BioStudies database

Ugis Sarkans

17/07/2019
Overview

- Introductory thoughts
- What is BioStudies
- Data access
- Data submissions
- Advanced use and future
Quick introduction about myself

- Computer science background (U of Latvia)
Quick introduction about myself

• Computer science background (U of Latvia)
• At EBI since 2000
• Run the Functional Genomics Software Development team
• Develop and run the BioStudies database
• Maintain software for ArrayExpress, and support ArrayExpress users (data access)
• Participate in building BioImage Archive (in collaboration with other teams)
<table>
<thead>
<tr>
<th>Team</th>
<th>ArrayExpress</th>
<th>BioStudies</th>
<th>Imaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Interface &amp; APIs</td>
<td><img src="image1" alt="User Interface &amp; APIs" /></td>
<td><img src="image2" alt="User Interface &amp; APIs" /></td>
<td><img src="image3" alt="User Interface &amp; APIs" /></td>
</tr>
<tr>
<td>Backend</td>
<td><img src="image4" alt="Backend" /></td>
<td><img src="image5" alt="Backend" /></td>
<td><img src="image6" alt="Backend" /></td>
</tr>
<tr>
<td>Submission tool</td>
<td><img src="image7" alt="Submission tool" /></td>
<td><img src="image8" alt="Submission tool" /></td>
<td><img src="image9" alt="Submission tool" /></td>
</tr>
<tr>
<td>Data</td>
<td><img src="image10" alt="Data" /></td>
<td><img src="image11" alt="Data" /></td>
<td><img src="image12" alt="Data" /></td>
</tr>
</tbody>
</table>
Good practices for research data management

• Describe experiment, samples, protocols, data files, analysis methods, …
• Adopt checklists, follow standards
• Adopt ontologies and “structured descriptions”
• Provide data in both human AND machine-readable form, use standard formats
• Submit data to appropriate databases
• QC data
• …
Good practices for research data management

Train online

Bringing data to life: data management for the biomolecular sciences

Managing and making the most of your data
Why share your data?
What happens to your data?
Giving data context, structure and meaning
Tools for data management planning
Myths and best practice
Contributors

Bringing data to life: data management for the biomolecular sciences

Webinar series 28 June – 2 August 2018

EMBL-EBI
Good practices for research data management

- Describe experiment, samples, protocols, data files, analysis methods, ...
- Adopt checklists, follow standards
- Adopt ontologies and “structured descriptions”
- Provide data in both human AND machine-readable form, use standard formats
- Submit data to appropriate databases
- QC data
- ...
- Result – more databases used, more data associated to a paper – difficult to track
Example

- Let’s look at this paper

  - [https://doi.org/10.1128%2FAEM.00675-10](https://doi.org/10.1128%2FAEM.00675-10)
Example

- Let’s look at this paper
  - https://doi.org/10.1128%2FAEM.00675-10

- What data are associated to this paper?
  - what data are in various community databases? which databases?
  - are all associations of the same nature?
  - what other data are attached to the paper?
Example - discussion

• “The nucleotide sequences … have been deposited in the GenBank database under accession numbers HQ438317 to HQ438435”

• “The microarray data reported here have been deposited in the Gene Expression Omnibus (GMO) database under accession number E-MEXP-2951”

• “Primers … were optimized or redesigned on the basis of the known gene sequences of B. longum subsp. infantis ATCC 15697 and B. longum subsp. longum DJO10A (GenBank accession number CP000605)”
**Supplemental material**

**Files in this Data Supplement:**

**Supplemental file 1** - Supplemental text and figure legends, MLST genes and primers (Table S1), descriptive evolutionary analysis of MLST data (Table S2), allelic profiles of 15 *B. longum* strains analyzed by MLST (Table S3), mobile element locations, sizes, and distributions (Table S4), GSEA analysis of COGs overrepresented in *B. longum* subsp. *infantis* (Table S5).

PDF file, 291K.

