Chapter 6:

SATORi: An R package for generic access and handling of genomic data

Erik B. van den Akker[1,2], Marian Beekman[2,3], Joris Deelen[2,3], P. Eline Slagboom[2,3] and Marcel J.T. Reinders[1]

1. The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
2. Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
3. Netherlands Consortium of Healthy Ageing

In preparation

1. Summary

SATORI is an R package offering standardized access to various types of big genomic datasets, enabling a rapid exploration, integration and mapping between different data types as well as to external genomic annotations. The package, vignette along with the example datasets used in this paper can be obtained from: bioinformatics.tudelft.nl/users/erik-van-den-akker

2. Introduction

A joint interpretation of life-science data is required for grasping the etiology of complex traits, however, this is challenging as the data types and quantities are ever increasing. The statistical platform R[1] provides some excellent tools for mapping and complex modeling of genomic data[2], and additionally offers a potent interface to several database engines[3]. However it lacks a uniform database design across all data sources, which would greatly ease the data integration necessary to solve complex life science problems.

Here we present an R package, called SATORi (Standardized Access To Omics in R), for accessing various big omics data types from R in a standardized way. SATORI organizes data by genomic location facilitating an easy annotation to genomic features, like genes, pathways or to other SATORI databases. Moreover, due to the standardized database design, SATORI provides generic ways for accessing genome-wide data enabling a rapid exploration of omics data, while keeping source code clear and understandable.

As a (running) example we applied SATORI to analyze methylation[4] and SNP data[5] assayed on two HapMap populations consisting of 30 trios of Caucasian (CEU) and Yoruban (YRI) origin. We show how easy data can be accessed and integrated with SATORI, while still having the power of the function set in R.

3. Implementation

The SATORI package employs the R plugin RSQLite to organize and access data by genomic location or measured entity ID, and subsequently, represents the associated data in a standardized and suitable format for performing analyses in R. SATORI introduces **`GVARdb`** objects to represent omics sources to R and different types of data sources are stored in specifically inherited subclasses of the **`GVARdb`** object. SATORI currently supports genotype data (**`GWASdb`**), methylation data (**`METHdb`**), and imputed genotype data (**`impGWASdb`**). Purpose-built constructors are defined for parsing raw data files of each supported data type to build the tables constituting the SATORI database (Figure 1A).

3.1 Accessing data stored in SATORI databases

Data in GVARdb objects can be accessed using the getGVAR function, which allows for intuitive queries mimicking normal matrix manipulation in R. For instance, entity IDs and sample IDs can be used to specify the composition of the output matrices.

As an illustration, methylation data is stored as methylated and unmethylated signals per CpG site per sample (matrices M and U in Fig. 1B, respectively). A query using a specific CpG entity returns a vector containing the methylated and unmethylated signals for that CpG site for all samples:

```
1> res_vc <- getGVAR(methdb,
gvarID="cg00000292")
```

