The *BIOSRutils* package

Facilitating integrated data analysis using *R*

Maarten van Iterson

Department of Molecular Epidemiology,
Leiden University Medical Center

September 20, 2016

BBMRI.nl
Biobanking and
BioMolecular resources
Research Infrastructure
The Netherlands

BBMRI.nl
Biobanking and
BioMolecular resources
Research Infrastructure
The Netherlands

Make the BIOS data collection ready-to-use from within *R*:

1. data easily and efficiently accessible
2. data preprocessed and quality controlled
3. easy linking between different data types and external annotations
4. make generation of the preprocessed data reproducible

BIOSRutils not a regular *R*-package

- on installation links to preprocessed datasets on the *virdir*
  - RNAseq datasets:

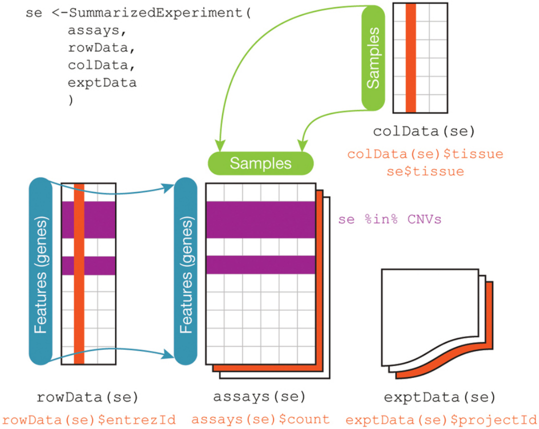
    containing exon/gene counts for both data freezes

    DNA methylation datasets:

    containing M- or beta-values per biobank or combined both data freezes
- provides a few helper-functions, e.g., querying the metadatabase
- workflows for generating the datasets
- example use cases, e.g., *How to run an epigenomewide association study (EWAS)*

datasets stored as a *Bioconductor SummarizedExperiment*

A comprehensive data structure for omics data [1]



```
se <-SummarizedExperiment(
      assays,
      rowData,
      colData,
      exptData
      )
```

[1] Huber, W. et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor.
Nat. Methods, 12(2):115–121

Preprocessing of the DNA methylation data

Input data:

- array-based DNA methylation measurements for 450k CpG's genomewide
- 6326 individuals across six biobanks
- 12652 raw data files (idat) with total size 96 GB

Output datasets:

containing M- or beta-values per biobank or combined for both data freezes

preprocessed and quality controlled

metadata and annotation

Preprocessing of the DNA methylation data

Steps involved:

1. reading of the data
2. sample level quality control and filtering[1]
3. probe level quality control and filtering
4. normalization and data transformation
5. sample identity checking
6. collecting metadata and annotation
7. construction of ready-to-use datasets

Several steps have been implemented in our *R*-package *Leiden450K* (`https://git.lumc.nl/molepi/Leiden450K`)

[1] van Iterson, M., Tobi, E. W., Slieker, R. C., den Hollander, W., Luijk, R., Slagboom, P. E., and Heijmans, B. T. (2014). MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics*, 30(23):3435–3437

BIOSRutils reproducible workflows a few examples

workflow:
`http://bios-vm.bbmrirp3-lumc.vm.surfsara.nl/BIOSRutils/PreparingDNAm.html`

interactive quality control apps:
`http://bios-vm.bbmrirp3-lumc.vm.surfsara.nl:8008/BIOSRutils/DNAm/LLS/`

sample identity checking:
`http://bios-vm.bbmrirp3-lumc.vm.surfsara.nl/BIOSRutils/DNAmSampleIdentityCheck.html`

DEMO: Use Case EWAS

Use Case: EWAS `http://bios-vm.bbmrirp3-lumc.vm.surfsara.nl/BIOSRutils/README.html#use-cases-ewas`

Future directions

- merge data sets in *MultiAssayExperiment*
- store large data using a HDF5-backend
- unify authentication of the different services e.g., VM, metadatabase, molgenis database access
- adding more use cases preferably by other users