RNA-seq data in ArrayExpress and recent database improvements

Gabriella Rustici
Functional Genomics Group

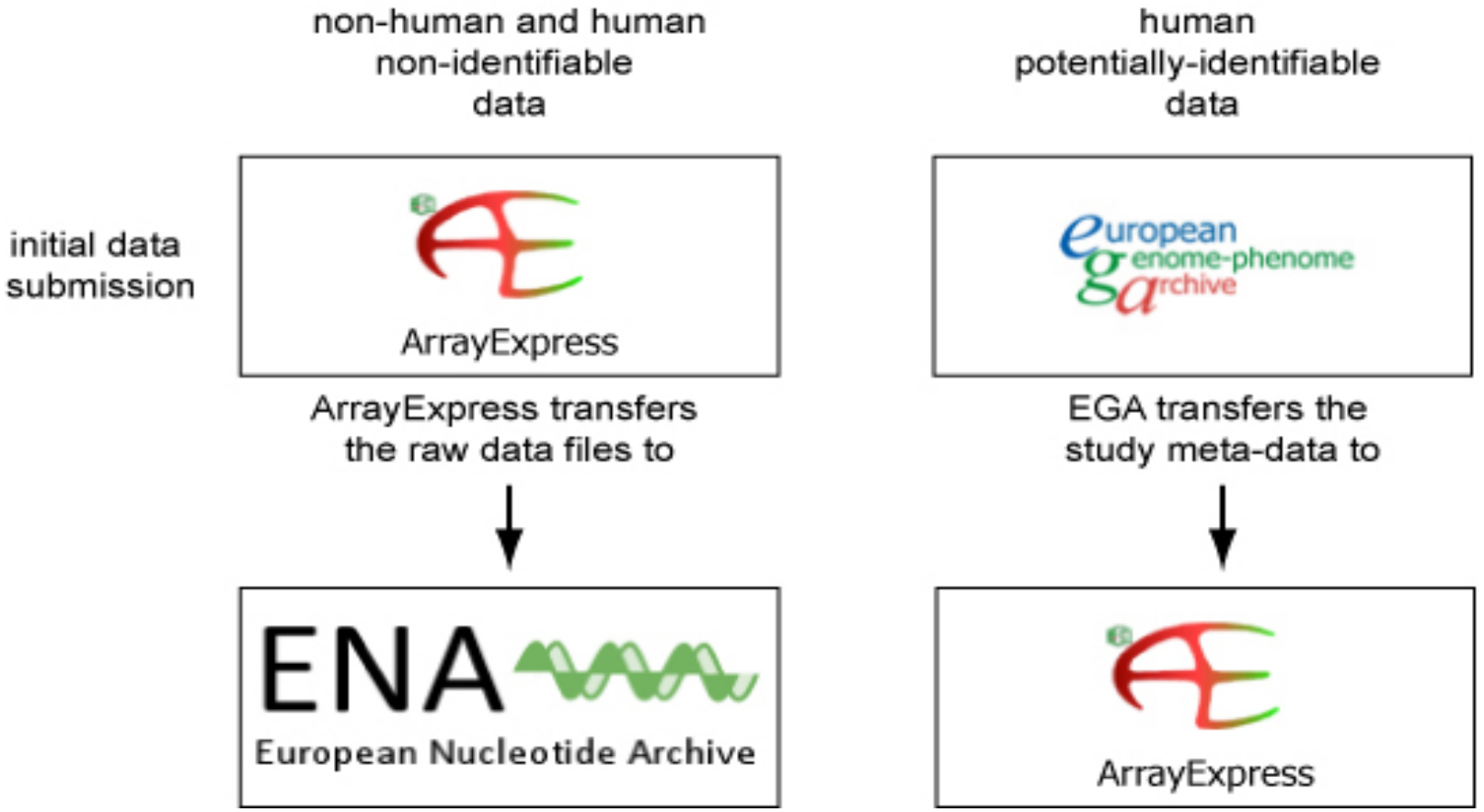
EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

