

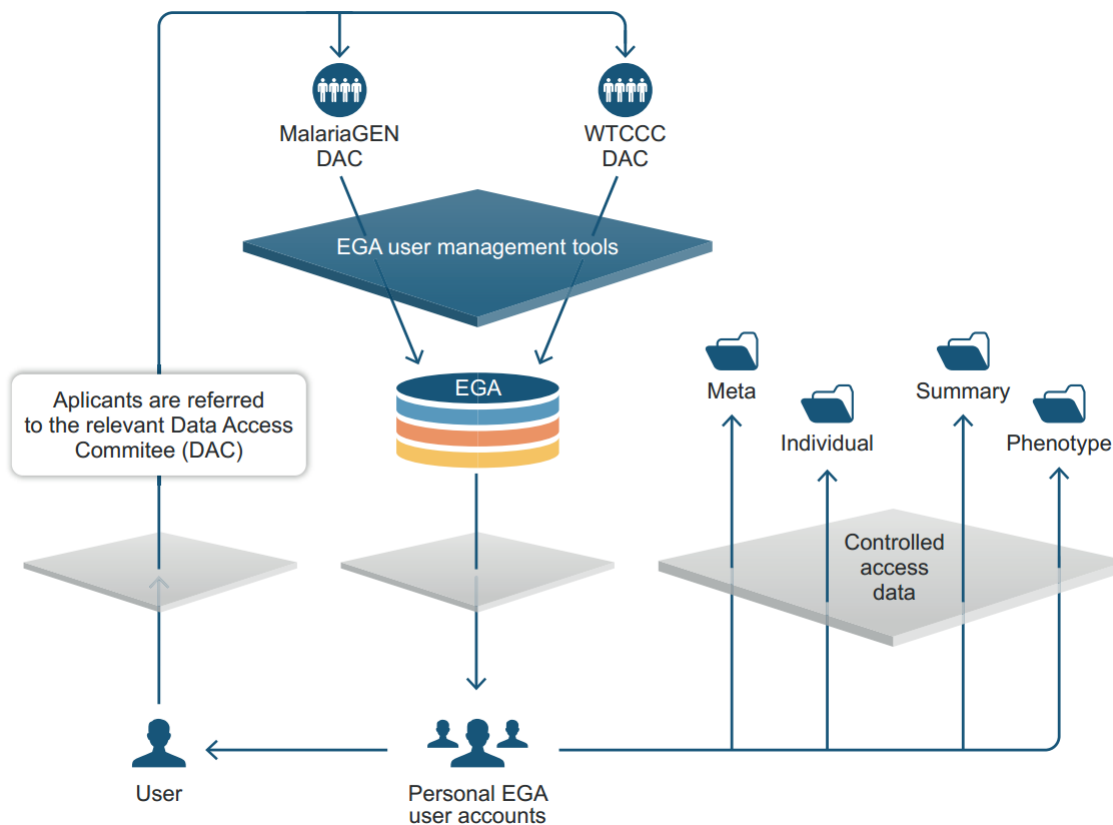
Reproducibility and discoverability at EGA

EOSCpilot workshop
September, 13th 2017



What is the EGA?

The EGA is a resource for permanent secure archiving and sharing of all types of potentially identifiable genetic and phenotypic data resulting from biomedical research projects.



Data is provided by research centers and health care institutions.

Access is controlled by Data Access Committees.

Data requesters are researchers from other research or health care institutions.

<https://ega-archive.org>

Project goal

The EGA was created by the EBI, in 2007, as an extension of the ENA...

Project goal:

To transform the EGA to a joint project (*in the context of ELIXIR Europe*) to have a real impact in the development of personalized medicine



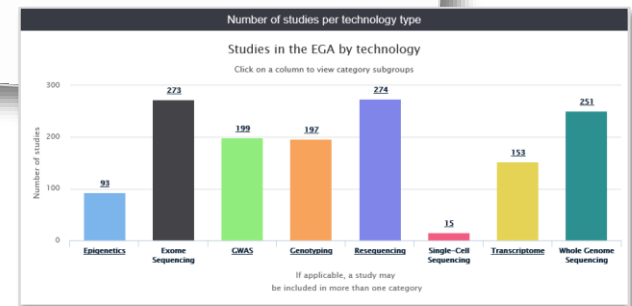
The EGA contains a variety of data

The EGA in numbers

- > 1,300 Studies
- 3,400 Datasets
- >800 Data providers
- >9,000 Data Requesters