**Supplemental file 2** - Location of MLST loci on the *B. longum* subsp. *infantis* ATCC 15697 genome (Fig. S1), determination of thresholds for absent, divergent, partially conserved, and conserved CDS (Fig. S2), relationship between log ratio and sequence identity (Fig. S3), split decomposition analyses of 17 *B. longum* strains based on concatenated sequences of seven loci (Fig. S4) and of individual loci (Fig. S5), and log ratio distribution of JCM7007/ATCC 15697 hybridizations (Fig. S6).

PDF file, 2 MB.

**Supplemental file 3** - CGH analyses of genes unique to ATCC 15697 (group 1; Fig. S7) or *B. longum* subsp. *infantis* (group 2; Fig. S8), of genes common to *B. longum* subspecies (group 3; Fig. S9), of *B. longum* subspecies fucosidases (Fig. S10), or of type 1 SBPs (Fig. S12), loci that are divergent in JCM7010 and conserved in other *B. longum* subsp. *infantis* strains (Fig. S11), and alignment of the HMO cluster on JCM1260 (Fig. S13).

PDF file, 24.5 MB.
BioStudies database: aim

Facilitate transparency and reproducibility of research, by aggregating all the data associated to a study, the data package, in a single place.

Latest publication:
Broad conservation of milk utilization genes in Bifidobacterium longum subsp. infantis as revealed by comparative genomic hybridization.

(PMID:20802066 PMCID:PMC2976205)

Data behind the paper

Find links to the data behind this article and supplemental files at BioStudies:
http://www.ebi.ac.uk/biostudies/studies/S-EPMC2976205/xr=true

BioStudies.

Functional Genomics Experiments in ArrayExpress (1)

Data cited in the article (1)

ArrayExpress:E-MEXP-2951 (1 citation)
Broad conservation of milk utilization genes in Bifidobacterium longum subsp. infantis as revealed by comparative genomic hybridization.

LoCasile RG, Desai P, Sela DA, Wilmer B, Mills DA.

1 Department of Viticulture and Enology, University of California, Davis, CA 95616, USA. 2 Department of Viticulture and Enology, Robert Mondavi Center for Wine and Food Science, Microbiology Graduate Group, University of California, Davis, California 95616, Department of Nutrition, Dietetics, and Food Sciences, Utah State University, Logan, Utah, 84322, Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, California 4

Accession
S-EPMC2976205

Abstract

Human milk oligosaccharides (HMOs) are the third-largest solid component of milk. Their structure renders them nondigestible to the host but liable to hydrolytic enzymes of the infant colonic microbiota. Bifidobacteria and, frequently, Bifidobacterium longum strains predominate the colonic microbiota of breast-fed infants. Among the three recognized subspecies of B. longum, B. longum subsp. infantis has higher levels of cell growth on HMOs and is associated with early colonization of the infant gut. The B. infantis ATCC 15697 genome features five distinct gene clusters with the predicted capacity to import milk oligosaccharides. Comparative genomic hybridizations (CGHs) were used to assess biomarkers among 15 B. longum strains exhibiting various HMO utilization phenotypes and host-Multilocus sequence typing provided taxonomic subspecies designations and grouped the strains between B. longum subsp. infantis and B. longum subsp. longum. CGH analysis determined that HMO utilization gene regions are exclusively conserved across all B. longum subsp. infantis strains capable of growth on HMOs and have diverged in B. longum subsp. longum strains that cannot grow on HMOs. These regions contain fucosidases, sialidases, glycosyl hydrolases, ABC transporters, and family 1 solute binding proteins and are likely needed for efficient metabolism of HMOs. Urea metabolism genes and their activity were exclusively conserved in B. longum subsp. infantis strains.
BioStudies data sources across the research life cycle

Pre-publication

- Europe PMC ingest
- Publisher collaboration (e.g., SourceData)

Post-publication

- Publisher system → BioStudies (e.g. JCB image viewer)
- Community DB → BioStudies (e.g. ArrayExpress)
Example - discussion

• Issues with automated data package generation
  • BioStudies entry automatically created – did not distinguish between data *generated* as a part of this study, and data *used* in this study as a reference
  • “deposited in the GenBank database under accession numbers HQ438317 *to* HQ438435” – a range of entries, not just 2 of them
• ...
Optimally data packages are created **before** publication

