GigaScience

# Experimenting with reproducibility in bioinformatics

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | GIGA-D-17-00317 |
| **Full Title:** | Experimenting with reproducibility in bioinformatics |
| **Article Type:** | Review |
| **Funding Information:** | Institut Pasteur (FR) — Not applicable<br>H2020 Health — Not applicable<br>Institut National des Sciences de l'Univers, Centre National de la Recherche Scientifique (FR) — Not applicable<br>Université Paris Diderot (FR) — Not applicable<br>Conny-Maeva Charitable Foundation — Not applicable<br>Cognacq-Jay Foundation — Not applicable<br>Orange — Not applicable<br>Fondation pour la Recherche Médicale — Not applicable<br>GenMed Labex — Not applicable<br>BioPsy Labex — Not applicable |
| **Abstract:** | Reproducibility has been shown to be limited in many scientific fields. This question is a fundamental tenet of the scientific activity, but the related issues of reusability of scientific data are poorly documented. Here, we present a case study of our attempt to reproduce a promising bioinformatics method [1] and illustrate the challenges to use a published method for which code and data were available. First, we tried to re-run the analysis with the code and data provided by the authors. Second, we reimplemented the method in Python to avoid dependency on a MATLAB licence and ease the execution of the code on HPCC (High-Performance Computing Cluster). Third, we assessed reusability of our reimplementation and the quality of our documentation. Then, we experimented with our own software and tested how easy it would be to start from our implementation to reproduce the results, hence attempting to estimate the robustness of the reproducibility. Finally, in a second part, we propose solutions from this case study and other observations to improve reproducibility and research efficiency at the individual and collective level. |
| **Corresponding Author:** | Yang-Min KIM<br>Institut Pasteur<br>Paris, Île-de-France FRANCE |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | Institut Pasteur |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Yang-Min KIM |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Yang-Min KIM<br>Jean-Baptiste Poline<br>Guillaume Dumas |
| **Order of Authors Secondary Information:** | |
| **Opposed Reviewers:** | Matan Hofree |

| | |
|---|---|
| | Broad Institute<br>mhofree@broadinstitute.org<br>our reproducibility work is based on his paper |
| | Trey Ideker<br>University of California San Diego Ideker Laboratory<br>tideker@ucsd.edu<br>our reproducibility work is based on his lab's paper |
| **Additional Information:** | |
| **Question** | **Response** |
| Are you submitting this manuscript to a special series or article collection? | No |
| **Experimental design and statistics**<br><br>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.<br><br>Have you included all the information requested in your manuscript? | No |
| If not, please give reasons for any omissions below.<br><br>as follow-up to "**Experimental design and statistics**<br><br>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.<br><br>Have you included all the information requested in your manuscript?<br><br>" | This review is especially based on the reproducibility issues. The illustrated method is directly based on the original paper by Hofree et al, 2013. We nevertheless described any specific variations. |
| **Resources**<br><br>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. | Yes |

| | |
|---|---|
| Have you included the information requested as detailed in our Minimum Standards Reporting Checklist? | |
| **Availability of data and materials**<br><br>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.<br><br>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist? | Yes |

# Experimenting with Reproducibility in Bioinformatics

## Authors

Yang-Min KIM [1,2,3,4,*] , Jean-Baptiste Poline [5,6] , Guillaume Dumas [1,2,3,4]

[1]Institut Pasteur, Human Genetics and Cognitive Functions Unit, Paris, France, [2]CNRS UMR 3571 Genes, Synapses and Cognition, Institut Pasteur, Paris, France, [3]University Paris Diderot, Sorbonne Paris Cité, Paris, France, [4]Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI, USR 3756 Institut Pasteur and CNRS), Paris, France, [5]Montreal Neurological Institute, Brain Imaging Center, Ludmer Center, McGill University, [6]Henry H. Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California, Berkeley, California, USA

*To whom correspondence should be addressed

Correspondence: yang-min.kim@pasteur.fr; jbpoline@gmail.com; guillaume.dumas@pasteur.fr

## Abstract

Reproducibility has been shown to be limited in many scientific fields. This question is a fundamental tenet of the scientific activity, but the related issues of reusability of scientific data are poorly documented. Here, we present a case study of our attempt to reproduce a promising bioinformatics method [1] and illustrate the challenges to use a published method for which code and data were available. First, we tried to re-run the analysis with the code and data provided by the authors. Second, we reimplemented the method in Python to avoid dependency on a MATLAB licence and ease the execution of the code on HPCC (High-Performance Computing Cluster). Third, we assessed reusability of our reimplementation and the quality of our documentation. Then, we experimented with our own software and tested how easy it would be to start from our implementation to reproduce the results, hence attempting to estimate the robustness of the reproducibility. Finally, in a second part, we propose solutions from this case study and other observations to improve reproducibility and research efficiency at the individual and collective level.

**Availability:** last version of StratiPy (Python) with two examples of reproducibility and dataset are available at GitHub [2].

## Keywords

- Reproducibility
- Robustness
- Reusability
- Network Based Stratification (NBS)

- Standard consensus dataset
- Cancer

# 1 Background

The collective endeavour of science depends on researchers being able to replicate the work of others. In a recent survey of 1,576 researchers, 70% of them admitted having difficulty in reproducing experiments proposed by other scientists [3]. For 50%, this reproducibility issue even concerns with their own experiments. Despite the growing attention on the replication crisis in science [4,5], this controversial subject is far from being new: already in the 17th century, scientists criticized the air pump invented by physicist Robert Boyle because it was too complicated and expensive to build [6].

Several concepts for reproducibility in computational science are closely associated [7,8]. Here we define them as mentioned by K. Whitaker [8]: obtaining the same results using same data and same code is ***Reproducibility***; if code is different, it is ***Robustness***. If we used different data but with the same code, it is ***Replicability***. lastly, using different data and different code is referred as ***Generalisability***. Here we will primarily elaborate on ***Reproducibility*** and ***Robustness***. Indeed, it takes great efforts and competence to overcome all the obstacles to reproduce successfully an experiment. The process is costly in resources, both in time and funding. In computational science, there are also many technical barriers ranging from unavailable data to hardware infrastructure [9]. Even when authors provide data and code, the outcome can vary either marginally or fundamentally [10]. Tackling irreproducibility in bioinformatics thus requires considerable effort beyond code and data availability. In most cases, there is a significant gap between apparent executable work (Fig 1 - i.e. above water portion of iceberg) and necessary effort in practice (Fig 1 - i.e. full iceberg). Such effort is nevertheless necessary to increase the consistency of the literature and efficiency of the scientific research process. Indeed, behind reproducibility hides reusability.

