相关信息。关于数据更新频率和具体下载方式，建议访问GO的官方网站（http://geneontology.org/）

三、如何运行GO，基本的调用函数/script？

基于library(clusterProfiler)的GO富集分析

四、一些工具帮助我们获得GO数据库，或者帮助我们进行GO的分析

自己从官网下载数据，自己注释，不仅麻烦，而且效率也不高，以下有一些可以简化GO分析数据的工具

DAVID（Database for Annotation, Visualization, and Integrated Discovery）：DAVID提供了一个综合的功能注释平台，可以将基因列表与GO术语进行关联，并提供功能富集分析、路径分析和可视化等功能。

Enrichr：Enrichr是一个在线工具，可以进行功能富集分析和可视化。它支持对基因列表进行GO功能富集分析，并提供了丰富的注释数据和可视化选项。

g:Profiler：g:Profiler是一个功能富集分析工具，支持多种物种和注释数据库。它可以将基因列表与GO术语进行关联，并提供详细的功能注释、富集分析和可视化结果。

WebGestalt：WebGestalt是一个功能富集分析工具，支持多种生物信息学数据库和注释资源。它可以进行GO功能富集分析、KEGG通路分析、疾病关联分析等，并提供交互式可视化结果。

clusterProfiler：clusterProfiler是一个基于R语言的功能富集分析包，专注于基因集合的GO和Pathway分析。它提供了丰富的功能注释数据和统计方法，并支持可视化结果的生成。（R语言包，明天单独介绍）这些工具都可以简化GO数据的分析过程，提供了用户友好的界面和丰富的功能选项。使用这些工具，研究人员可以更轻松地进行基因功能注释和功能富集分析，从而加快研究的进展和发现有意义的生物学结论。

基因富集的一些概念，GO库，本体学，三大类：BP、CC、MF

之前是先构建一个非模式生物的ORGpackage，然后通过这个来进行enrich

总的来看，做GO富集主要是两个准备：①目标探究基因，gene\_list;②可参考的GO富集库（物种对应性、基因名称对应性），这样才具有实际意义和可操作性。

从深层次来看难在于理解GO富集库的建立，从.GO文件来看，里面包括gene id和GO term，以及GO的注释分功能性描述和类别。那么基因与基因产物的注释基于大数据，离不开数据库的支持和验证实验保证基因产物认识的正确性。那么我们基于基因组注释得到基因产物，再通过基因产物注释得到GO富集库，这样建立其由基因到表型/功能的认识，那么我们就可以探究我们感兴趣的功能。

问题1：GO term的编号意义？在GO库提供详细的解释，便于人们查找基因产物信息吗，另外基因与基因之间是相互联系的，可以通过基因产物的联系去认识这种联系，所以去建立网络化的信息联系，在数据架构上是如何实现的呢？尤其是由上到下的方向性，难道这个在GO term的数字上体现？

问题2：GO 库提供了很好的资源，在实际分析和运用中存在的那些不足，在那些运用场景可以得到优化使用。基于这个问题，也是研究发展方向和热点思考，以及出现多种的GO分析方式我们选哪一个问题的回答。带着这个问题，我们应该去对比mulea和traditional clusterprofilter的区别，又有那些优势？

