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

25. III 2020

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 Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

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.

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

COMPUTER SCIENCE

Accessible Reproducible Research

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

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.

between two types of acute leukemia, based on

guage that can produce all of the text, figures,

Zarządzanie danymi w bioinformatyce

- Dostępne i powtarzalne wyniki badań (Accessible and Reproducible Research)
- RREnvironment, RRSystem, 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 | Literate Program... | Pweave example galle... | Sage Synapse: Con... | Sage Synapse: Con... | Project Jupyter | A... | Home | test-1 | Access : Repeatabi...

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

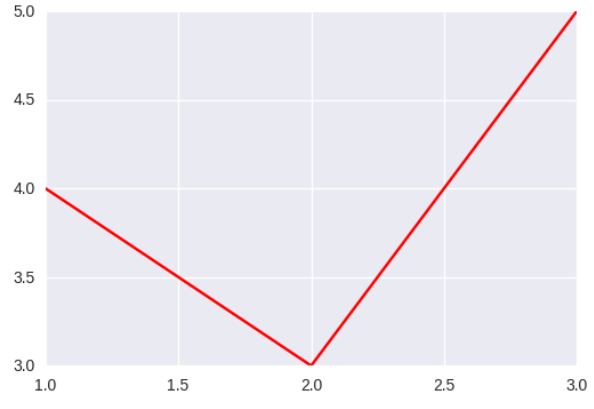
Code CellToolbar

```
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>]
```



