PhenData: data retrieval from IMPC database

Natalja Kurbatova, Jeremy Mason

Modified: 08 September, 2014. Compiled: September 11, 2014

PhenData is an R package that allows easy access to the phenotyping data produced by the International Mouse Phenotyping Consortium (IMPC). For more information about the IMPC project, please visit `http://www.mousephenotype.org`.

The IMPC implements a standardized set of phenotyping protocols to generate data. These standardized protocols are defined in the International Mouse Phenotyping Resource of Standardised Screens, IMPReSS (`http://www.mousephenotype.org/impress`). Programmatically navigating the IMPC raw data APIs and the IMPReSS SOAP APIs can be technically challenging – PhenData makes the process easier for R users.

The intended users of the PhenData package are familiar with R and would like easy access to the IMPC data. The data retrieved from PhenData can be directly used by PhenStat – an R package that encapsulates the IMPC statistical pipeline, available at `http://www.bioconductor.org/packages/release/bioc/html/PhenStat.html`).

PhenData functions

The idea of PhenData is to systematically explore the IMPC dataset's multiple dimensions until the correct combination of filters has been selected and then download that data.

The PhenData functions can be divided into two logical groups: functions to retrieve lists of IMPC database objects, for example *getPhenCenters* and functions to obtain datasets from IMPC database, like *getIMPCDataset* and *getIMPCTable*.

"List" functions

There are "print" and "get" options for each one function in this group. "print" function prints the ID and name of the objects within a list. "get" function returns the list of IMPC IDs. There are following objects in IMPC database that can be obtained through appropriate functions:

- IMPC **phenotyping centers**: *getPhenCenters*, *printPhenCenters*.

```
> library(PhenData)
> getPhenCenters()
> printPhenCenters()
```

- IMPC **pipelines**: *getPipelines*, *printPipelines*. Both functions have two arguments: *PhenCenterName* and *excludeLegacyPipelines* to exclude legacy pipelines from the list with default value set to TRUE.

```
> library(PhenData)
> getPipelines("WTSI")
> printPipelines("WTSI")
```

- IMPC **procedures** (sometimes called **screens**) that are run for a specified phenotyping center and pipeline: *getProcedures*, *printProcedures*. Both functions have two arguments: *PhenCenterName* and *PipelineID*.

```
> library(PhenData)
> getProcedures("WTSI","MGP_001")
> printProcedures("WTSI","MGP_001")
```

- IMPC **parameters** that are measured within specified procedure for a pipeline run by phenotyping center: *getParameters*, *printParameters*. Functions have three arguments: *PhenCenterName*, *PipelineID* and *ProcedureID*.

```
> library(PhenData)
> getParameters("WTSI","MGP_001","IMPC_CBC_001")
> printParameters("WTSI","MGP_001","IMPC_CBC_001")
```

- Genetic backgrounds or, in other words, **strains** from which the knockout mice were derived for a specific combination of pipeline, procedure and parameter for a phenotyping center: *getStrains*, *printStrains*. Strain's ID is MGI (`http://www.informatics.jax.org/`) ID or temporary ID if the MGI is not assigned yet. There are following arguments for both functions: *PhenCenterName*, *PipelineID*, *ProcedureID* and *ParameterID*.

```
> library(PhenData)
> getStrains("WTSI","MGP_001","IMPC_CBC_001","IMPC_CBC_003_001")
> printStrains("WTSI","MGP_001","IMPC_CBC_001","IMPC_CBC_003_001")
```

- **Genes** that are reported for a specified combination of parameter, procedure, pipeline and phenotyping center: *getGenes*, *printGenes*. Gene's ID is MGI (`http://www.informatics.jax.org/`) ID or temporary ID if the MGI is not assigned yet. There are following arguments: *PhenCenterName*, *PipelineID*, *ProcedureID*, *ParameterID*, *StrainID* is an optional argument.

```
> library(PhenData)
> getGenes("WTSI","MGP_001","IMPC_CBC_001","IMPC_CBC_003_001")
> printGenes("WTSI","MGP_001","IMPC_CBC_001",
+ "IMPC_CBC_003_001","MGI:2159965")
```

- **Alleles** that are processed for a specified combination of parameter, procedure, pipeline and phenotyping center: *getAlleles*, *printAlleles*. Allele's ID is MGI (`http://www.informatics.jax.org/`) ID or temporary ID if the MGI is not assigned yet. There are following arguments: *PhenCenterName*, *PipelineID*, *ProcedureID*, *ParameterID* and *StrainID*, which is an optional argument.

```
> library(PhenData)
> getAlleles("WTSI","MGP_001","IMPC_CBC_001",
+ "IMPC_CBC_003_001","MGI:5446362")
> printAlleles("WTSI","MGP_001","IMPC_CBC_001","IMPC_CBC_003_001")
```

- **Zygosities** (homozygous, heterozygous and hemizygous) for mice that were measured for a gene/allele for a specified combination of parameter, procedure, pipeline and phenotyping center: *getZygosities*. There are following arguments of the *getZygosities* function: *PhenCenterName*, *PipelineID*, *ProcedureID*, *ParameterID*, *StrainID*, which is an optional argument, *GeneID*, which is an optional argument and finally *AlleleID*, which is also an optional argument.

```
> library(PhenData)
> getZygosities("WTSI","MGP_001","IMPC_CBC_001","IMPC_CBC_003_001",
+ StrainID="MGI:5446362",AlleleID="EUROALL:64")
```

"Dataset" functions

The PhenData package "dataset" functions allow to obtain IMPC datasets where potential phenodeviants data and control data are matched together according to the internal IMPC rules. There are two functions in this group: *getIMPCTable* and *getIMPCDataset*.

Function *getIMPCTable* creates a table with combination of objects to define IMPC datasets according to the parameters that have been passed to the function. Table is stored using comma separated format in the file specified by user. Function's arguments are:

- *fileName* – defines name of the file where to save resulting table with IMPC objects, mandatory argument with default value set to "PhenData_IMPC";
- *PhenCenterName* – IMPC phenotyping center;
- *PipelineID* – IMPC pipeline ID;
- *ProcedureID* – IMPC procedure ID;
- *ParameterID* – IMPC parameter ID;
- *AlleleID* – allele ID;
- *StrainID* – strain ID;
- *multipleFiles* – flag: "FALSE" value to get all records into one specified file; "TRUE" value (default) to split records across multiple files named starting with "fileName";
- *recordsPerFile* – number that specifies how many records to write into one file with default value set to 10000.

Example of usage:

```
> library(PhenData)
> getIMPCTable("./IMPCData.csv","WTSI","MGP_001","IMPC_CBC_001")
```

All possible combinations are stored now into the file "IMPCData.csv" in comma separated format. There are six columns in the saved file: "Phenotyping Center", "Pipeline", "Screen/Procedure", "Parameter", "Allele" and "Function to get IMPC Dataset". The last one column "Function to get IMPC Dataset" contains the prepared R code to call the function *getIMPCDataset* with appropriate parameters and to obtain the dataset.

For example, take the first row and the last column with dataset function call:

```
> library(PhenData)
> IMPC_dataset1 <- getIMPCDataset("WTSI","MGP_001","IMPC_CBC_001",
+ "IMPC_CBC_003_001","MGI:4431644")
```

Now IMPC dataset is obtained and ready to use with PhenStat, for example:

```
> library(PhenStat)
> testIMPC1 <- PhenList(dataset=IMPC_dataset1,
+ testGenotype="MDTZ",
+ refGenotype="+/+",
+ dataset.colname.genotype="Colony")
```