The R package **bigstatsr**: Memory- and Computation-Efficient Tools for Big Matrices

useR!2017 lightning talk

Florian Privé (@privefl)

July 6, 2017

About

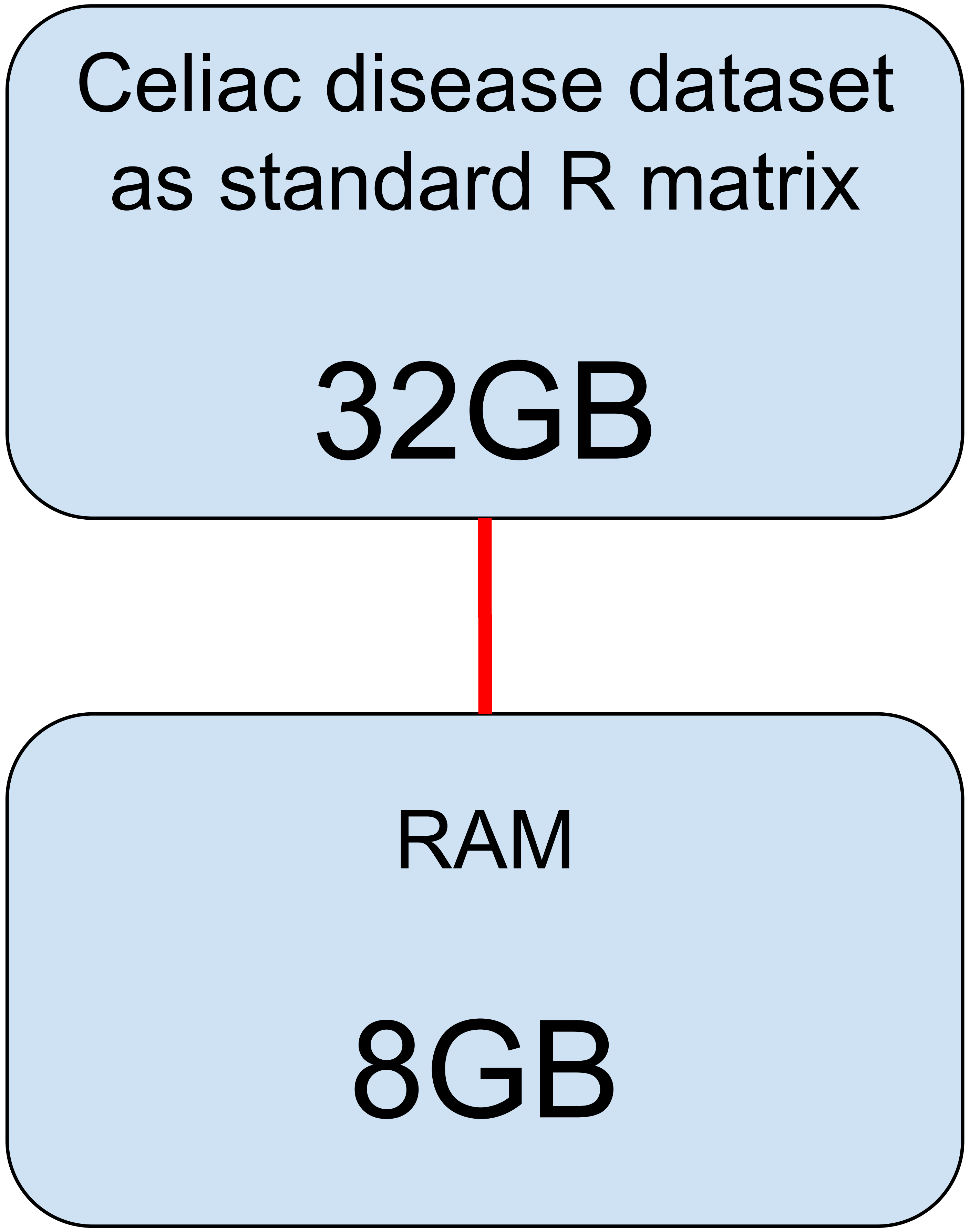
I'm a PhD Student (2016-2019) in **Predictive Human Genetics** in Grenoble.

