[38]:
```
%load_ext autoreload
%autoreload 2

%matplotlib inline
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

Descarga de Metadatos

Primer *Notebook* a ejecutar para hacer una descarga de datos del servidor de Genomic Data Commons (GDC). Se obtiene información de todos los casos que cumplan el filtro indicado en el fichero de configuración llamado 'cases info'. Para cada uno de esos casos se obtienen los metadatos de las imágenes histólogicas disponibles y de ficheros de datos ómicos. Los datos se guardarán en la ruta indicada en la configuración (*data_path*). Los ficheros de salida son:

- cases.csv
- slides_metadata.csv
- rnaseq_metadata.csv

Packages

[6]:
```
from gdc.metadata import get_rnaseq_metadata
from gdc.metadata import get_slides_metadata
from gdc.metadata import get_cases
```

[7]:
```
import requests
import json
import yaml
import pandas as pd
import io
import os
```

[8]:
```
pd.options.display.max_columns = 100
```

Config

[27]:
```
with open('conf/user_conf.yaml', 'r') as f:
    conf = yaml.load(f)
```

Cases Data

Descarga la información indicada para cada uno de los casos. Se ha seleccionado utilizar todos los datos del proyecto TCGA-PAAD (Pancreatic Adenocarcinoma).

[16]:
```
conf['cases_info']
```

[16]:
```
{'project_id': 'TCGA-PAAD'}
```

Los campos a obtener para cada uno de los pacientes también son configurables. Se puede ver una lista de los campos disponibles en la siguiente URL: https://docs.gdc.cancer.gov/API/Users_Guide/Appendix_A_Available_Fields/. Se puede obtener

información demográfica del paciente, su consumo de alcohol y/o cigarrillos asi como datos del diagnóstico.

[20]:
```
conf['fields']['cases']
```

[20]:
```
{'cases': ['primary_site', 'disease_type'],
 'project': ['program.name', 'project_id'],
 'demographic': ['gender', 'race'],
 'diagnoses': ['age_at_diagnosis',
  'tissue_or_organ_of_origin',
  'primary_diagnosis',
  'tumor_stage',
  'morphology'],
 'exposures': ['height', 'weight', 'alcohol_history', 'cigarettes_per_day']}
```

La función *get_cases* obtiene un DataFrame con todos los casos y las columnas indicadas mediante una llamada a la API de GDC.

[17]:
```
cases_df = get_cases(conf['cases_info'], conf['fields']['cases'])
```

Guarda la salida en un CSV

[10]:
```
cases_df.to_csv(os.path.join(conf['data_path'], 'cases.csv'), sep='|', index=False)
```

Muestra de la salida:

[18]:
```
cases_df.sample(5)
```

[18]:

	case_id	primary_site	disease_type	program_name	project_id	gender	race	age_at_diagnosis	tissue_or_organ_of_origin	primary_diagnosis	tumor_stage	morphology	height	weight	alcohol_history	cigarettes_per_day
142	TCGA-IB-7889	Pancreas	Ductal and Lobular Neoplasms	TCGA	TCGA-PAAD	female	white	31393	Pancreas, NOS	Infiltrating duct carcinoma, NOS	stage iib	8500/3	NaN	NaN	No	NaN
171	TCGA-HZ-7289	Pancreas	Adenomas and Adenocarcinomas	TCGA	TCGA-PAAD	male	white	28174	Head of pancreas	Adenocarcinoma, NOS	stage iib	8140/3	NaN	NaN	No	NaN
86	TCGA-HZ-8003	Pancreas	Ductal and Lobular Neoplasms	TCGA	TCGA-PAAD	female	white	23868	Head of pancreas	Infiltrating duct carcinoma, NOS	stage iib	8500/3	NaN	NaN	Not Reported	NaN
143	TCGA-Z5-AAPL	Pancreas	Ductal and Lobular Neoplasms	TCGA	TCGA-PAAD	female	white	27152	Pancreas, NOS	Infiltrating duct carcinoma, NOS	stage iia	8500/3	NaN	NaN		
19	TCGA-FB-AAPP	Pancreas	Ductal and Lobular Neoplasms	TCGA	TCGA-PAAD	male	white	26239	Head of pancreas	Infiltrating duct carcinoma, NOS	stage iib	8500/3	NaN	NaN		

