SPRINT:

a Simple Parallel R INTerface to High Performance Computing (HPC) and a Parallel R function Library

Muriel Mewissen
Division of Pathway Medicine, University of Edinburgh, UK

useR!2010, Gaithersburgh, 21st July 2010

Talk Outline

- Motivation: Pathway Biology
- High-throughput technologies, HPC & R
- SPRINT:
  - Functionality and Releases
  - Architecture
  - Performance
- Future work

Division of PATHWAY MEDICINE

epcc

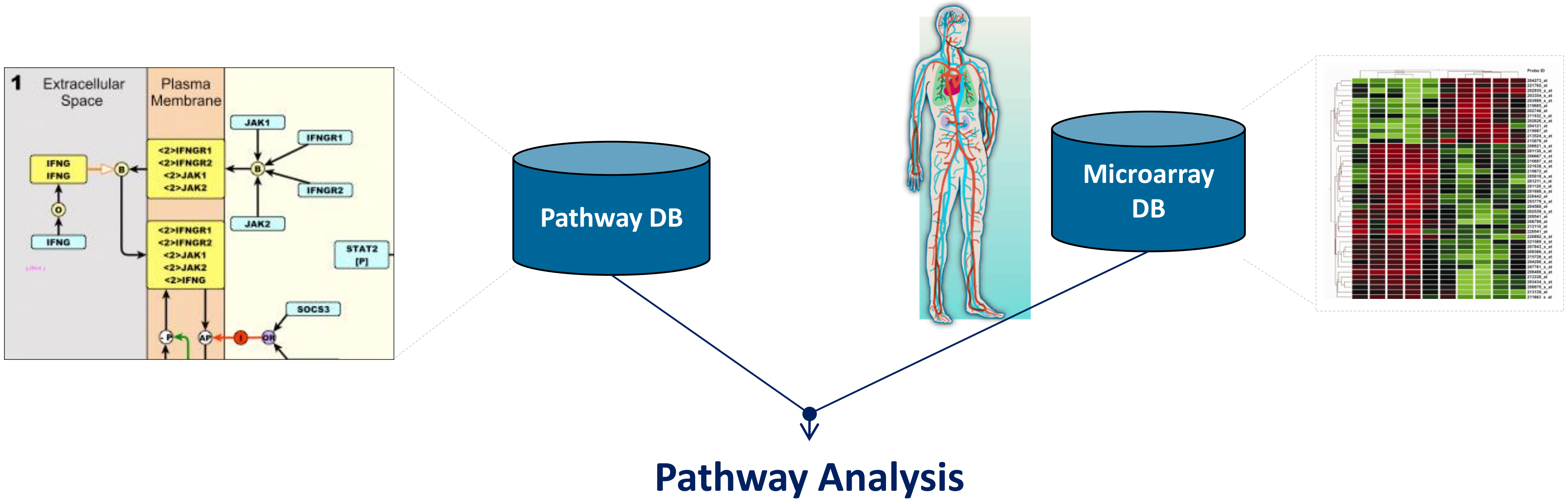
What is Pathway Biology?

Pathway biology is….

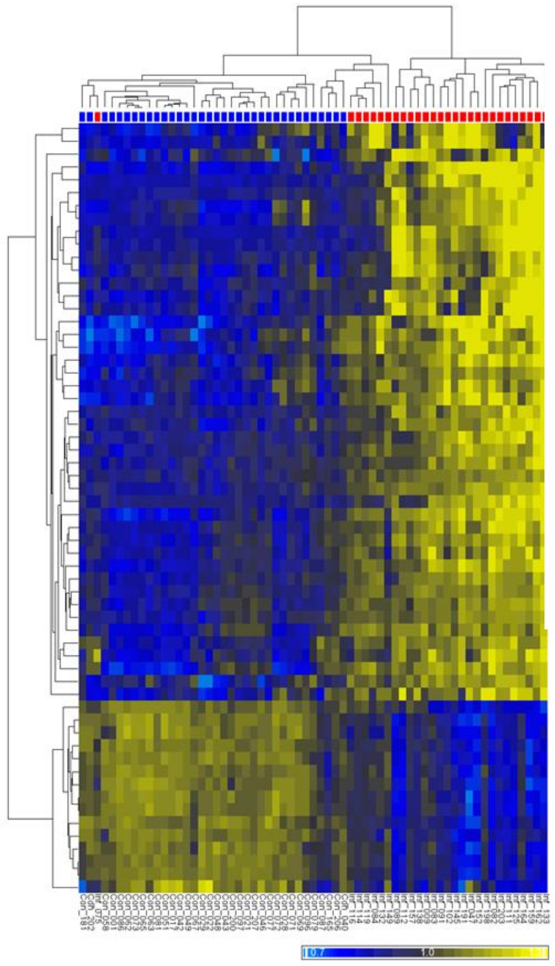
A systems biology approach for understanding a biological process