# 2 Reproducibility and Robustness in bioinformatics: a case study

## 2.1 Reproducibility: from MATLAB to MATLAB, OS and environment

Our team studies Autism Spectrum Disorders (ASD), a group of neurodevelopmental disorders well known for its heterogeneity. One of the current challenges of our research is to uncover homogeneous subgroups of patients (i.e. stratification) with more precise clinical outcomes, improving their prognosis and treatment [11,12]. An interesting stratification method was recently proposed in the field of cancer research [1], where the authors proposed to combine genetic profiles of patients' tumours with protein-protein interaction networks to uncover meaningful homogeneous subgroups, a method called Network Based Stratification (NBS).

Before using this NBS method on our data, we studied the method by reproducing results from the original study. We are very grateful to the main authors who kindly provided online all the related data and code, and gave us invaluable input upon request. The authors of this study thus should not be blamed for the difficulty that we experienced in attempting to reproduce and to make more robust their study, as they did more to help reproduce their results than is generally done. Despite their help we experienced a number of difficulties that we document here, hoping that this report will help future researchers to improve the reproducibility of results and reusability of research products.

The first step of our project was to execute the original method code with the given data: reproducibility. The programming code was written in MATLAB, an interpreted language originally developed for linear algebra computations which is easier and faster to write as well as more readable than compiled language such as C, making our reproducibility attempt easier. To improve execution speed, the original authors used a library for MATLAB using executable compiled code MEX file [13] callable from MATLAB: MTIMESX [14], a library with compiled code allowing acceleration of large matrix multiplication. MEX files however are specific to the architecture and have to be recompiled for each Operating System (OS). The original MEX file was initially developed for Linux. Since our lab was using Mac OS X Sierra, the compilation of this MEX file into a mac64 binary required a new version of MTIMESX. It was also necessary to install and to configure properly OpenMP [15], a development library for parallel computing. After this, the original MATLAB code was successfully run in our environment.

These issues are classic, but may not be overcome by researchers with little experience in compilation or installation issues. For these reasons alone, many individuals may turn down the opportunity of reusing code.

The next part will focus on code re-implementation, a procedure, which can help understanding the method, but can be even more costly.

## 2.2 Robustness: from MATLAB to Python, language and organization

To fully master the method, adapt it to our data, and ease its reuse, we developed a complete open source toolkit of genomic stratification in Python [2]. Python is also an interpreted programming language, but contrary to MATLAB is free of use and has a GPL-compatible license [16], which fosters both robustness and generalizability. Recoding in another language in a different environment will lead to be some unavoidable problems such as variation in low level libraries (e.g. glibc): it is likely that the outcomes will vary even if the same algorithm is implemented, and therefore should indicate how sensitive the proposed method results are with respect to these settings. In addition, we rely on Python packages to perform visualization or linear algebra computations (e.g. Matplotlib, SciPy, NumPy), and results may depend on these packages versions. Python is currently in a transitional period between two major versions 2 and 3. We chose to write the code in Python 3, which is the current recommendation.

### 2.2.1 Metadata and File formats

Even if the original code could be run, we had to handle several file formats to check and understand the structure of the original data. For instance the data on patients with cancer data was provided by The Cancer Genome (TCGA) [17] and made available in a MATLAB *.mat* file format. Thanks to SciPy, Python can load all versions through v7.2 MATLAB files. To read v7.3 *.mat* files, we however needed an HDF5 Python library. Moreover, the original authors had denoted download dates of patients' data of TCGA, thereby clarifying source of data. But in the absence of structural metadata, it was not always obvious how to interpret patients' dataset variables (e.g. patient ID, gene ID, phenotype). Fig 2 shows an analogy between robustness issues and road transport: driving in a different environment (e.g. OS), we attempt to obtain identical results (i.e. to reach the same location) using the same input data of TCGA (i.e. gasoline). But there is a difficult to transfer proposed method (i.e. engine) from one programming language to another (i.e. MATLAB and Python roads).

### 2.2.2 Codes and parameters

Once the environment and file format issues were resolved, the code was finally executable with genetic data. Unfortunately, several attempts produced error messages. Alternatively, "unexpected" results were obtained: e.g. during the application of hierarchical clustering, we used the clustering tools of SciPy. Both SciPy and MATLAB (MathWorks) functions offer seven linkage methods, however, SciPy's default option (single method) [18] differs from MATLAB's default option (UPGMA or average method) [19], which was used in the original study. Another key example is the value of one of the most important parameters of the method, the graph regulator factor, which was not clarified in the original paper. We believed that this factor had a constant value of 1.0 until we found in the code that during iterations, its value was changing and converged to a high optimal value (~1800). Therefore, we initially obtained very different results from the original NBS (Fig 3). We observed heterogeneous subgroups instead of obtaining homogeneous clusters. No or little explanation on the parameter choices can explain variability in the results as we explored the possible parameters range. Moreover, during our attempts to run the original code to understand the causes of the errors, we realized that some parts of the code were not run anymore (e.g. discarded work, remaining traces of debugging) which made the attempt to understand the implementation harder.

To allow others to reproduce our results, we created documentation and tutorials for the associated Python package *StratiPy* [2]. Readers are thus invited to experiment with reproducibility too by generating the two confusion matrices of Fig 3. They will use the different tools described in the following: GitHub, Docker, and Jupyter/IPython notebook.

### 2.2.3 Jupyter/IPython

During the re-cording process, we used an enhanced Python interpreter to debug: IPython, an interactive shell supporting both Python 2 and 3. Since the dataset is large and the execution takes a significant amount of time, we used IPython to re-run interactively some sub-sections of the script, which is one of the most helpful features. IPython can be integrated in the web

interface Jupyter Notebook, offering an advanced structure for mixing code and documentation. For instance, verifying intermediate results by plotting helped us to better understand the original code. While the Jupyter/IPython notebook was therefore initially convenient, it does not scale well and is not well adapted to versioning. However, ability of mixing code with document text is very useful for tutorials: a user of the code can read documentation (docstring), text explanations, and see how to run the code, explore parameters and visualize results in the browser. Our work on NBS, as related here, can be reproduced with a Jupyter/IPython notebook available on our GitHub [2]. You can find more examples and several helpful links on this "gallery of interesting Jupyter Notebooks" [20], which even contains a section about "Reproducible academic publications".

