TCGA 下载和提取临床数据

一、数据库：TCGA

二、内容：下载临床数据，提取临床数据

三、癌症数据：宫颈鳞状细胞癌 CESC

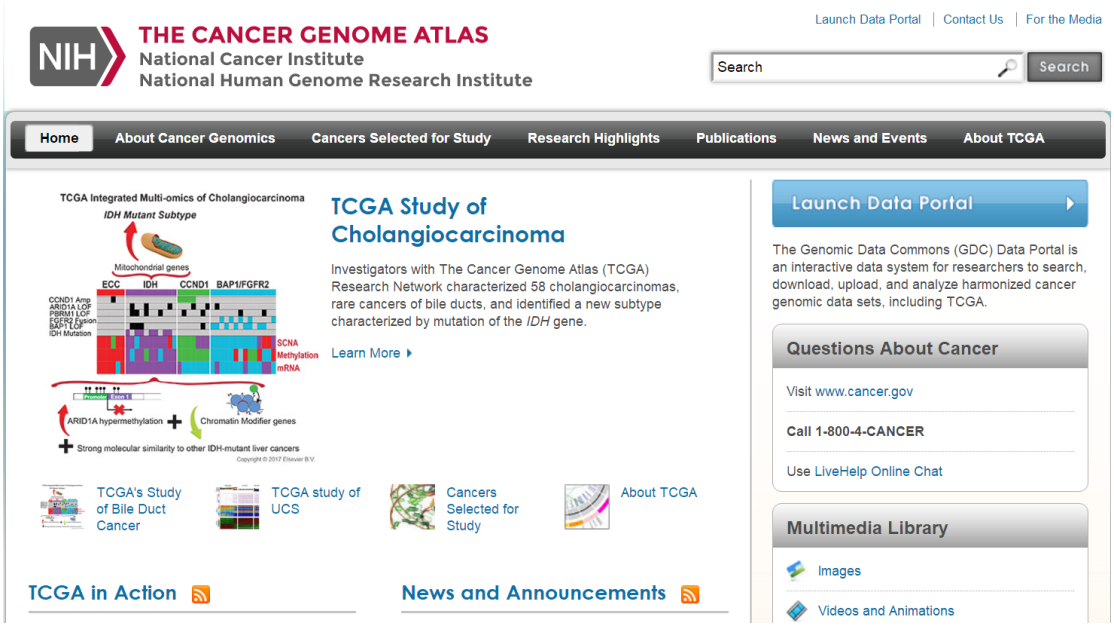
四、方法：

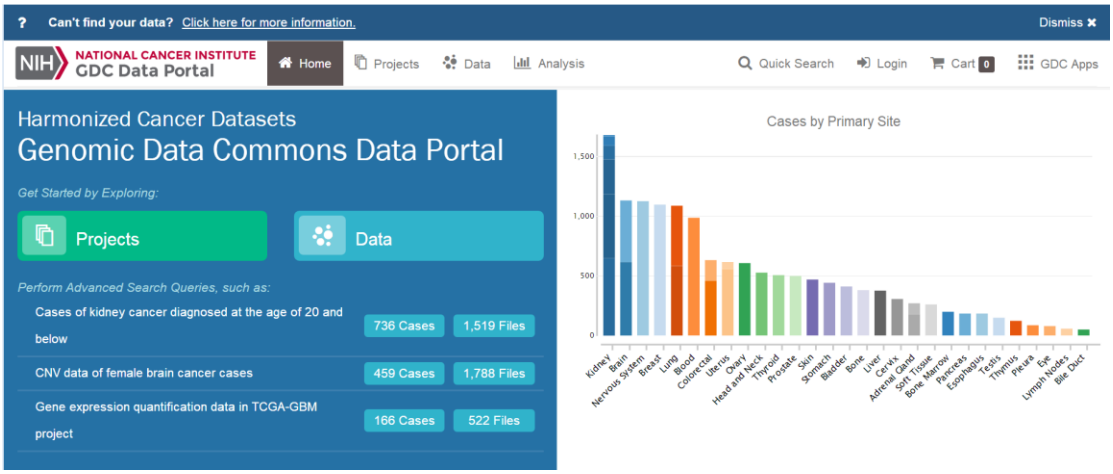
1、可视化下载 XML 原始文件

2、perl 脚本提取 XML 文件的临床信息，得到临床数据

五、步骤

1、登陆 TCGA 数据库官方网站，https://cancergenome.nih.gov/ ，点击"Launch