Introduction to SPRINT

Parallel Computing with R using SPRINT
on post-genomic data

What is **SPRINT**

Simple **P**arallel **R** **INT**erface

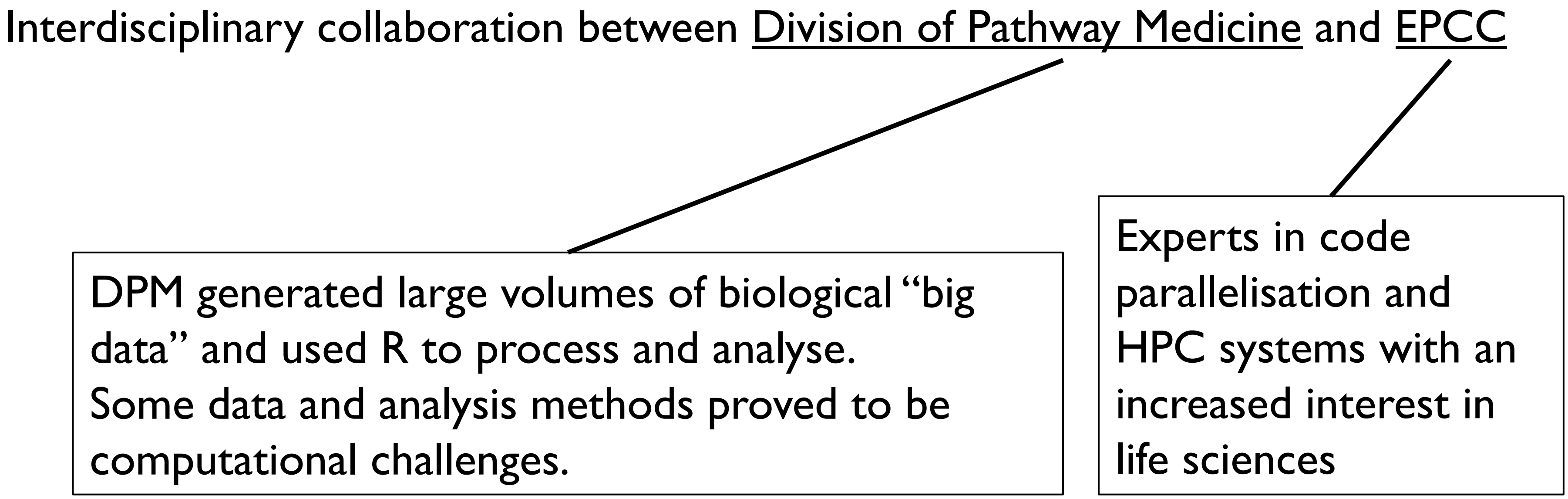
R package (www.r-sprint.org, R-Forge, CRAN?)

A collection of R functions we've already parallelised

No HPC programming experience required for use

Not a commercial product…

Some historical context



Interdisciplinary collaboration between Division of Pathway Medicine and EPCC

DPM generated large volumes of biological "big data" and used R to process and analyse.
Some data and analysis methods proved to be computational challenges.

Experts in code parallelisation and HPC systems with an increased interest in life sciences

Funded as research since ~2004 in order to make HPC accessible to R's post-genomic users

Why use SPRINT (or other parallelisation solutions)

Problems and limitations with R when used with “big data”

- Analysis takes too long (CPU limitations)

- Analysis fails due to memory usage (RAM limitations)

Don’t use SPRINT (or other parallelisation solutions)

- Reduce data size

  criteria for this are not driven by biology

- Process in batches

  doesn’t work well if entities are not independent, and doesn’t fix CPU time issues

- Choose alternative analyses

  compromise between analysis strategy and performance

- Do it outside of R

  takes time to implement, may not be helpful for automated workflows, requires expertise