```
143    Not Reported                NaN
19              Yes                NaN
```

Files Metadata

Slides Metadata

A continuación se obienen los metadatos de las imágenes histológicas. Además del filtro de pacientes que se aplica en los casos también se puede filtrar el tipo de *slide*. Las opciones diponibles son:

- *Tissue Slide*: muestras almezandas mediante congelación.
- *Diagnostic Slide*: tejidos embebidos en parafina y fijados con formalina (FFPE).

Las muestras FFPE son las más utlizadas por los médicos para el diagnóstico y el tejido se converva mejor, sin embargo los químicos en los que se almacenan afectan a los ácididos nucleicos, no siendo válidos para anális de ADN o RNA.

Las muestras congeladas son más baratas de almacenar y suelen ser utilizar después de la cirujía para determinar si los bordes del tumor están limpios, es decir, si el tumor ha sido completamente eliminado. Por tanto sobre estas muestras se recogen datos sobre el porcentaje de las células afectadas por el tumor. Otra ventaja de estas imágenes es que es muy común que se haya recogido también su información genética, lo que permitiría un análisis de correlación entre la morfología del tejido tumoral y la experesión genética.

En este caso se han seleccionado ambos tipos para mostrar la diferencia de los datos disonibles pero para el proyecto se utilarán únicamente tejidos congelados por los motivos expuestos.

[61]:
```
conf['slide_types']
```

[61]:
```
['Tissue Slide', 'Diagnostic Slide']
```

[33]:
```
slides_df = get_slides_metadata(conf['cases_info'], conf['fields']['slides'],␣
 ↪experimental_strategies=conf['slide_types'])
```

[24]:
```
slides_df.to_csv(os.path.join(conf['data_path'], 'slides_metadata.csv'), sep='|',␣
 ↪index=False)
```

Muestra de tejidos congelados

[34]:
```
slides_df[slides_df['experimental_strategy'] == 'Tissue Side'].sample(3)
```

[34]:
```
                                file_id       case_id         sample_id  \
264  9ab2c9cf-0170-47c3-9857-fc8f65270835  TCGA-IB-A5SO  TCGA-IB-A5SO-01A
138  38cd9f74-9697-4ae7-a7c4-b25eb921610b  TCGA-HV-A7OP  TCGA-HV-A7OP-01A
413  8ceee345-b355-4b2a-a08b-097beca25d84  TCGA-HZ-A77O  TCGA-HZ-A77O-01A

                   slide_id    data_type experimental_strategy data_format  \
264  TCGA-IB-A5SO-01A-01-TSA  Slide Image          Tissue Slide         SVS
138  TCGA-HV-A7OP-01A-01-TS1  Slide Image          Tissue Slide         SVS
413  TCGA-HZ-A77O-01A-01-TS1  Slide Image          Tissue Slide         SVS

     file_size                    file_name primary_site  \
264     366.10  TCGA-IB-A5SO-01A-01-TSA.svs     Pancreas
138     207.88  TCGA-HV-A7OP-01A-01-TS1.svs     Pancreas
```

```
413     160.72  TCGA-HZ-A77O-01A-01-TS1.svs     Pancreas

                    disease_type    sample_type  is_ffpe  \
264  Ductal and Lobular Neoplasms  Primary Tumor    False
138  Ductal and Lobular Neoplasms  Primary Tumor    False
413  Ductal and Lobular Neoplasms  Primary Tumor    False

     percent_normal_cells  percent_stromal_cells  percent_tumor_cells  \
264                   8.0                   65.0                 25.0
138                   0.0                   25.0                 70.0
413                   0.0                    0.0                100.0

     percent_tumor_nuclei
264                  50.0
138                  70.0
413                  75.0
```

Muestra de tejidos FFPE

Se observa que no las columnas *percent_normal_cells, percent_stromal_cells, percent_tumor_cells, percent_tumor_nuclei* están siempre a nulo ya que no se recogen estos datos.

[36]:
```
slides_df[slides_df['experimental_strategy'] == 'Diagnostic Slide'].sample(3)
```

[36]:
```
                                file_id       case_id         sample_id  \
151  928a0146-f9ab-4f05-8bb4-08c5af116755  TCGA-HV-AA8X  TCGA-HV-AA8X-01Z
93   1571f1c2-e6b5-41cf-a61d-69faac8ffea6  TCGA-HZ-7924  TCGA-HZ-7924-01Z
419  c5ccbe59-3312-4522-ba7d-d57999499844  TCGA-RB-A7B8  TCGA-RB-A7B8-01Z

                    slide_id    data_type experimental_strategy data_format  \
151  TCGA-HV-AA8X-01Z-00-DX1  Slide Image      Diagnostic Slide         SVS
93   TCGA-HZ-7924-01Z-00-DX1  Slide Image      Diagnostic Slide         SVS
419  TCGA-RB-A7B8-01Z-00-DX1  Slide Image      Diagnostic Slide         SVS

     file_size                    file_name primary_site  \
151     242.98  TCGA-HV-AA8X-01Z-00-DX1.svs     Pancreas
93       61.94  TCGA-HZ-7924-01Z-00-DX1.svs     Pancreas
419    2578.13  TCGA-RB-A7B8-01Z-00-DX1.svs     Pancreas

                    disease_type    sample_type  is_ffpe  \
151  Ductal and Lobular Neoplasms  Primary Tumor     True
93   Ductal and Lobular Neoplasms  Primary Tumor     True
419  Adenomas and Adenocarcinomas  Primary Tumor     True

     percent_normal_cells  percent_stromal_cells  percent_tumor_cells  \
151                   NaN                   NaN                  NaN
93                    NaN                   NaN                  NaN
419                   NaN                   NaN                  NaN