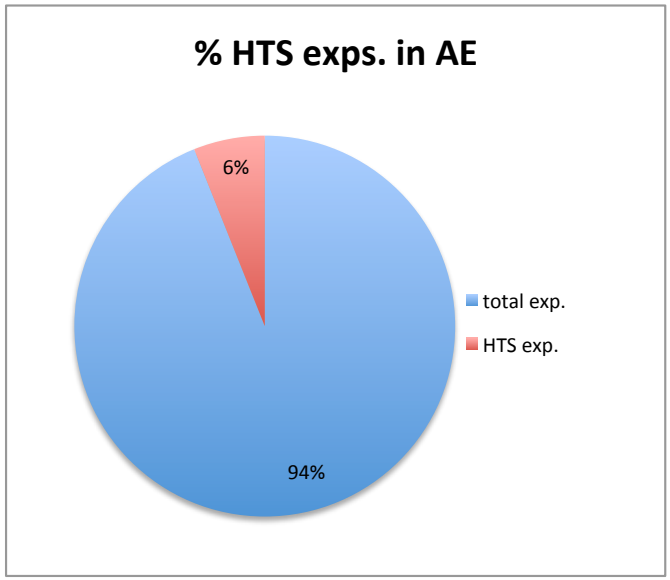
HTS data in ArrayExpress and Atlas

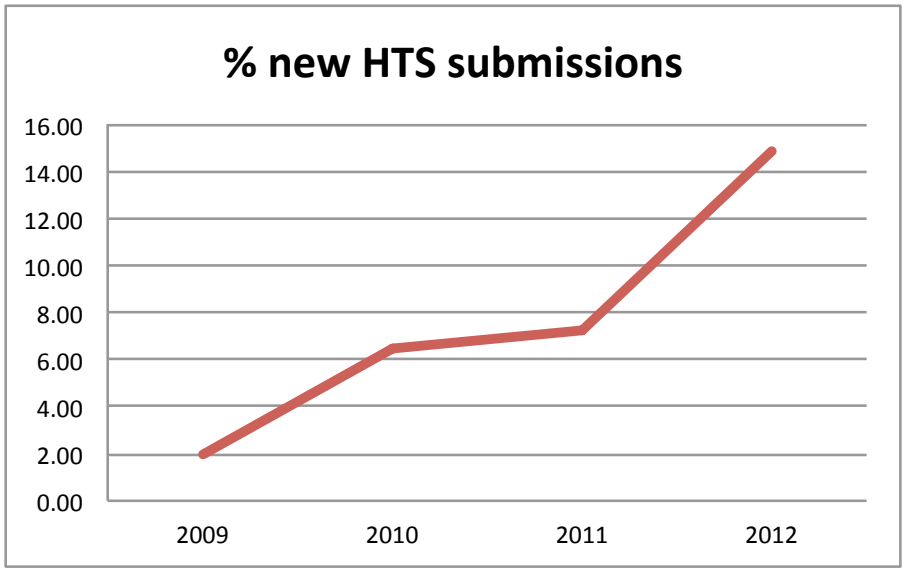
- We archive functional genomics data derived from HTS experiments in the ArrayExpress Archive
- Focus on RNA-seq and ChIP-seq experiments
- We are developing robust internal pipelines for RNA-seq data processing and quality assessment
- We display gene expression changes measured by RNA-seq in the Expression Atlas

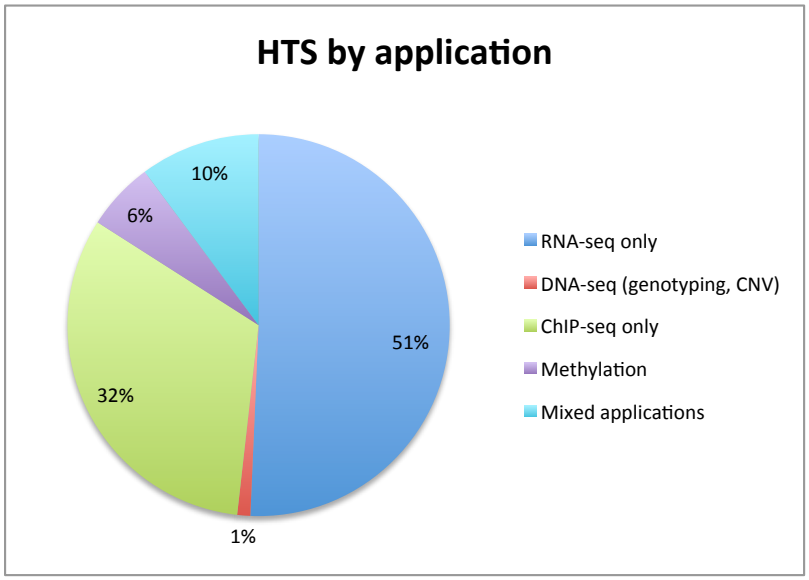
Types of data that can be submitted to AE



