BRCA data exploration

Urminder Singh

May 2, 2019

BRCA data exploration

This is a short lesson on how to explore the mutation data at TCGA using TCGABiolinks and maftools libraries. Using study metadata is crucial to the analysis of TCGA datasets. I have provided some functions which user can use to format TCGA metadata and easily use that in analysis. I will particulary look at BRCA data at TCGA but the methods and functions provided here could easily be extended to any cancer type.

Downloading TCGA Metadata

For all the data deposited in TCGA, there is associated clinical metadata. Clinical metadata could be critical in understanding the data from different perspectives. For example, the clinical data has information about tumor stage, gender, age etc. all of which could be helpful in data analysis from various perspectives.

TCGAbiolinks provides the function "*GDCquery_clinic*" to download the clinical metadata. The clinical metadata is separated into two categories i.e. Clinical and Biospecimen. More information on this is available at https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/clinical.html

A simple example to use *GDCquery_clinic* function is shown below:

```
clinicalBRCA <- GDCquery_clinic(project = "TCGA-BRCA", type = "clinical")
biospecimenBRCA <- GDCquery_clinic(project = "TCGA-BRCA", type = "Biospecimen")

print(head(clinicalBRCA))
```

submitter_id classification_of_tumor last_known_disease_status 1 TCGA-3C-AAAU not reported not reported 2 TCGA-3C-AALI not reported not reported 3 TCGA-3C-AALJ not reported not reported 4 TCGA-3C-AALK not reported not reported 5 TCGA-4H-AAAK not reported not reported 6 TCGA-5L-AAT0 not reported not reported updated_datetime primary_diagnosis 1 2018-09-06T13:49:20.245333-05:00 Lobular carcinoma, NOS 2 2018-09-06T13:49:20.245333-05:00 Infiltrating duct carcinoma, NOS 3 2018-09-06T13:49:20.245333-05:00 Infiltrating duct carcinoma, NOS 4 2018-09-06T13:49:20.245333-05:00 Infiltrating duct carcinoma, NOS 5 2018-09-06T13:49:20.245333-05:00 Lobular carcinoma, NOS 6 2018-09-06T13:49:20.245333-05:00 Lobular carcinoma, NOS tumor_stage age_at_diagnosis vital_status morphology days_to_death 1 stage x 20211 alive 8520/3 NA 2 stage iib 18538 alive 8500/3 NA 3 stage iib 22848 alive 8500/3 NA 4 stage ia 19074 alive 8500/3 NA 5 stage iiia 18371 alive 8520/3 NA 6 stage iia 15393 alive 8520/3 NA days_to_last_known_disease_status created_datetime state 1 NA NA released 2 NA NA released 3 NA NA released 4 NA NA released 5 NA NA released 6 NA NA released days_to_recurrence diagnosis_id tumor_grade 1 NA 8cfb8afb-b915-5255-865b-a5923f47b351 not reported 2 NA 8cafc022-585f-54a1-a7d4-cfa632b3991e not reported 3 NA 63d85b81-8eba-5f17-8552-92babf137c00 not reported 4 NA 8c90b19d-54f7-5788-a5eb-49abe239ef0b not reported 5 NA 81c406cd-ad2d-552f-b543-8b2b03044686 not reported 6 NA ebff6a7b-3b6c-5f71-86b0-1bdd3b78edd4 not reported tissue_or_organ_of_origin days_to_birth progression_or_recurrence 1 Breast, NOS -20211 not reported 2 Breast, NOS -18538 not reported 3 Breast, NOS -22848 not reported 4 Breast, NOS -19074 not reported 5 Breast, NOS -18371 not reported 6 Breast, NOS -15393 not reported prior_malignancy site_of_resection_or_biopsy days_to_last_follow_up 1 not reported Breast, NOS 4047 2 not reported Breast, NOS 4005 3 not reported Breast, NOS 1474 4 not reported Breast, NOS 1448 5 not reported Breast, NOS 348 6 not reported Breast, NOS 1477 cigarettes_per_day weight alcohol_history alcohol_intensity bmi 1 NA NA NA NA NA 2 NA NA NA NA NA 3 NA NA NA NA NA 4 NA NA NA NA NA

