The CGDS-R library

Anders Jacobsen

January 2, 2015

Contents

1 Introduction

This package provides a basic set of R functions for querying the Cancer Genomic Data Server (CGDS) hosted by the Computational Biology Center (cBio) at the Memorial Sloan-Kettering Cancer Center (MSKCC). This service is a part of the cBio Cancer Genomics Portal, `http://www.cbioportal.org/`.

In summary, the library can issue the following types of queries:

- `getCancerStudies()` : What cancer studies are hosted on the server? For example, TCGA glioblastoma or TCGA ovarian cancer.
- `getGeneticProfiles()` : What genetic profile types are available for cancer study X? For example, mRNA expression or copy number alterations.
- `getCaseLists()` : what case sets are available for cancer study X? For example, all samples or only samples corresponding to a given cancer subtype.

- `getProfileData()`: Retrieve slices of genomic data. For example, a client can retrieve all mutation data for PTEN and EGFR in TCGA glioblastoma.
- `getClinicalData()`: Retrieve clinical data (e.g. patient survival time and age) for a given cancer study and list of cases.

Each of these functions will be briefly described in the following sections. The last part of this document includes some concrete examples of how to access and plot the data.

The purpose of this document is to give the reader a quick overview of the `cgdsr` package. Please refer to the corresponding R manual pages for a more detailed explanation of arguments and output for each function.

2 The CGDS R interface

2.1 `CGDS()` : Create a CGDS connection object

Initially, we will establish a connection to the public CGDS server hosted by Memorial Sloan-Kettering Cancer Center. The function for creating a CGDS connection object requires the URL of the CGDS server service, in this case `http://www.cbioportal.org/public-portal/`, as an argument.

```
> library(cgdsr)
> # Create CGDS object
> mycgds = CGDS("http://www.cbioportal.org/public-portal/")
```

The variable `mycgds` is now a CGDS connection object pointing at the URL for the public CGDS server. This connection object must be included as an argument to all subsequent interface calls. Optionally, we can now perform a set of simple tests of the data returned from the CGDS connection object using the `test` function:

```
> # Test the CGDS endpoint URL using a few simple API tests
> test(mycgds)

getCancerStudies...  OK
getCaseLists (1/2) ...  OK
getCaseLists (2/2) ...  OK
getGeneticProfiles (1/2) ...  OK
getGeneticProfiles (2/2) ...  OK
getClinicalData (1/1) ...  OK
getProfileData (1/5) ...  OK
getProfileData (2/5) ...  OK
getProfileData (3/5) ...  OK
getProfileData (4/5) ...  OK
getProfileData (5/5) ...  OK
```

2.2 `getCancerStudies()` : Retrieve a set of available cancer studies

Having created a CGDS connection object, we can now retrieve a data frame with available cancer studies using the `getCancerStudies` function:

```
> # Get list of cancer studies at server
> getCancerStudies(mycgds)[,c(1,2)]
```

```
                    cancer_study_id
1                     laml_tcga_pub
2                         laml_tcga
3                        acyc_mskcc
4                          acc_tcga
5              blca_mskcc_solit_2014
6              blca_mskcc_solit_2012
7                          blca_bgi
8                     blca_tcga_pub
9                         blca_tcga
10                         lgg_tcga
11                        brca_bccrc
12                        brca_broad
13                      brca_sanger
14                    brca_tcga_pub
15                        brca_tcga
16               cellline_ccle_broad
17                        cesc_tcga
18                    chol_nccs_2013
19                     chol_nus_2012
20                coadread_genentech
21                coadread_tcga_pub
22                    coadread_tcga
23                   coadread_mskcc
24                       esca_broad
25                        esca_tcga
26                        escc_icgc
27                  gbm_tcga_pub2013
28                      gbm_tcga_pub
29                          gbm_tcga
30                       hnsc_broad
31                         hnsc_jhu
32                        hnsc_tcga
33                    hnsc_tcga_pub
34                    chol_jhu_2013
35                    kich_tcga_pub
36                        kich_tcga
37                         kirc_bgi
38                    kirc_tcga_pub
39                        kirc_tcga
40                        kirp_tcga
41                     lihc_amc_prv
```

