

Architektura dużych projektów bioinformatycznych

Bartek Wilczyński

bartek@mimuw.edu.pl

<http://www.mimuw.edu.pl/~bartek>

Wykład 5. - Reprodukowalne badania naukowe
- Od systemów LI(M)S do Galaxy
27. III 2019

Problem?



"Think this is bad? You should see the inside of my head."

Analysis

Nature Genetics **41**, 149 - 155 (2009)

Published online: 28 January 2008 | doi:10.1038/ng.295

Repeatability of published microarray gene expression analyses

See associated Correspondence: [Baggerly, *Nature* **467**, 401 \(September 2010\)](#)

John P A Ioannidis^{1,2,3}, David B Allison⁴, Catherine A Ball⁵, Issa Coulibaly⁴, Xiangqin Cui⁴, Aedín C Culhane^{6,7}, Mario Falchi^{8,9}, Cesare Furlanello¹⁰, Laurence Game¹¹, Giuseppe Jurman¹⁰, Jon Mangion¹¹, Tapan Mehta⁴, Michael Nitzberg⁵, Grier P Page^{4,12}, Enrico Petretto^{11,13} & Vera van Noort¹⁴

Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data are publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis. Repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered.

To read this story in full you will need to login or make a payment (see right).

MORE ARTICLES LIKE THIS

These links to content published by NPG are automatically generated.

ARTICLE LINKS

- ▶ Figures and tables
- ▶ Supplementary info

ARTICLE TOOLS

- ✉ Send to a friend
- 📄 Export citation
- 📄 Export references
- 🔒 Rights and permissions
- 📄 Order commercial reprints

SEARCH PUBMED FOR

- ▶ John P A Ioannidis
- ▶ David B Allison
- ▶ Catherine A Ball
- ▶ Issa Coulibaly
- ▶ Xiangqin Cui
- ▶ Aedín C Culhane
- ▶ [more authors of this article](#)

I want to buy this article via ReadCube

Rent: \$4.99*

Purchase: \$9.99*

*Printing and sharing restrictions apply

Purchase now

I want to subscribe to *Nature Genetics*

Subscribe now

Personal subscribers to *Nature Genetics* can view this article. To do this, associate your subscription with your registration via the [My Account](#) page. If you already have an active subscription, [login here](#) to your nature.com account. View our [privacy policy](#) and [use of cookies](#).

If you do not have access to the article you require, you can purchase the article (see below) or access it through a [site license](#). Institutions can add additional archived content to their license at any time. [Recommend](#) site license access to your institution.

[Login via your institution](#)

[Login via OpenAthens](#)

Email:

Password:

COMPUTER SCIENCE

Accessible Reproducible Research

Jill P. Mesirov

Scientific publications have at least two goals: (i) to announce a result and (ii) to convince readers that the result is correct. Mathematics papers are expected to contain a proof complete enough to allow knowledgeable readers to fill in any details. Papers in experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension.

Over the past ~35 years, computational science has posed challenges to this traditional paradigm—from the publication of the four-color theorem in mathematics (1), in which the proof was partially performed by a computer program, to results depending on computer simulation in chemistry, materials science, astrophysics, geophysics, and climate modeling. In these settings, the scientists are often sophisticated, skilled, and innovative programmers who develop large, robust software packages.

As use of computation in research grows, new tools are needed to expand recording, reporting, and reproduction of methods and data.



between two types of acute leukemia, based on language that can produce all of the text, figures,

Zarządzanie danymi w bioinformatyce

- Dostępne i powtarzalne wyniki badań
(Accessible and Reproducible Research)
- RREnvironment, RRSysystem, RRPublishing
- Systemy LI(M)S, Zarządzanie protokołami
(Labkey, BASE, etc...)
- Literate programming (WEB, sweave, etc...)
- Systemy RRS – MeV, GenePattern,
GenomeSpace, Galaxy

RREnvironment - środowiska

- Środowiska pozwalające na programowanie w sposób powtarzalny, pozwalające na dzielenie się wynikami
- Wcześniej często w oparciu o pliki sesji/notatniki lokalnie
- Obecnie najczęściej w oparciu o www
- Jupyter notebooks
- Rstudio, Eclipse
- Matlab notebooks, mathematica notebooks

Jupyter notebook

knuthweb.pdf x Literate Program... x Pweave example galle... x Sage Synapse: Con... x Sage Synapse: Con... x Project Jupyter | A... x Home x test-1 x Access : Repeatabi... x

localhost:8888/notebooks/test-1.ipynb

literate programming example

jupyter test-1 Last Checkpoint: 09/30/2016 (autosaved)

