EMBL-EBI Bioinformatics resources for exploring functional genomics data

Discover functional genomics data with ArrayExpress

Laura Huerta, PhD

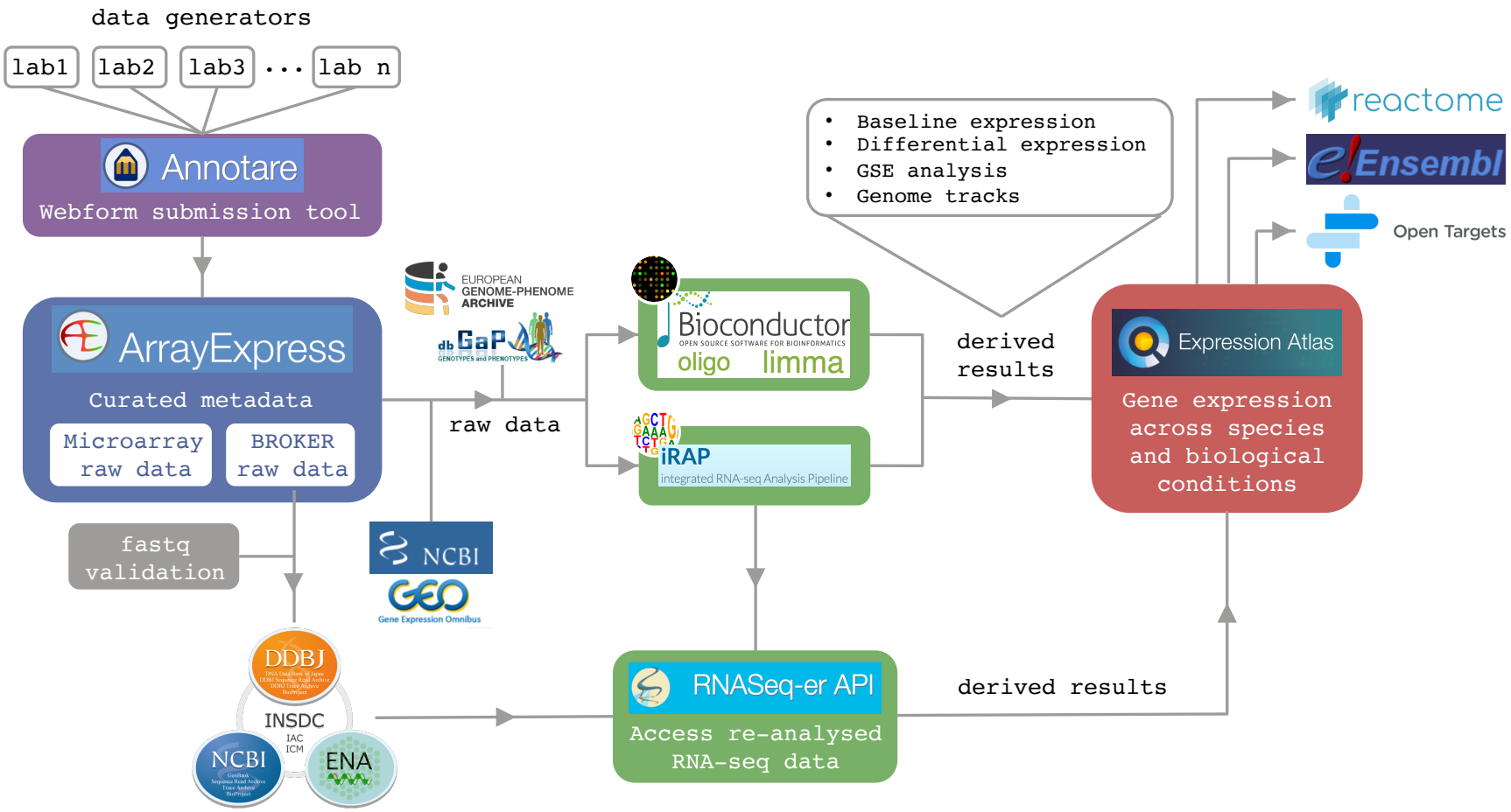
Senior Scientific Curator

lauhuema@ebi.ac.uk

9 November 2017

EMBL-EBI

Functional genomics resources at EMBL-EBI



Outline of the session

- ✓ What is ArrayExpress?
  - ✓ Data standards: MIAME & MINSEQE
  - ✓ Data format: MAGE-TAB
    - ✓ Hands-on exercise
  - ✓ Experimental variable
    - ✓ Hands-on exercise
- ✓ Data mining with ArrayExpress
    - ✓ Hands-on exercise

What is ArrayExpress?

commentary

One-stop shop for microarray data

Is a universal, public DNA-microarray database a realistic goal?

Alvis Brazma, Alan Robinson,
Graham Cameron
and Michael Ashburner

NATURE | VOL 403 | 17 FEBRUARY 2000 | www.nature.com

"The EBI, in collaboration with the German Cancer Reasearch Centre, is developing ArrayExpress"

Martin Vingron

Alvis Brazma

What is ArrayExpress?

commentary

One-stop shop for microarray data

Is a universal, public DNA-microarray database a realistic goal?

Alvis Brazma, Alan Robinson, Graham Cameron and Michael Ashburner

Of the techniques that are being used to obtain the massive data sets of the molecules of life, the most visible is the DNA sequencing of the human genome. Following on from the publication of the human chromosome 22 sequence[1], a rough draft of the whole human genome should be available by the spring. But such advances can create the false impression that everything about life at the molecular level will soon be understood.

In reality, genome projects simply transfer digital information from DNA to computer file; this genetic 'parts-list' is a long way from providing an understanding of function. It took hundreds of years to advance from a fairly detailed understanding of human anatomy to any real understanding of function. Knowing the genome sequence and experiment looking at 40,000 genes from 10 different samples, under 20 different conditions, produces at least 8,000,000 pieces of information. Currently, these data are scattered among various independent Internet sites, or may not be publicly available at all, although conclusions drawn from the data will have been published. Details about how experiments were carried out are often incomplete. Yet the amount of information being produced in this way is set to explode as the cost of microarray technology falls.

The need for a public repository

It is time to create a public repository for microarray data, with standardized annotation (see Box 2, overleaf). But this is a complex and ambitious project, and is one of the biggest challenges that bioinformatics has yet faced. Major difficulties stem from the detail required to describe the conditions of an experiment, and the relative and imprecise

One difficulty concerns the inherent fuzziness of gene-expression data. Essentially all current expression measurements are relative: we can tell which genes are expressed differently in an experiment only in comparison with another experiment, or in relation to another gene in the same experiment. Such methods tell us little about how many copies of a messenger RNA are present. Moreover, the transcription levels reported are an average over the whole cell population sampled.

Consequently, gene-expression measurements from different technologies, or even from the same technology but from different laboratories, may not be quantitatively comparable. Two steps should allow data from different sources to be compared. First, relatively raw data should be stored to obviate any variation owing to, say, data-normalization methods. Second, standard sets of control probes and samples should be designed and used in experiments to give reference points

"It's time to create a public repository for microarray data, with standardized annotation" (2000)

What is ArrayExpress?

commentary

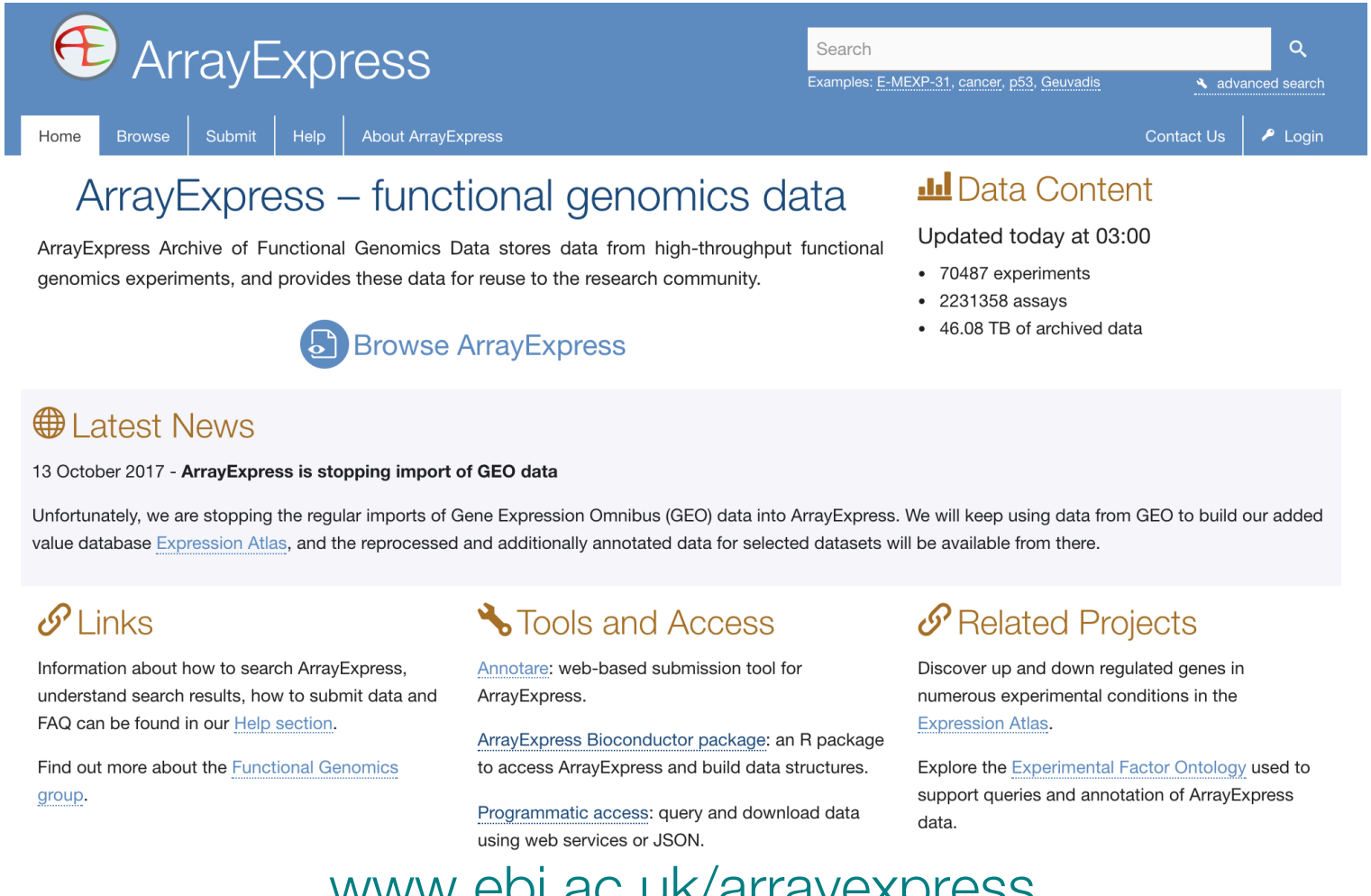
One-stop shop for microarray data

Is a universal, public DNA-microarray database a realistic goal?

Important tasks to be undertaken include:

1. agreement on the essential information that should be reported for a microarray experiment
2. definition of ontologies and an extensible, structured document format to capture these data and their semantics
3. production of a database to store these documents
4. development of tools for searching documents in a database and using the semantic context to allow comparisons and sophisticated queries.