42	lihc_riken
43	lihc_tcga
44	luad_broad
45	luad_tcga_pub
46	luad_tcga
47	luad_tsp
48	lusc_tcga_pub
49	lusc_tcga
50	dlbc_tcga
51	mpnst_mskcc
52	mbl_broad
53	mbl_icgc
54	mbl_pcgp
55	skcm_broad_dfarber
56	mm_broad
57	cellline_nci60
58	npc_nusingapore
59	ov_tcga_pub
60	ov_tcga
61	paad_icgc
62	paad_tcga
63	thca_tcga_pub
64	pcpg_tcga
65	prad_broad_2013
66	prad_broad
67	prad_mskcc
68	prad_tcga
69	prad_mskcc_2014
70	prad_mskcc_cheny1_organoids_2014
71	prad_mich
72	sarc_mskcc
73	sarc_tcga
74	skcm_broad
75	skcm_tcga
76	skcm_yale
77	scco_mskcc
78	sclc_clcgp
79	sclc_jhu
80	stad_pfizer_uhongkong
81	stad_tcga_pub
82	stad_tcga
83	stad_utokyo
84	stad_uhongkong
85	thca_tcga
86	ucs_tcga
87	ucec_tcga
88	ucec_tcga_pub

	name
1	Acute Myeloid Leukemia (TCGA, NEJM 2013)
2	Acute Myeloid Leukemia (TCGA, Provisional)

```
3                                Adenoid Cystic Carcinoma (MSKCC, Nature Genetics 2013)
4                                    Adrenocortical Carcinoma (TCGA, Provisional)
5                                          Bladder Cancer (MSKCC, Eur Urol 2014)
6                                               Bladder Cancer (MSKCC, JCO 2013)
7                             Bladder Urothelial Carcinoma (BGI, Nature Genetics 2013)
8                                    Bladder Urothelial Carcinoma (TCGA, Nature 2014)
9                                    Bladder Urothelial Carcinoma (TCGA, Provisional)
10                                   Brain Lower Grade Glioma (TCGA, Provisional)
11                           Breast Invasive Carcinoma (British Columbia, Nature 2012)
12                                Breast Invasive Carcinoma (Broad, Nature 2012)
13                               Breast Invasive Carcinoma (Sanger, Nature 2012)
14                                 Breast Invasive Carcinoma (TCGA, Nature 2012)
15                                 Breast Invasive Carcinoma (TCGA, Provisional)
16                        Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012)
17 Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA, Provisional)
18       Cholangiocarcinoma (National Cancer Centre of Singapore, Nature Genetics 2013)
19           Cholangiocarcinoma (National University of Singapore,Nature Genetics 2012)
20                                Colorectal Adenocarcinoma (Genentech, Nature 2012)
21                                     Colorectal Adenocarcinoma (TCGA, Nature 2012)
22                                     Colorectal Adenocarcinoma (TCGA, Provisional)
23                    Colorectal Adenocarcinoma Triplets (MSKCC, Genome Biology 2014)
24                            Esophageal Adenocarcinoma (Broad, Nature Genetics 2013)
25                                         Esophageal Carcinoma (TCGA, Provisional)
26                             Esophageal Squamous Cell Carcinoma (ICGC, Nature 2014)
27                                                 Glioblastoma (TCGA, Cell 2013)
28                                                Glioblastoma (TCGA, Nature 2008)
29                                         Glioblastoma Multiforme (TCGA, Provisional)
30                           Head and Neck Squamous Cell Carcinoma (Broad, Science 2011)
31                  Head and Neck Squamous Cell Carcinoma (Johns Hopkins, Science 2011)
32                            Head and Neck Squamous Cell Carcinoma (TCGA, Provisional)
33                            Head and Neck Squamous Cell Carcinoma (TCGA, in revision)
34     Intrahepatic Cholangiocarcinoma (Johns Hopkins University, Nature Genetics 2013)
35                                       Kidney Chromophobe (TCGA, Cancer Cell 2014)
36                                            Kidney Chromophobe (TCGA, Provisional)
37                         Kidney Renal Clear Cell Carcinoma (BGI, Nature Genetics 2012)
38                                 Kidney Renal Clear Cell Carcinoma (TCGA, Nature 2013)
39                                 Kidney Renal Clear Cell Carcinoma (TCGA, Provisional)
40                            Kidney Renal Papillary Cell Carcinoma (TCGA, Provisional)
41                               Liver Hepatocellular Carcinoma (AMC, Hepatology 2014)
42                           Liver Hepatocellular Carcinoma (RIKEN, Nature Genetics 2012)
43                                   Liver Hepatocellular Carcinoma (TCGA, Provisional)
44                                           Lung Adenocarcinoma (Broad, Cell 2012)
