# WHAT ABOUT DATA?

@Chris\_Evelo
Department of Bioinformatics – BiGCaT
Maastricht University

### Open data versus closed data







## Typically we want to:



"And that's why we need a computer."

## FAIR data

Findable

Accessible

Interoperable

### Reusable

## Find it through a FAIRport



#### What is Data FAIRport

Data FAIRport is an open initiative of representatives from the worlds of research infrastructure and policy, publishing, the semantic web and life sciences research. Data FAIRport was founded in January of 2014. It's vision is to realise and enable a situation where valuable scientific data is 'FAIR' in the sense of being Findable, Accessible, Interoperable and Re-usable. Data FAIRport engages all 'Enablers' in the field of data publishing and re-use. It will not develop any solutions that are already in place and does not <u>read</u> more.

#### Join Data FAIRport!

We are an open initiative and are very happy for anyone to join us in achieving 'fair' data. Please contact us <u>here</u>.

#### What's happening

To see what's happening in Data FAIRport have a look here.

### But first we need to capture it

NATURE GENETICS | COMMENTARY OPEN

-< 🖂 🖨

### Toward interoperable bioscience data

Affiliations | Corresponding author

*Nature Genetics* **44**, 121–126 (2012) | doi:10.1038/ng.1054 Published online 27 January 2012

## A common standard

<b></b> i	<b>sa</b> commons		Fr	om the same iosharing	e people who brought you: mibbi	Social media
⇒ <sup>Ir</sup> ht	ttp://www.sciencemagazinedigital.org/	AghaKouchak et al. sciencemagazine/20140328_2?pg=42#	use data from #pg42 http://www.nature.com/a	recent articles/sda	pub in Sci ata20141 #opendata	Data
The fac	e ISA Commons is a gro ilitate standards-comp inc	owing community that u liant <b>collection, curatic</b> reasingly diverse set of	ses the <b>ISA metada</b> <b>on, management</b> ar f life science domain	n <b>ta trac</b> nd <b>reus</b> ns.	<b>cking framewo</b> se of datasets i	<b>rk</b> to n an
i <b>c</b> da		Orford effectarch centre Pacific Northwest retrock Laboretory Pacific Northwest Pacific N	average of a second sec	HARV	ARD f Public Health	VARD MEDICA OOL
	Benetics		Encidements		L'	•
<	How to cite us! Towards interoperable bioscience data.	The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again	The ISA software suite	Standa	rdizing data, introducing ISA- Tab-Nano	>
	READ THE PAPERI	READ THE PAPER!	Bioinformatics, 2010	Na	READ THE PAPER!	

### Generic Study Capture Framework Data input / output



## The study description points to:

The raw data	<ul> <li>And is the raw data provenance</li> </ul>
The data processing protocols	<ul> <li>Which together is the processed data provenance</li> </ul>
The processed data	<ul> <li>Typically two levels: clean and interpreted</li> </ul>
The integrated, evaluated data	<ul> <li>For which we need knowledge based models</li> </ul>



### **dbNP** Architecture



Faculty of Health, Medicine and Life Sciences

### **Epigenetics DNA Methylation Pipeline**





### **Process the data...**



#### Faculty of Health, Medicine and Life Sciences

### Fairports: connecting to other data

We both need Study Capturing

#### 🖉 Commons Repository - Mozilla Firefo: - 🗆 × Eile Edit View History Bookmarks Tools Help Commons Repository +(a) sagebase.org/commons/repository.php ☆ マ C Sagebionetworks 🔎 🏦 🔮 👀 http://www.bigcat.un... 🙍 Most Visited 🥘 Getting Started 🔊 Latest Headlines 🔅 Import to Mendeley Sage Info Commons Home > Commons > Repository **Data Repository** Data Release 4.0 - April 2011 Go to Repository Sage Bionetworks has established a catalogue of datasets for use in integrative genomics analysis and building predictive computational disease models. The goal is to collate, curate and host these datasets for use by the entire research community. Later this year the Repository will be hosted within a new interactive software platform designed to facilitate the active sharing and evolution of datasets, disease models and computational tools. Datasets are listed in the repository under one of three categories; (A) data currently available for download, (B) pending datasets that will soon be made available, and (C) other notable datasets that have been identified and/or are currently being generated. Data available for download is publicly available. Access the datasets by registering on the Repository Access Page and then confirming agreement with the user license. Licenses vary by data package and include requirements to properly cite and acknowledge the data and model sources in all publications resulting from use of these datasets. Datasets in the Sage Bionetworks Repository are selected based on their utility for modeling techniques. Global coherent datasets (GCDs) are the most powerful and contain three layers of information: genome-wide DNA variation, genome-wide intermediate traits and phenotypes. Intermediate traits are typically gene expression profiles, but may also include proteomic, metabolomic, RNA-seq and other molecular data. Although current Sage Bionetworks efforts are focused on providing access to global coherent datasets, we will also offer other useful genomic datasets containing phenotypic traits in conjunction with either DNA variation or intermediate trait data as they become available. Data is in a datapackage under the parameters described here. Go to Repository For more information you can also: Read about Sage Bionetworks' Strategy for community-based integrative genomics modeling. Download the terms of use for Sage Bionetworks Repository data, tools and models as well as the quidelines for the data curation process.

