EMBL-EBI workshop: Exploring plant gene expression in Expression Atlas

https://www.ebi.ac.uk/~lauhuema/workshop/fg_Copenhagen2018

Laura Huerta, PhD

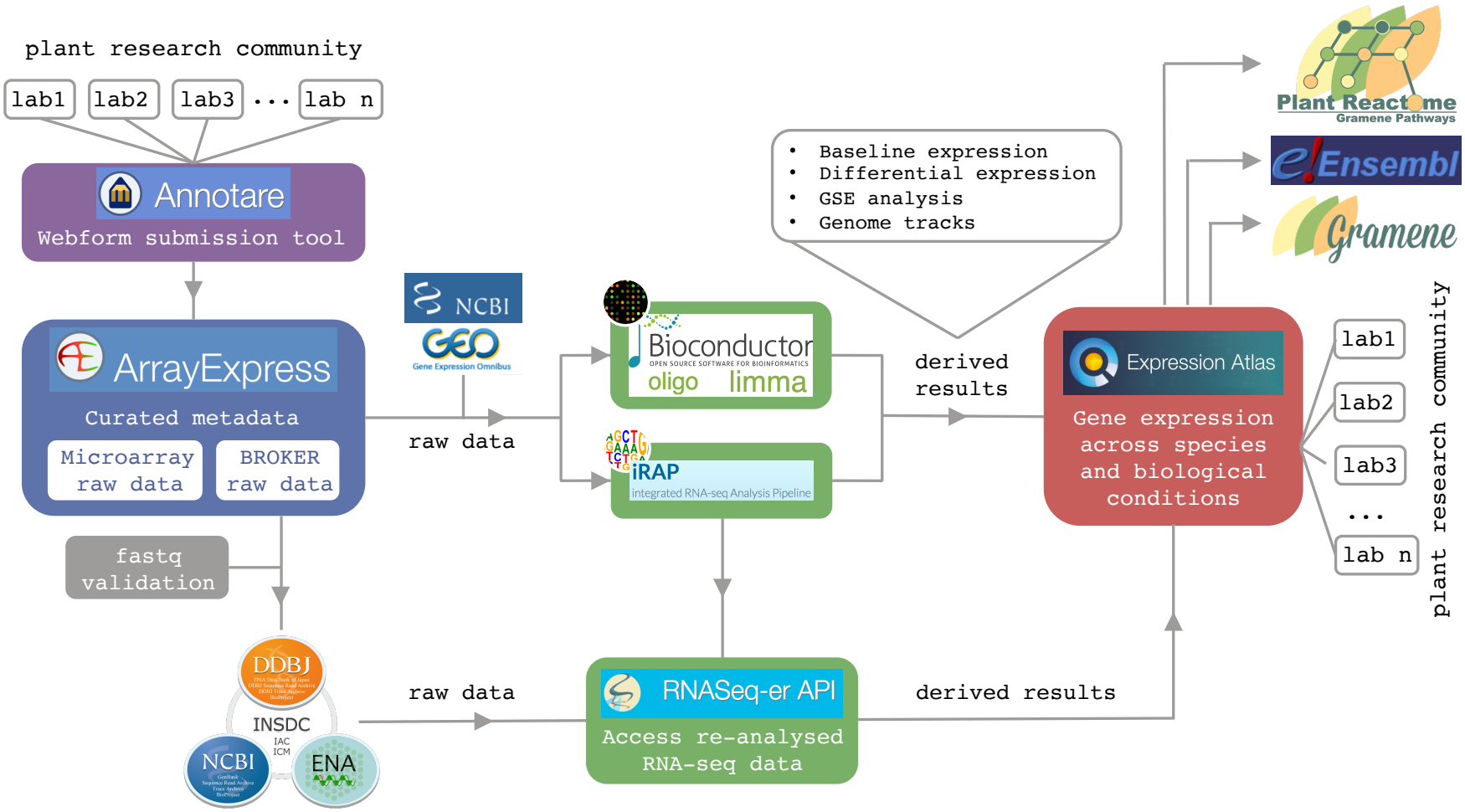
Senior Scientific Curator

lauhuema@ebi.ac.uk

Copenhagen, 10 September 2018

EMBL-EBI

Functional genomics resources at EMBL-EBI



EMBL-EBI workshop: Exploring plant gene expression in Expression Atlas

Data reproducibility: standards and ontologies

Laura Huerta, PhD

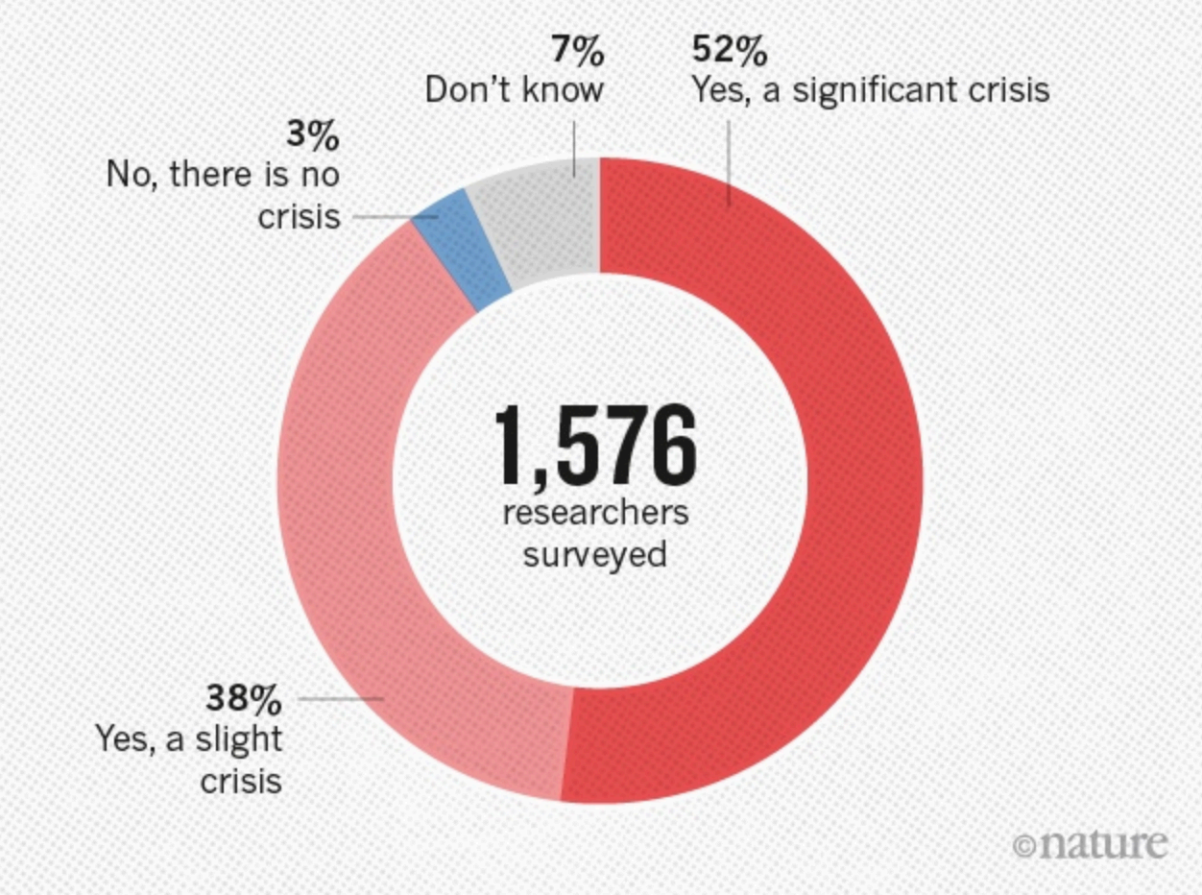
Senior Scientific Curator

lauhuema@ebi.ac.uk

Copenhagen, 10 September 2018

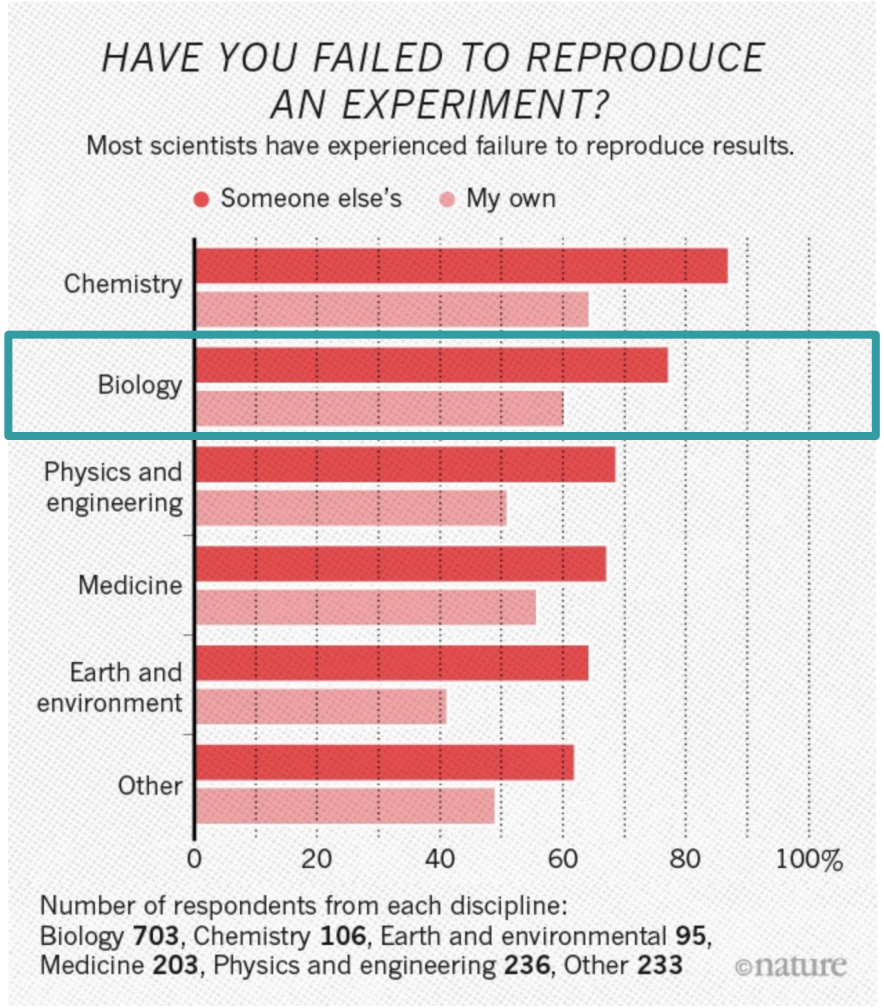
EMBL-EBI

Is there a reproducibility crisis in science?

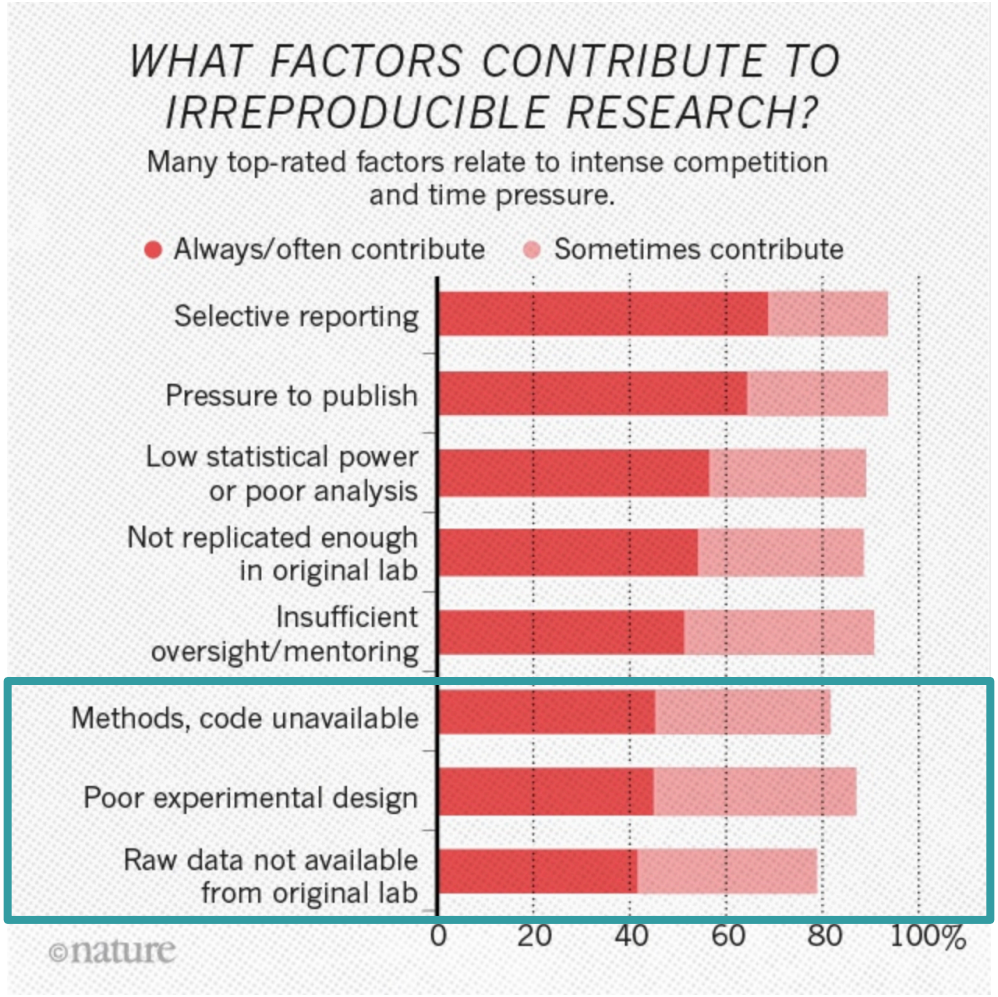


Nature 533, 452–454 (26 May 2016); doi:10.1038/533452a

Reproducibility crisis in science



Reproducibility crisis in science



Reproducibility: developing data standards

Metadata

Data standards in functional genomics

Microarray data

Nature Genetics **29**, 365 - 371 (2001)
