## How to convert your data into QTL viewer Data:

### Step 1: Generate the following files in TAB separated format (no headers)

#### features.txt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LAYOUT | | | | |
|  | NAME | DATA TYPE | EXAMPLE | NOTE |
| 1 | feature\_id | STRING | ENMUSG00000019966 |  |
| 2 | group\_id | STRING | ENMUSG00000019966 |  |
| 3 | chrom | CHAR (2 characters) | 10 |  |
| 4 | location | FLOAT | 100.01563 | location is in Mb |
| 5 | name | STRING | Kitl |  |
| 6 | description | STRING | Kit ligand |  |

Example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ENSMUSG00000000001 | ENSMUSG00000000001 | 3 | 108.126713 | Gnai3 | guanine nucleotide binding protein |
| ENSMUSG00000000049 | ENSMUSG00000000049 | 11 | 108.378875 | Apoh | apolipoprotein H |
| ENSMUSG00000000088 | ENSMUSG00000000088 | 9 | 57.5268828 | Cox5a | cytochrome c oxidase subunit Va |

***NOTE***: if genes, group\_id = feature\_id,

if pQTL, the group\_id will be gene\_id while the feature\_id will be protein id

#### markers.txt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LAYOUT | | | | |
|  | **NAME** | **DATA TYPE** | **EXAMPLE** | **NOTE** |
| 1 | marker\_id | STRING | rs6412653 |  |
| 2 | chrom | CHAR (2 characters) | X |  |
| 3 | location | FLOAT | 11.232 | location is in Mb |

Example:

|  |  |  |
| --- | --- | --- |
| 1\_677642 | 1 | 0.677642 |
| rs48774772 | 9 | 57.522172 |
| 1\_32749486 | 1 | 32.7494861 |

#### lod.txt - dataset of LOD scores in FLOATING point, features (rows) x markers (columns)

**M**

**F**

**M is the number of markers**

**F is the number of features**

***NOTE***: ALL values should be of type FLOAT, if there is no value, leave that cell empty

**FOR THE COEEFICIENT EFFECT PLOT (not required)**

#### strains.txt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LAYOUT | | | | |
|  | **NAME** | **DATA TYPE** | **EXAMPLE** | **NOTE** |
| 1 | strain\_id | STRING | A |  |
| 2 | name | STRING | A/J |  |
| 3 | description | STRING |  |  |

Example:

|  |  |  |
| --- | --- | --- |
| A | A/J |  |
| B | C57BL/6J |  |
| C | 129S1/SvImJ |  |

#### coef\_[strain\_id].txt

The easiest way to import the coefficient data is to create one file per strain (N). Rows would be features (F) and columns would be markers (M).

**M**

**F**

**N**

***NOTE***: ALL values should be of type FLOAT, if there is no value, leave that row empty

**F** is the number of features

**M** is the number of markers

**N** in the number of strains

Example: If you have 2 strains with IDs of ‘A’ and ‘B’, you would have 2 files called ‘coef\_A.txt’ and ‘coef\_B.txt’. The reason for this format is for making the importing from the scripts easier.

**FOR THE FACTORIAL VIEWER (NOT REQUIRED)**

#### samples.txt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LAYOUT | | | | |
|  | **NAME** | **DATA TYPE** | **EXAMPLE** | **NOTE** |
| 1 | sample\_id | STRING | Mouse142 |  |
| 2 | name | STRING | Mouse 142 |  |
| 3 | description | STRING |  |  |

Example:

|  |  |  |
| --- | --- | --- |
| Mouse142 | Mouse 142 |  |
| 4sd56 | Mouse #1 | My favorite mouse |
| abds | ABDS | Unique mouse in the experiment |

#### factors.txt - factors (examples are sex, diet, tissue, etc)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LAYOUT | | | | |
|  | **NAME** | **DATA TYPE** | **EXAMPLE** | **NOTE** |
| 1 | factor\_id | STRING | tissue |  |
| 2 | Name | STRING | Tissue |  |
| 3 | description | STRING | The type of tissue |  |

Example:

|  |  |  |
| --- | --- | --- |
| Sex | Sex | M,F are present |
| Diet | Diet | Low fat, high fat, chow diets are used |

#### phenotypes.txt – phenotypes with samples (rows) and factors (columns)

**R**

**S**

**R** is the number of factors

**S** in the number of samples

Example: 2 factors (sex, diet), 6 samples

|  |  |
| --- | --- |
| F | LowFat |
| F | HighFat |
| F | Chow |
| M | Chow |
| M | HighFat |
| M | LowFat |

#### genotypes.txt – genotypes, markers (rows) and samples(columns)

**S**

**M**

**R** is the number of factors

**S** in the number of samples

Example: 6 samples, many markers (omitted for brevity)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CC | AA | BB | AB | CC | AB |
| CB | AA | CB |  | CB | BB |
| CC | AA | CB | AB | CC | BB |
| BC | AC | CB | AB | CC | AB |
| CC | AA | CB | AB | CC | BB |

#### expression.txt – expression values, features (rows) and samples(columns)

**S**

**F**

**F** is the number of features

**S** in the number of samples

Example: 6 samples, many features (omitted for brevity)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| -0.26623 | 1.7633 | 1.23753 | 0.21211 | -0.12122 | 0.78667 |
| 0.876322 | 1.12122 | 1.87212 |  | 1.32322 | 1.82162 |
| 1.98872 | 0.29903 | -0.672833 | 0.21642 | 1.24422 | -1.93722 |
| 0.23232 | -0.233223 | 0.323232 | 0.217652 | -1.232332 | 0.828821 |
| 1.7832786 | 1.872563 | -1.23222 | -0.81221 | 0.165212 | 1.34651 |

***NOTE***: ALL values should be of type FLOAT, if there is no value, leave that cell empty

### Step 2: Run the following Python script:

qtl\_viewer/utils/txt\_to\_hdf5.py

|  |  |  |  |
| --- | --- | --- | --- |
| PARAMETER | REQUIRED | DESCRIPTION | NOTE |
| -o, --out | Yes | Output file |  |
| -d, --dataset | Yes | Name of the dataset | Example: adipose |
| -f, --features | Yes | Name of features file |  |
| -l, --lod | Yes | Name of lod scores file |  |
| -m, --markers | Yes | Name of markers file |  |
| --dfA | Yes | Degrees of freedom autosomal |  |
| --dfX | Yes | Degrees of freedom, X chrom |  |
| -c, --csv | No | Use csv file instead of TSV | Not fully tested |
| --datasetname | No | The display name of the dataset | Adipose Tissue instead of adipose |
| --coef | No | Name of the ALL coefficient files | \*\_[strain\_id].txt works nice |
| --strains | No | Name of the strains file |  |
| --samples | No | Name of the samples file |  |
| --factors | No | Name of the factors file |  |
| --phenotypes | No | Name of phenotypes file |  |
| --genotypes | No | Name of genotypes file |  |
| --expression | No | Name of expression data file |  |

Example:

python qtl\_viewer/utils/txt\_to\_hdf5.py

--out proteomics.h5

--dataset proteomics\_do192

--datasetname “Proteomics 192”

--dfA 7

--dfX 14

--features proteomics\_features.txt

--lod proteomics\_lod.txt

--markers proteomics\_markers.txt

--strain proteomics\_strains.txt

--coef proteomics\_coef\_\*.txt

--factors proteomics\_factors.txt

--phenotypes proteomics\_phenotypes.txt

--genotypes proteomics\_genotypes.txt

--samples proteomics\_samples.txt

--expression proteomics\_expression.txt