## 2.3 Reproducibility of Robustness: from Python to Python

Besides Jupyter/IPython notebook, we used versioning tools like the git code version control system (VCS) to document the development of our Python code. Git is arguably one of the most powerful VCS, allowing easy development of branches and helping us to work together as a distributed team (Paris, Berkeley) on the same project. This project, *StratiPy*, is hosted on GitHub, a web-based Git repository hosting service [2]. While the original code was not available on GitHub, the main authors shared their code on a website. This should be sufficient for our purpose, but makes it less easy to collaborate on code. While working on our GitHub repository, several researchers from all over the world contacted us about our robustness experiment. Not only GitHub supports a better organization of projects, it also facilitates the collaboration of open-source software projects, thanks to several social network functions [21]. We tried to comply with open source coding standards and to learn how to efficiently use Git and GitHub. Both required considerable efforts on the short-term but brought clear benefits on the long-term, especially regarding collaboration and debugging.

We then attempted to re-run and reproduce the results we obtained on another platform. While the Python code was developed under Mac OS X Sierra (10.12) we used an Ubuntu 16.04.1 (Xenial) computer to test the Python implementation. Some additional issues emerged. First, our initial documentation was not complete enough to know which packages were required and how to launch the code. Second, the code was very slow to the extent that it was impractical to run it on a laptop because the Numpy package had not been compiled with BLAS (Basic Linear Algebra Subprograms), low-level routines performing basic vector and matrix operations. Last, there was (initially) no easy way to check whether the results obtained on a different architecture were the expected ones. We added documentation and tests on the results files md5sum to solve this. To summarize, although the reuse and reproducibility of the results of the developed package were improved, these were far from being optimal.

# 3 Potential solutions: from local to global

## 3.1 Act locally: simple practices and available tools

Given the observed difficulties, we draw some conclusions on this reproducibility case study experiment and suggest some practices and tools. In addition to this guidance, scientists are strongly encouraged to follow detailed advice of Wilson et al. such as modularizing and re-using code, unit testing, document design, data management, and project organization [4,22]. Sandve and colleagues [23] also suggest to keep the data provenance with recording all intermediate results.

### 3.1.1 Environment

Container technologies such as Docker [24], Vagrant [25], or Singularity [26,27] (easily works in cluster environments) are becoming a standard solution to installation issues. These rely however on competencies that we think few biologists possess today. Also, while the container will encapsulate everything needed for the software execution, it could be hard to develop in a container. For instance, running Jupyter/IPython notebooks in Docker's container requires certain knowledge of computer science (e.g. advanced port forwarding), which can become a discouraging task. Therefore, we decided to propose two options in our example implementation of reproducibility: 1) a step-by-step process to follow in a Jupyter/IPython notebook; or 2) a Docker container ready to be built and run. Nevertheless, mastering Docker –or other container tools– will become an important skill for computational reproducible researchers.

### 3.1.2 Metadata

Standard metadata are vital for an efficient documentation of both data and software. In our example, we still lack the standard lexicon to document the data as well as documenting the software: e.g. using HDF5 file instead of *.mat* file is more suitable to store patient's' data. We however aim to follow the recommendations by Stodden *et al.* [28]: *"Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements".* The more comprehensive is the metadata description, the more likely the reuse will be both efficient and appropriate [29].

### 3.1.3 Write readable code

Anyone who has spent time to understand someone else's code would advise some simple basic rules to help make the code readable and understandable.

First, the structure of the program should be clear and easily accessible. Second, good concise code documentation and naming convention will help readability. Third, the code should not contain left-overs of previously tested solutions. When a solution takes a long time to compute, an option to store it locally can be proposed. Using standard coding and documentation conventions

(e.g. PEP 8 and PEP 257 in Python [30,31]) with detailed comments and references of papers makes the code more accessible. When an algorithm from another paper is used, any modification should be explained and discussed in the paper as well as in the code. All these remarks are not necessarily obvious especially if the developer is working on her/his own, and to some extent "writes for her/himself". We advocate for researchers to write code "for their colleagues", hence, the opinion and notice of co-working or partner laboratories should be very helpful. Furthermore, the collaboration between researchers working on different environments can more easily isolate reproducibility problems. In the future, journals may consider review of code as part of the standard review process.

#### 3.1.4 Test the code

To check if the code is yielding a correct answer, software developers associate test suites (unit tests or integration tests) with their software. While we developed only a few tests in this project, we realize that this has a number of advantages, such as checking if the software installation seems correct, check if updates in the operating system impact the results, etc. This does not in general validate the method, but at least provides a basic check. In our case, we propose to check for the integrity of the data and for the results of some key processing.

## 3.2 Think globally: from education to community standards

### 3.2.1 Training the new generation of scientists to digital tools and practices

Unlike theoretical and academic courses and projects, software testing systems are well developed in industry since software quality is not the priority in Academia [32]. For a student, discovering and learning this core system of reproducibility, possibly during an internship in cooperation with industry, is a great opportunity for her/his future. Furthermore, as Internet applications in science are growing, networks of scientists and developers are forming and provide learning opportunities on the development practices. For instance, software developers have recently adopted "agile" practices and fast prototyping, test based development, etc. Some of these ideas and practices can —and should— be adapted to scientific software development.

The training in coding is still too limited for biologists. Often, it is self-training, from searching answers on Stack Overflow or equivalent. Despite efforts by organizations such as Software [33] or Data Carpentry [34] and the growing demand for 'data scientists' in life science, university training on coding practices is still not enough generalized. The difficulty to access and understand code may lead to applying code blindly without checking the validity of the results: often, scientists may prefer to believe that the results are correct because of the time that would be needed to check the validity of the results. Mastering a package such that results are truly understood can take a long time, as it was the case in our experiment.

Academia could instruct young scientists best practices for reproducibility. For instance, Hothorn and Leisch organized a reproducibility workshop gathering mostly PhD students and young postdocs specialized in bioinformatics and biostatistics. Then they evaluated 100 random sample papers from *Bioinformatics [5]*. Their study revealed how such a workshop can raise young

scientists awareness about *"what makes reproduction easy or hard at first hand"*. Indeed, they found out that only a third of the original papers and two-thirds for applications notes had given access to the source code of software used.

### 3.2.2 Standard consensus dataset and workflow system

We propose here that bioinformatics methods publications are systematically accompanied with a test dataset, code source and some basic tests. As the method is tested on new datasets, the number of tests of the method would increase in number and cover a wider range of applications. We give a first example with our NBS re-implementation. We develop below how this could generalize and what would be the benefit for the scientific community. In a sense, we propose to use the software development test framework idea but apply it to the scientific context.

