Facilitating reproducible research:

from the perspective of a public data archive, ArrayExpress, at EMBL-EBI

Amy Tang

Functional Genomics Curation and Training Project Leader

EMBL-EBI

EMBL-EBI

Talk outline

- What is ArrayExpress?
- Facilitating reproducibility
  1. Submission timing is key
  2. The ideal world: standards, and ontologies.
  3. When standards fail --- why we need curators
  4. When the dust settles --- meta-data updates
- Further challenges

ArrayExpress at EMBL-EBI

www.ebi.ac.uk/arrayexpress (daily release at 6am UK time)

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis

Advanced

Home | Browse | Submit | Help | About ArrayExpress | Feedback | Login

ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

Browse ArrayExpress

Data Content

Updated today at 06:30

- 61020 experiments
- 1835198 assays
- 39.28 TB of archived data

Latest News

10 September 2015 - **ArrayExpress submission now supports "double-blind" peer review**

"Double-blind" review happens when the manuscript authors and reviewers remain anonymous to one another. It is bias during the review process, and some journals have started supporting it. ArrayExpress submission tool "Annotare review by allowing submitters to enable an anonymity feature, which ensures reviewers only have access to redacted versions of the experiments they're reviewing, with sensitive fields such as "Contact" removed from the meta-data. Please refer to this guide on submitter anonymity for further details.

- Submission
- Archiving
- Data mining

"PubMed for data sets"

Find us on the Wellcome Genome Campus

EMBL-EBI

wellcome trust sanger institute

Genomics england

elixir

CTTV Centre for Therapeutic Target Validation

ArrayExpress data sources

Direct submission

from researchers
(curated)

Automatic import

from NCBI Gene Expr.
Omnibus
(no curation)