- Links from paper to the data package can be created
- Opportunity to have better structured supplementary materials/data
BioStudies data sources across the research life cycle

Pre-publication
- manual BioStudies deposition: links + orphan data
- data w/o community repository (e.g., imaging)
- BioStudies ingest pipeline for multi-omics data (EBI-wide submissions, ArrayExpress)
- collaboration with on-going projects: facilitating data sharing and publishing

Post-publication
- Europe PMC ingest
- Publisher system → BioStudies (e.g. JCB image viewer)
- Publisher collaboration (e.g., SourceData)
- Community DB → BioStudies (e.g. ArrayExpress)
Data access
Browse view – facets

S-BSST244  •  22 March 2019  •  1 link  •  6 files
Supplementary Dataset for "Pore-scale hydrodynamics influence the spatial evolution of bacterial biofilms in a microfluidic porous network"

S-BSST245  •  21 March 2019  •  1 link  •  1 file
Metabolomics-driven exploration of the chemical drug space to predict combination antimicrobial therapies.

S-SUBS3  •  14 March 2019
Ischaemic sensitivity of human tissue by single cell RNA seq

S-BSST243  •  12 March 2019  •  4 links  •  2 files
Genome-wide CRISPR-Cas screen identifies CRNKL-1 as a nuclear retention factor of HIV-1 unspliced RNA

S-BSST241  •  7 March 2019  •  1 link  •  5 files
Have the cake and eat it: Optimising non-destructive DNA metabarcoding of macroinvertebrate samples for freshwater biomonitoring

S-BSST240  •  6 March 2019  •  1 link  •  3 files
PyBioNetFit and the Biological Property Specification Language

S-BSST242  •  5 March 2019  •  1 link  •  1 file
Duration of protection from live attenuated vs. sub unit HSV-2 vaccines in the guinea pig model of genital herpes: reassessing efficacy using endpoints from clinical trials
A chemical-genetic interaction map of small molecules using high-throughput imaging in cancer cells

Marco Breitig 1, Felix A. Klein 2, Wolfgang Huber 3, Michael Boutros 1.

1 German Cancer Research Center (DKFZ), Division of Signalling and Functional Genomics and Heidelberg University, Dept. Cell and Molecular Biology, Heidelberg, Germany
2 European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany

Accession
S-BSMS-PGPC1

Abstract
Small molecules often affect multiple targets, elicit off-target effects and induce genotype-specific responses. Chemical genetics, the mapping of the genotype-dependence of a small molecule's effects across a broad spectrum of phenotypes, can identify novel mechanisms of action. It can also reveal unexpected effects and could thereby reduce high attrition rates of small molecule development pipelines. Here, we used high-content screening and image analysis to measure effects of 1280 pharmacologically active compounds on complex phenotypes in isogenic cancer cell lines which harbor activating or inactivating mutations in key oncogenic signaling pathways. Using multiparametric chemical-genetic interaction analysis, we observed phenotypic gene–drug interactions for more than 183 compounds, with many affecting phenotypes other than cell growth. We created a resource termed the Pharmacogenetic Phenome Compendium (PGPC), which enables exploration of drug mode-of-action, detection of potential off-target effects and the generation of hypotheses on drug combinations and synergism. For example, we demonstrate that MEK inhibitors amplify the viability effects and show that the EGFR inhibitor tyrophostin AG556 has off-target effects. This demonstrates how combining multiparametric phenotyping with additional mechanistic data can lead to repositioning opportunities.

Data availability
Complementary views on this dataset are available through different repositories. The image data files are available as a spreadsheet describing individual files.
Example – a SourceData dataset in BioStudies

cohesin

Dr. Ireneusz, Bernard Lithman 1, Mr. Tomasz Bajekowski 1, Barnabas Szakal 2, Ms. Ewa Pilarsczyk 3, Dr. Ewa Maciaszczyk-Dziubinska 1, Dr. Dana Branzei 2, Prof. Robert Wysocz 1