$\boxed{\text{Disease} \sim \text{DNA mutations}}$

Very large genotype matrices

- currently: 15K x 300K, celiac disease
- soon: 500K x 800K, UK Biobank

Problem I had



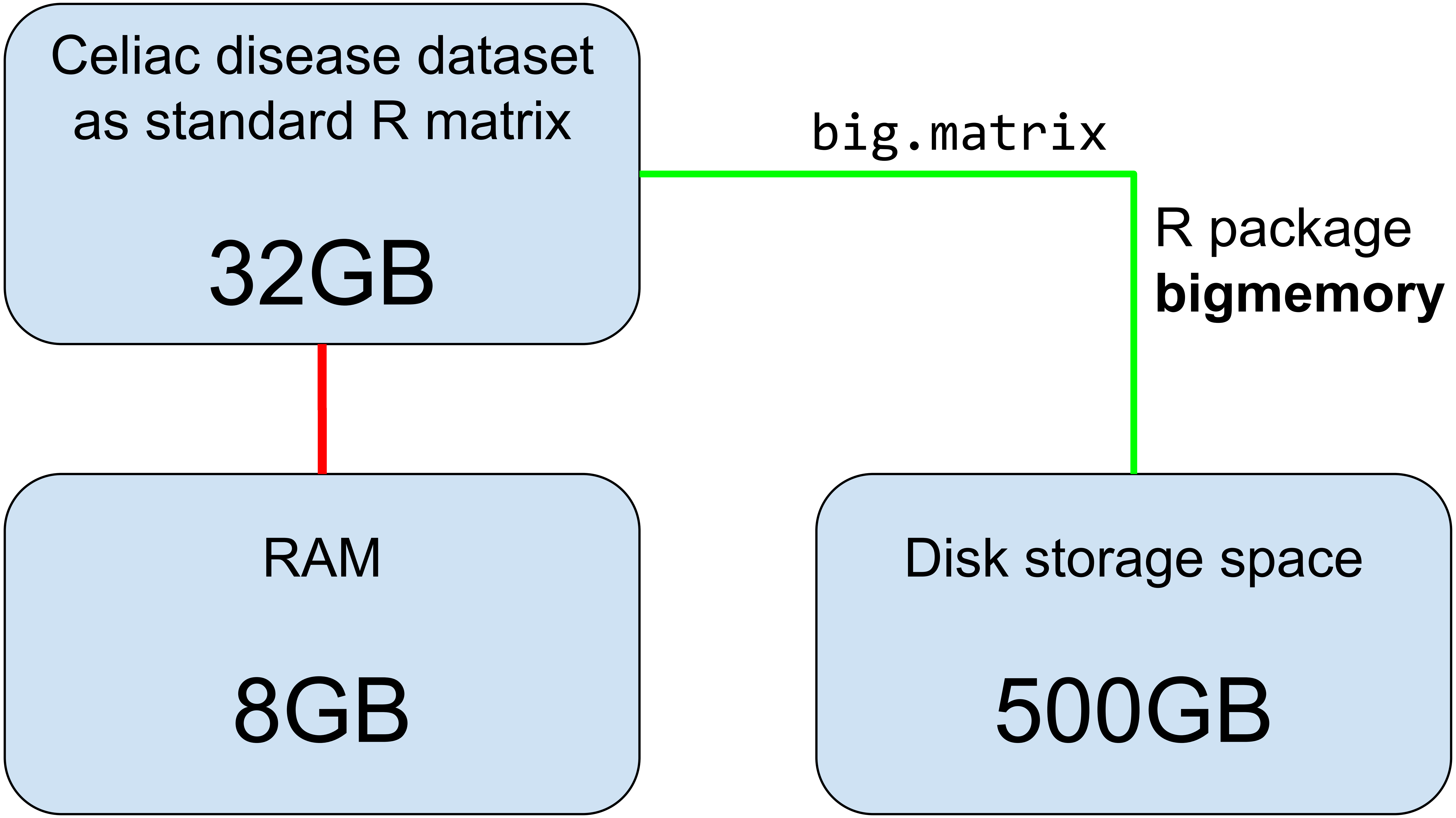
Celiac disease dataset as standard R matrix