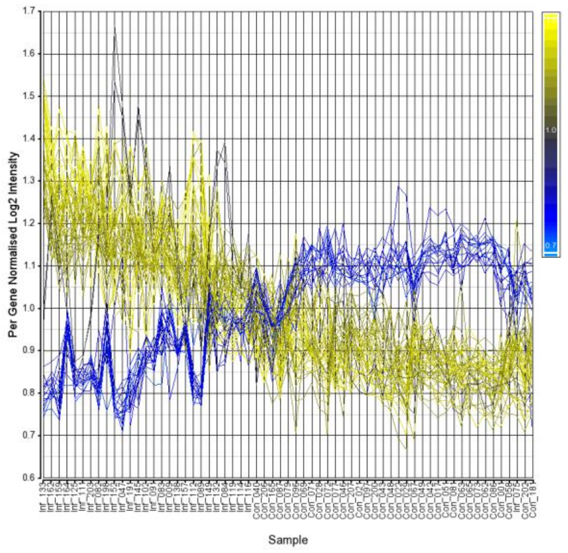
- empirically by functional association of multiple gene products & metabolites
- computationally by defining networks of cause-effect relationships.

➔ Pathway Models link molecular; cellular; whole organism levels.



Differentially expressed genes in neonates control vs Infected (FDR $p>1\times10^{-5}$, FC±4)