45                                        Lung Adenocarcinoma (TCGA, Nature 2014)
46                                        Lung Adenocarcinoma (TCGA, Provisional)
47                                         Lung Adenocarcinoma (TSP, Nature 2008)
48                                     Lung Squamous Cell Carcinoma (TCGA, Nature 2012)
49                                     Lung Squamous Cell Carcinoma (TCGA, Provisional)
50                    Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (TCGA, Provisional)
51                  Malignant Peripheral Nerve Sheath Tumor (MSKCC, Nature Genetics 2014)
52                                                 Medulloblastoma (Broad, Nature 2012)
```

```
53                                                Medulloblastoma (ICGC, Nature 2012)
54                                                Medulloblastoma (PCGP, Nature 2012)
55                                             Melanoma (Broad/Dana Farber, Nature 2012)
56                                            Multiple Myeloma (Broad, Cancer Cell 2014)
57                                             NCI-60 Cell Lines (NCI, Cancer Res. 2012)
58                          Nasopharyngeal Carcinoma (Singapore, Nature Genetics 2014)
59                               Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011)
60                               Ovarian Serous Cystadenocarcinoma (TCGA, Provisional)
61                                       Pancreatic Adenocarcinoma (ICGC, Nature 2012)
62                                       Pancreatic Adenocarcinoma (TCGA, Provisional)
63                                       Papillary Thyroid Carcinoma (TCGA, Cell 2014)
64                                  Pheochromocytoma and Paraganglioma (TCGA, Provisional)
65                                     Prostate Adenocarcinoma (Broad/Cornell, Cell 2013)
66                           Prostate Adenocarcinoma (Broad/Cornell, Nature Genetics 2012)
67                                      Prostate Adenocarcinoma (MSKCC, Cancer Cell 2010)
68                                            Prostate Adenocarcinoma (TCGA, Provisional)
69                                  Prostate Adenocarcinoma CNA study (MSKCC, PNAS 2014)
70                                  Prostate Adenocarcinoma Organoids (MSKCC, Cell 2014)
71                            Prostate Adenocarcinoma, Metastatic (Michigan, Nature 2012)
72                                           Sarcoma (MSKCC/Broad, Nature Genetics 2010)
73                                                        Sarcoma (TCGA, Provisional)
74                                         Skin Cutaneous Melanoma (Broad, Cell 2012)
75                                         Skin Cutaneous Melanoma (TCGA, Provisional)
76                                       Skin Cutaneous Melanoma (Yale, Nature Genetics 2012)
77                         Small Cell Carcinoma of the Ovary (MSKCC, Nature Genetics 2014)
78                                       Small Cell Lung Cancer (CLCGP, Nature Genetics 2012)
79                                  Small Cell Lung Cancer (Johns Hopkins, Nature Genetics 2012)
80                                Stomach Adenocarcinoma (Pfizer and UHK, Nature Genetics 2014)
81                                           Stomach Adenocarcinoma (TCGA, Nature 2014)
82                                           Stomach Adenocarcinoma (TCGA, Provisional)
83                                 Stomach Adenocarcinoma (U Tokyo, Nature Genetics 2014)
84                                     Stomach Adenocarcinoma (UHK, Nature Genetics 2011)
85                                                  Thyroid Carcinoma (TCGA, Provisional)
86                                           Uterine Carcinosarcoma (TCGA, Provisional)
87                                Uterine Corpus Endometrial Carcinoma (TCGA, Provisional)
88                               Uterine Corpus Endometrioid Carcinoma (TCGA, Nature 2013)
```

Here we are only showing the first two columns, the cancer study ID and short name, of the result data frame. There is also a third column, a longer description of the cancer study. The cancer study ID must be used in subsequent interface calls to retrieve case lists and genetic data profiles (see below).

2.3 `getGeneticProfiles()` : Retrieve genetic data profiles for a specific cancer study