32GB

RAM

8GB

Solution I found



Celiac disease dataset as standard R matrix

32GB

big.matrix

R package **bigmemory**

RAM

8GB

Disk storage space

500GB

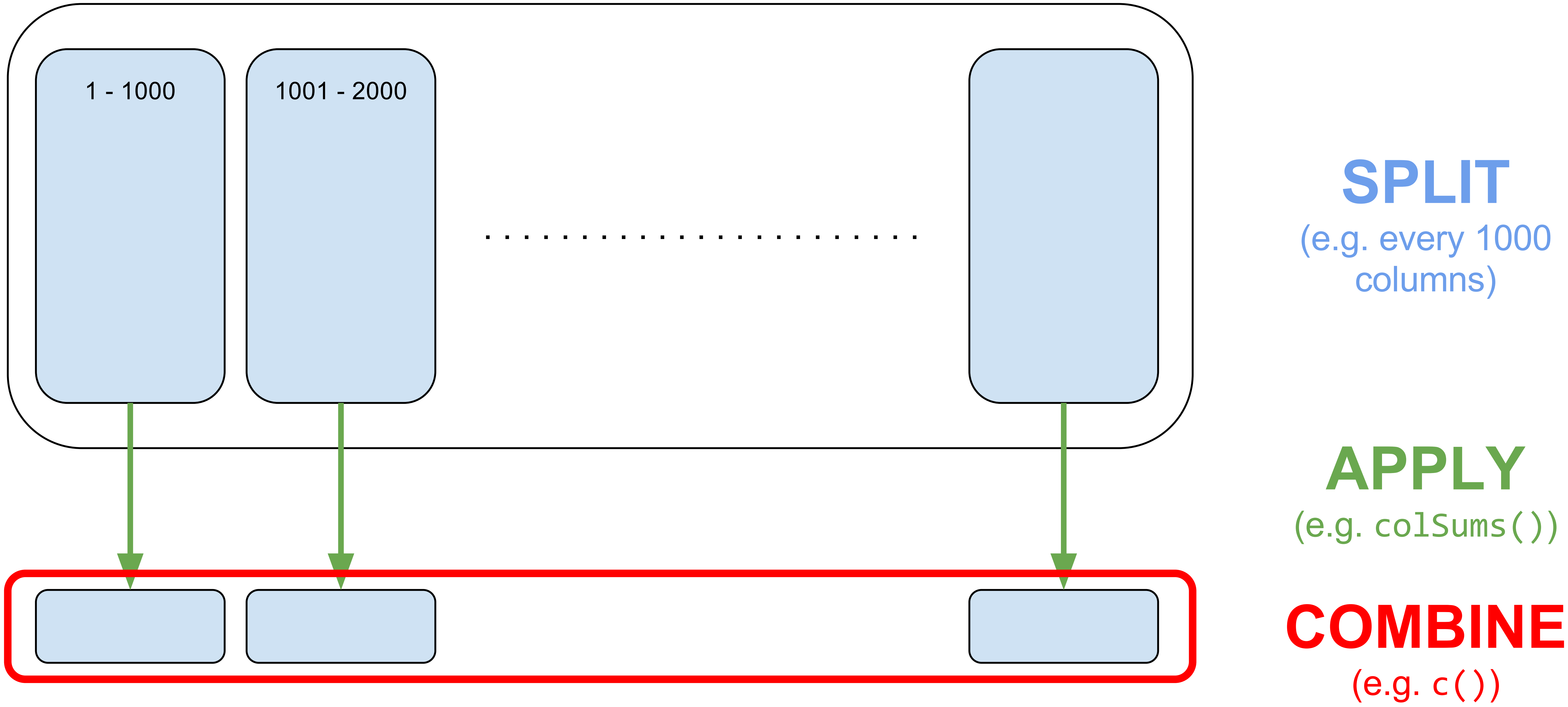
Michael J. Kane, John Emerson, Stephen Weston (2013).