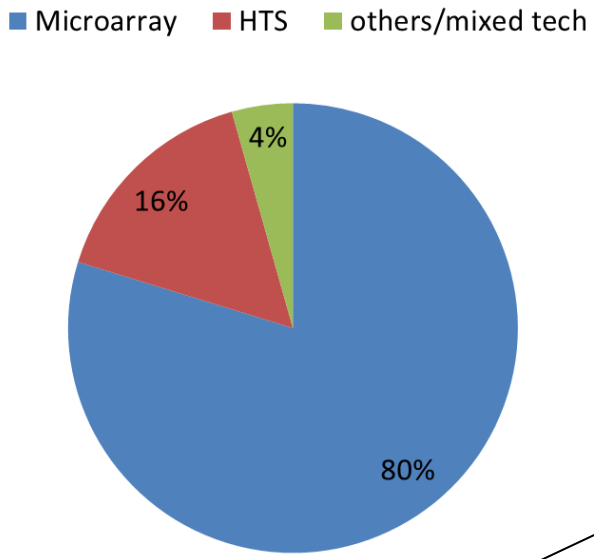
Gene Expression Omnibus

ArrayExpress

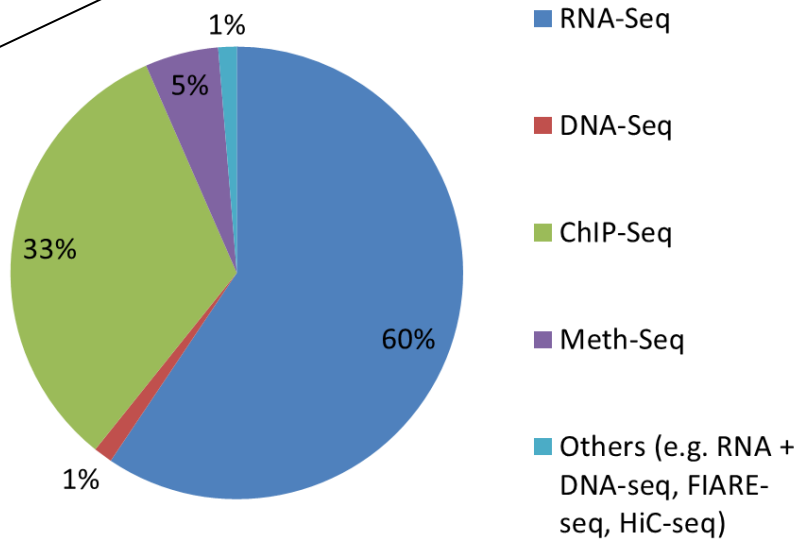
EMBL-EBI

Data volume in ArrayExpress

~61600 experiments, ~16% direct submissions, the rest imported



Microarray vs HTS



RNA-, DNA-, ChIP-seq breakdown

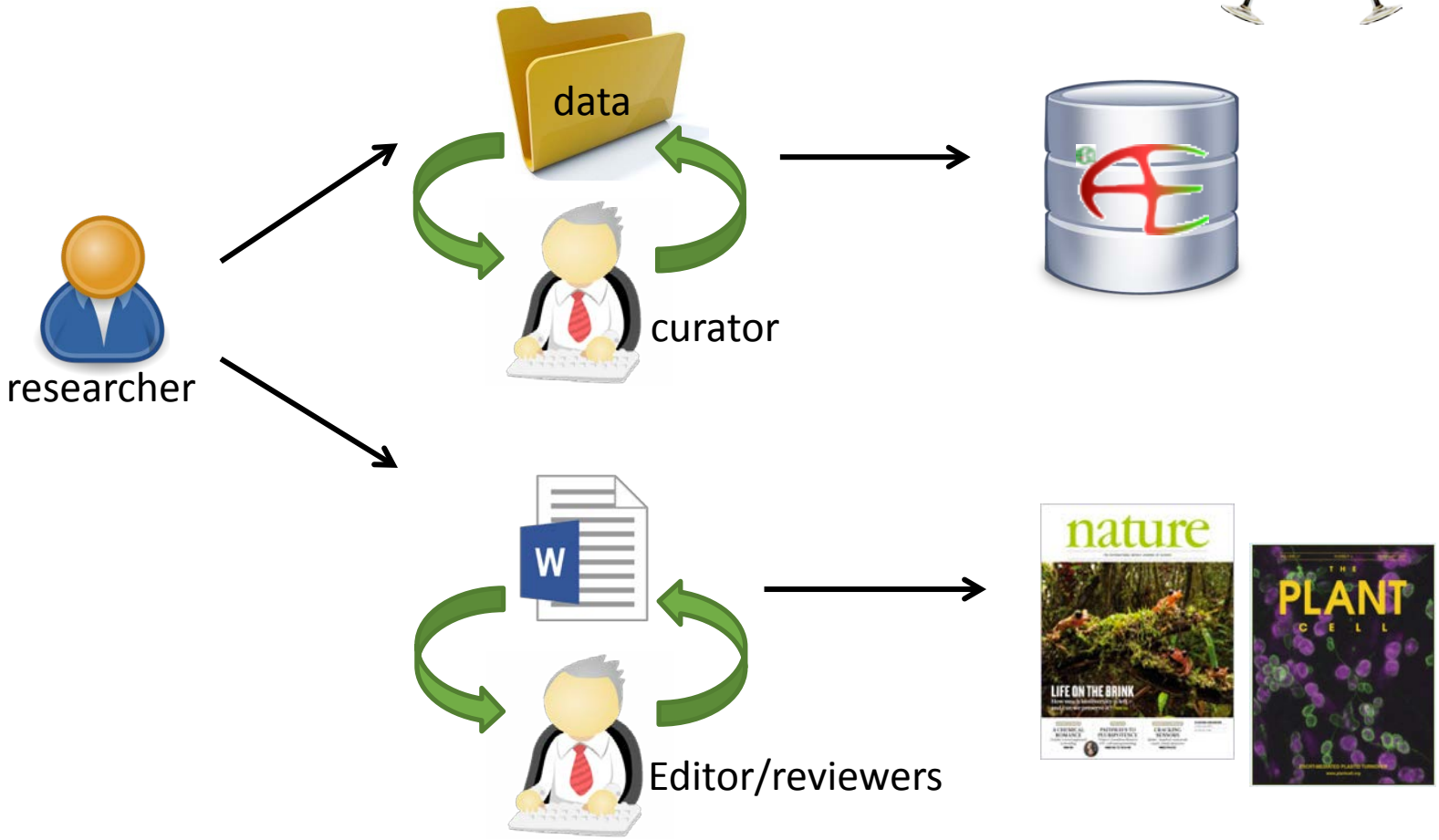
(Pie charts as of 13 Nov 2015)