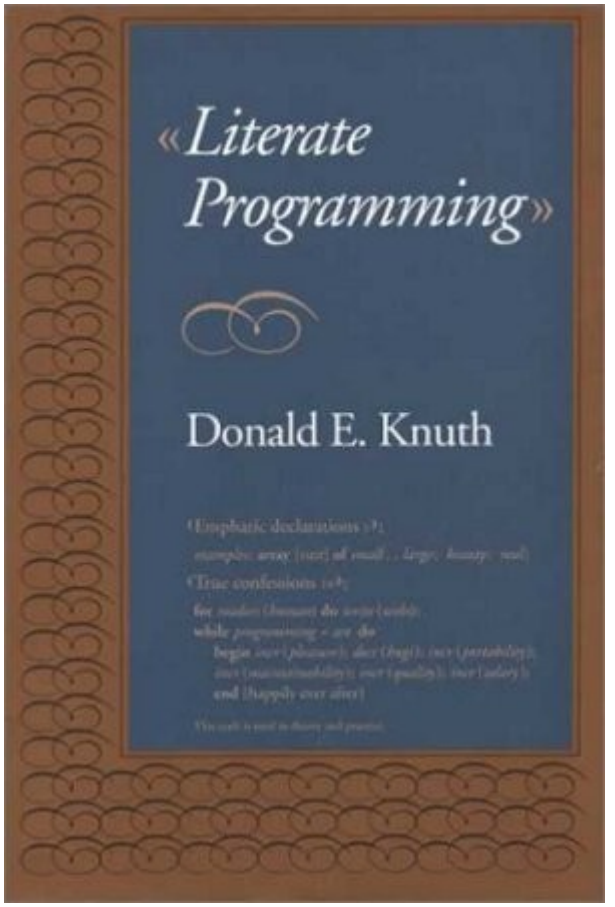

```
In [6]: import seaborn as sns
```

```
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/roślin/organizmów 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 wyspecjalizowanych:

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

BioMed Central

Software

Open Access

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

Johan Vallon-Christersson[1,2], Nicklas Nordborg[3], Martin Svensson[3] and Jari Häkkinen*[1]

Address: [1]Department of Oncology, Clinical Sciences, Lund University, SE-221 84 Lund, Sweden, [2]CREATE