This function queries the CGDS API and returns the available genetic profiles, e.g. mutation or copy number profiles, stored about a specific cancer study. Below we list the current genetic profiles for the TCGA glioblastoma cancer study:

```
> getGeneticProfiles(mycgds,'gbm_tcga')[,c(1:2)]
```

```
                       genetic_profile_id
1                       gbm_tcga_mutations
2               gbm_tcga_RPPA_protein_level
3                gbm_tcga_methylation_hm450
4                 gbm_tcga_methylation_hm27
5                         gbm_tcga_log2CNA
6                  gbm_tcga_rna_seq_v2_mrna
7  gbm_tcga_rna_seq_v2_mrna_median_Zscores
8              gbm_tcga_mrna_median_Zscores
9                gbm_tcga_mrna_U133_Zscores
10                      gbm_tcga_mrna_U133
11                         gbm_tcga_gistic
12                           gbm_tcga_mrna
                               genetic_profile_name
1                                        Mutations
2              protein/phosphoprotein level (RPPA)
3                               Methylation (HM450)
4                                Methylation (HM27)
5                           Log2 copy-number values
6                 mRNA expression (RNA Seq V2 RSEM)
7       mRNA Expression z-Scores (RNA Seq V2 RSEM)
8            mRNA Expression z-Scores (microarray)
9  mRNA Expression z-Scores (U133 microarray only)
10          mRNA expression (U133 microarray only)
11    Putative copy-number alterations from GISTIC
12                    mRNA expression (microarray)
```

Here we are only listing the first two columns, genetic profile ID and short name, of the resulting data frame. Please refer to the R manual pages for a more extended specification of the arguments and output.

2.4 getCaseLists() : Retrieve case lists for a specific cancer study

This function queries the CGDS API and returns available case lists for a specific cancer study. For example, within a particular study, only some cases may have sequence data, and another subset of cases may have been sequenced and treated with a specific therapeutic protocol. Multiple case lists may be associated with each cancer study, and this method enables you to retrieve meta-data regarding all of these case lists. Below we list the current case lists for the TCGA glioblastoma cancer study:

```
> getCaseLists(mycgds,'gbm_tcga')[,c(1:2)]
```

```
                case_list_id                              case_list_name
1     gbm_tcga_3way_complete                  All Complete Tumors
2               gbm_tcga_all                           All Tumors
3         gbm_tcga_sequenced                     Sequenced Tumors
4               gbm_tcga_cna                           Tumors CNA
5           gbm_tcga_log2CNA                 Tumors log2 copy-number
```

```
6   gbm_tcga_methylation_hm27          Tumors with methylation data (HM27)
7  gbm_tcga_methylation_hm450         Tumors with methylation data (HM450)
8               gbm_tcga_mrna   Tumors with mRNA data (Agilent microarray)
9    gbm_tcga_rna_seq_v2_mrna           Tumors with mRNA data (RNA Seq V2)
10        gbm_tcga_mrna_U133 Tumors with mRNA data (U133 microarray only)
11             gbm_tcga_rppa                         Tumors with RPPA data
12           gbm_tcga_cnaseq          Tumors with sequencing and CNA data
```

Here we are only listing the first two columns, case list ID and short name, of the resulting data frame. Please refer to the R manual pages for a more extended specification of the arguments and output.

2.5 getProfileData() : Retrieve genomic profile data for genes and genetic profiles

The function queries the CGDS API and returns data based on gene(s), genetic profile(s), and a case list. The function only allows specifying a list of genes and a single genetic profile, or oppositely a single gene and a list of genetic profiles. Importantly, the format of the output data frame depends on if a single or a list of genes was specified in the arguments. Below we are retrieving mRNA expression and copy number alteration genetic profiles for the NF1 gene in all samples of the TCGA glioblastoma cancer study:

```
> getProfileData(mycgds, "NF1", c("gbm_tcga_gistic","gbm_tcga_mrna"), "gbm_tcga_all")[c(1:

                gbm_tcga_gistic gbm_tcga_mrna
TCGA.02.0001.01              -1            NA
TCGA.02.0003.01               0            NA
TCGA.02.0006.01               0            NA
TCGA.02.0007.01               0            NA
TCGA.02.0009.01               0            NA
```

We are here only showing the first five rows of the data frame. In the next example, we are retrieving mRNA expression data for the MDM2 and MDM4 genes:

```
> getProfileData(mycgds, c("MDM2","MDM4"), "gbm_tcga_mrna", "gbm_tcga_all")[c(1:5),]

                MDM2 MDM4
TCGA.02.0001.01   NA   NA
TCGA.02.0003.01   NA   NA
TCGA.02.0006.01   NA   NA
TCGA.02.0007.01   NA   NA
TCGA.02.0009.01   NA   NA
```

We are again only showing the first five rows of the data frame.