Accesses almost as if the matrix were in memory

- in **R**: accesses with `[`, as standard R matrices,
- in **Rcpp**: accesses single elements with `X(i, j)`, as standard Rcpp matrices.

Split-(par)Apply-Combine Strategy

Apply standard R functions to big matrices (in parallel)



strategy coined by Hadley Wickham (2011)

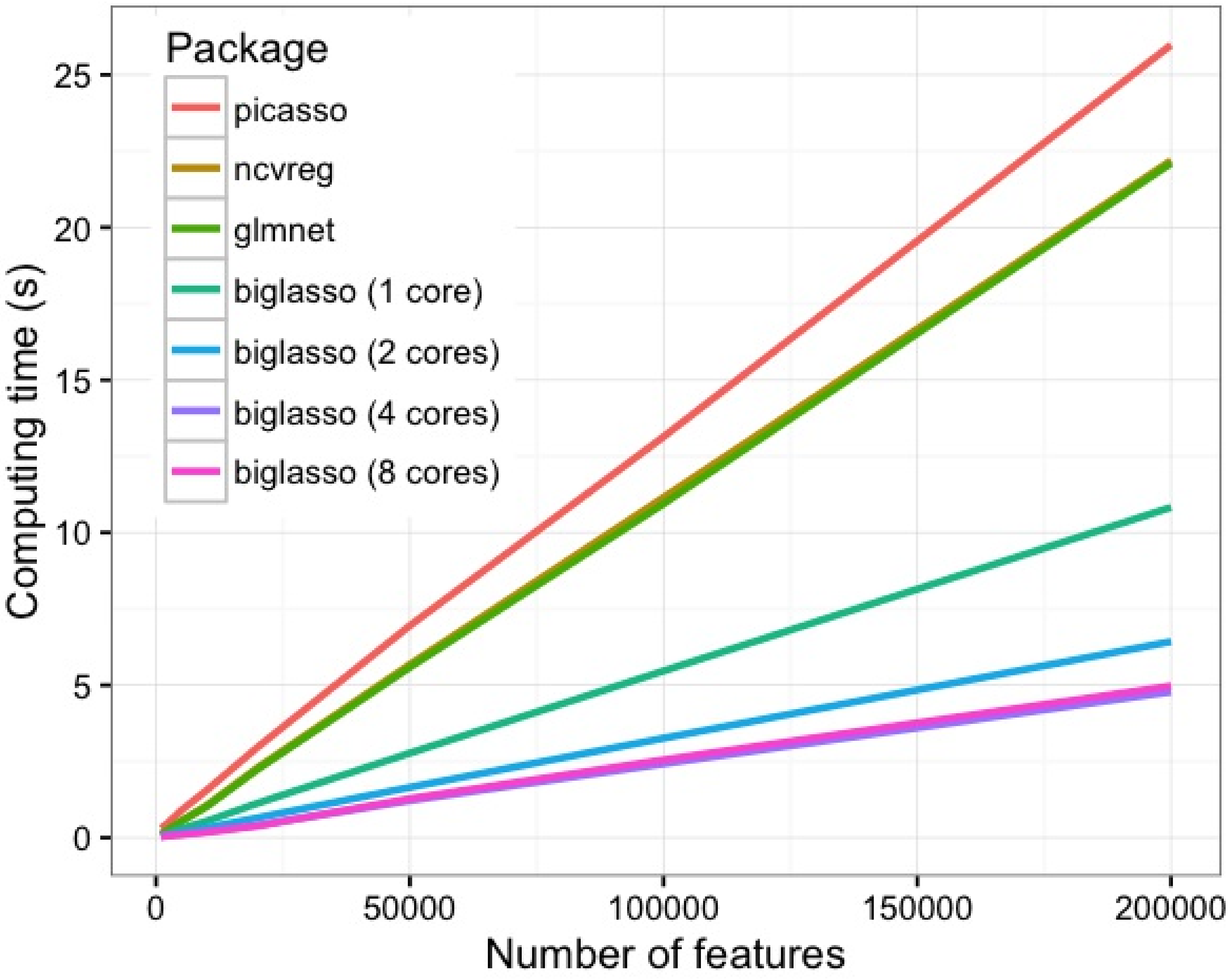
Matrix operations

- (cross-)products with matrices/vectors
- special tricks for handling scaling (vignette and blog post)

