benefits of open science

lennart martens

lennart.martens@vib-ugent.be computational omics and systems biology group VIB / Ghent University, Ghent, Belgium







Open science creates very many opportunities but it does take some work to get it right

- Open science should not serve to police scientists, but to provide opportunities
- Creating an open data ecosystem is not trivial, but with a bit of work, it can certainly be done
- Scientists should not ask themselves whether there will be open science, but rather what they will be able to do thanks to open science



Why should we share our work?

Our biggest challenge is to make it work!

What can we do with open science?

Of sedimentation, opportunity, and (an absence of) dragons



Why should we share our work?

Our biggest challenge is to make it work!

What can we do with open science?

Of sedimentation, opportunity, and (an absence of) dragons



We usually think we need open science to prevent bad things from happening

- While open science helps prevent some cases of fraud or low quality work being published, it is certainly not a panacea (cfr. peer review)
- Simultaneously, fraud is regularly detected:
 - in the absence of the source data
 - from papers published in closed access journals
 - without any of the code or metadata available
- Why should we define the use of open science through an application with negative connotation?



Instead, we should rather focus on the good that comes from open science

- Open science makes the work accessible to anyone
- Open science allows people to build much more efficiently on previous work
- Open science helps maximize the usefulness of each individual research effort
- Data tend to have a (much!) longer shelf life than our (limited) interpretations
- Open science fosters creativity, and stimulates revolutionary research



Why should we really have open science?

Our biggest challenge is to make it work!

What can we do with open science?

Of sedimentation, opportunity, and an absence of dragons



Making open science work requires a bit of effort from every scientist

- The data that is obtained should be accompanied by the associated metadata
- The code that is written should be understandable, documented, and hosted at a reliable site
- The protocols should be provided clearly and in full
- The interpretations should be clearly linked to the data (*full provenance*)
- Everything should be licensed in a permissible way



Any open data exchange ecosystem requires standardization

1. Data and metadata generation





Masuzzo, Trends in Cell Biology, 2014

But scientists are human, and very fallible - especially when extra effort is required





But scientists are human, and very fallible - especially when extra effort is required





The arrival of user-friendly submission tools did not nothing to reverse reporting trends





Manual curation of submissions, equivalent to restrictive policing, did help





Mid-to long-term storage of data is expensive and thus very difficult to fund and maintain

S NCBI Resources 🗹 How To 🖂	Sign in to NCBI
We are sorry, but the page you requested is no longer available.	
Peptidome	
Peptidome was a public repository that archived tandem mass spectrometry peptide and protein identification data generated by the scientific community. This is in archival mode. All data may be obtained from the Peptidome FTP site.	s repository is now offline and
Due to budgetary constraints NCBI has discontinued the Peptidome Repository. All existing data and metadata files will continue to be made available from ou ftp://ftp.ncbi.nih.gov/pub/peptidome/ indefinitely. Those files are named according to their Peptidome accession number, allowing cited data to be identified and Peptidome studies have been made publicly available at the PRoteomics IDEntifications (PRIDE) database. A map of Peptidome to Pride accessions may be /pub/peptidome/peptidome-pride map.txt.	ur ftp server a d downloaded. All of the found at <u>ftp://ftp.ncbi.nih.gov</u>
If you have any specific questions, please feel free to contact us at info@ncbi.nlm.nih.gov.	
Archived Peptidome Studies	
PSE101 Comparative analysis of five different preparation methods for identification of complex protein mixtures in yeast	
PSE102 Proteomic analysis of the intestinal epithelial cell response to enteropathogenic Escherichia coli	
PSE103 Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry	
PSE104 Head-to-head comparison of serum fractionation techniques	
PSE105 Quality control metrics for LC-MS feature detection tools demonstrated in yeast	
PSE106 Improving reproducibility and sensitivity in identifying human proteins by shotaun proteomics	

http://www.ncbi.nlm.nih.gov/peptidome

Slotta, Nature Biotechnology, 2009; Csordas, Proteomics, 2013; Martens, Proteomics, 2013