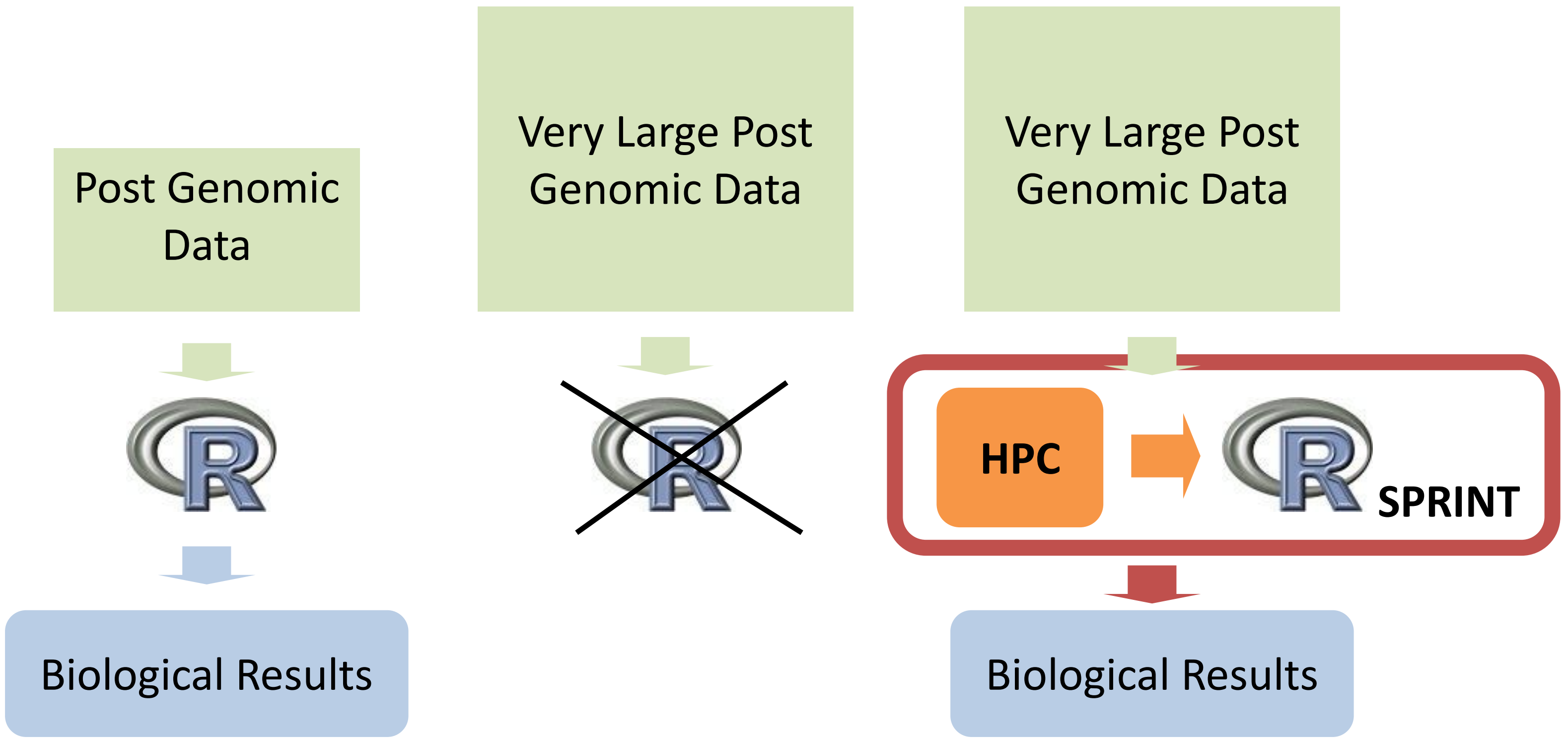




- High throughput approaches to mapping and understanding host-response to infection.
- Targeting the host NOT the “bug” as anti-infective strategy

Story starts at the bed side.

High throughput data requires high throughput analysis



Using all or many genes (Exons, SNPs, …) will either crash or be very slow:

- ➢ Space limitation ("…cannot allocate vector of size…")
- ➢ Time limitation

Issues with Existing Parallel R packages

Parallel building blocks:

- Bespoke implementation each time
- Difficult to program:

  require scientist to also be a parallel programmer!
- Rmpi, rpvm, nws & sleigh

Task farms:

- Require substantial changes to existing scripts
- No data dependencies allowed:

  Can't be used to solve certain class of problems
- SNOW, R/Parallel, papply, BioPara, taskPR

SPRINT: Simple Parallel R INTerface

Aims to overcome limitations on data size and analysis time by providing easy access to HPC for all R users

SPRINT:

- An intelligent HPC harness:
  - Implemented in C & MPI
  - Scalable (RAM & CPU), portable and flexible
- R parallel function library:
  - Popular functions, complex functions, open to contributions
  - Optimized
- User Friendly:
  - Aimed at biologists and biostatisticians
  - Minimum changes, R interface