The EGA in Volume

- >4 Petabytes

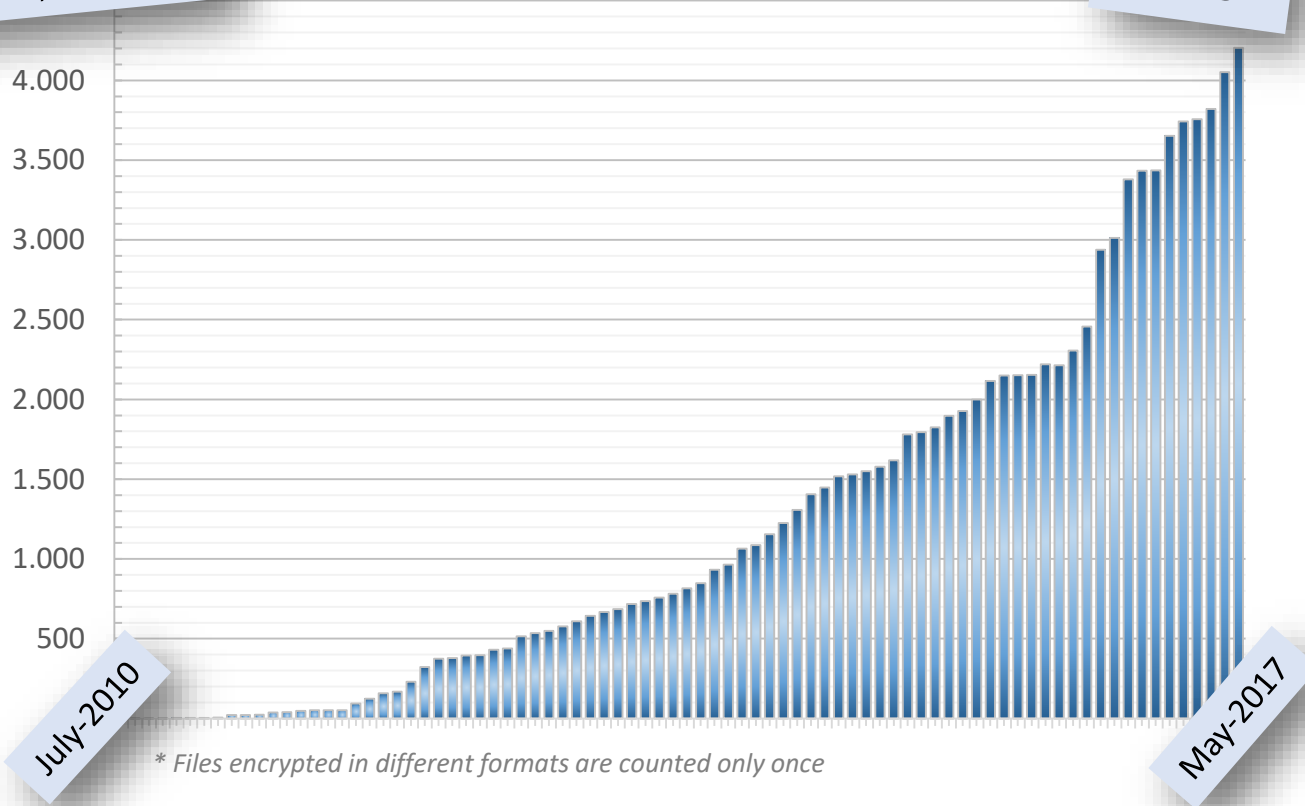


* Updated Sept, 8th 2017

The EGA contains a growing amount of data

~1,000,000 files*

>4 PB*



* Files encrypted in different formats are counted only once



The EGA is part of many international projects



The EGA is a key partner of ELIXIR

- Ongoing projects:
 - EXCELERATE WP9
 - 2 Human Data Implementation Studies
 - Beacon 2017
 - Rare diseases Visualization
- Finished:
 - EGA as a joint-venture
 - OncoTrack
 - TraIT
- **EGA as CORE Resource**