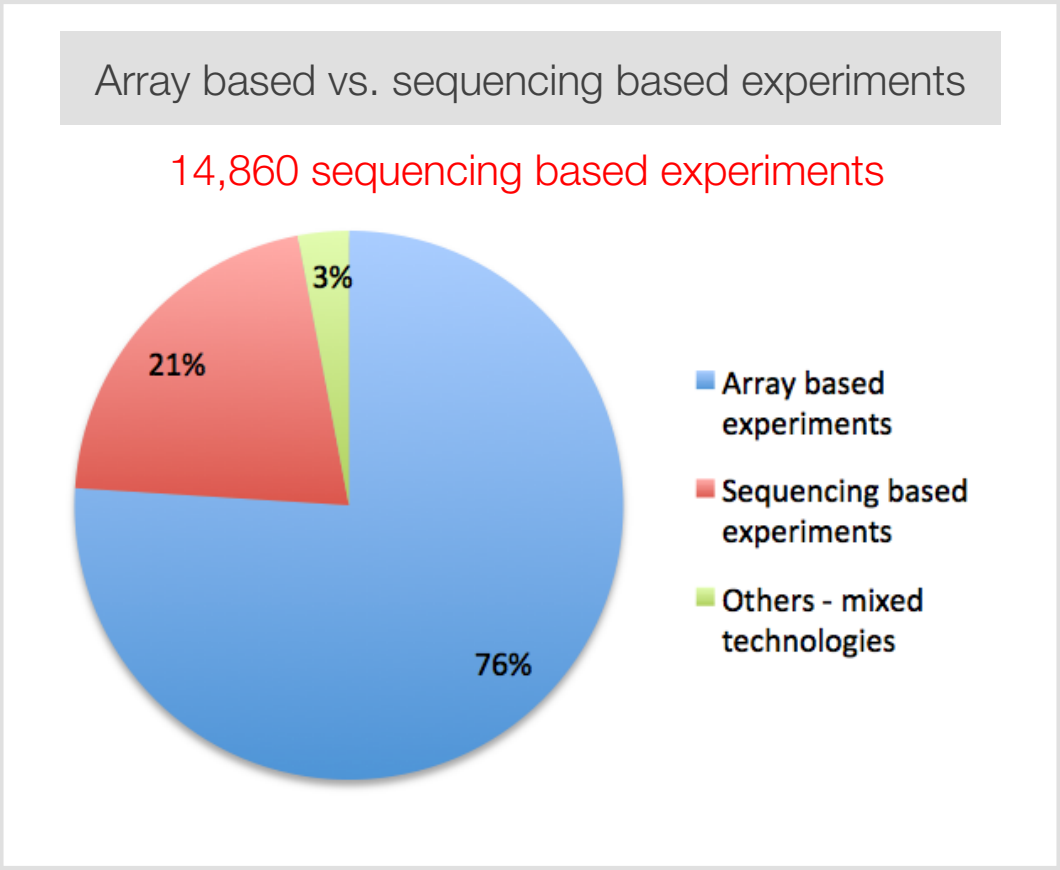
ArrayExpress: functional genomics data archive



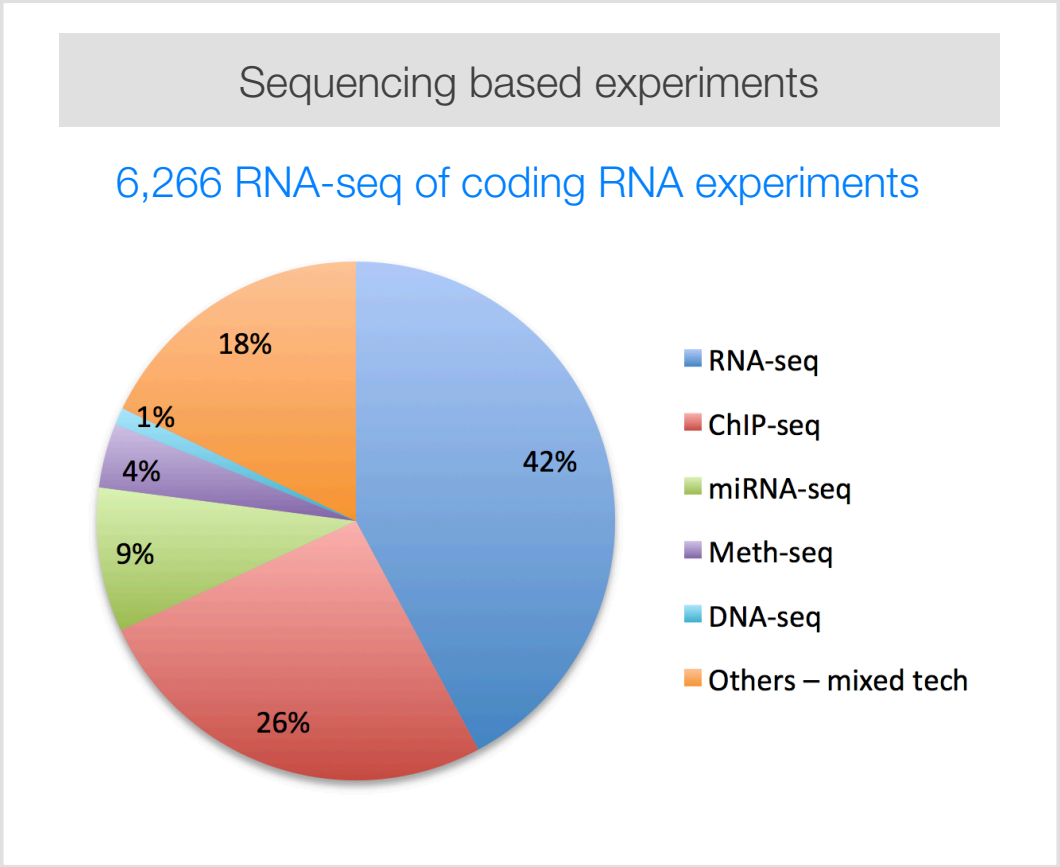
www.ebi.ac.uk/arrayexpress

EMBL-EBI

ArrayExpress: functional genomics data archive



ArrayExpress: functional genomics data archive



ArrayExpress: functional genomics data archive

Experiment types in ArrayExpress

Experiment Type	Definition (EFO term accession)	Example
4C	EFO_0007690	E-MTAB-2180
antigen profiling	EFO_0000747	E-MTAB-3606
ATAC-seq	EFO_0007045	E-MTAB-3972
Bisulfite-seq	EFO_0003753	E-MTAB-1042
Capture-C	EFO_0007691	E-MTAB-4845
ChIP-chip by array	EFO_0002760	E-MTAB-2804
ChIP-chip by SNP array	EFO_0002764	E-GEOD-22306
ChIP-chip by tiling array	EFO_0002762	E-MTAB-1402
ChIP-seq	EFO_0002692	E-MTAB-3631
CLIP-seq	EFO_0003143	E-MTAB-1371
comparative genomic hybridization by array	EFO_0000749	E-MTAB-2293
DNA-seq	EFO_0002693	E-MTAB-3109
FAIRE-seq	EFO_0004428	E-MTAB-3199

www.ebi.ac.uk/arrayexpress/help/experiment_types.html

ArrayExpress: ELIXIR core data resources

ELIXIR Core Data Resources

ELIXIR Core Data Resources are a set of European data resources of fundamental importance to the wider life-science community and the long-term preservation of biological data.

Identification of the ELIXIR Core Data Resources involves a careful evaluation of the multiple facets of the data resources. Indicators used in the evaluation are grouped into five categories:

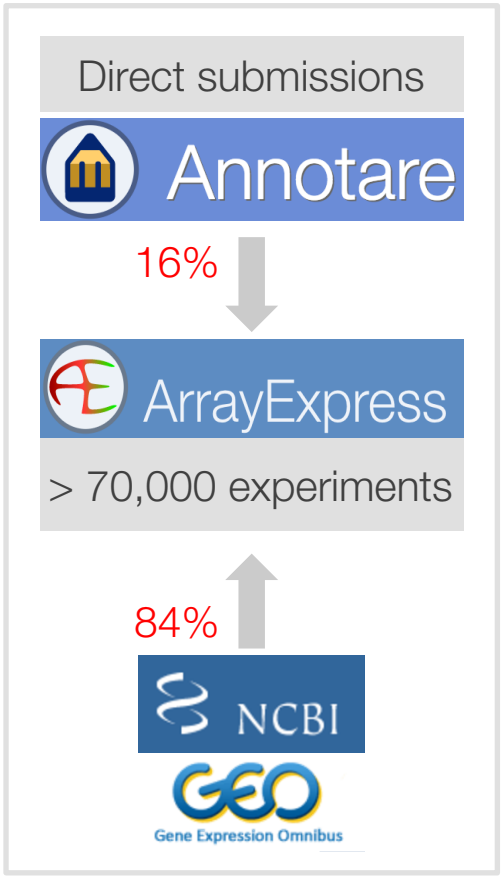
- Scientific focus and quality of science
- Community served by the resource
- Quality of service
- Legal and funding infrastructure, and governance
- Impact and translational stories

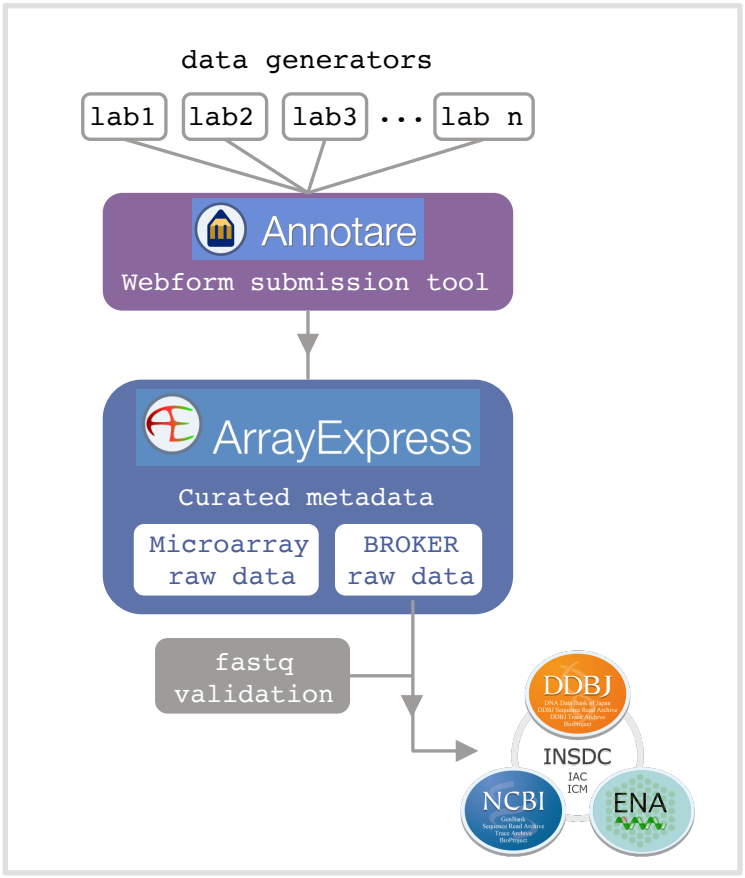
The details of the selection criteria are described in the F1000R ELIXIR track article 'Identifying ELIXIR Core Data Resources'. For an overview, watch the webinar about the Core Data Resources.

The initial Core Data Resource list was defined in July of 2017. The list will be reviewed regularly - further rounds of selection are planned, going forward.