doi:10.1038/ng1201-365

Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma[1], Pascal Hingamp[2], John Quackenbush[3], Gavin Sherlock[4], Paul Spellman[5], Chris Stoeckert[6], John Aach[7], Wilhelm Ansorge[8], Catherine A. Ball[4], Helen C. Causton[9], Terry Gaasterland[10], Patrick Glenisson[11], Frank C.P. Holstege[12], Irene F. Kim[4], Victor Markowitz[13], John C. Matese[4], Helen Parkinson[1], Alan Robinson[1], Ugis Sarkans[1], Steffen Schulze-Kremer[14], Jason Stewart[15], Ronald Taylor[16], Jaak Vilo[1] & Martin Vingron[17]

"*raw data is not enough to interpret the results and to verify the conclusions based on microarray data analysis*"

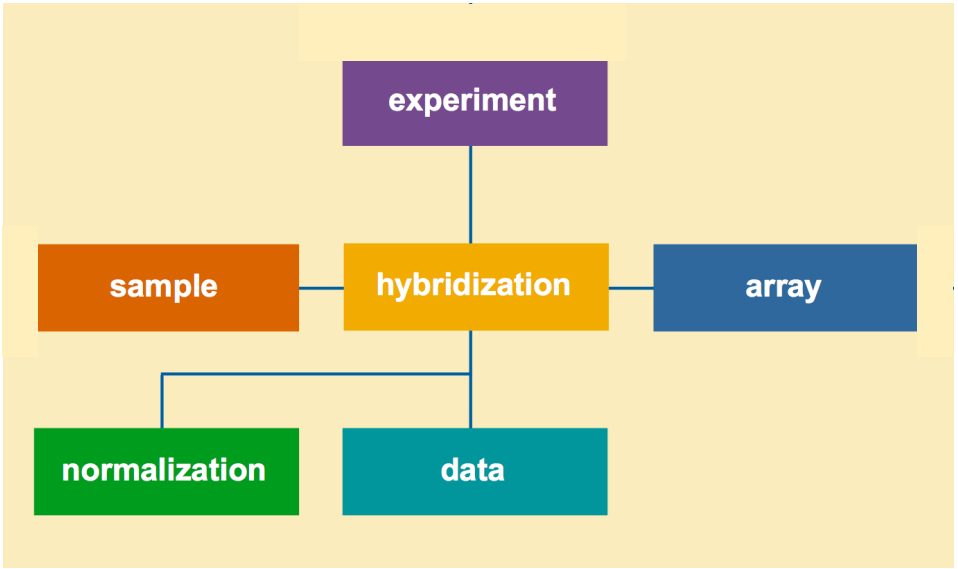
Data standards in functional genomics

"MIAME describes the data and metadata that authors must provide to support conclusions drawn from a microarray investigation, so that the data obtained in the investigation can be interpreted unambiguously and the investigation can be reproduced."