A schematic overview of workflow system is shown in Fig 4. The core of this system would be a standard consensus dataset used to validate methods. For instance in the field of machine learning, standard image databases are widely used for training and testing (e.g. MNIST for handwritten digits [35]). In the case of our proposal, data could be classified in general categories such as binary, text, image (A, B, C in Fig 4 b), and with sub-categories to introduce criteria such as size, quantitative/qualitative, discrete/continuous using a tagging system (e.g. A-2, B-1, C-5 in Fig 4 b). Dataset could be issued from simulations or from acquisition, and would validate a method on a particular component. This workflow system will help scientists that cannot release their data because of privacy issues (Fig 4 a.1) (although these can often be overcome) but also give access to data and tests to a wide community. Scientists can use only standard data at the beginning of the project. And if there is no appropriate data, they have to suggest a new standard data.

Roughly, we divide those who interact with scientific software or analysis code in two categories. First, the authors ("A") who propose a method and need to verify its validity and usefulness with public and/or their own – often private – data. Second, the users ("U", e.g. developers, engineers, bioinformaticians) who need to test and evaluate the proposed methods with other data.

When authors launch a research about a method, this method must belong to a general category of methods (e.g. classification, regression) and could have a reproducibility profile, which will progressively be built by authors and users (Fig 4 b.3, b.4). Even the information of which method does or does not work with a standard data is a crucial information for future work. During optimization of project, the programming code with guide should be accessible although authors do not publish. To achieve current work, they can also post on preprint servers such as bioRxiv [36,37] associated with a GitHub repository by digital object identifiers (DOI).

Furthermore, users who test and approve reproducibility on original or new data could be credited and recognized by the scientific and developer communities (i.e. Stack Overflow, GitHub). This workflow system thus could facilitate the gathering of diverse users of the science community.

# 4 Conclusion and perspective

Across the scientific fields, the reproducibility issue is seen as a growing concern. Before reusing a published method, we attempted to reproduce the initial results and recoded the method to have a deep understanding of it. The investment in time to verify a previously published method can be more important than the work needed to publish a new paper. Despite the willingness of the authors to share their tool and help us in our work, we have faced computational reproducibility and robustness problems due to compatibility between environments, programming languages and software versions, choice of parameters, etc. In addition to individual effort to write well documented and readable code, we recommend to use online repositories and tools to help other scientists in their exploration of the method: Docker for environment standardization, GitHub for code version management, and Jupyter notebooks for demonstration and tutorial [20,21,24]. Scientists are strongly encouraged to adopt such practices, not only for writing code but also manuscripts [4]. At the community level, we should enhance the cooperation between academic education and industry to foster a new generation of well-trained scientists in software development. For instance, Academia-Industry Software Quality & Testing summit (AISTQ) has organised conference in order to encourage collaboration between Academia and Industry [38]. This is also in line with the work of the Software or Data carpentry organization [33,34]. Here, we propose a workflow system where the community uses standard datasets to validate tools. The proposed method success on data profile will be evaluated continuously with new datasets. Eventually, data and software can be versioned and cited to give credit to the individuals who have contributed to these building blocks of Science. This workflow is not merely a reproducibility validation tool, it is an attempt to make research product more reusable by the community using online platforms and open source tools, beyond the publication of a PDF file. Such system could be seen as a generalisation of already existing workflow systems such as Galaxy or GATK, integrating data provenance [39,40]. Some top-down initiatives already provide some incentives for such a process i.e. Horizon 2020 (H2020) [41] project of the European Commission (EC) mandates open access of research data, while respecting security and liability. H2020 supports OpenAIRE [42], a technical infrastructure of the open access, which allows the interconnection between projects, publications, datasets, and author information across Europe. Thanks to common guidelines, OpenAIRE interoperates with other web-based *generalist* scientific data repositories such as Zenodo, hosted by CERN, which allows combining data and GitHub repository using DOI. The Open Science Framework also hosts data and software for a given project [43]. Respecting standard guidelines to transparently communicate the scientific work is a key step towards tackling irreproducibility and insures a robust scientific endeavor.

## Key points

- Main barrier for reproducibility is in the lack of compatibility between environments, programming languages, software versions, etc.

- At the individual level, the key is in research practices such as proper code and data documentation and exploitation of online repositories and collaborative tools.
- At the community level, we propose a workflow system where standard consensus datasets are used to validate new methods and foster their generalizability.

# Declarations

## Ethics approval and consent to participate

We used the uterine endometrial carcinoma data as they were downloaded on January 1st, 2013 from the The Cancer Genome Atlas (TCGA) portal by Hofree and colleagues [1].

## Consent for publication

Not applicable

## Availability of data and material

Last version of StratiPy (Python) with two examples of reproducibility and dataset are available at GitHub [2].

Zenodo DOI: https://doi.org/10.5281/zenodo.1042546

## Competing interests

The authors declare that they have no competing interests.

## Funding

This work was supported by:

- Institut Pasteur (http://dx.doi.org/10.13039/501100003762)
- H2020 Societal Challenges (http://dx.doi.org/10.13039/100010676)
- Centre National de la Recherche Scientifique (http://dx.doi.org/10.13039/501100004794)
- Université Paris Diderot (http://dx.doi.org/10.13039/501100005736)
- Conny-Maeva Charitable Foundation
- Cognacq-Jay Foundation
- Orange (http://dx.doi.org/10.13039/501100003951)
- Fondation pour la Recherche Médicale (http://dx.doi.org/10.13039/501100002915)
- GenMed Labex
- BioPsy Labex.

## • Authors' contributions

Y-M. K., J-B. P., and G.D. wrote the manuscript, Y-M. K. and G.D. developed the StratiPy module.

All authors read and approved the final manuscript.

## • Acknowledgements

We thank Thomas Rolland and Freddy Cliquet for sharing their advices and comments.

## • Authors' information

**Yang-Min KIM** [1,2,3,4,*] is a PhD student at Human Genetics and Cognitive Functions unit at the neuroscience department of the Institut Pasteur in Paris. Her research on next-generation sequencing data and biological networks is focused on stratification of patients with Autism.

*Keywords:* Autism Spectrum Disorder; Network; Protein-Protein Interaction; Personalize Medicine; Network-Based Stratification; Computational Biology.

**Jean-Baptiste Poline** [5,6] is is a researcher at Henry H. Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California, Berkeley, California, USA. His research focuses on neuroimaging methods, imaging-genetic biostatistics and neuroinformatics.

*Keywords:* Neuroinformatics; Statistical methods; Brain imaging; Imaging genomics.

**Guillaume Dumas** [1,2,3,4] is research fellow of the Human Genetics and Cognitive Functions unit at the neuroscience department of the Institut Pasteur in Paris. His interdisciplinary work is at the cross-road of social psychology, cognitive neuroscience, and system biology.

*Keywords:* Open Science; Complex Systems; Computational Biology; Cognitive Science; Social Neuroscience; Autism Spectrum Disorder.