1 University of Wroclaw, 2 IFOM (PIRC Institute of Molecular Oncology)

Accession
S-SCDT-EMBOJ-2017-98732

Abstract
DNA damage tolerance (DDT) mechanisms facilitate replication resumption and completion when DNA replication is blocked by bulky DNA lesions. In budding yeast, Template switching (TS) via the Rad18/Rad5 pathway is a favored DDT pathway that involves usage of the sister chromatid as a template to bypass DNA lesions in an error-free recombination-like process. Here, we establish that the Sfn2 family translocase Irc5 is a novel factor that promotes TS and averts single-stranded DNA persistence during replication. We demonstrate that during replication stress, Irc5 enables replication progression by assisting enrichment of cohesin complexes, recruited in an Scx2/Scx4-dependent fashion, near blocked replication forks. This allows efficient formation of sister chromatid junctions that are crucial for error-free DNA lesion bypass. Our results support the notion of a key role of cohesin in the completion of DNA synthesis under replication stress, and reveal that the Rad18/Rad5-mediated DDT pathway is linked to cohesin enrichment at sites of perturbed replication via the SNF2 family translocase, Irc5.

Publication
The EMBO Journal doi: 10.15252/emboj.201798732

Funding
National Science Center (Poland): 2013/11/D/NZ2/02696
Italian Association for Cancer Research: IG 18976
European Research Council Consolidator Grant: 682190

Figure 1

<table>
<thead>
<tr>
<th>Figure 1-A</th>
<th>SOURCEDATA SmartFigure</th>
</tr>
</thead>
</table>

Table: Entities

<table>
<thead>
<tr>
<th>text</th>
<th>role</th>
<th>type</th>
<th>tax_id</th>
<th>external_id</th>
<th>external_name</th>
<th>database_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL612</td>
<td>component</td>
<td>cell</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>irs5</td>
<td>intervention</td>
<td>gene</td>
<td>599292</td>
<td>858699</td>
<td>IRC5</td>
<td>NCBI gene</td>
</tr>
<tr>
<td>MMS</td>
<td>intervention</td>
<td>molecule</td>
<td>None</td>
<td>CHEBI:25265</td>
<td>methyl methanesulfonate</td>
<td>ChEBI</td>
</tr>
</tbody>
</table>
Search results for **gut**

1 – 20 of 14,620 results

- **S-EPMC2126836** - 1 January 1996 - 10 links
  - *eat-5 and unc-7 represent a multigene family in Caenorhabditis elegans involved in cell-cell coupling.*
  - ... are unsynchronized. Injection of carboxyfluorescein into muscles of the posterior pharynx demonstrates that all pharyngeal muscles are dye-coupled in wild-type animals; in *eat-5* mutants, however, muscles of the anterior pharynx are no longer dye-coupled to posterior pharyngeal muscles. We show...

- **S-EPMC3485723** - 1 January 2012 - 16 links - 2 files
  - *Changes in the gut microbiome of the sea lamprey during metamorphosis.*
  - ...-S-EPMC3485723 PMC3485723 Changes in the gut microbiome of the sea lamprey during metamorphosis. Vertebrate metamorphosis is often marked by dramatic morphological and physiological changes of the alimentary tract, along with major shifts in diet following development from larva to adult. Little...

- **S-EPMC3339417** - 1 January 2012 - 3 links - 2 files
  - *Exceptionally preserved Cambrian trilobite digestive system revealed in 3D by synchrotron-radiation X-ray tomographic microscopy.*
  - ... Synchrotron-radiation X-ray tomographic microscopy enabled three-dimensional internal recordings at sub-micrometre resolution. The specimen provides the first unambiguous evidence for a J-shaped anterior gut and the presence of a crop with a constricted alimentary tract in the Trilobite. Moreover...
How does the system know about synonyms and other relations?

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.

EFO (version 3) is now available - please see here for more info

Browse

Browse EFO with EBI's OLS or The NCBO BioPortal (external). You can also search EFO using the search box, above. The EFO source ontology can be viewed on GitHub at https://github.com/EBISPOT/efo.

Submit

Submit new terms or report bugs using our GitHub issue tracker or if you prefer to e-mail us please contact efo-users [at] ebi.ac.uk. For EFO announcements and monthly updates please join the efo-users public mailing list efo-users public mailing list.