Reproducible

Raw data

Metadata



Data standards in functional genomics

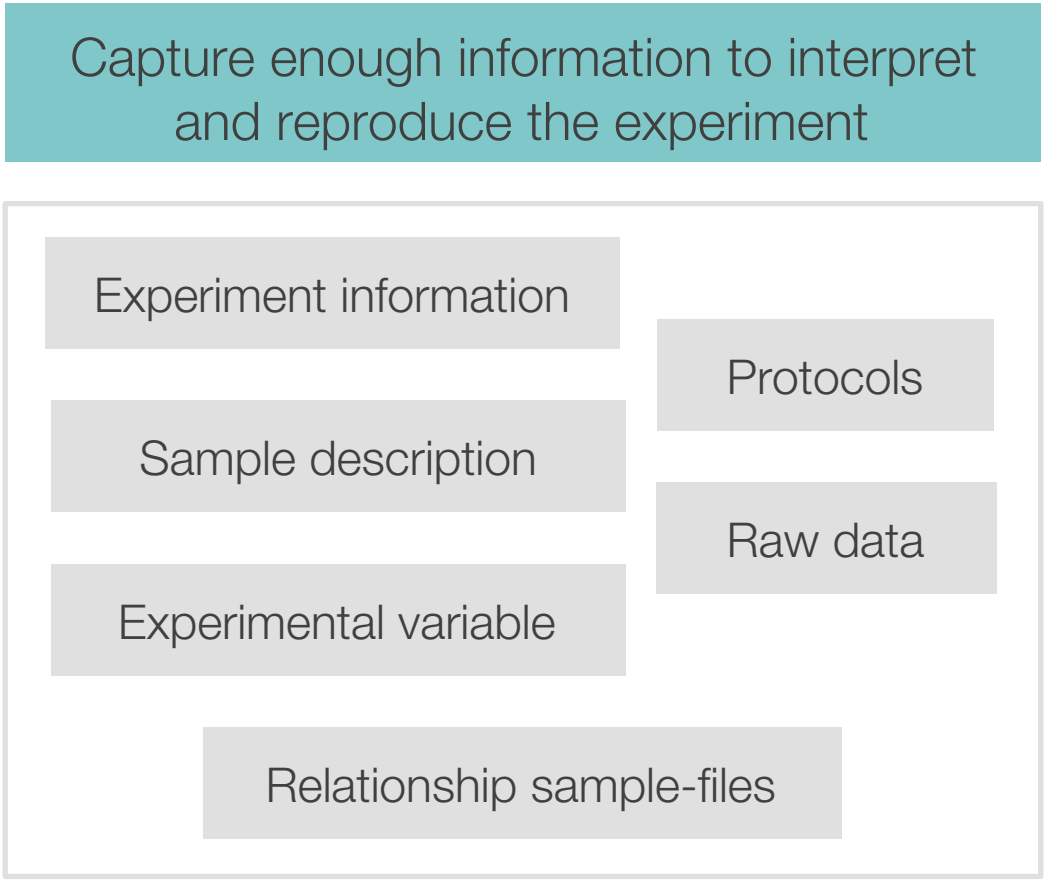
RNA-sequencing

FGED
Functional Genomics Data Society

1. **The description of the biological system, samples, and the experimental variables being studied**:
   - "compound" and "dose" in dose-response experiments or "antibody" in ChIP-Seq experiments, the organism, tissue, and the treatment(s) applied.
2. **The sequence read data for each assay**:
   - read sequences and base-level quality scores for each assay; FASTQ format is recommended, with a description of the scale used for quality scores.
3. **The 'final' processed (or summary) data for the set of assays in the study**:
   - the data on which the conclusions in the related publication are based, and descriptions of the data format.
4. **General information about the experiment and sample-data relationships**:
   - a summary of the experiment and its goals, contact information, any associated publication, and a table specifying sample-data relationships.
5. **Essential experimental and data processing protocols**:
   - how the nucleic acid samples were isolated, purified and processed prior to sequencing, a summary of the instrumentation used, library preparation strategy, labelling and amplification methodologies, alignment algorithms and data filtering plus data processing & analysis protocols.

http://fged.org/projects/minseqe

Data standards in functional genomics



How do we store all that information?

MAGE-TAB specification

1. *Investigation Description Format*
2. *Sample and Data Relationship Format*
3. *Raw data files*

Investigation
Description
Format

Accession: E-MTAB-9999
Title: "Transcription profiling of...
Description: "In this experiment...
Contacts: "r.e.searcher@lab...
Protocol: "Growth protocol...
Citation: "Dynamics of...

IDF describes the experiment

Sample and
Data
Relationship
Format

Characteristics of the samples

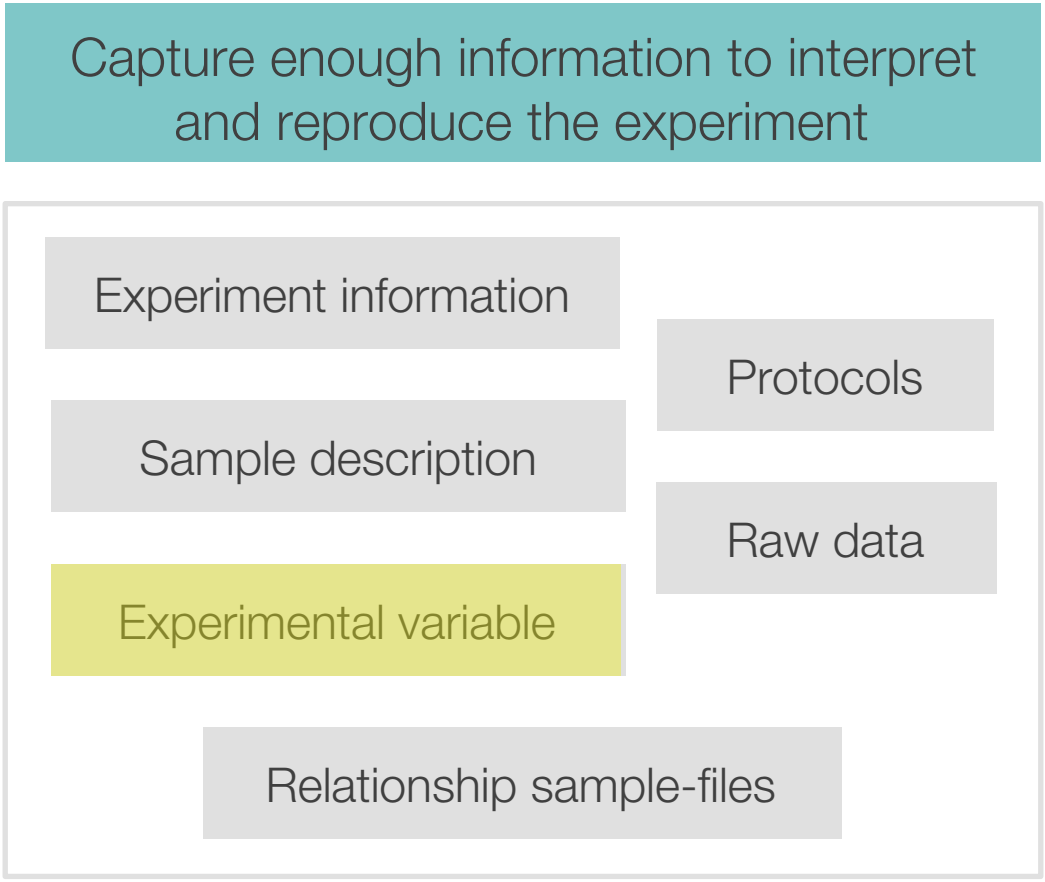
Experimental variables

Relationship samples -> files

Sample Name	Attributes	Assay Name	File
Sample 1	skin, epithelial cell	Hyb_sample1	S1.CEL
Sample 2	skin, epithelial cell	Hyb_sample2	S2.CEL
...			