File Edit View Insert Cell Kernel Help Python 3

Code CellToolbar

```
In [1]: print 10
File "<ipython-input-1-0d5b6a5ad0a3>", line 1
      print 10
          ^
SyntaxError: Missing parentheses in call to 'print'
```

```
In [2]: print(10)
10
```

```
In [3]: import pylab
```

```
In [13]: pylab.plot([1,2,3],[4,3,5],"x-r")
Out[13]: [<matplotlib.lines.Line2D at 0x7f94032252b0>]
```

x	y
1	4
2	3
3	5

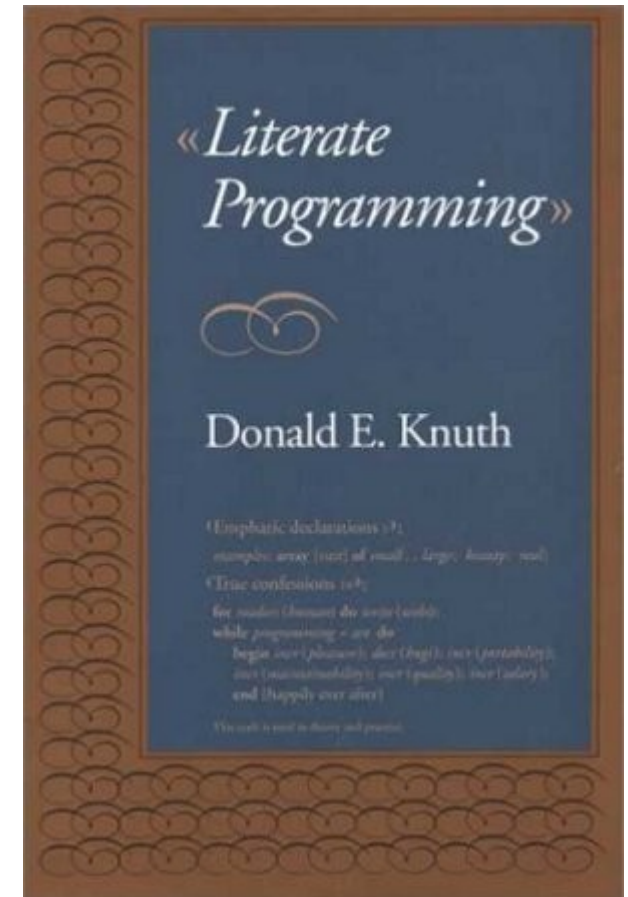
```
In [6]: import seaborn as sns
```

```
In [9]: pylab.show()
```

```
In [10]: %matplotlib inline
```

Literate programming

- Idea “programowania literackiego” wprowadzona przez Donalda Knuth'a w książce w 1984.
- Program komputerowy jest “zanurzony” w tekście opisującym co robi i dlaczego
- Oryginalny pakiet do TeX'a, WEB powstał jeszcze w latach 80tych
- Obecnie wiele implementacji: Sweave dla R, PyWEB i Pweave dla Pythona



Pakiety LIMS

- Laboratory Information Management Software
- Zyskuje popularność od lat 80tych XX w. (era PC)
- Obecnie w zasadzie dwie możliwości:
 - desktop (w oparciu o lokalną bazę danych)
 - Klient-serwer, zwykle przez sieć www, choć istnieją jeszcze rozwiązania z lokalną bazą danych w modelu klient-serwer

Budowa pakietu LIMS

- Baza danych odczynników
- Baza przeprowadzonych analiz
- Linie komórkowe, populacje zwierząt laboratoryjnych
- Eksperymenty przeprowadzane na bieżąco
- Artykuły i rysunki, wyniki częściowe
- Coraz większe rozmiary danych pośrednich

Pakiety LIMS

- Rynek rozczłonkowany pomiędzy bardzo wiele rozwiązań komercyjnych, często bardzo specjalizowanych:

Accelrys LIMS from Accelrys
AgiLIMS from AgiLab
ApolloLIMS from Common Cents Systems, Inc
benchsys benchsys
Biotracker from Ocimum Bio Solutions
Biotracker Lite from Ocimum Bio Solutions
CaliberLIMS from Caliber Technologies Pvt. Ltd.
Care Med LIS from Medcare International
CCLAS from Ventyx, an ABB company, formerly Mincom (company)
Clarity LIMS from GenoLogics Life Sciences
CloudLIMS from CloudLIMS
Cyberlab in Cloud (Toplab) from Megaweb websites: Megaweb and TOPLAB
Darwin from Thermo Fisher Scientific
ELab from LabLynx
Element LIMS from Promium
eQMS::LIMS Pardus d.o.o. [1]
Exemplar Biomarker Discovery from Sapio Sciences
Exemplar Dx LIMS from Sapio Sciences
Exemplar Research LIMS from Sapio Sciences
Galileo from Thermo Fisher Scientific
LABAsistan from Tenay Medical Software

LABbase from Analytik Jena
LabPlus PRÉVENTION EXPERT CONSEIL INC. (PEC)
LabSoft LIMS from Computing Solutions Inc.
LABVANTAGE from LABVANTAGE Solutions
Labware from LabWare
LABWORKS from PerkinElmer
Labway-LIMS from Ambidata Digital Innovation Solutions & Consulting
LDMS from Frontier Science and Technology Research Foundation
Matrix Gemini from Autoscribe
MetaField Lab from Agile Frameworks, LLC
Nautilus from Thermo Fisher Scientific
ProlabQ from Open-Co
readyLIMS from Analytik Jena
Result Point from Accelerated Technology Laboratories, Inc
SampleManager from Thermo Fisher Scientific
Sample Master from Accelerated Technology Laboratories, Inc
Select Agent Inventory (SAI) Management System Foxspire
SchuyLab from Schuyler House, Inc
SIMATIC IT Unilab from Siemens
SLims from Genohm
Solution Laboratoire from Limseo
SmartLims from SmartSoft, Inc
STARLIMS from STARLIMS Corporation
StrainControl Laboratory Manager from DNA Globe
TITAN from Accelerated Technology Laboratories, Inc
TremoLAB from Binsol S.A. www.binsol.com.ar
Watson from Thermo Fisher Scientific
webLIMS from LabLynx
WinLIMS from QSI Corporation N
NuGenesis from Waters Corporation www.Waters.com

Od niedawna również oprogramowanie open source

- Labkey Server (apache license)
- MISO (GPLv3)
- BIKA LIMS (AGPLv3)
- Typowy model to komercyjna firma rozwijająca oprogramowanie i świadcząca usługi wsparcia
- Często dużo tańsze od rozwiązań komercyjnych, ale wymagające większego know-how na miejscu, popularne w dużych instytucjach

Pakiety typu open lab notebook

... there is a URL to a laboratory notebook that is freely available and indexed on common search engines. It does not necessarily have to look like a paper notebook but it is essential that all of the information available to the researchers to make their conclusions is equally available to the rest of the world

—Jean-Claude Bradley



Dla eksperymentów mikromacierzowych

- System BASE

BMC Bioinformatics



Software

Open Access

BASE - 2nd generation software for microarray data management and analysis

Johan Vallon-Christersson^{1,2}, Nicklas Nordborg³, Martin Svensson³ and Jari Häkkinen^{*1}

Address: ¹Department of Oncology, Clinical Sciences, Lund University, SE-221 84 Lund, Sweden, ²CREATE Health Strategic Centre for Translational Cancer Research, Lund University, SE-221 84 Lund, Sweden and ³Department of Theoretical Physics, Lund University, Sölvegatan 14a, SE-223 62 Lund, Sweden

Email: Johan Vallon-Christersson - johan.vallon-christersson@med.lu.se; Nicklas Nordborg - nicklas@thep.lu.se; Martin Svensson - martin@thep.lu.se; Jari Häkkinen^{*} - jari.hakkinen@med.lu.se

^{*} Corresponding author

Published: 12 October 2009

BMC Bioinformatics 2009, **10**:330 doi:10.1186/1471-2105-10-330

Received: 23 June 2009

Accepted: 12 October 2009

Do eksperymentów sekwencjonowania

- Galaxy server

The screenshot displays the Galaxy web interface. At the top, a navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A sidebar on the left lists various tools under categories like 'Get Data', 'Text Manipulation', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analyses', 'Genome Diversity', and 'NGS TOOLBOX BETA'. The main content area features a central banner that reads 'Try Galaxy on the Cloud' with a subtext 'Now you can have a personal Galaxy within the infinite Universe'. To the right of the banner is a 'Tweets' section showing three tweets from the Galaxy Project (@galaxyproject). On the far right, a 'History' panel shows 'Unnamed history' with 0 bytes and a message indicating the history is empty. The bottom of the page contains logos for Penn State, Johns Hopkins University, TACC, and iPlant Collaborative, along with text describing the Galaxy Team's affiliation and the infrastructure provided by the iPlant Collaborative and TACC.

Galaxy

https://usegalaxy.org

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

New high performance job execution options are available! See [the wiki](#) for more information.

Tools

search tools

Get Data

Lift-Over

Text Manipulation

Convert Formats

FASTA manipulation

Filter and Sort

Join, Subtract and Group

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

Genome Diversity

NGS TOOLBOX BETA

Phenotype Association

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Peak Calling

NGS: RNA-seq

NGS: Picard (beta)

NGS: Variant Analysis

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).