Example: computation of correlation of a 100,000 x 5000 matrix

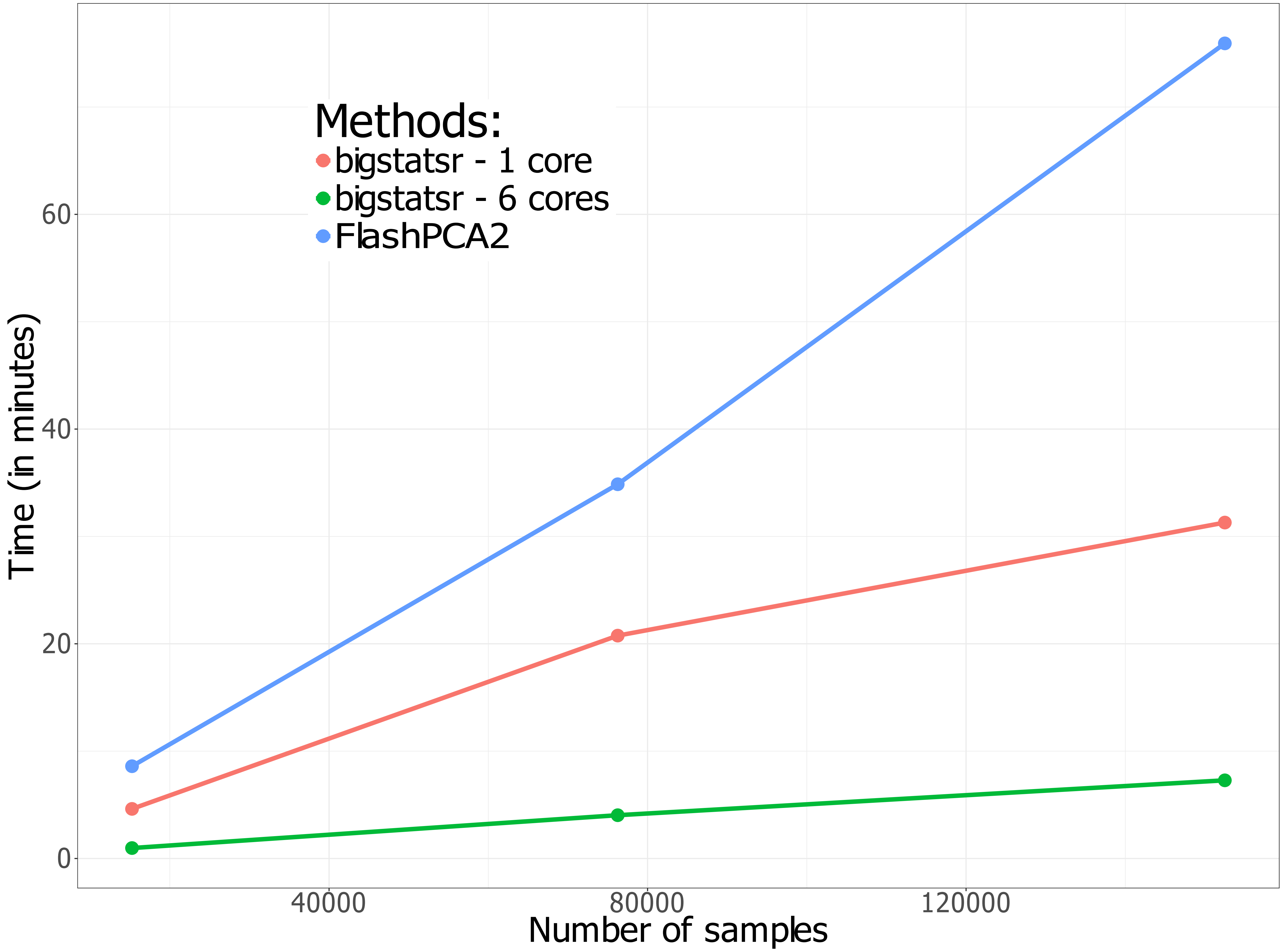
- `cor`: 22 minutes
- `big_cor`: 1 minute

Sparse linear models: **biglasso**



Zeng, Y., and Breheny, P. (2017).