EMBL-EBI

Submission timing

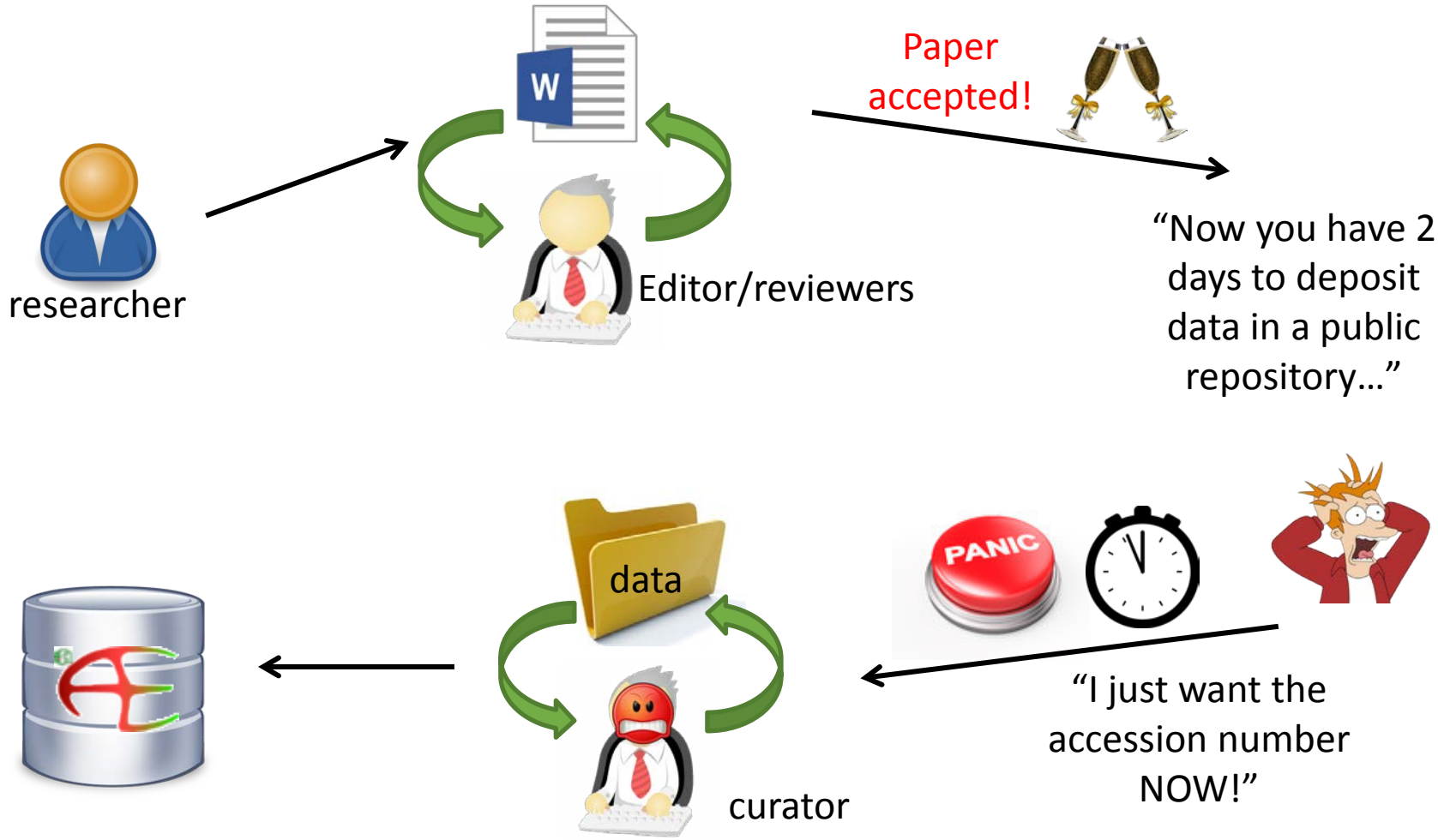
Submit to ArrayExpress – expected timing

Submission → Curation/Review → Publication



EMBL-EBI

Submit to ArrayExpress – common scenario



Submitting to ArrayExpress – when?

As explained, our paper is conditionally accepted in Science and they want the accession numbers before the proof stage. Since the data files and metadata have now been submitted could you please assign and send us an accession number?

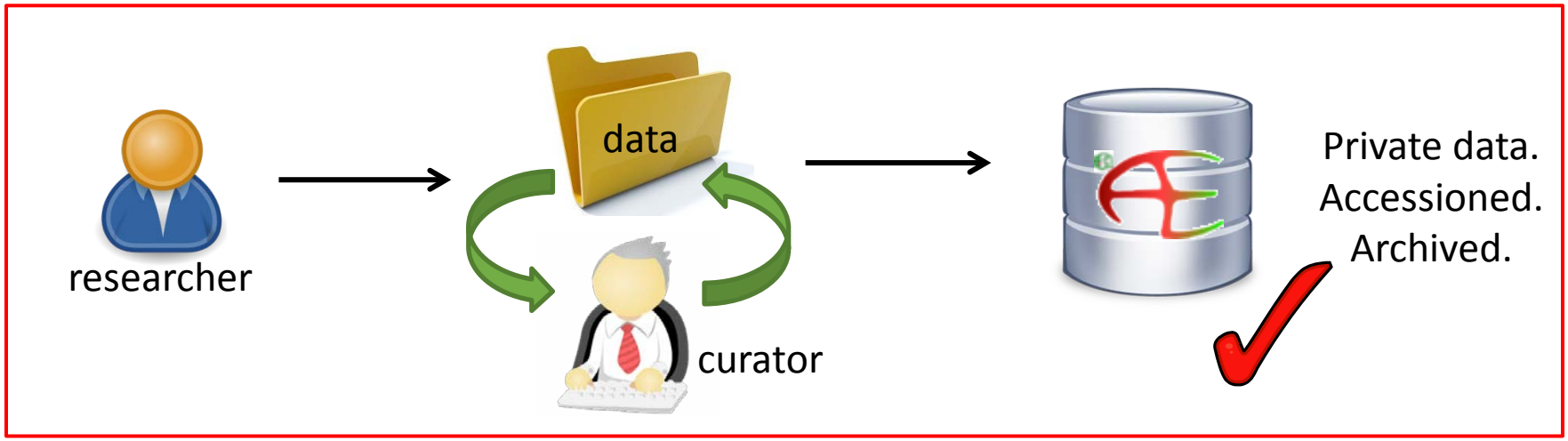
"The issue is that we need an accession number as soon as possible."