SPRINT User Requirement Survey

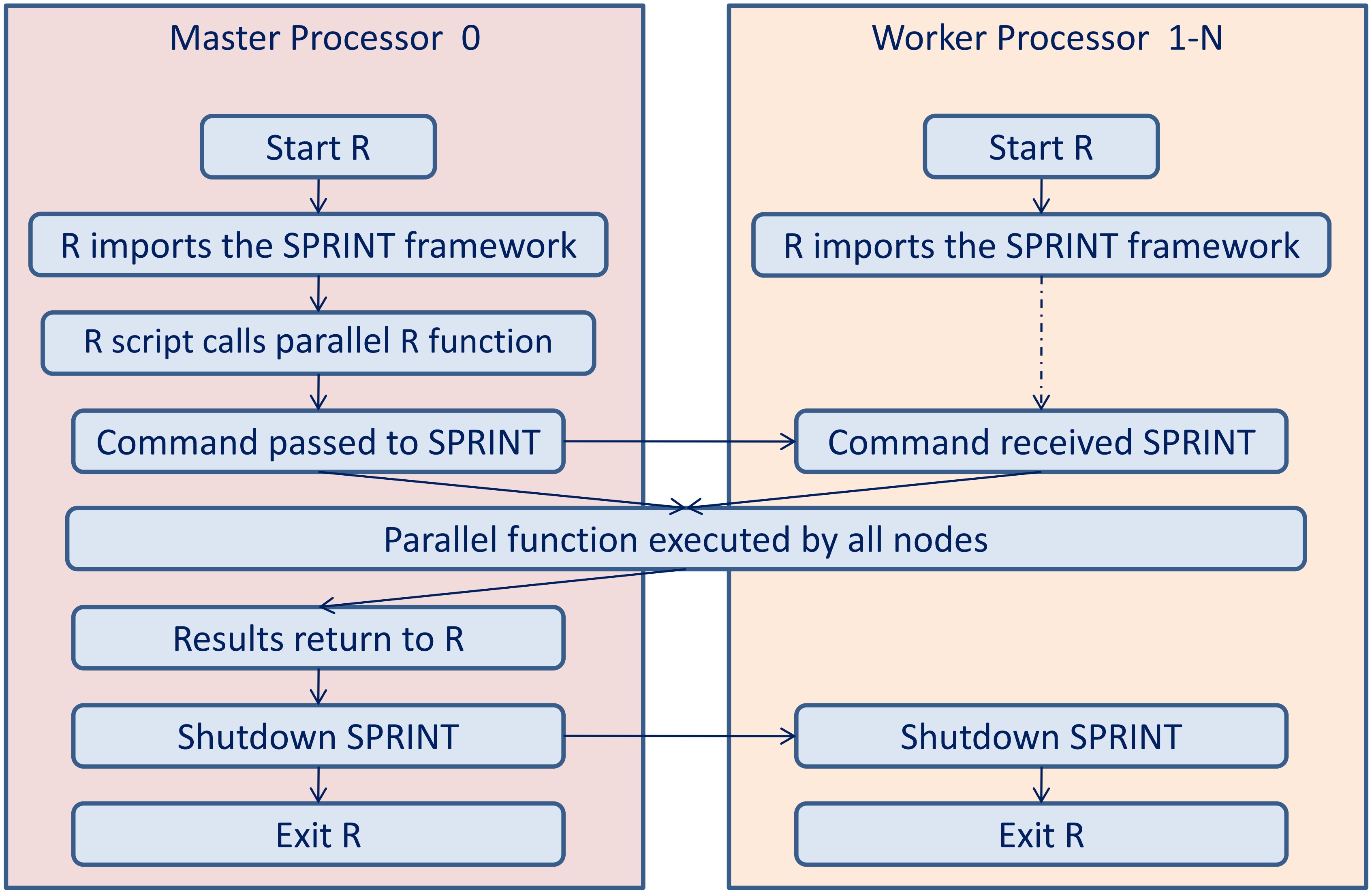
Function	# requested	SPRINT function	Release
Standard R functions	15		

Permutation, bootstrapping	10	pmaxT()
		pboot()
			Beta 2 (Jun 2010)
			Beta 4 (TBC)

Machine learning algorithms	9	ppam()	Beta 3 (Soon)
Correlation functions	8	pcor()	Beta 1 (Jan 2010)
Normalisation	8		
Standard Statistics	7		
Matrix operations	7		
Other	12		

- No GUI
- Full report available at www.r-sprint.org

SPRINT Architecture



Code Modification

```
data(golub)
smallgd <- golub[1:100,]
classlabel <- golub.cl

resT <- mt.maxT(smallgd, classlabel, test="t", side="abs")

quit(save="no")
```