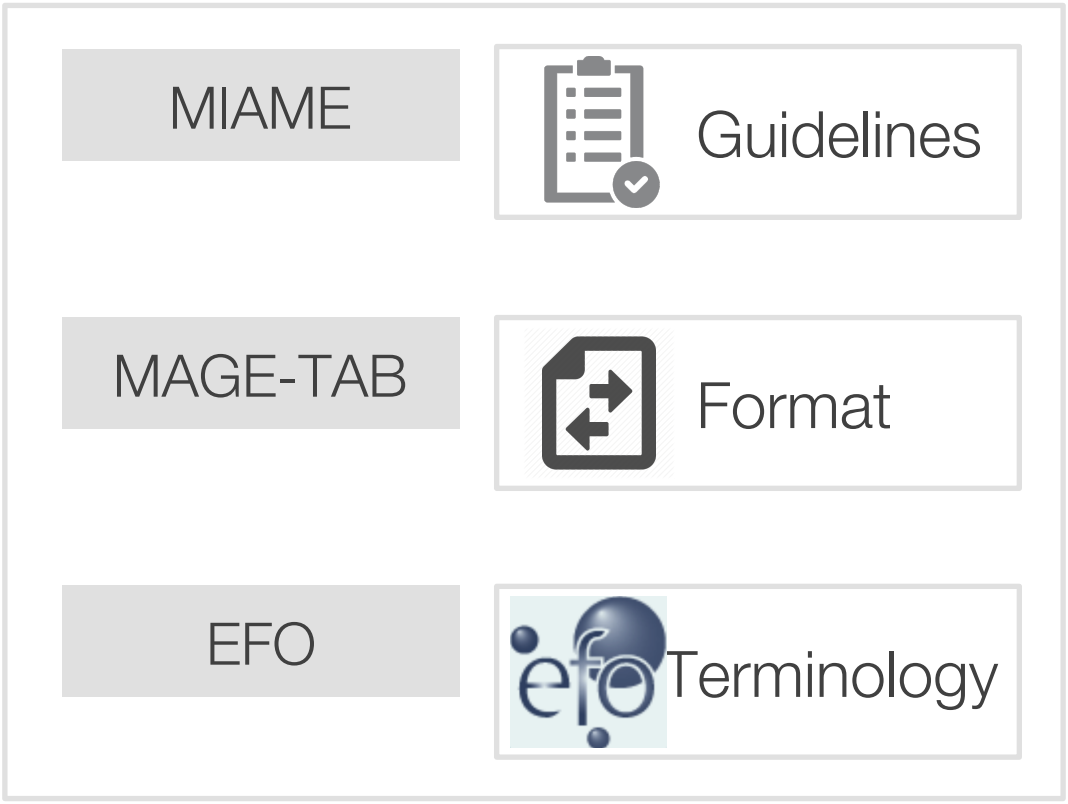
www.elixir-europe.org/platforms/data/core-data-resources

ArrayExpress: archive for direct submissions





How standards are used in ArrayExpress



ArrayExpress: MIAME compliant

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*"

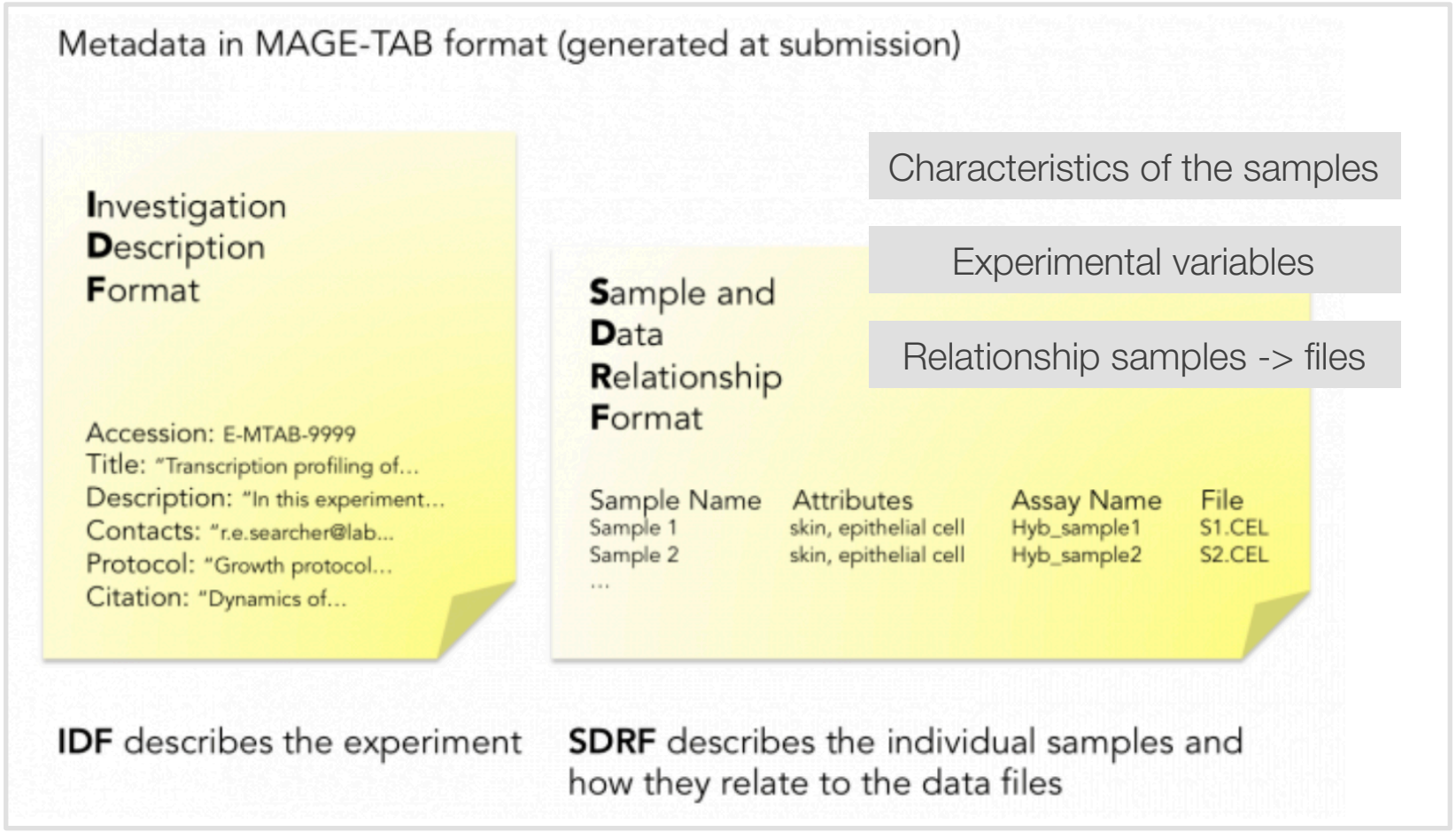
Metadata is stored in MAGE-TAB format

Microarray data

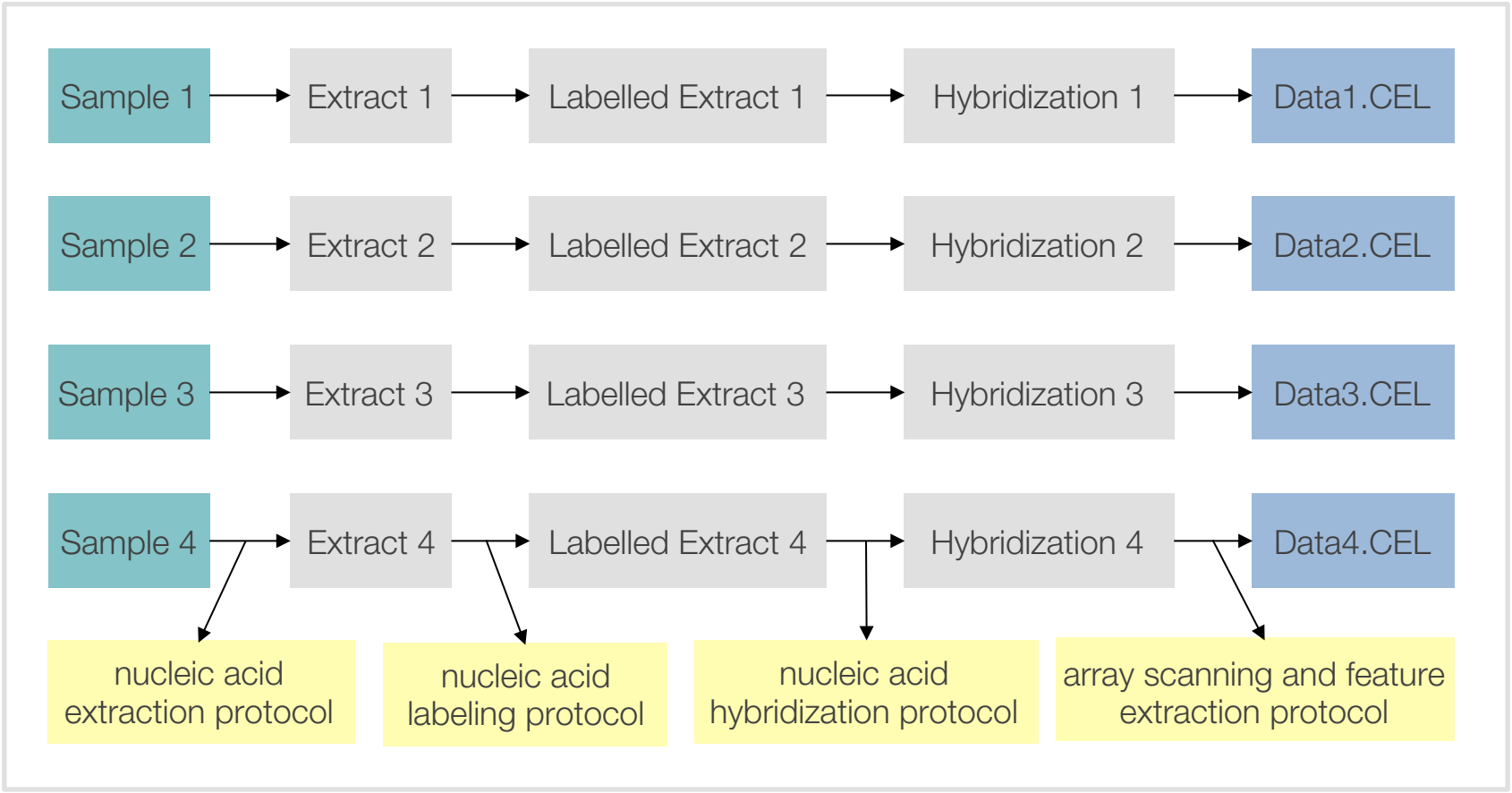
MAGE-TAB specification defines 4 types of files:

1. *Investigation Description Format (IDF)*
2. *Sample and Data Relationship Format (SDRF)*
3. *Array Design Format (ADF)*
4. *Raw and processed data files*

