## TCGA数据观察和总结

更新至2015年2月5日

### TCGA数据下载

#### Data matrix

TCGA提供多种下载数据的方式，可以下载特定样本的特定molecular data和clinical data，也可以同时下载多个病人甚至一种癌症中所有病人的分子数据。在TCGA主界面中点击”launch data portal”后会出现所有癌症当前的样本量以及每种癌症最近一次更新数据的时间。点击进入癌症后比如LAML的“cases with data”能看到一个data matrix，其中包含了该癌症中所有病人的clinical and molecular data。通过点击相应样本后build archive就能得到下载链接。

#### Bulk download

TCGA提供该种数据下载方式，不需要通过选择病人来build archive等过程，就可以直接下载某一种或几种癌症中，不同类型数据和相应的平台下所测得的所有病人的数据。

#### Broad institute

Broad institute定期处理和发布level-3和level-4的数据，网址为gdac.broadinstitue.org/runs。其中analyses\_\_[###]为level-4数据，stddata\_[###]为level-3数据。其功能相当于bulk download，而用户可以根据需求，直接下载相应文件。

##### Dashboards

可在gdac.broadinstitute.org中“Dashboards”下载最新的level-3和level-4数据，分别对应“standard data”和“standard analyses”。Dashboards->standard data可以下载level-3数据，Dashboards->standard analyses可以下载level-4数据。而具体应该下载哪些文件，应根据所需要的数据类型来看。

##### Data

Data下主要描述的是level-3数据。在Data->Levels and Types中清晰全面的介绍了所所有类型的数据的层次，以及相应解释的链接。在Data->Workflow Graph中显示了所有level-3数据及相应的简单处理过程，但是每个标签不能点击，因此不能看到相应文件的具体处理方法。而在Dashboards->Full Samples Report中，我们可以得到所有癌症最新情况。同时，我们也应该了解从GDAC中究竟下载了什么样的数据。

GDAC首先需要过滤从DCC(TCGA存储数据)镜像过来的样本，利用所谓的校订(redaction)、重复过滤(replicate filtering)和拉黑名单(blacklisting)。所存储的地方被称为Firehose。

##### Analyses

在Analyses –> Release Notes中，主要描述每个季度Broad institute GDAC都在哪些分析流程中做了相应的修改。

##### 注意

标注有”FFPE”的样本，根据

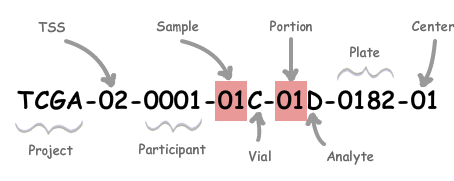
<http://gdac.broadinstitute.org/runs/sampleReports/latest/UCEC_FFPE_Cases.html>

中的解释，“FFPE (formalin fixed paraffin embedded) samples are not suitable for molecular analysis because the RNA and DNA are trapped in the nucleic acid-protein cross linking from the fixation process.”。因此可以不必下载标有FFPE的所有文件。

### TCGA的barcode注释

TCGA中每个病人都有一个unique barcode，用于记录与该病人相关的所有信息，包括clinical data和molecular data。一般来说，TCGA中有每个病人的癌症组织，但不一定有相应的癌旁组织。下面是TCGA中barcode方面的详细介绍。

The Barcode of samples are important:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Identifier for** | **Value** | **Value description** | **Possible values** |
| Project | Project name | TCGA | TCGA project | TCGA |
| [TSS](https://wiki.nci.nih.gov/x/IJhXAg) | Tissue source site | 02 | GBM (brain tumor) sample from MD Anderson | See [Code Tables Report](https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=Tissue%20Source%20Site) |
| [Participant](https://wiki.nci.nih.gov/x/8ZdXAg) | Study participant | 0001 | The first participant from MD Anderson for GBM study | Any alpha-numeric value |
| [Sample](https://wiki.nci.nih.gov/x/BZhXAg) | Sample type | 01 | A solid tumor | Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See [Code Tables Report](https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=Sample%20type)for a complete list of sample codes |
| [Vial](https://wiki.nci.nih.gov/x/KJhXAg) | Order of sample in a sequence of samples | C | The third vial | A to Z |
| [Portion](https://wiki.nci.nih.gov/x/_ZdXAg) | Order of portion in a sequence of 100 - 120 mg sample portions | 01 | The first portion of the sample | 01-99 |
| [Analyte](https://wiki.nci.nih.gov/x/dZRXAg) | Molecular type of analyte for analysis | D | The analyte is a DNA sample | See [Code Tables Report](https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=portion%20analyte) |
| [Plate](https://wiki.nci.nih.gov/x/95dXAg) | Order of plate in a sequence of 96-well plates | 0182 | The 182nd plate | 4-digit alphanumeric value |
| [Center](https://wiki.nci.nih.gov/x/2wlhAg) | [Sequencing](https://wiki.nci.nih.gov/x/IJZXAg) or [characterization](https://wiki.nci.nih.gov/x/CJZXAg) center that will receive the aliquot for analysis | 01 | The Broad Institute[GCC](https://wiki.nci.nih.gov/x/CJZXAg) | See [Code Tables Report](https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=center) |

## 数据处理

### Broad institute

根据<http://gdac.broadinstitute.org/Analyses-DAG.html可以知道Broad> institute都利用了什么样的pipeline处理和分析相关数据集。下面，根据不同类型的分子数据进行总结。

#### 预处理

##### 删除部分样本

GDAC会将从TCGA中出现TSS/BCR subject link出现问题的样本，或者是该样本实际上和tissue是不对应等情况，一般称为redaction。有时候，一个样本可能有好几个整份(aliquot)，因此GDAC也要根据一定的原则取其中之一存储在Firehose中。而拉黑(blacklisting)也是处理多整份(aliquot)时选择其中之一存储的一种方式。

