

Overlap Death Match!

Toby Dylan Hocking
toby.hocking@mail.mcgill.ca

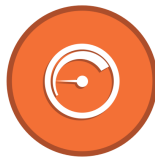


VS



Bioconductor

VS



data.table

12 February 2015

The contenders

Round 1: timings on genomic data

Round 2: accuracy on genomic data

And the champion is...

Google Summer of Code

Brief history of statistical computing

- 1957 FORTRAN by John Backus (IBM).
- 1972 C by Dennis Ritchie (Bell Labs).
- 1976 S by John M Chambers (Bell Labs).
- 1983 C++ by Bjarne Stroustrup (Bell Labs).
- 1993 S exclusively licensed to StatSci/MathSoft.
R by Ross Ihaka and Robert Gentleman
(Univ Auckland, New Zealand).
- 1998 Association of Computing Machinery Software
System award to John M Chambers for “the S
system, which has forever altered the way people
analyze, visualize, and manipulate data.”

Source: Wikipedia, R-FAQ.

R = interactive, graphical, programming with data

What is R? (Source: R-FAQ)

“R is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files.”

interactive command line (versus compiled).

graphical publication-quality plots.

programming with data `data.frame` which represents a tabular data set.

Selected Bioconductor project history

Version	Release	Packages	Depends	Firsts/notes
1.0	1 May 2001	15	R 1.5	
2.3	22 Oct 2008	294	R 2.8	IRanges
2.5	28 Oct 2009	352	R 2.10	findOverlaps
2.6	23 Apr 2010	389	R 2.11	GenomicRanges
3.0	14 Oct 2014	934	R 3.1	TESTED



Bioconductor = R packages with compiled C code.

Source: Wikipedia “Bioconductor,” archived packages e.g.
<http://www.bioconductor.org/packages/2.5>

Selected bedtools project history

VERSION 1.1, 04/23/2009. Initial release.

BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (2010).

Version 2.14.1-3 (2-Nov-2011)

...

2.14.3-1

TESTED ubuntu

...

Version 2.17.0 (3-Nov-2012)

TESTED guillimin

...

Version 2.22.1 (1 Jan 2015)

TESTED github



= command line C++ program.

Source: bedtools RELEASE_HISTORY file.

data.table is “just like a data.frame”



data.table = an R package with compiled C code.

Version: 1.0, 2006-04-12

Author: Matt Dowle

Title: Just like a data.frame but without rownames,
up to 10 times faster, up to 10 times less memory

data.table is an “extension of data.frame”



data.table = an R package with compiled C code.

Version: 1.9.4, 2014-10-02

Author: M Dowle, T Short, S Lianoglou, A Srinivasan
with contributions from R Saporta, E Antonyan

Title: Extension of data.frame

Description: Fast aggregation of large data
(e.g. 100GB in RAM), fast ordered joins,
fast add/modify/delete of columns
by group using no copies at all,
list columns and a fast file reader (fread)...

data.table supports overlap joins

data.table/README.md

Changes in v1.9.4 (on CRAN 2 Oct 2014)

NEW FEATURES

...

Overlap joins (#528) is now here, finally!!

Except for type="equal" and maxgap and minoverlap arguments, everything else is implemented.

Check out ?foverlaps and the examples there on its usage.

This is a major feature addition to data.table.

[https://github.com/Rdatatable/data.table/wiki/talks/
EARL2014_OverlapRangeJoin_Arun.pdf](https://github.com/Rdatatable/data.table/wiki/talks/EARL2014_OverlapRangeJoin_Arun.pdf)

Summary of contenders, demo



Bioconductor

data.table

Since	2009	2009	2014
Language	C++	R/C	R/C
Versions Tested	2.14.3 2.17.0 2.22.1	3.0	1.9.4

The contenders

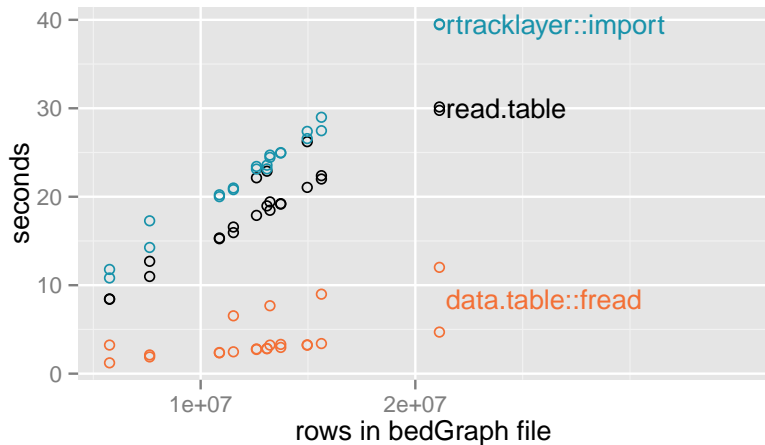
Round 1: timings on genomic data

Round 2: accuracy on genomic data

And the champion is...