Metadata is stored in MAGE-TAB format



Metadata is stored in MAGE-TAB format



Metadata is stored in MAGE-TAB format

SRDF and IDF files

Hands-on activity

Go to experiment E-GEOD-49515 and download and open the corresponding SDRF and IDF files

www.ebi.ac.uk/arrayexpress

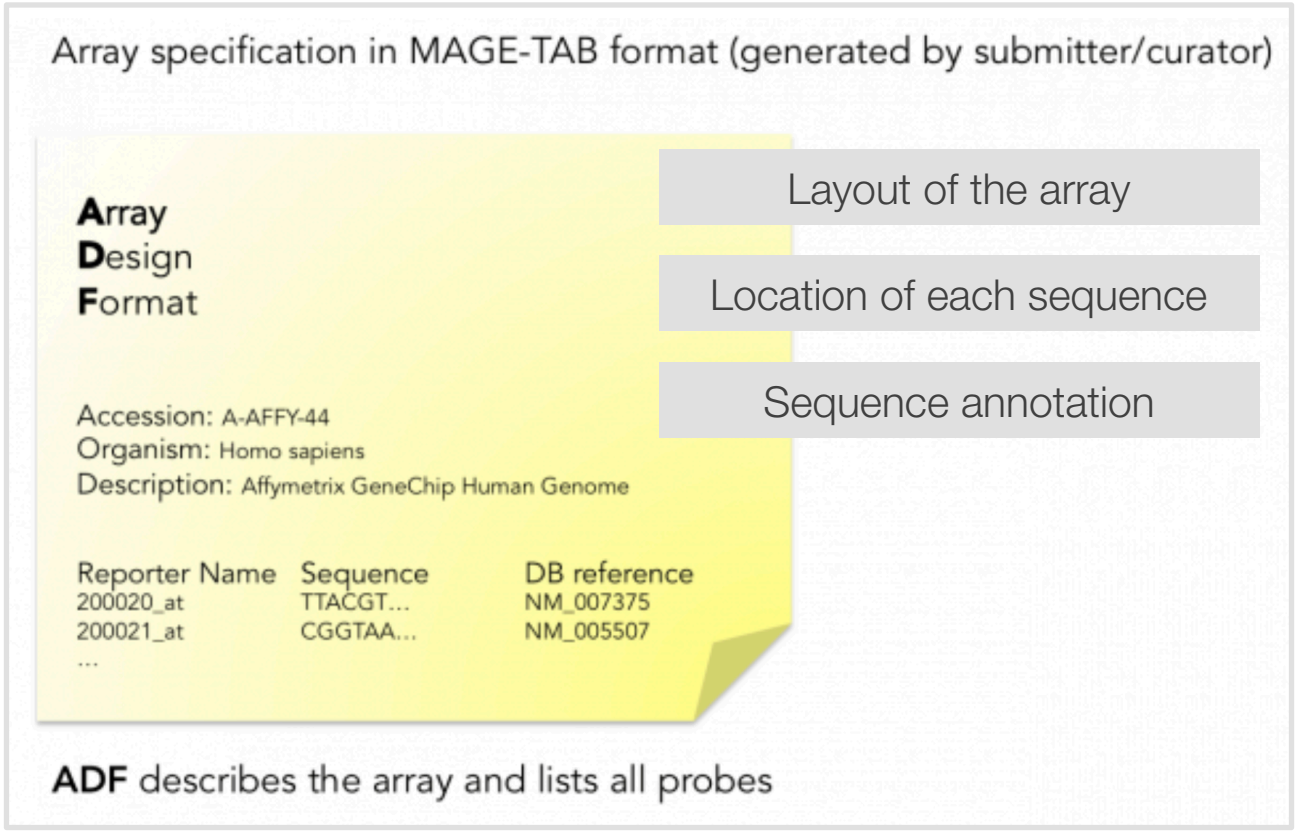
Metadata is stored in MAGE-TAB format

Microarray data

MAGE-TAB specification defines 4 types of files:

1. *Investigation Description Format (IDF)*
2. *Sample and Data Relationship Format (SDRF)*
3. *Array Design Format (ADF)*
4. *Raw and processed data files*

Metadata is stored in MAGE-TAB format



www.ebi.ac.uk/arrayexpress/help/adf_submissions_overview.html

Metadata is stored in MAGE-TAB format

Array Design (ADF) file

Hands-on activity

Go to experiment E-GEOD-49515 and download and open the ADF file

www.ebi.ac.uk/arrayexpress

Metadata is stored in MAGE-TAB format

Microarray data

MAGE-TAB specification defines 4 types of files:

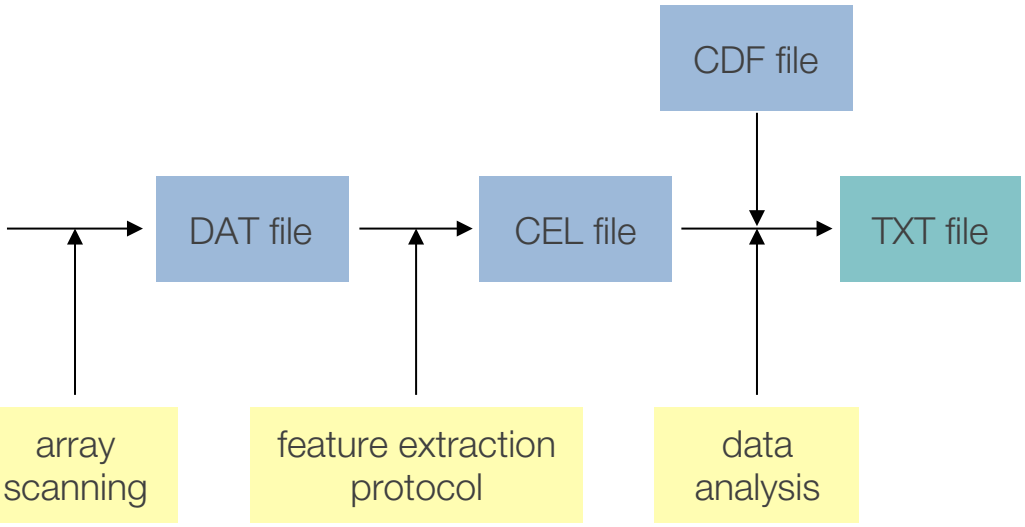
1. *Investigation Description Format (IDF)*
2. *Sample and Data Relationship Format (SDRF)*
3. *Array Design Format (ADF)*
4. *Raw and processed data files*

Raw and processed data files

Microarray data

Raw data: "native" files generated by the microarray scanner software.

Hybridised GeneChip



Common microarray raw data files

Manufacturer	Typical raw data format	How to open / Analysis software examples
Affymetrix	*.CEL*	R packages (affy, limma, oligo...)
Agilent	*feature extraction file* (tab-delimited text file per hybridisation)	Spreadsheet software (Excel, OpenOffice, etc.)
GenePix (scanner)	*.gpr* (tab-delimited text file per hybridisation)	Spreadsheet software (Excel, OpenOffice, etc.)
Illumina	*.idat*	R packages (e.g. illuminaio)
Illumina	*txt* (tab-delimited text matrix for all samples)	Spreadsheet software (Excel, OpenOffice, etc.)
Nimblegen	NimbleScan, *pair* (tab-delimited text matrix for all samples)	Spreadsheet software (Excel, OpenOffice, etc.)

www.ebi.ac.uk/fg/annotare/help/accepted_raw_ma_file_formats.html

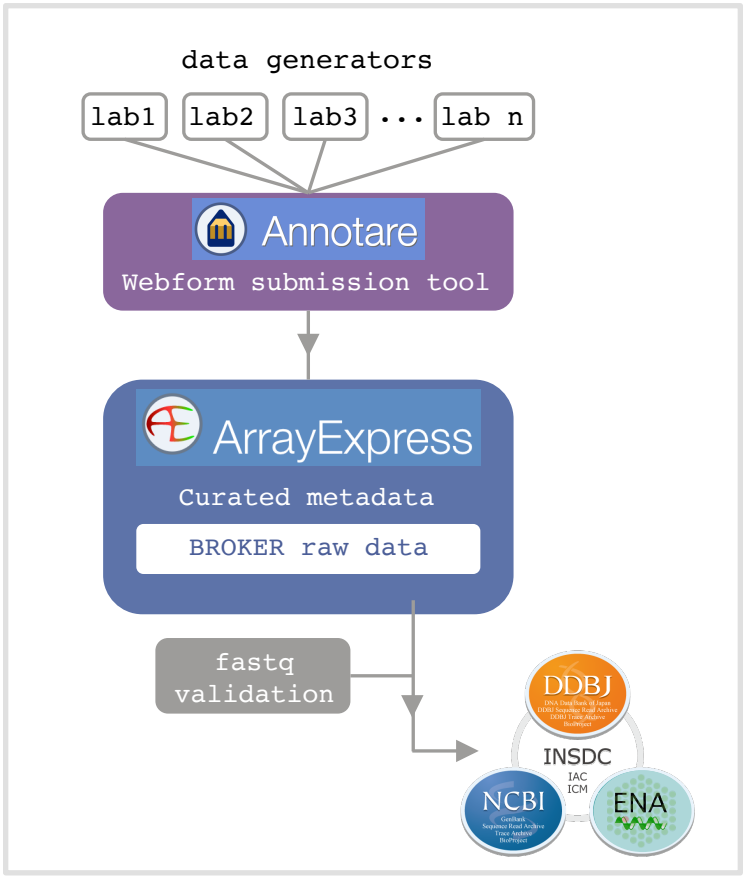
Metadata is stored in MAGE-TAB format

RNA-seq data

MAGE-TAB specification defines 3 types of files:

1. *Investigation Description Format (IDF)*
2. *Sample and Data Relationship Format (SDRF)*
3. ~~*Array Design Format (ADF)*~~
4. *Raw and processed data files*

Raw data is deposited at ENA



Raw data – stable archiving at INSDC

- ✓ INSDC (3 partners) runs the sequence read archive (SRA)
- ✓ ArrayExpress brokers raw NGS data to ENA & GEO to NCBI SRA

INSDC – International Nucleotide Sequence Database Collaboration

Accession number conventions

E-ERAD-475 RNA-seq of zebrafish developmental stages

ERP014517 at ENA

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model		
PRJEB12982	SAMEA3892004	ERS1079138	ERX1512938	ERR1442561	7955	Danio rerio	Illumina HiSeq 2500	PAIRED	File 1 File 2
PRJEB12982	SAMEA3892005	ERS1079139	ERX1512939	ERR1442562	7955	Danio rerio	Illumina HiSeq 2500	PAIRED	File 1 File 2
PRJEB12982	SAMEA3892006	ERS1079140	ERX1512940	ERR1442563	7955	Danio rerio	Illumina HiSeq 2500	PAIRED	File 1 File 2
							Illumina HiSeq 2500	PAIRED	File 1 File 2

ENA objects	Accession examples
study	ERP014517
samples	ERS1079138
experiments (RNA-seq lib)	ERX1512938
runs	ERR1442561

SRA objects	DDBJ	ENA	NCBI
study	DRP	ERP	SRP
sample	DRS	ERS	SRS
experiment	DRX	ERX	SRX
run	DRR	ERR	SRR

