Scientific Workflow Systems for accessible, reproducible research

Peter van Heusden and Alan Christoffels

Email: pvh@sanbi.ac.za
South African National Bioinformatics Institute
University of the Western Cape
Bellville, South Africa

UNIVERSITY of the WESTERN CAPE

SANBI
South African National Bioinformatics Institute

eResearch Africa 2013 / Cape Town

UNIVERSITY of the WESTERN CAPE

Outline

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Hi

Hi!

I'm Peter

and I'm a bioinformaticist.

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Hi

Hi!

I'm Peter

and I'm a bioinformaticist.

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Hi

Hi!

I'm Peter

and I'm a bioinformaticist.

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

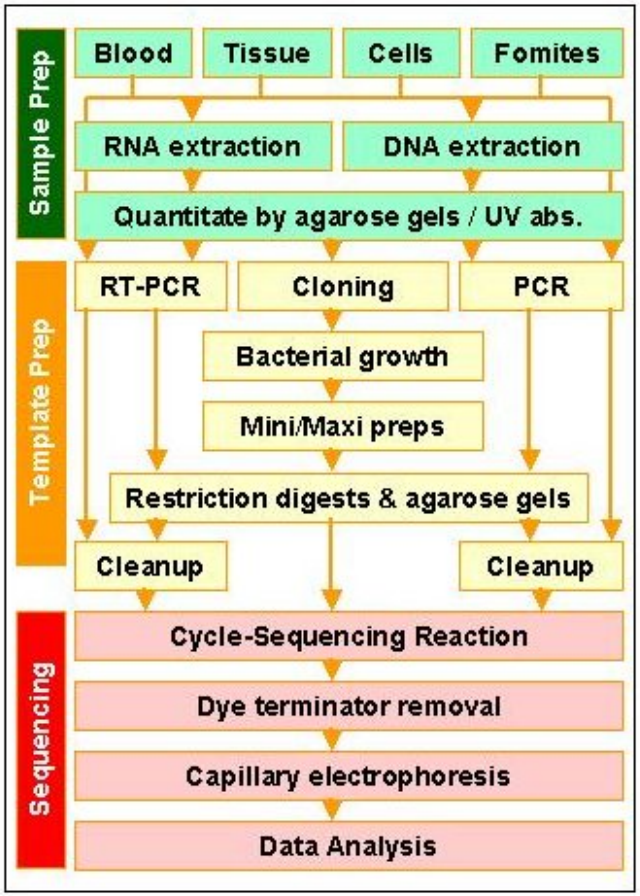
SANBI

from SANBI

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

What do we do?