And as responsible caretaker of your stuff, you will sometimes need to take action too

Google code	
Project has moved	
What happened?	
Project "peptide-shaker" has moved to another location on the Internet.	
 Your options: View the project at: <u>http://compomics.github.io/projects/peptide-shaker.html</u> <u>Search the web</u> for pages about "peptide-shaker". 	
If you are the project's administrator, you can update the new project URL <u>here</u> .	
	<u>Terms</u> - <u>Privacy</u> - <u>Project Hosting Help</u> Powered by <u>Google Project Hosting</u>

http://peptide-shaker.googlecode.com Vaudel, *Nature Biotechnology*, 2015



Maintaining what you do is not trivial, not cheap, and considered unwise by senior PIs



http://peptide-shaker.googlecode.com

Vaudel, Nature Biotechnology, 2015



Why should we really have open science?

Our biggest challenge is to make it work!

What can we do with open science?

Of sedimentation, opportunity, and an absence of dragons



The PRIDE database was started to allow (orthogonal) re-analysis of proteomics data

Application for Training at the EMBL-EBI EU Marie Curie Training Site 1. Name of Applicant (please print) Please indicate with whom you would like to undertake training specifying reasons for 4. LENNART MARTENS your choice(s) and the benefits you would expect from the training provided. Restrict the text to this page only. 2. Name of the supervisor of your Ph.D. work University (or Institute), and Country where you Building on the successes achieved by genome sequencing efforts and the resulting whole-genome databases, the research focus in the life sciences these days has shifted Prof. Dr. Joël Vandekerckhove towards the proteomes of cells and tissues. As traditional protein identification Ghent University (UGent) techniques such as two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) are currently joined by a number of highly sensitive peptide-based (or gel-free) Department of Medical Protein Chemistry techniques, mass-spectrometry based data on protein identification is quickly A. Baertsoenkaai 3 becoming available. Yet this data is currently lacking a standardized format as well as B-9000 Ghent centralized storage. Whereas, for instance, authors of gene-sequencing related articles Belgium are typically required to submit their findings to public databases prior to publication for the inspection and benefit of the scientific community at large, no such system exists in the field of proteomics today. As a result, many researches simply publish their list of identifications in a proprietary format as supplementary information to their publication. On top of this, these formats are often not readily machine-readable. 3. Please give the dates during which you would wi Building such a well-structured centralized repository of proteomics data therefore is this must between three to twelve months for eli extremely important for the continued growth and long-term success of (large-scale) around two months to process). protein identification strategies, the flagships of which are currently the Human Proteomics Organization (HUPO) projects. Another pivotal element in proteomics research today consists of cross-linkages Start date (dd/mm/yy): between the different protein sequence databases as well as links from these sequence 01/09/2003 databases to repositories of higher-order information such as the gene ontology database. I would very much like to be involved in defining and developing the data structure and submission/query tools for such a centralized, standardized and extensively linked proteomics data repository. Such data management skills will become ever more important for researchers in the field of proteomics as many labs are implementing massive high-throughput proteomics techniques.

The experience of the ERI will allow me to study detabase design and implementation

Martens, EuPA Open Proteomics, in press



A large amount of post-consumer MS data is collected in public databases such as PRIDE



The identified peptides reported by the community proved highly informative



Foster, Proteomics, 2011; Colaert, Nature Methods, 2011; Barsnes, Proteomics 2011, Vandermarliere, Proteomics 2013; Degroeve, Bioinformatics 2013



Meanwhile, we took an interest in the unidentified part of the data





We built the ReSpin pipeline to enable fast re-processing of proteomics data in new ways





First of all, we had to understand the data to allow reliable re-interpretation



http://compomics.github.io/projects/pride-asa-pipeline.html Hulstaert, Journal of Proteomics, 2013



Then we had to make it easy to re-analyze these data with multiple search algorithms