5 NA NA NA NA NA 6 NA NA NA NA NA years_smoked exposure_id height gender 1 NA 72f0be98-dffa-5d35-88fe-f9ca774d6db0 NA female 2 NA 63b59ac3-ccc2-5590-bff7-673f15713369 NA female 3 NA 05defab8-b347-540a-8950-8a180faeb67e NA female 4 NA a97db788-0772-5d32-878c-d2080d979c37 NA female 5 NA e80a24a3-d0fc-5067-a47d-83ac937af2f0 NA female 6 NA 47c9213f-b2b8-5297-a0b6-21bce2cfa3f8 NA female year_of_birth race 1 1949 white 2 1953 black or african american 3 1949 black or african american 4 1959 black or african american 5 1963 white 6 1968 white demographic_id ethnicity 1 cee0a94c-1d9e-5650-a500-a6b021fe138d not hispanic or latino 2 583a1ee5-3175-523e-ba1f-75a30a3e1e41 not hispanic or latino 3 9619a908-1684-547f-a407-6adf93e15b8d not hispanic or latino 4 e54b1469-fffc-5291-a8e3-df2092ab5f34 not hispanic or latino 5 64a204f0-b380-5962-a927-84af1841b6d6 not hispanic or latino 6 776bb6f9-f5c8-57b6-bf4d-4014b8025a06 hispanic or latino year_of_death treatment_id therapeutic_agents 1 NA 2d88df62-dc75-5c01-b249-3b914cd7380a NA 2 NA 89c2d475-7048-52d3-8dc0-1425330d35ee NA 3 NA a1ecb0cf-fa4a-58df-8c6d-c1baaad53f7e NA 4 NA 35fe07bb-5a79-5549-9a08-a851d9aa3de1 NA 5 NA 68417a03-5e66-535d-ac9a-fa6ffb98b571 NA 6 NA 38f1dbba-d771-539b-91e7-5b9761c0f592 NA treatment_intent_type treatment_or_therapy bcr_patient_barcode disease 1 NA NA TCGA-3C-AAAU BRCA 2 NA NA TCGA-3C-AALI BRCA 3 NA NA TCGA-3C-AALJ BRCA 4 NA NA TCGA-3C-AALK BRCA 5 NA NA TCGA-4H-AAAK BRCA 6 NA NA TCGA-5L-AAT0 BRCA

```
head(biospecimenBRCA)
```