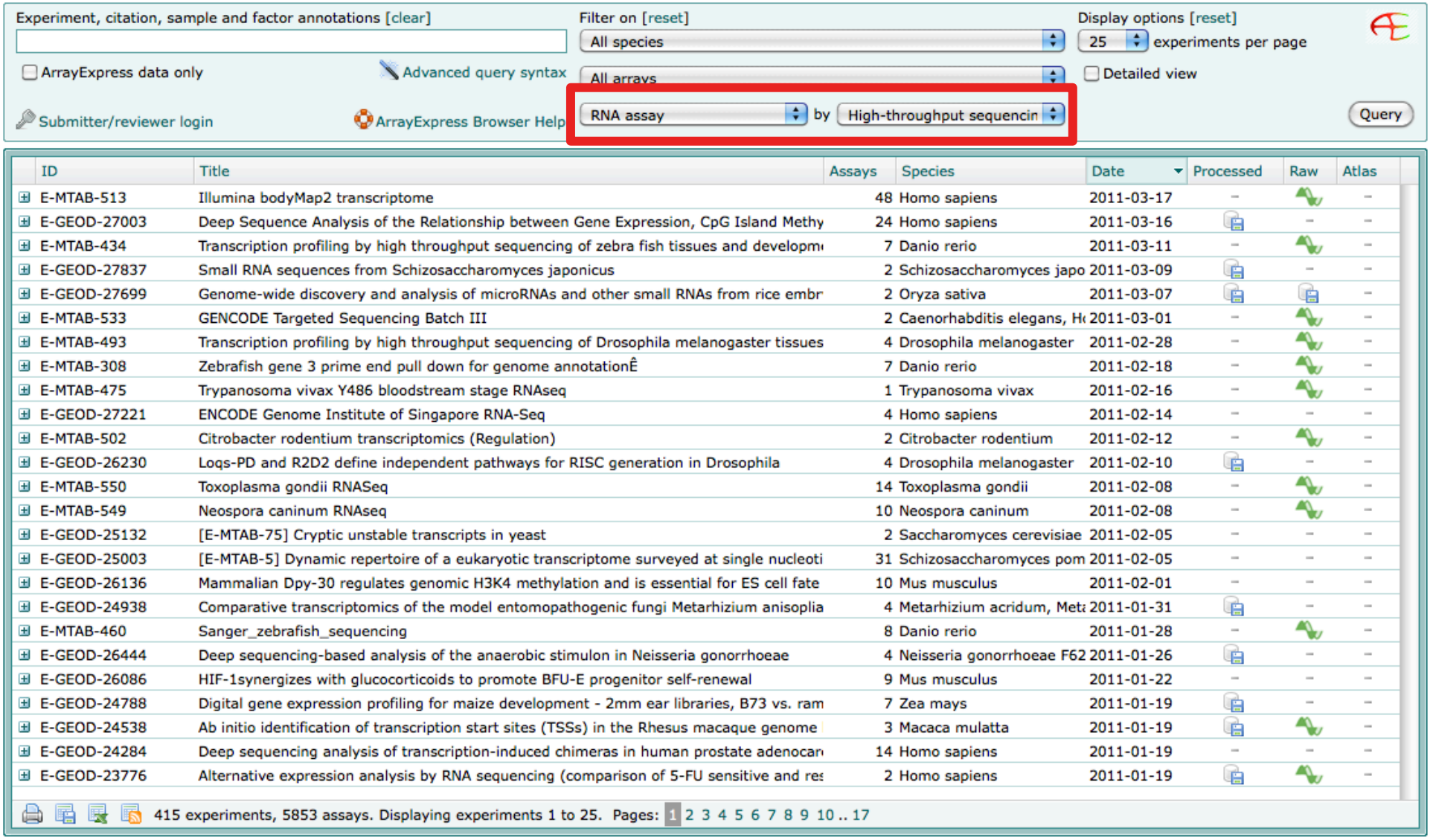
Some stats


% HTS exps. in AE
6%
94%
total exp.
HTS exp.


% new HTS submissions
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
2009
2010
2011
2012


HTS by application
10%
6%
51%
32%
1%
RNA-seq only
DNA-seq (genotyping, CNV)
ChIP-seq only
Methylation
Mixed applications