			SearchGUI 2.6.5				
e Edit Help							
Input & Output							
Spectrum File(s)		Add Clear					
Search Settings		Add Edit					
Output Folder	EMBL_proteomics_tra	Browse					
Pre Processing (beta)						
	msconvert	ß	msconvert File Conversion - ProteoWizard web page	0			
Search Engines							
V	X!Tandem	<i>l</i> # é ∆	X!Tandem Search Algorithm - XITandem web page	0			
	MyriMatch	ay 🛆	MyriMatch Search Algorithm - MyriMatch web page	0			
	MS Amanda	NS Amanda 🛛 🕸 🏟 👌 MS Amanda Search Algorithm - <u>MS Amanda web page</u>					
	MS-GF+	MS-GF+ # 🔹 👌 MS-GF+ Search Algorithm - <u>MS-GF+ web page</u>					
	OMSSA	ASSA at the constant of the co					
	Comet	$R \land$	Comet Search Algorithm - <u>Comet web page</u>	0			
	Tide	<i>1</i> 1 € ∆	Tide Search Algorithm - Tide web page	0			
	Andromeda	11	Andromeda Search Algorithm - Andromeda web page	0			
ost Processing							
	Two was	<i>ltt</i> ≤ ∆	PeptideShaker - <u>Visualize the results in PeptideShaker</u>	0			
2	Place cite SearchO	II as Vaudal of a	Proteomics 2011-11/5/206-0	Start the Search			

http://compomics.github.io/projects/searchgui.html Vaudel, Proteomics, 2011



And finally, a means to collate, process and validate the results from our re-analyses









http://compomics.github.io/projects/peptide-shaker.html Vaudel, Nature Biotechnology, 2015



Incidentally, anyone can reprocess public data with PeptideShaker and about ten mouseclicks

Welcome to P	eptideShaker 1.8	.1	×								
<u>In</u>	New Project	Open Project									
			PRIDE Reshake								– a ×
		67329	Elle Edit Help PRIDE Projects (2236)								
	Start Search	PRIDE Reshake	Accession 1 02050174 2 020501860 3 020501860 4 020502880 5 020502880 5 02050280 5 02050280 9 02050280 9 02050284 10 02050282 11 02050292 11 02050292 11 02050292 12 02050295 12 02050005 12 020500000 12 02050000000 12 020500000000000000	The Decision August 51 Reveal the Nex effects decision in teerdification of proteins starty associated with Diabet DFL enderfloation of a service the reveal with the service of the service term of the service term of the memory of the service term of the service term of the memory of the service term of terms of the service of the service term of the service term of terms of the service term of the service term of terms of the service of the service term of terms of term of the service of the service term of the service term of terms of term of the service of the service term of the service term of terms of term of the service of the service term of the service term of terms of term of terms of the service term of terms of term of terms of term of terms of ter	Tea Bringou Baradou Baradou Baradou Baragou Baragou Baragou Baradou Baradou Baradou	Beddes Weich College	Tasses entryp color color voire heat - precessing flad conception flad	Final analysis of the second s	Instruments LTG Oxforge LTG Oxforge LTG Oxforge LTG Oxforge Velos LTG Oxforge Velos Disothe	Koravis Force	Date Date INL 201943-17 • PLETE 201943-16 • INL 201943-15 • INLETE 201943-15 •
Settings & Help	Ope PRIDE Data	Selection	11 20100000 14 201000000 15 201000000 16 201000000 17 201000000 18 201000000 19 201000000 19 201000000 20 201000000 21 2010000000 22 2010000000 23 20100000000 24 20100000000 25 20100000000 26 201000000000 27 20100000000 28 20100000000000000000000000000000000000	Consider a large set of the	Hatanodal Balancia Balancia Balancia Balancia Technical Balancia Balancia Balancia Balancia Balancia Balancia Balancia	Monocontrol transmission and transmissio	Bood on calana and the basis tasa tasa tasa tasa tasa tasa tasa	Caracterisation (Linearco) Consensational (Linearco) Consensational (Linearco) (Linearco) Consensational (Linearco) Researco) (Linearco) Resear	A Decembre Constraine O Exactive O Exactive O Exactive O Exactive O Exactive D O Definity Vites O Exactive U10 01 Exactive Exactive U10 01 Exactive	20 CCM 10 CCM 10 CCM 11 CCM 12 PAPE 12 PAPE 12 PAPE 13 PAPE 14 PAPE 15 PAPE 16 PAPE 16 PAPE 16 PAPE 16 PAPE 16 PAPE 16 PAPE	Butto 2010/2014 PLCTE 2016/8014 PMLTE 2016/8014 PML 2016/8014
PRIDE Public Data			Select a project to see the pro	pert details. For more details duit the Accession Infis. Project Span	10 L					Brows	# Public Data / Access Private Data
0			Assays for P00002305 (72)	764	Datases	Seedes	Topues	PTMs	Instrumenta	#Podeina #Peo	obdea #Spectra
omics		Private Data	1 0.000 1 0.000 2 0.000 3 0.000 4 0.000 4 0.000 5 0.000 6 0.000 7 0.000 8 0.000 9 0.000 10 0.000 10 0.000 10 0.000 10 0.000 10 0.000 10 0.000 11 0.000 12 0.000 13 0.000 14 0.000 15 0.000 16 0.000 17 0.000 18 0.000 19 0.000 10 0.000 11 0.000 12 0.000 12 0.000 13 0.000 14 0.000 15 0.000						00 00 00 00	Image Image <th< td=""><td></td></th<>	
				6							
			4497 - 4497		Sec. 2007 Sec. 2					Download Constant Con	
			Her you have found the no	one Pepilishaker as <u>"audel et al.</u> Kalues Biotechnol. 2015.]	I formatic peak lots, now data and PREE ANL.					Download	Rostnake PROE Das





