Reproducibility and discoverability at EGA

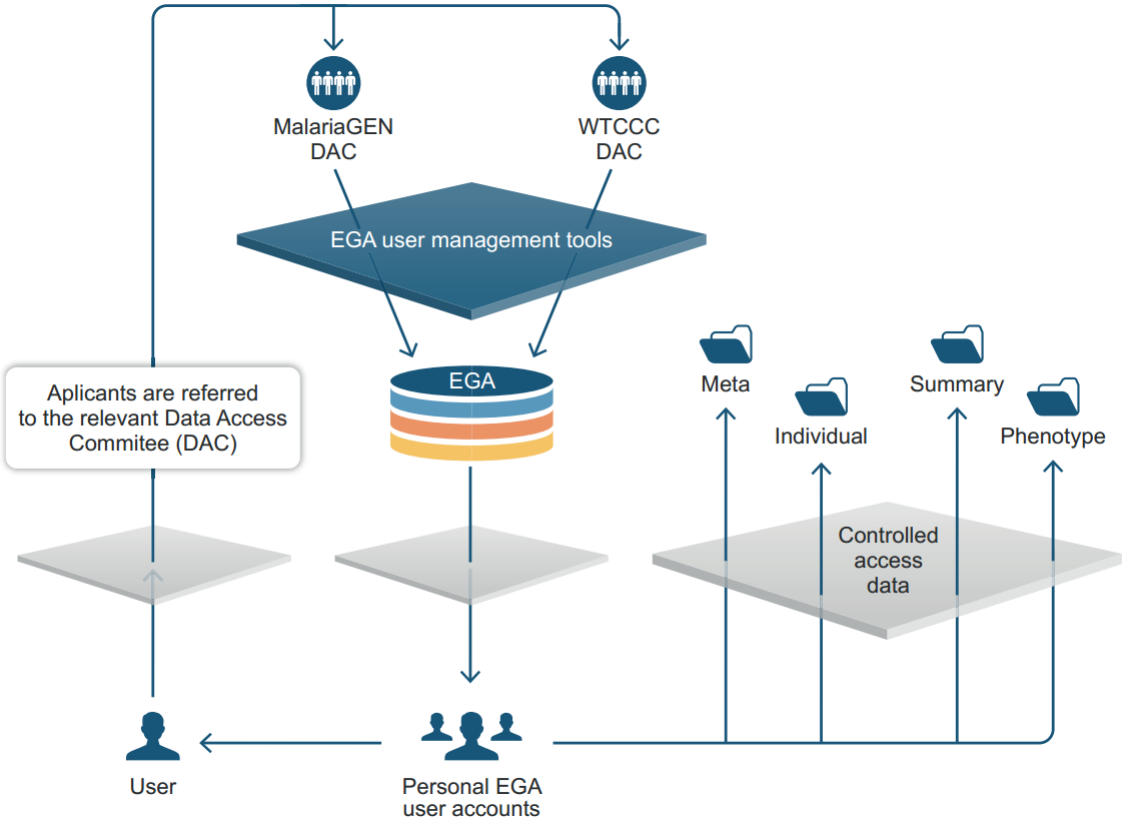
EOSCpilot workshop

September, 13th 2017

Obra Social "la Caixa"
Unión Europea
Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"
EXCELENCIA SEVERO OCHOA
Generalitat de Catalunya
CRG Centre for Genomic Regulation
EMBL-EBI
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

