Lab10

GDC Data Lab

By

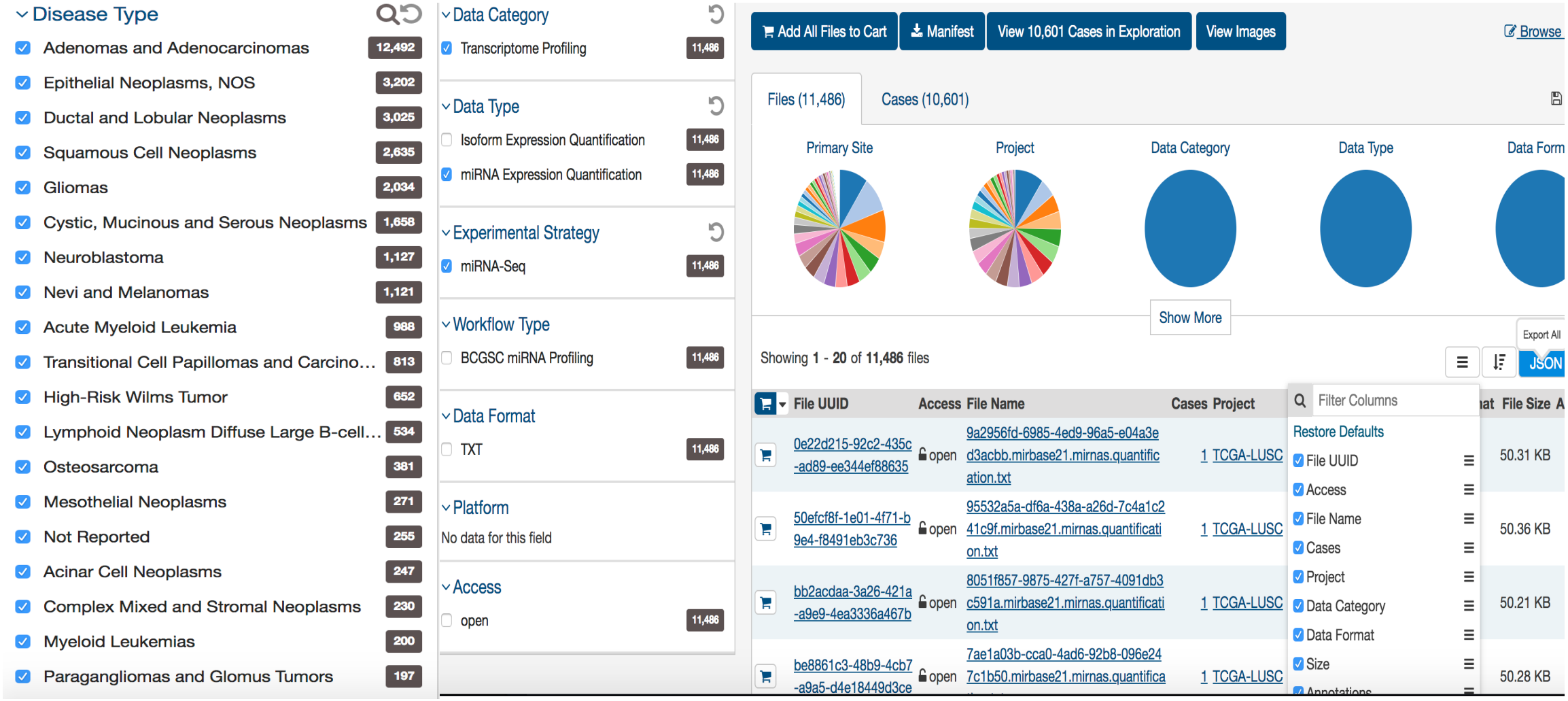
Siva Sai Krishna Marthy

Jianan gao

Xin Wang

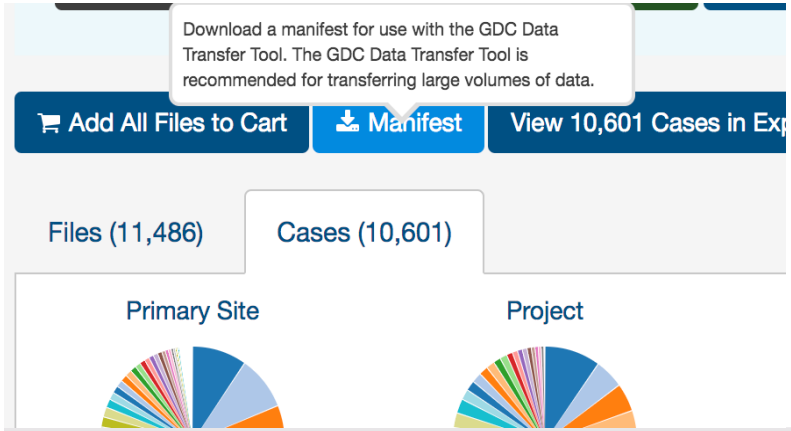
https://github.com/appleshine77/GDC-data-lab.git

Part1: Data download, integration and preprocess



Disease Type selects all in case page, and in file page, only select miRNA, there are totally 11,486 files and 10,601 cases;