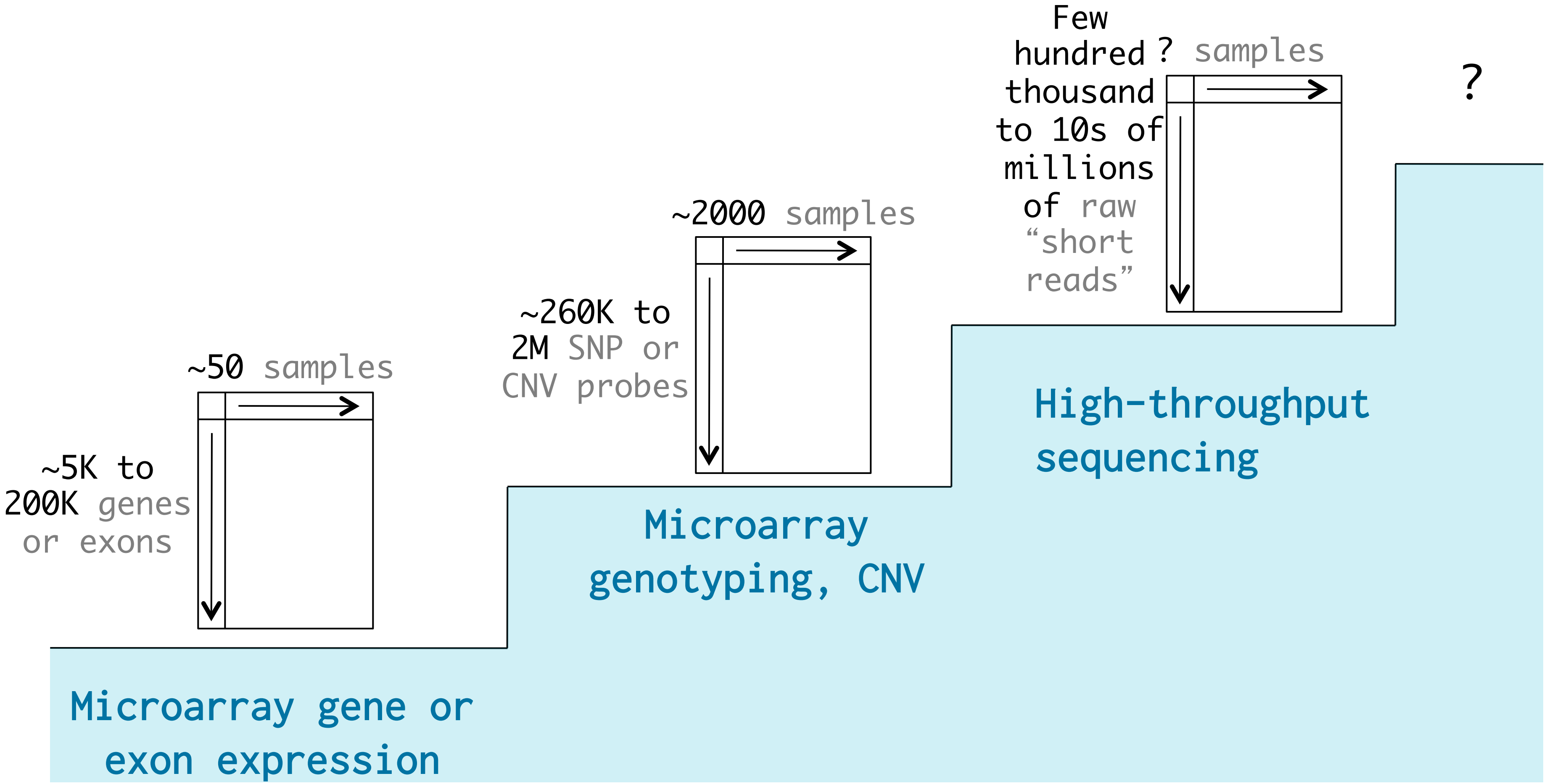
R user survey

For results, see www.r-sprint.org -> Documents

The most frequently mentioned problems were:

- apply() operations
- Correlation computations
- Linear models
- Bootstrapping
- Random Forest classification
- Matrix operations
- Bayesian inference (especially MCMC samplers)

Example issues – data type



~50 samples

~5K to 200K genes or exons

Microarray gene or exon expression

~2000 samples

~260K to 2M SNP or CNV probes

Microarray genotyping, CNV

? samples

Few hundred thousand to 10s of millions of raw “short reads”

High-throughput sequencing

?