Data Portal"进入数据库页面，或者直接登陆数据库网站：https://portal.gdc.cancer.gov/ 。进到数据库网站，点击“Data”，进入可视化选择页面。



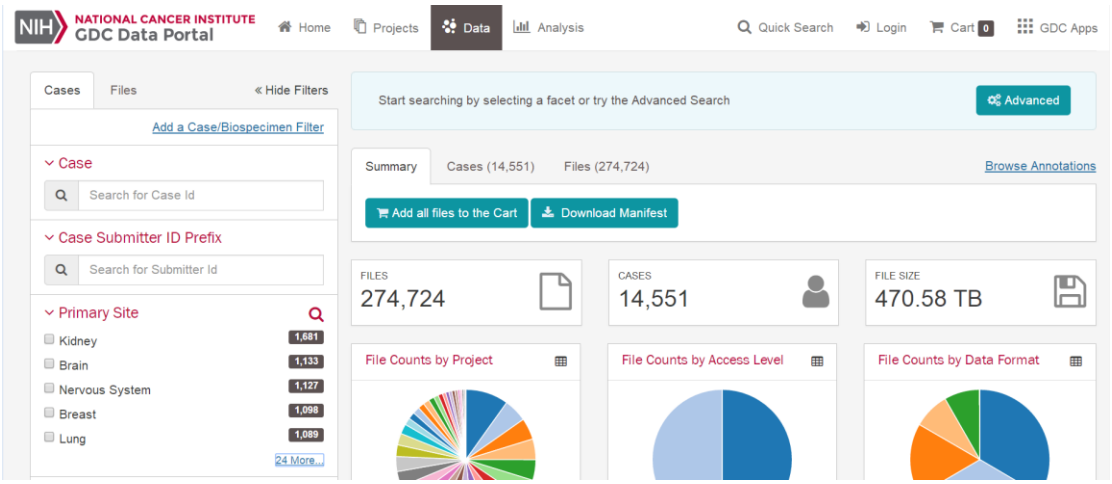


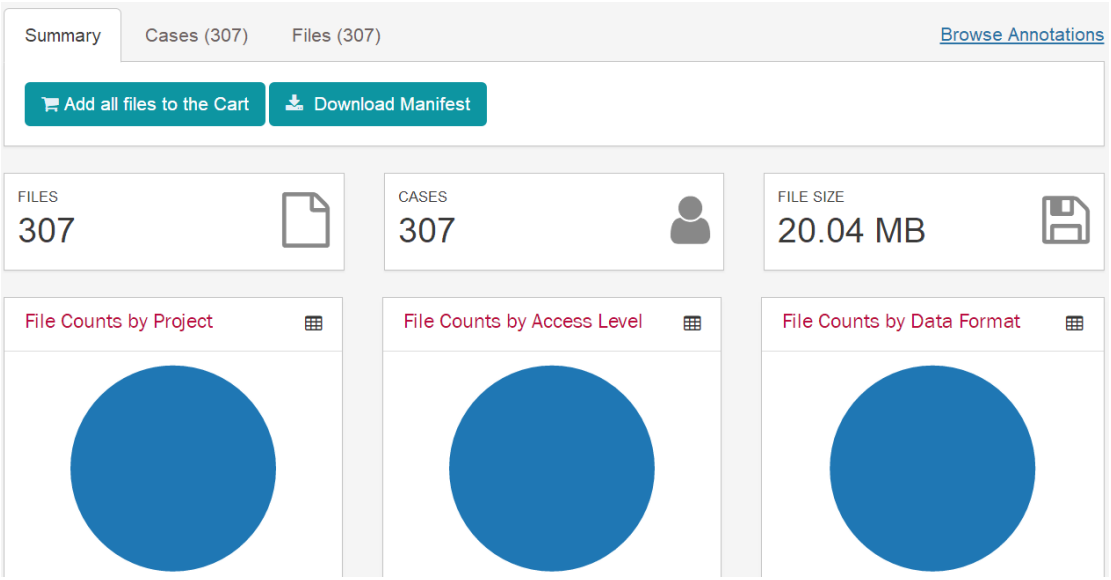
2、选择的方法：CASE 选项框依次选择——"Primary Site"-Cervix——"Cancer Program"-TCGA——"Project"-TCGA-CESC——其他默认即可