sample_type_id updated_datetime 1 01 2018-11-15T21:38:54.195821-06:00 2 11 2018-11-15T21:38:54.195821-06:00 3 01 2018-11-15T21:10:03.529893-06:00 4 01 2018-11-15T21:38:54.195821-06:00 5 10 2018-11-15T21:38:54.195821-06:00 6 01 2018-11-15T21:10:03.529893-06:00 time_between_excision_and_freezing oct_embedded tumor_code_id 1 NA true NA 2 NA true NA 3 NA No NA 4 NA true NA 5 NA false NA 6 NA No NA submitter_id intermediate_dimension 1 TCGA-BH-A0C3-01A NA 2 TCGA-BH-A0C3-11A NA 3 TCGA-BH-A0C3-01Z NA 4 TCGA-BH-A0HQ-01A NA 5 TCGA-BH-A0HQ-10A NA 6 TCGA-BH-A0HQ-01Z NA sample_id is_ffpe 1 21ba28ff-89f6-4f02-a135-821efc4f42f8 FALSE 2 82e6dc7b-fe63-4fc6-af9f-68dc76c2cb88 FALSE 3 bb18ed04-1c02-41f5-af7a-d82980d185f3 TRUE 4 ef48e806-e31c-4a11-afe8-dcc232357329 FALSE 5 c14eb1f6-791e-491f-8b60-f9856daf77b8 FALSE 6 6b4f016a-8a55-49fd-9331-855a5fde317e TRUE pathology_report_uuid created_datetime 1 3A54CF6E-AFDB-4609-A827-77D75BB376A7 2 3 2018-05-17T12:14:28.274820-05:00 4 A76A272F-675E-4E56-8761-96B71419A012 5 6 2018-05-17T12:12:29.643720-05:00 tumor_descriptor sample_type state current_weight 1 NA Primary Tumor released NA 2 NA Solid Tissue Normal released NA 3 NA Primary Tumor released NA 4 NA Primary Tumor released NA 5 NA Blood Derived Normal released NA 6 NA Primary Tumor released NA composition time_between_clamping_and_freezing shortest_dimension 1 NA NA NA 2 NA NA NA 3 NA NA NA 4 NA NA NA 5 NA NA NA 6 NA NA NA tumor_code tissue_type days_to_sample_procurement freezing_method 1 NA Not Reported NA NA 2 NA Not Reported NA NA 3 NA Not Reported 0 NA 4 NA Not Reported NA NA 5 NA Not Reported NA NA 6 NA Not Reported 0 NA portions 1 1289952000, 1300752000, 21, 11, 30, NA, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, NA, NA, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, NA, NA, NA, NA, NA, NA, Repli-G (Qiagen) DNA, DNA, RNA, TCGA-BH-A0C3-01A-21W, TCGA-BH-A0C3-01A-21D, TCGA-BH-A0C3-01A-21R, NA, 2.29, 1.85, NA, NA, NA, released, released, released, released, 2018-11-27T09:46:29.784386-06:00, NA, NA, NA, TCGA-BH-A0C3-01A-21W-A14O-09, NA, c069b0fa-e30b-4d26-92ab-63e9e670da3b, 0.5, 23, released, released, released, released, 2018-11-27T09:46:29.784386-06:00, 2018-11-27T09:46:29.784386-06:00, 2018-11-27T09:46:29.784386-06:00, 2018-11-27T09:46:29.784386-06:00, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, TCGA-BH-A0C3-01A-21D-A12N-01, TCGA-BH-A0C3-01A-21D-A12R-05, TCGA-BH-A0C3-01A-21D-A12Q-09, TCGA-BH-A0C3-01A-21D-A12M-02, NA, NA, NA, NA, 40aac58d-9d3d-4cc3-ab4b-c02b28b8fc5f, fee0245d-1292-4a5f-88b4-79606996a536, ec57ee0f-949e-4eee-91c2-dd129d657065, d3b5b2f1-3a99-4245-afad-40047c4c44ae, 0.16, 0.16, 0.08, 0.16, 23, 23, 23, 23, released, released, 2018-11-27T09:46:29.784386-06:00, 2018-11-27T09:46:29.784386-06:00, NA, NA, NA, NA, NA, NA, TCGA-BH-A0C3-01A-21R-A12O-13, TCGA-BH-A0C3-01A-21R-A12P-07, NA, NA, a150022e-d6a8-4d2d-8d21-57b85790c180, 52a863fe-b210-456b-9c2d-545277362b65, 0.16, 0.16, 23, 23, b39eba23-b29f-4b51-8de2-43ae2c2a440b, 058979d0-fc17-4a31-959a-2a7c47bad097, 1fcee90a-0d2c-4011-96c2-c6898e3088b0, NA, 0.16, 0.16, NA, UV Spec, UV Spec, W, D, R, TCGA-BH-A0C3-01A-21, TCGA-BH-A0C3-01A-11-A13C-20, released, released, 3598b43b-ac89-4800-b6bf-0390680b35ff, 42fa3207-2c4f-4499-99b8-fe5275232eed, FALSE, FALSE 2 1292371200, 23, 130, 2018-09-06T13:49:20.245333-05:00, NA, 2018-09-06T13:49:20.245333-05:00, 2018-09-