Example issues –analysis/processing

Measuring distances or similarity between all possible pairs of genes (or exons, SNPs, short read sequences)

Resampling strategies (permutations, bootstrapping, MCMC)

Simulation, optimisation

Data processing

Example issues – parallelisation strategy

By parallelisation approach:

1. **Task farming**

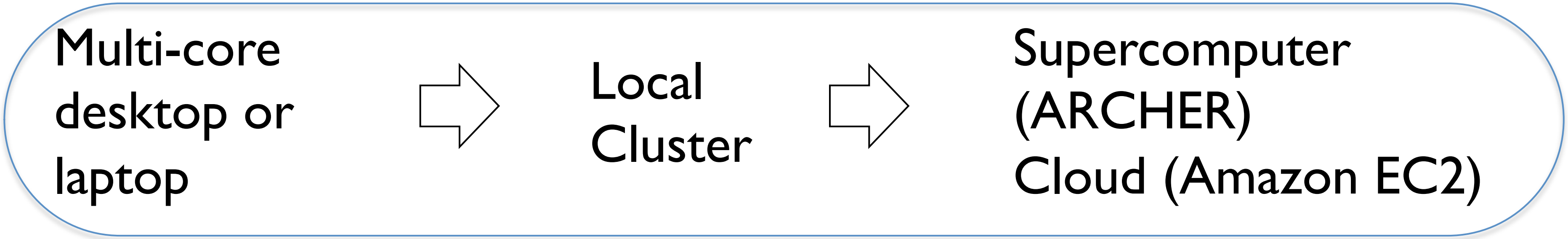
Large numbers of independent R tasks that can be parceled out to individual processors without any problem, e.g. papply(), pboot(), prandomForest()

2. Dependent-data

These are more difficult to implement, as each processor needs to know what part o the work has already been done by another processor. E.g. pcor(), pstringdistmatrix()

Why use SPRINT (or other parallelisation solutions)

Can we use High Performance Computing?



Multi-core desktop or laptop

Local Cluster

Supercomputer (ARCHER)
Cloud (Amazon EC2)

Carefully designed parallel apps - scalable

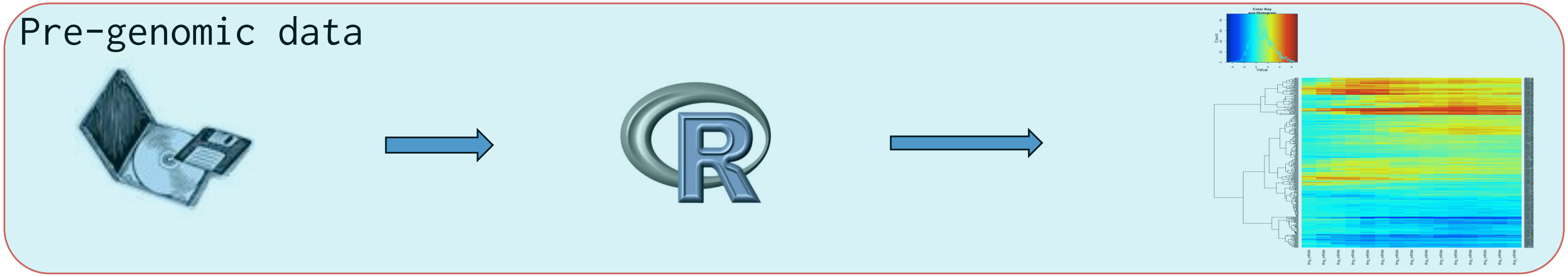
More CPUs/ memory - Faster analyses
Larger datasets