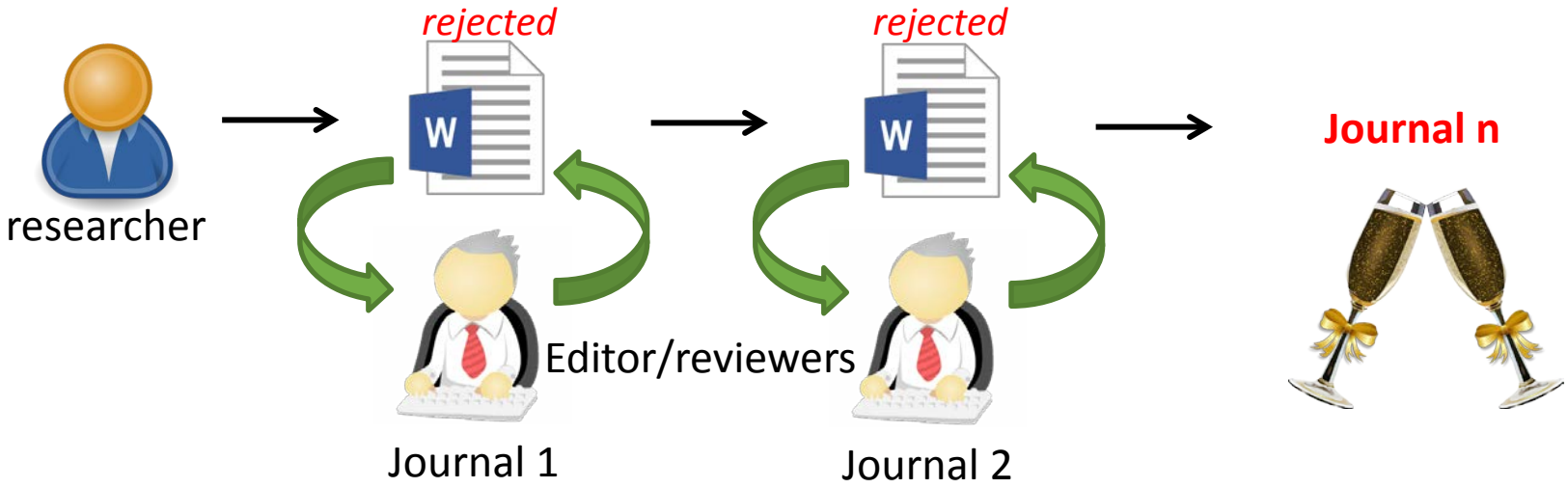
Dear Sir/Madam,

I have done a MicroArray submission today entitled: "xxxxx." I would like to request you for providing me its accession number today, as I intend to submit the concerned manuscript by today.

Many thanks for your time and considerations.

Submit to ArrayExpress – do it early, park it





Standards and ontologies

Data standards – where we're at in 2001…

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]

Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

Introduction

After genome sequencing, DNA microarray analysis[1] has become cult, because at present, microarrays do not measure gene expression levels in any objective units. In fact, most measurements report

Data standards

- MIAME = Minimal Information About a Microarray Experiment (http://www.mged.org/Workgroups/MIAME/miame_2.0.html)

- MINSEQE = Minimal Information about a high-throughput Nucleotide SEQuencing Experiment (http://www.mged.org/minseqe)

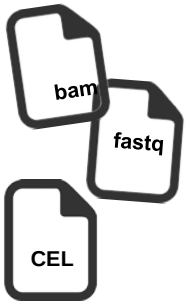
- Aim: capture information from every stage of an experiment

IDEA!

- Bio background
- Experiment aim
- Who's submitting
- Main experimental factor/variable