Try Galaxy on the Cloud

Now you can have a personal Galaxy within the infinite Universe

Tweets

Galaxy Project @galaxyproject 41m
Fall 2014 Galaxy User Group Grand Oust (GUGGO) Events Report: Tools, User Group, RAD-Seq: [bit.ly/1vquYOm](#) #usegalaxy @Biogenouest

Galaxy Project @galaxyproject 31 Oct
Next Generation Data Analysis Workshop, Dec 5-8, 2014 @UCRiverside [bit.ly/ucrworkshops](#) #usegalaxy

Galaxy Project @galaxyproject 31 Oct
Research Specialist, Institute for Cyber-Enabled Research, Michigan State University, United States [bit.ly/13pqd0T](#) #usegalaxy

History

search datasets

Unnamed history

0 bytes

This history has been deleted

This history is empty. You can [load your own data](#) or [get data from an external source](#)

PENNSTATE

JOHNS HOPKINS UNIVERSITY

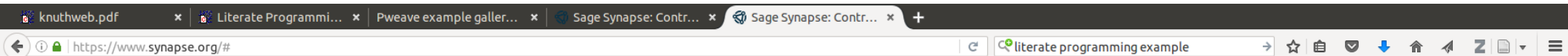
TACC

iPlant Collaborative

The Galaxy Team is a part of the [Center for Comparative Genomics and Bioinformatics](#) at Penn State, and the [Department of Biology](#) and at [Johns Hopkins University](#).

This instance of Galaxy is utilizing infrastructure generously provided by the [iPlant Collaborative](#) at the [Texas Advanced Computing Center](#), with support from the [National Science Foundation](#).

Synapse by SAGE

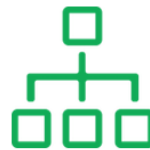


Search B Bartek Wilczynski (barwil)



Organize your digital research assets

Create a free Synapse Project to store your research data, code, and results.



Get credit for your research

Mint a DOI for your work - and describe exactly what you did using Synapse provenance.



Collaborate

Share your Project with your collaborators, or make it Public!

WELCOME BACK, BARWIL

My Dashboard

Synapse @SageSynapse
A great resource in Alzheimer's Disease:
[nature.com/articles/sdata...](https://www.nature.com/articles/sdata...) and the associated Synapse
portal: [synapse.org/#!/Synapse:syn2...](https://www.synapse.org/#!/Synapse:syn2...) @ScientificData

Human whole genome genotype and transcriptome data...
Previous genome-wide association studies (GWAS),
conducted by our group and others, have identified loci that
harbor risk variants for neurodegenerative diseases.

HELP

Getting Started

Synapse is an open source software platform that data scientists can use to carry out, track, and communicate their research in real time.

[Open the guide](#)

Programmatic Clients

Synapse is designed to easily integrate into your current work. That's why we've created the following clients so that you can interact with all of Synapse's functionality programmatically. Create projects, upload & download files, generate provenance, query, create wikis and more all from the comfort of your own code.

Don't see your language of choice here? Check out our [full REST API documentation](#)

R client

Python client

Command line client

Java client

EXPLORE HOW RESEARCHERS ARE USING SYNAPSE

Standardy depozycji danych

- MIAME – Minimum Information About a Microarray Experiment (Brazma et al. 2001)
- Standardowe repozytoria danych ArrayExpress (EU), GeneExpression Omnibus (USA)

1 MI Standards

- 1.1 MIAPPE, Minimum Information About a Plant Phenotyping Experiment
- 1.2 MIAME, gene expression microarray
- 1.3 MINI: Minimum Information about a Neuroscience Investigation
 - 1.3.1 MINI: Electrophysiology
- 1.4 MIARE, RNAi experiment
- 1.5 MIACA, cell based assay
- 1.6 MIAPPE, proteomic experiments
- 1.7 MIMIX, molecular interactions
- 1.8 MIAPAR, protein affinity reagents
- 1.9 MIABE, bioactive entities
- 1.10 MIGS/MIMS, genome/metagenome sequences
- 1.11 MIFlowCyt, flow cytometry
- 1.12 Minimum Information about a Flow Cytometry Experiment
- 1.13 MISFISHIE, In Situ Hybridization and Immunohistochemistry Experiments
- 1.14 MIAPA, Phylogenetic Analysis
- 1.15 MIRAGE, Glycomics
- 1.16 MIAO, ORF
- 1.17 MIAMET, METabolomics experiment
- 1.18 MIAFGE, Functional Genomics Experiment
- 1.19 MIRIAM, Minimum Information Required in the Annotation of Models
- 1.20 MIASE, Minimum Information About a Simulation Experiment
- 1.21 CIMR, Core Information for Metabolomics Reporting