We then automated everything on our Pladipus custom grid engine



http://compomics.github.io/projects/pladipus.html Verheggen, Journal of Proteome Research, 2016



Combining data sets brings clear benefits



Vizcaíno, Nature Biotechnology, 2014; Wilhelm, Nature, 2014; Kim, Nature, 2014



We already went hunting for translated IncRNAs, but we found very few (< 1%)





Volders, NAR, 2013; Volders, NAR, 2015

We were able to help confirm expression of small ORFs across human tissues



#PSMs per tissue per sORFs with more than 5 occurences



Data growth is increasing in PRIDE; more, and ever bigger data sets are submitted daily



Number of submissions

June 2012 - May 2015



Vaudel, Proteomics, 2015

Projections for data growth in PRIDE are therefore quite impressive



Courtesy of Dr. Juan Antonio Vizcaíno, Proteomics Team Leader, EMBL-EBI



All of the available data simply bristle with opportunity





Why should we really have open science?

Our biggest challenge is to make it work!

What can we do with open science?

Of sedimentation, opportunity, and an absence of dragons



A sociologist's take on our efforts towards (orthogonal) data reuse

"This desire to reactivate data is widespread, and Klie et al. are not alone in wanting to show that 'far from being places where data goes to die' (Klie et al., 2007: 190), such data collections can be mined for valuable information that could not be obtained in any other way."

"In attempting to **reactivate sedimented data** in order to enable its re-use, their first step was ..."

"... they are experiments in seeing, in furnishing ways of seeing how data on proteins could become re-usable, could be reactivated as **collective property rather than the by-product of publication**."



One of the big ideas in science will be the (ortogonal-) re-use of (big) public data



And now think about open science, and imagine the opportunities

- What could you do with open science? What could you study? What could you learn?
- What opportunities would present themselves, if...
 - All data (in your field) were available online
 - All algorithms (in your field) were available online
 - All publications (in your field) were open access
- Most of these opportunities are not little steps forward; instead they promise to be revolutionary!





J.R.R. Tolkien, A Conversation with Smaug

Here is treasure of unlimited size, with all dragons chased away – *now what will you do?*



https://www.flickr.com/photos/fantasy-art-and-portraits/2884954207 (CC BY-NC-SA 2.0)















