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

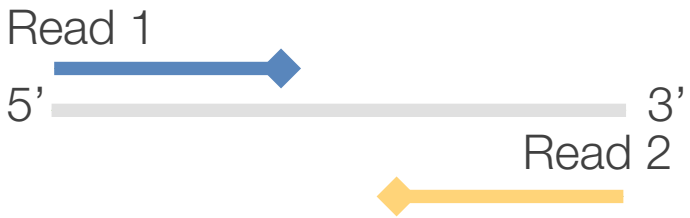
Single-end and paired-end sequencing

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
```

ArrayExpress: Experimental variable

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

Transcription profiling of blood from smokers, non-smokers and former smokers to identify gene expression signature for cigarette smoke exposure response

- organism > Homo sapiens
- age > 22, 57, 43, 39 year, etc.
- sex > male OR female
- organism part > blood
- disease > normal
- ethnic group > Caucasian OR African American
- clinical history> smoker, former smoker OR non-smoker

ArrayExpress: Experimental variable

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

Transcription profiling of blood from smokers (with or without COPD), non-smokers and former smokers to identify gene expression signature for cigarette smoke exposure response

- organism > Homo sapiens
- age > 22, 57, 43, 39 year, etc.
- sex > male OR female
- organism part > blood
- ethnic group > Caucasian OR African American
- disease > normal OR COPD
- clinical history> smoker, former smoker OR non-smoker

ArrayExpress: Experimental variable

MIAME & MAGE-TAB

Hands-on activity

Go to experiment E-GEOD-49515, find the experimental variable and the groups of samples compared. Download the SDRF file and figure out which file belongs to each category

www.ebi.ac.uk/arrayexpress

ArrayExpress: Experimental variable

	A	B	C
1	Source Name	Array Data File	FactorValue [disease]
2	GSM1200316 1	GSM1200316_PAN-PBMC-S03.CEL	pancreatic cancer
3	GSM1200315 1	GSM1200315_PAN-PBMC-S02.CEL	pancreatic cancer
4	GSM1200314 1	GSM1200314_PAN-PBMC-S01.CEL	pancreatic cancer
5	GSM1200313 1	GSM1200313_NOR-PBMC-S10.CEL	normal
6	GSM1200312 1	GSM1200312_NOR-PBMC-S09.CEL	normal
7	GSM1200311 1	GSM1200311_NOR-PBMC-S08.CEL	normal
8	GSM1200310 1	GSM1200310_NOR-PBMC-S07.CEL	normal
9	GSM1200309 1	GSM1200309_NOR-PBMC-S06.CEL	normal
10	GSM1200308 1	GSM1200308_NOR-PBMC-S05.CEL	normal
11	GSM1200307 1	GSM1200307_NOR-PBMC-S04.CEL	normal
12	GSM1200306 1	GSM1200306_NOR-PBMC-S03.CEL	normal
13	GSM1200305 1	GSM1200305_NOR-PBMC-S02.CEL	normal
14	GSM1200304 1	GSM1200304_NOR-PBMC-S01.CEL	normal
15	GSM1200303 1	GSM1200303_HCC-PBMC-S10.CEL	hepatocellular carcinoma
16	GSM1200302 1	GSM1200302_HCC-PBMC-S09.CEL	hepatocellular carcinoma
17	GSM1200301 1	GSM1200301_HCC-PBMC-S08.CEL	hepatocellular carcinoma
18	GSM1200300 1	GSM1200300_HCC-PBMC-S07.CEL	hepatocellular carcinoma
19	GSM1200299 1	GSM1200299_HCC-PBMC-S06.CEL	hepatocellular carcinoma
20	GSM1200298 1	GSM1200298_HCC-PBMC-S05.CEL	hepatocellular carcinoma
21	GSM1200297 1	GSM1200297_HCC-PBMC-S04.CEL	hepatocellular carcinoma
22	GSM1200296 1	GSM1200296_HCC-PBMC-S03.CEL	hepatocellular carcinoma
23	GSM1200295 1	GSM1200295_HCC-PBMC-S02.CEL	hepatocellular carcinoma
24	GSM1200294 1	GSM1200294_HCC-PBMC-S01.CEL	hepatocellular carcinoma
25	GSM1200293 1	GSM1200293_GAS-PBMC-S03.CEL	gastric cancer
26	GSM1200292 1	GSM1200292_GAS-PBMC-S02.CEL	gastric cancer
27	GSM1200291 1	GSM1200291_GAS-PBMC-S01.CEL	gastric cancer

ArrayExpress: Experimental variable

MIAME & MAGE-TAB

Hands-on activity

Annotate sample metadata following MIAME standard in MAGE-TAB format

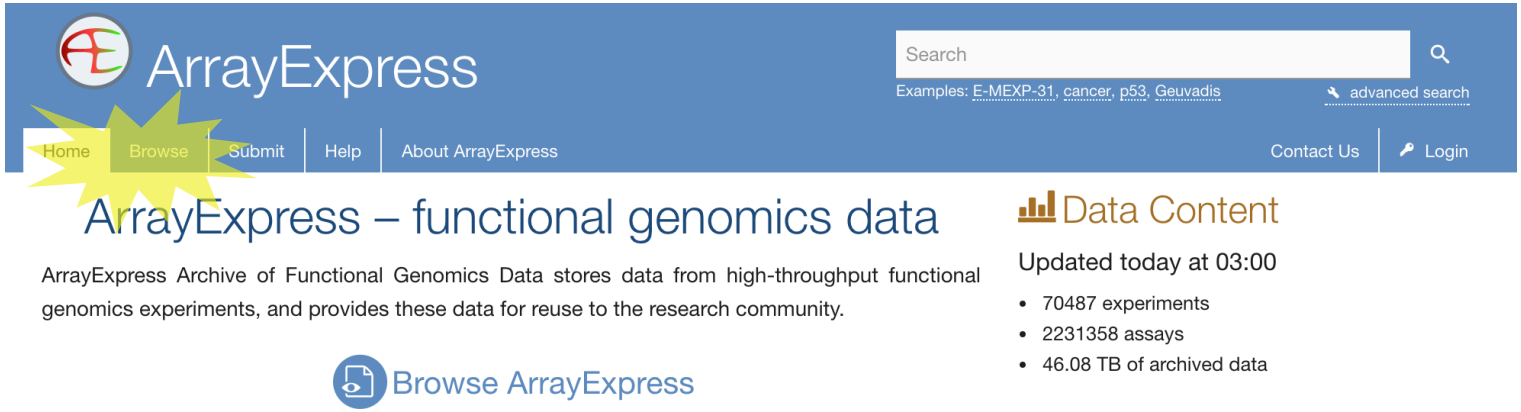
In pairs

www.ncbi.nlm.nih.gov/geo

Outline of the session

- ✓ What is ArrayExpress?
  - ✓ Data standards: MIAME & MINSEQE
  - ✓ Data format: MAGE-TAB
    - ✓ Hands-on exercise
  - ✓ Experimental variable
    - ✓ Hands-on exercise
- ✓ Data mining with ArrayExpress
    - ✓ Hands-on exercise

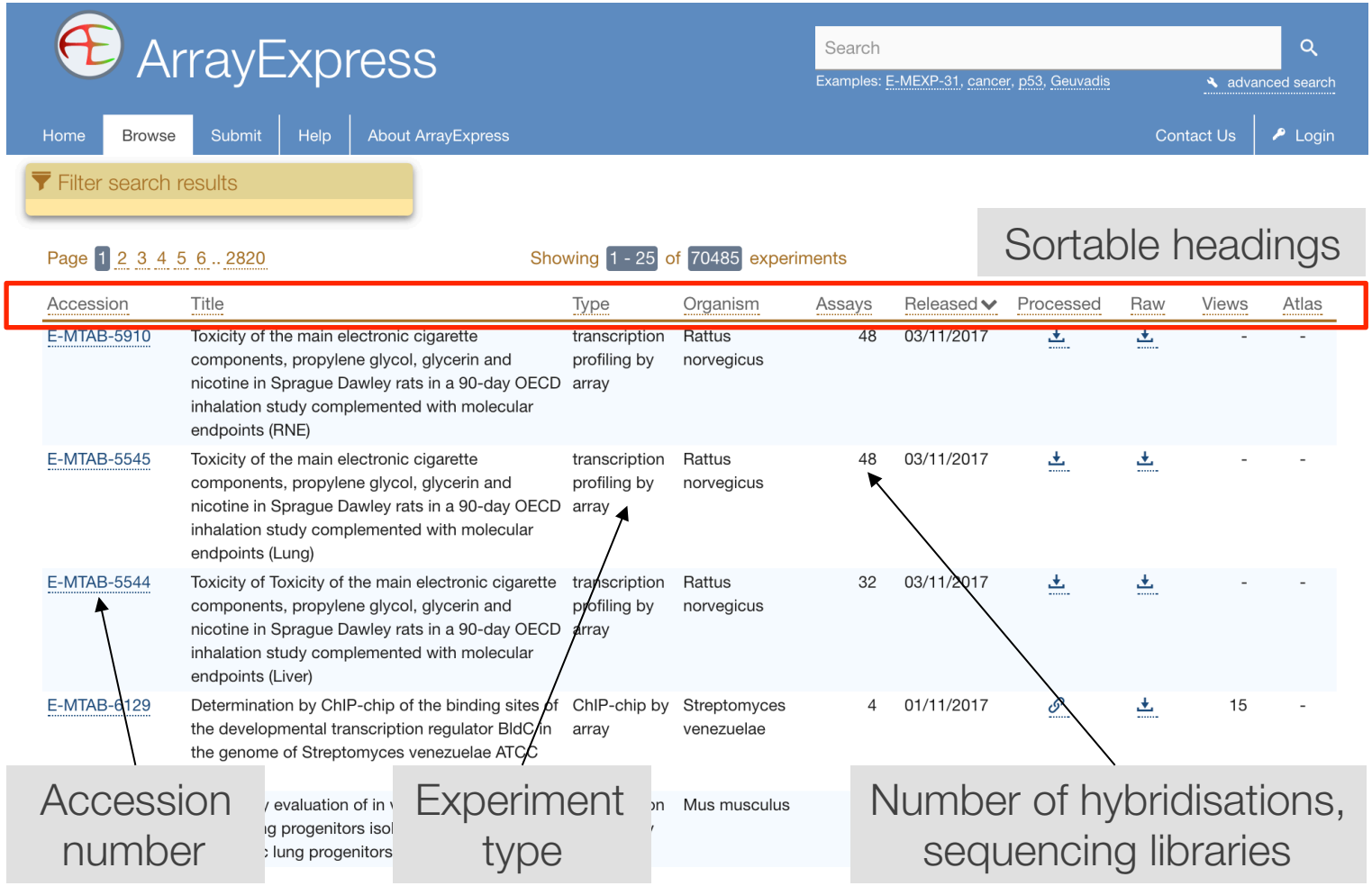
Data mining with ArrayExpress



www.ebi.ac.uk/arrayexpress

EMBL-EBI

Data mining with ArrayExpress



Data mining with ArrayExpress

Browse ArrayExpress

Hands-on activity

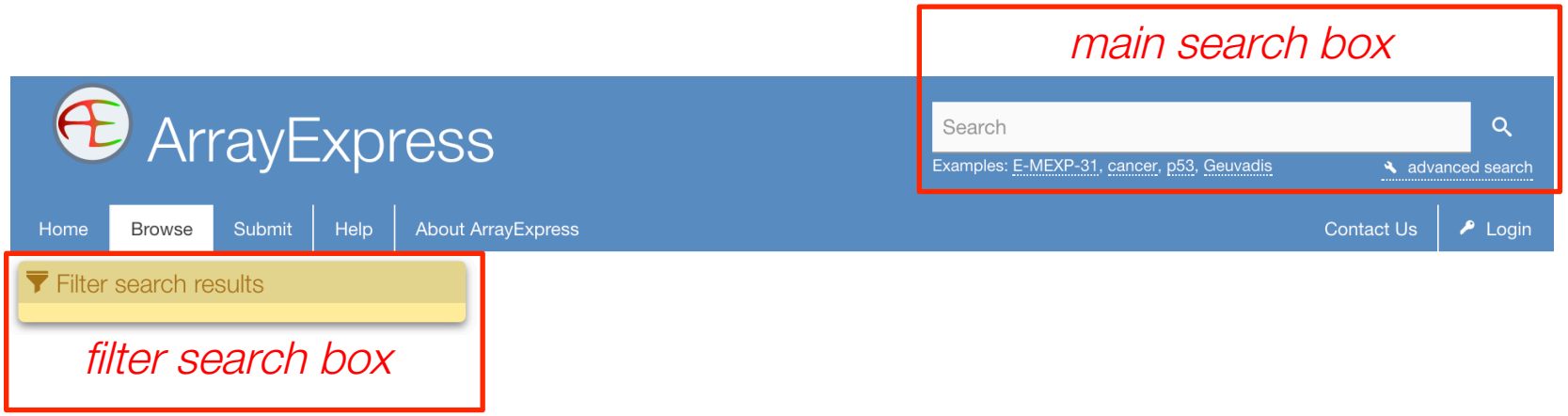
Find the biggest experiments (in terms of number of assays). Now display the top 100 most viewed experiments on one page

www.ebi.ac.uk/arrayexpress/experiments/browse.html

Data mining with ArrayExpress

Data mining: Discover datasets to analyse yourself

Find RNA-seq experiments related human autoimmune disease

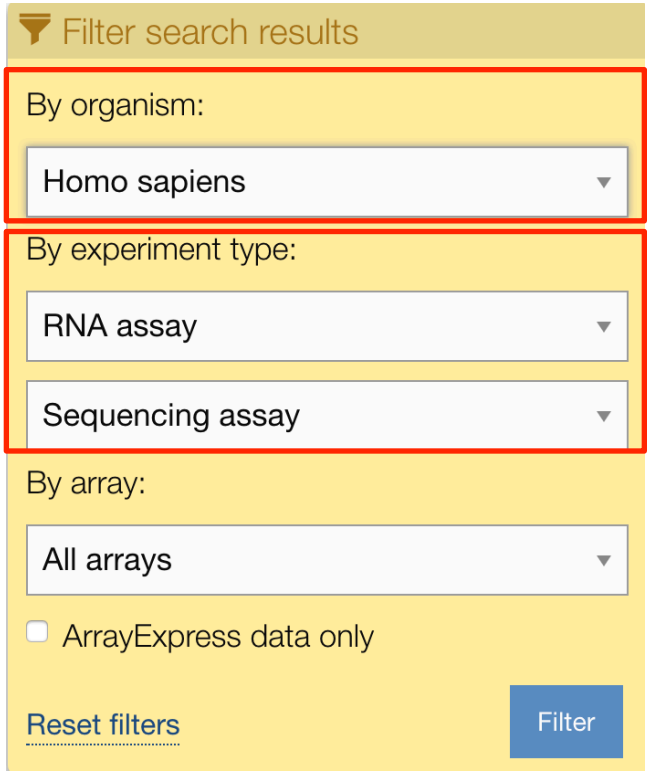


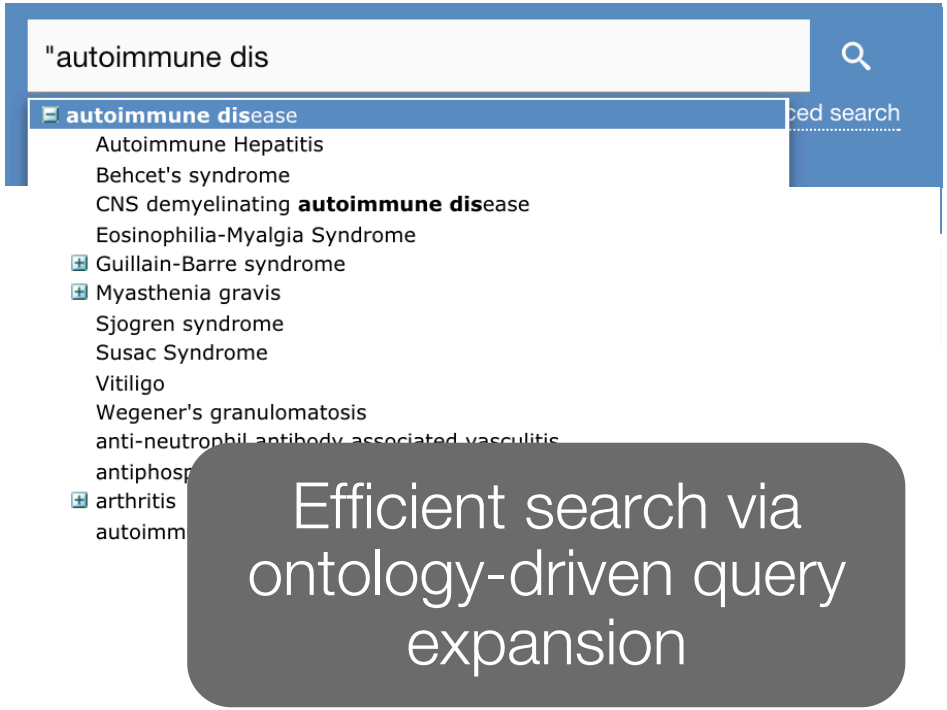
www.ebi.ac.uk/arrayexpress/browse.html

Data mining with ArrayExpress

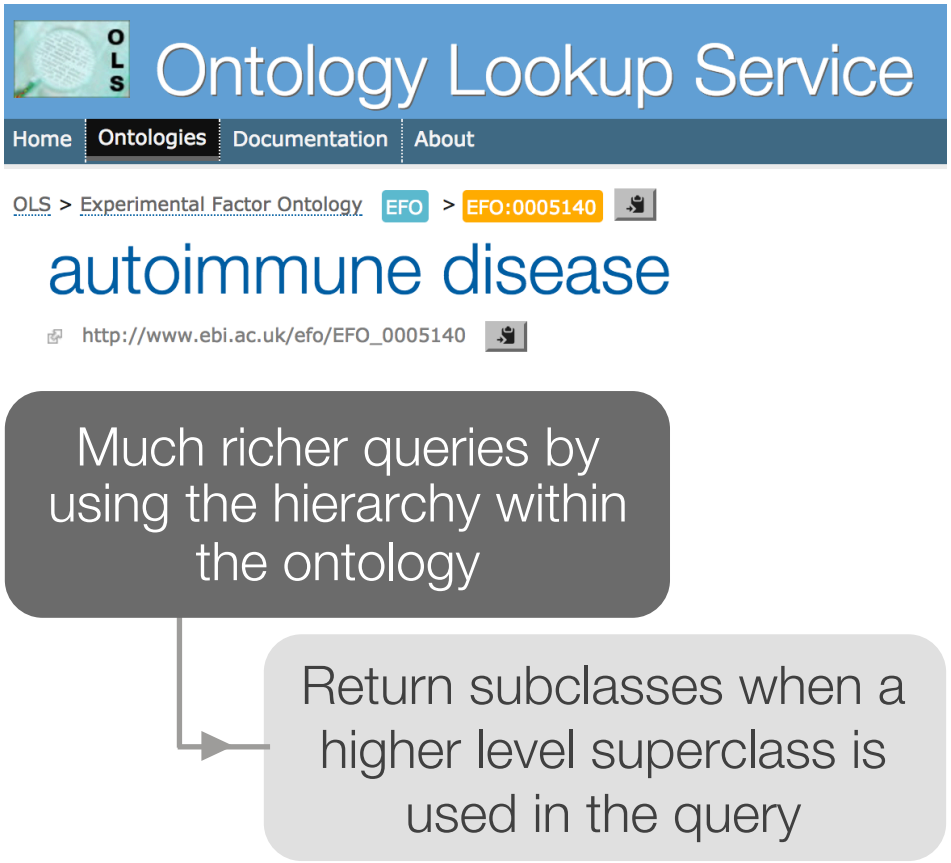
Data mining: Discover datasets to analyse yourself

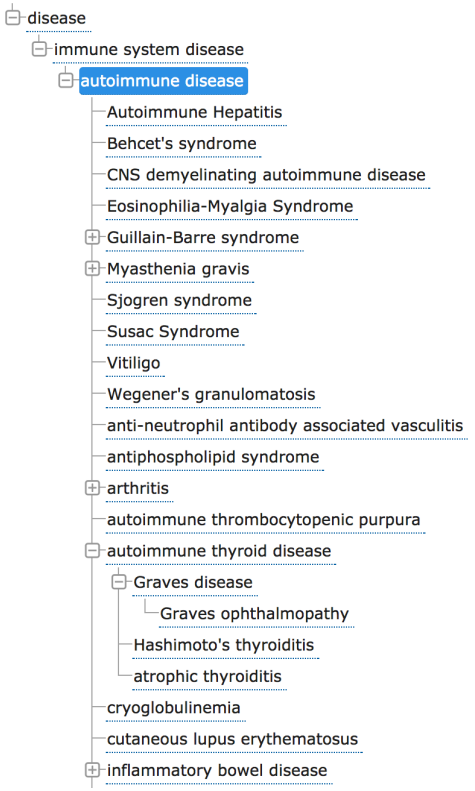
Find RNA-seq experiments related human autoimmune disease





Ontology-driven query expansion





www.ebi.ac.uk/ols/ontologies/efo

Data mining with ArrayExpress

Data mining: Discover datasets to analyse yourself

Find RNA-seq experiments related human autoimmune disease

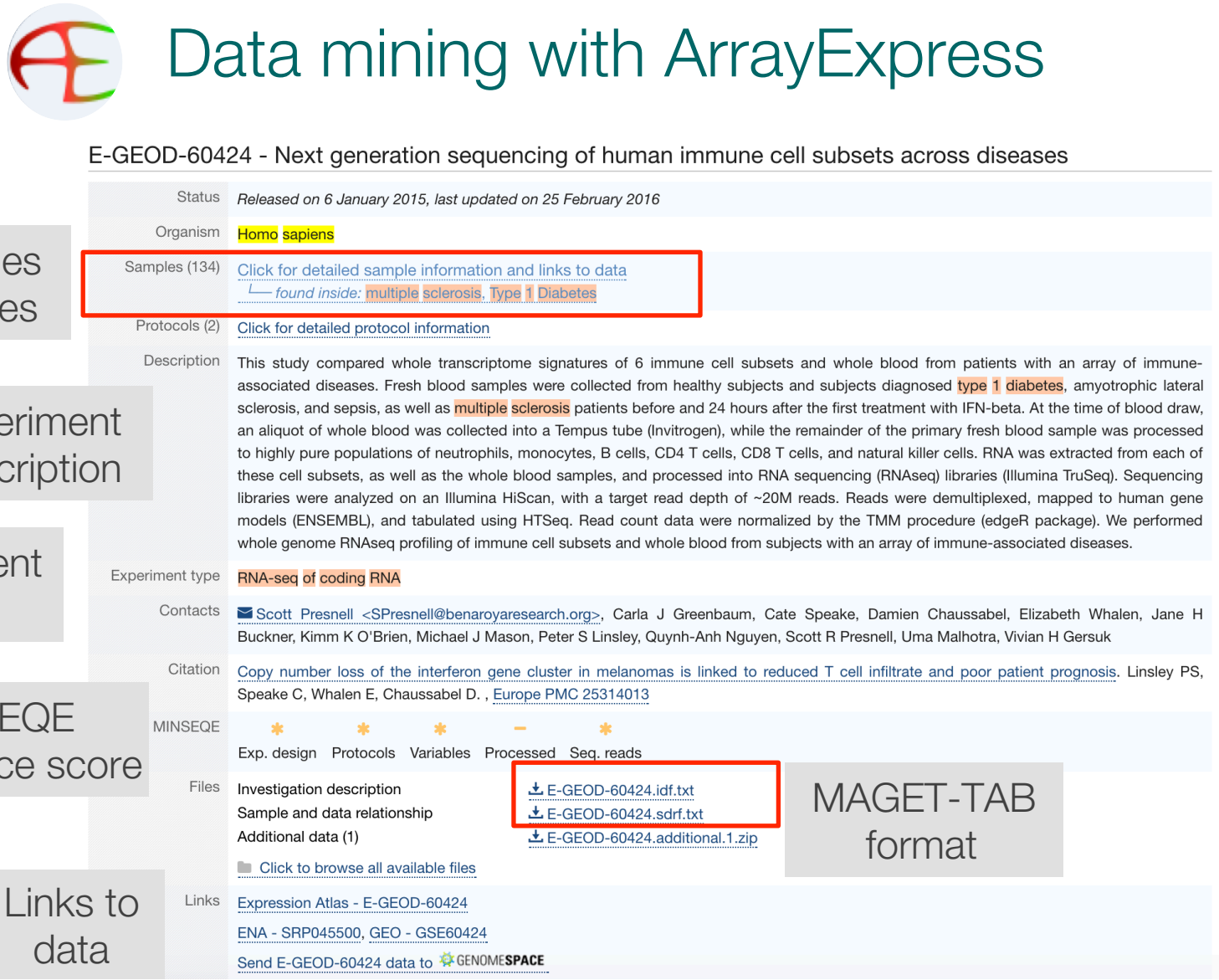
Search results for "autoimmune disease"

Filtered by organism Homo sapiens, experiment type "sequencing assay", experiment type "rna assay"

Page 1 2 — Showing 1 - 25 of 36 experiments — Page size 25 50 100 250 500

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Views	Atlas
E-GEOD-60424	Next generation sequencing of human immune cell subsets across diseases	RNA-seq of coding RNA	Homo sapiens	134	06/01/2015	-		3054	