Sample annotation

wet/dry lab protocols



Raw data files

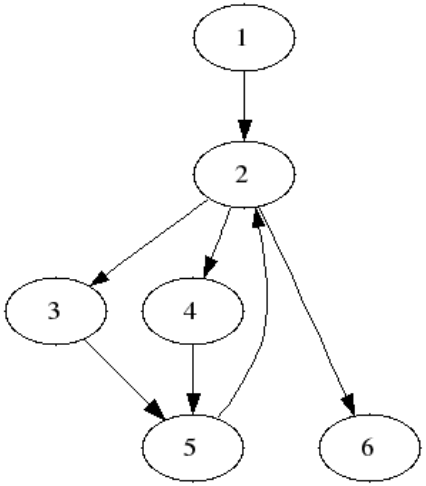
Processed files

Microarray only

Array design annotation

EMBL-EBI

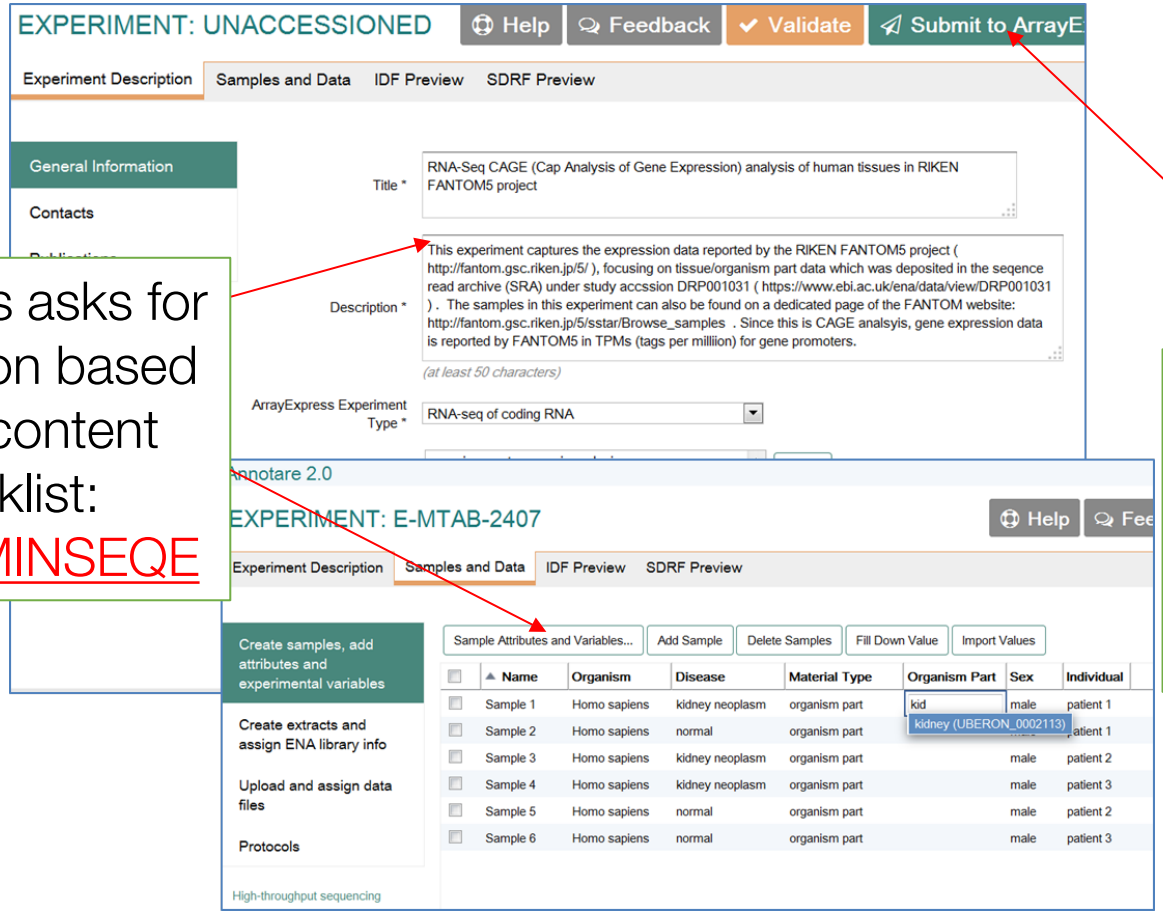
Various ways submitters tried to comply with our standards



EMBL-EBI

Fulfil standards with submission tool “Annotare”

http://www.ebi.ac.uk/fgpt/annotare_help/



Webforms asks for information based on the content checklist: MIAME/MINSEQE

Automatically generates MAGE-TAB format spreadsheet

EMBL-EBI

Sample annotation

- State the “obvious”