Files 选项框依次选择——"Data Category"-Clinical——其他默认即可

这是右边可以得到 Cases 数目 307 个，Files 数目 307 个，大小是 20.04M

说明：Case 是样本的数据，Files 是文件数目，在 mRNA 的数据时，经常出现 Cases 的数目和 Files 的数目是不相等的，这是因为，一个样本可能有多份数据。





3、点击"Add all files to the cart"，然后进入右上角的"Cart"进入数据展示和下载页面

说明："Cart"是 TCGA 数据库类似购物车的一个工具，里面是我们选到的数据界面。

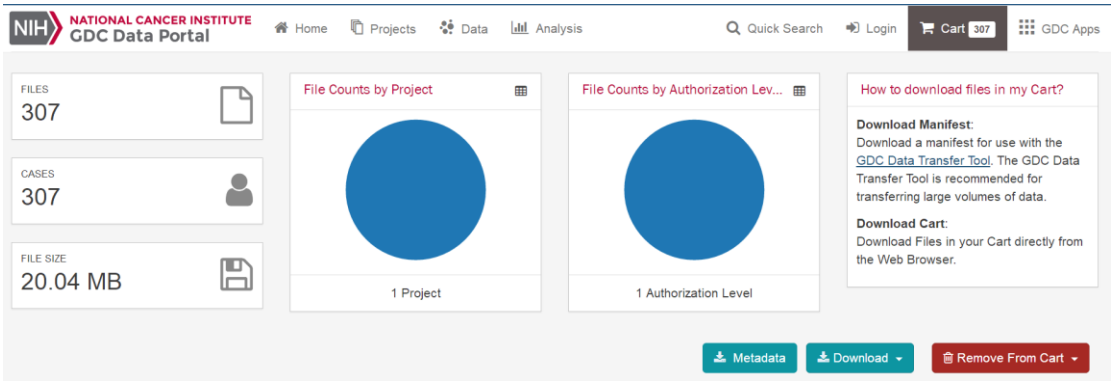
4、在“Cart"页面中，我们需要下载 3 个数据：Metadata、"Download"-Manifest、Cart