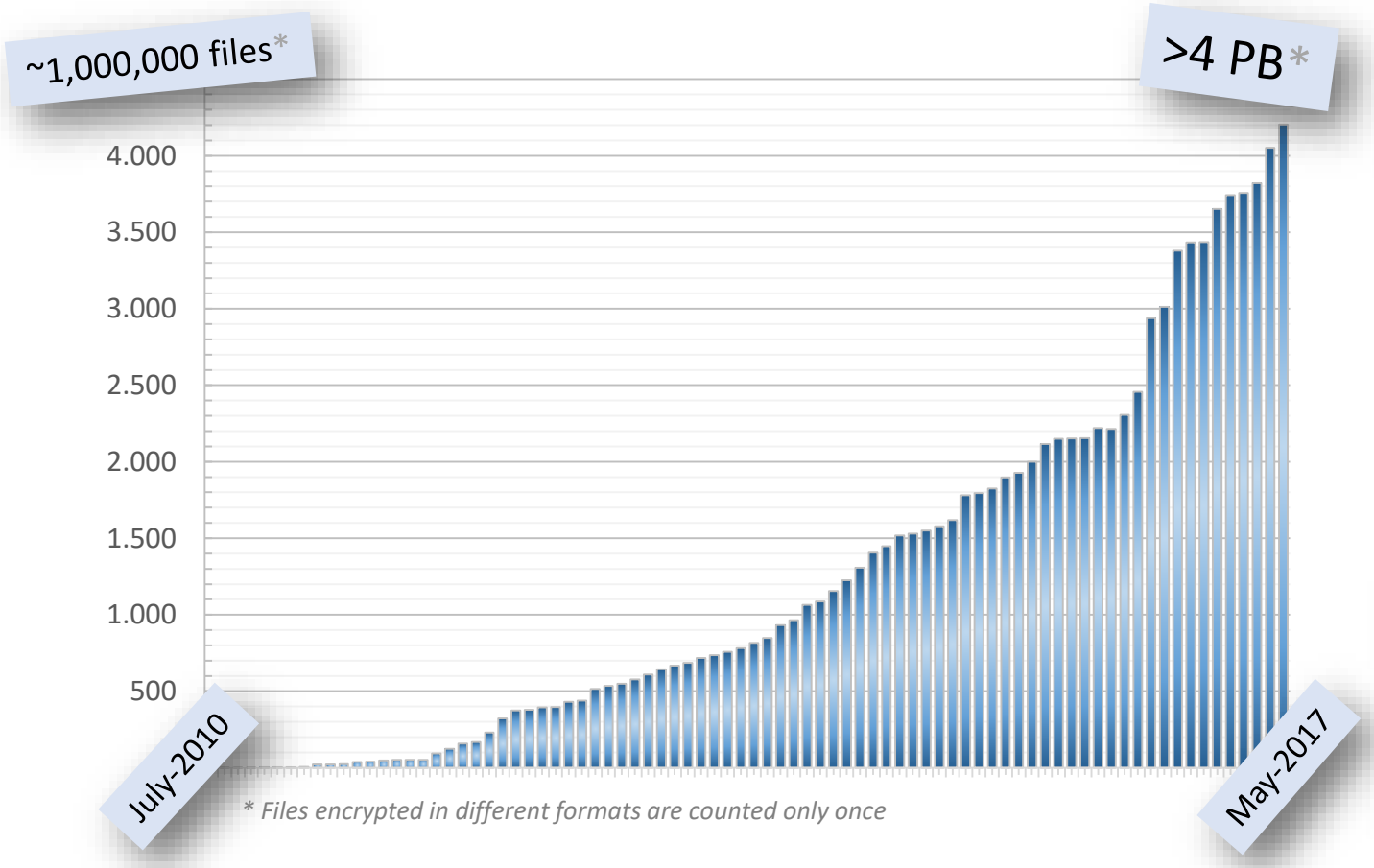
- > 1,300 Studies
- 3,400 Datasets
- >800 Data providers
- >9,000 Data Requesters