SPRINT

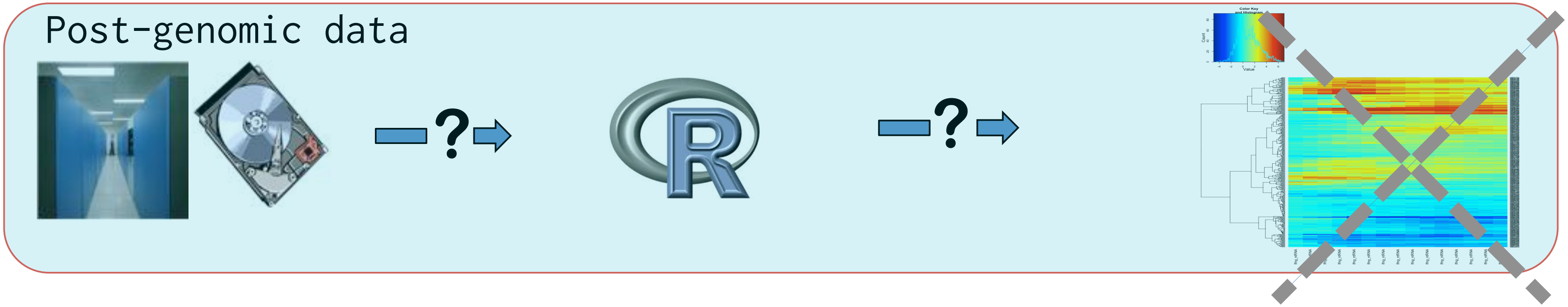
Therefore, we planned SPRINT to

- provide easy access to parallelised functions
- be Open Source
- be scalable and tackle both CPU and RAM problems
- tackle complex dependent-data problems

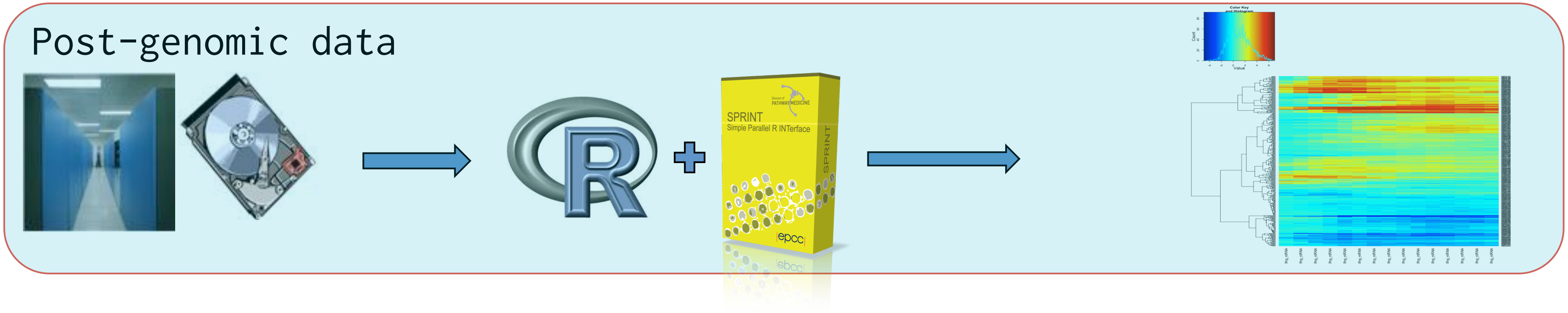
Or simply put...