e.g. mouse strain, dose of drug, bio. reps vs tech reps

- Include confounding variables

e.g. sex/gender in clinical studies

- Avoid context-specific acronyms

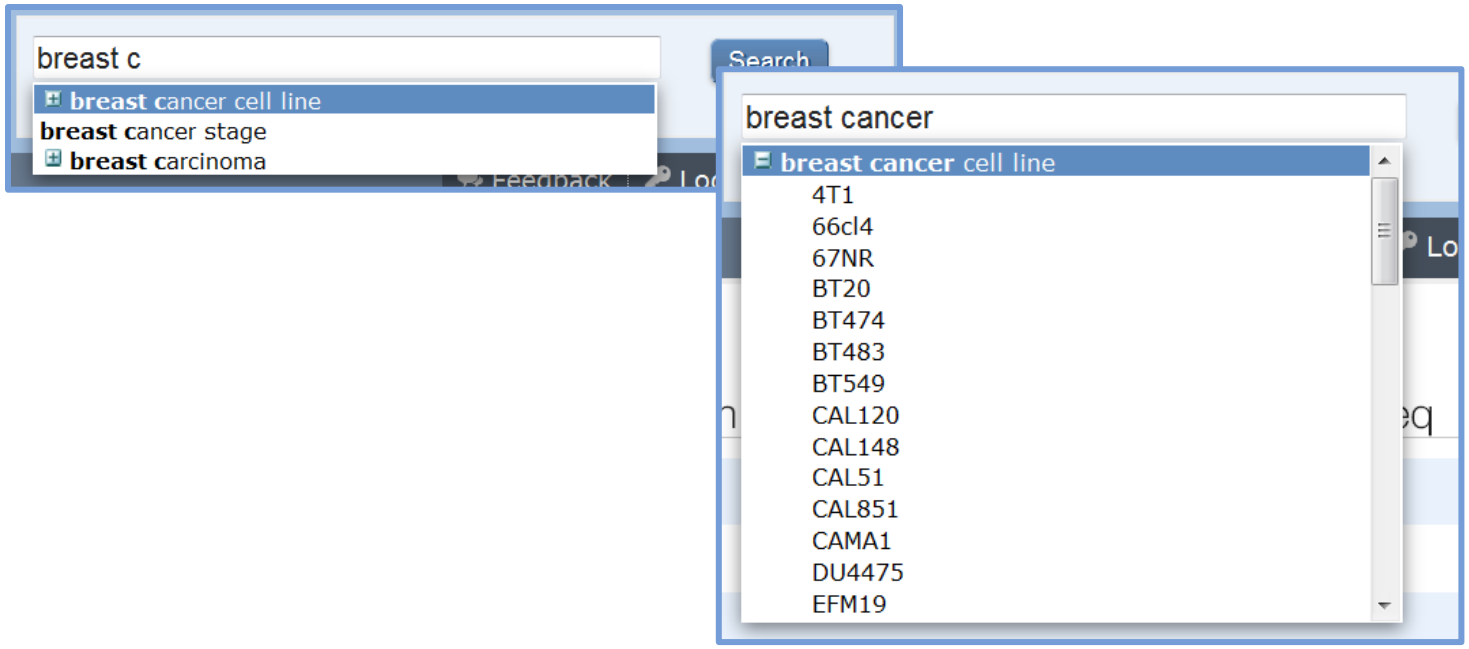
e.g. “m” =  ?  ?  ? or 

- Consult submitters to define required meta-data for emerging technologies

e.g. single-cell RNA-seq

Annotate samples with ontology terms

- Ontology = controlled vocabulary with hierarchical relationships
- We mainly use Experimental Factor ontology (EFO) Experimental Factor Ontology



- We request new terms regularly

Annotate samples with ontology terms – why?

Avoid acronyms and typographical errors

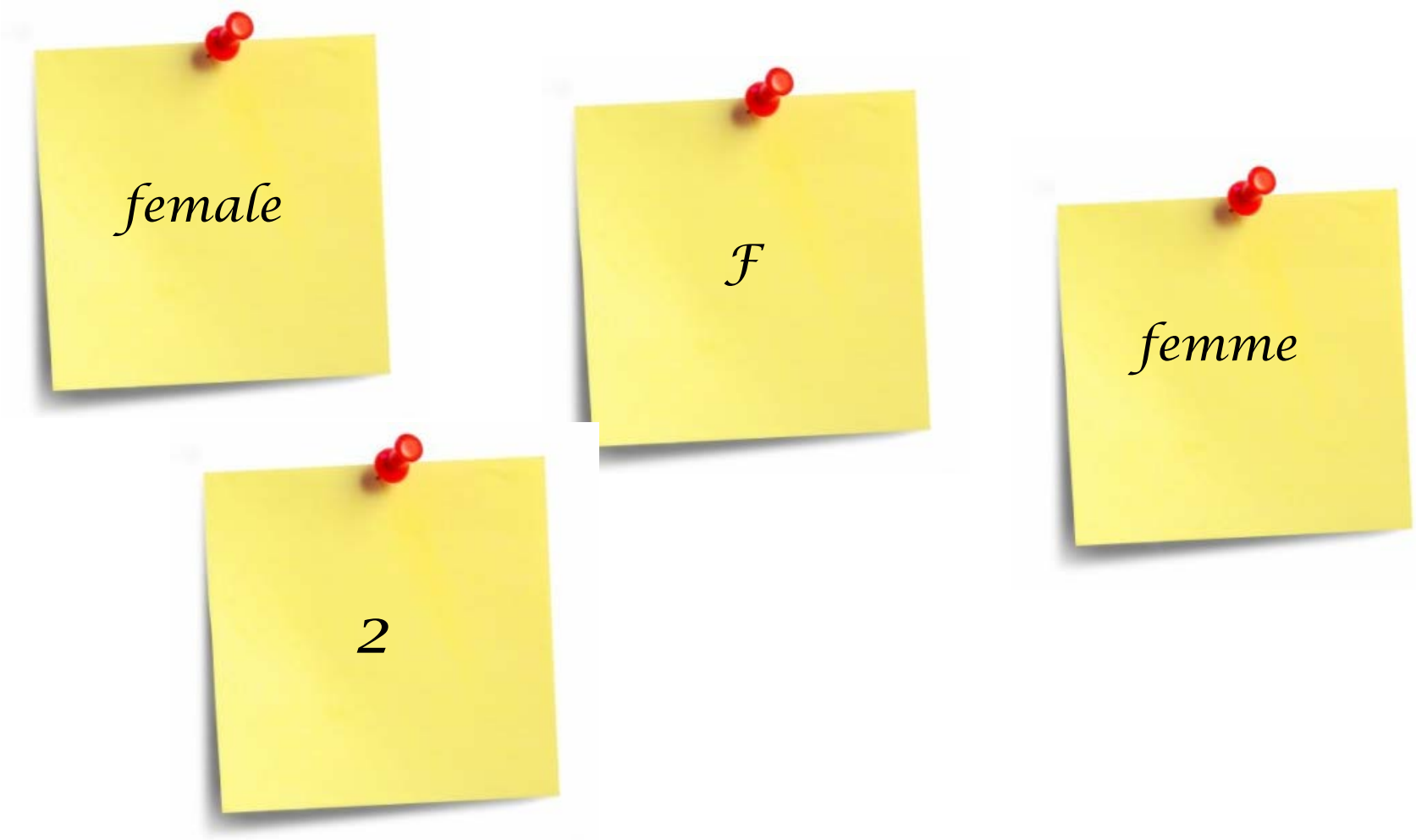
- Choose terms from dropdown, e.g. NCBI Taxonomy species