# Reference

1. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat. Methods [Internet]. 2013 [cited 2015 Aug 12];10. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866081/

2. Stratipy: Graph regularized nonnegative matrix factorization (GNMF) in Python [Internet]. GHFC; 2017. Available from: https://github.com/GHFC/Stratipy

3. Baker M. 1,500 scientists lift the lid on reproducibility. Nat. News. 2016;533:452.

4. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good enough practices in scientific computing. PLOS Comput. Biol. 2017;13:e1005510.

5. Hothorn T, Leisch F. Case studies in reproducibility. Brief. Bioinform. 2011;12:288–300.

6. Shapin S, Schaffer S. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life (New in Paper). Princeton University Press; 2011.

7. Peng RD. Reproducible research in computational science. Science. 2011;334:1226–7.

8. Whitaker K. Showing your working: a how to guide to reproducible research [Internet]. 2017. Available from: https://figshare.com/articles/Showing_your_working_a_how_to_guide_to_reproducible_research/5443201

9. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat. Rev. Genet. 2012;13:667–72.

10. Herndon T, Ash M, Pollin R. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Camb. J. Econ. 2014;38:257–79.

11. Bourgeron T. From the genetic architecture to synaptic plasticity in autism spectrum disorder. Nat. Rev. Neurosci. 2015;16:551–63.

12. Loth E, Spooren W, Ham LM, Isaac MB, Auriche-Benichou C, Banaschewski T, et al. Identification and validation of biomarkers for autism spectrum disorders. Nat. Rev. Drug Discov. 2016;15:70–73.

13. Introducing MEX Files - MATLAB & Simulink - MathWorks France [Internet]. [cited 2017 Aug 18]. Available from: https://fr.mathworks.com/help/matlab/matlab_external/introducing-mex-files.html?requestedDomain=www.mathworks.com

14. Tursa. MTIMESX - Fast Matrix Multiply with Multi-Dimensional Support - File Exchange - MATLAB Central [Internet]. 2009 [cited 2017 Apr 24]. Available from: http://fr.mathworks.com/matlabcentral/fileexchange/25977-mtimesx-fast-matrix-multiply-with-multi-dimensional-support

15. tim.lewis. Specifications [Internet]. OpenMP. [cited 2017 Aug 18]. Available from: http://www.openmp.org/specifications/

16. Python Software Foundation. History and License — Python 3.6.1 documentation [Internet]. 2017 [cited 2017 Apr 24]. Available from: https://docs.python.org/3/license.html#licenses-and-acknowledgements-for-incorporated-software

17. TCGA [Internet]. Cancer Genome Atlas - Natl. Cancer Inst. [cited 2017 Apr 24]. Available from: https://cancergenome.nih.gov/

18. Eads. Hierarchical clustering (scipy.cluster.hierarchy) — SciPy v0.19.0 Reference Guide [Internet]. 2007 [cited 2017 Apr 24]. Available from: https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

19. Hierarchical Clustering - MATLAB & Simulink - MathWorks France [Internet]. [cited 2017 Apr 24]. Available from: https://fr.mathworks.com/help/stats/hierarchical-clustering-12.html

20. A gallery of interesting Jupyter Notebooks · jupyter/jupyter Wiki [Internet]. [cited 2017 Aug 18]. Available from: https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks

21. Blischak JD, Davenport ER, Wilson G. A Quick Introduction to Version Control with Git and GitHub. PLoS Comput Biol. 2016;12:e1004668.

22. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. Best Practices for Scientific Computing. Eisen JA, editor. PLoS Biol. 2014;12:e1001745.

23. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. Bourne PE, editor. PLoS Comput. Biol. 2013;9:e1003285-4.

24. Boettiger C. An introduction to Docker for reproducible research, with examples from the R environment. ACM SIGOPS Oper. Syst. Rev. 2015;49:71–9.

25. Introduction [Internet]. Vagrant HashiCorp. [cited 2017 Oct 13]. Available from: https://www.vagrantup.com/intro/index.html

26. Singularity | Singularity [Internet]. [cited 2017 Oct 13]. Available from: http://singularity.lbl.gov/

27. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLOS ONE. 2017;12:e0177459.

28. Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, et al. Enhancing reproducibility for computational methods. Science. 2016;354:1240–1.

29. Hill SL. How do we know what we know? Discovering neuroscience data sets through minimal metadata. Nat. Rev. Neurosci. 2016;17:735–6.

30. PEP 8 -- Style Guide for Python Code [Internet]. Python.org. [cited 2017 Aug 21]. Available from: https://www.python.org/dev/peps/pep-0008/

31. PEP 257 -- Docstring Conventions [Internet]. Python.org. [cited 2017 Aug 21]. Available from: https://www.python.org/dev/peps/pep-0257/

32. Jasny BR, Wigginton N, McNutt M, Bubela T, Buck S, Cook-Deegan R, et al. Fostering reproducibility in industry-academia research. Science. 2017;357:759–61.

33. Software Carpentry [Internet]. Softw. Carpentry. [cited 2017 Aug 22]. Available from: http://software-carpentry.org//index.html

34. Data Carpentry [Internet]. Data Carpentry. [cited 2017 Aug 22]. Available from: http://www.datacarpentry.org/

35. MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges [Internet]. [cited 2017 Aug 23]. Available from: http://yann.lecun.com/exdb/mnist/

36. Bourne PE, Polka JK, Vale RD, Kiley R. Ten simple rules to consider regarding preprint submission. PLOS Comput. Biol. 2017;13:e1005473.

37. Preprints in biology. Nat. Methods. 2016;13:277–277.

38. Academia – Industry Software Quality & Testing summit - ISTQB® International Software Testing Qualifications Board [Internet]. [cited 2017 Aug 23]. Available from: http://www.istqb.org/special-initiatives/istqb-conference-network-2istqb-conference-network-academia/academia-%E2%80%93-industry-software-quality-testing-summit.html

39. Kanwal S, Khan FZ, Lonie A, Sinnott RO. Investigating reproducibility and tracking provenance – A genomic workflow case study. BMC Bioinformatics. 2017;18:337.