Part1: Data download, integration and preprocess



gdc_manifest.2018-10-21.txt

id	filename	md5	size	state
0e22d215-92c2-435c-ad89-ee344ef88635	9a2956fd-6985-4ed9-96a5-e04a3ed3acbb.mirbase21.mirnas.quantification.txt	db8bf73e8e380425980b9df4ffe49d42	50309	released
50efcf8f-1e01-4f71-b9e4-f8491eb3c736	95532a5a-df6a-438a-a26d-7c4a1c241c9f.mirbase21.mirnas.quantification.txt	5fadff056f52ba9278ffc370e13e7328	50362	released
bb2acdaa-3a26-421a-a9e9-4ea3336a467b	8051f857-9875-427f-a757-4091db3c591a.mirbase21.mirnas.quantification.txt	0e278be8f5caf3cc822a47631c10c287	50206	released
be8861c3-48b9-4cb7-a9a5-d4e18449d3ce	7ae1a03b-cca0-4ad6-92b8-096e247c1b50.mirbase21.mirnas.quantification.txt	a10935d970d86448e955a497866f9256	50277	released
0435e383-6ec2-46c4-ba15-713a2447b0d1	37be9876-e459-48e9-b9ba-b47d02c17ae7.mirbase21.mirnas.quantification.txt	26683c027ed3504836f79b1ccaf28090	50426	released
f7e9016d-2d0b-40d1-a330-09ae95684813	69c2dd2f-9df4-4f79-9d18-577ad2d53dfd.mirbase21.mirnas.quantification.txt	82a9b5796efe1f00f7fdda3ca8b8471a	50307	released
07fa22d0-99b1-4364-9b93-d937005a2416	7df584d3-5b11-4987-8227-6d809e8305af.mirbase21.mirnas.quantification.txt	51550cb760688a6242cb3e6a301d370b	50150	released
d6e60c88-6acc-45a0-abe8-e8bcbfccf369	36ff000c-51b2-4e3d-9300-4c536d0ae141.mirbase21.mirnas.quantification.txt	96b72016cddb3f0cb627abf64c909c5f	50392	released
2c5eacad-0a4f-43e6-ad8f-dae2ffe4b0c6	cb606a2b-3c67-44c8-866a-2422e7ac369f.mirbase21.mirnas.quantification.txt	6343156b82a3deaa4ef11098d7e2d1c5	50297	released
5f06f29b-ddc7-43ae-a2d0-7df4349f1507	1f39d8bd-9c1f-4dcf-9de0-62e3966b014f.mirbase21.mirnas.quantification.txt	65eed838d29779714f8c2ce8575225b7	50331	released
94a5d869-a95f-4dca-9f5a-59aaaeb39670	f7ca0295-d9da-4a0d-8923-8a2053aedfe3.mirbase21.mirnas.quantification.txt	cbcc855aa63622a6bad762ac2a66a2ca	50230	released
314186da-49ad-4db7-b3c2-c1e0b262f4f8	d3394983-097e-45dd-b449-0dded9446dcd.mirbase21.mirnas.quantification.txt	e41ac4c76f0c8c490f01d8a0c4459152	50267	released
ece4bd52-0c7b-4cb8-ab6c-fbe10f97eb36	ad29232f-e5e5-4206-8ab3-cf45ac0c49e1.mirbase21.mirnas.quantification.txt	1618d7a51dcbfcd8cd3c4af088d544bb	50424	released
c7cf5c61-54b6-41ea-8c8e-294211f2b7fa	b1475bd4-e50b-41bf-a2dd-19bcd035805b.mirbase21.mirnas.quantification.txt	ba3a546250b44cdbd9e4210061eae37e	50207	released
5984e36f-e65b-4095-a1f9-c848bc0fe08a	944e5ba8-7c95-4b21-81eb-c02b4eb0bbed.mirbase21.mirnas.quantification.txt	162d0e4f1bf2092cb32ea9dedda3a22b	50321	released
ac3a3cb1-b6ce-41a8-af04-132012e48df3	064d7af1-f631-42aa-bfa6-23afabdee202.mirbase21.mirnas.quantification.txt	f1da8f5fbbe352d74c6f6312673f5a48	50340	released
ae4cd42a-c88a-4e9d-8a19-f93a9ade179a	e44b8d48-2216-462c-8b83-c220c4c6d2ee.mirbase21.mirnas.quantification.txt	fde3b7bc7791af6d70635884cbcd01a0	50103	released
54057a72-d718-4ba7-9e04-68036c16a79f	48053f52-5a54-417c-abe6-190e236a47df.mirbase21.mirnas.quantification.txt	32c5d6dce8fba6d739ae158fda5c9e4d	50126	released
fe335358-a468-49b1-89ec-31e1b5e084bd	72bb1711-cec9-4de1-ad36-80b806dbcdde.mirbase21.mirnas.quantification.txt	7d6a4f51faed7dc9380b876ebb5a90d8	50276	released

Manifest download

Part1: Data download, integration and preprocess

```
XINs-MacBook-Pro:miRNA xin$ python -3 ../../src/check.py
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:16: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
  from . import multiarray
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:34: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
  from . import umath
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/_internal.py:204: DeprecationWarning: Overriding __e
q__ blocks inheritance of __hash__ in 3.x


/Users/xin/anaconda/lib/python2.7/site-packages/pandas/core/reshape/util.py:76: DeprecationWarning: reduce() no
t supported in 3.x; use functools.reduce()
  return reduce(_compose2, funcs)
/Users/xin/anaconda/lib/python2.7/site-packages/pandas/io/pytables.py:1488: DeprecationWarning: Overriding __eq
__ blocks inheritance of __hash__ in 3.x
  class IndexCol(StringMixin):
/Users/xin/anaconda/lib/python2.7/site-packages/pandas/io/pytables.py:1790: DeprecationWarning: Overriding __eq
__ blocks inheritance of __hash__ in 3.x
  class DataCol(IndexCol):
/Users/xin/anaconda/lib/python2.7/site-packages/pandas/io/stata.py:720: DeprecationWarning: Overriding __eq__ b
locks inheritance of __hash__ in 3.x
  class StataMissingValue(StringMixin):
[2018-10-21 11:58:03,798 - GDC - INFO] ====start checking====
[2018-10-21 11:58:46,772 - GDC - INFO] successful downloads
[2018-10-21 11:58:46,773 - GDC - INFO] ====check finished====
```