Organism
Escherichia coli O157
Escherichia coli O158
Escherichia coli O159
Escherichia coli O160

Avoid ambiguity – standardise terminology

- Developmental stage (different nomenclature systems)
- Mouse phenotype (IMPC project)
- Plant ontology (various plant “treatments”)

How hard is it to agree on how we say “female”?



EMBL-EBI

How many ways can you say “female”?

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

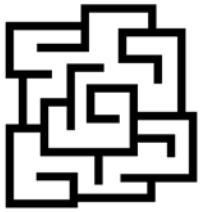
How many ways can you say “male”?

37 year old male
600 yr. old male
adult male
bull
castrated male
cm
dioecious male
diploid male
drone
engorged male
fertile male
four males mixed
individual male
initial phase male
m
make
makle
mal e
male
male (7-2872)
male (7-3074)
male (m-a)
male (m-o)
male caucasian
male child
male fertile
male fetus
male plant
male, 8 weeks old
male, castrated
male, pooled
males
man
men
normale male
ram
rooster
s1 male sterile
sex: male
six males mixed
stallion
steer
sterile male
strictly male
tetraploide male
type i males
type ii males
virgin male
winged and wingless males
young male
male (note: this sample was originally provided as a \female\ 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 male individual)

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

Protocols

- Include full protocols
  - Quoting pre-published papers – can't verify.
  - Quoting paper 1 → paper 2 → paper n : trail lost!
  - "Where is the protocol in the 50-page PDF!?"



  - Often too generic:

"Standard protocol" ;
"… carried out according to manufacturer's instructions"

Protocols

- Include supporting files, e.g.
  - Very detailed protocols (run over pages)
  - spike-in sequences in single-cell RNA-seq
  - gtf file for transcript quantification
  - scripts used to process the data
  - README file describing supporting files

Raw data – rigorous check

Correct array design assignment?

Fastq files

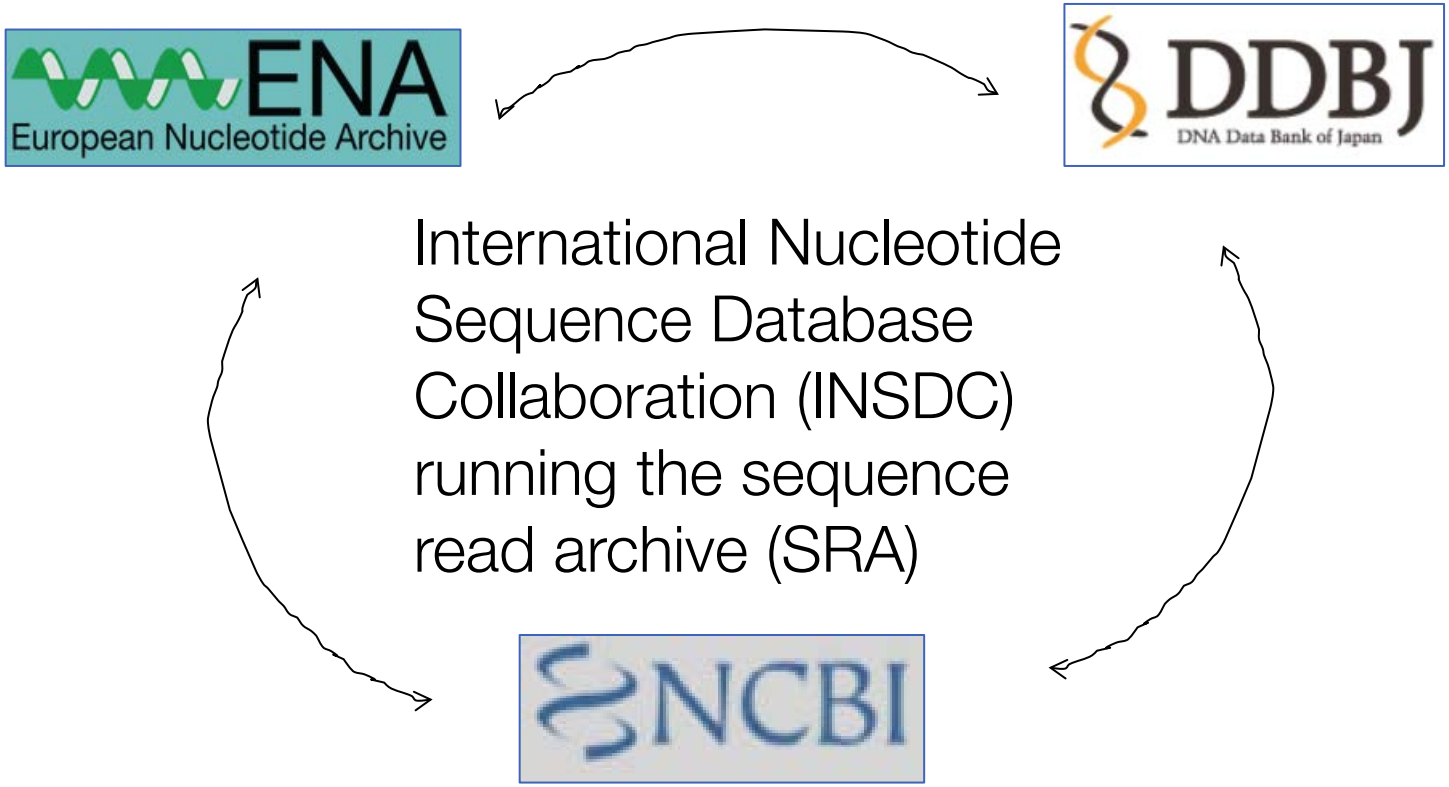
- Trimmed reads? (read length = 1bp!)
- Check mate pairs (if paired end)