40. Karim MR, Michel A, Zappa A, Baranov P, Sahay R, Rebholz-Schuhmann D. Improving data workflow systems with cloud services and use of open data for bioinformatics research. Brief. Bioinform. [Internet]. [cited 2017 Jul 31]; Available from: https://academic.oup.com/bib/article/doi/10.1093/bib/bbx039/3737318/Improving-data-workflow-systems-with-cloud

41. Open Research Data in Horizon 2020 [Internet]. [cited 2017 Aug 23]. Available from: https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

42. Open Access in Horizon 2020 - EC funded projects [Internet]. [cited 2017 Aug 23]. Available from: https://www.openaire.eu/edocman?id=749&task=document.viewdoc

43. Foster ED, Deardorff A. Open Science Framework (OSF). J. Med. Libr. Assoc. JMLA. 2017;105:203–6.

# Figure legends

**Figure 1: Hidden reproducibility issues like underwater iceberg.** Scientific journals readers have the impression that they can almost see the full work of method. But in reality, articles do not take into account adjustment and configuration for significant

replication in most cases. Therefore, there is a significant gap between apparent executable work (i.e. above water portion of iceberg) and necessary effort in practice (i.e. full iceberg).

**Figure 2: Analogy between robustness issues and road transport.** The aim is to achieve same output (i.e. to reach the same location) using published methods (i.e. engine). Despite the same input data (i.e. gasoline), we obtained different results due to different programming languages —e.g. MATLAB and Python— (i.e. different roadways) and environments (i.e. different vehicles).

**Figure 3: Normalized confusion matrices between original and replicated results.** Before (a) and after (b) applying appropriate value of graph regularization factor on NBS method. Each row or column corresponds to a subgroup of patients (here three subgroups). The diagonal elements show the frequency of correct classifications for each subgroup: a high value indicates a correct prediction.

**Figure 4: Working principles of Workflow system with private data.** Figure 4a shows a classical workflow: (a.1) Authors ("A") take private data; (a.2) Authors publish their method and corresponding outputs/results; (a.3) Users ("U") having their own data find a relevant paper but will be lost in the labyrinth of reproducibility. Figure 4b shows workflow with standard consensus dataset: (b.1) If authors work with their own data, they must identify corresponding standard data tag(s) (e.g. A-2); (b.2) Users choose a method category (e.g. "Classification"); (b.3) Reproducibility profile with standard data is progressively built with method upgrade throughout publication of the first version. Color corresponds to method category depending score and bar length corresponds to progression of replication test; (b.4) Users can test proposed method with other data standards and thus participate to enhancement of the reproducibility profile; (b.5) Thanks to the collective work on testing, the method could be optimized and authors can upgrade their initial paper (versioning).