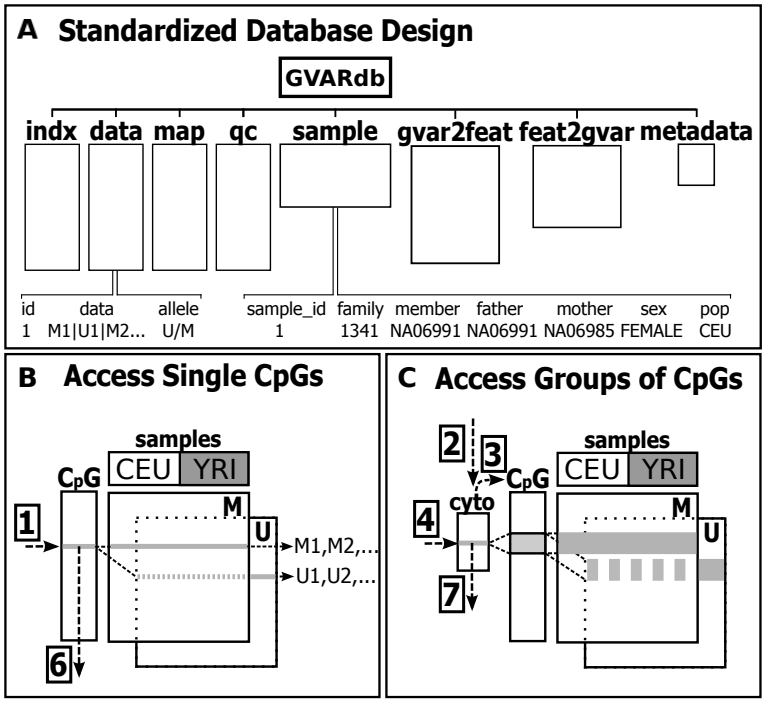

FIGURE 1: DESIGN AND FUNCTIONALITY OF SATORI. [A] An overview of the SQL tables composing a SATORI **GVARdb** object, including tables for storing data (data), mappings to the genome (map), mappings to genomic features (gvar2feat) and sample information (sample). Data points are collapsed per measured entity into single strings across all samples allowing a standardized storage and handling across data types (left). The sample table can be used to store for instance covariates or the familial relationships (right). **[B]** A database storing (for example) methylation data can be thought of as two big matrices containing the methylated (M) and unmethylated (U) signals for all measured CpGs for all samples. Using, for example, code snippet 1, one can access the data of a particular CpG, while with code snippet 6, one can find differentially methylated CpG sites. **[C]** Genomic features obtained from, for example, UCSC (cyto) are mapped to the database, after which data can be accessed by feature ID. See code snippets 2,3,4 on how to do this using SATORI.

Several wrappers for getGVAR have been defined to query data associated to particular (pre-defined) genomic features. For that, the SATORI database needs to be made aware of an entity-to-feature mapping (Fig. 1B), which can be created and stored using the `mapGVAR2FEAT` function.

For instance, suppose we want to access methylation data per cytoband. The genomic locations of cytobands can be downloaded from UCSC to R and mapped to the ***GVARdb*** *object using:*

```
2> cyto <- getFEATfromUCSC
("cytoBand","hg18")
3> mapGVAR2FEAT(methdb,cyto)
```

From now on, data can be queried using a specific cytoband designation, which would return a matrix containing the methylated and unmethylated signals for the CpG sites situated within the requested genomic feature for all samples (Fig. 1C):

```
4> res_mat <- getGVARbyFeat
(methdb,"chr1q21.3")
```

Conceptually, there is no difference between mapping to annotations downloaded from UCSC, to annotation databases in R storing transcript locations (Carlson, et al., 2013), or to other SATORI databases.

For example, to map SNP data stored in a SATORI's ***GWASdb*** *object to our methylation data within a 10kb region around the CpGs, also use:*

```
5> mapGVAR2FEAT
(methdb,gwasdb,flanking=10000)
```

From now on, data of SNPs positioned within 10kb from a CpG site can be queried with the same CpG IDs used to query the methylation data.

3.2 Operations on SATORI databases

Custom R functions can be applied to **GVARdb** objects using `dbGVARapply`, which applies a user-defined R function to each entry of the **GVARdb** object. This is done by sequentially loading and applying the function to chunks of the database (enabling parallel execution or handling big data sets).

As an illustration, differentially methylated CpG sites (DM-CpGs) between the two HapMap populations can be found by applying a Wilcoxon signed-rank test to the ***METHdb*** *object (Fig. 1B)*

```
6> res_wilcox <-
dbGVARapply(methdb,FUN=wilcox)
```

SATORI also allows applying functions to groups of entries using `dbFEATapply`. This function makes use of a previously stored mapping to iteratively load and analyze grouped data associated to genomic features.

For example, to investigate whether the previously identified DM-CpGs are confined to certain cytobands, we re-compute the number of DM-CpGs per region given a p-value threshold and report these as percentages per cytoband (Fig 1C):

```
7> res_p <- dbFEATapply
(methdb,FUN=perc,p=1e-5)
```

4 Application

Previous code illustrated the ease with which methylation and SNP data could be integrated using the SATORI package. To illustrate the seamless integration in R we investigate the relationship between CpGs and their neighboring SNPs within the CEU and YRI population by calculating CpG-SNP correlations.

To perform this computation, we define a function that queries the data from a ***METHdb*** *and a* ***GWASdb*** *given a CpG ID, and computes CpG-SNP pair correlations using all shared samples:*

```
# Find shared samples:
8> sampID <- intersect(colnames(methdb),
colnames(gwasdb))
# Get SNP data, for instance cpgID =
"cg00000292":
9> snp <- getGVARbyFeat
(gwasdb,cpgID,sampID)[[1]]
```

```
# Get the beta values (M/
(M+U+100)) from the methylated
data. This indicated to getGVAR
by the "as.B=T" argument
10> MR <- getGVAR
(methdb,cpgID,sampID,as.B=T)[1,]
# Compute (with pairwise.
complete observations)
11> cor(t(snp),MR,use="p")
```

Genome plots of chr21q22.11 (Figure 2) show that DM-CpGs coincide with significant CpG-SNP correlations in YRI, suggesting a potential crosstalk between the two data types on this location.