Download

Download the latest release of EFO in OWL format. There is an OBO format version and an inferred OWL view. Read the latest Release Notes. You can also get previous releases by release day e.g. http://www.ebi.ac.uk/efo/releases/v2017-10-16/efo.owl
Queries – combining search terms

Use AND, OR, NOT, and brackets

Ontology expansion still works
Queries – brackets

“Homo sapiens” instead of human can be used

OR can be skipped
Other project-specific facets

1 - 20 of 1,201,549 results

Sort by: Released

- S-EPMC65302955 · 1 January 2019 · 7 links · 1 file
  The Human Mesenteric Lymph Node Microbiome Differentiates Between Crohn's Disease and Ulcerative Colitis.

- S-EPMC65303422 · 1 January 2019 · 5 links · 1 file
  Safety and efficacy of denatuxizumab mafodotin + temozolomide in patients with EGFR-amplified, recurrent glioblastoma: results from an international phase I multicenter trial.

- S-EPMC6530252 · 1 January 2019 · 3 links · 1 file
  Optical coherence tomography-verified morphological correlates of high-intensity coronary plaques on non-contrast T1-weighted magnetic resonance imaging in patients with stable coronary artery disease.

- S-EPMC65302953 · 1 January 2019 · 3 links · 1 file
  Incidence of Arthritis/Arthralgia in Inflammatory Bowel Disease with Long-term Vedolizumab Treatment: Post-Hoc Analyses of the GEMINI Trials.

- S-EPMC65303128 · 1 January 2019 · 4 links
  No Difference between Spray Dried Milk and Native Whey Supplementation with Strength Training.

OR can be skipped

"Homo sapiens" instead of human can be used
Data submissions
Data submission tool

- Enables flexible annotation of a dataset, associated files and links
- Provides both a web form and tab-delimited file upload functionality
- Provides configurable templates for different domains and projects
Register

Back to Log in

Registration

Please provide the information below to create a new account. Note that both the Full name and ORCID fields are optional.

Full name

Email address
manager

Password
******

ORCID (use 4x4 digit format)
e.g. 1892-5647-2571-9436

Verify you are a human

☑️ I'm not a robot

I have read and agree to the Privacy Notice and Terms of Use, including the limited processing of personal data.

REGISTER
<table>
<thead>
<tr>
<th>Accession</th>
<th>Version</th>
<th>Latest Title</th>
<th>First Released</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-BSST503</td>
<td>2</td>
<td>dgigcf Godzilla dsfg sdgh sdfg sdfg fg change</td>
<td>24 Apr 2018</td>
<td></td>
</tr>
<tr>
<td>S-BSST357</td>
<td>2</td>
<td>Fixed test as it was long and bad, please check this</td>
<td>08 Mar 2018</td>
<td></td>
</tr>
<tr>
<td>S-BSST364</td>
<td>1</td>
<td>a temporally submission - by 11:52 8 march of 2018</td>
<td>08 Mar 2018</td>
<td></td>
</tr>
<tr>
<td>S-BSST161</td>
<td>2</td>
<td>test - 4drdrdrd rrrdrdrdrdr rrdrdrdrdrd rgytrgyrtrgyrtygyrnuhu</td>
<td>27 Feb 2018</td>
<td></td>
</tr>
<tr>
<td>S-BSST351</td>
<td>1</td>
<td>title is 50 characters title is 50 characters title</td>
<td>26 Feb 2018</td>
<td></td>
</tr>
<tr>
<td>S-BSST644</td>
<td>1</td>
<td>QSAR for predicting up regulation of PXR nuclear receptor as MIE of steatosis AOP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-BSST47</td>
<td>1</td>
<td>S-BSST47 13-Jun-2018 16:13:16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-BSST405</td>
<td>1</td>
<td>Study Title</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
List of my data files

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Progress</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data storage V6_AK_UKN4</td>
<td>xlsx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UKN3a raw new.xlsx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data storage V6_AK_UKN4_b</td>
<td>xlsx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data storage V6_AK_UKN4_c</td>
<td>xlsx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data storage V6_AK_UKN4_d</td>
<td>xlsx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data storage V6_AK_UKN4_e</td>
<td>xlsx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data storage V6_AK_UKN4_f</td>
<td>xlsx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>valid.xlsx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing-variable-column.xlsx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-MEXP-700 asf_xts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>valid.xls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell_assay_data_form_SH_ATP_120p_BIOT.xlsx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sero</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPS1a FOC-BOA_FBA and 2nd CSS.xlsx</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
List of my data files

FTP Upload

Your secret directory is: ...