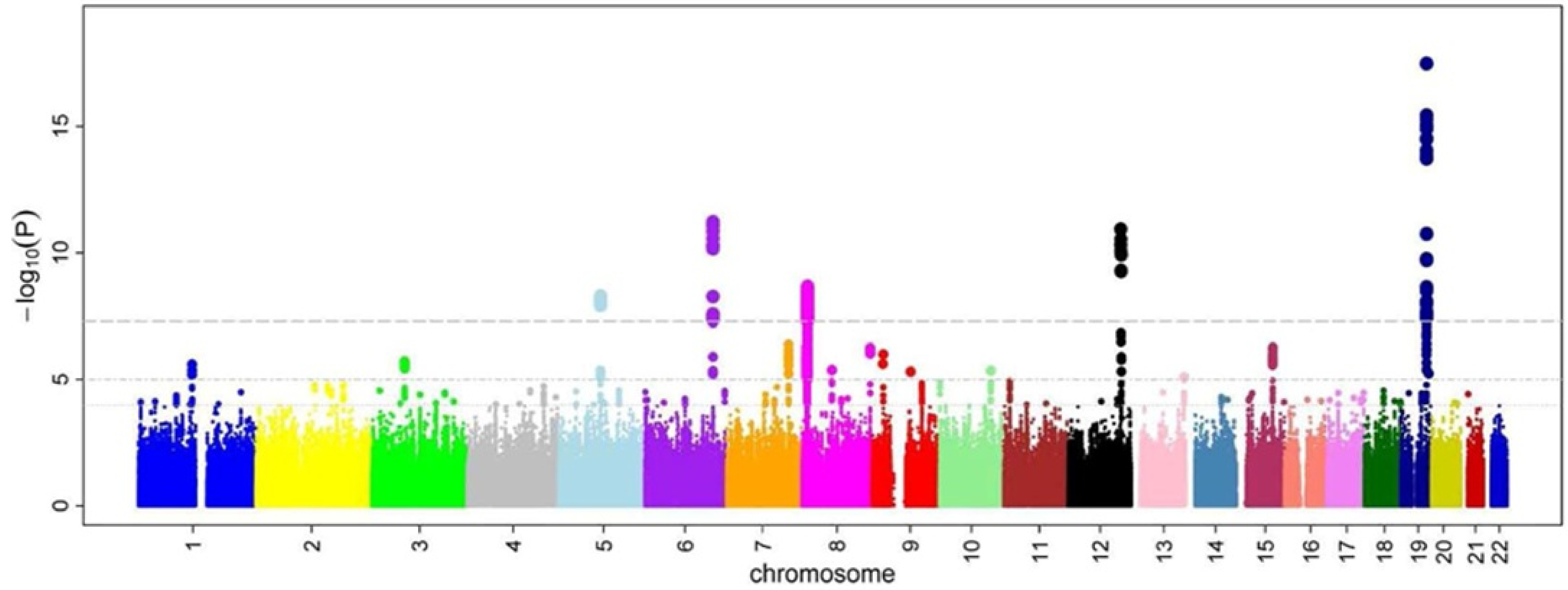
Partial Singular Value Decomposition



based on R package **RSpectra**

Test association of each variable with an outcome

In genetics, this is called a Genome-Wide Association Study (GWAS)



I'm now able
to run algorithms
on 100GB of data

R Packages

bigmemory	`big.matrix` object
bigstatsr	Statistical functions for `big.matrix` objects **to be used by any field**
bigsnpr	Specific functions for SNP arrays

Paper in preparation: "Efficient management and analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr".

Contributors are welcomed!

Thanks!

Package's website: https://privefl.github.io/bigstatsr/

Twitter and GitHub: @privefl

Presentation available online: https://goo.gl/nNg0hw

Slides created via the R package **xaringan**.