2.6 getClinicalData() : Retrieve clinical data for a list of cases

The function queries the CGDS API and returns available clinical data (e.g. patient survival time and age) for a given case list. Results are returned in a

data frame with a row for each case and a column for each clinical attribute. The available clinical attributes are:

- `overall_survival_months`: Overall survival, in months.
- `overall_survival_status`: Overall survival status, usually indicated as "LIVING" or "DECEASED".
- `disease_free_survival_months`: Disease free survival, in months.
- `disease_free_survival_status`: Disease free survival status, usually indicated as "DiseaseFree" or "Recurred/Progressed".
- `age_at_diagnosis`: Age at diagnosis.

Below we retrieve clinical data for the TCGA ovarian cancer dataset (only first five cases/rows are shown):

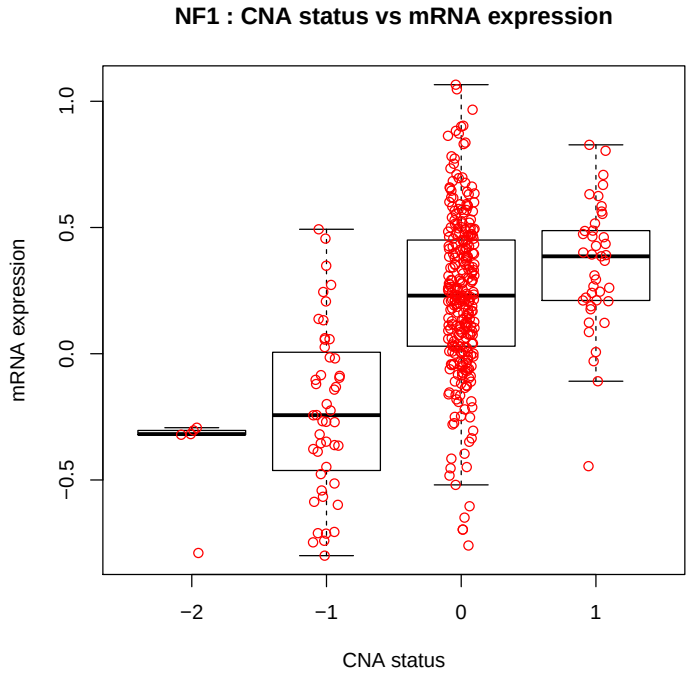
```
> getClinicalData(mycgds, "ova_all")[c(1:5),]

data frame with 0 columns and 5 rows
```

3 Examples

3.1 Example 1: Association of NF1 copy number alteration and mRNA expression in glioblastoma

As a simple example, we will generate a plot of the association between copy number alteration (CNA) status and mRNA expression change for the NF1 tumor suprpressor gene in glioblastoma. This plot is very similar to Figure 2b in the TCGA research network paper on glioblastoma (McLendon et al. 2008). The mRNA expression of NF1 has been median adjusted on the gene level (by globally subtracting the median expression level of NF1 across all samples).

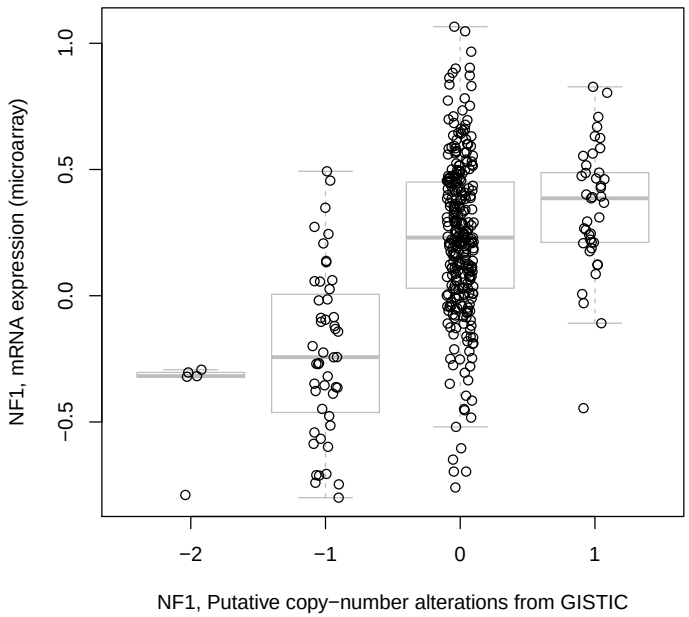
```
> df = getProfileData(mycgds, "NF1", c("gbm_tcga_gistic","gbm_tcga_mrna"), "gbm_tcga_all")
> head(df)

                gbm_tcga_gistic gbm_tcga_mrna
TCGA.02.0001.01              -1            NA
TCGA.02.0003.01               0            NA
TCGA.02.0006.01               0            NA
TCGA.02.0007.01               0            NA
TCGA.02.0009.01               0            NA
TCGA.02.0010.01               0            NA

> boxplot(df[,2] ~ df[,1], main="NF1 : CNA status vs mRNA expression", xlab="CNA status",
> stripchart(df[,2] ~ df[,1], vertical=T, add=T, method="jitter",pch=1,col='red')
```