     percent_tumor_nuclei
151                   NaN
93                    NaN
```

```
419                    NaN
```

El 97% de las muestras FFPE pertenecen a tejidos tumorales.

[37]:
```
slides_df.groupby(['experimental_strategy', 'sample_type']).size()
```

[37]:
```
experimental_strategy  sample_type
Diagnostic Slide       Primary Tumor          203
                       Solid Tissue Normal      6
Tissue Slide           Metastatic               1
                       Primary Tumor          219
                       Solid Tissue Normal     37
dtype: int64
```

RNA-Seq Metadata

Por último se obtienen los nombres de los ficheros de expresión genética, en ese caso también se puede seleccionar qué tipo de fichero queremos, las opciones son:

- HTSeq - Counts
- HTSeq - FPKM
- HTSeq - FPKM-UQ

El primero tiene la cuenta en bruto de las lecturas mapeadas a cada gen. Los otros dos están normalizados respecto al número de cuentas, la diferencia entre ellos es que *UQ* utiliza el percentil 75% mientras que el otro la cuenta total.

[66]:
```
conf['rnaseq_types']
```

[66]:
```
['HTSeq - Counts', 'HTSeq - FPKM-UQ', 'HTSeq - FPKM']
```

[68]:
```
rnaseq_df = get_rnaseq_metadata(conf['cases_info'], conf['fields']['rnaseq'],␣
↪workflow_types=conf['rnaseq_types'])
```

[69]:
```
rnaseq_df.to_csv(os.path.join(conf['data_path'], 'rnaseq_metadata.csv'), sep='|',␣
↪index=False)
```

Observar que para cada una de las muestras se obtienen los 3 ficheros de RNA-Seq.

[76]:
```
rnaseq_df.sort_values('sample_id').head(6)
```

[76]:
```
                                file_id       case_id         sample_id  \
71   e2567946-b4e5-408a-950b-3bb6c130b2a1  TCGA-2J-AAB1  TCGA-2J-AAB1-01A
184  6deb2016-8321-465d-b4ab-05d92a4c04c0  TCGA-2J-AAB1  TCGA-2J-AAB1-01A
216  012007d2-bc82-4a37-a123-ff30e18629b8  TCGA-2J-AAB1  TCGA-2J-AAB1-01A
396  53aaf772-2876-46eb-8efe-1bbd4c5b6df6  TCGA-2J-AAB4  TCGA-2J-AAB4-01A
350  0fed1108-f65c-45e0-9695-66fb085119c7  TCGA-2J-AAB4  TCGA-2J-AAB4-01A
479  9f2e5dd4-df50-4efe-bcbb-a919a02f1114  TCGA-2J-AAB4  TCGA-2J-AAB4-01A

                                      rnaseq_id  \
71    caf9cab4-f98f-46bd-a75d-0eb1e9c6c9ea_count
184    caf9cab4-f98f-46bd-a75d-0eb1e9c6c9ea_fpkm
```

```
216  caf9cab4-f98f-46bd-a75d-0eb1e9c6c9ea_uqfpkm
396    057aa9ac-f22c-4c11-a44d-ad52ae59b4cf_fpkm
350   057aa9ac-f22c-4c11-a44d-ad52ae59b4cf_count
479  057aa9ac-f22c-4c11-a44d-ad52ae59b4cf_uqfpkm

                        data_type experimental_strategy data_format  \
71   Gene Expression Quantification               RNA-Seq         TXT
184  Gene Expression Quantification               RNA-Seq         TXT
216  Gene Expression Quantification               RNA-Seq         TXT
396  Gene Expression Quantification               RNA-Seq         TXT
350  Gene Expression Quantification               RNA-Seq         TXT
479  Gene Expression Quantification               RNA-Seq         TXT

     file_size                                file_name    workflow_type  \
71        0.25   TCGA-2J-AAB1-01A_HTSeq-Counts.txt.gz   HTSeq - Counts
184       0.51     TCGA-2J-AAB1-01A_HTSeq-FPKM.txt.gz     HTSeq - FPKM
216       0.51  TCGA-2J-AAB1-01A_HTSeq-FPKM-UQ.txt.gz  HTSeq - FPKM-UQ
396       0.52     TCGA-2J-AAB4-01A_HTSeq-FPKM.txt.gz     HTSeq - FPKM
350       0.26   TCGA-2J-AAB4-01A_HTSeq-Counts.txt.gz   HTSeq - Counts
479       0.53  TCGA-2J-AAB4-01A_HTSeq-FPKM-UQ.txt.gz  HTSeq - FPKM-UQ

    primary_site                            disease_type    sample_type
71      Pancreas  Cystic, Mucinous and Serous Neoplasms  Primary Tumor
184     Pancreas  Cystic, Mucinous and Serous Neoplasms  Primary Tumor
216     Pancreas  Cystic, Mucinous and Serous Neoplasms  Primary Tumor
396     Pancreas           Adenomas and Adenocarcinomas  Primary Tumor
350     Pancreas           Adenomas and Adenocarcinomas  Primary Tumor
479     Pancreas           Adenomas and Adenocarcinomas  Primary Tumor
```