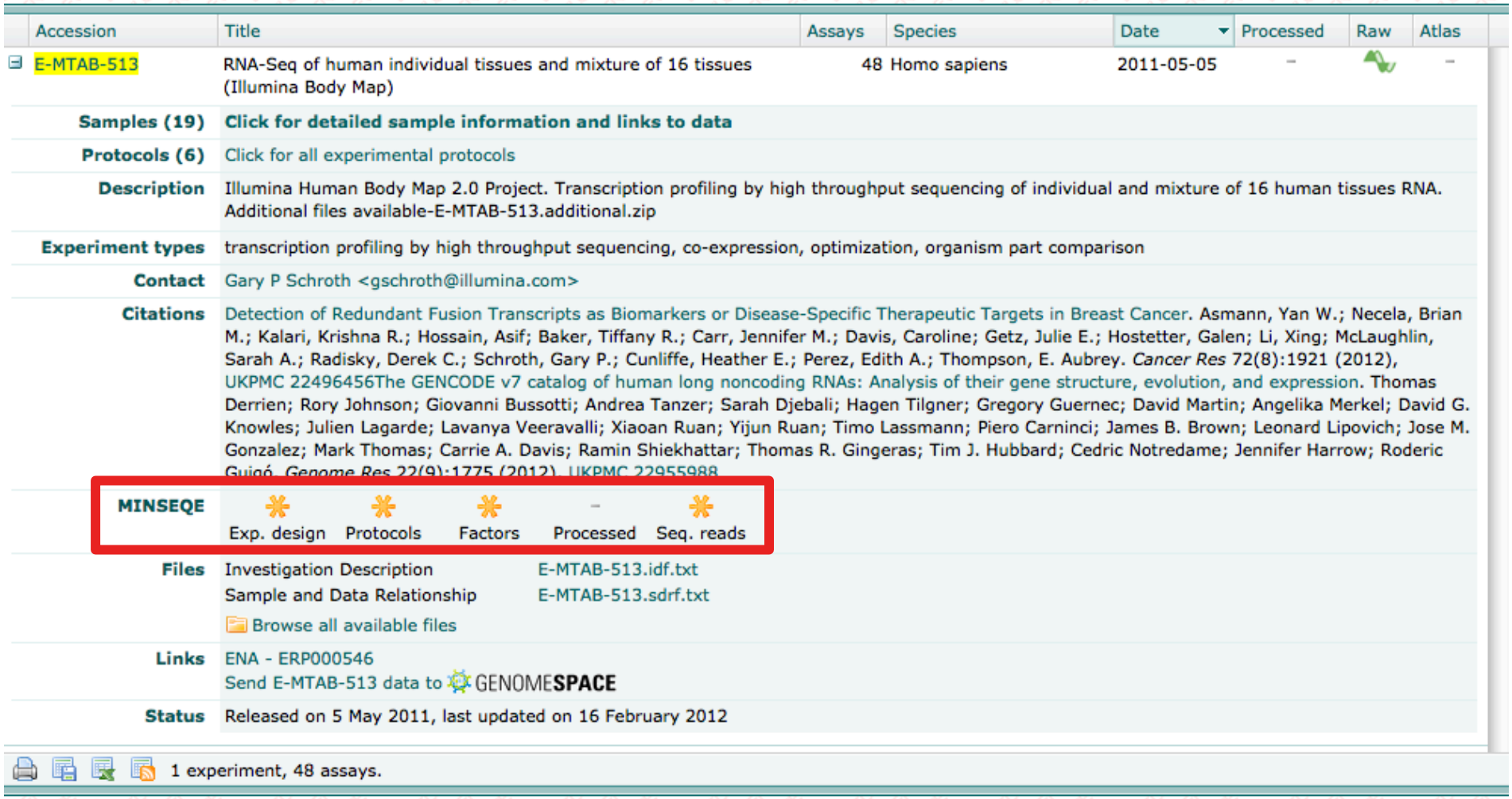
RNA-seq data in ArrayExpress



ID	Title	Assays	Species	Date	Processed	Raw	Atlas
E-MTAB-513	Illumina bodyMap2 transcriptome	48	Homo sapiens	2011-03-17	–		–
E-GEOD-27003	Deep Sequence Analysis of the Relationship between Gene Expression, CpG Island Methy	24	Homo sapiens	2011-03-16		–	–
E-MTAB-434	Transcription profiling by high throughput sequencing of zebra fish tissues and developm	7	Danio rerio	2011-03-11	–		–
E-GEOD-27837	Small RNA sequences from Schizosaccharomyces japonicus	2	Schizosaccharomyces japo	2011-03-09		–	–
E-GEOD-27699	Genome-wide discovery and analysis of microRNAs and other small RNAs from rice embr	2	Oryza sativa	2011-03-07			–
E-MTAB-533	GENCODE Targeted Sequencing Batch III	2	Caenorhabditis elegans, H	2011-03-01	–		–
E-MTAB-493	Transcription profiling by high throughput sequencing of Drosophila melanogaster tissues	4	Drosophila melanogaster	2011-02-28	–		–
E-MTAB-308	Zebrafish gene 3 prime end pull down for genome annotationÊ	7	Danio rerio	2011-02-18	–		–
E-MTAB-475	Trypanosoma vivax Y486 bloodstream stage RNAseq	1	Trypanosoma vivax	2011-02-16	–		–
E-GEOD-27221	ENCODE Genome Institute of Singapore RNA-Seq	4	Homo sapiens	2011-02-14	–	–	–
E-MTAB-502	Citrobacter rodentium transcriptomics (Regulation)	2	Citrobacter rodentium	2011-02-12	–		–
E-GEOD-26230	Loqs-PD and R2D2 define independent pathways for RISC generation in Drosophila	4	Drosophila melanogaster	2011-02-10		–	–
E-MTAB-550	Toxoplasma gondii RNASeq	14	Toxoplasma gondii	2011-02-08	–		–
E-MTAB-549	Neospora caninum RNAseq	10	Neospora caninum	2011-02-08	–		–
E-GEOD-25132	[E-MTAB-75] Cryptic unstable transcripts in yeast	2	Saccharomyces cerevisiae	2011-02-05	–	–	–
E-GEOD-25003	[E-MTAB-5] Dynamic repertoire of a eukaryotic transcriptome surveyed at single nucleoti	31	Schizosaccharomyces pom	2011-02-05	–	–	–
E-GEOD-26136	Mammalian Dpy-30 regulates genomic H3K4 methylation and is essential for ES cell fate	10	Mus musculus	2011-02-01	–	–	–
E-GEOD-24938	Comparative transcriptomics of the model entomopathogenic fungi Metarhizium anisoplia	4	Metarhizium acridum, Meta	2011-01-31		–	–
E-MTAB-460	Sanger_zebrafish_sequencing	8	Danio rerio	2011-01-28	–		–
E-GEOD-26444	Deep sequencing-based analysis of the anaerobic stimulon in Neisseria gonorrhoeae	4	Neisseria gonorrhoeae F62	2011-01-26		–	–
E-GEOD-26086	HIF-1synergizes with glucocorticoids to promote BFU-E progenitor self-renewal	9	Mus musculus	2011-01-22	–	–	–
E-GEOD-24788	Digital gene expression profiling for maize development - 2mm ear libraries, B73 vs. ram	7	Zea mays	2011-01-19		–	–
E-GEOD-24538	Ab initio identification of transcription start sites (TSSs) in the Rhesus macaque genome	3	Macaca mulatta	2011-01-19			–
E-GEOD-24284	Deep sequencing analysis of transcription-induced chimeras in human prostate adenocar	14	Homo sapiens	2011-01-19	–	–	–
E-GEOD-23776	Alternative expression analysis by RNA sequencing (comparison of 5-FU sensitive and res	2	Homo sapiens	2011-01-19			–

415 experiments, 5853 assays. Displaying experiments 1 to 25. Pages: 1 2 3 4 5 6 7 8 9 10 .. 17

MINSEQE score



Accession	Title	Assays	Species	Date	Processed	Raw	Atlas
E-MTAB-513	RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)	48	Homo sapiens	2011-05-05	–		–

Samples (19) Click for detailed sample information and links to data

Protocols (6) Click for all experimental protocols

Description Illumina Human Body Map 2.0 Project. Transcription profiling by high throughput sequencing of individual and mixture of 16 human tissues RNA. Additional files available-E-MTAB-513.additional.zip

Experiment types transcription profiling by high throughput sequencing, co-expression, optimization, organism part comparison

Contact Gary P Schroth <gschroth@illumina.com>

Citations Detection of Redundant Fusion Transcripts as Biomarkers or Disease-Specific Therapeutic Targets in Breast Cancer. Asmann, Yan W.; Necela, Brian M.; Kalari, Krishna R.; Hossain, Asif; Baker, Tiffany R.; Carr, Jennifer M.; Davis, Caroline; Getz, Julie E.; Hostetter, Galen; Li, Xing; McLaughlin, Sarah A.; Radisky, Derek C.; Schroth, Gary P.; Cunliffe, Heather E.; Perez, Edith A.; Thompson, E. Aubrey. *Cancer Res* 72(8):1921 (2012), UKPMC 22496456The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Thomas Derrien; Rory Johnson; Giovanni Bussotti; Andrea Tanzer; Sarah Djebali; Hagen Tilgner; Gregory Guernec; David Martin; Angelika Merkel; David G. Knowles; Julien Lagarde; Lavanya Veeravalli; Xiaoan Ruan; Yijun Ruan; Timo Lassmann; Piero Carninci; James B. Brown; Leonard Lipovich; Jose M. Gonzalez; Mark Thomas; Carrie A. Davis; Ramin Shiekhattar; Thomas R. Gingeras; Tim J. Hubbard; Cedric Notredame; Jennifer Harrow; Roderic Guigó. *Genome Res* 22(9):1775 (2012). UKPMC 22955988

MINSEQE

Exp. design	Protocols	Factors	Processed	Seq. reads
✱	✱	✱	–	✱

Files
Investigation Description E-MTAB-513.idf.txt
Sample and Data Relationship E-MTAB-513.sdrf.txt
Browse all available files

Links ENA - ERP000546
Send E-MTAB-513 data to GENOMESPACE

Status Released on 5 May 2011, last updated on 16 February 2012

1 experiment, 48 assays.