Check the successful download by check.py

Part1: Data download, integration and preprocess

https://gdc.cancer.gov/access-data/gdc-data-transfer-tool

Links to the binary distributions for supported platforr

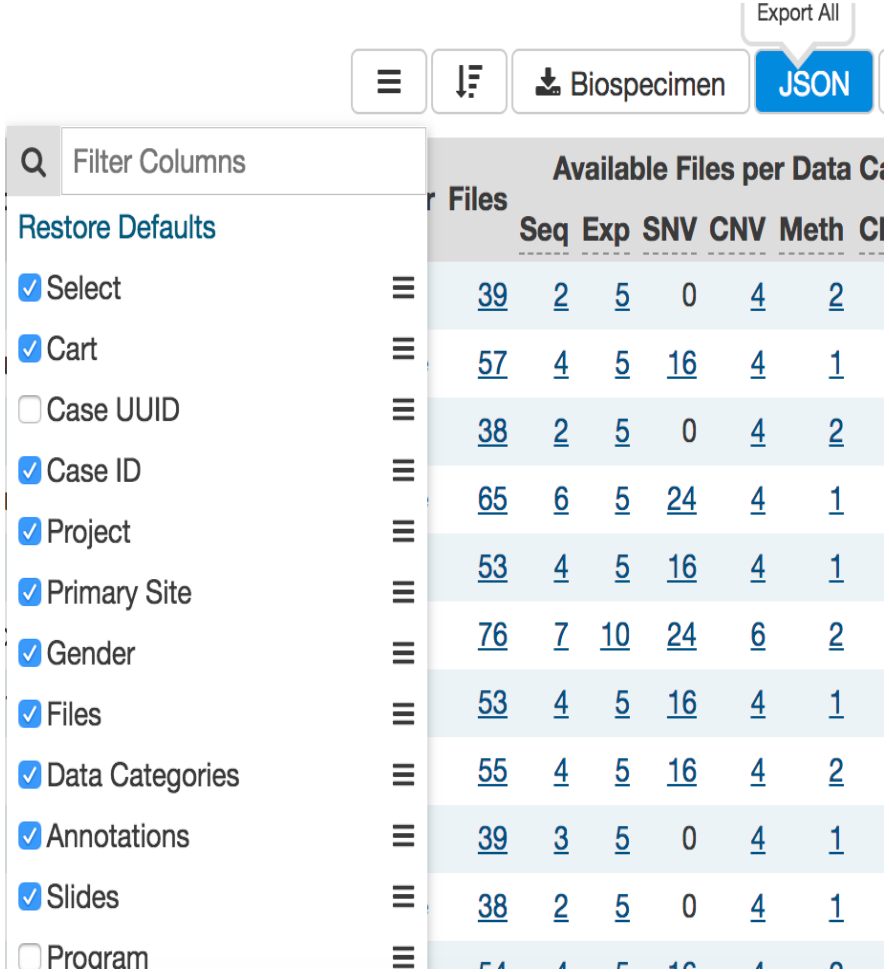
- gdc-client_v1.3.0_Windows_x64.zip
- gdc-client_v1.3.0_Ubuntu14.04_x64.zip
- gdc-client_v1.3.0_OSX_x64.zip

```
XINs-MacBook-Pro:miRNA xin$ ../../gdc-client download -m ../gdc_manifest.2018-10-21.txt
[
100% [##############################################################################] Time: 0:00:04   0.24  B/s
100% [##############################################################################] Time: 0:00:04   0.25  B/s
100% [##############################################################################] Time: 0:00:04   0.24  B/s
100% [##############################################################################] Time: 0:00:04   0.23  B/s
100% [##############################################################################] Time: 0:00:04   0.22  B/s
100% [##############################################################################] Time: 0:00:04   0.24  B/s
100% [##############################################################################] Time: 0:00:03   0.26  B/s
100% [##############################################################################] Time: 0:00:03   0.26  B/s

100% [##############################################################################] Time: 0:00:01  39.65 kB/s
100% [##############################################################################] Time: 0:00:00  58.05 kB/s
100% [##############################################################################] Time: 0:00:00  67.05 kB/s
100% [##############################################################################] Time: 0:00:00  62.62 kB/s
100% [##############################################################################] Time: 0:00:00  63.52 kB/s
100% [##############################################################################] Time: 0:00:00  72.58 kB/s
100% [##############################################################################] Time: 0:00:00  72.68 kB/s
100% [##############################################################################] Time: 0:00:00  84.89 kB/s
100% [##############################################################################] Time: 0:00:00  73.23 kB/s
100% [##############################################################################] Time: 0:00:00  70.54 kB/s
Successfully downloaded: 11486
```