Figure 1

Click here to download Figure fig1_iceberg_lzw.tif

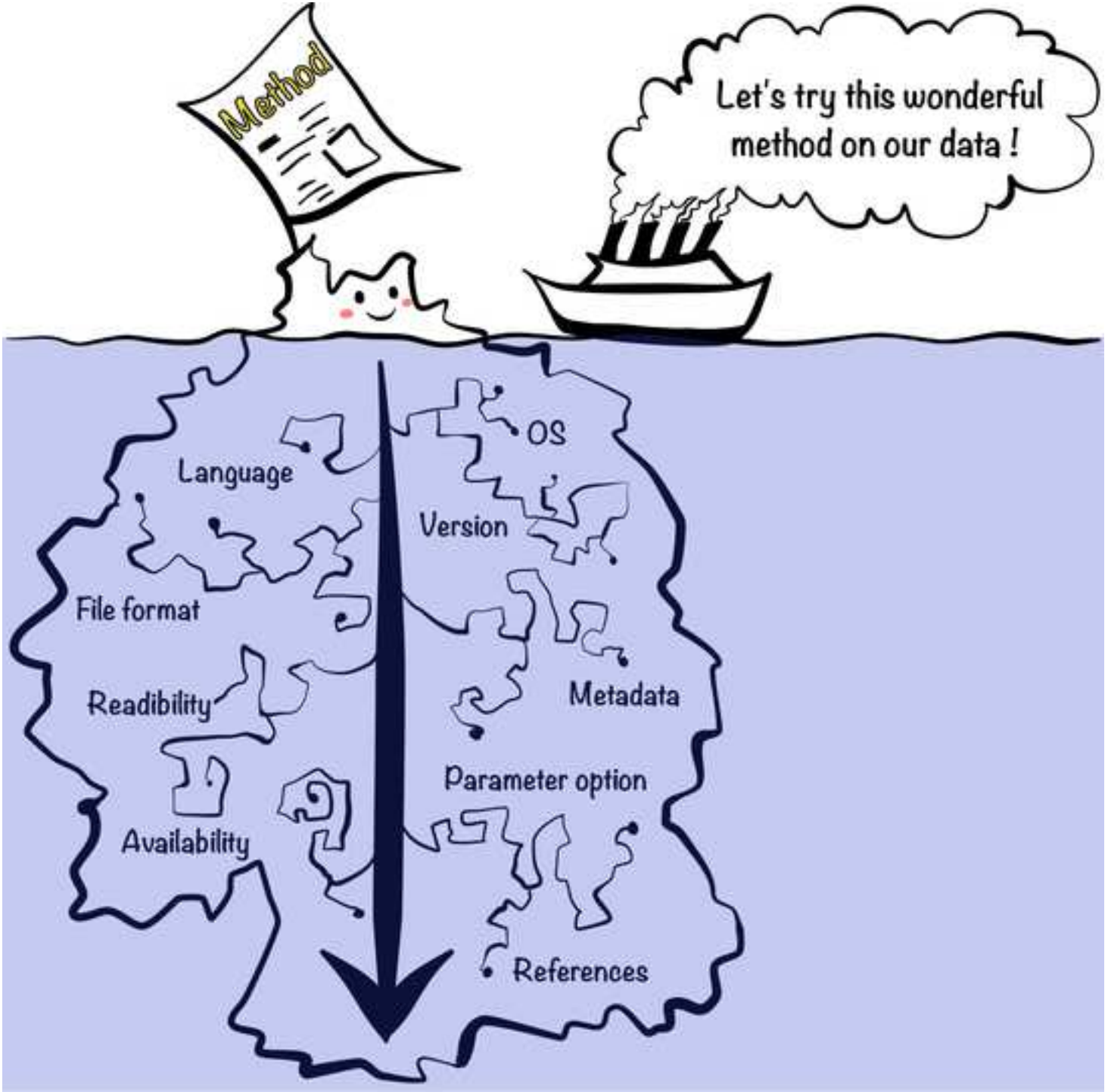


Figure 2

Click here to download Figure fig2_cars_lzw.tif

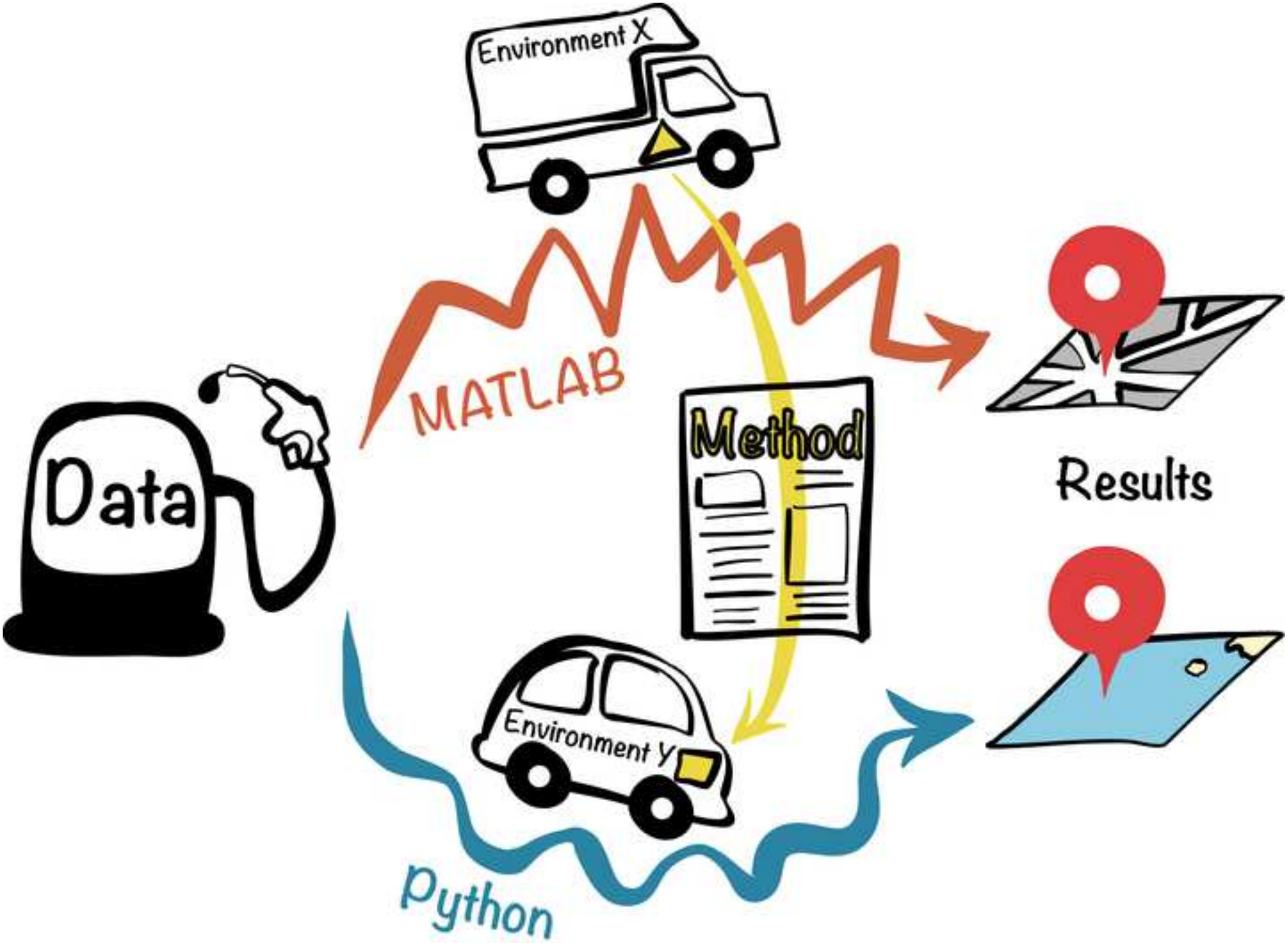


Figure 3

Click here to download Figure fig3_confusion_matrices_lzw.tif

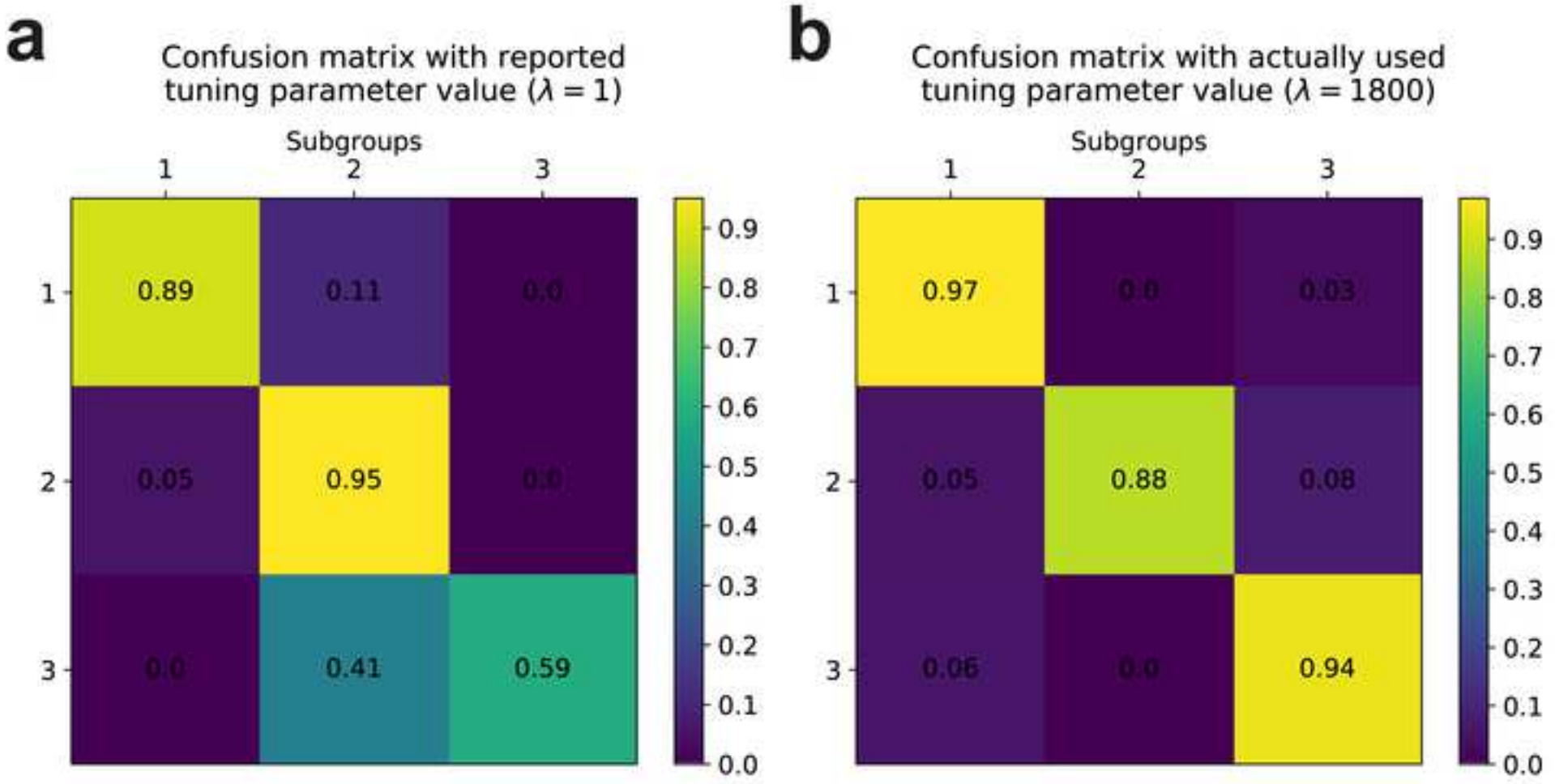
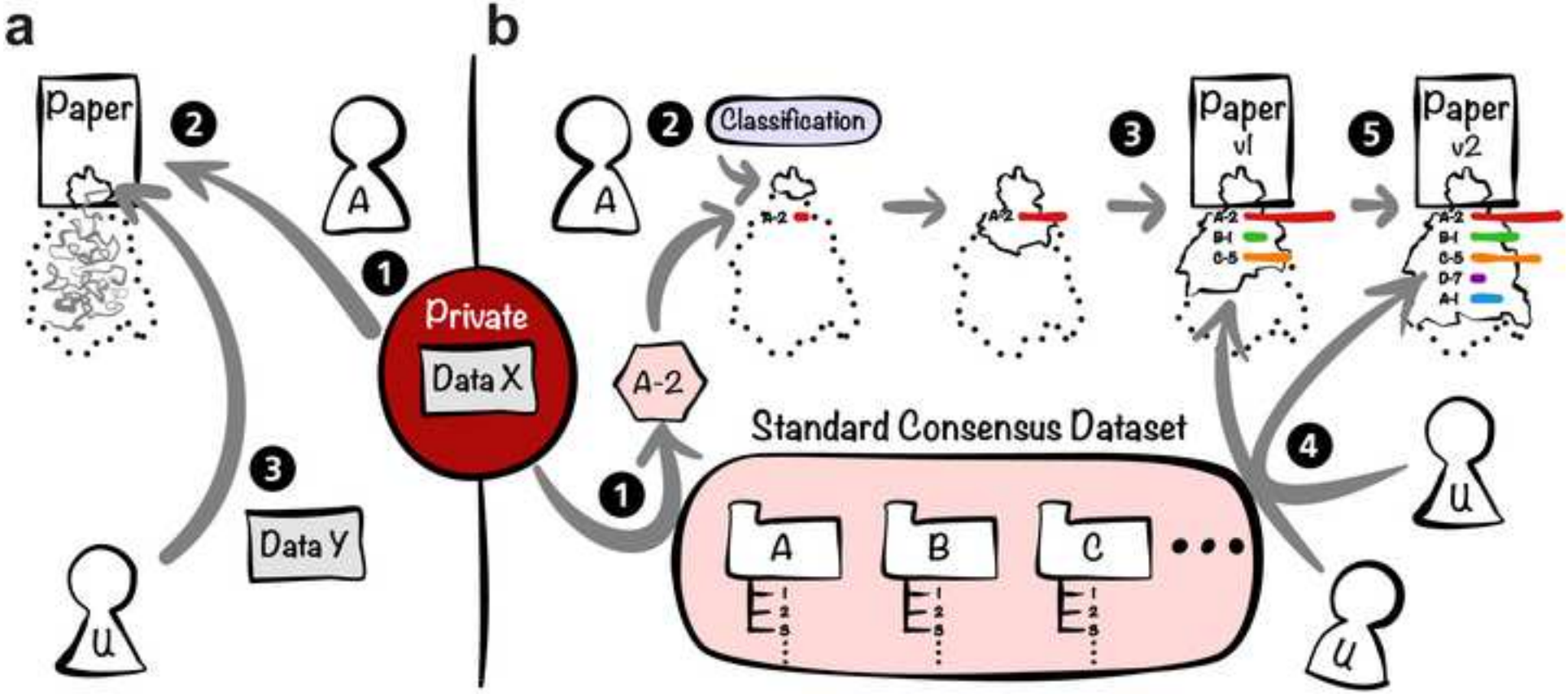


Figure 4

Click here to download Figure fig4_workflow_lzw.tif



*Paris, November 6, 2017*

*Human Genetics and*
*Cognitive Functions unit*

From : *Yang-Min KIM*

Phone: *+33 1 40 61 36 54*
Fax: *+33 1 40 61 34 21*
Email: *yang-min.kim@pasteur.fr*

Dear Editors,

Please find enclosed a copy of our manuscript entitled "***Experimenting with Reproducibility in Bioinformatics***" by Kim, Poline, and Dumas (bioRxiv 143503; doi: https://doi.org/10.1101/143503) that we would like to be considered for publication in *GigaScience*.

The main subject of our manuscript is, starting from a case study, to understand the possible causes of bioinformatic irreproducibility and to propose some practical solutions at both individual and collective levels.

One of our current research challenge is to classify patients with autism into more homogeneous subgroups using genetic stratification. Cancer research has recently proposed Network Based Stratification (NBS), to stratify tumours into meaningful subset (Hofree N. *et al* 2013). Reported original results should be reproducible by our team before applying NBS on our data. Despite original data and code were available, we faced several obstacles at several scales driven by different environments, programming languages (MATLAB and Python), or algorithm parameterization.

As the number of published increases every year, the capacity for the scientific community to check and reuse the results described in these article is not following this increase. The reproducibility crisis is now well described in many scientific fields and in particular in biology but the number of studies takling this issue remains very scarce in the literature. While reproducibility of results appears as the key for scientific progress, it is not yet generally considered as important as novelty.

In this experiment on reproducibility, we 1) attempt to reproduce initial results 2) attempt to estimate the robustness of our own results 3) extract lessons from these attempts, using a specific and concrete example. In addition, we propose to readers two options to reproduce the figure 3 using our NBS library (https://github.com/GHFC/Stratipy) with either a Jupyter Notebook or a Docker container. We argue that the bioinformatics community should collaborate with other communities and we propose a workflow system where standard datasets are used to validate tools continuously. We think that these observations and suggestions should be of interest to bioinformatics community and we hope to contribute to solutions to the so-called reproducibility crisis.

We certify that this manuscript has not been published elsewhere.
We look forward to hearing from you.

Yang-Min KIM, on the behalf of all the authors.



Institut Pasteur
25-28 rue du Docteur Roux
75724 Paris Cedex 15