MINSEQE guidelines

- Minimal Information about a high-throughput Nucleotide SEQuencing Experiment

- The proposed guidelines for MINSEQE are:
  1. General information about the experiment and sample-data relationships
  2. The description of the biological system, samples, and the experimental variables being studied
  3. Essential experimental and data processing protocols
  4. The sequence read data for each assay
  5. The processed data for the set of assays in the study

- For the last 2 years, MINSEQE scores are no worse than MIAME scores. There may be differences between HTS and microarray submissions in terms of what is absent (e.g., processed data for HTS).

Sample-data relationship file

Accession	Title	Assays	Species	Date	Processed	Raw	Atlas
E-MTAB-513	RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)	48	Homo sapiens	2011-05-05	-		-

Samples (19) **Click for detailed sample information and links to data**

Protocols (6) Click for all experimental protocols

Description Illumina Human Body Map 2.0 Project. Transcription profiling by high throughput sequencing of individual and mixture of 16 human tissues RNA. Additional files available-E-MTAB-513.additional.zip

Experiment types transcription profiling by high throughput sequencing, co-expression, optimization, organism part comparison

Contact Gary P Schroth <gschroth@illumina.com>

Citations Detection of Redundant Fusion Transcripts as Biomarkers or Disease-Specific Therapeutic Targets in Breast Cancer. Asmann, Yan W.; Necela, Brian M.; Kalari, Krishna R.; Hossain, Asif; Baker, Tiffany R.; Carr, Jennifer M.; Davis, Caroline; Getz, Julie E.; Hostetter, Galen; Li, Xing; McLaughlin, Sarah A.; Radisky, Derek C.; Schroth, Gary P.; Cunliffe, Heather E.; Perez, Edith A.; Thompson, E. Aubrey. *Cancer Res* 72(8):1921 (2012), UKPMC 22496456The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Thomas Derrien; Rory Johnson; Giovanni Bussotti; Andrea Tanzer; Sarah Djebali; Hagen Tilgner; Gregory Guernec; David Martin; Angelika Merkel; David G. Knowles; Julien Lagarde; Lavanya Veeravalli; Xiaoan Ruan; Yijun Ruan; Timo Lassmann; Piero Carninci; James B. Brown; Leonard Lipovich; Jose M. Gonzalez; Mark Thomas; Carrie A. Davis; Ramin Shiekhattar; Thomas R. Gingeras; Tim J. Hubbard; Cedric Notredame; Jennifer Harrow; Roderic Guigó. *Genome Res* 22(9):1775 (2012), UKPMC 22955988

MINSEQE Exp. design, Protocols, Factors, - Processed, Seq. reads

Files Investigation Description E-MTAB-513.idf.txt
Sample and Data Relationship E-MTAB-513.sdrf.txt
Browse all available files

Links ENA - ERP000546
Send E-MTAB-513 data to GENOMESPACE

Status Released on 5 May 2011, last updated on 16 February 2012

1 experiment, 48 assays.

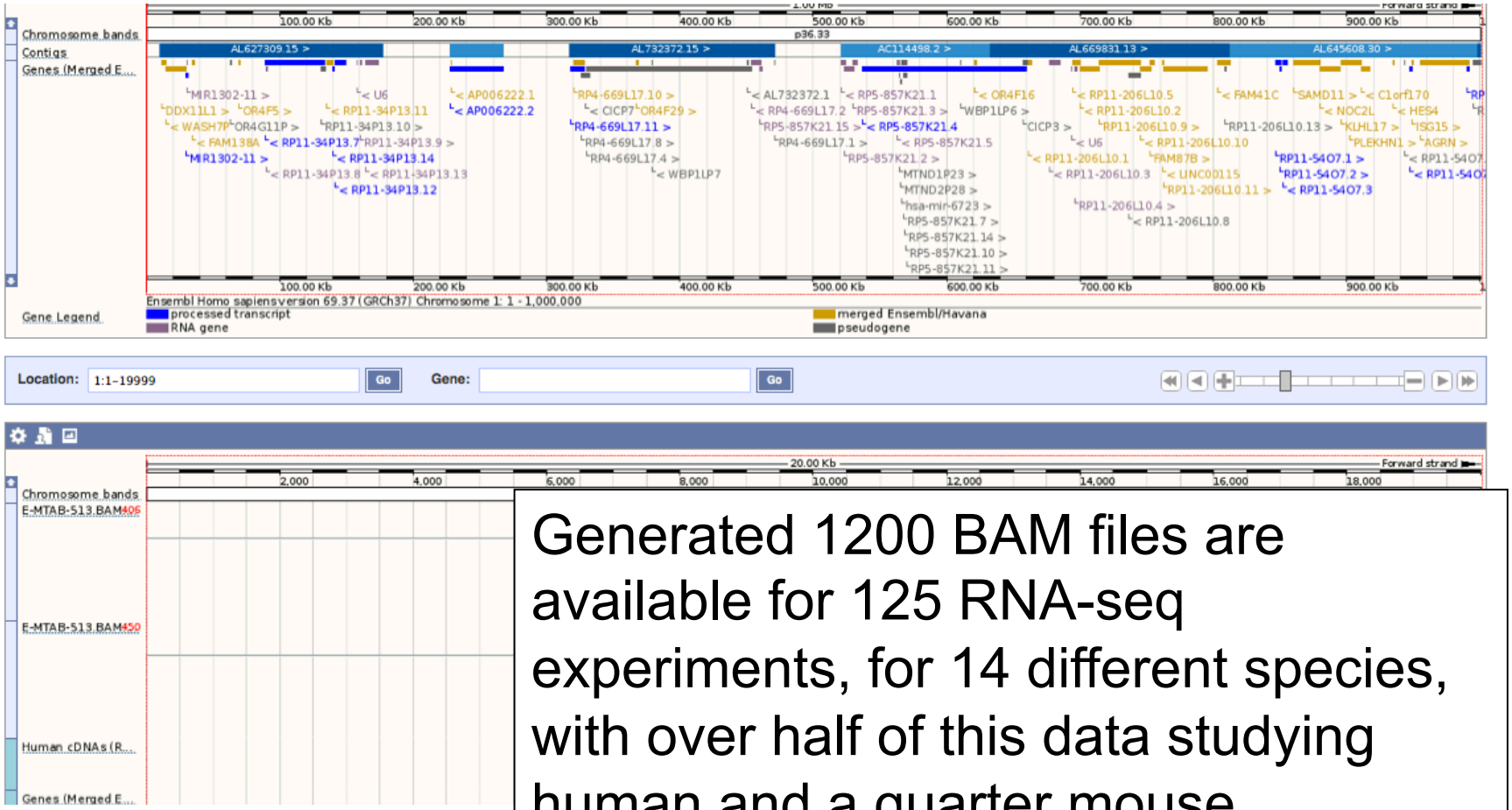
SDRF for RNA-seq experiment

Experiment E-MTAB-513

RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map) (19 samples)