E-MTAB-1568	RNA-seq of coding RNA in human plasma cells isolated from individuals with colorectal cancer, ulcertaive colitis and normal individuals	RNA-seq of coding RNA	Homo sapiens	9	29/03/2013	-		1078	-
E-GEOD-46579	A blood based 12-miRNA signature of Alzheimer patients	RNA-seq of non coding RNA	Homo sapiens	70	17/06/2013	-		824	-
E-GEOD-57945	Core Ileal Transcriptome in Pediatric Crohn Disease	RNA-seq of coding RNA	Homo sapiens	359	24/07/2014			623	-
E-MTAB-4304	RNA-Seq analysis of human intact and damaged osteoarthrtitic cartilage following total knee replacement	RNA-seq of coding RNA	Homo sapiens	22	06/04/2016	-		487	-
E-GEOD-83139	Single cell RNA-seq of human pancreatic endocrine cells from Juvenile, adult control and type 2 diabetic donors	RNA-seq of coding RNA	Homo sapiens	635	05/07/2016	-		328	-

Data mining with ArrayExpress



Data mining with ArrayExpress

E-GEOD-60424 - Next generation sequencing of human immune cell subsets across diseases

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

Page 1 2 3 4 5 6 .. 11 | Showing 1 - 25 of 268 rows | Page size 25 50 100 250 500

		Sample Attributes									Links to Data	
Source Name	name	age	cell type	disease	individual	organism	ethnic group	sex	clinical information	cell type	disea	ENA	FASTQ
GSM1479433 1		32 (year)	whole blood	normal	44	Homo sapiens	hispanic	female		whole blood	norm		
GSM1479433 1		32 (year)	whole blood	normal	44	Homo sapiens	hispanic	female		whole blood	norm		
GSM1479434 1		52 (year)	whole blood	multiple sclerosis	31	Homo sapiens	white	female	smoker	whole blood	multi		
GSM1479434 1		52 (year)	whole blood	multiple sclerosis	31	Homo sapiens	white	female	smoker	whole blood	multi		
GSM1479435 1		52 (year)	whole blood	multiple sclerosis	33	Homo sapiens	white	female	smoker	whole blood	multi		
GSM1479435 1		52 (year)	whole blood	multiple sclerosis	33	Homo sapiens	white	female	smoker	whole blood	multi		
GSM1479436 1		24 (year)	whole blood	Type 1 Diabetes	34	Homo sapiens	white	female		whole blood	Type		
GSM1479436 1		24 (year)	whole blood	Type 1 Diabetes	34	Homo sapiens	white	female		whole blood	Type		
GSM1479437 1		27 (year)	whole blood	Type 1 Diabetes	37	Homo sapiens	white	female		whole blood	Type		
GSM1479437 1		27 (year)	whole blood	Type 1 Diabetes	37	Homo sapiens	white	female		whole blood	Type		
GSM1479438 1		32 (year)	neutrophils	normal	44	Homo sapiens	hispanic	female		neutrophils	norm		
GSM1479438 1		32 (year)	neutrophils	normal	44	Homo sapiens	hispanic	female		neutrophils	norm		
		(year)	monocytes	normal			ispanic	female		monocytes	norm		
		(year)	monocytes	normal			ispanic	female					
		(year)	B-Cells	normal			ispanic	female					
		(year)	B-Cells	normal			ispanic	female					

Unique sample name for each biological replicate

Sample attributes describing the source material

Experimental variables

Links to download data

Data mining with ArrayExpress

Data mining: Discover datasets to analyse yourself

Find RNA expression arrays from human patients with diabetes

Search results for **diabetes**

Filtered by organism **Homo sapiens**, experiment type **"array assay"**, experiment type **"rna assay"**

Page **1** 2 3 4 5 6 .. 11 Showing **1 - 25** of **253** experiments Page size **25** 50 100 250 500

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Views	Atlas
E-MTAB-2976	Global gene expression of bone marrow multipotent mesenchymal stromal cells isolated from type 1 diabetes patients and healthy donors	transcription profiling by array	Homo sapiens	8	01/12/2014		-	272	-
E-MTAB-2902	Transcriptional profiling by array of microRNAs of human peripheral blood mononuclear cells - PBMC - from control individuals for Type 2 diabetes mellitus patients	microRNA profiling by array	Homo sapiens	9	01/04/2015	-		255	-
E-MTAB-2899	Transcriptional profiling by array of microRNAs from human peripheral blood mononuclear cells - PBMC - of Type 2 diabetes mellitus patients	microR profilin array							
E-MTAB-2896	Transcriptional Profiles in Peripheral Blood Mononuclear Cells - PBMC - from individual controls for Type 2 Diabetes Mellitus	transc profilin array							

Exact match to search term

Matched EFO synonyms to search term

Matched EFO child term of search term

ArrayExpress – Advanced search

Search in all experimental fields, e.g. experiment description, protocols, publication title…

Limit your search only to experiments in which “diabetes” is the value of the experimental variable

Search results for evv:diabetes

Filtered by organism Homo sapiens, experiment type "rna assay", experiment type "sequencing assay"

3 experiments

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Views	Atlas
E-MTAB-5061	Single-cell RNA-seq analysis of human pancreas from healthy individuals and type 2 diabetes patients	RNA-seq of coding RNA from single cells	Homo sapiens	3514	22/09/2016			3091	
E-MTAB-5060	Whole-islet RNA-sequencing analysis of human pancreas from healthy individuals and type 2 diabetes patients	RNA-seq of coding RNA	Homo sapiens	7	22/09/2016			903	