Bam files

- Contain “all” reads (mapped + unmapped)

NGS raw data – stable archiving at INSDC

- Standard SRA-format fastq
- Daily data exchange/mirroring



When standards aren’t met…

The gatekeepers, aka the curators

Amy Tang

Maria Keays

Melissa Burke

Anja Fullgrabe

Laura Martinez

Satu Koskinen

Curators/Bioinformaticians

How submitters see us ....


Amy Tang


Maria Keays


Melissa Burke


Anja Fullgrabe


Laura Martinez


Satu Koskinen

Curators/Bioinformaticians

Curator: the “gap filler”

Sample	genotype	treatment
sample 1	DT	1
sample 2	DT	1
sample 3	DT	2
sample 4	DT	2
sample 5	DS	1
sample 6	DS	1
sample 7	DS	2
sample 8	DS	2

- Genotype of what gene? What is “DT”?
- What is treatment “1” ?
- Any controls?

Curator: the “cleaner”

Sample Attributes and Variables... | Add Sample | Delete Samples | Fill Down Value | Import Values

	Name	Organism	Material Type	Description
	Sample 1	Homo sapiens	cell	islet cells isolated from 23-year-old male with type I diabetes
	Sample 2	Homo sapiens	cell	islet cells isolated from 23-year-old male with type I diabetes
	Sample 3	Homo sapiens	cell	islet cells isolated from 28-year-old female with type I diabetes
	Sample 4	Homo sapiens	cell	islet cells isolated from 24-year-old female with type I diabetes
	Sample 5	Homo sapiens	cell	islet cells isolated from 21-year-old male with type I diabetes
	Sample 6	Homo sapiens	cell	islet cells isolated from 25-year-old male with type I diabetes

Sample Attributes and Variables... | Add Sample | Delete Samples | Fill Down Value | Import Values

	Name	Material Type	Organism	Disease	Cell Type	Sex	Age (year)
	Sample 1	cell	Homo sapiens	type I diabetes mellitus	islet cell	male	23
	Sample 2	cell	Homo sapiens	type I diabetes mellitus	islet cell	male	22
	Sample 3	cell	Homo sapiens	type I diabetes mellitus	islet cell	female	28
	Sample 4	cell	Homo sapiens	type I diabetes mellitus	islet cell	female	24
	Sample 5	cell	Homo sapiens	type I diabetes mellitus	islet cell	male	21
	Sample 6	cell	Homo sapiens	type I diabetes mellitus	islet cell	male	25

Curator: the “investigator”

Sample Name	Sex
Patient A	male
Patient A	female
Patient B	male
Patient B	male

- Organism: Equus caballus
- Age: 70 (years)

- Sex: male
- Organism part: endometrium

- Lung sample from one mouse
- Sex = “<u>mixed</u> sex”.

When the dust (finally) settles…

Contact info and citation

- Contact update – people move!

- Citation update

- Release policy: we crawl literature for published ArrayExpress accessions; turn private data sets to public if the paper is out.

Europe PubMed Central

EMBL-EBI

Continuing challenges

- We have no power to reject data sets

- Submitters don't treat us like journal editors/reviewers
- Deposition may be mandatory, standards aren't
- Non-obvious errors will pass without suspicion
- Secure funding to run archival databases

Faces behind ArrayExpress

Robert Petryszak

Amy Tang

Maria Keays

Melissa Burke

Anja Fullgrabe

Curators/Bio-informaticians

Satu Koskinen

Laura Martinez

Ugis Sarkans

Ahmed Ali

Nikolay Kolesnikov (“Mr Annotare”)

Catherine Snow

Miroslaw Dylag

Data management

web