SDRF describes individual samples and how they relate to data files

Data standards in functional genomics



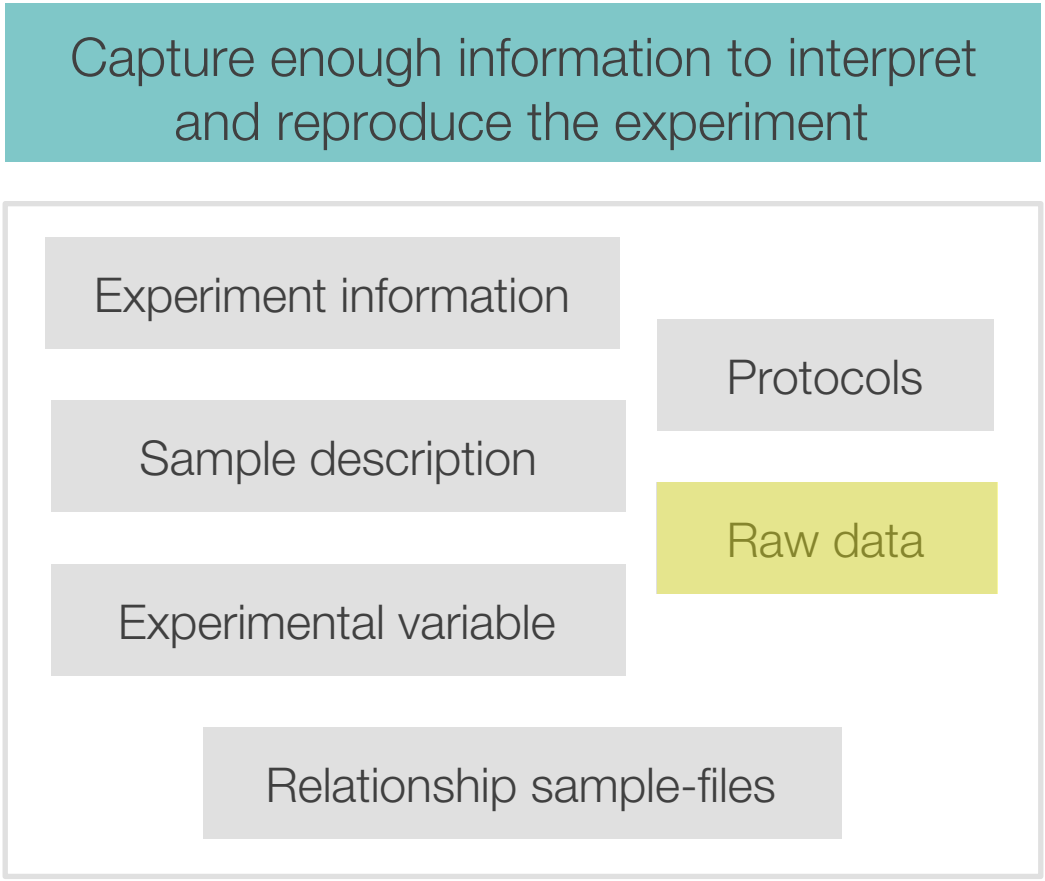
Experimental variable

Experimental variable → It is the main factor that you are investigating

Transcriptome profiling of short-term response to chilling stress in resistant and susceptible rice seedlings

- organism > *Oryza sativa Japonica Group*
- organism part > shoot
- cultivar > Thaibonnet OR Volcano
- age > 15 day
- genotype > wild type
- phenotype > cold stress sensitive OR tolerant
- environmental stress > none OR cold temperature regimen

Data standards in functional genomics



FASTQ format: RNA-seq raw data files

FASTQ file = FASTA + Quality

A FASTQ file is the common file format for sharing sequencing read data combining both the sequence and an associated per base quality score.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)

For each read:

1. @ Read ID
2. Nucleotide sequence of the read
3. +
4. Quality score for each nucleotide of the read

en.wikipedia.org/wiki/FASTQ_format

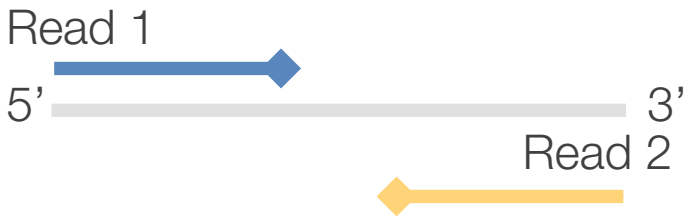
Single-end and paired-end RNA-seq

Single-end sequencing

my_sequence.fastq

```
@HWI-BRUNOP16X_0001:1:1:1466:1018#0/1
AAGGAAGTGCTTGTCTGGCTAACACAGCNAGNCACGTGAC
+
aVfbe`^^^_TTTSSdffffdfffabbZbbfebafbbbbb
```

Paired-end sequencing



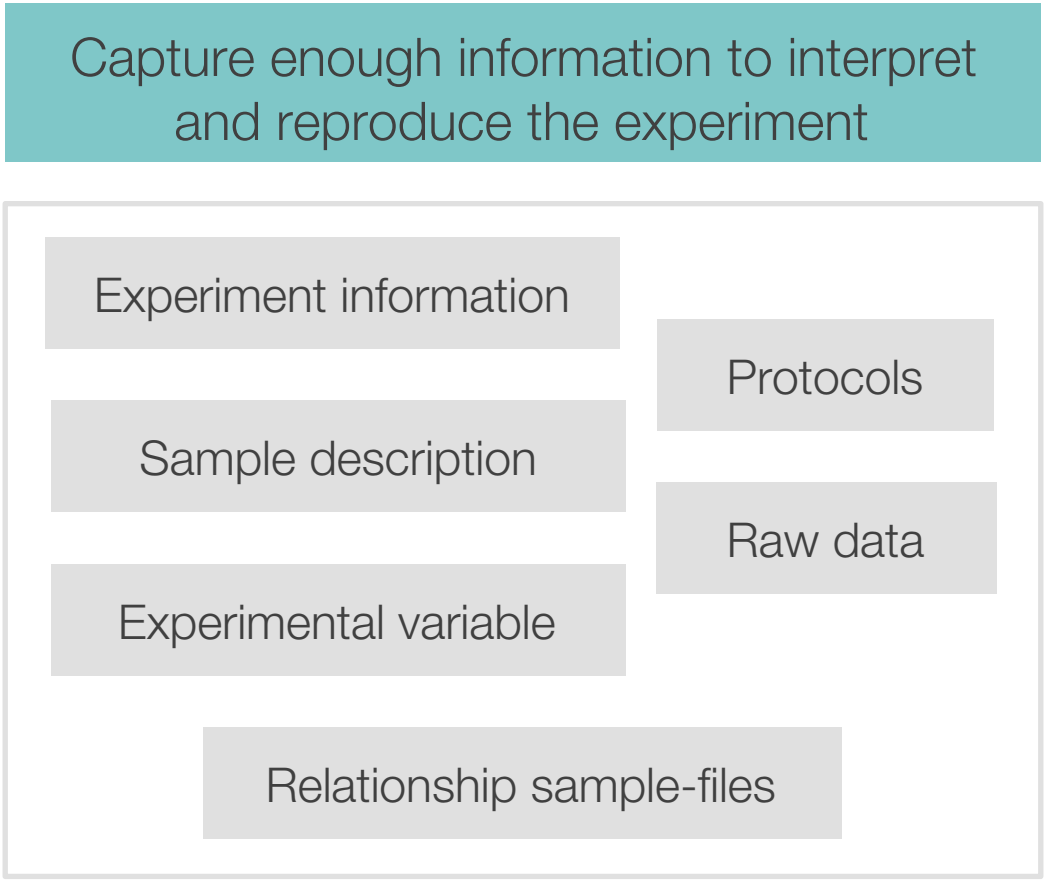
my_sequence_1.fastq

```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/1
NAAATTTCGAATTTCTGTGAAGTAAGCATCTTCTTTGTCAT
+
BJJGGKIINN^^^^^QQNTUQOOTTTRTOTY^^Y^\\^^^
```

my_sequence_2.fastq

```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/2
AACCCACACAGGAGAGCAGCCTTACAGATGCAAATACTGTG
+
]K___fffffggghgegggggdgggggfgggggeggggghh
```

Data standards in functional genomics



Advantages of using data standards

Authoring more complete and standardised metadata…

- Will aid dataset:
  - Discovery
  - Exploration
  - Integration
  - Secondary use
- Will aid communication of scientific results
- Will aid knowledge management within research organizations
- Will make your data more FAIR

Findable | Accessible | Interoperable | Re-usable

Authoring better metadata in spreadsheets

1. Choose good names for things → Ontologies

Arabidopsis thaliana
Columbia-0

Columbia-0 is a ...