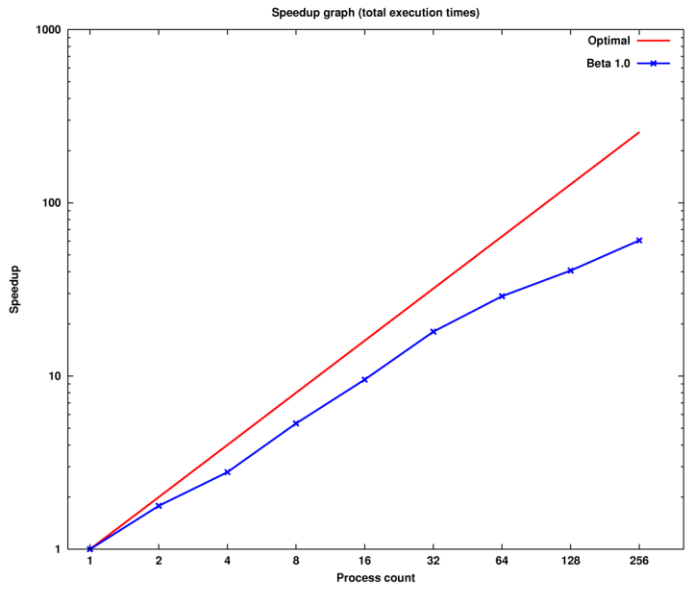
```
library("sprint")

data(golub)
smallgd <- golub[1:100,]
classlabel <- golub.cl

resT <- pmaxT(smallgd, classlabel, test="t", side="abs")

pterminate()

quit(save="no")
```

Division of PATHWAY MEDICINE

SPRINT Pearson Correlation: pcor()

epcc

- Parallel implementation of cor() .
- uses ff package: memory-efficient storage of large data on disk and fast access functions (available from CRAN). Implements fast memory mapped access to flat files.
- ff objects can be created, stored, used and removed, almost like standard R RAM objects.
- Allows to process datasets that do not fit into CPU physical memory.
- ff objects are perfect to read the same data from many R processes.



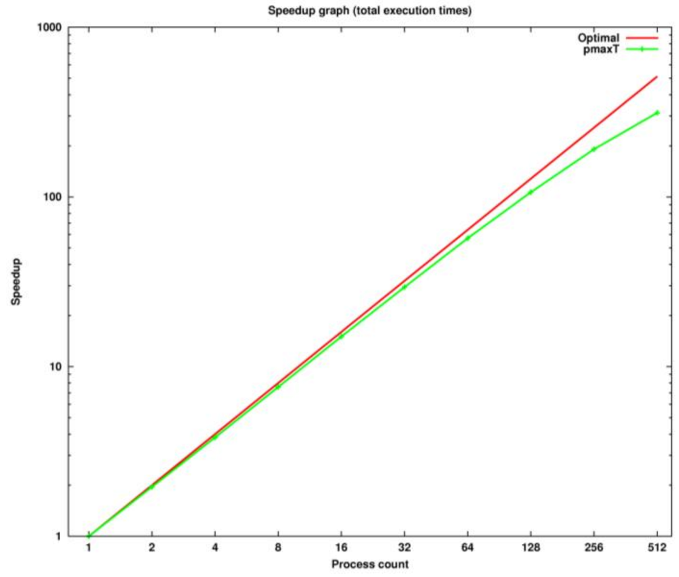
Input Matrix Size	Output Matrix Size	Serial Run Time	Parallel Run Time
11,000 x 320 **26.85 MB**	**0.9 GB**	**63.18 secs**	**4.76 secs**
22,000 x 320 **53.7 MB**	**3.6 GB**	**Insufficient memory**	**13.87 secs**
35,000 x 320 **85.44 MB**	**9.12 GB**	**Crashed**	**36.64 secs**
45,000 x 320 **109.86 MB**	**15.08 GB**	**Crashed**	**42.18 secs**

Benchmark on HECToR - UK National Supercomputing Service on 256 cores.

SPRINT Permutation Test: pmaxT()

- Parallel implementation of mt.maxT() from multtest package (available from CRAN).

