Small tools for bioinformatics

sambamba, pfff, once-only, bio-vcf

Pjotr Prins, Artem Tarasov, Konstantin Tretyakov

Bioinformatics Open Source Conference (BOSC) 2014
Stating the problem

Large scale data acquisition in research has led to fundamental challenges in

- scaling of calculations
- storage and full data integration
- data exploration and visualisation
Hefty challenges

- Tool integration and versioning
- Workflow management
- Manage changing environment and tools, provenance
- Local, parallel, cluster, map-reduce, Cloud, super, GRID computing
- Data integration, annotation
- Visualisation and user interfaces
- Challenges; duplication of effort
So why is bioinformatics so often ‘not invented here’ and opting for ‘monolithic’ solutions, despite a long history of tools in the spirit of Unix?

- Technology requires it? (deployment)
- Bioinformaticians and organisations want ‘control’?
- Biologists ask for it?
The MANIFESTO builds on the Unix computer tradition

- Provide ‘small tools’
- that can be used in a modular and pluggable way
- to create efficient computational solutions
- where individual parts can be easily replaced
- 50 stars, 35 forks, 26 signed

https://github.com/pjotrp/bioinformatics/README.md
MANIFESTO

- ‘Small tool’ should do smallest possible task really well
- FOSS published source code (FSF license)
- Command line interface and pipes (if possible)
- Sane error handling, transparent and reproducible
- Automated testing
- Software packaging
- Anti-fragile (abide by rules of evolution)

https://github.com/pjotrpbioinformatics/README.md
**Pfff example**

- Fastest file Hash generator on the planet (C)
- Pfff is an MD5 replacement for large data
- Sampling fingerprints reduces IO
- Flat performance characteristic
- Adoption by those who find IO is a bottleneck
- Possible because md5sum is a ‘small tool’

https://github.com/pfff/pfff (Konstantin Tretyakov)
Sambamba example

- Fastest SAM/BAM parser on the planet (D)
- Drop-in replacement of samtools/Picard
- Great speed and comes with powerful filtering
- Used in pipelines around the world, incl. Illumina and Harvard
- Adding CRAM support and more integration options
- Possible because samtools is a ‘small tool’

https://github.com/lomereiter/sambamba (Artem Tarasov)
Bio-vcf

- Fastest VCF parser on the planet (Ruby)
- Can access any VCF format
- Expressive filtering and evaluation language
- Can calculate and rewrite VCF
- Can output RDF/tabular/LaTeX/JSON
- Possible because snpsft is a ‘small tool’

https://github.com/pjotrp/bioruby-vcf (Pjotr Prins)
MANIFESTO

Modules, plugins, packages...
Design software to be a component that can be wired up
Design software for replacement
Design software for failure
Sign the manifesto!

https://github.com/pjotrp/bioinformatics/README.md
NIH - CPAN, Ruby gems, Pypy, Homebrew, Galaxy toolshed...

GNU Guix is packaging done right

Dependency and versioning are solved problems

All users can install software, without conflicts

True reproducible software installations

This way, a small tool can be hosted anywhere
Small tools are game changers:

- GNU Guix (software deployment)
- Pfff (instant file comparison)
- Sambamba (parsing/filtering/rewriting SAM/BAM)
- bio-vcf (parsing/filtering/rewriting VCF)
- bio-table (parsing/filtering/rewriting tabular data)
- once-only (run commands only once - pfff on inputs)

https://github.com/pjotrp