- background
- strain
- line
- cultivar
- accession
- plant line
- variation
- subspecies
- genotype
- accessions
- genetic background
- ecotype/background
- genotype/ecotype
- accession name
- ecotype

Authoring better metadata in spreadsheets

2. Be consistent! → Ontologies

Pick one term and stick to it

- drought conditions
- drought treated plant
- water deficit
- drought
- D1
- water stress
- water-limited
- drought treatment
- drought stress
- WS
- water deprivation
- water deficit stress
- drought treated sample
- Drought
- drought environment
- drought treated

Authoring better metadata in spreadsheets

3. Keep things simple → Single concept in a cell

genotype	ecotype	genotype
wild type Col-0	Col-0	wild type

treatment	compound	dose	concentration
ABA 25 μM	abscisic acid	25	micromolar

organism part	organism	organism part
Arabidopsis inflorescence	Arabidopsis thaliana	inflorescence

Authoring better metadata in spreadsheets

4. Make it a rectangle

	A	B	C	D
1	0 hour		1 hour	
2	control		sodium chloride	
3	wild type	bzip1, bzip53 double knockout	wild type	bzip1, bzip53 double knockout
4	GSM1824868	GSM1824880	GSM1824871	GSM1824883
5	GSM1824867	GSM1824879	GSM1824870	GSM1824882
6	GSM1824866	GSM1824878	GSM1824869	GSM1824881
7				
8	3 hour		6 hour	
9	sodium chloride		sodium chloride	
10	wild type	bzip1, bzip53 double knockout	wild type	bzip1, bzip53 double knockout
11	GSM1824874	GSM1824886	GSM1824877	GSM1824889
12	GSM1824873	GSM1824885	GSM1824876	GSM1824888
13	GSM1824872	GSM1824884	GSM1824875	GSM1824887

Authoring better metadata in spreadsheets

	A	B	C	D	E
1	Sample Name	genotype	compound	time	unit
2	GSM1824866	wild type	control	0	hour
3	GSM1824867	wild type	control	0	hour
4	GSM1824868	wild type	control	0	hour
5	GSM1824869	wild type	sodium chloride	1	hour
6	GSM1824870	wild type	sodium chloride	1	hour
7	GSM1824871	wild type	sodium chloride	1	hour
8	GSM1824872	wild type	sodium chloride	3	hour
9	GSM1824873	wild type	sodium chloride	3	hour
10	GSM1824874	wild type	sodium chloride	3	hour
11	GSM1824875	wild type	sodium chloride	6	hour
12	GSM1824876	wild type	sodium chloride	6	hour
13	GSM1824877	wild type	sodium chloride	6	hour
14	GSM1824878	bzip1, bzip53 double knockout	control	0	hour
15	GSM1824879	bzip1, bzip53 double knockout	control	0	hour
16	GSM1824880	bzip1, bzip53 double knockout	control	0	hour
17	GSM1824881	bzip1, bzip53 double knockout	sodium chloride	1	hour
18	GSM1824882	bzip1, bzip53 double knockout	sodium chloride	1	hour
19	GSM1824883	bzip1, bzip53 double knockout	sodium chloride	1	hour
20	GSM1824884	bzip1, bzip53 double knockout	sodium chloride	3	hour
21	GSM1824885	bzip1, bzip53 double knockout	sodium chloride	3	hour
22	GSM1824886	bzip1, bzip53 double knockout	sodium chloride	3	hour
23	GSM1824887	bzip1, bzip53 double knockout	sodium chloride	6	hour
24	GSM1824888	bzip1, bzip53 double knockout	sodium chloride	6	hour
25	GSM1824889	bzip1, bzip53 double knockout	sodium chloride	6	hour

Authoring better metadata in spreadsheets

5. Save data in plain text files

TSV file format

tab-separated values

- delimited text file that uses tab to separate values
- stores tabular data in plain text
- each line is a data record
- each record consists of one or more fields, separated by tab character

CSV file format

comma-separated values

- delimited text file that uses comma to separate values
- stores tabular data in plain text
- each line is a data record
- each record consists of one or more fields, separated by commas

Authoring better metadata in spreadsheets

Data standards

Hands-on activity

Let's become a biocurator,
at least for 15 min…

In pairs

Biocuration: improving data quality

Biocuration → Translation and integration of information relevant to biology into a database or resource

- Raw data
- Metadata

Raw data

- unprocessed data files
- **Microarray:** files from the scanner (e.g. Affymetrix CEL files, Agilent feature extraction *txt* files, Illumina *idat* files)
- **Sequecing:** raw sequence read files (e.g. FASTQ files)

Metadata

- Experiment description
- Experiment title
- Sample annotation
- Protocols
- Publication details (if any)
- Author contact details

Biocuration: improving data quality

Making data reusable

PeerJ

✔ PEER-REVIEWED

XA21-specific induction of stress-related genes following *Xanthomonas* infection of detached rice leaves

Data Availability

The following information was supplied regarding data availability:

The National Center for Biotechnology Information Sequence Read Archive (SRA)

BioProject ID PRJNA250865.

Metadata

7 samples x 3 biological replicates

RNA sequencing sample treatment summary

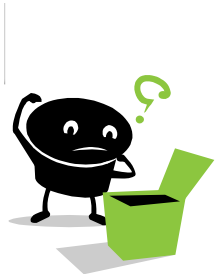
Table summarizes the experimental setup including the genotypes, time of treatment, and type of treatment used for samples used in RNA sequencing. There were three replicates for each sample for a total of 21 sequenced samples.

	A	B	C
1	**Genotype**	**Time (hours)**	**Treatment**
2	Kitaake	0	None
3	EFR::XA21::GFP	0	None
4	EFR::XA21::GFP	0.5	500 nM elf18
5	EFR::XA21::GFP	1	500 nM elf18
6	EFR::XA21::GFP	3	500 nM elf18
7	EFR::XA21::GFP	6	500 nM elf18
8	EFR::XA21::GFP	12	500 nM elf18

Biocuration: improving data quality

Making data reusable

Raw data



- Genotype?
- Treatment?
- Time?

SRX873376: Other Sequencing of Japanese rice
1 ILLUMINA (Illumina HiSeq 2000) run: 29.9M spots, 9G bases, 5.3Gb downloads

Submitted by: DOE JOINT GENOME INSTITUTE (JGI)

Study: Oryza sativa Japonica strain:EFR-XA21 | cultivar:Kitaake Transcriptome or Gene expression
PRJNA250865 • SRP054056 • All experiments • All runs
show Abstract

Sample: Oryza sativa cv. Kitaake EFR-XA21
SAMN03003383 • SRS843490 • All experiments • All runs
Organism: Oryza sativa Japonica Group

Library:
Name: NTBA
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: RT-PCR
Layout: PAIRED

Spot descriptor:
1 forward 151 reverse

Runs: 1 run, 29.9M spots, 9G bases, 5.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR1799213	29,893,272	9G	5.3Gb	2015-02-20

Biocuration: improving data quality

Making data reusable

Raw data

Metadata

Run	FactorValue[genotype]	FactorValue[compound]	FactorValue[dose]	Unit[concentration unit]	FactorValue[time]	Unit[time unit]
SRR1799194	wild type	none	0	nanomolar	0	hour
SRR1799197	wild type	none	0	nanomolar	0	hour
SRR1799193	wild type	none	0	nanomolar	0	hour
SRR1799210	EFR:XA21:GFP	none	0	nanomolar	0	hour
SRR1799203	EFR:XA21:GFP	none	0	nanomolar	0	hour
SRR1799205	EFR:XA21:GFP	none	0	nanomolar	0	hour
SRR1799199	EFR:XA21:GFP	elf18	500	nanomolar	0.5	hour
SRR1799204	EFR:XA21:GFP	elf18	500	nanomolar	0.5	hour
SRR1799212	EFR:XA21:GFP	elf18	500	nanomolar	0.5	hour
SRR1799208	EFR:XA21:GFP	elf18	500	nanomolar	1	hour
SRR1799202	EFR:XA21:GFP	elf18	500	nanomolar	1	hour
SRR1799198	EFR:XA21:GFP	elf18	500	nanomolar	1	hour
SRR1799213	EFR:XA21:GFP	elf18	500	nanomolar	3	hour
SRR1799201	EFR:XA21:GFP	elf18	500	nanomolar	3	hour
SRR1799196	EFR:XA21:GFP	elf18	500	nanomolar	3	hour
SRR1799209	EFR:XA21:GFP	elf18	500	nanomolar	6	hour
SRR1799206	EFR:XA21:GFP	elf18	500	nanomolar	6	hour
SRR1799195	EFR:XA21:GFP	elf18	500	nanomolar	6	hour
SRR1799207	EFR:XA21:GFP	elf18	500	nanomolar	12	hour
SRR1799200	EFR:XA21:GFP	elf18	500	nanomolar	12	hour
SRR1799211	EFR:XA21:GFP	elf18	500	nanomolar	12	hour