E-GEOD-60424	Next generation sequencing of human immune cell subsets across diseases	RNA-seq of coding RNA	Homo sapiens	134	06/01/2015	-		3163	

Export table in Tab-delimited format | Export matching metadata in XML format | Subscribe to RSS feed matching this search

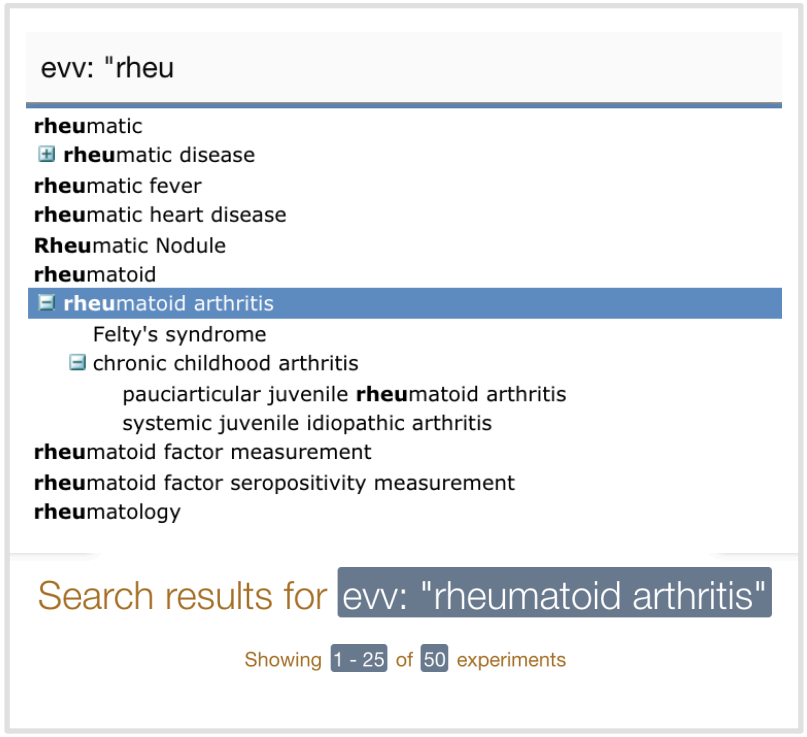
ArrayExpress – Advanced search

Field name	Search scope	Example use case
accession	Experiment primary or secondary accession	accession:E-MTAB-1234
array	Array design accession or name	array:A-AFFY-33
ev (or ef)	Experimental variable (or factor), the name of the main variable under study in an experiment. E.g. if the variable is "sex" in a human study, the researchers would be comparing between male and female samples, and "sex" is not merely an attribute the samples happen to have. Has EFO expansion.	ev:genotype
evv (or efv)	The value of an experimental variable (or factor). E.g. The values for "genotype" factor can be "wild type", "p53-/-". Has EFO expansion.	evv:"wild type"
expdesign	Experiment design type, related to the questions being addressed by the study, e.g. "time series design", "stimulus or stress design", "genetic modification design". Has EFO expansion.	expdesign:"time series"
exptype	Experiment type, related to the assay technology used. See the full list of experiment types in ArrayExpress. Has EFO expansion.	exptype:"RNA-seq of coding RNA"
gxa	Presence/absence of an ArrayExpress experiment in the Expression Atlas. Use values "true" and "false" respectively.	gxa:true
pmid	PubMed identifier for a publication.	pmid:16553887
sa	Sample attribute values. Has EFO expansion.	sa:fibroblast
sac	Sample attribute category. Find experiments that have a specific sample attribute defined, e.g. "age", "strain". Has EFO expansion.	sac:age
organism	Species of the samples. Can use common name (e.g. "mouse") or binomial nomenclature/Latin names (e.g. "Mus musculus"). Has EFO expansion.	organism:"homo sapiens"

ArrayExpress – Advanced search

Data mining: Discover datasets to analyse yourself

Find experiments comparing healthy (normal) and rheumatoid arthritis patients



ArrayExpress – Advanced search

Data mining: Discover datasets to analyse yourself

Find experiments comparing healthy (normal) and rheumatoid arthritis patients

evv:"rheumatoid arthritis" AND evv:normal

19 experiments

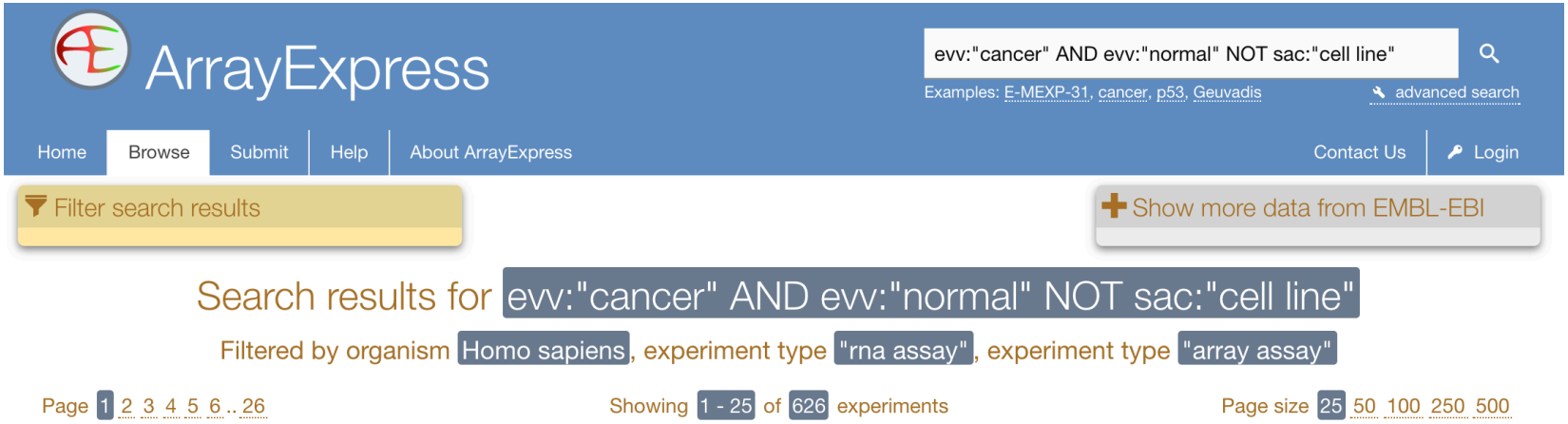
evv:"rheumatoid arthritis" AND (evv:normal OR evv:healthy)

Showing 1 - 25 of 28 experiments

ArrayExpress – Advanced search

Data mining: Discover datasets to analyse yourself

Find RNA expression arrays from human cancer samples
excluding all experiments performed on cell lines



Let's try ArrayExpress

Browsing ArrayExpress

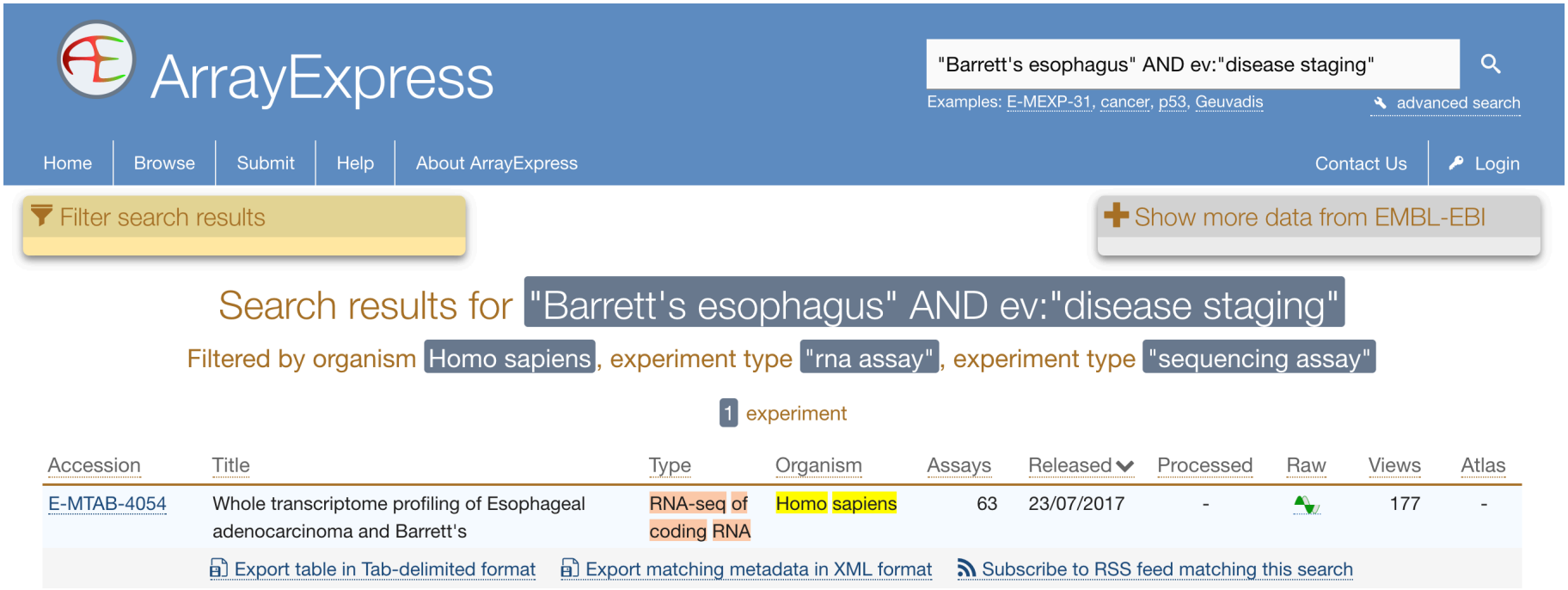
Hands-on activity

Discover interesting datasets related to your research area

In pairs

www.ebi.ac.uk/arrayexpress

Let's try ArrayExpress



Let's try ArrayExpress

E-MTAB-4054 - Whole transcriptome profiling of Esophageal adenocarcinoma and Barrett's

Status	*Submitted on 11 November 2015, released on 23 July 2017, last updated yesterday*
Organism	Homo sapiens
Samples (51)	Click for detailed sample information and links to data
Protocols (4)	Click for detailed protocol information
Description	RNA-seq was performed on esophageal adenocarcinoma (EAC), Barrett's without dysplasia, Barrett's with low-grade dysplasia (LGD) and normal squamous esophagus tissue to find early alterations in the transcriptome level turning Barrett's dysplastic.
Experiment types	RNA-seq of coding RNA, disease state design
Contact	Jesper LV Maag <j.maag@garvan.org.au>
Citation	Novel Aberrations Uncovered in Barrett's Esophagus and Esophageal Adenocarcinoma Using Whole Transcriptome Sequencing. Maag J, Fisher OM, Levert-Mignon AJ, Kaczorowski DC, Thomas ML, Hussey D, Watson D, Wettstein A, Bobryshev YV, Edwards M, Dinger ME, Lord RV. , Europe PMC 28751461
MINSEQE	* Exp. design – Protocols – Variables – Processed – Seq. reads
Files	Investigation description E-MTAB-4054.idf.txt Sample and data relationship E-MTAB-4054.sdrf.txt Click to browse all available files
Links	ENA - ERP013206 Send E-MTAB-4054 data to GENOMESPACE

Let's try ArrayExpress

E-MTAB-4054 - Whole transcriptome profiling of Esophageal adenocarcinoma and Barrett's

Display summary

Export table in Tab-delimited format

Page 1 2 3 4 5 6 Showing 76 - 100 of 126 rows Page size 25 50 100 250 500

Source Name	Protocol REF	Performer	Assay Name	Comment[technical replicate group]	Technology Type	Comment[ENA_EXP
N09	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N09_R		sequencing assay	ERX1220581
N10	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N10_R		sequencing assay	ERX1220582
N10	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N10_R		sequencing assay	ERX1220582
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_Human_CCGTCC_R	group1	sequencing assay	ERX1220524
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_Human_CCGTCC_R	group1	sequencing assay	ERX1220524
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_R	group1	sequencing assay	ERX1220525
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_R	group1	sequencing assay	ERX1220525
