Data Elements for the QTL Viewer

The QTL Viewer utilizes R and several different libraries in order to calculate the data for various types of QTL projects. The following sections will explain each element in detail.

Please note that some data element must be pre-computed.

R Environment Overview

The following elements should be contained within the R environment. These can be in one or multiple RData and/or Rds files.

Element	Description
`ensembl_release`	the numerical version of Ensembl
`genoprobs`	the genotype probabilities
`K`	the kinship matric
`map`	list of one element per chromosome, with the genomic position of each marker
`markers`	marker names and positions

The following element is a *special* element. A good practice is to keep a one to one matching between dataset and Rds file.

Element	Description
`dataset.*`	where * should be a very short, unique and informative name. This element will contain most of the data and will be detailed in the section below.

Exact case of element and variable names is very important.

Other meta data can be included in the RData file as long as there are no conflicting names.

Elements

ensembl.version

`R data type: numeric`

This specifies the genome release version for the genomic marker positions and

for annotations attached to molecular phenotypes IF any, i.e. mRNA. Please see the documentation at Ensembl for build and release information.

genoprobs

`R data type: list, calc_genoprobs`

This is the genotype probabilities and must be supplied by the user. This is a list with one element per chromosome of **N** * **K** * **Mj** arrays, where:

- **N** represents the number of samples (i.e. mice)
- **K** represents the number of strains (i.e. founder strains)
- **Mj** represents the number of markers on chromosome **j**

`rownames(genoprobs)` are the same value of the sample id column in the samples element

`colnames(genoprobs)` are strains, for the founder strains they are symbols A,B,C,D,E,F,G,H

`dimnames(genoprobs[[j]])` are marker names on chromosome `j`

May be produced by qtl2convert::probs_doqtl_to_qtl2. Please see the documentation of R/qtl2geno.

K

`R data type: list`

A list of kinship matrices, with one element per chromosome of **N** * **N** matrices, where:

- **N** represents the number of samples

`rownames(K)` are the same value of the sample id column in the samples element

`colnames(K)` are the same value of the sample id column in the samples element

May be produced by qtl2geno::calc_kinship(genoprobs, type="loco"). Please see the documentation of R/qtl2geno.

map

`R data type: list`

This is a list with one element per chromosome of named numeric vector. Elements of the vector are positions along the chromosome in Mb units. Element names are marker names and must match the dimnames of genoprobs.

Users can download maps for MUGA platforms or for 69k pseudomarker grid.

May be produced by qtl2convert::map_df_to_list. Please see the documentation of R/qtl2geno.

markers

`R data type: tibble`

Marker information containing the following information:

- **marker_id** character string, unique name of the marker
- **chr** character string, the chromosome
- **pos** numeric, position in Mbp

The dataset.* Element

The environment must contain at least one object of this type, multiple are allowed. The * should be a very short, unique and informative name. It is for internal use only and will not appear in the QTL Viewer interface.

The main purpose of the dataset.* element is to store multiple datasets per RData file with informative information regarding the data.

The dataset.* element is a list that should contain the following named elements:

Element	Description
`annot.**datatype**`	annotations, where datatype is one of **mrna**, **protein**, **phos** or **phenotype**
`annot_samples`	annotation data for the samples
`covar_info`	*(optional)* information describing the covariates
`data`	either a matrix containing data or a list containing several kinds of data
`datatype`	one of **mrna**, **protein**, **phos** or **phenotype**
`display_name`	name of the dataset, for QTL Viewer display purposes
`lod_peaks`	*(optional)* a list of LOD peaks over a certain threshold

annot_datatype

`R data type: tibble`

The annot_***datatype*** element will have different data and column names depending on whether this is a **mrna**, **protein**, **phos** or **phenotype** dataset.