Download GDC-data-transfer tool and download gdc_manifest file with the tool

Part1: Data download, integration and preprocess

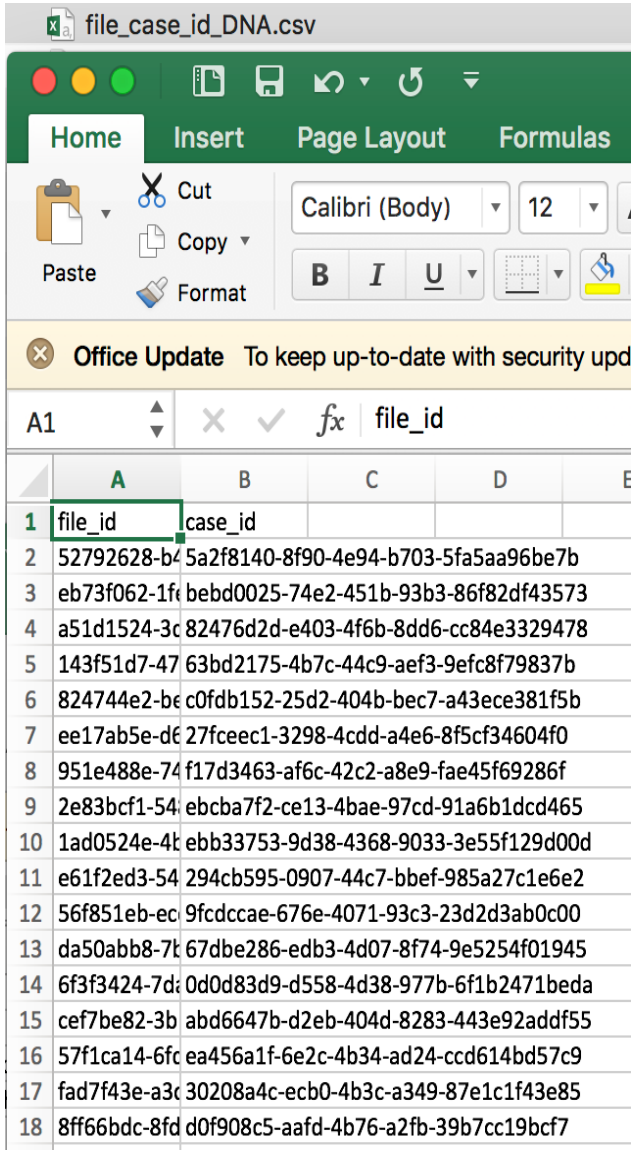


Click ≡ on the tab , and check all the following items, then click on the JSON tab to download the case ids for the files.

Part1: Data download, integration and preprocess

```
XINs-MacBook-Pro:miRNA xin$ python -3 ../../src/parse_file_case_id.py
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:16: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
  from . import multiarray
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/__init__.py:34: DeprecationWarning: CObject type is
not supported in 3.x. Please use capsule objects instead.
  from . import umath
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/core/_internal.py:204: DeprecationWarning: Overriding __e
q__ blocks inheritance of __hash__ in 3.x
  class dummy_ctype(object):
/Users/xin/anaconda/lib/python2.7/site-packages/numpy/lib/mixins.py:63: DeprecationWarning: Overriding __eq__ b
```

Parse file case id by parse_file_case_id.py



	A	B
1	file_id	case_id
2	52792628-b4	5a2f8140-8f90-4e94-b703-5fa5aa96be7b
3	eb73f062-1f	bebd0025-74e2-451b-93b3-86f82df43573
4	a51d1524-3c	82476d2d-e403-4f6b-8dd6-cc84e3329478
5	143f51d7-47	63bd2175-4b7c-44c9-aef3-9efc8f79837b
6	824744e2-be	c0fdb152-25d2-404b-bec7-a43ece381f5b
7	ee17ab5e-d6	27fceec1-3298-4cdd-a4e6-8f5cf34604f0
8	951e488e-74	f17d3463-af6c-42c2-a8e9-fae45f69286f
9	2e83bcf1-54	ebcba7f2-ce13-4bae-97cd-91a6b1dcd465
10	1ad0524e-4b	ebb33753-9d38-4368-9033-3e55f129d00d
11	e61f2ed3-54	294cb595-0907-44c7-bbef-985a27c1e6e2
12	56f851eb-ec	9fcdccae-676e-4071-93c3-23d2d3ab0c00
13	da50abb8-7b	67dbe286-edb3-4d07-8f74-9e5254f01945
14	6f3f3424-7d	0d0d83d9-d558-4d38-977b-6f1b2471beda
15	cef7be82-3b	abd6647b-d2eb-404d-8283-443e92addf55
16	57f1ca14-6f	ea456a1f-6e2c-4b34-ad24-ccd614bd57c9
17	fad7f43e-a3	30208a4c-ecb0-4b3c-a349-87e1c1f43e85
18	8ff66bdc-8fd	d0f908c5-aafd-4b76-a2fb-39b7cc19bcf7

Part1: Data download, integration and preprocess

```
[XINs-MacBook-Pro:miRNA xin$ python -3 ../../src/request_meta.py
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/bindings/openssl/binding.py:14: Deprecation
Warning: CObject type is not supported in 3.x. Please use capsule objects instead.
  from cryptography.hazmat.bindings._openssl import ffi, lib
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/primitives/asymmetric/dsa.py:150: Deprecati
onWarning: Overriding __eq__ blocks inheritance of __hash__ in 3.x
  class DSAParameterNumbers(object):
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/primitives/asymmetric/dsa.py:189: Deprecati
onWarning: Overriding __eq__ blocks inheritance of __hash__ in 3.x
  class DSAPublicNumbers(object):
/Users/xin/anaconda/lib/python2.7/site-packages/cryptography/hazmat/primitives/asymmetric/dsa.py:227: Deprecati
onWarning: Overriding __eq__ blocks inheritance of __hash__ in 3.x
  class DSAPrivateNumbers(object):
/Users/xin/anaconda/lib/python2.7/site-packages/OpenSSL/crypto.py:463: DeprecationWarning: Overriding __eq__ bl
ocks inheritance of __hash__ in 3.x
  class X509Name(object):
```