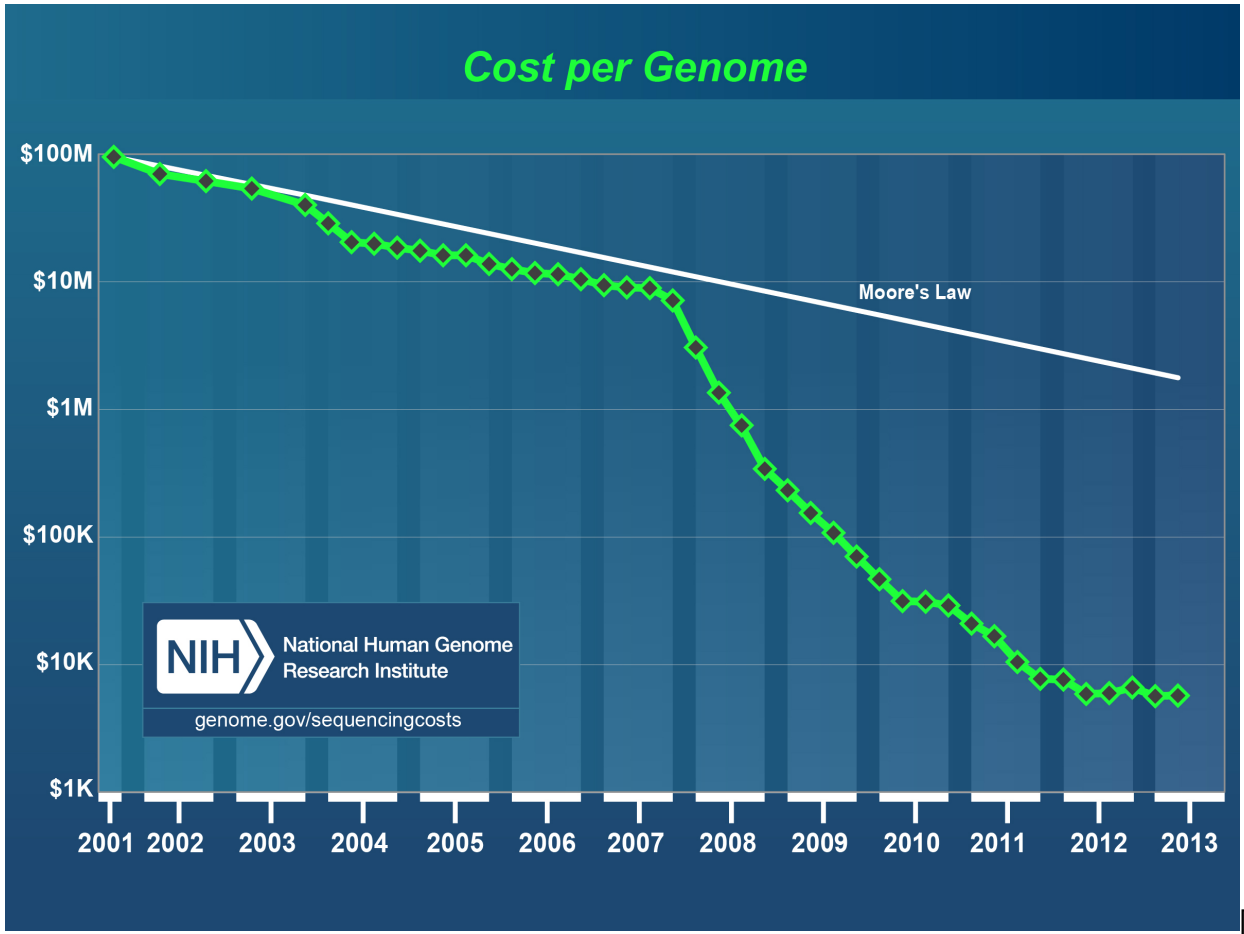


[Abizar Lakdawalla, 2007]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

But wait. . .

