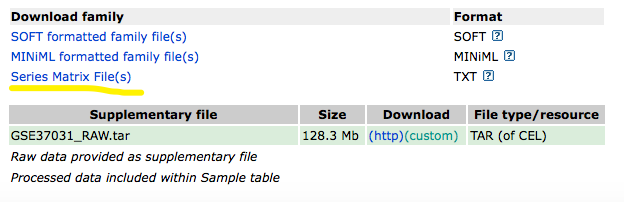
GSE、GPL文件均可在这里直接搜编号：

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi

GSE文件：



txt（依赖ftp，mac不能下）

为别人做好的基因Xsample的表达谱矩阵（R可直接导入txt）

GSE的txt文件：第一列ID是与不同芯片相关的特定ID

需下载对应的GPL文件



下载表格或者在线打开再全选复制

在网站上搜的GPL可能不全，不如在R里用GEOquery包下载全。

GEOquery可以直接下GSE、GPL文件，科学上网会更好。

R GEOquery包的下载：坎坷

install.packages不行

if (!requireNamespace("BiocManager", quietly = TRUE))

install.packages("BiocManager")

BiocManager::install("GEOquery")

options(BioC\_mirror="https://mirrors.tuna.tsinghua.edu.cn/bioconductor")

不行

BiocManager::install("GEOquery")

BJ：

现状：mac装了biocmanager但不能装包本体

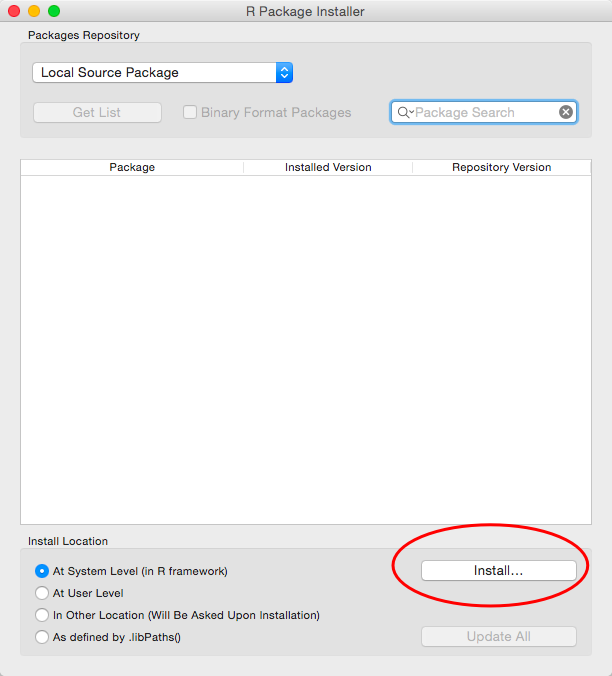
cran没有这个包

github可以打包下载，但慢

瓜瓜：GEOquery\_2.54.1.tgz R版本太低不能用

GEOquery-2.47.4.tar.gz 不知为什么反正装不上

以前如meta\_4.9-0.tar可以本地直接安装 如下



简单粗暴的流程：

https://www.jianshu.com/p/bc5d885ff4cf

GSE4100

GPL570

原始数据等下载：

http://www.360doc.com/content/17/0711/11/19913717\_670504853.shtml

GEOquery教程：

https://blog.csdn.net/twentyonepilots/article/details/86494193 （query直接出ID转换

http://www.bio-info-trainee.com/bioconductor\_China/software/GEOquery.html

差异分析教程：

https://blog.csdn.net/tuanzide5233/article/details/83541443

整个matrix log2可以直接得到log2的矩阵

log2之后 相减即可得到fold change

如果表达丰度的数值在50以内，通常是经过log2转化的。如果数字在几百几千，则是未经转化的。因为2的几十次方已经非常巨大，如果2的几百次方，则不符合实际情况。

对于是否需要标准化的问题，可以通过boxplot函数观察一下样本表达丰度值的分布是否整齐进行判断

基因的不同编号：

gene symbol： TP53

gene ID：7157=Entrez Gene ID=pubmed上找的ID=spot id

Gene Name：tumor protein p53

其它：

Ensembl:ENSG00000141510 即Ensembl数据库的ID编号，MIM:191170是来源于OMIM数据库（Online Mendelian Inheritance in Man ,人类孟德尔遗传在线数据库）的编号。Vega:OTTHUMG00000162125来自Vega数据库（Vertebrate Genome Annotation，脊椎动物基因组注释 ）

最后一般希望得到symbol，GENE ID转symbol最常用

GENBANK accession： NM啥啥啥 RNA片段 即GB\_ACC

比如：GSE59034

最左列在DAVID里应该是AFFYM EXTON ID，有的有AFFYM EXTON ID但没有NM啥啥啥，则无法在DAVID中匹配

基因ID转换：

R的clusterProfiler包/AnnotationDbi 包

https://blog.csdn.net/weixin\_40739969/article/details/89354167 （几种方法

<https://www.jianshu.com/p/0695a6a3c51f>

Bioconductor包/ org.Hs.eg.db包

http://www.bio-info-trainee.com/1089.html

https://blog.csdn.net/weixin\_40739969/article/details/103186027

https://blog.csdn.net/weixin\_40739969/article/details/89203971

DAVID

DAVID里GB\_ACC转基因名

https://www.sohu.com/a/132828786\_652735

可以做各种gene ID与SYMBLE的转换（效率不高）（但不能直接转芯片专属的ID ，BJ：GI list可以，我？）

https://david.ncifcrf.gov/conversion.jsp?VFROM=NA

直接粘贴基因编码或symbol在框里 每个换行

GENLIST



DAVID疑问：

https://www.dxy.cn/bbs/newweb/pc/post/36294035?source=rss

其它在线注释工具：

https://www.sohu.com/a/228051812\_419916

http://www.360doc.com/content/19/0317/10/49128372\_822156568.shtml

DEseq包

https://www.jianshu.com/p/a27dce71f6ea

火山图详解

https://www.jianshu.com/p/b16bbba4833c

带例子的火山图教程

https://genehub.wordpress.com/2018/04/25/r-volcano-plot/

limma

PCA

https://www.jianshu.com/p/8073226cdbd7

<https://www.jianshu.com/p/9226b777ae86>

<https://blog.csdn.net/u014801157/article/details/24372385> （各种包）

**差异倍数(fold change)**

fold change翻译过来就是倍数变化，假设A基因表达值为1，B表达值为3，那么B的表达就是A的3倍。一般我们都用count、TPM或FPKM来衡量基因表达水平，所以基因表达值肯定是非负数，那么fold change的取值就是(0, +∞).

为什么我们经常看到差异基因里负数代表下调、正数代表上调？因为我们用了log2 fold change。当expr(A) < expr(B)时，B对A的fold change就大于1，log2 fold change就大于0（见下图），B相对A就是上调；当expr(A) > expr(B)时，B对A的fold change就小于1，log2 fold change就小于0。通常为了防止取log2时产生NA，我们会给表达值加1（或者一个极小的数），也就是log2(B+1) - log2(A+1). 【需要一点对数函数的基础知识】