files_meta.tsv

sample_type

Done Replace

cases.0.samples.0.portions.0.analytes.0.aliquots.0.aliquot_id	data_type	cases.0.samples.0.sample_type	file_name	cases.0.samples.0.submitter_id	cases.0.samples.0.tissue_type	cases.0.samples.0.tumor_descriptor	file_id	data_category	cases.0.submitter_id	cases.0.samples.0.sample_id	cases.0.case_id	id	cases.0.samples.0.portions.0.analytes.0.aliquots.0.submitter_id
4aaff894-0a42-4b20-9290-759a34e6f248	miRNA Expression Quantification	Primary Tumor	9a2956fd-6985-4ed9-96a5-e04a3ed3acbb.mirbase21.mirnas.quantification.txt	TCGA-68-A59I-01A	0e22d215-92c2-435c-ad89-ee344ef88635	Transcriptome Profiling	TCGA-68-A59I	8856291a-fef5-4008-b6cc-636aa8975905	0f8a76a1-4c3e-42e1-9df3-95aa3ae8e676	0e22d215-92c2-435c-ad89-ee344ef88635	TCGA-68-A59I-01A-11H-A263-13		
f4a61521-e301-4b9a-8958-e483136011d1	miRNA Expression Quantification	Primary Tumor	95532a5a-df6a-438a-a26d-7c4a1c241c9f.mirbase21.mirnas.quantification.txt	TCGA-22-5479-01A	50efcf8f-1e01-4f71-b9e4-f8491eb3c736	Transcriptome Profiling	TCGA-22-5479	63157691-5ad8-4440-be22-c01a2614c9d0	fadd20a1-90ca-4d99-a128-9730d02a7bd6	50efcf8f-1e01-4f71-b9e4-f8491eb3c736	TCGA-22-5479-01A-31H-1948-13		
874c3ea2-4ea4-467d-8c31-49ec735bd4f2	miRNA Expression Quantification	Primary Tumor	8051f857-9875-427f-a757-4091db3c591a.mirbase21.mirnas.quantification.txt	TCGA-43-6771-01A	bb2acdaa-3a26-421a-a9e9-4ea3336a467b	Transcriptome Profiling	TCGA-43-6771	12b448e3-2d58-4e68-92e1-dd484742cd99	b0818bee-c32c-405c-ba09-956091fbe09b	bb2acdaa-3a26-421a-a9e9-4ea3336a467b	TCGA-43-6771-01A-11H-1819-13		
ad9ebe65-3dce-4b52-909c-99b07343d814	miRNA Expression Quantification	Primary Tumor	7ae1a03b-cca0-4ad6-92b8-096e247c1b50.mirbase21.mirnas.quantification.txt	TCGA-37-4141-01A	be8861c3-48b9-4cb7-a9a5-d4e18449d3ce	Transcriptome Profiling	TCGA-37-4141	bc1b8610-9e9d-4fbf-9698-82ceca1e935f	5d3d3918-7da8-4147-90a3-dc015be1dfbe	be8861c3-48b9-4cb7-a9a5-d4e18449d3ce	TCGA-37-4141-01A-02T-1557-13		
4eedeдac-03b6-4072-b518-b6760da4e656	miRNA Expression Quantification	Primary Tumor	37be9876-e459-48e9-b9ba-b47d02c17ae7.mirbase21.mirnas.quantification.txt	TCGA-NC-A5HH-01A	0435e383-6ec2-46c4-ba15-713a2447b0d1	Transcriptome Profiling	TCGA-NC-A5HH	3ac19224-4535-4f04-bb3d-7de9025542c2	b5865f1e-604d-487d-8a19-dc0c887203ed	0435e383-6ec2-46c4-ba15-713a2447b0d1	TCGA-NC-A5HH-01A-11H-A26V-13		
e15c3854-997c-4197-877f-99ac7695cb3a	miRNA Expression Quantification	Primary Tumor	69c2dd2f-9df4-4f79-9d18-577ad2d53dfd.mirbase21.mirnas.quantification.txt	TCGA-22-5478-01A	f7e9016d-2d0b-40d1-a330-09ae95684813	Transcriptome Profiling	TCGA-22-5478	9fc48d9d-cfb3-4c04-b918-878c05c57a35	4daf4a91-bc36-40c8-8fca-ea61b6706775	f7e9016d-2d0b-40d1-a330-09ae95684813	TCGA-22-5478-01A-01T-1634-13		
e824558d-7170-43a3-b483-e9fb02d9cf16	miRNA Expression Quantification	Primary Tumor	7df584d3-5b11-4987-8227-6d809e8305af.mirbase21.mirnas.quantification.txt	TCGA-22-1011-01A	07fa22d0-99b1-4364-9b93-d937005a2416	Transcriptome Profiling	TCGA-22-1011	ad8d4ef5-5c51-4342-983d-ef912fda745a	18058bfb-a008-410f-8e04-7c2b45a856b7	07fa22d0-99b1-4364-9b93-d937005a2416	TCGA-22-1011-01A-01T-1557-13		