Google Summer of Code

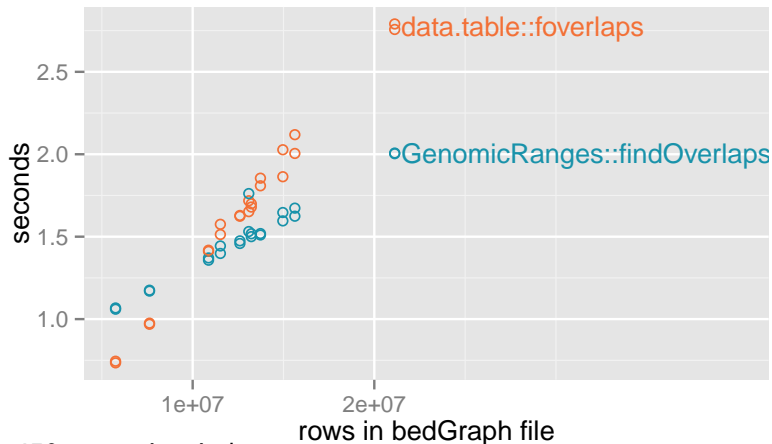
data.table::fread faster than read.table for bedGraph files



```
read.table(file, sep=, colClasses=, nrow=) vs  
fread(file)
```

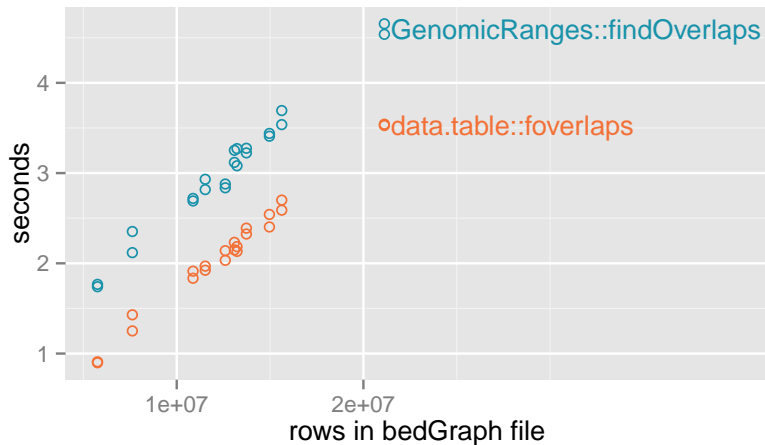
Only compute overlapping indices

```
GenomicRanges::findOverlaps(bedGraph.gr, windows.gr)  
data.table::foverlaps(bedGraph.dt, windows.dt,  
                      nomatch=0L, which=TRUE)
```



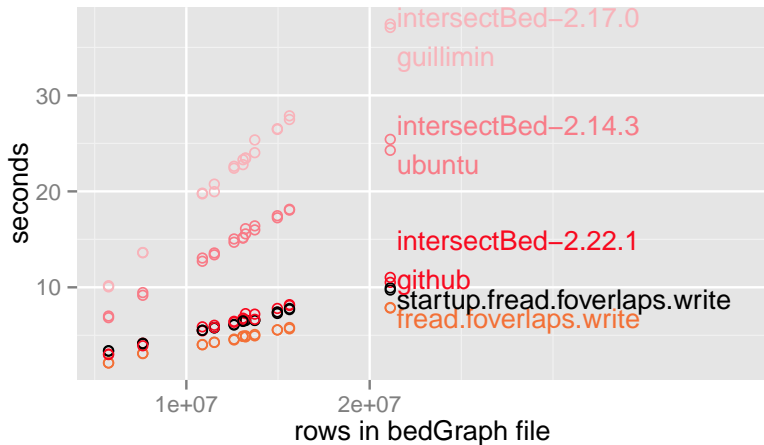
≈ 450 genomic windows.