Alternatively, the generic `cgdsr plot()` function can be used to generate a similar plot:

```
> plot(mycgds, "gbm_tcga", "NF1", c("gbm_tcga_gistic","gbm_tcga_mrna"), "gbm_tcga_all", sk
```

```
[1] TRUE
```



3.2 Example 2: MDM2 and MDM4 mRNA expression levels in glioblastoma

In this example, we evaluate the relationship of MDM2 and MDM4 expression levels in glioblastoma. mRNA expression levels of MDM2 and MDM4 have been median adjusted on the gene level (by globally subtracting the median expression level of the individual gene across all samples).

```
> df = getProfileData(mycgds, c("MDM2","MDM4"), "gbm_tcga_mrna", "gbm_tcga_all")
> head(df)
```

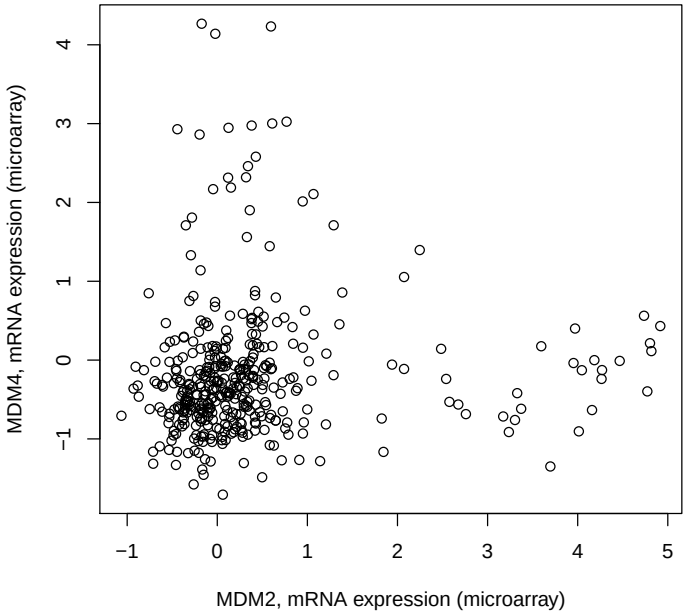
```
                MDM2 MDM4
TCGA.02.0001.01   NA   NA
TCGA.02.0003.01   NA   NA
TCGA.02.0006.01   NA   NA
TCGA.02.0007.01   NA   NA
TCGA.02.0009.01   NA   NA
TCGA.02.0010.01   NA   NA
```

```
> plot(df, main="MDM2 and MDM4 mRNA expression", xlab="MDM2 mRNA expression", ylab="MDM4 m
```

Alternatively, the generic `cgdsr plot()` function can be used to generate a similar plot:

```
> plot(mycgds, "gbm_tcga", c("MDM2","MDM4"), "gbm_tcga_mrna" ,"gbm_tcga_all")
```

```
[1] TRUE
```



3.3 Example 3: Comparing expression of PTEN in primary and metastatic prostate cancer tumors

In this example we plot the mRNA expression levels of PTEN in primary and metastatic prostate cancer tumors.

```
> df.pri = getProfileData(mycgds, "PTEN", "prad_mskcc_mrna", "prad_mskcc_primary")
> head(df.pri)
```

```
            PTEN
PCA0001 9.467183
PCA0002 9.041528
PCA0003 8.511305
PCA0004       NA
PCA0005 9.413217
PCA0006       NA
```

```
> df.met = getProfileData(mycgds, "PTEN", "prad_mskcc_mrna", "prad_mskcc_mets")
> head(df.met)
```

```
            PTEN
PCA0182 7.486938
PCA0183       NA
PCA0184 7.578755
PCA0185       NA
PCA0186       NA
PCA0187 8.756132
```

```
> boxplot(list(t(df.pri),t(df.met)), main="PTEN expression in primary and metastatic tumor
> stripchart(list(t(df.pri),t(df.met)), vertical=T, add=T, method="jitter",pch=1,col='red'
```