cases_meta.tsv

primary_site

Done Replace

submitter_aliquot_ids.2	submitter_aliquot_ids.39	submitter_portion_ids.9	submitter_aliquot_ids.38	submitter_id	submitter_aliquot_ids.28	exposures.0.bmi	submitter_aliquot_ids.24	submitter_aliquot_ids.25	submitter_aliquot_ids.26	submitter_aliquot_ids.27	submitter_aliquot_ids.20	submitter_aliquot_ids.21	submitter_aliquot_ids.22
submitter_aliquot_ids.23	slide_ids.9	demographic.year_of_birth	demographic.submitter_id	diagnoses.0.classification_of_tumor	diagnoses.0.progression_or_recurrence	case_id	diagnoses.0.state	submitter_aliquot_ids.30	aliquot_ids.47	diagnoses.0.days_to_last_follow_up	exposures.0.height	demographic.ethnicity	aliquot_ids.17
aliquot_ids.16	aliquot_ids.15	aliquot_ids.14	aliquot_ids.13	diagnoses.0.tumor_grade	aliquot_ids.11	aliquot_ids.10	submitter_aliquot_ids.37	submitter_aliquot_ids.36	submitter_aliquot_ids.35	submitter_aliquot_ids.34	submitter_aliquot_ids.33	submitter_aliquot_ids.32	submitter_aliquot_ids.31
aliquot_ids.18	disease_type	demographic.state	slide_ids.10	submitter_analyte_ids.12	submitter_analyte_ids.13	submitter_analyte_ids.10	submitter_analyte_ids.11	submitter_analyte_ids.14	exposures.0.state	submitter_aliquot_ids.29	sample_ids.5	exposures.0.submitter_id	updated_datetime
submitter_slide_ids.9	submitter_slide_ids.8	submitter_slide_ids.3	submitter_slide_ids.2	submitter_slide_ids.1	submitter_slide_ids.0	submitter_slide_ids.7	submitter_slide_ids.6	submitter_slide_ids.5	submitter_slide_ids.4	exposures.0.alcohol_history	submitter_portion_ids.2	created_datetime	demographic.gender
state	demographic.race	analyte_ids.10	analyte_ids.11	analyte_ids.12	analyte_ids.13	analyte_ids.14	diagnoses.0.created_datetime	submitter_aliquot_ids.15	exposures.0.exposure_id	portion_ids.3	portion_ids.2	submitter_aliquot_ids.11	portion_ids.4
submitter_aliquot_ids.13	analyte_ids.9	analyte_ids.6	analyte_ids.7	analyte_ids.4	analyte_ids.5	analyte_ids.2	analyte_ids.3	analyte_ids.0	analyte_ids.1	portion_ids.1	portion_ids.0	portion_ids.9	submitter_aliquot_ids.17
diagnoses.0.prior_malignancy	analyte_ids.8	submitter_aliquot_ids.16	portion_ids.5	submitter_aliquot_ids.10	submitter_aliquot_ids.46	portion_ids.7	submitter_aliquot_ids.47	aliquot_ids.44	aliquot_ids.45	aliquot_ids.46	portion_ids.6	aliquot_ids.40	aliquot_ids.41
aliquot_ids.42	aliquot_ids.43	exposures.0.updated_datetime	submitter_aliquot_ids.6	submitter_portion_ids.0	portion_ids.8	aliquot_ids.2	diagnoses.0.age_at_diagnosis	diagnoses.0.vital_status	submitter_aliquot_ids.12	submitter_aliquot_ids.19			

Get Meta data by request_meta.py

Part1: Data download, integration and preprocess

```
n_labels = df['label'].value_counts()
print(n_labels)
print(df.shape)
```

```
2.0     1104
4.0     1081
22.0    1000
0.0      691
13.0     550
6.0      532
23.0     515
29.0     500
5.0      499
1.0      459
20.0     452
12.0     450
17.0     418
33.0     408
16.0     309
14.0     301
7.0      231
3.0      185
21.0     179
35.0     157
11.0     131
34.0     129
28.0     119
18.0     118
19.0     117
27.0     103
```

```
labels = {0: 0, 'Colon': 1, 'Breast': 2, 'Esophagus': 3, nan: 60, 'Kidney': 4, 'Prostate gland': 5,
          'Brain': 6, 'Adrenal gland': 7, 'Thymus': 8, 'Lymph nodes': 9,
          'Base of tongue': 10, 'Other and unspecified parts of tongue': 11, 'Skin': 12,
          'Corpus uteri': 13, 'Blood': 14, '0': 15, 'Cervix uteri': 16, 'Bladder': 17,
          'Retroperitoneum and peritoneum': 18, 'Larynx': 19, 'Stomach': 20, 'Pancreas': 21,
          'Bronchus and lung': 22, 'Thyroid gland': 23, 'Tonsil': 24, 'Rectum': 25,
          'Hypopharynx': 26, 'Hematopoietic and reticuloendothelial systems': 27,
          'Heart, mediastinum, and pleura': 28, 'Ovary': 29, 'Uterus, NOS': 30,
          'Other and ill-defined sites in lip, oral cavity and pharynx': 31,
          'Eye and adnexa': 32, 'Liver and intrahepatic bile ducts': 33,
          'Connective, subcutaneous and other soft tissues': 34, 'Testis': 35,
          'Rectosigmoid junction': 36, 'Other and unspecified parts of mouth': 37,
          'Other and ill-defined sites': 38, 'Floor of mouth': 39,
          'Other endocrine glands and related structures': 40, 'Lip': 41,
          'Small intestine': 42, 'Gum': 43, 'Oropharynx': 44,
          'Peripheral nerves and autonomic nervous system': 45,
          'Bones, joints and articular cartilage of other and unspecified sites': 46,
          'Unknown primary site': 47, 'Palate': 48,
          'Bones, joints and articular cartilage of limbs': 49,
          'Other and unspecified parts of biliary tract': 50, 'Meninges': 51,
          'Gallbladder': 52,
          'Spinal cord, cranial nerves, and other parts of central nervous system': 53,
          'Other and unspecified male genital organs': 54}
```

- Generate the miRNA matrix for all the files with labeled normal or tumor.