Biocuration: improving data quality

Sample	Treatment
Sample A_control	Compound X
Sample B_treated	Compound X
Sample C_control	control
Sample D_treated	Compound X

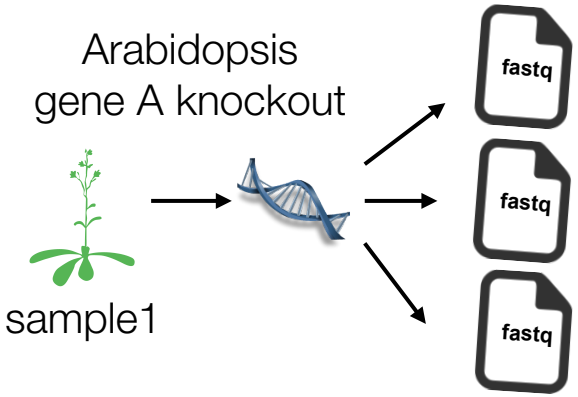
We compared knockouts against wild-type plants

Sample	Genotype
Sample 1	gene X KO
Sample 2	gene X KO
Sample 3	gene X KO

Sample	Organism part
Sample D	flower
Sample D	shoot
Sample D	root

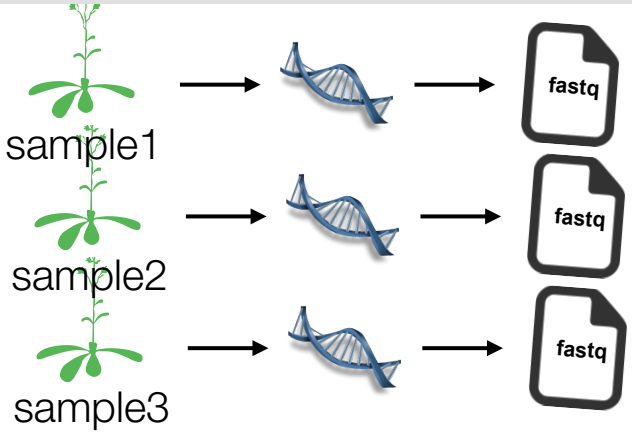
Biocuration: improving data quality

Technical replicates



Sample	Seq. run	File
Sample 1	run_assay_1	1.fq.gz
Sample 2	run_assay_2	2.fq.gz
Sample 3	run_assay_3	3.fq.gz

Biological replicates



Sample	Seq. run	File
Sample 1	run_assay_1	1.fq.gz
Sample 1	run_assay_2	2.fq.gz
Sample 1	run_assay_3	3.fq.gz

Annotation with ontology terms

Ontology → A systematic way to name and organise entities, establishing relationships between the entities

- Controlled vocabulary
- Hierarchy (relationship)

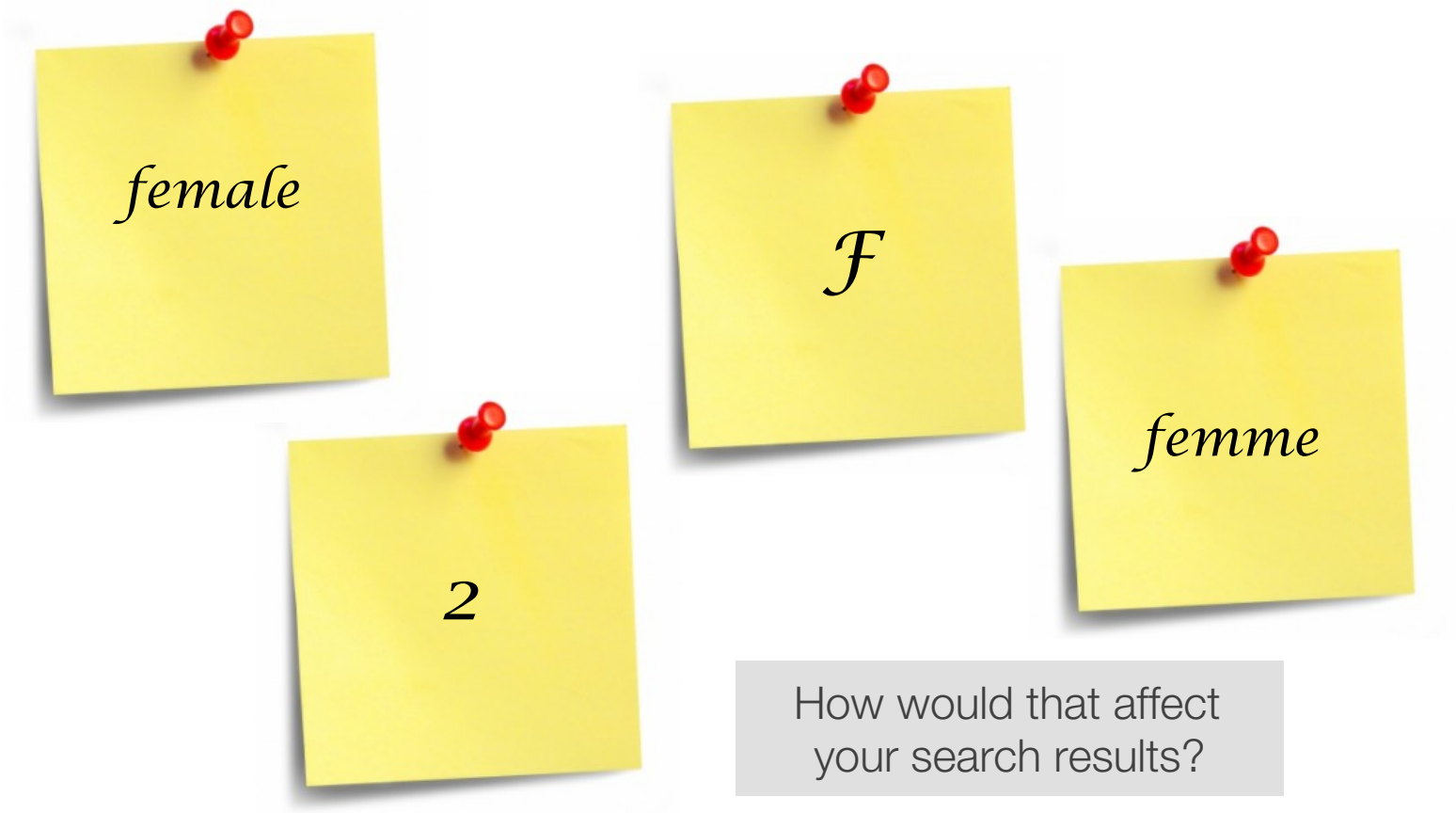
Representing experimental variables with EFO

The **Experimental Factor Ontology** (EFO) provides a systematic description of many experimental variables available in EBI databases, and for external projects such as the NHGRI GWAS catalog. It combines parts of several biological ontologies, such as UBERON anatomy, ChEBI chemical compounds, and Cell Ontology. The scope of EFO is to support the annotation, analysis and visualization of data handled by many groups at the EBI and as the core ontology for Open Targets. EFO is developed by the EMBL-EBI Samples, Phenotypes and Ontologies Team (SPOT). We also add terms for external users when requested. If you are new to ontologies, there is a short introduction on the subject available and a blog post by James Malone on what ontologies are for.

www.ebi.ac.uk/efo

Why we annotate samples using ontologies?

How many ways can you say “female”?



How would that affect your search results?

Why we annotate samples using ontologies?