The screenshot shows the ELIXIR website's 'Human Data Use Case' page. The navigation bar includes 'ABOUT US', 'SERVICES', 'PLATFORMS', 'USE CASES', 'EVENTS', 'NEWS', and 'INTRANET'. The page title is 'Human Data Use Case'. A sidebar lists categories: 'Human Data', 'Rare Diseases', 'Marine Metagenomics', and 'Plant Sciences'. The main content area contains an introduction to the use case, a list of what it does, and a section on how it is organized. The 'How the Use Case is organised' section mentions Serena Scollen (ELIXIR Hub), Jordi Rambla (ELIXIR Spain), and Thomas Keane (EMBL-EBI). Below the text are three portrait photos of these individuals with their names and affiliations.

USE CASES

- Human Data
- Rare Diseases
- Marine Metagenomics
- Plant Sciences

Human Data Use Case

DNA and RNA sequencing have become increasingly important in medical and translational research. The data generated from these techniques has led to a huge demand for secure means to store, transfer and analyse the human biomedical data that has been consented for research.

The Use Case takes the [European Genome-phenome Archive \(EGA\)](#) as its primary data source, access to which is controlled. The EGA allows an authorised user to search sequenced material, patient samples stored in biobanks, and the metadata around patients (their illnesses, treatments, outcomes). It also queries national search engines on behalf of the users. Datasets can then be downloaded into an EGA compatible cloud or cluster local to the researcher.

The Human Data Use Case extends and generalises the system of access authorisation and secure data transfer developed in the EGA. It aims to provide a framework for the secure submission, archiving, dissemination and analysis of human biomedical data across Europe.

What the Use Case does

- Provides life scientists with a sustainable infrastructure for the storage, coordination and distribution of human data. It will provide standardised tools that researchers can use to discover and access this data.
- Brings together experts from ELIXIR's Nodes and external partners to develop a long-term management policy for human data.
- Ensures that human data in ELIXIR services is handled in accordance with the appropriate legal framework.

How the Use Case is organised

Serena Scollen is Head of Human Genomics and Translational Data at the [ELIXIR Hub](#). The work of the Human Data Use Case is defined in [Work Package 9](#) of the ELIXIR EXCELERATE project. The Work Package is led by Jordi Rambla and Thomas Keane.



Serena Scollen
(ELIXIR Hub)



Jordi Rambla
(ELIXIR Spain)



Thomas Keane
(EMBL-EBI)

Reproducibility crisis

The image shows a screenshot of a Nature journal article page. The article title is "Assessing the validity and reproducibility of results" by Lauren A. Sugden, Michael R. Tackett, and Yiannis A. Savva. The article is dated 24 April 2013. The page includes a navigation bar with links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Volume 496, Issue 7446, Editorial, and Article. The article's abstract discusses the reproducibility crisis and the need for rigorous methods. The page also features a sidebar with a search bar and a "Research" button.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Volume 496 | Issue 7446 | Editorial | Article

Journal List > Bioinformatics > PMC3810853

Announcement: I

24 April 2013

PDF | Rights & Permissions

Over the past year, *Nature* has and reproducibility of published research. The problems arise in laboratories that do not exert sufficient scrutiny over their research information for other researchers.

Assessing the validity and reproducibility of results

Lauren A. Sugden,¹ Michael R. Tackett,² Yiannis A. Savva

Author information | Article notes | Copyright and License information

Abstract

Motivation: Validation and reproducibility of results are essential for the scientific process. Recent embarrassing incidents involving the irreproducibility of results highlight the importance of this issue and the need for rigorous methods.

Results: Here, we describe an existing statistical method and its utility for assessing the reproducibility of validation of adenosine deaminase acting on RNA (ADAR)-mediated gene editing. We also describe a statistical method for planning validation experiments with confidence limits, which, for a fixed total number of experiments, allows for a more rigorous approach to the study.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive | Volume 515 | Issue 7525 | Editorial | Article

Journals unite for reproducibility

Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science.

05 November 2014

PDF | Rights & Permissions

Reproducibility, rigour, transparency and independent verification are cornerstones of the scientific method. Of course, just because a result is reproducible does not make it right, and just because it is not reproducible does not make it wrong. A transparent and rigorous approach, however, will almost always shine a light on issues of reproducibility. This light ensures that science moves forward, through independent verifications as well as the course corrections that come from refutations and the objective examination of the resulting data.



Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome

Daniel Garijo¹, Sarah Kinnings², Li Xie³, Lei Xie⁴, Yinliang Zhang⁵, Philip E. Bourne^{3*}, Yolanda Gil^{6*}

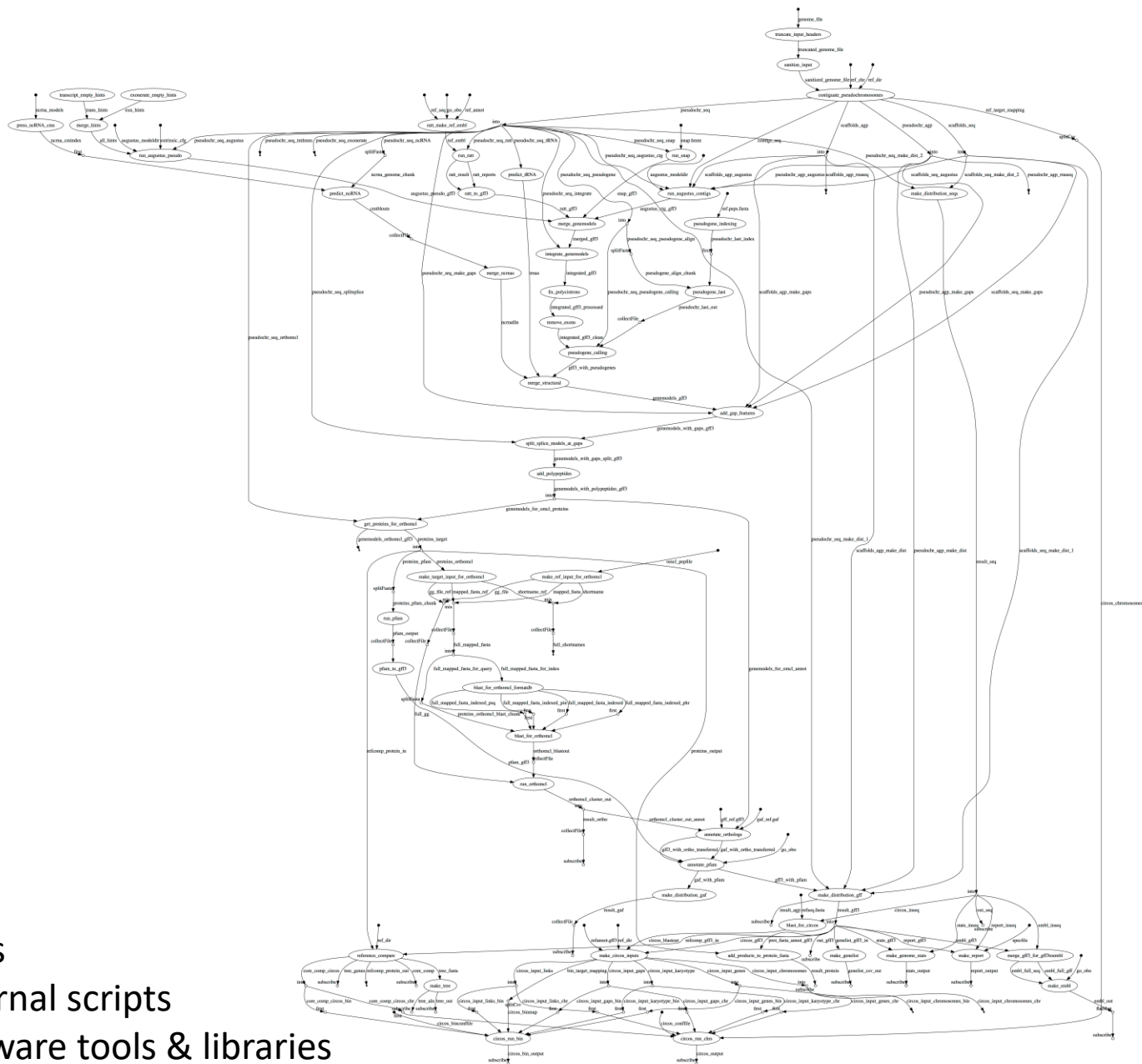
1 Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain, **2** Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, United States of America, **3** Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, United States of America, **4** Department of Computer Science, Hunter College, The City University of New York, New York, New York, United States of America, **5** School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China, **6** Information Sciences Institute and Department of Computer Science, University of Southern California, Los Angeles, California, United States of America

To replicate the result of a typical
computational biology paper
requires 280 hours.