The miRNA seq that comes from totally so many different cancers based on “primary site” in cases meta

- Remove some primary site type which counts are less than 100, and then only 26 primary site left
- Normal tissue is labeled with 0, and other primary sites labeled the different values based on dictionary in gen_miRNA_matrix.py
- Finally, there still have over 10,000 samples

Part2: Prediction using KNN

```
data_dir ="/Users/xin/Downloads/EE542/Labs/lab10/data/"
data_file = data_dir + "miRNA_matrix.csv"
df = pd.read_csv(data_file)
```

sa-mir-941-	hsa-mir-941-	hsa-mir-941-	hsa-mir-942	hsa-mir-943	hsa-mir-944	hsa-mir-95	hsa-mir-9500	hsa-mir-96	hsa-mir-98	hsa-mir-99a	hsa-mir-99b	label
0	0	0	13	0	0	4	0	1	84	762	75212	1
0	0	0	13	0	2	94	0	79	530	12392	233865	2
0	0	0	101	0	2078	44	0	92	338	782	88949	4
0	0	0	72	0	3	18	0	83	374	9898	153229	5
0	0	0	24	0	0	5	0	2	24	461	871177	6
0	0	0	3	0	1	1	0	27	114	395	37435	7
0	0	0	35	1	600	5	0	80	97	490	12794	6
0	0	0	18	0	0	11	0	240	211	885	72173	12
0	0	0	43	2	0	36	0	157	185	2228	146318	13
0	0	0	60	0	975	88	0	182	174	4536	43652	14
0	0	0	10	0	38	8	0	11	92	3055	32007	0
0	0	0	4	0	0	1	0	94	133	3079	52183	16
0	0	0	91	1	310	3	0	150	133	442	29625	7
0	0	0	15	0	2	6	0	21	114	280	42735	6
0	0	0	42	0	837	30	0	56	253	726	48645	4
0	0	0	8	0	0	2	0	5	256	3483	261715	17

```
# split the data to train and test set
X_train, X_test, y_train, y_test = train_test_split(X_data, y_data, test_size=0.3, random_state=0)

#standardize the data.
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

- Read data from miRNA matrix
- Scatter data to train set and test set 7:3

Part2: Prediction using KNN

```
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
num_features=800
pca = PCA(n_components=num_features, whiten=True)
pca.fit(X_train)
X_train = pd.DataFrame(pca.transform(X_train))
X_test = pd.DataFrame(pca.transform(X_test))
print("PCA reduction")
print("feature numbers: "+str(num_features))
#print("explained_variance_ratio_: \n"+str(pca.explained_variance_ratio_))
print("explained_variance_ratio_.cumsum: \n"+str(pca.explained_variance_ratio_.cumsum()))
```

```
PCA reduction
feature numbers: 800
explained_variance_ratio_.cumsum:
[0.08024552 0.11213829 0.14088122 0.16403317 0.18027623 0.19439016
 0.20674242 0.21807625 0.22671984 0.234468   0.2417881  0.24870134
 0.25542148 0.26137246 0.26716042 0.27258638 0.27793182 0.28295968
 0.28767314 0.2920368  0.29624207 0.30035979 0.30441073 0.30826549
 0.31186178 0.31539722 0.3188568  0.32218002 0.32537988 0.32846721
 0.3314421  0.3342966  0.33710044 0.33986321 0.34253281 0.34517575
 0.34776262 0.35033472 0.35283313 0.35530327 0.35770834 0.36010286
 0.3624359  0.36468992 0.366889   0.36907654 0.37121741 0.37329379
 0.37536765 0.37741505 0.37944168 0.38144251 0.38343012 0.38540043
```

- Feature selection using PCA and t-SNE
- Visualize the result – we can see that reducing dimension from 1883 to 800 has already been able to keep 89% information
- We can not use the higher dimension because our sample number is 8788 after data processing

```
  ......

 0.87386774 0.87418761 0.87450662 0.87482541 0.87514312 0.87546015
 0.87577579 0.87609111 0.87640541 0.87671878 0.87703196 0.87734436
 0.87765576 0.87796593 0.87827549 0.87858431 0.87889263 0.8792
 0.87950651 0.87981199 0.88011685 0.88042108 0.88072333 0.88102454
 0.88132522 0.88162546 0.88192437 0.8822216  0.88251842 0.88281364
 0.88310788 0.88340136 0.8836945  0.88398741 0.88427956 0.88457093
 0.88486063 0.88515005 0.88543808 0.88572572 0.88601299 0.88629756
 0.88658203 0.8868655  0.88714781 0.88742977 0.88771032 0.88798906
 0.88826589 0.88854103 0.88881472 0.88908791 0.88936033 0.88963194
 0.88990166 0.89017089]
```

Part2: Prediction using KNN

```
print("t-SNE evaluation")
tsne = TSNE(n_components=2, verbose=1, perplexity=40, n_iter=250)
new = tsne.fit_transform(X_train)
features = pd.DataFrame(new)
```

```
t-SNE evaluation
[t-SNE] Computing 121 nearest neighbors...
[t-SNE] Indexed 6151 samples in 0.006s...
[t-SNE] Computed neighbors for 6151 samples in 0.869s...
[t-SNE] Computed conditional probabilities for sample 1000 / 6151
[t-SNE] Computed conditional probabilities for sample 2000 / 6151
[t-SNE] Computed conditional probabilities for sample 3000 / 6151
[t-SNE] Computed conditional probabilities for sample 4000 / 6151
[t-SNE] Computed conditional probabilities for sample 5000 / 6151
[t-SNE] Computed conditional probabilities for sample 6000 / 6151
[t-SNE] Computed conditional probabilities for sample 6151 / 6151
[t-SNE] Mean sigma: 0.090937
```