The FTP server is ftp-private.ebi.ac.uk, FTP user is bsftp, and the password is bsftp1. You can use any FTP client such as FileZilla, or issue FTP commands via a command line, e.g.,

```
ftp ftp-private.ebi.ac.uk
Name: bsftp
Password: bsftp1
ftp> cd xx/xxxxxx-xxxxxx-xxxxxx-xxxxxx-xxxxxx
ftp> put testfile
```

The Aspera ascp command line client is distributed as part of the 'Aspera Connect' high-performance transfer browser plug-in which can be downloaded from http://www.asperasoft.com/downloads/connect. Your command for uploading a directory can be:

```
ascp -P33001 -i <path to asperaweb_id_dsa.openssh file> -d <local directory to upload> bsaspera@fasp-beta.ebi.ac.uk:dropbox/<secret directory>
```

where

- `-P33001` and `bsaspera@fasp-beta.ebi.ac.uk` defines port, user and server for Aspera connection
- `-i` tells the server to use ssh for Aspera connections
- `asperaweb_id_dsa.openssh` is the public key for Aspera connection. `<path to asperaweb_id_dsa.openssh file>` might be:
  1. on Linux: `<aspera connect installation directory>/etc/asperaweb_id_dsa.openssh`
  2. on Mac OS X: `<aspera connect installation directory>/asperaweb_id_dsa.openssh`
  3. on Windows: "%UserProfile%\AppData\Local\Programs\Aspera\Aspera Connect\etc\asperaweb_id_dsa.openssh"
- `<secret directory>` looks like xx/xxxxxx-xxxxxx-xxxxxx-xxxxxx-xxxxxx
Submitting a single study

New submission
Please fill in the form below. The Check tab on the left-hand side lists those fields still incomplete or incorrect. Use the Add tab to quickly add new rows or tables.

Please note: All fields are required unless with a dashed outline or marked as "Optional".

Title (at least 25 characters long)

Release date
Choose a date

Description (at least 50 characters long)

Annotations
Provide any additional details that may help discover or interpret the study.
Available keys: Organism, Experimental design, Experimental factor, Organ, Cell type.

Key
Value

Add Annotations row
Submitting a single study – scrolling down..

**Contacts (at least one is required)**

Add the contact details for the authors involved in the study.

*More columns: Role, Address, Department, Funding.*

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>E-mail</th>
<th>Organisation</th>
<th>ORCID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td><a href="mailto:uqit.sarkans+57@gmail.com">uqit.sarkans+57@gmail.com</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Add Contacts row**

**Links**

Provide pointers to data held in external databases or to related information on the web. Compact URIs from identifiers.org are supported. URLs must include the scheme, e.g. "http://".

<table>
<thead>
<tr>
<th>Column</th>
<th>Pointer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>URL or preXID</td>
<td></td>
</tr>
</tbody>
</table>

**Add Links row**

**Files**

List the data files for the study and describe their respective scopes.

<table>
<thead>
<tr>
<th>Column</th>
<th>Path</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Add Files row**

- Home
  - Data storage V6_AK_UKN4.xlsx
  - UKN3a_raw_new.xlsx
  - Data storage V6_AK_UKN4_b.xlsx

[SUBMIT]
Using identifiers.org for link entry

**Links**
Provide pointers to data held in external databases or to related information on the web. Compact URLs from identifiers.org are supported. URLs must include the scheme, e.g., "http://".