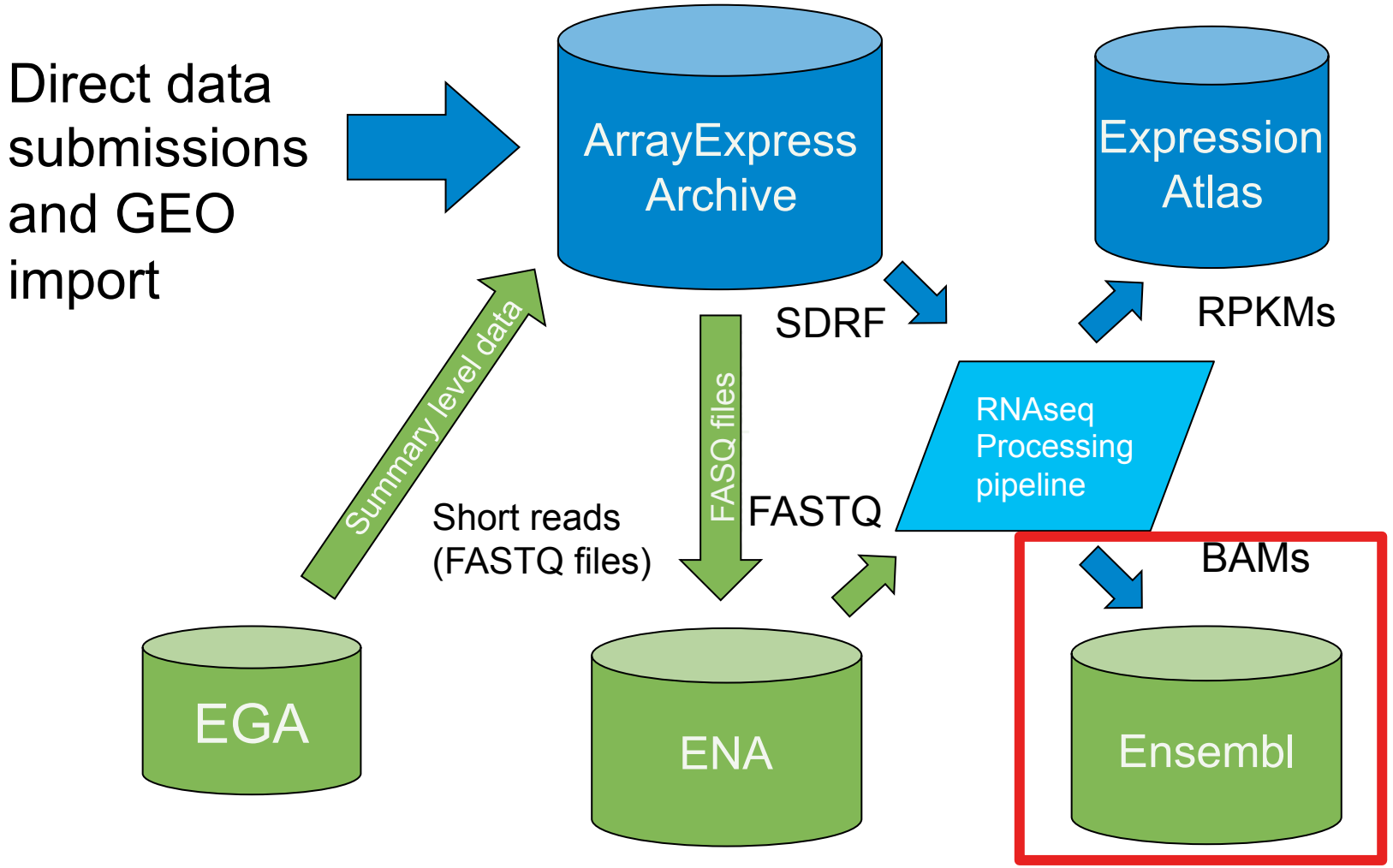
	Sample Characteristics					Factor Values		Links to Data
Source Name	Organism	Age	(unit)	OrganismPart	Sex	ORGANISMPART	LIBRARYPREP	ENA
HCT20142	Homo sapiens	60	years	kidney	female	kidney	mRNA-Seq	
HCT20142	Homo sapiens	60	years	kidney	female	kidney	mRNA-Seq	
HCT20142	Homo sapiens	60	years	kidney	female	kidney	mRNA-Seq	
HCT20143	Homo sapiens	77	years	heart	male	heart	mRNA-Seq	
HCT20143	Homo sapiens	77	years	heart	male	heart	mRNA-Seq	
HCT20143	Homo sapiens	77	years	heart	male	heart	mRNA-Seq	
HCT20144	Homo sapiens	37	years	liver	male	liver	mRNA-Seq	
HCT20144	Homo sapiens	37	years	liver	male	liver	mRNA-Seq	
HCT20144	Homo sapiens	37	years	liver	male	liver	mRNA-Seq	
HCT20145	Homo sapiens	65	years	lung	male	lung	mRNA-Seq	
HCT20145	Homo sapiens	65	years	lung	male	lung	mRNA-Seq	
HCT20145	Homo sapiens	65	years	lung	male	lung	mRNA-Seq	
HCT20146	Homo sapiens	86	years	lymph node	female	lymph node	mRNA-Seq	
HCT20146	Homo sapiens	86	years	lymph node	female	lymph node	mRNA-Seq	
HCT20146	Homo sapiens	86	years	lymph node	female	lymph node	mRNA-Seq	
HCT20147	Homo sapiens	73	years	prostate	male	prostate	mRNA-Seq	
HCT20147	Homo sapiens	73	years	prostate	male	prostate	mRNA-Seq	
HCT20147	Homo sapiens	73	years	prostate	male	prostate	mRNA-Seq	