Start from data.frame, convert to GRanges/data.table,
compute overlapping indices, select data



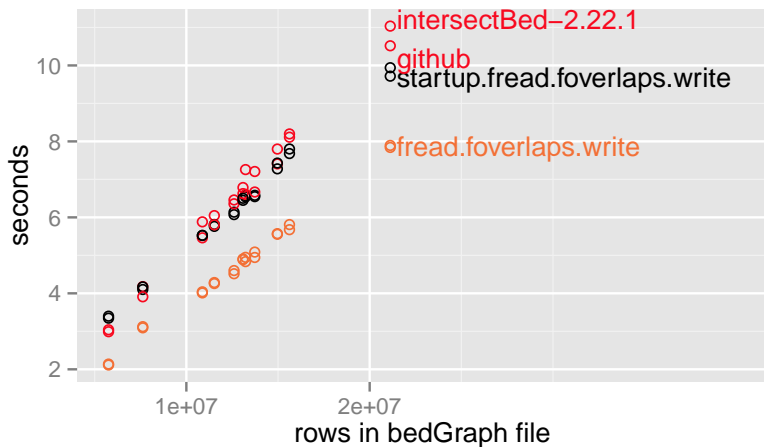
≈ 450 genomic windows.

Read files, overlap, write file



≈ 450 genomic windows.

Read files, overlap, write file



≈ 450 genomic windows.

The contenders

Round 1: timings on genomic data

Round 2: accuracy on genomic data

And the champion is...

Google Summer of Code

`data.table::foverlaps` always gives the same result as
`GenomicRanges::findOverlaps`



... but are they correct?

Need chromStart+1 in R!

chipseq.bedGraph

```
chrom chromStart chromEnd coverage
chr1      0      200      1
chr1     200     300      2
```

Expected versus computed coverage for

- ▶ 4 test windows (columns) and
- ▶ 5 overlap methods (rows).

chromStart	200	199	0	0
chromEnd	1000	1000	200	201
expected	2	1,2	1	1,2
findOverlaps	1,2	1,2	1,2	1,2
findOverlaps+1	2	1,2	1	1,2
foverlaps	1,2	1,2	1,2	1,2
foverlaps+1	2	1,2	1	1,2
intersectBed-2.22.1	2	1,2	1	1,2

The contenders

Round 1: timings on genomic data

Round 2: accuracy on genomic data

And the champion is...

Google Summer of Code

All packages have similar speed, accuracy

Caveats:

- ▶ need most recent versions!
- ▶ need `chromStart+1` in R!

Bonus points

+1 to

- ▶ **bedtools** for binary files (BAM).
- ▶ **data.table** for `fread`
(read text files faster than `read.table`).
- ▶ **bedtools** and **GenomicRanges::findOverlaps** for
`-f/minoverlap`
(not yet implemented in `data.table::foverlaps`).
- ▶ **GenomicRanges::findOverlaps** for `maxgap`
(not yet implemented in `data.table::foverlaps`).
- ▶ **bedtools** for native support for 0-based `chromStart` of
`bed/bedGraph` files (need to use `chromStart+1` in R).

The contenders

Round 1: timings on genomic data

Round 2: accuracy on genomic data

And the champion is...

Google Summer of Code

Google Summer of Code (GSOC)

Student gets \$5000 for writing open source code for 3 months.

Feb Admins for open source organizations e.g. R, Bioconductor apply to Google.

Mar Mentors suggest projects for each org.
Students submit project proposals to Google.
Google gives funding for n students to an org.

April The top n students get \$500 and begin coding.

July Midterm evaluation, pass = \$2250.

Aug Final evaluation, pass = \$2250.

November Orgs get \$500/student mentored.

I have participated as an **admin** and **mentor** for the R project.

What makes a good GSOC project?

Coding projects should:

- ▶ Result in free/open-source software.
- ▶ Be 3 months of full time work for a student.
- ▶ Include writing documentation and tests.
- ▶ Not include original research.

Examples:

- ▶ Louis/Mathieu can be **admins** for MUGQIC org.
- ▶ Warren/Stephan can be **mentors** for a project to write a new R package for methylation analysis.
- ▶ Robert/Dan can be **mentors** for a project to implement new features in Gemini.
- ▶ Any undergrad/master/PhD candidates (at McGill or not) can be **students**.

Thanks for your attention!

Any questions? `toby.hocking@mail.mcgill.ca`

Source code for benchmarks is at
<https://github.com/tdhock/datatable-foverlaps>

old intersectBed -sorted is much slower

“-sorted Invoke a memory-efficient algorithm for very large files.”

Version: 2.14.3-1

```
$ time intersectBed -a windows.bed \  
  -b chip-seq.bedGraph -sorted > overlap.bedGraph  
real 154m10.653s  
user 153m2.188s  
sys  0m20.204s
```

```
$ time intersectBed -a windows.bed \  
  -b chip-seq.bedGraph > overlap.bedGraph  
real 0m8.733s  
user 0m7.932s  
sys  0m0.728s
```