06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, NA, NA, NA, NA, NA, NA, Repli-G (Qiagen) DNA, RNA, DNA, TCGA-BH-A0C3-11A-23W, TCGA-BH-A0C3-11A-23R, TCGA-BH-A0C3-11A-23D, NA, 1.83, 1.98, NA, NA, NA, released, released, released, released, 2018-11-27T11:13:44.305228-06:00, NA, NA, NA, TCGA-BH-A0C3-11A-23W-A14O-09, NA, 809ff824-d6a5-42cb-a28f-ddb019ecad1b, 0.5, 23, released, released, 2018-11-27T11:13:44.305228-06:00, 2018-11-27T11:13:44.305228-06:00, NA, NA, NA, NA, NA, NA, TCGA-BH-A0C3-11A-23R-A12P-07, TCGA-BH-A0C3-11A-23R-A12O-13, NA, NA, 65923c68-2c86-4b4d-abe2-5320c08bb68f, fe8c1b0d-f7b8-4e58-a728-55d367798adf, 0.17, 0.17, 23, 23, released, released, released, released, 2018-11-27T11:13:44.305228-06:00, 2018-11-27T11:13:44.305228-06:00, 2018-11-27T11:13:44.305228-06:00, 2018-11-27T11:13:44.305228-06:00, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, TCGA-BH-A0C3-11A-23D-A12R-05, TCGA-BH-A0C3-11A-23D-A12Q-09, TCGA-BH-A0C3-11A-23D-A12M-02, TCGA-BH-A0C3-11A-23D-A12N-01, NA, NA, NA, NA, 3669fca6-4284-488e-9c6c-4d96885fac9c, 128e802b-64d9-42d9-8a22-59c8ce061eb4, f0f6753c-220c-41da-b82b-4f18d0920547, fccfb1a9-bb7b-47fc-aee6-9512cea51b43, 0.16, 0.08, 0.16, 0.16, 23, 23, 23, 23, 1c136836-e2b2-4510-8d29-a8b369cd617a, b1d4ea1a-2d31-4d8f-a68a-9de922b0e7ff, cb8b0bad-b60e-4c91-b7d0-33f8f90d22ff, NA, 0.17, 0.16, NA, UV Spec, UV Spec, W, R, D, TCGA-BH-A0C3-11A-23, released, 9c72579e-0c59-49fd-9df1-a9aa66e3f3b2, FALSE 3 ed0b7894-e22d-432e-a521-1a86254d6214 4 1277769600, 11, 30, 2018-09-06T13:49:20.245333-05:00, NA, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, NA, NA, NA, NA, NA, NA, NA, NA, Repli-G (Qiagen) DNA, RNA, Repli-G X (Qiagen) DNA, DNA, TCGA-BH-A0HQ-01A-11W, TCGA-BH-A0HQ-01A-11R, TCGA-BH-A0HQ-01A-11X, TCGA-BH-A0HQ-01A-11D, NA, 1.76, NA, 1.85, NA, NA, NA, NA, released, released, released, released, released, released, 2018-11-27T12:40:03.513897-06:00, 2018-11-27T12:40:03.513897-06:00, NA, NA, NA, NA, NA, NA, TCGA-BH-A0HQ-01A-11W-A051-08, TCGA-BH-A0HQ-01A-11W-A050-09, NA, NA, 13e0b1a5-6834-4152-a792-20747d3c9655, f03af67f-3119-4ee4-a4b0-227d36f493ba, 0.5, 0.5, 23, 23, released, released, 2018-11-27T12:40:03.513897-06:00, 2018-11-27T12:40:03.513897-06:00, NA, NA, NA, NA, NA, NA, TCGA-BH-A0HQ-01A-11R-A035-13, TCGA-BH-A0HQ-01A-11R-A034-07, NA, NA, 5a604ff4-28d4-4863-bb27-06d38eb63a0f, 856452d1-5480-4921-bf5b-4c6f625a99c0, 0.14, 0.14, 23, 23, released, released, 2018-11-27T12:40:03.513897-06:00, 2018-11-27T12:40:03.513897-06:00, NA, NA, NA, NA, NA, NA, TCGA-BH-A0HQ-01A-11X-A049-09, TCGA-BH-A0HQ-01A-11X-A048-08, NA, NA, 00a31122-f2d5-4d09-b165-1231703d8dc2, 4189472a-47ff-414e-8695-b888f21f69ba, 0.5, 0.5, 23, 23, released, released, released, released, released, 2018-11-27T12:40:03.513897-06:00, 2018-11-27T12:40:03.513897-06:00, 2018-11-27T12:40:03.513897-06:00, 2018-11-27T12:40:03.513897-06:00, 