BAM files visualization in Ensembl

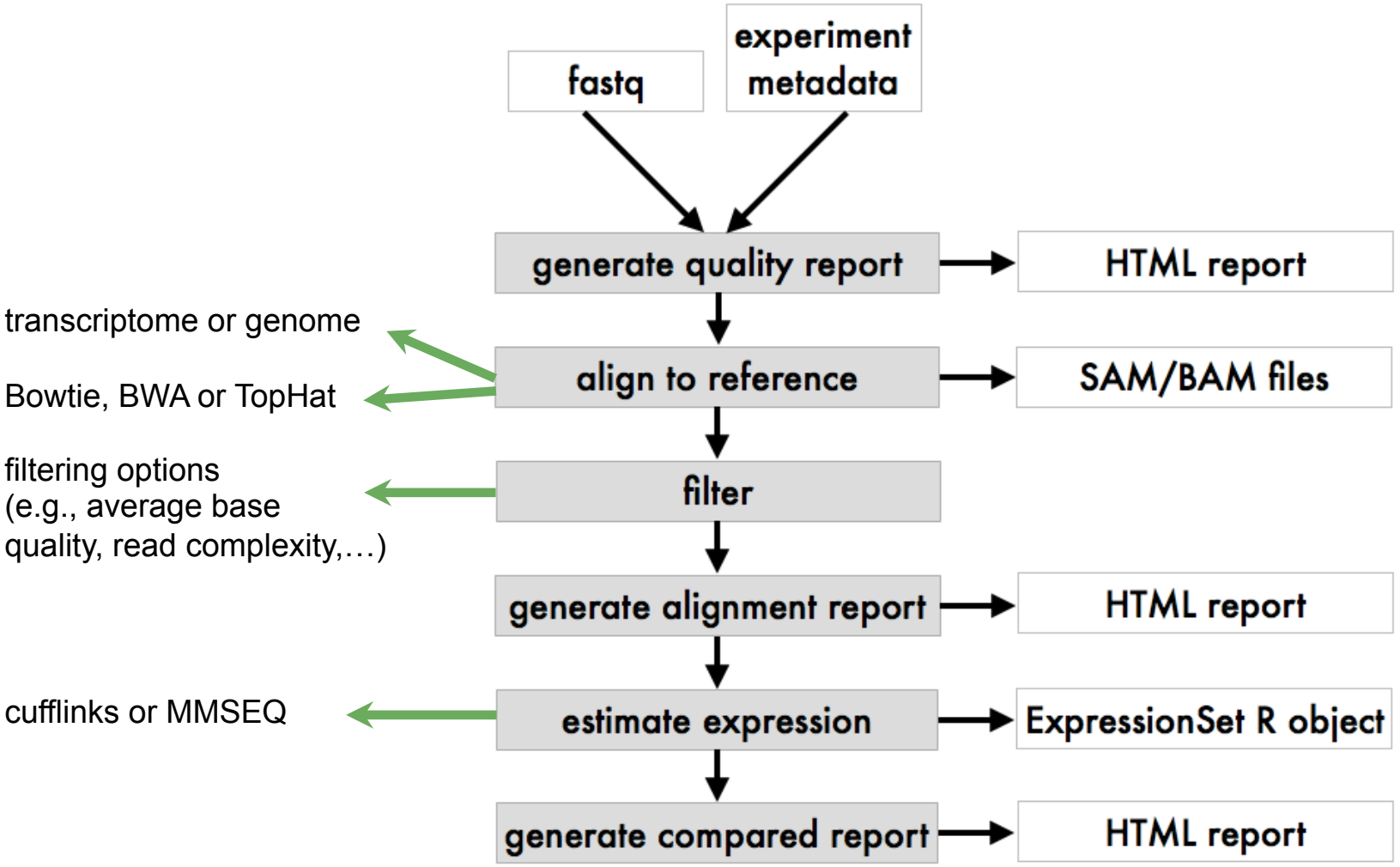


Generated 1200 BAM files are available for 125 RNA-seq experiments, for 14 different species, with over half of this data studying human and a quarter mouse