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf.female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, spayed
dioecious female	female, worker	pseudohermaprhoditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",

Courtesy of N. Silvester, European Nucleotide Archive, EMBL-EBI

Ontology annotation

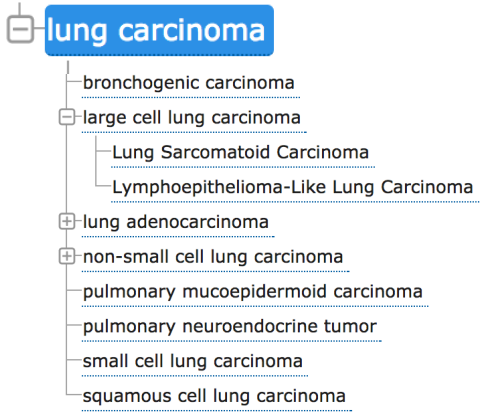
Ontology → Powerful search

Ontology → Data integration

Smart search

Efficient search via ontology-driven query expansion

Complex queries



Ontology annotation: adding value to your data

https://www.ncbi.nlm.nih.gov/geo

GSE81089 Search

Series GSE81089	Query DataSets for GSE81089
Status	Public on Jun 01, 2016
Title	Next Generation Sequencing (RNAseq) of non-small cell lung cancer
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	Cancer testis antigens (CTAs) are of clinical interest as biomarkers and present valuable targets for immunotherapy. To comprehensively characterize the CTA landscape of non-small cell lung cancer (NSCLC), we compared RNAseq data of 199 NSCLC tissues to the normal transcriptome of 142 samples from 32 different normal organs. Of 232 CTAs currently annotated in the CTdatabase, 96 were confirmed in NSCLC. To obtain an unbiased CTA profile of NSCLC, we applied stringent criteria on our RNAseq data set and defined 90 genes as CTAs, of which 55 genes were not annotated in the CTdatabase. Cluster analysis revealed that CTA expression is histology-dependent and concurrent expression is common. Immunohistochemistry confirmed tissue specific protein expression of selected genes. Furthermore, methylation was identified as a regulatory mechanism of CTA expression based on independent data from the Cancer Genome Atlas. The proposed prognostic impact of CTAs in lung cancer, was not confirmed, neither in our RNAseq-cohort nor in an independent meta-analysis of 1117 NSCLC cases.
Overall design	Fresh frozen tumor tissue from 199 patients diagnosed with NSCLC and surgically treated 2006-2010 at the Uppsala University Hospital, Uppsala, Sweden and 19 paired normal lung tissues. Clinical data were retrieved from the regional lung cancer registry. Several of the new CTAs are poorly characterized Sample characteristics values represent; pTNM: decided by Hans Brunnström, pathologist in Lund Spring 2013 Stage according to pTNM: 1=1a 2=1b 3=2a 4=2b 5=3a 6=3b 7=IV Histology diagnosis spring 2013 HB: 1=squamous cell cancer 2=AC unspecified 3=Large cell/ NOS Surgery date: the date when sample arrived at Patologen UAS Age: age when surgery was performed Vital date: day of death or latest contact Dead: 0=no 1= yes Smoking history : 1=current 2=ex >1year 3=never WHO performance status: Performance status 0-4 Please note that the L608T_2122, L771T_1 data columns (in the processed data files) are associated with L608T and L771T samples, respectively.
Contributor(s)	Djureinovic D, Hallström BM, Horie M, Mattsson JS, La Fleur L, Fagerberg L, Brunnström H, Lindskog C, Madjar K, Rahnenführer J, Ekman S, Ståhle E, Koyi H, Brandén E, Edlund K, Hengstler JG, Lambe M, Saito A, Botling J, Pontén F, Uhlén M, Micke P
Citation missing	*Has this study been published? Please login to update or notify GEO.*

EMBL-EBI

Ontology annotation: adding value to your data

https://www.ncbi.nlm.nih.gov/geo

GSE81089 Search

Sample GSM2142642

Status	Public on Jun 01, 2016
Title	matched sample_L511N
Sample type	SRA
Source name	Human non-malignant tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511N

non-small cell lung carcinoma

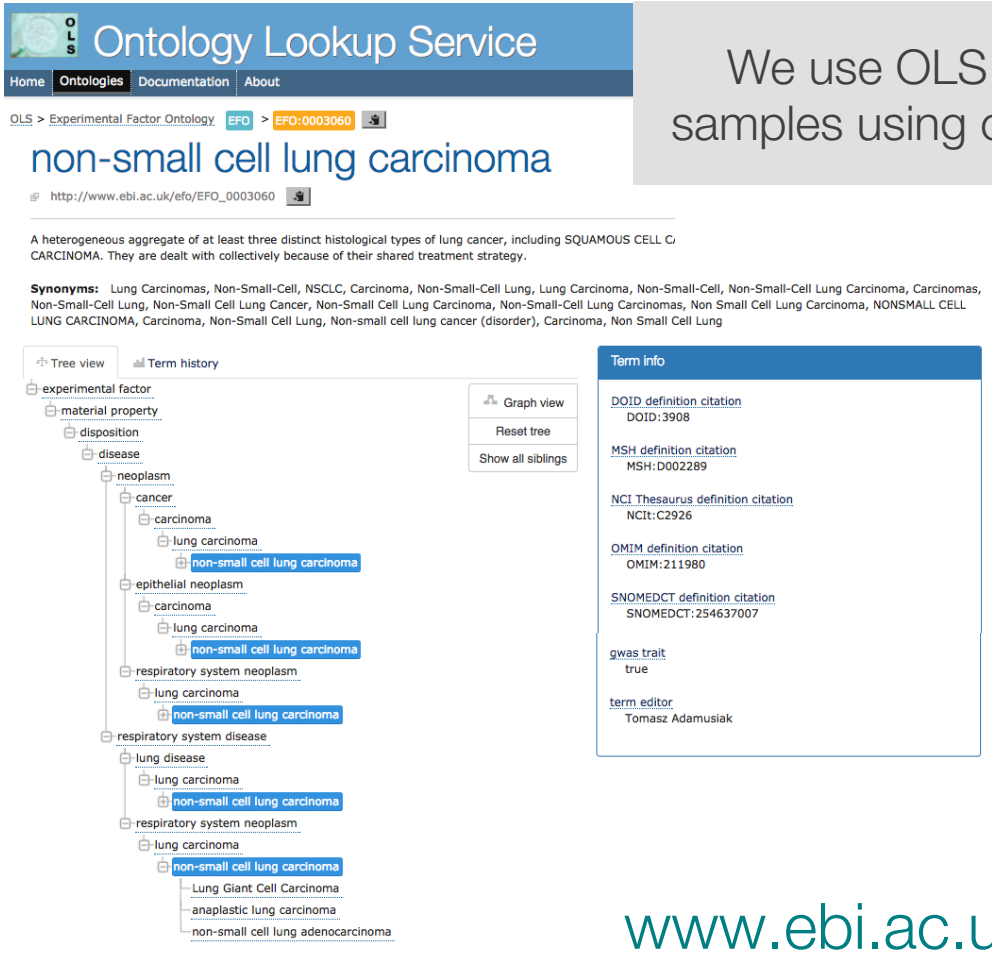
Sample GSM2142481

Status	Public on Jun 01, 2016
Title	L511T
Sample type	SRA
Source name	Human NSCLC tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511T stage tnm: 5 histology: 2 surgery date: 2007-04-05 age: 56 gender: female vital date: 2013-04-28 dead: 0 smoking: 3 ps who: 0

non-small cell lung carcinoma

Disease?

Ontology annotation: adding value to your data



We use OLS to annotate samples using ontology terms

www.ebi.ac.uk/ols/index

Ontology annotation: adding value to your data

https://www.ncbi.nlm.nih.gov/geo

GSE81089 Search

Sample GSM2142642

Status	Public on Jun 01, 2016
Title	matched sample_L511N
Sample type	SRA
Source name	Human non-malignant tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511N

lung

Sample GSM2142481

Status	Public on Jun 01, 2016
Title	L511T
Sample type	SRA
Source name	Human NSCLC tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511T stage tnm: 5 histology: 2 surgery date: 2007-04-05 age: 56 gender: female vital date: 2013-04-28 dead: 0 smoking: 3 ps who: 0

lung

Tissue?