For **mrna**, the following fields are required:

Field	Description
`gene_id`	character string, Ensembl gene id
`symbol`	character string, Symbol of the gene

Field	Description
chr	character string, chromosome
start	numeric, position in Mbp
end	numeric, position in Mbp

For **protein**, all **mrna** fields *PLUS* the following field:

Field	Description
protein_id	character string, Ensembl protein id

For **phos**, all **protein** fields *PLUS* the following field:

Field	Description
phos_id	character string, Phosphopetide ID

For **phenotype**, the following fields are required:

Field	Description
data_name	character string, phenotype id
short_name	character string, short descriptive name
category	character string, category if any
description	character string, phenotype description
is_id	logical, should only be 1 TRUE
is_pheno	logical, is this an actual phenotype
is_numeric	logical, is this a numeric field
omit	logical, T to omit, F to include
use_covar	character string, colon seperated covar values

Also for **phenotype**, the following fields are legacy fields (not required):

Field	Description
units	character string, measureing units
R_name	character string, name used by R
R_category	character string, category used by R
is_date	logical, does this contain a date
is_factor	logical, is this a factor
factor_levels	character string, ":" separated values
is_covar	logical, is this a covariate
is_derived	logical, is this phenotype derived

Extra information in the tibble will be ignored by the QTL Viewer.

annot_samples

`R data type: tibble`

Annotations for the samples in this dataset. The unique identifying column is **sample_id**. We use a regular expression to determine the unique **sample_id** column. Examples that work are mouse.id, mouse_id, sample.id, sample_id. There should be a unique value for **sample_id** in every row.

For the purpose of doing certain scans, there will need to be other columns that match the information stored in the covar_info element.

covar_info

`R data type: tibble`

This element controls how we scan and interact with the RData object. The following columns must be present:

Field	Description
`sample_column`	name of the column in the **annot_samples** element
`display_name`	QTL Viewer uses this to display a nice name
`interactive`	TRUE for an interactive covariate, must also set lod_peaks if TRUE. If FALSE, lod_peaks value should be NA. This controls whether or not interactive scans are performed for a particular covariate.
`primary`	which covariate to display preselected in the Effect Plot
`lod_peaks`	named tibble in the lod_peaks element

data

`R data type: matrix or list`

This element is either a matrix or a list.

If it is a matrix, there is one and only set of data for this dataset.

If it is a list, each named element in the list should be a matrix with the following controlled vocabulary for the names:

- **rz**
- **norm**
- **raw**
- **log**
- **transformed**

Each matrix will contain numerical data with samples (rows) by annotations (columns).

datatype

`R data type: character`

This will be used to identify the type of dataset. This is a controlled vocabulary consisting of the following values:

- **mrna**
- **protein**
- **phos**
- **phenotype**

Based upon the value of this element, the QTL Viewer will treat the data as accordingly.

display_name

`R data type: character`

This will be used to display the name of the dataset to the user in the QTL Viewer. This will be used in a dropdown menu to switch among the datasets.

lod_peaks

`R data type: list`

This is a list with each value in the list being either **additive** (the default) or one of the interactive covariates (if set in covar_info). The **additive** values should always be present.

The covar_info element should have values with interactive set to TRUE and lod.peaks set to the name of the element in this list.

Depending on the value of datatype (**mrna**, **protein**, **phos**, **phenotype**), the annotation column identifier will match to the appropriate column in the annot_*datatype* element.

The following shows the required fields in each tibble.

If datatype is **mrna**, the following fields are required:

Field	Description
`gene_id`	the Ensembl gene identifier in the annot_mrna element
`marker_id`	the marker identifier in the markers element
`lod`	the lod score

If datatype is **protein**, the following fields are required:

Field	Description
`protein_id`	the Ensembl protein id in the annot_protein element
`marker_id`	the marker identifier in the markers element
`lod`	the lod score

If datatype is **phos**, the following fields are required:

Field	Description
`phos_id`	the Phosphopeptide identifer
`marker_id`	the marker identifier in the markers element
`lod`	the lod score

If datatype is **phenotype**, the following fields are required:

Field	Description
`data_name`	the unique identifier in the annot_phenotype element
`marker_id`	the marker identifier in the markers element
`lod`	the lod score