说明：

Metadata：最后一次随访的临床数据

Manifest：样本注释文件，主要用于 Data Transfer Tool 工具下载数据时使用

Cart：压缩包，包含所有的 XML 文件，也就是临床数据的压缩包文件。

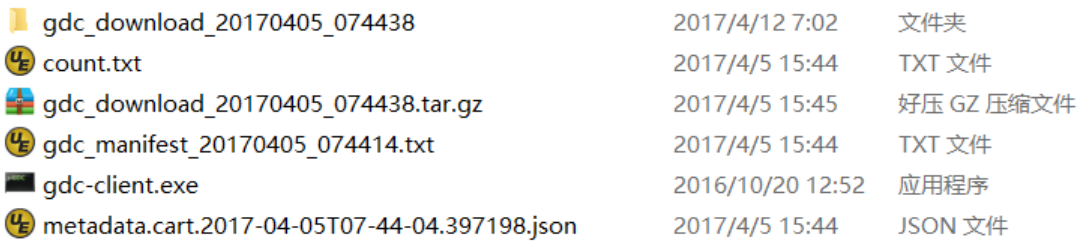


5、TCGA 数据库在数据下载有规定：让 Cart 文件夹大于 50M 时（这个依据网络情况，和下载用户数目），只能通过 Data Transfer Tool 工具进行下载。我们这里的 Cart 时 20.04M，

一般情况可以直接下载压缩包。注：后面 mRNA 的内容，我们会讲解如何使用 Data Transfer Tool 工具下载数据。

File Name	Cases	Project	Data Type	Data Format	Size
nationwidechildrens.org_clinical.TCGA-2W-A8YY.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	75 KB
nationwidechildrens.org_clinical.TCGA-4J-AA1J.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	78 KB
nationwidechildrens.org_clinical.TCGA-BI-A0VR.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	53 KB
nationwidechildrens.org_clinical.TCGA-BI-A0VS.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	57 KB
nationwidechildrens.org_clinical.TCGA-BI-A20A.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	66 KB
nationwidechildrens.org_clinical.TCGA-C5-A0TN.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	71 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BE.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	70 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BF.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	70 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BI.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	83 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BJ.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	72 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BK.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	68 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BL.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	76 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BM.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	58 KB
nationwidechildrens.org_clinical.TCGA-C5-A1BN.xml	1	TCGA-CESC	Clinical Supplement	BCR XML	65 KB

6、下载好所有需要的数据之后，我们需要用 perl 脚本提取文件里面的临床数据。我们首先把 gdc_download_20170405_074438.tar.gz 这个压缩包解压，解压得到 307 个文件夹，也就是一本样本一个临床数据文件夹。



7、把 307 个文件夹、MANIFEST.txt、get_clinical.pl 脚本放在一起，我们在 CMD 里面输入代码"perl get_clinical.pl MANIFEST.txt"，按回车，脚本文件开始运行，运行完就可以得到我们需要的 clinical.txt

Id	futime	fustat	age	gender	race	grade	clinical_sta	clinical_T	clinical_M	clinical_N	patholog	patholog	patholog	patholog	cancerTyp
TCGA-MY	1066	0	42	FEMALE	BLACK OR	G3	Stage IB1	unknow	unknow	unknow	unknow	T1b1	MX	N0	CESC
TCGA-Q1	499	0	64	FEMALE	WHITE	G3	unknow	unknow	unknow	unknow	unknow	TX	MX	NX	CESC
TCGA-HG	773	0	38	FEMALE	WHITE	G2	Stage IB2	unknow	unknow	unknow	unknow	T1b2	M0	N0	CESC
TCGA-Q1	483	0	45	FEMALE	WHITE	G1	Stage IB1	unknow	unknow	unknow	unknow	T1b	MX	N0	CESC
TCGA-VS	442	0	42	FEMALE	unknow	G1	Stage IIB	unknow	unknow	unknow	unknow	T2b	MX	NX	CESC
TCGA-DS	376	0	47	FEMALE	WHITE	G2	Stage IB	unknow	unknow	unknow	unknow	T1b1	M0	N1	CESC
TCGA-FU	954	0	47	FEMALE	WHITE	G2	Stage IB2	unknow	unknow	unknow	unknow	T1b2	MX	N1	CESC

```
use strict;
use warnings;
#use File::Basename;
use XML::Simple;
#use Data::Dumper;

my @dirs=glob("*");
open(WF,">clinical.txt") or die $!;
if(-d $dir){
      opendir(RD,"$dir") or die $!;
      while(my $xmlfile=readdir(RD)){
```

博淼生物 项目部