```
print(features)
```

	0	1
0	-0.972481	-0.445210
1	0.697620	-2.116041
2	-0.782807	-0.510700
3	0.611465	2.169794
4	-0.048844	2.251833
5	-0.007524	-1.655136
6	-0.291251	-0.689894
7	0.017664	-1.646448
8	-0.730924	-0.151445
9	-0.947784	1.730283
10	-1.055320	-2.054043
11	-1.086948	0.247150
12	0.656100	2.440754
13	-0.108890	-3.028151
14	-0.472456	-0.835887
15	-0.244564	-0.809207
16	0.128819	1.124375
17	0.328129	1.508422
18	-0.805383	-2.538891
19	-0.571301	-4.143396
20	-1.211597	0.818011
21	0.162854	-1.594920
22	-0.197186	-2.877308
23	1.445303	-0.139750
24	0.598879	2.575904
25	-0.526559	0.121936
26	-0.633962	0.442991
27	1.095082	-1.749704
28	0.636407	1.886242
29	-0.506113	-4.005802
...	...	...

- Feature selection using PCA and t-SNE
- Visualize the result – we can see that reducing dimension from 1883 to 800 has already been able to keep 89% information

Part2: Prediction using KNN

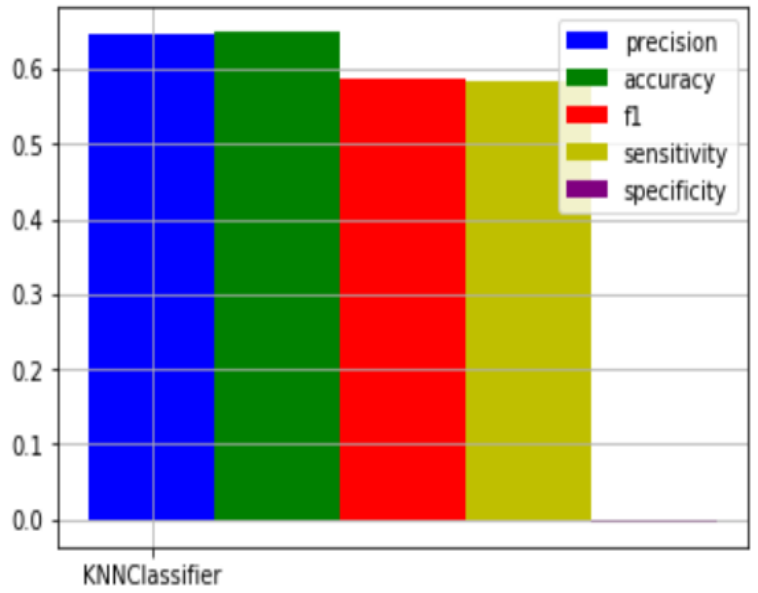
```
models = {
    'KNeighborsClassifier': KNeighborsClassifier(),
}
tuned_parameters = {
    'KNeighborsClassifier': {'n_neighbors': [3, 5, 7, 10]},
}
```

```
clf = GridSearchCV(models[key], tuned_parameters[key], scoring=None,  refit=True, cv=10)
clf.fit(X_train,y_train)
y_test_predict = clf.predict(X_test)
precision = precision_score(y_test, y_test_predict, average='micro')
accuracy = accuracy_score(y_test, y_test_predict)
f1 = f1_score(y_test, y_test_predict, average='micro')
recall = recall_score(y_test, y_test_predict, average='micro')
specificity = specificity_score(y_test, y_test_predict)
scores[key] = [precision,accuracy,f1,recall,specificity]
# reture the best model which K hyper-parameter can get the highest performance
best_model = clf.best_estimator_
```

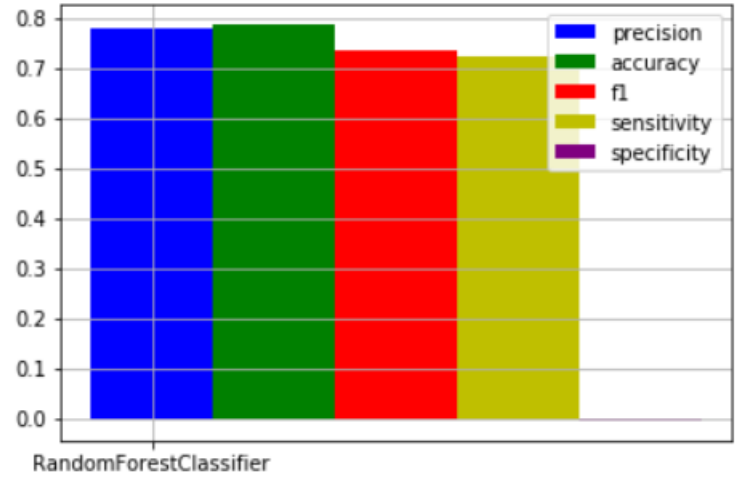
- Hyper-parameter tuning – we select K = {3, 5,7,10} to try to get the best K using grid search CV
- Score average set ‘micro’ for multiclass, which will return the total ratio

Part2: Prediction using KNN

['KNNClassifier']



['RandomForestClassifier']




```
In [6]: print(scores)
```

```
{'KNNClassifier': [0.6448927653055788, 0.6485780615206036, 0.5867057788654338, 0.584273255563632, -0.0036629130058161118]}
```

- Predict and output evaluation metrics and ROC curve

https://github.com/appleshine77/GDC-data-lab.git