Ontology annotation: adding value to your data

https://www.ncbi.nlm.nih.gov/geo → GSE81089 Search

Sample GSM2142642	
Status	Public on Jun 01, 2016
Title	matched sample_L511N
Sample type	SRA
Source name	Human non-malignant tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511N

non-tumor

Sample GSM2142481	
Status	Public on Jun 01, 2016
Title	L511T
Sample type	SRA
Source name	Human NSCLC tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511T stage tnm: 5 histology: 2 surgery date: 2007-04-05 age: 56 gender: female vital date: 2013-04-28 dead: 0 smoking: 3 ps who: 0

tumor

Sampling site?

Ontology annotation: adding value to your data

https://www.ncbi.nlm.nih.gov/geo

GSE81089 Search

Sample GSM2142642

Status	Public on Jun 01, 2016
Title	matched sample_L511N
Sample type	SRA
Source name	Human non-malignant tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511N

individual 511

Sample GSM2142481

Status	Public on Jun 01, 2016
Title	L511T
Sample type	SRA
Source name	Human NSCLC tissue
Organism	Homo sapiens
Characteristics	tumor (t) or normal (n): L511T
	stage tnm: 5
	histology: 2
	surgery date: 2007-04-05
	age: 56
	gender: female
	vital date: 2013-04-28
	dead: 0
	smoking: 3
	ps who: 0

individual 511

Same patient?

Data and metadata in public repositories

E-GEOD-81089 - RNA-seq of 199 non-small cell lung carcinoma patients and 19 paired normal lung tissues

Status	*Released on 1 June 2016, last updated on 22 August 2017*
Organism	Homo sapiens
Samples (218)	Click for detailed sample information and links to data
Protocols (3)	Click for detailed protocol information
Description	Cancer testis antigens (CTAs) are of clinical interest as biomarkers and present valuable targets for immunotherapy. To comprehensively characterize the CTA landscape of non-small cell lung cancer (NSCLC), we compared RNAseq data of 199 NSCLC tissues to the normal transcriptome of 142 samples from 32 different normal organs. Of 232 CTAs currently annotated in the CTdatabase, 96 were confirmed in NSCLC. To obtain an unbiased CTA profile of NSCLC, we applied stringent criteria on our RNAseq data set and defined 90 genes as CTAs, of which 55 genes were not annotated in the CTdatabase. Cluster analysis revealed that CTA expression is histology-dependent and concurrent expression is common. Immunohistochemistry confirmed tissue specific protein expression of selected genes. Furthermore, methylation was identified as a regulatory mechanism of CTA expression based on independent data from the Cancer Genome Atlas. The proposed prognostic impact of CTAs in lung cancer, was not confirmed, neither in our RNAseq-cohort nor in an independent meta-analysis of 1117 NSCLC cases. Fresh frozen tumor tissue from 199 patients diagnosed with NSCLC and surgically treated 2006-2010 at the Uppsala University Hospital, Uppsala, Sweden and 19 paired normal lung tissues. Clinical data were retrieved from the regional lung cancer registry. Several of the new CTAs are poorly characterized Sample characteristics values represent; pTNM: decided by Hans Brunnstrom, pathologist in Lund Spring 2013 Stage according to pTNM: 1=1a 2=1b 3=2a 4=2b 5=3a 6=3b 7=IV Histology diagnosis spring 2013 HB: 1=squamous cell cancer 2=AC unspecified 3=Large cell/ NOS Surgery date: the date when sample arrived at Patologen UAS Age: age when surgery was performed Vital date: day of death or latest contact Dead: 0=no 1= yes Smoking history : 1=current 2=ex >1year 3=never WHO performance status: Performance status 0-4 Please note that the L608T_2122, L771T_1 data columns (in the processed data files) are associated with L608T and L771T samples, respectively.
Experiment type	RNA-seq of coding RNA
Contacts	Bjorn M Hallstrom <bjorn.hallstrom@scilifelab.se>, Akira Saito, Cecilia Lindskog, Dijana Djureinovic, Elisabeth Stahle, Eva Branden, Fredrik Ponton, Hans Brunnstrom, Hirsh Koyi, Jan G Hengstler, Johan Botling, Johanna S Mattsson, Jorg Rahnenfuhrer, Karolina Edlund, Katrin Madjar, Linn Fagerberg, Linnea La Fleur, Masafumi Horie, Mathias Uhlen, Mats Lambe, Patrick Micke, Simon Ekman
Citation	Profiling cancer testis antigens in non-small-cell lung cancer. Djureinovic D, Hallstrom BM, Horie M, Mattsson JS, La Fleur L, Fagerberg L, Brunnstrom H, Lindskog C, Madjar K, Rahnenfuhrer J, Ekman S, Stahle E, Koyi H, Branden E, Edlund K, Hengstler JG, Lambe M, Saito A, Botling J, Ponten F, Uhlen M, Micke P. , Europe PMC 27699219
MINSEQE	Exp. design ✱ Protocols ✱ Variables ✱ Processed − Seq. reads ✱
Files	Investigation description E-GEOD-81089.idf.txt Sample and data relationship E-GEOD-81089.sdrf.txt Additional data (1) E-GEOD-81089.additional.1.zip Click to browse all available files
Links	ENA - SRP074349, GEO - GSE81089 Send E-GEOD-81089 data to GENOMESPACE

EMBL-EBI

Data and metadata in public repositories

E-GEOD-81089 - RNA-seq of 199 non-small cell lung carcinoma patients and 19 paired normal lung tissues

Display full sample-data table | Export table in Tab-delimited format

Page 1 2 3 4 5 6 .. 54 | Showing 76 - 100 of 1340 rows | Page size 25 50 100 250 500

	Sample Attributes									Links to Data	
Source Name	organism	individual	sex	age	organism part	disease	sampling site	clinical information	s	ENA	FASTQ
GSM2142480 1	Homo sapiens	L504	female	65 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142481 1	Homo sapiens	L511	female	56 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	non-smoker	tu		
GSM2142481 1	Homo sapiens	L511	female	56 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	non-smoker	tu		
GSM2142642 1	Homo sapiens	L511	female	56 *(year)*	lung	non-small cell lung carcinoma	non-malignant tissue	non-smoker	n		
GSM2142642 1	Homo sapiens	L511	female	56 *(year)*	lung	non-small cell lung carcinoma	non-malignant tissue	non-smoker	n		
GSM2142482 1	Homo sapiens	L529	female	79 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142482 1	Homo sapiens	L529	female	79 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142483 1	Homo sapiens	L530	female	76 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	smoker	tu		
GSM2142483 1	Homo sapiens	L530	female	76 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	smoker	tu		
GSM2142484 1	Homo sapiens	L531	male	74 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142484 1	Homo sapiens	L531	male	74 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142485 1	Homo sapiens	L532	female	69 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142485 1	Homo sapiens	L532	female	69 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142643 1	Homo sapiens	L532	female	69 *(year)*	lung	non-small cell lung carcinoma	non-malignant tissue	non-smoker	n		
GSM2142643 1	Homo sapiens	L532	female	69 *(year)*	lung	non-small cell lung carcinoma	non-malignant tissue	non-smoker	n		
GSM2142486 1	Homo sapiens	L534	male	66 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142486 1	Homo sapiens	L534	male	66 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142487 1	Homo sapiens	L535	male	65 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142487 1	Homo sapiens	L535	male	65 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142488 1	Homo sapiens	L538	male	56 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	smoker	tu		
GSM2142488 1	Homo sapiens	L538	male	56 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	smoker	tu		
GSM2142489 1	Homo sapiens	L539	male	78 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142489 1	Homo sapiens	L539	male	78 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	ex-smoker	tu		
GSM2142490 1	Homo sapiens	L541	female	70 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	non-smoker	tu		
GSM2142490 1	Homo sapiens	L541	female	70 *(year)*	lung	non-small cell lung carcinoma	tumor tissue	non-smoker	tu		

Data and metadata in public repositories

- Open, online access to experimental datasets
- Annotation of datasets with adequate metadata
- Use of controlled terms in metadata annotations -> ontologies
- Mechanisms to search for metadata to find relevant experimental results

Open access | Metadata | Ontologies | Discovery

EMBL-EBI workshop: Exploring plant gene expression in Expression Atlas

Data reproducibility: standards and ontologies

Laura Huerta, PhD

Senior Scientific Curator

lauhuema@ebi.ac.uk

Copenhagen, 10 September 2018

EMBL-EBI