LNC14	P-MTAB-47385	Dominik C. Kaczorowski	LNC_14_Human_GTGAAA_R	group10	sequencing assay	ERX1220566
[illegible]	[illegible]	[illegible]	[illegible]	group10	sequencing assay	ERX1220566
[illegible]	[illegible]	[illegible]	[illegible]	group10	sequencing assay	ERX1220567
[illegible]	[illegible]	[illegible]	[illegible]	group10	sequencing assay	ERX1220567
[illegible]	[illegible]	[illegible]	[illegible]	group11	sequencing assay	ERX1220575
[illegible]	[illegible]	[illegible]	[illegible]	group11	sequencing assay	ERX1220575
[illegible]	[illegible]	[illegible]	[illegible]	group11	sequencing assay	ERX1220576
[illegible]	[illegible]	[illegible]	[illegible]	group11	sequencing assay	ERX1220576
[illegible]	[illegible]	[illegible]	[illegible]	group12	sequencing assay	ERX1220579
N08	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N08_Human_CCGTCC_R	group12	sequencing assay	ERX1220579
N08	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N08_R	group12	sequencing assay	ERX1220580
N08	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N08_R	group12	sequencing assay	ERX1220580
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_Human_ATGTCA_R	group2	sequencing assay	ERX1220526
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_Human_ATGTCA_R	group2	sequencing assay	ERX1220526
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_R	group2	sequencing assay	ERX1220527
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_R	group2	sequencing assay	ERX1220527
B07	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B07_Human_GATCAG_R	group3	sequencing assay	ERX1220528
B07	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B07_Human_GATCAG_R	group3	sequencing assay	ERX1220528

12 samples (2 normal, 5 Barrett's esophagus without dysplasia and 5 esophageal adenocarcinoma)

Let's try ArrayExpress

E-MTAB-4054 - Whole transcriptome profiling of Esophageal adenocarcinoma and Barrett's

Display summary

Export table in Tab-delimited format

Page 1 2 3 4 5 6 Showing 76 - 100 of 126 rows Page size 25 50 100 250 500

Source Name	Protocol REF	Performer	Assay Name	Comment[technical replicate group]	Technology Type	Comment[ENA_EXP
N09	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N09_R		sequencing assay	ERX1220581
N10	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N10_R		sequencing assay	ERX1220582
N10	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N10_R		sequencing assay	ERX1220582
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_Human_CCGTCC_R	group1	sequencing assay	ERX1220524
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_Human_CCGTCC_R	group1	sequencing assay	ERX1220524
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_R	group1	sequencing assay	ERX1220525
B05	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B05_R	group1	sequencing assay	ERX1220525
LNC14	P-MTAB-47385	Do			equencing assay	ERX1220566
LNC14	P-MTAB-47385	Do			equencing assay	ERX1220566
LNC14	P-MTAB-47385	Do			equencing assay	ERX1220567
LNC14	P-MTAB-47385	Dominik C. Kaczorowski	LNC_14_R	group10	sequencing assay	ERX1220567
N04	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N04_Human_AGTTCC_R	group11	sequencing assay	ERX1220575
N04	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N04_Human_AGTTCC_R	group11	sequencing assay	ERX1220575
N04	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N04_R	group11	sequencing assay	ERX1220576
N04	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N04_R	group11	sequencing assay	ERX1220576
N08	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N08_Human_CCGTCC_R	group12	sequencing assay	ERX1220579
N08	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N08_Human_CCGTCC_R	group12	sequencing assay	ERX1220579
N08	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N08_R	group12	sequencing assay	ERX1220580
N08	P-MTAB-47385	Dominik C. Kaczorowski	LNC_N08_R	group12	sequencing assay	ERX1220580
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_Human_ATGTCA_R	group2	sequencing assay	ERX1220526
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_Human_ATGTCA_R	group2	sequencing assay	ERX1220526
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_R	group2	sequencing assay	ERX1220527
B06	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B06_R	group2	sequencing assay	ERX1220527
B07	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B07_Human_GATCAG_R	group3	sequencing assay	ERX1220528
B07	P-MTAB-47385	Dominik C. Kaczorowski	LNC_B07_Human_GATCAG_R	group3	sequencing assay	ERX1220528

12 samples were sequenced twice

Let's try ArrayExpress

Disease cohort	Number of samples	Number of technical replicates
normal	17	2
Barrett's esophagus non-dysplastic	14	5
Barrett's esophagus low-grade dysplasia	8	0
esophageal adenocarcinoma	12	5
Total number of samples	**51**	12
Number of files (paired-end libraries)	102	24
Total number of files		**126**

EMBL-EBI Bioinformatics resources for exploring functional genomics data

Discover functional genomics data with ArrayExpress

Laura Huerta, PhD

Senior Scientific Curator

lauhuema@ebi.ac.uk

9 November 2017

EMBL-EBI