Pre-genomic data



Post-genomic data



Post-genomic data

Other solutions

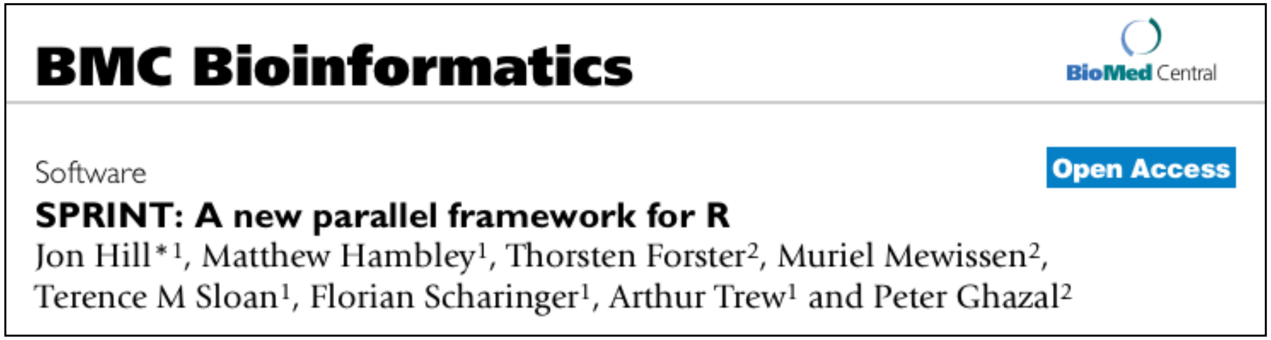
See Dirk Eddelbuettel's pretty comprehensive list*:

http://cran.r-project.org/web/views/HighPerformanceComputing.html

R now directly supports parallelisation with package "parallel", and if you can do your own system administration and code implementation, this and other packages can be more flexible than SPRINT

*SPRINT is not currently on there, because CRAN requires support of the OpenMPI (rather than MPICH) standard…we'll have that sorted out before summer 2014.

SPRINT Publications



BMC Bioinformatics

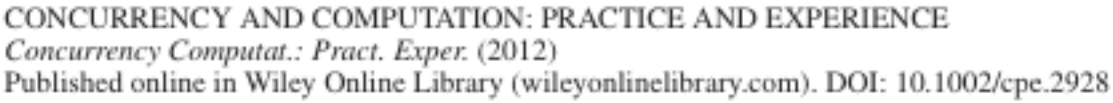
BioMed Central

Software

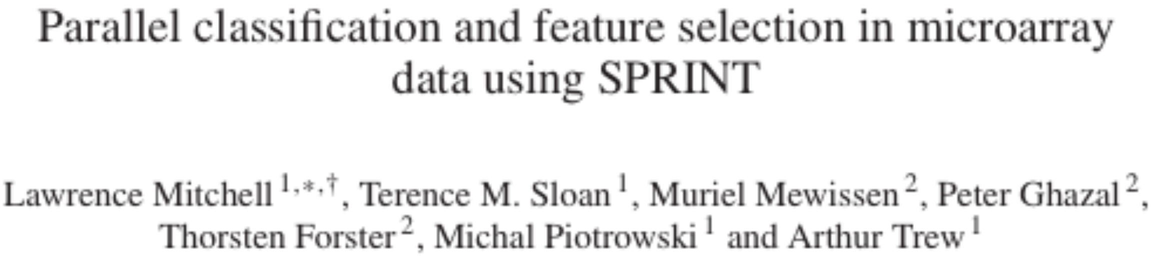
Open Access

SPRINT: A new parallel framework for R

Jon Hill*[1], Matthew Hambley[1], Thorsten Forster[2], Muriel Mewissen[2], Terence M Sloan[1], Florian Scharinger[1], Arthur Trew[1] and Peter Ghazal[2]



CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. (2012)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.2928



Parallel classification and feature selection in microarray data using SPRINT

Lawrence Mitchell[1,*,†], Terence M. Sloan[1], Muriel Mewissen[2], Peter Ghazal[2], Thorsten Forster[2], Michal Piotrowski[1] and Arthur Trew[1]