≈1.7 months!

What's wrong with computational workflows?: Complexity

- Dozens of dependencies (binary tools, compilers, libraries, system tools, etc)
- Experimental nature of academic SW tends to be difficult to install, configure and deploy
- Heterogeneous executing platforms and system architecture (laptop→supercomputer)



70 tasks

55 external scripts

39 software tools & libraries

* Companion parasite genome annotation pipeline, Steinbiss et al., DOI: 10.1093/nar/gkw292

Comparison of the Companion pipeline annotation of *Leishmania infantum* genome executed across different platforms *

Platform	Amazon Linux	Debian Linux	Mac OSX
<i>Number of chromosomes</i>	36	36	36
<i>Overall length (bp)</i>	32,032,223	32,032,223	32,032,223
<i>Number of genes</i>	7,781	7,783	7,771
<i>Gene density</i>	236.64	236.64	236.32
<i>Number of coding genes</i>	7,580	7,580	7570
<i>Average coding length (bp)</i>	1,764	1,764	1,762
<i>Number of genes with multiple CDS</i>	113	113	111
<i>Number of genes with known function</i>	4,147	4,147	4,142
<i>Number of t-RNAs</i>	88	90	88

* Di Tommaso P, et al., *Nextflow enables computational reproducibility*, Nature Biotech, 2017 (publication pending)

nextflow

- A framework for computational workflows
- It provides a DSL to simplify the writing complex parallel workflows
- Enables transparent deployment on multiple platforms
- Built-in integration with containers technology

- Easy installation
- Use existing tools and scripts
- Implicit parallelization
- Simplified deployment
- Lightweight, self-contained

Easiness

nextflow

containers



HPC clusters
and cloud

Reproducibility



versioning

the EGA EOSCpilot project

The EGA EOSCpilot project: GOALS

1. Make easier to reproduce results archived at EGA
2. Avoid repeated reprocessing of the data with modern tools
3. Make artifacts involved easier to discover (FAIR)

Results reproducibility

- EGA stores both raw and secondary analysis data
- We will like to make very simple to get the published/archived from the raw data
 - Given the reproducibility crisis, ensuring exactitude is very desirable
 - Link data to the pipelines and tools used to analyze them
- Pipeline and tool repositories using stable identifiers are required

Remastered results

- Once raw data is downloaded many users will up to date them by processing against current references and using popular pipelines
 - This means tons of wasted resources to get the same results: *human, computational and time resources*
- We would like to generate reproducible pipelines, run them and get the results back to the EGA
 - Thus users could choose to get the originals, the remastered or both
- We need to actually check the popularity of such “service”
 - Maybe we just need to leverage work done by previous users

Make data more discoverable

- EGA is already honoring some FAIR principles
 - Findable, Accessible (\pm), Interoperable (\pm), Re-usable
- As we expand the number of artifacts related to the data archived at EGA, we are increasing the need to describe and link such objects
- We would like to leverage the process of generating the previously described artifacts to gather metadata that would be exposed through the right tools and services.



Some other attributes to mention

- Most of the data involved is under controlled access (not open), thus, security restrictions apply
 - A description of the required environment is a potential byproduct of the pilot
- Using Singularity instead of Docker to avoid using root privileges at an HPC facility



Success criteria

- Obvious:
 - Actually reproduce results
 - Get the processing artifacts permanently archived and a proposal for linking them to data
 - Get an updated version of the results
 - Have a pilot FAIR solution working
- Most important:
 - Learn about the pros and cons of the ideas

credits



Evan Floden, CRG



Emilio Palumbo, CRG



Maria Chatzou, CRG



Pablo Prieto, CRG



Cedric Notredame, CRG

THANKS!

Core organizations:



Additional sources:



And infrastructure support from the following sources:

