# Facilitating BBMRI-omics downstream analysis using *R*

## The *BBMRIomics* package

Maarten van Iterson

Department of Molecular Epidemiology,
Leiden University Medical Center

March 29, 2017

# Making BBMRI-omics data ready-to-use:

1. data sets easily and efficiently accessible
2. data sets preprocessed and quality controlled
3. sample metadata and feature annotation should be provided
4. easy linking across BBMRI-omic data types and external genomic data/annotations (ENCODE, ROADMAP, etc.)

*It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data.*
`vita.had.co.nz/papers/tidy-data.pdf`

*Data scientists spend most of their time cleaning data.*
`whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/`

# *BBMRIomics* not a regular *R*-package

1. on installation links to preprocessed datasets
   - RNAseq datasets:

     gene counts per biobank or combined
   - DNA methylation datasets:

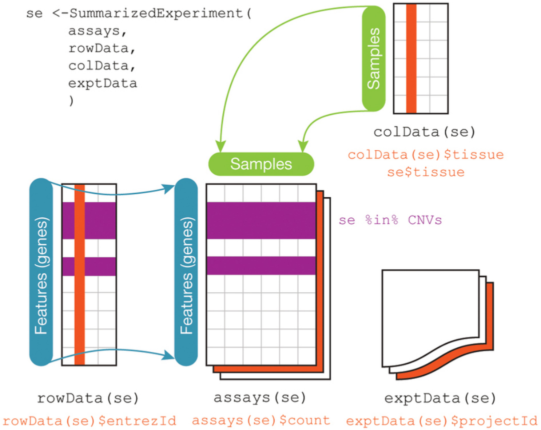
     M- or beta-values per biobank or combine
   - metabolomics data:

     overlap with BIOS
2. helper-functions, e.g., importing imputed genotype data
3. example use cases, e.g., *How to run an EWAS*
4. workflows for generating the datasets

`http://bios-vm.bbmrirp3-lumc.surf-hosted.nl/BBMRIomics`

# 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

# Advantages of *SummarizedExperiment*

1. reduces errors during reordering or subsetting operations:
   `se <- se[feature_index, sample_index])`
2. aid integrative analysis, i.e., matching experiments according to overlap of genomic regions:
   `hits <- findOverlaps(se, roi)`
3. easily extendable i.e., adding slots or use on disk storage (*HDF5Array* package)

## Data already available as *SummarizedExperiment*:

- recount (collection of RNA seq datasets, i.e. GEUVADIS):
  `https://jhubiostatistics.shinyapps.io/recount/`
- Expression Atlas (subset of EBI-EMBL ArrayExpress):
  `http://www.ebi.ac.uk/gxa/home`
- TCGAbiolinks (Access to The Cancer Genome Atlas (TCGA))
  `http://bioconductor.org/packages/TCGAbiolinks/`

# Preprocessing of the DNA methylation data

Input data:

- array-based DNA methylation measurements for 450k CpG's genomewide
- $>$ 4000 individuals across six biobanks
- $>$ 8000 raw data files (idat) with total size $\approx$ 100Gb

Output datasets:

containing M- or beta-values per biobank or combined

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 *DNAmArray* (`https://github.com/molepi/DNAmArray`)

[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

# Preprocessing of the RNA sequencing data

Input data:

- RNA from whole blood, Illumina HiSeq 2000, STAR aligner[1]
- $>$ 4000 individuals across six biobanks
- $>$ 8000 raw data files (fastq) with total size $\approx$ 10Tb
- $>$ 4000 bam-files ($\approx$ 0.5Tb)

Output datasets:

containing read counts per biobank or combined

preprocessed and quality controlled

metadata and annotation

[1] Zhernakova, D. et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.*, 49(1):139–145

# Preprocessing of the RNA sequencing data

Steps involved:

1. sample level quality control and filtering
   - rerun if total number of reads $< 15M$
2. sample identity checking[1]
3. collecting metadata and annotation
4. construction of ready-to-use datasets

[1] Westra, H. J., Jansen, R. C., Fehrmann, R. S., te Meerman, G. J., van Heel, D., Wijmenga, C., and Franke, L. (2011). MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics*, 27(15):2104–2111

# Current released datasets

A maximal set of unrelated individuals for which RNA sequencing and DNA methylation data could be generated

| | Pass QC | Unrelated |
|---|---|---|
| RNA | 4456 | 3560 |
| DNAm | 6121 | 4453 |
| Overlap | 4250 | 3435 |

- GoNL (trio's), twins, (unexpected family relations, replicates and longitudinal measurements)
- not all have (imputed)genotypes or a complete set of phenotypes available

| | Total | Overlap |
|---|---|---|
| Metabolomics (BRAINSHAKE) | 23729 | 3880 |

# Future

- More datasets i.e. RNA/DNAm specific for the GoNL-subset
- Update of all data to genome build GHRC38
- requests?