…more at r-sprint.org

SPRINT Now

Applications

- Integrated analyses of merged health data (next-gen sequencing, clinical data)
- Clinician-ready tool for classification-biomarkers
- Loosely support third-party use of SPRINT
- No longer able to be reactive to R-community posed problems…rarely fundable

Access

- Local installations (EPCC clusters, CRUK, HPC Wales,…)
- Central installations (ARCHER, previously HECToR)
- Personal installations (multicore deskotps and laptops)

Availability

- Linux (*nix).
- Apple OSX

Dissemination

Talks, training courses, workshops etc.

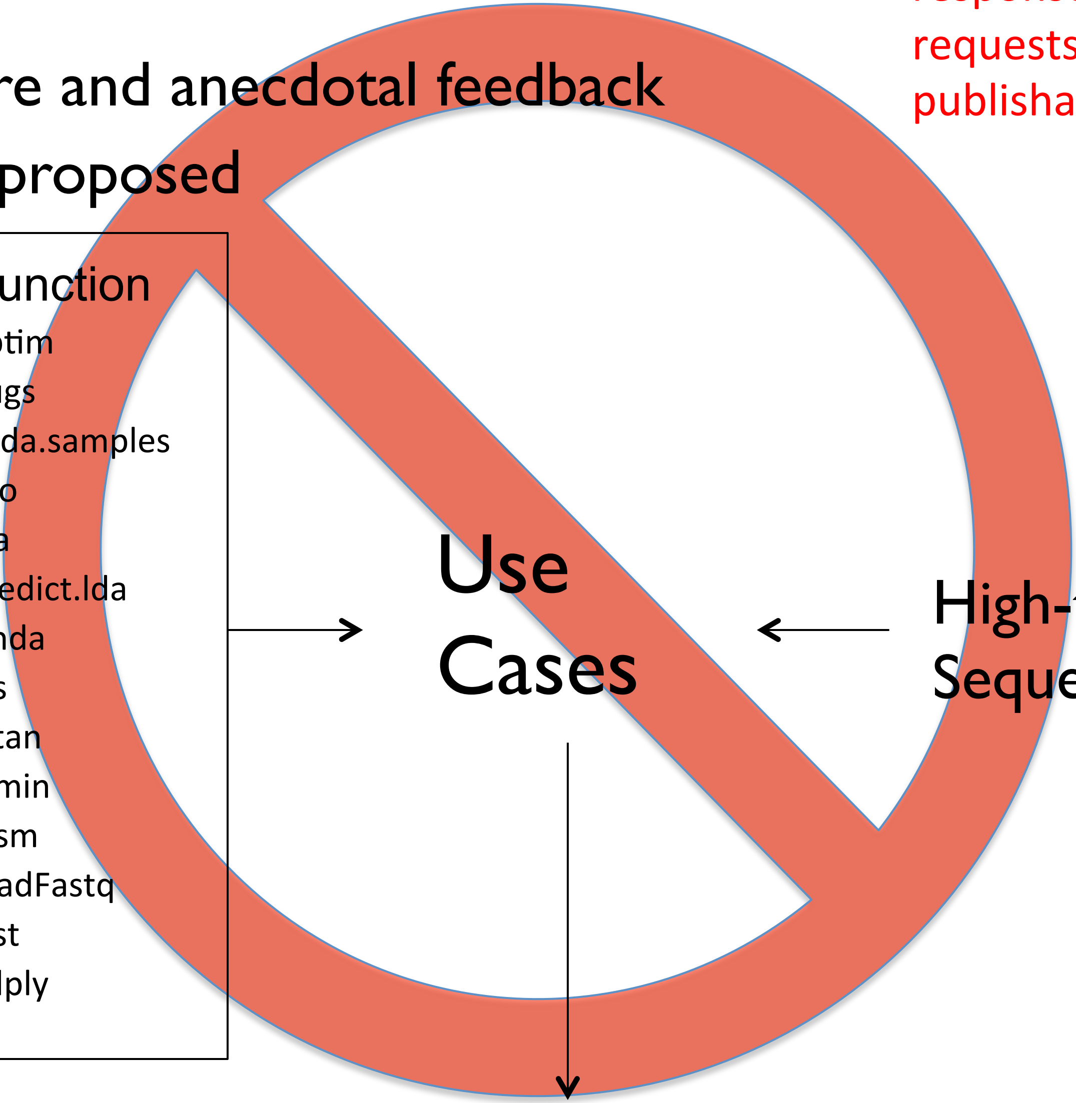
SPRINT Now

No longer able to operate in response to crowd-sourced requests…usually lack of publishable use case or data.

Questionnaire and anecdotal feedback

37 functions proposed

Package	Function
Stats	optim
R2OpenBugs	bugs
rjags	coda.samples
VGAM	cao
MASS	lda
MASS	predict.lda
rrcov	Linda
GenABEL	ibs
	rstan
	nlmin
	msm
ShortRead	readFastq
	dist
plyr	ddply



Use Cases

High-throughput Sequence analysis

Development Time ~ 3 Months/ function

All of SPRINT

EPCC

- Eilidh Troup
- Luis Cebamanos
- Terence Sloan (PI)

DPM

- Thorsten Forster
  - Peter Ghazal (PI)

Former Contributors and Funders

Muriel Mewissen
Savaas Petrou
Michal Piotrowski
Jon Hill
Florian Scharinger
Laurence Baldwin
Bartek Dobrzelecki
Lawrence Mitchell
Kevin Robertson
Andy Turner

BBSRC
bioscience for the future

wellcome trust

www.software.ac.uk

nag

edikt
e-science data information
& knowledge transformation

PRACE

David Henty
Irina Nazarova
Catherine Inglis

How to interact with the SPRINT Team

r-sprint.org

sprint@ed.ac.uk

- Installation queries