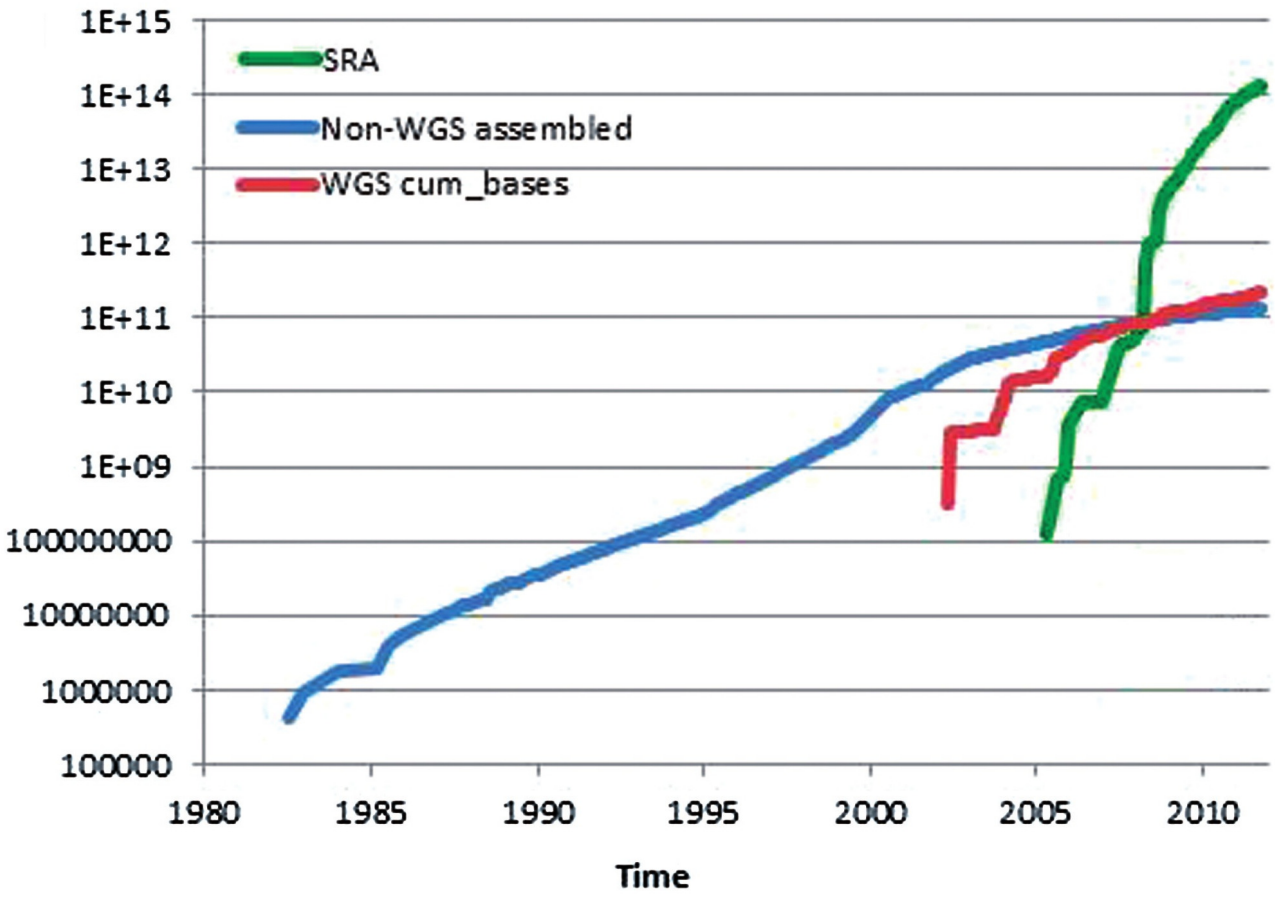


[Wetterstrand, 2013]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Data is cheap and plentiful



[Karsch-Mizrachi et al., 2012]

SANB South African Nation Bioinformatics Institu

UNIVERSITY of the WESTERN CAPE

Plunging cost of sequencing changes bioinformatics

"[Y]ou see people collecting information and then having to put a lot more energy into the analysis of the information than they have done in getting the information in the first place. The software is typically very idiosyncratic since there are very few generic tools that the bench scientist has for collecting and analyzing and processing the data. This is something that we computer scientists could help fix by building generic tools for the scientists." Jim Gray [Anthony J. G. Hey et al., 2009]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

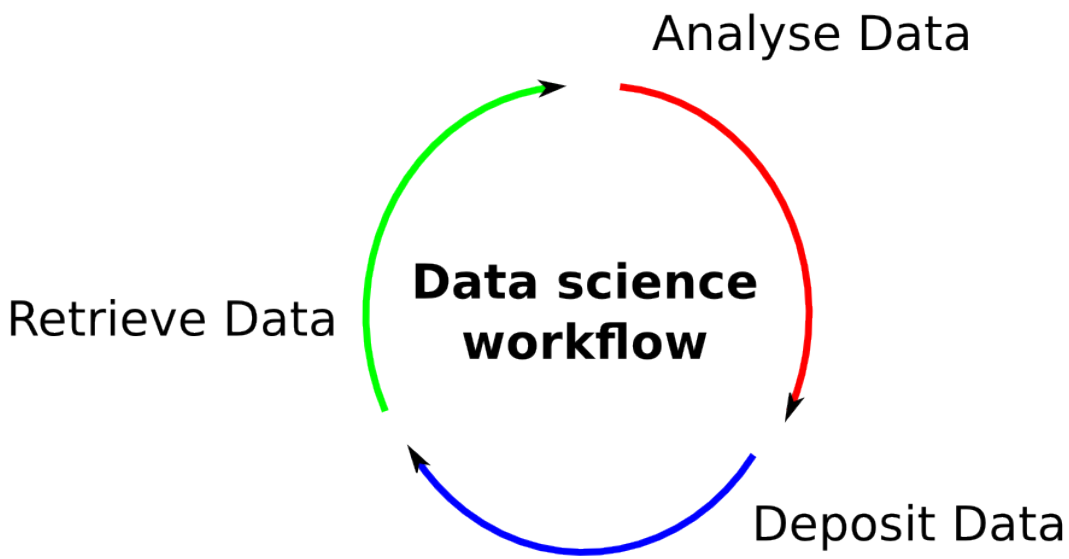
Plunging cost of sequencing changes bioinformatics

"[Y]ou see people collecting information and then having to put a lot more energy into the analysis of the information than they have done in getting the information in the first place. The software is typically very idiosyncratic since there are very few generic tools that the bench scientist has for collecting and analyzing and processing the data.
This is something that we computer scientists could help fix by building generic tools for the scientists." Jim Gray [Anthony J. G. Hey et al., 2009]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