Benchmark on HECToR – UK National Supercomputing Service on 512 cores for 150,000 permutations of 6102 x 76 matrix



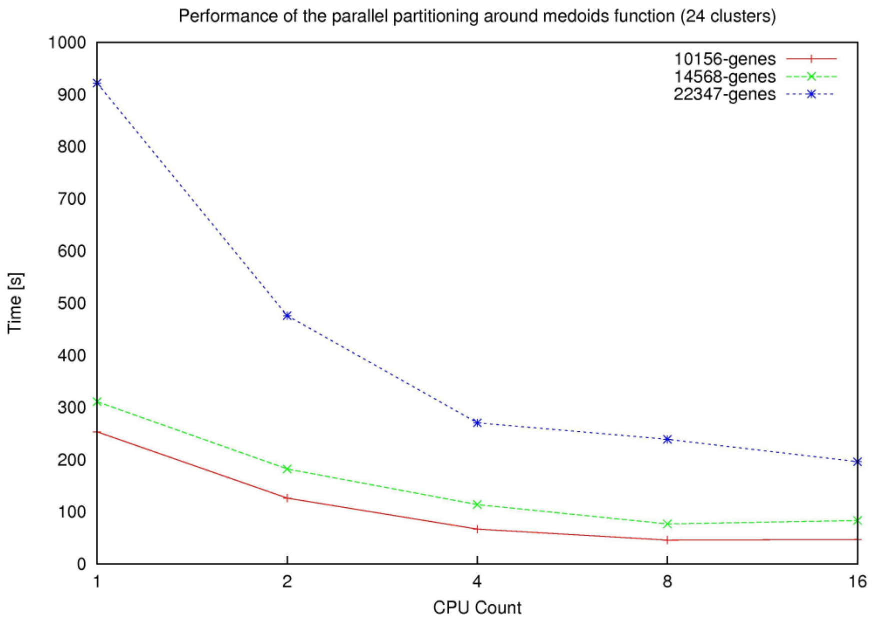
Input Matrix Size	# Permutations	Serial Run Time (estimated)	Parallel Run Time
36,612 x 76	**500,000**	**6 hrs**	**73.18 secs**
36,612 x 76	**1,000,000**	**12 hrs**	**146.64 secs**
36,612 x 76	**2,000,000**	**23 hrs**	**290.22 secs**
73,224 x 76	**500,000**	**10 hrs**	**148.46 secs**
73,224 x 76	**1,000,000**	**20 hrs**	**294.61 secs**
73,224 x 76	**2,000,000**	**39 hrs**	**591.48 secs**

Benchmark on HECToR - UK National Supercomputing Service on 256 cores.

SPRINT Clustering: ppam()

- Parallel implementation of pam() from cluster package (available from CRAN).
- Optimisation of serial version through memory and data storage management.
- Increase capacity by using external memory (ff objects).

Benchmark on a shared memory cluster with 8 dual-core 2.6GHz AMD Opteron processors with 2GB of RAM per core.



Input Data Size	# Clusters	Serial Run Time Pam()	Parallel Run Time Ppam()
2400	**12**	**11.3 secs**	**1.1 secs**
2400	**24**	**52.5 secs**	**2.2 secs**
4800	**12**	**83.3 secs**	**4.4 secs**
4800	**24**	**434.7 secs**	**15.9 secs**
10000	**12**	**17 mins**	**22.3 secs**
10000	**24**	**99 mins**	**77.1 secs**
22374	**24**	**Insufficient memory**	**270.5 secs**

What next?

- SPRINT future releases:
  - Other distance metrics, bootstrapping, clustering, apply functionality,...
- Open source project **for** and **by** the R community:
  - Tell us what functionality you want
  - Add your own functions to SPRINT
- Started in biology but statistics methods can be apply to any subject.

Division of Pathway Medicine and Edinburgh Parallel Computing Centre at University of Edinburgh.

DPM Team:

- Prof. Peter Ghazal
- Thorsten Forster
- Muriel Mewissen

http://www.r-sprint.org

EPCC Team:

- Terry Sloan
- Michal Piotrowski
- Savvas Petrou
- Bartek Dobrzelecki
- Jon Hill
- Florian Scharinger

This work was supported by the Wellcome Trust grant [086696/Z/08/Z], by an HECToR computational science and engineering support award (dCSE) administered by NAG Ltd and edikt2.