- Bug reporting
- Use-case queries

https://r-forge.r-project.org/projects/sprint/

- Function development
- Bug Fixes

2014

Parallel Optimisation of Bootstrapping in R. Sloan TM, Piotrowski M, Forster T, Ghazal P. arXiv.org pre-publication January 2014.

2013

Embedded systems for global e-Social Science: Moving computation rather than data. Lloyd A. et al. 2013. Future Generation Computer Systems Volume 29, Issue 5, July 2013.

Exploiting Parallel R in the Cloud with SPRINT. Piotrowski M. et al. Methods Inf Med. 2013

2012

Parallel classification and feature selection in microarray data using SPRINT. Mitchell L. et al. 2012. Concurrency and Computation: Practice and Experience.

2011

Optimisation and parallelisation of the partitioning around medoids function in R. Piotrowksi M. et al. BILIS 2011, Jul 2011.

Optimization of a parallel permutation testing function for the SPRINT R package, S. Petrou et al, Concurrency and Computation: Practice and Experience, Jun 2011.

A parallel random forest classifier for R, L. Mitchell et al, HPDC 2011, Jun 2011.

Managing and Analysing Genomic Data using HPC and Clouds, B. Dobrzelecki et al, book chapter in Grid and Cloud Database Management, G. Aloisio & S. Fiore, Springer, 2011.

2010

"SPRINT: a Simple Parallel INTerface to High Performance Computing and a Parallel R Function Library", M. Mewissen et al., useR! The R User Conference 2010, pp 104, R Foundation for Statistical Computing.

Optimization of a parallel permutation testing function for the SPRINT R Package, S. Petrou et al, HPDC 2010 Proceedings, Jun 2010.

2008

SPRINT: A new parallel framework for R, J. Hill et al, BMC Bioinformatics, Dec 2008.