The EGA in Volume

- >4 Petabytes

** Updated Sept, 8th 2017*

The EGA contains a growing amount of data



International Cancer Genome Consortium

IRDiRC INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

UK 10K

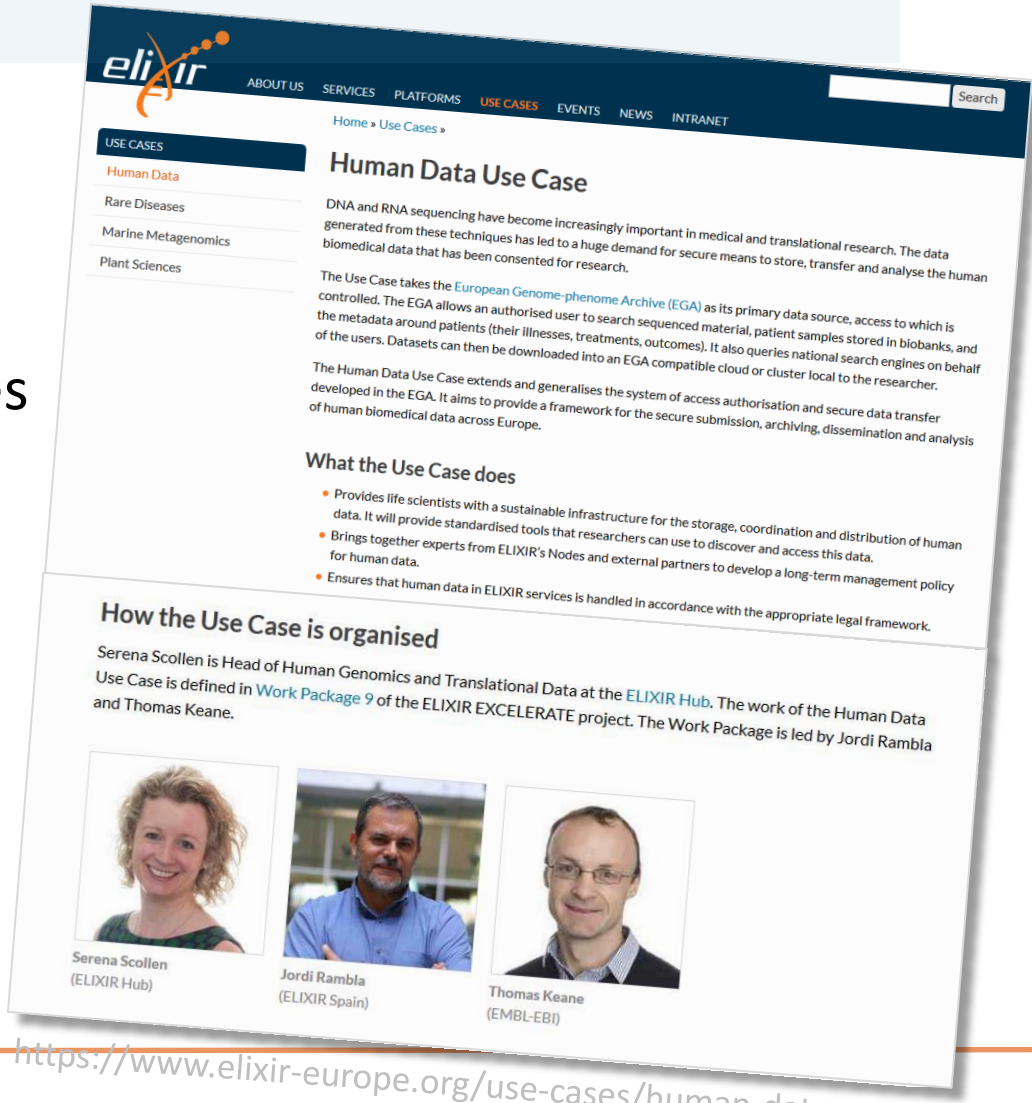
wellcome trust sanger institute

IHEC

The EGA is part of many international projects

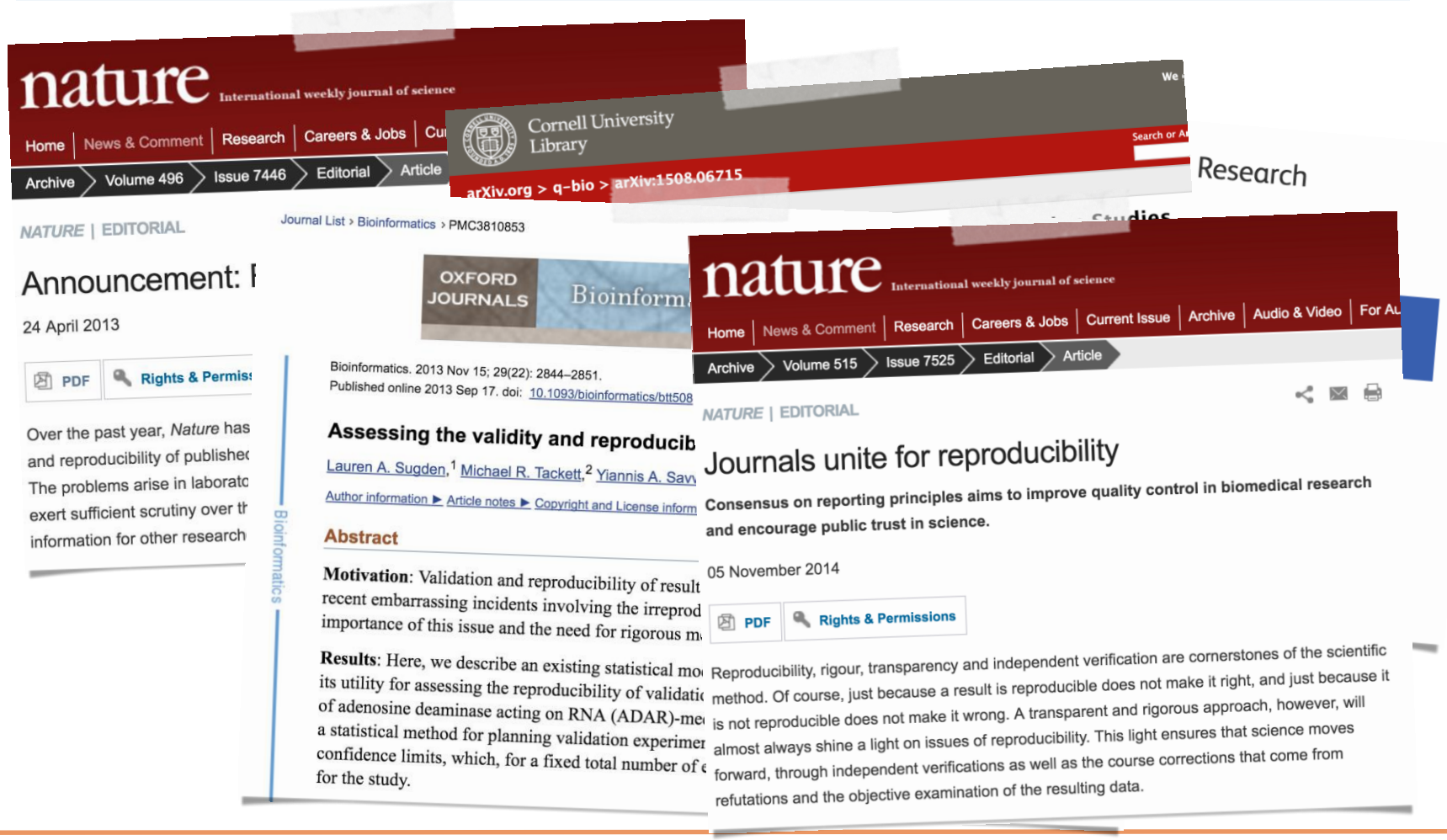
The EGA is a key partner of ELIXIR

- 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



https://www.elixir-europe.org/use-cases/human-data

Reproducibility crisis



OPEN ACCESS Freely available online

PLOS ONE

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

Daniel Garijo[1], Sarah Kinnings[2], Li Xie[3], Lei Xie[4], Yinliang Zhang[5], Philip E. Bourne[3]*, Yolanda Gil[6]*