Health Strategic Centre for Translational Cancer Research, Lund University, SE-221 84 Lund, Sweden and [3]Department 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

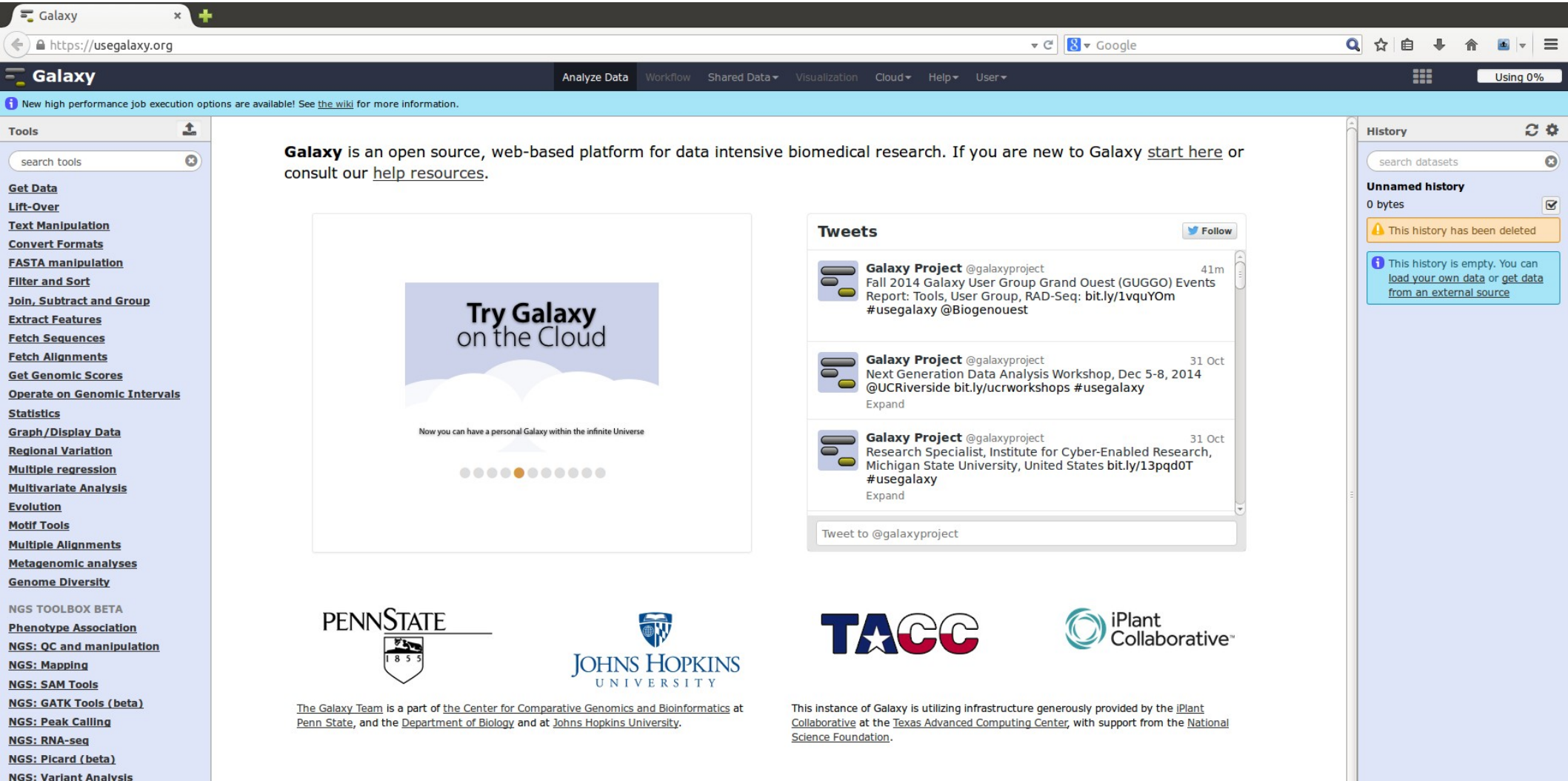
Published: 12 October 2009

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

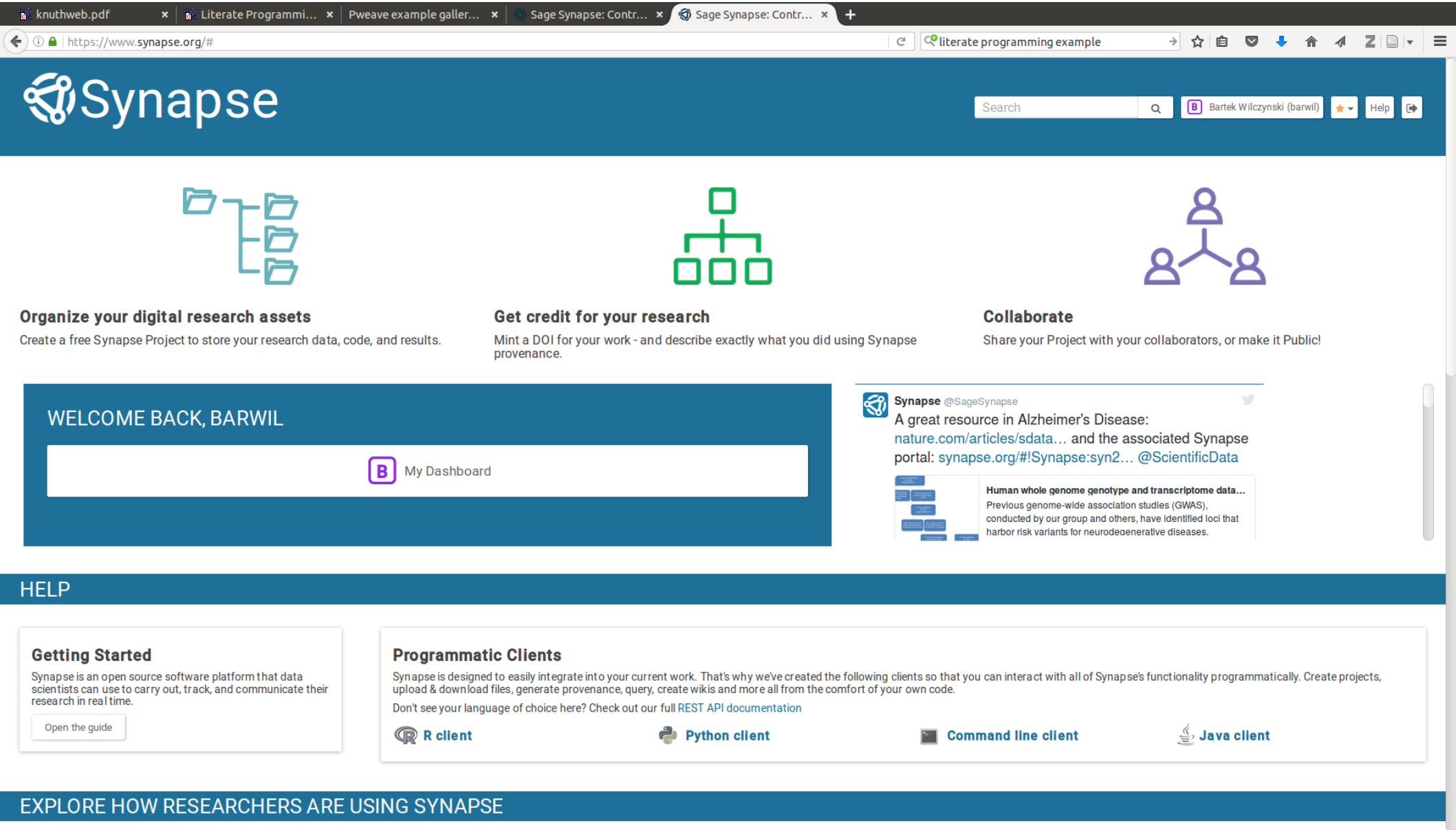
Received: 23 June 2009
Accepted: 12 October 2009

Do eksperymentów sekwencjonowania

- Galaxy server



Synapse by SAGE



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 MIAPE, 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