#### **Contribute to Community-Based Data Acquisition**

Tell us about a dataset you have generated or one you are interested in accessing through this repository:

I would like to see this dataset in the Sage Bionetworks Repository.

(nlassa nrovida sufficient detail for us to locate

· .

### BiGCaT approach to Bioinformatics for Integrative Systems Biology





http://www.wikipathways.org/index.php/Pathway:WP430



**DUTCH TECHCENTRE FOR LIFE SCIENCES** 



#### NEWS

- Pan-European Imaging infrastructure gains momentum
- DTL Focus meeting: NGS Data Storage and Sharing
- CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy
- Proposal on Open Discovery and Exchange of (big) data in Life sciences funded for 2 years
- News overview

#### **EVENTS**

- 08.04.2014 | Netherlands Bioinformatics Conference 2014
- 09.04.2014 | Join ELIXIR-NL session at NBIC 2014
- 14.04.2014 | DTL Focus meeting: Proteomics bioinformatics
- 15.04.2014 | DTL Focus Meeting: 'NGS Production Pipelines'
- Events overview



#### DUTCH TECHCENTRE FOR LIFE SCIENCES (DTL)

DTL Headlines:

- ZonMw and NWO-ALW award 35 projects access to DTL technology hotels
- Netherlands joins international ELIXIR Consortium on biological data

The Dutch Techcentre for Lifesciences (DTL) focuses on the high-end technologies that drive next generation life science research in the clinical & health, nutrition, agro-genomics and industrial microbiology sectors. DTL **aims** to establish a world-class research infrastructure enabling the Netherlands to remain at the forefront of biology research and innovation.

DTL is deeply embedded in the Dutch life sciences community and has been **founded** with strong involvement from universities, university medical centres, research institutes, science policy makers & funders and industry. Their joint **mission** is to create a platform in DTL where they can bundle their efforts to establish an integrated research infrastructure that is cross-technology, cross-sector, cost effective and that stimulates international collaboration.

DTL focuses its efforts to develop, combine and apply technologies in three closely interlinked **DTL**|programmes:

- DTL|Data
- DTL|Technologies
- DTL|Learning

Over 100 **expert groups** collaborate within these programmes to share and provide access to their advanced research facilities and expertise.

As part of its data activities, DTL hosts the Dutch node of **ELIXIR**, the European initiative that builds a sustainable international infrastructure for biological information.

Announcing the 2013 DREAM8.5 Challenges Announcing the final results of the 2013 DREAM8 Challenges More information about previous DREAM Challenges About DREAM Challenges About Synapse



### Announcing the 2013 DREAM8.5 Challenges

We are pleased to announce three new DREAM8.5 challenges. Best performers in all DREAM8.5 Challenges will be invited to present at the 2014 DREAM conference (date and location to be determined). We are also working to establish publishing partners for each of these challenges. The DREAM8.5 Challenges are now open for registration, and will begin active problem-solving in late 2013 or early 2014.

Click on a link below to read the Challenge detail and register for a DREAM8.5 Challenge.

#### Alzheimer's Disease Big Data DREAM Challenge #1

In the first of what will be a series of Alzheimer's Disease (AD) Big Data Challenges, participants will utilize data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Data will consist of cognitive, imaging, biological, and whole genome sequencing data on cohorts of volunteers, who range from cognitively normal, mild cognitive impairment and dementia. Participants will analyze the data to solve two sub-challenges: (1) Build a model that best predicts change over time in AD cognitive scores using all available test and adjacent data, and (2) Build a model that best predicts discordance between biomarkers suggestive of amyloid perturbations and lack of cognitive impairment. These models will be used to better understand the biomolecular mechanisms leading to Alzheimer's disease, and ultimately to develop new therapies. We expect to announce a publishing partner for this Challenge shortly.

### ICGC-TCGA-DREAM Somatic Mutation Calling Challenge

Working with technology partners Google and Annai, we will provide 9 terabytes of raw human sequence data derived from pairs of normal and tumor tissue (from prostate and pancreas). Approved participants will analyze the data to solve two sub-challenges: (1) build a model that accurately predicts cancer mutations that alter a single nucleotide in the genome (single nucleotide variants, SNVs) (2) Build a model that accurately predicts cancer mutations that alter the order of a large stretch of the genome (i.e. a structural variation, SV), such as a rearrangement, inversion or copy-number aberration. Improving the algorithms that correctly identify these variations is important because these variations provide key genetic data which can be used by predictive models to guide personalized cancer therapies. *Nature Publishing Group* enthusiastically welcomes the opportunity to consider for publication work that achieves best performance in this Challenge.

### The Rheumatoid Arthritis Responder Challenge

Participants will have access to whole genome genotype data (2.2 million SNPs) and clinical data collected from 2,000 individuals with Rheumatoid Arthritis who have been treated with anti-TNF therapy. Up to one third of these patients fail to enter clinical remission. Participants will use these data to solve two sub-challenges: (1) Build a model that best predicts treatment response as measured by the change in disease activity score (DAS28) in response to anti-TNF therapy, and (2) Build a model that best predicts poor responders as defined by specific criteria (yet to be specified). The winning model will be the one that can predict the largest pottion of this sample subset (10% of the population) with a positive predictive value greater than a predetormined.