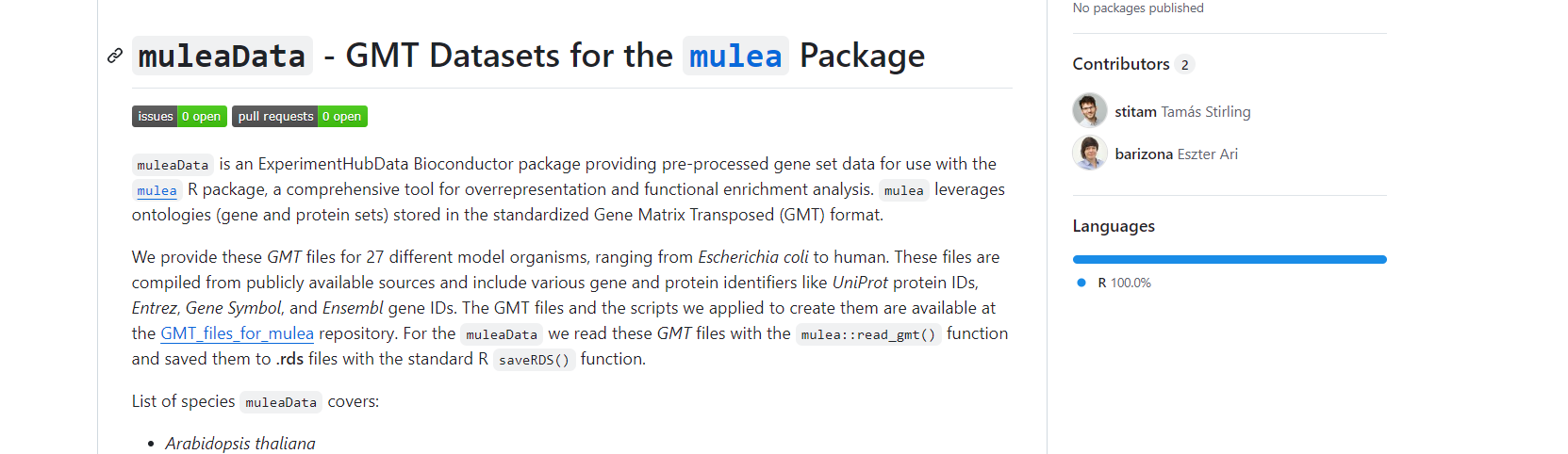
现在有一个新的tool，那就是mulea。

这个推文介绍<https://mp.weixin.qq.com/s/KRI7PujFDle0cZjyx0oOCw>

先来学习一下这个mulea

<https://github.com/ELTEbioinformatics/mulea?tab=readme-ov-file#usage>

gsea或者gsva所需要的gmt文件：是什么，怎么处理，以及应用 <https://mp.weixin.qq.com/s/VRy_A5fHOtCt1rMSQnVJVw>



muleaData提供了拟兰介的gmt文件，我们可否通过Cer与Ara的同源比对，获得自己的Cer的gmt文件，然后用于到mulea的ora训练

讲如何得到gmt很详细，看看数据结构 <https://github.com/ELTEbioinformatics/mulea?tab=readme-ov-file#formatting-the-results-of-a-differential-expression-analysis>

# Creating the ORA model using the GMT variable

ora\_model <- ora(gmt = tf\_ontology\_filtered,

# Test set variable

element\_names = target\_set,

# Background set variable

background\_element\_names = background\_set,

# p-value adjustment method

p\_value\_adjustment\_method = "eFDR",

# Number of permutations

number\_of\_permutations = 10000,

# Number of processor threads to use

nthreads = 2,

# Setting a random seed for reproducibility

random\_seed = 1)

# Running the ORA

ora\_results <- run\_test(ora\_model)

##########################################

empirical false discovery rate (eFDR) method

传统的基因集富集分析——通常仅限于少数本体，并且不考虑基因集或术语的相互依赖性，从而导致p值过度校正

通过其渐进式 eFDR 和支持各种本体来解决传统工具的局限性

mulea主打一个可整合多数据库的个性化分析策略

mulea与 Enrichr、g:Profiler 和 clusterProfiler 等工具具有功能相似性，但它具有几个独特的优势。

值得注意的是，与这些工具相比，mulea实现了更合适的多重检验校正方法 (eFDR)，使分析更敏感，可以检测到常用生物本体中的显著富集。

此外，mulea还通过包含 27 个物种为广泛的生物体提供了预定义的基因集。

因此，mulea超越了现有的基因集集合 MSigDB，后者仅限于人类和小鼠。这种更广泛的物种覆盖范围增强了mulea在各种研究环境中的适用性。

此外，mulea允许用户使用专用功能整合自己的本体，使他们能够利用来自不同来源的数据集并将分析范围扩展到默认选项之外。

这些独特的功能使mulea成为研究人员进行功能富集分析的多功能且用户友好的资源。

我们开始试试给的sample文件长什么样子？