<table>
<thead>
<tr>
<th>Column</th>
<th>Pointer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>anonymous.E-MTAB-1</td>
<td></td>
</tr>
</tbody>
</table>

**Files**
List the data files for the study and describe their respective scopes.

<table>
<thead>
<tr>
<th>Column</th>
<th>Path</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please provide a valid value
What is identifiers.org?

Identifiers.org is an established resolving system that enables the referencing of data for the scientific community, with a current focus on the Life Sciences domain. It handles persistent identifiers in the form of URLs and CURIES. This allows the referencing of data in both a location-independent and resource-dependent manner. The provision of resolvable identifiers (URLs) fits well with the Semantic Web vision, and the Linked Data initiative.

Resolve Compact Identifiers (prefix:identifier), eg: CHEBI:36927

Enter a prefix:identifier and press Enter
“Direct” submissions

- Describe your study in a tab-delimited (e.g., Excel) file
- Useful for deposition of many similarly structured datasets (at once, or over time)
“Direct” submissions

Format description available from https://www.ebi.ac.uk/biostudies/submit

- Description
- Data file
- Links

This layout is useful if there are many files to describe. Alternatively, use a layout similar to that of the Study block, where a single block describes a single file:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>155</td>
<td>File</td>
<td>lib.txt</td>
</tr>
<tr>
<td>156</td>
<td>Type</td>
<td>library file</td>
</tr>
<tr>
<td>158</td>
<td>Format</td>
<td>tab-delimited text</td>
</tr>
<tr>
<td>159</td>
<td>Description</td>
<td>Bioneeer haploid deletion library v.2 modified</td>
</tr>
</tbody>
</table>

Use the Links section if you want to include and describe arbitrary hyperlinks. Similarly as for Files, use a horizontal or a vertical layout. This illustrates the vertical layout:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>165</td>
<td>Link</td>
</tr>
<tr>
<td>166</td>
<td><a href="http://idr-demo.openmicroscopy.org/webclient/?show=screen-3">http://idr-demo.openmicroscopy.org/webclient/?show=screen-3</a></td>
</tr>
<tr>
<td>167</td>
<td>Type</td>
</tr>
<tr>
<td>168</td>
<td>Raw data</td>
</tr>
<tr>
<td>169</td>
<td>Data format</td>
</tr>
<tr>
<td>170</td>
<td>Evotec/PerkinElmer Opera Flex</td>
</tr>
</tbody>
</table>

There is an alternative way to define Links that should be used when the Study refers to records in bioinformatics resources that the BioStudies database knows about - see <here> for a list of types of identifiers that can be used. The user
Advanced use & future

• Data sharing in a project
• ArrayExpress & BioStudies
• Imaging data & BioStudies
Data sharing in a project

• Use BioStudies for sharing & easy data publishing in future
  • Set up a new project
  • Design useful data & metadata formats
  • Manage access rights

• *(Currently cannot be done without involving the BioStudies team)*
Designing your own data sharing format

- Attributes - which attributes will help users looking for a dataset like yours to find and interpret it?
- Files - which attributes need to be attached to individual files so that meaningful downstream processing could happen?
- Links - is there a need to link to existing data in other databases?
- Sections - is the study that you are describing very complex, and/or generates many data files, so that some hierarchical organization would be beneficial for clarity?
BioStudies access rights for use in an ongoing project

- **Upload new dataset**
- **Edit/delete own dataset**
- **Read all datasets**
- **Read public datasets**
- **“read-only” project account**
- **Individual accounts**
- **BioStudies – project data**
- **everybody**
ArrayExpress migration to BioStudies – coming soon:

Current display on ArrayExpress

E-MTAB-5324 - Expression analysis of Pea3 overexpression in SH-SY5Y cell line

- Source Name
- Attributes: genotype, phenotype
- Variables: genotype, phenotype
- Links to Data: Processed

Cell line: SH-SY5Y
Cell type: neural cell
Genotype: pCDNA3 empty vector, pCDNA3-mPea3-VP16 vector

ORCID: Data claiming
You can sign in to ORCID to claim your data
Remember me on this computer

Similar Studies:
- Transcription profiling of effected genes by PAD2 and compound AB9 [E-MTAB-408]
- Study of SPEN overexpression in T47D cells [E-MTAB-312]
ArrayExpress migration to BioStudies – coming soon:

Experimental Designs
replicate design

Experimental Factors
genotype, phenotype

MIAME Score
Platforms Protocols Variables Processed Raw
* * * * * -

Experiment Protocols
show table

Source Characteristics
hide table
Table: source_chars

<table>
<thead>
<tr>
<th>genotype</th>
<th>phenotype</th>
<th>No. of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCDNA3 empty vector</td>
<td>wild type</td>
<td>3</td>
</tr>
<tr>
<td>pCDNA3-mPea3-VP16 vector</td>
<td>Pea3 overexpression</td>
<td>3</td>
</tr>
</tbody>
</table>

Organism
Homo sapiens

Disease
neuroblastoma

Cell line
SH-SY5Y

Cell type
neural cell

Genotype
pCDNA3 empty vector, pCDNA3-mPea3-VP16 vector
Figure from: Sero JE, Bakal C. Multiparametric Analysis of Cell Shape Demonstrates that β-PIX Directly Couples YAP Activation to Extracellular Matrix Adhesion. Cell Syst. 2017 Jan;4(1) 84-96.e6.
Imaging data in BioStudies: what we are aiming for

- Store all image data;
- Help explore the multi-dimensional image space.
## Imaging data in BioStudies – coming soon:

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Thumbnail</th>
<th>Plate</th>
<th>Row</th>
<th>Col</th>
<th>Well Name</th>
<th>CH4</th>
<th>Gene Target/Treatment</th>
<th>Cell Area</th>
<th>Nucleus Area</th>
<th>Nucleus/Cell Area</th>
<th>Neighbor Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>002002-1.tif</td>
<td>849 KB</td>
<td><img src="image1.png" alt="Image 1" /></td>
<td>20097272[5177]</td>
<td>2</td>
<td>B2</td>
<td>CD44</td>
<td>PFKP</td>
<td>NULL</td>
<td>largenuclei</td>
<td>highnuc/cellarea</td>
<td>highNeighborFraction</td>
<td>highlocaldensity</td>
</tr>
<tr>
<td>002002-1.tif</td>
<td>1.1 MB</td>
<td><img src="image2.png" alt="Image 2" /></td>
<td>20097272[5187]</td>
<td>2</td>
<td>B2</td>
<td>YAP/TAZ(rabbit,C-terepitope)</td>
<td>PFKP</td>
<td>NULL</td>
<td>largenuclei</td>
<td>highnuc/cellarea</td>
<td>highNeighborFraction</td>
<td>highlocaldensity</td>
</tr>
<tr>
<td>002002-1.tif</td>
<td>1.6 MB</td>
<td><img src="image3.png" alt="Image 3" /></td>
<td>20097301[5184]</td>
<td>2</td>
<td>B2</td>
<td>CD44</td>
<td>ADRBK1</td>
<td>NULL</td>
<td>largenuclei</td>
<td>NULL</td>
<td>highNeighborFraction</td>
<td>highlocaldensity</td>
</tr>
<tr>
<td>002002-1.tif</td>
<td>1.1 MB</td>
<td><img src="image4.png" alt="Image 4" /></td>
<td>20097301[5185]</td>
<td>2</td>
<td>B2</td>
<td>YAP/TAZ(rabbit,C-terepitope)</td>
<td>ADRBK1</td>
<td>NULL</td>
<td>largenuclei</td>
<td>highnuc/cellarea</td>
<td>highNeighborFraction</td>
<td>highlocaldensity</td>
</tr>
<tr>
<td>002002-10.tif</td>
<td>1.1 MB</td>
<td><img src="image5.png" alt="Image 5" /></td>
<td>20097272[5177]</td>
<td>2</td>
<td>B2</td>
<td>CD44</td>
<td>PFKP</td>
<td>NULL</td>
<td>largenuclei</td>
<td>highnuc/cellarea</td>
<td>highNeighborFraction</td>
<td>highlocaldensity</td>
</tr>
</tbody>
</table>
Thanks!

- www.ebi.ac.uk/biostudies
- biostudies@ebi.ac.uk
- Any questions?