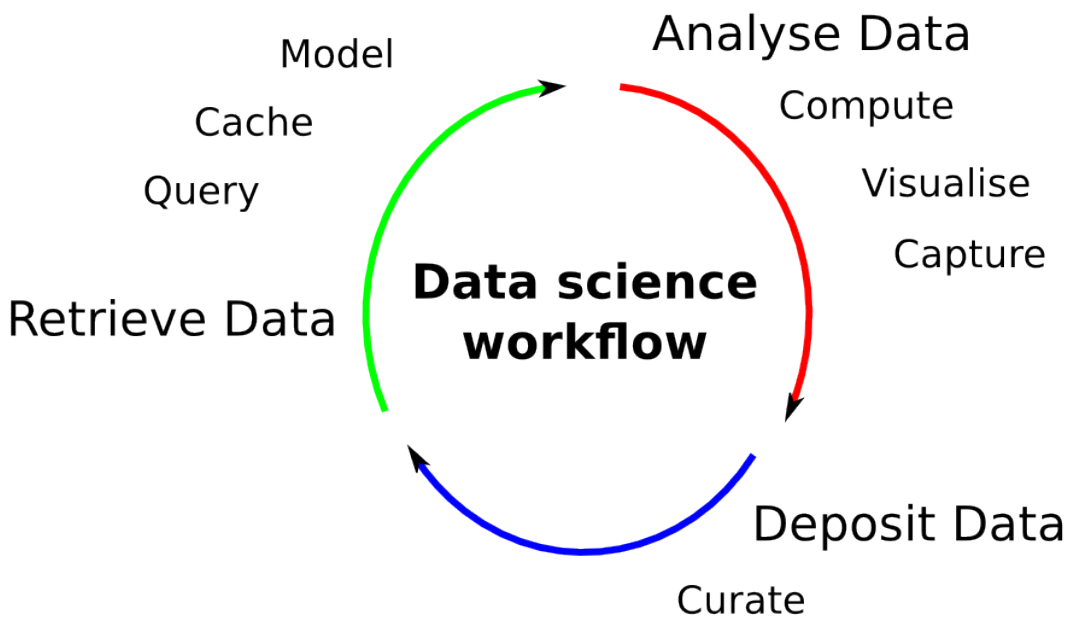
Bioinformatics analysis



SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

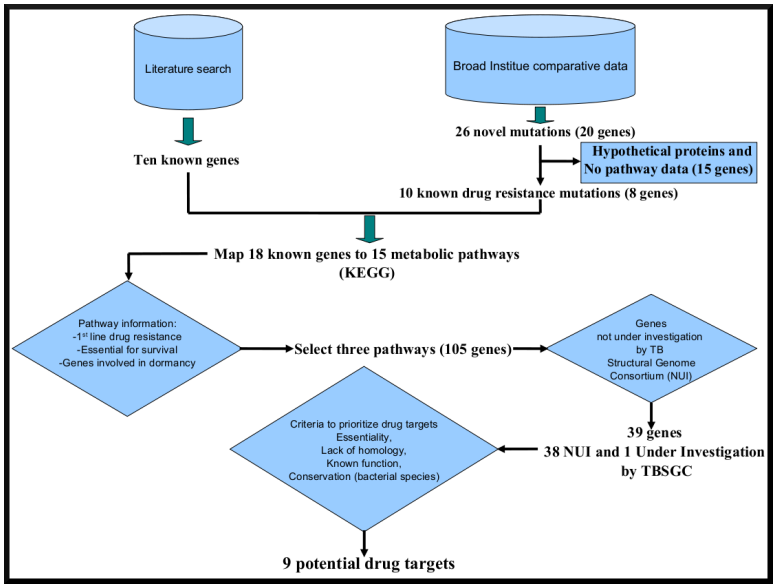
Bioinformatics analysis



SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Dr Cloete's hunt for a TB targetting compound



- PhD student at SANBI, Ruben Cloete, mined existing knowledge about M. tuberculosis to find potential drug targets for curing the disease
- Analysis proceeds through a series of queries, transforms and filters
- Collection of tasks in analysis comprises a bioinformatics workflow

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

How workflows are implemented

```
print "The dictionary of sample and genomic files", Variables.dict_of_sample_fastqfiles

for each_sample in Variables.dict_of_sample_fastqfiles.keys():

        #count number of genomic files N for each sample. N - number of amplicons per sample
        N=len(Variables.dict_of_sample_fastqfiles[each_sample])

        if N == 0: continue
        #create bash file for each amplicon/genomic file to be submitted in array job
        counter=0
        list_of_toolname_fastm_files=[]

        #get time stamp for appending in job name
        timestamp = str(datetime.now()).replace(".", "").replace("-", "").replace(" ","").replace(":",""
        print "Sample name", each_sample
        for ref_align in Variables.dict_of_sample_fastqfiles[each_sample]:

                print "sample gene file ", ref_align
```

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

The problem with our workflows

- Our workflows are specified at a low level, in terms of filenames and commands to execute
- Workflow scripts run analyses in the same way as when they are executed by hand by a researcher
- Last decade has seen change, but not enough:

Language	Perl
Execution	On server
Data storage	Files

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

The problem with our workflows

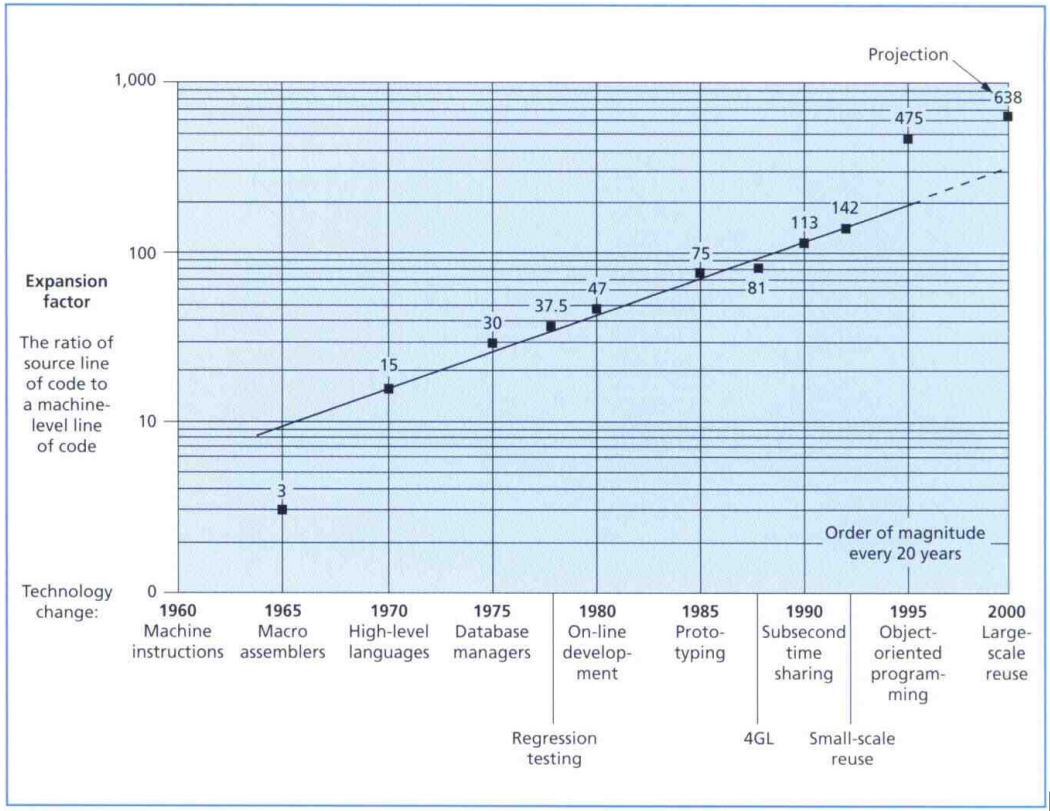
- Our workflows are specified at a low level, in terms of filenames and commands to execute
- Workflow scripts run analyses in the same way as when they are executed by hand by a researcher
- Last decade has seen change, but not enough:

Language	Python
Execution	On cluster / grid
Data storage	Files / SQL db

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

A better toolset



[Bernstein, 1997]

New toolsets drive software engineering productivity

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Scientific workflow management systems

"Scientific workflows have emerged to tackle the problem of excessive complexity in scientific experiments and applications. They provide a high-level declarative way of specifying what a particular in silico experiment modelled by a workflow is set to achieve, not how it will be executed." [Taverna project, 2009]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