2018-11-27T12:40:03.513897-06:00, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, TCGA-BH-A0HQ-01A-11D-A044-08, TCGA-BH-A0HQ-01A-11D-A033-02, TCGA-BH-A0HQ-01A-11D-A036-01, TCGA-BH-A0HQ-01A-11D-A032-05, TCGA-BH-A0HQ-01A-11D-A045-09, NA, NA, NA, NA, NA, 9c4da4d7-a1b3-4c95-a808-6f8eceaf3ecd, c888dc10-e062-4e97-8dff-1ecd04dae66f, e84b732c-f8a5-4f4c-933e-ea539fff5d48, 247b89d7-05c2-49ca-8f96-08786b03a511, 33d16efa-5e83-4e18-93f2-be2fc32ccbac, 0.07, 0.15, 0.15, 0.15, 0.07, 23, 23, 23, 23, 23, bf1c746c-fb30-4cb5-9795-a42b8fae3c23, 69530b24-2ad6-49c6-bb16-65da83137428, f963acc8-7df0-44ec-8927-29fe4184bd0e, e8fcb6eb-d4d0-4be2-82b1-f136e1e7cfe5, NA, 0.14, NA, 0.15, NA, UV Spec, NA, UV Spec, W, R, X, D, TCGA-BH-A0HQ-01A-11, released, a379bdff-7c8d-4dea-934d-a84211baf11c, FALSE 5 1278547200, 01, 200, 2018-09-06T13:49:20.245333-05:00, NA, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, 2018-09-06T13:49:20.245333-05:00, NA, NA, NA, NA, NA, NA, DNA, Repli-G X (Qiagen) DNA, Repli-G (Qiagen) DNA, TCGA-BH-A0HQ-10A-01D, TCGA-BH-A0HQ-10A-01X, TCGA-BH-A0HQ-10A-01W, 2.03, NA, NA, NA, NA, NA, released, released, released, released, released, released, released, 2018-11-27T12:11:57.452067-06:00, 2018-11-27T12:11:57.452067-06:00, 2018-11-27T12:11:57.452067-06:00, 2018-11-27T12:11:57.452067-06:00, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, TCGA-BH-A0HQ-10A-01D-A037-01, TCGA-BH-A0HQ-10A-01D-A046-08, TCGA-BH-A0HQ-10A-01D-A047-09, TCGA-BH-A0HQ-10A-01D-A031-02, NA, NA, NA, NA, e417b9bb-683e-40fe-b2e7-0829a2644cc3, bbd634d7-6b6c-4dde-aac8-82f68bb6a7f5, 20e5b563-bec2-436b-b9b1-3be85a249dec, 843e9ea6-203f-43fc-8ce0-a4f4f5265ebe, 0.16, 0.08, 0.08, 0.16, 23, 23, 23, 23, released, released, 2018-11-27T12:11:57.452067-06:00, 2018-11-27T12:11:57.452067-06:00, NA, NA, NA, NA, NA, NA, TCGA-BH-A0HQ-10A-01X-A054-09, TCGA-BH-A0HQ-10A-01X-A052-08, NA, NA, d2cb9c50-daee-48b0-93e0-38754c88c0a6, 2ebb7129-2d21-4583-944b-c4efb12124c5, 0.5, 0.5, 23, 23, released, released, 2018-11-27T12:11:57.452067-06:00, 2018-11-27T12:11:57.452067-06:00, NA, NA, NA, NA, NA, NA, TCGA-BH-A0HQ-10A-01W-A055-09, TCGA-BH-A0HQ-10A-01W-A053-08, NA, NA, 006ff264-ef7d-4021-a879-77080c0440cf, c5632337-1721-4f35-b1f7-be3fd4bc2a28, 0.5, 0.5, 23, 23, dbf00f52-a117-4ab4-a901-c473c59afc15, 600f0b6e-c3ea-428c-94b7-264f8dcf3ea0, f1df1493-c229-406c-b111-83cd80582ca6, 0.16, NA, NA, UV Spec, NA, NA, D, X, W, TCGA-BH-A0HQ-10A-01, released, 9b93040b-a786-4cdd-aa3f-52920e7624c3, FALSE 6 093f0a51-a8ef-4ea0-9aad-758e867ad222 preservation_method days_to_collection initial_weight longest_dimension 1 1335 160 NA 2 1335 200 NA 3 FFPE NA NA NA 4 962 110 NA 5 962 NA NA 6 FFPE NA NA NA distance_normal_to_tumor biospecimen_anatomic_site 1 NA NA 2 NA NA 3 NA NA 4 NA NA 5 NA NA 6 NA NA diagnosis_pathologically_confirmed distributor_reference 1 NA NA 2 NA NA 3 NA NA 4 NA NA 5 NA NA 6 NA NA method_of_sample_procurement passage_count biospecimen_laterality 1 NA NA NA 2 NA NA NA 3 NA NA NA 4 NA NA NA 5 NA NA NA 6 NA NA NA growth_rate catalog_reference 1 NA NA 2 NA NA 3 NA NA 4 NA NA 5 NA NA 6 NA NA

```
2+2
```

```
[1] 4
```