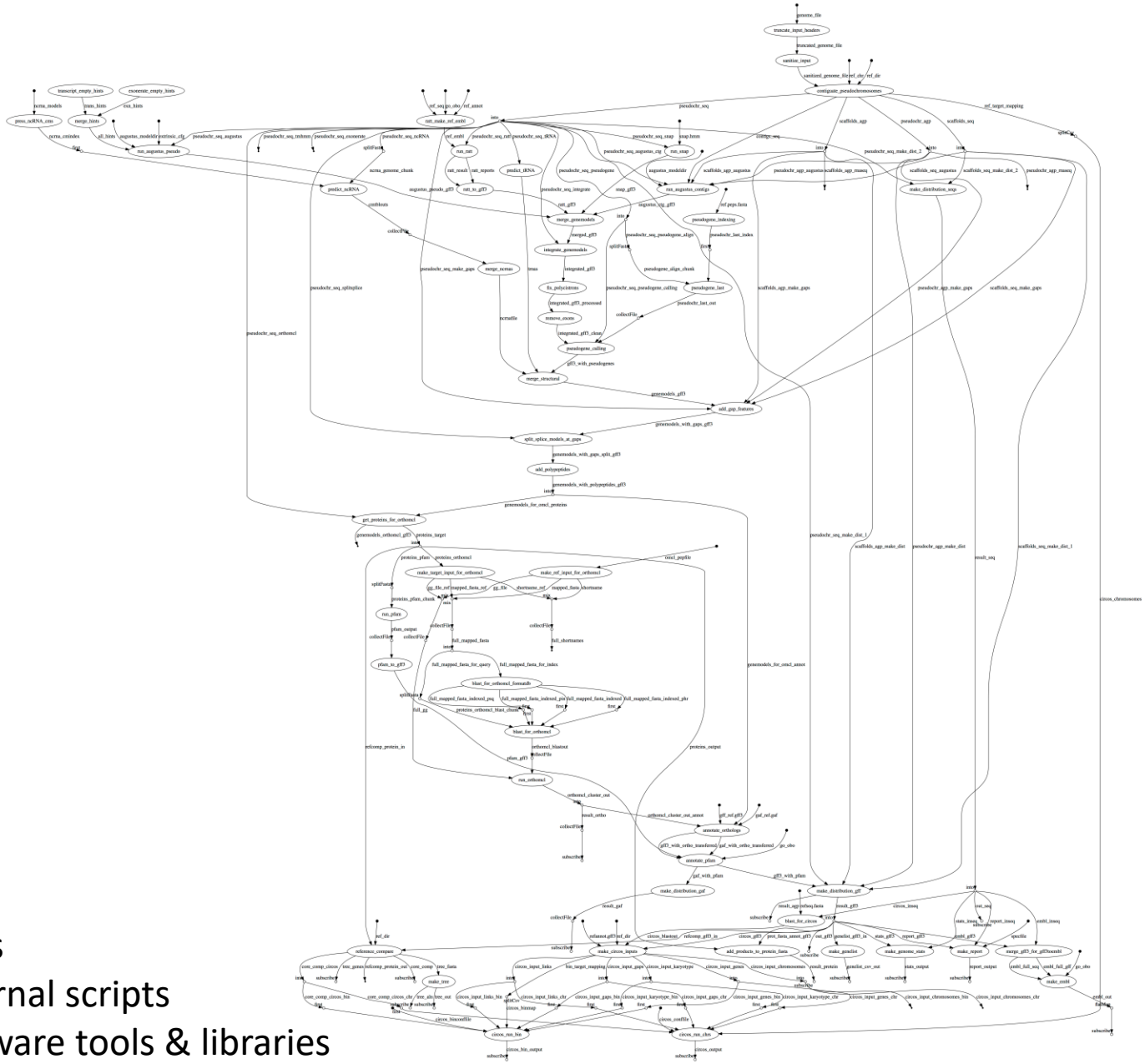
1 Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain, **2** Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, United States of America, **3** Skaggs School of Pharmacy and Pharmaceutical Sciences, University 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, LosAngeles, 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
Number of chromosomes	36	36	36
Overall length (bp)	32,032,223	32,032,223	32,032,223
Number of genes	7,781	7,783	7,771
Gene density	236.64	236.64	236.32
Number of coding genes	7,580	7,580	7570
Average coding length (bp)	1,764	1,764	1,762
Number of genes with multiple CDS	113	113	111
Number of genes with known function	4,147	4,147	4,142
Number of t-RNAs	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 an 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 (±), Interoperable (±), 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.

DTL
DUTCH TECHCENTRE FOR LIFE SCIENCES

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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Success criteria

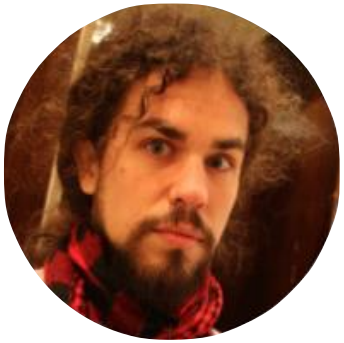
- Obvious:
  - Actually reproduce results
  - Get the processing artifacts permanently archived and a proposal for linking them to data
  - Get an updated version of the results
  - Have a pilot FAIR solution working
- Most important:
  - Learn about the pros and cons of the ideas

credits

Evan Floden, CRG

Emilio Palumbo, CRG

Maria Chatzou, CRG



Pablo Prieto, CRG

Cedric Notredame, CRG

THANKS!

Core organizations:

CRG Centre for Genomic Regulation

EMBL

Additional sources:

Generalitat de Catalunya

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Obra Social "la Caixa"

elixir

HORIZON 2020

SEVENTH FRAMEWORK PROGRAMME

EXCELENCIA SEVERO OCHOA

FNIH Foundation for the National Institutes of Health

Accelerating Medicines Partnership

RD-Connect

And infrastructure support from the following sources:

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

upf. Universitat Pompeu Fabra Barcelona

inb Spanish National Bioinformatics Institute

https://ega-archive.org/support