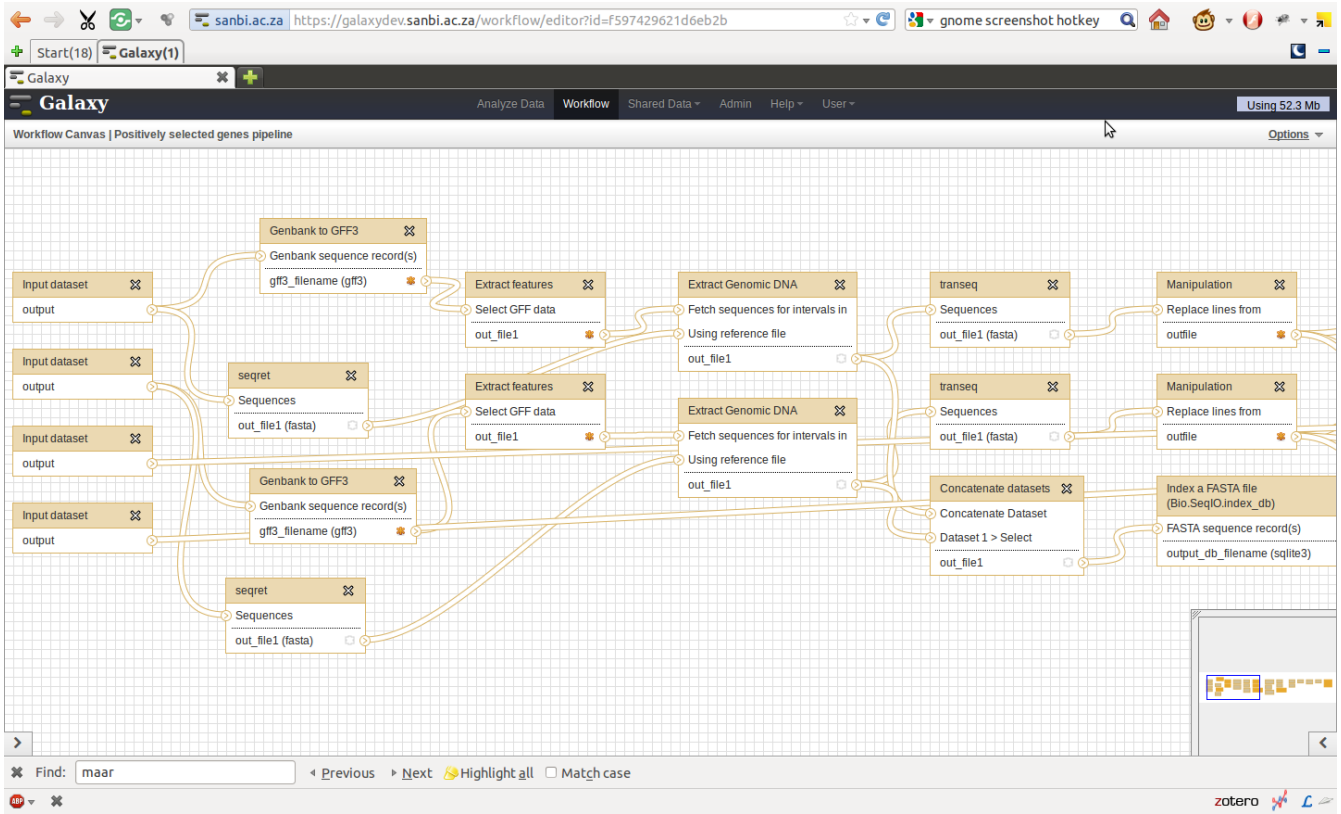
Scientific workflow management systems (2)

- Scientific workflow management systems (SciWMSs) are data-flow oriented
- Expressing workflows in terms of data dependencies allows for flexible mapping to computational resources
- SciWMSs are frameworks, not APIs: they implement a workflow pattern, user fills in the blanks of what needs doing
  - "A framework is a set of cooperating classes that make up a reusable design for a specific class of software" [Gamma et al., 1993]
  - "If applications are hard to design, and toolkits are harder, then frameworks are hardest of all. A framework designer gambles that one architecture will work for all applications in the domain." [Gamma et al., 1993]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

A graphical workflow specification (Galaxy)



[Goecks et al., 2010]

A textual workflow specification (bpipe)

```
REFERENCE="reference.fa"
PICARD_HOME="/usr/local/picard-tools/"

...

index = {
    exec "samtools index $input"
}

call_variants = {
    exec "samtools mpileup -uf $REFERENCE $input | bcftools view -bvcg - > $output"
}

Bpipe.run {
    align + sort + dedupe + index + call_variants
}
```

[Sadedin et al., 2012]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Why don't we use scientific workflow management systems already?

"Contemporary workflow platforms fall short of adequately supporting rapid deployment into the user applications that consume them, and legacy application codes need to be integrated and managed." — Goble and de Roure in [Anthony J. G. Hey et al., 2009]

- SciWMSs are often not available on the computing platforms that scientists use.
- SciWMS support for the workflow patterns that scientists use is sometimes poor
- Bioinformaticists use a large number of tools, and a SciWMS for bioinformatics must have rich tool support
- The software adoption process in science is not a straightforward one

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Barriers to SciWMS availability

- Some are commercial products (e.g. Pipeline Pilot)
- Can be hard to install (e.g. Kepler [Altintas et al., 2004] and Taverna [Hull et al., 2006])
- Frameworks require long-running daemons that require sysadmin buy in to run
  - and sysadmins and scientists live in different worlds

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

SciWMS support for scientific workflow patterns

- Current tool of choice for scientific workflow composition is a combination of manual effort and general purpose scripting languages
- As software frameworks SciWMSs implement particular workflow patterns
- These might not match workflow requirements: e.g. Galaxy lacks sub-workflow support or iteration across a collection
- Lack of adequate support for workflow patterns results in "hacks" that re-introduce needless complexity into the workflow specification

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Tool support in SciWMSs

- Scientific workflows make use of a large and evolving software collection (e.g. see [Eagle Genomics, 2013])
- SciWMSs must support:
  1. Rich set of tools "out of the box"
  2. Straightforward procedures to add new tools
- Some SciWMSs (such as bpipe) closely resemble scripting languages and replicate the "command line" model of tool invocation
- On the other extreme, Taverna is service oriented and expects to tools to be made available as web services
- If adding a tool is complex, scientists typically won't do it
- A partial solution is supporting a community collection of tool definitions (e.g. Galaxy toolshed)

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Software adoption in scientific communities

"'End-user developers' commonly create scientific software, but they are often unaware of or ignore traditional software engineering standards, leaving trust in their coding expertise potentially misplaced."

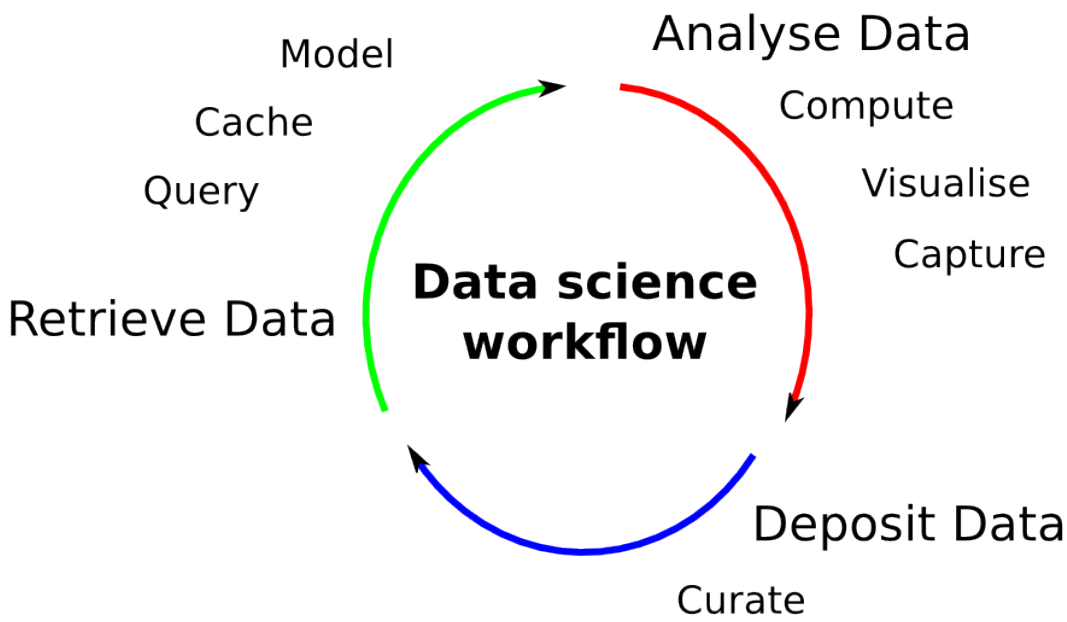
— [Joppa et al., 2013]

- Scientific software adoption follows trends within the scientific community, not necessarily software engineering best practice.
- While training is necessary, it is not sufficient to ensure adoption of new tools: building communication channels between different communities is key.
- Evolving scientific workflow practice is a social problem not just a technical one.

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Bioinformatics analysis (revisited)



SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Workflow re-use and reproducible results

"all software is wrong, although some of it is useful, and all results are approximate" — C. Titus Brown [C. Titus Brown, 2013]

- Current scientific work process is focused on end products (papers, theses) that cannot be easily reproduced.
- In 2010 survey of 20 most highly cited journals found that only 3 required source code to be submitted along with papers.
- Even then: "Workflows are often difficult to author, using languages that are at an inappropriate level of abstraction and expecting too much knowledge of the underlying infrastructure. The reusability of a workflow is often confined to the project it was conceived in – or even to its author – and it is inherently only as strong as its components." — Goble and de Roure [Anthony J. G. Hey et al., 2009]

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

SciWMS for reproducable research

- Results of scientific computing are not final "truth" but best-guess approximations (with associated uncertainty)
- SciWMSs that provide features that enhance the reproducibility of research aid in the collaborative critique and elaboration of results
- Critical features for reproducible research include:
  - Provenance recording: recording data on the provenance of workflow products (but how and what should we record?)
  - Abstract workflow components: liberate workflows from the lab and allow scientists to "run the components of the workflow wherever the data is." [C. Titus Brown, 2011]
  - Bring data closer to publications: publication should combine data, text and workflow e.g. Galaxy Pages, Vistrails' LaTeX integration

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Workflow re-use

- If we can repeat a workflow, we can re-use it
- Workflows will evolve: workflows seldom are "one size fits all"
- If workflows evolve we need change management and versioning
  - Textual workflow specifications can be managed by traditional source code control tools (e.g. **git**)
  - In SciWMSs that use graphical specifications Vistrails provides an example of how a "version tree" can be represented and navigated
  - But:

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Workflow re-use

- If we can repeat a workflow, we can re-use it
- Workflows will evolve: workflows seldom are "one size fits all"
- If workflows evolve we need change management and versioning
  - Textual workflow specifications can be managed by traditional source code control tools (e.g. **git**)
  - In SciWMSs that use graphical specifications Vistrails provides an example of how a "version tree" can be represented and navigated
  - But: most SciWMSs currently have poor change management support

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

In conclusion

- In bioinformatics data is becoming plentiful and cheap
  - this is changing the cost structure of research, since data analysis is still expensive
- Data analysis is large done using workflows involving human effort and scripting languages
  - these workflows are fragile, not reproduceable and hard to understand and re-use
- SciWMSs provide frameworks for higher level specification of scientific workflow
  - these frameworks are not widely used due to technical and community limitations
  - despite these limitations SciWMSs offer an essential path towards reproduceable and re-useable scientific workflows

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

In conclusion

- In bioinformatics data is becoming plentiful and cheap
  - this is changing the cost structure of research, since data analysis is still expensive
- Data analysis is large done using workflows involving human effort and scripting languages
  - these workflows are fragile, not reproduceable and hard to understand and re-use
- SciWMSs provide frameworks for higher level specification of scientific workflow
  - these frameworks are not widely used due to technical and community limitations
  - despite these limitations SciWMSs offer an essential path towards reproduceable and re-useable scientific workflows

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

In conclusion

- In bioinformatics data is becoming plentiful and cheap
  - this is changing the cost structure of research, since data analysis is still expensive
- Data analysis is large done using workflows involving human effort and scripting languages
  - these workflows are fragile, not reproduceable and hard to understand and re-use
- SciWMSs provide frameworks for higher level specification of scientific workflow
  - these frameworks are not widely used due to technical and community limitations
  - despite these limitations SciWMSs offer an essential path towards reproduceable and re-useable scientific workflows

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Thanks

Workflows for biological sequence analysis are discussed by the "Pipelines collaboration"

Research on SciWMS supported by the MRC and Prof Christoffels

Professor Alan Christoffels

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE
RESPICE PROSPICE

Bibliography I

Abizar Lakdawalla. File:Sequencing workflow.jpg - wikipedia, the free encyclopedia, Apr. 2007. URL http://en.wikipedia.org/wiki/File:Sequencing_workflow.jpg.

I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. Kepler: an extensible system for design and execution of scientific workflows. In *16th International Conference on Scientific and Statistical Database Management, 2004. Proceedings*, pages 423– 424. IEEE, June 2004. ISBN 0-7695-2146-0. doi: 10.1109/SSDM.2004.1311241.

Anthony J. G. Hey, Stewart Tansley, and Kristen Michelle Tolle. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, Oct. 2009. ISBN 9780982544204. URL http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

L. Bernstein. Software investment strategy. *Bell Labs Technical Journal*, 2(3):233, 1997. ISSN 10897089.

C. Titus Brown. Data intensive science, and workflows, Dec. 2011. URL http://ivory.idyll.org/blog/data-intensive-science-and-workflows.html.

C. Titus Brown. Thoughts on the assemblathon 2 paper, 2013. URL http://ivory.idyll.org/blog/thoughts-on-assemblathon-2.html.

Eagle Genomics. The elements of bioinformatics, 2013. URL http://elements.eaglegenomics.com/.

E. Gamma, R. Helm, R. Johnson, and J. Vlissides. Design patterns: Abstraction and reuse of object-oriented design. In O. M. Nierstrasz, editor, *ECOOP' 93 — Object-Oriented Programming*, number 707 in Lecture Notes in Computer Science, pages 406–431. Springer Berlin Heidelberg, Jan. 1993. ISBN 978-3-540-57120-9, 978-3-540-47910-9.

J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8), 2010.

SANBI South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE

Bibliography II

D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34(suppl 2):W729, 2006. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538887/.

L. N. Joppa, G. McInerny, R. Harper, L. Salido, K. Takeda, K. O'Hara, D. Gavaghan, and S. Emmott. Troubling trends in scientific software use. *Science*, 340(6134):814–815, May 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1231535. URL http://www.sciencemag.org/content/340/6134/814. PMID: 23687031.

I. Karsch-Mizrachi, Y. Nakamura, and G. Cochrane. The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 40(D1):D33–D37, Jan. 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr1006. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3244996/. PMID: 22080546 PMCID: PMC3244996.

S. P. Sadedin, B. Pope, and A. Oshlack. Bpipe: A tool for running and managing bioinformatics pipelines. *Bioinformatics*, Apr. 2012. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts167. URL http://bioinformatics.oxfordjournals.org/content/early/2012/04/11/bioinformatics.bts167.

Taverna project. Why use workflows?, 2009. URL http://www.taverna.org.uk/introduction/why-use-workflows/.

K. Wetterstrand. DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP), 2013. URL http://www.genome.gov/sequencingcosts/.

SANBI
South African National Bioinformatics Institute

UNIVERSITY of the WESTERN CAPE