cgData:
UCSC Cancer Genomics Browser Data Specification

v1.22, June 5, 2012

Kyle Ellrott, Brian Craft, Mark Diekhans, Mary Goldman, Teresa Swatloski, Singer Ma, Chris Wilks and Jingchun Zhu

Table of Contents

1) Introduction

The UCSC Cancer Genomics browser (https://genome-cancer.soe.ucsc.edu) displays genomic data and the associated clinical information for cohorts of samples. We developed cgData (UCSC **C**ancer **G**enomics Browser **DATA** specification) to support our large genomic data repository. External data that meets this specification can be easily ingested to and visualized by the UCSC Cancer Genomics Browser.

1.1) Data overview

The primary data files necessary for each data set are:

- Genomic data
- Clinical data
- Clinical feature description (optional)
- probeMap
- sampleMap

Each primary data file is also required to have an associated meta data file. Meta data includes information like the genome assembly or how the primary data files are connected. The meta data file is required to be in JSON format.

Example:
my_genomic_data
my_genomic_data.json

2) Genomic data

2.1) genomicMatrix data file

The genomic matrix is a tab-separated file containing the genomic values for each sample for each probe. The matrix is arranged so that the columns are samples and the rows are the probe names. The first line in the file defines the sample names and the first column defines the probe names. If there is no genomic value for a particular sample/probe, the value is written as ‘NA’.

The probe names are consistent with the probe names in the corresponding *probeMap* and the sample names are consistent with the sample identifiers in the corresponding *sampleMap*.

Example:

```
NAME	TCGA-CH-5748-01A-11D-1574-01	TCGA-CH-5748-01A-11D-1574-01
ELMO2	0.162208	0.577708
CREB3L1	1.338	-0.4835
RPS11	0.044063	-0.25806
```

2.2) genomicMatrix meta data file

Required fields are in bold. Other fields are optional.

type: “genomicMatrix”
name: name of the genomic data set
shortTitle: short title of the genomic data set
longTitle: longer description of the genomic data set
wrangler**:** person who wrangled the data
:probeMap: name of the probeMap file
:dataSubType: type of genomic data:

- cna: DNA copy number aberration
- geneExp: gene expression
- miRNAExp: miRNA expression
- protein: protein activity
- DNAMethylation: DNA methylation
- siRNAViability: cell viability under siRNA knockdown screen
- RPPA: protein activity
- PARADIGM: UCSC Paradigm pathway analysis pathway activity
- PARADIGM.pathlette: UCSC Paradigm pathlette pathway analysis pathway activity

dataProducer: url of publication or, for unpublished data, the lab that produced the data
version: yyyy-mm-dd
:sampleMap: name of the sampleMap file
redistribution: true/false (whether bulk downloads of the dataset is allowed)
platform: experimental platform (such as IlluminaHiSeq)
articleTitle: title of the published article
citation: article citation
url: url of the published article

Example:

```
{
"name": "brca_illumina_20110101",
"type": "genomicMatrix",
"shortTitle": "TCGA BRCA expression data",
"longTitle":" TCGA breast cancer gene expression data",
":dataSubType": "geneExp",
“author”: “Chris Szeto”,
“group”: “tcga”,
":probeMap": "affyU133A",
":sampleMap": "tcgaSamples",
“version”: ”2011-01-01”,
“redistribution”: true,
"dataProducer": "url for TCGA DCC",
}
```

2.3) genomicSegment data file

Segmented data, typically copy number data, is best represented as a genomicSegment file rather than a genomic matrix. It is a 6 column tab-separated format:

sampleID: sample identifiers that correspond to the ID in the sampleMap
chrom: name of the chromosome (e.g. chr3, chrY), please note, we use chrX, chrY, chrM, not chr23, chr24. Chr#_random data is not displayed in the browser but may be part of the file.
start: starting position in the chromosome or scaffold.
end: ending position in the chromosome or scaffold.

strand: either +:forward; -:reverse; or **.** for both strands.
score: genomic data value

See Notes for more information about genomic coordinates.

Example:

```
TCGA-CH-5748-01A-11D-1574-01    chr3    19981039    55760837    .    -0.0029
TCGA-CH-5748-01A-11D-1574-01    chr3    55761580    55765804    .    -1.1186
TCGA-CH-5748-01A-11D-1574-01    chr3    55770566    68828854    .    0.0132
```

2.4) genomicSegment meta data file

Required fields are in bold. Other fields are optional and may be added as desired.

type: “genomicSegment”
name: name of the genomic data
shortTitle: short title of the genomic data set
longTitle: longer description of the genomic data set
:assembly: the genomic assembly for the segment set
:dataSubType: type of genomic data:

- cna: DNA copy number aberration
- geneExp: gene expression
- miRNAExp: miRNA expression
- protein: protein activity
- DNAMethylation: DNA methylation
- siRNAViability: cell viability under siRNA knockdown screen
- RPPA: protein activity
- PARADIGM: UCSC Paradigm pathway analysis pathway activity
- PARADIGM.pathlette: UCSC Paradigm pathlette pathway analysis pathway activity

dataProducer: url of publication or, for unpublished data, the lab that produced the data
version: yyyy-mm-dd
:sampleMap: name of the sampleMap file
redistribution: true/false (whether bulk downloads of the dataset is allowed)
platform: experimental platform (such as SNP6)
articleTitle: title of the published article
citation: article citation
url: url of the published article

Example:

```
{
"name": "brca_illumina_20110101",
"type": "genomicSegment",
"shortTitle": "TCGA BRCA copy number variation",
"longTitle":" TCGA breast cancer copy number variation data",
":dataSubType": "cna",
“author”: “Chris Szeto”,
“group”: “tcga”,
":assembly": "hg18",
":sampleMap": "tcgaSamples",
“version”: ”2011-01-01”,
“redistribution”: true,
"dataProducer": "url for TCGA DCC",
}
```

3) Clinical data

3.1) clinialMatrix data file

The clinical matrix is a tab-separated file containing the clinical values for each sample for each clinical parameter/feature. The matrix is arranged so that the columns are the clinical parameters and the rows are samples. The first line defines the clinical parameter name and the first column defines the sample names. The very first column must be titled ‘sampleID’.

The sample names need to be consistent with the sample identifiers in the corresponding sampleMap. If there is no genomic value for a particular sample/probe, the value should be left blank.

Example:

```
sampleID	center	erStatus
ABCD-1EFG-2JKL-3OPQ	Broad	+
ABCD-EFGH-JKLM-NOPQ	UCSC	+
```

3.2) clinicalMatrix meta data file

Required fields are in bold. Other fields may be added as desired.

type: “clinicalMatrix”

name: name of the clinical data

:sampleMap: name of the sampleMap file

version: yyyy-mm-dd

:clinicalFeature**:** name of the clinicalFeature file, if available

Example:

```
{
"type": "clinicalMatrix",
“name”:”TCGABRCAClinicalMatrix”,
"version":"2012-05-31",
“:clinicalFeature”:”TCGABRCAClinicalFeature”,
":sampleMap": "tcgaBRCASamples"
}
```

4) Clinical Feature Description

4.1) clinicalFeature data file

The clinical feature file contains additional details about the clinical features in the clinical matrix file. This file is not required, but can help the cancer browser display clinical information in a more meaningful way. The clinical feature file uses a three-column tab-separated format; the first column is the name of clinical feature, the second is a key word from our cgData vocabulary, and the third is the value. Here are some of the most common cgData key words:

cgData key words:

shortTitle	optional	Short text description of feature
longTitle	optional	Longer text description of feature
valueType	required	"category" or "float" (whether this is a continuous feature like age or a discrete feature like ER status)
state	optional	Valid states of ‘category’ valueType.
stateOrder	optional	The order with which the states should be sorted. Comma-separated list

Example:

```
ER      shortTitle      estrogen receptor status
ER      longTitle       estrogen receptor positivity (positive, negative, intermediate)
ER      valueType       category
ER      state   positive
ER      state   negative
ER      state   intermediate
ER      stateOrder      positive,intermediate,negative,

GI50_BIBW2992   shortTitle      GI50 of BIBW29920
GI50_BIBW2992   longTitle       GI50 of BIBW29920 in -log(M) concentration
GI50_BIBW2992   valueType       float
```

4.2) clinicalFeature meta data file

Required fields are in bold. Other fields may be added as desired.

type: “clinicalFeature”
name: name of the clinical feature file (30 character limit)
:clinicalMatrix: associated clinical matrix
version:yyyy-dd-mm

Example:

```
{
“name”: “TCGABRCAClinicalFeature”,
“shortTitle”:“TCGA BRCA Public clinical feature Information”,
“longTitle”: “TCGA BRCA Public clinical feature Information”,
“type”: “clinicalFeature”,
“:clinicalMatrix”: “TCGABRCAClinicalMatrix”,
"version": 2012-05-31
}
```

5) sampleMap

5.1) sampleMap file

Some clinical information is mapped to the patient (like age), while other clinical information is mapped to the sample (like sample type). This results in a tree-like structure where multiple slide/section/aliquot ids can share the same sample ids and multiple sample ids share the same patient id. For example in the TCGA data set:

```
Patient1 – Sample1 – Aliquot1-1
          \  Sample2 – Aliquot2-1
                       \   Aliquot2-2
```

The sampleMap expresses this data in a two-column, parent-child tab-separated file. If an identifier does not have a parent or a child, it is specified as "self self". A sample must be specified in the sampleMap in order to be displayed on the browser.

Example (with parent-child relationships):

```
TCGA-AA-0001          TCGA-AA-0001
TCGA-AA-0001          TCGA-AA-0001-01A
TCGA-AA-0001          TCGA-AA-0001-11A
TCGA-AA-0001-11A      TCGA-AA-0001-11A-0001
TCGA-AA-0001-11A      TCGA-AA-0001-11A-0002
TCGA-AA-0001-11A-0001     TCGA-AA-0001-11A-0001
TCGA-AA-0001-11A-0002     TCGA-AA-0001-11A-0002
```

Example (without parent-child relationships):

```
M001   M001
M002   M002
M003   M003
```

5.2) Metadata (in JSON file) of a sampleMap file

Required fields are in bold. Other fields may be added as desired.

type: “sampleMap”
name: name of the sampleMap
version:yyyy-mm-dd

Example:

```
{
"type": "sampleMap",
"name": "TCGABRCASample",
"version":2012-05-31,
"shortTitle": "TCGA breast cancer project identifier map",
"longTitle": "TCGA breast cancer project identifier map"
}
```

6) probeMap

6.1) probeMap data file

A probeMap id a tab-separated file that connects probes from a microarray platform to their genomic coordinates. The probeMap file must have mapping from probe to HUGO gene names through the aliasList.

Title rows or any other comments must be preceded by a “#”. Required fields are in bold.

name: name of the probe
aliasList: comma separated alias names, such as HUGO names.
chrom: name of the chromosome (e.g. chr3, chrY), please note, we use chrX, chrY, chrM, not chr23, chr24. Chr#_random data is not displayed in the browser but may be part of the file.
chromStart: starting position of the probe in the chromosome or scaffold.
chromEnd: ending position of the probe in the chromosome or scaffold.
strand: either +:forward; -:reverse; or **.** for both strands

See Notes for more information about genomic coordinates and the aliasList.

(The following fields are optional as a group)
thickStart: starting position of translation. Use NULL if unknown or unsure.
thickEnd : ending position of translation. Use NULL if unknown or unsure.
blockCount: number of blocks (exons)
blockSizes: comma-separated list of the block sizes. The number of items in this list should correspond to blockCount.
blockStarts: comma-separated list of block starts. All of the blockStart positions should be calculated relative to chromStart. The number of items in this list should correspond to blockCount.

Example:

RP11-243P9	TTTY14,CD24	chrY	19532644	19869336	.
RP11-109F19	PRKY	chrY	7282039	7451740	.
RP11-478I15		chrY	17602121	17781835	.
RP11-88F4		chrY	49745385	49840479	.
RP11-182H20	RBMY1A3P,TTTY20	chrY	9092794	9255172	.
RP11-214M24	TTTY17A,TTTY17B,TTTY17C	chrY	25642524	25939317	.
RP11-20H21	TTTY14,CD24	chrY	19527090	19870478	.
RP11-91N9	USP9Y	chrY	13351811	13513105	.
RP11-386L3		chrY	14451443	14621963	.

6.2) probeMap meta data file

Required fields are in bold. Other fields may be added as desired.

type: “probeMap”
name: name of the probeMap
:assembly: the genomic assembly for the probeMap
version:yyyy-mm-dd

Example:

```
{
"type": "probeMap",
"name": "affyU133_hg18",
"version":"2012-05-31",
“:assembly”: “hg18”,
“shortTitle”: "affyU133A microarray platform probe information",
“longTitle”: "affyU133A microarray platform probe information mapped to hg18 assembly,
information downloaded from GEO GPL1234 record and processed."
}
```

7) Notes

7.1) Genomic Positions are One-Based and INCLUSIVE

Genomic coordinates in cgData files use one-based coordinates notation. In addition start and end positions are both INCLUSIVE coordinates.

Please note that UCSC Genome Browser (not UCSC Cancer Browser) uses zero-based half inclusive and half-exclusive coordinates. For example cgData [1,10] equals to UCSC genome browser’s [0,10), both representing the beginning 10 bases of a region.

cgData example
chr1 1 10 → include first base
UCSC genome browser bed
chr1 1 10 → skip the first base (position = 0)

7.2) Genomic Positions Use chrStart <= chrEnd Coordinates

cgData requires chromStart<=chromEnd, and use “strand” to specify forward or reverse strand.
strand = ”+” forward strand
strand = ”-” reverse strand
strand = “.” both strands

7.3) Genesets view requires aliasList in probeMap files

Cancer Browser’s genesets view completely relies on the mapping between probe and alias; we do not use genomic coordinates to map probes dynamically. While any probe can be mapped to any alias, the mappings must be made explicit. For example, if a data provider does not include an alias (for example TP53) in the probeMap file, the cancer browser will not display any data for that alias in the genesets view.

8) FAQ

8.1) Why does some information in the metadata files have a ":" before it?

The ":" indicates that this is a common piece of information that may be connected through multiple datasets. For instance, many datasets may share the same probeMap, sampleMap or assembly. We use this linking to help us gain a larger picture of the entire repository.

8.2) What assemblies are supported by cgData?

We currently support hg17/NCBI35, hg18/NCBI36 and hg19/GRCh37.