此外，FFPE(formalin fixed paraffin embedded)由于在fixation过程中核苷酸-蛋白交联而不适用于继续分析其分子数据。

##### Preprocessors

1. mRNA expression

大概意思是：不同平台的选择样本最多的。

1. mRNAseq

用Illumina Hiseq/GA的RSEM或者RPKM，如果一个样本有两个平台的测量结果取Hiseq。所以我们的问题是，两个平台的结果混合起来是能用的？

1. miRseq

取RKM值，同样是Hiseq/GA取Hiseq。

1. methylation

分成两步预处理。我们可以处理原始文件，根据其中提到的步骤进行处理。

1. 删除比对到X或者Y染色体上的probe
2. 删除Gene symbol是NA的probe
3. 删除超过5%样本值为NA的probe
4. 如果一个probe比对到不止一个gene上，则分成相应gene多个行
5. 由于每个probe在所有样本中比对相同的位点和gene上，所以每行只需要留下beta值
6. Clinical数据

根据standard workflow可知，merge\_clinical数据转到了clinical\_pick\_tier\_1中。尚不清楚这两个文档的区别。

#### Level-3 stddata

根据<http://gdac.broadinstitute.org/stddata-DAG.html> 中的DAG，每种数据都有详细的相应详细的框图，虽然没有详细的介绍，但是有通过名称几乎都可以辨认。其中一些标有”preprocess”字眼的文件，可以在Data->Full Samples Report中看到。

##### Mutation数据

Level-3中有关mutation的数据有两个，maf文件和wig文件。Maf文件描述了每个病人call出来的mutation以及相应性质(properties)，其annotation可在<https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+%28MAF%29+Specification>中查到。Wiggle文件描述了dense, continuous data such as sequence coverage，可以参考<https://wiki.nci.nih.gov/display/TCGA/Wiggle+format>

<http://en.wikipedia.org/wiki/Deep_sequencing>

和<https://wiki.nci.nih.gov/display/TCGA/Wiggle+Format+Specification#WiggleFormatSpecification-Wiggle(WIG)filespecification>。

##### DNA甲基化数据

预处理后的甲基化数据在上文已经描述过。

##### CNV数据

Level-3都是segmented data，那么这些data是如何计算得到的？

##### RNA-Seq数据

Illumina HiSeq/GA两个平台经常被拿到一起用，那么是否可以一起用？

##### miRNA-Seq数据

Illunima Hiseq/GA两个平台，是否可以融合一起来使用？机器的区别在于reads长度、测序深度等吧？

##### mRNA表达数据

有多个平台，各个平台之间有不同，测量的信号应该也用不同的 数学公式来表达？

##### clinical信息(level-1&2)

首先，clinical data包括clinical data和biospeciman data。其中，clinical data主要包括病人的人口统计信息(demographic information)、治疗信息(treatment information)和生存信息(survival data)等；biospecimen data主要描述了从病人得到的组织样本是如何被Biospeciman core resource center(BCR)处理的。

根据Data levels and Data types可以知道，Clinical data主要包括biotab(level-2)临床信息，每个病人单独显示的信息xml(level-1)，一些病人有pathological reports。我们选择biotab格式进行研究，并选择其中较为重要的词条(CDE term)作为备选。Biotab文件分成两类，分别以biospeciman和clinical作为关键字。Biospeciman关键字文件有11个，分别以aliquot, analyte, cqcf, diagnostic\_slides, normal\_control, portion, protocol, sample, shipment\_portion, slide和tumor sample；clinical关键字文件一般有8个，分别是cqcf, drug, follow\_up\_v4.0, follow\_up\_v4.0\_nte, nte, omf\_v4.0, patient和radiation。下面就按照这两大类关键字中相应文件介绍。在介绍之前，有必要了解我们应以哪种方式理解clinical信息。在观察biotab文件时，每个term一般来说对应一个CDE的ID，有两个资源可以用于理解该term具体含义。

1. BCR data dictionary

<https://tcga-data.nci.nih.gov/docs/dictionary/>

其中描述了当前clinical data中已经有的所有term所指含义

1. Clinical Data Elements (CDE) browser

<https://cdebrowser.nci.nih.gov/CDEBrowser/>

包含TCGA中还尚未包括的term所指含义。

截止到2015年2月，TCGA只给出了每个病人初次确诊的年费而没有精确的月日。我们不能用样本处理的信息来反推病人的临床信息。在omf\_v4.0文件中，只有少量样本提供了surgical resection时间。在drug文件中，描述了TCGA中一些病人的用药信息。可以通过follow\_up\_v4.0文件提取病人的生存信息。

#### Level-4 analyses

##### Mutation

GDAC利用MutSig1.5，MutSig2.0和MutSigCV来计算significantly mutated genes。

###### MutSig

<http://www.broadinstitute.org/cancer/cga/mutsig>

该链接非常清楚的解释了MutSig几个版本的软件，包括1.0, 1.5, S2N和CV的区别。当前最新的软件为MutSigCV。

<http://www.broadinstitute.org/cancer/cga/mutsig_run>

该链接解释了如何使用MutSigCV。其中”CV”指coviriate，是mutsig其他版本算法软件中没有使用过的一种数据，主要描述了？

###### IntOGen

IntOGen采用了Oncodrive-FM来计算candidate driver gene；用oncodriveClust来计算在一个protein中，gene层次的mutation是不是显著聚集。

##### CNV

GDAC利用Gistic2，根据segmented value来计算出significant focal or arm level amplification/deletion。

###### Gistic2

利用segmented value来计算focal and arm-level amplification/deletions。

###### OncodriveCIS

##### Methylation