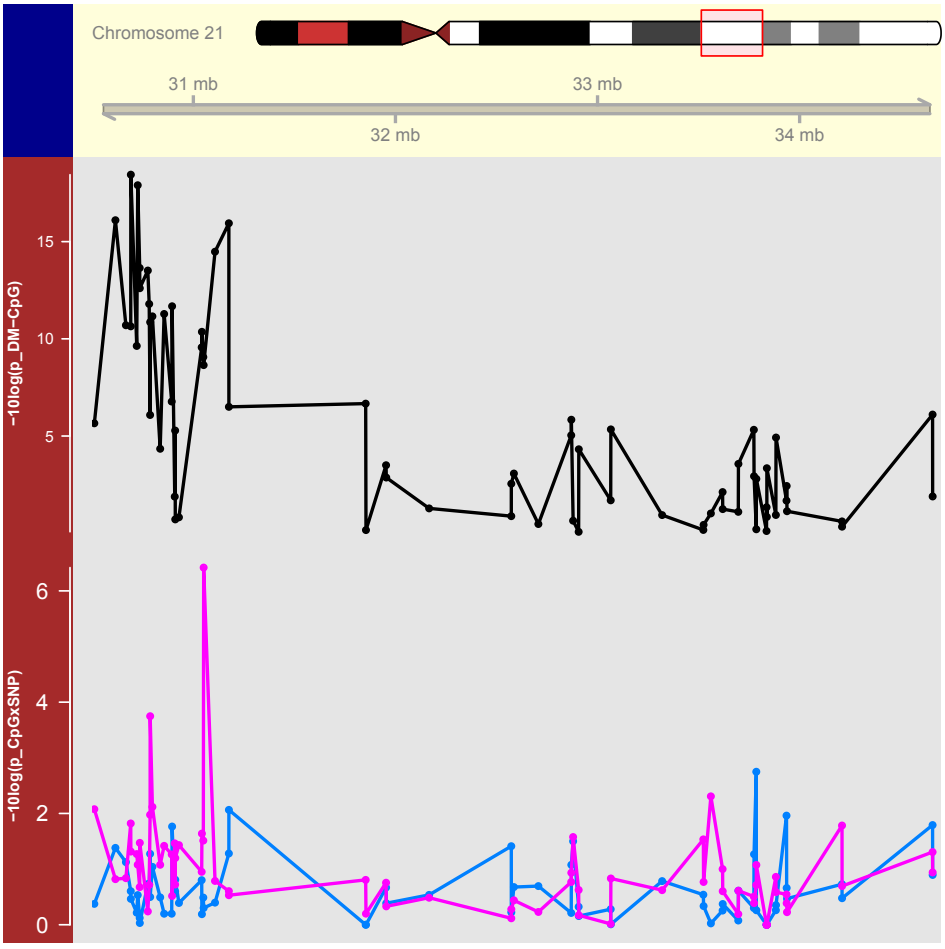

FIGURE 2: A VISUALIZATION OF THE RESULTS COMPUTED WITH AID OF SATORI ON CHR21Q22.11. In black, the significance (-10log(p)) and relative positioning of DM-CpGs between the YRI and CEU populations are depicted. Below in pink (YRI) and blue (CEU) correlations between CpG levels and genotypes in cis are displayed, where each point represents the most significant correlation (-10log(p)) with a SNP within 10kb. Detailed code for performing the analyses and drawing genome plots is provided in the supplemental materials.

5. Conclusions

With the examples in this paper we illustrated the merits of a generic access to various big data sources, which greatly simplifies integrative analyses of life science data. With SATORI the user can quickly explore data and test novel hypotheses by mapping the data to external annotations sources or applying user-defined functions to the data. Hence SATORI is a useful tool in the integrative analysis of omics datasets.

6. Acknowledgements

Funding: The research leading to these results has received funding from the Medical Delta (COMO) and the European Union's Seventh Framework Programme (FP7/2007-2011) IDEAL-ageing under grant agreement n° 259679. This study was supported by a grant from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology, the Netherlands Consortium for Healthy Ageing (Grant 050-060-810), all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO) and by Unilever Colworth.

7. References

1. R-Core-Team. R: A Language and Environment for Statistical Computing. (2013).
2. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
3. D. A. James & Falcon, S. RSQLite: SQLite interface for R. (2013).
4. Fraser, H.B., Lam, L.L., Neumann, S.M. & Kobor, M.S. Population-specificity of human DNA methylation. *Genome Biol* **13**, R8 (2012).
5. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-96 (2003).