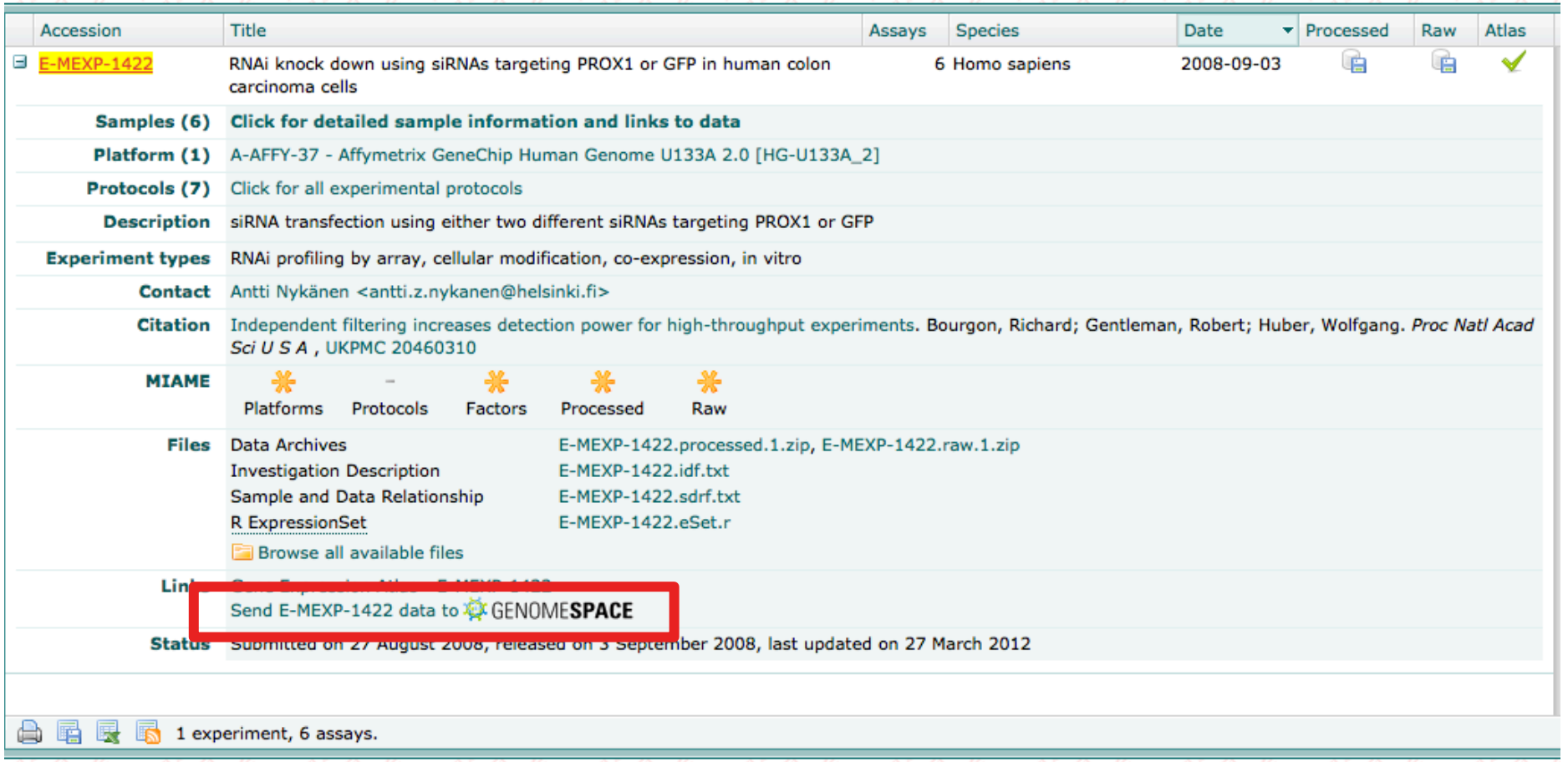
RNA-seq processing pipeline



ArrayExpressHTS pipeline



Links with data analysis tools

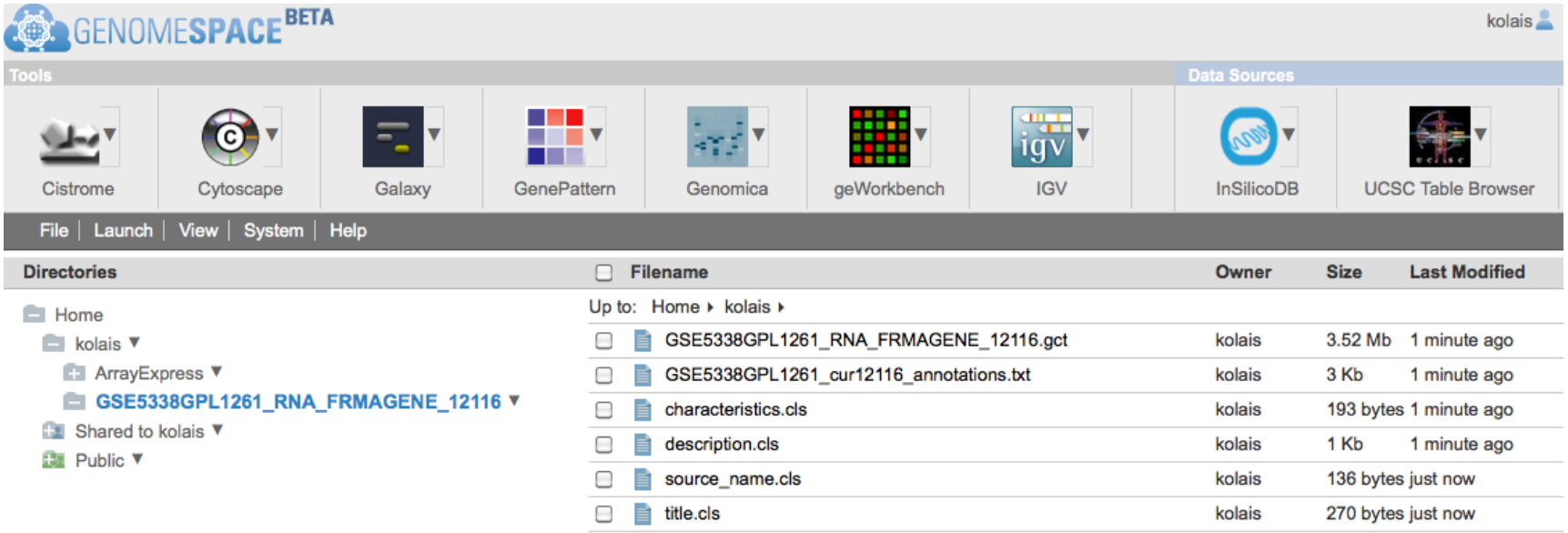


Accession	Title	Assays	Species	Date	Processed	Raw	Atlas
E-MEXP-1422	RNAi knock down using siRNAs targeting PROX1 or GFP in human colon carcinoma cells	6	Homo sapiens	2008-09-03			

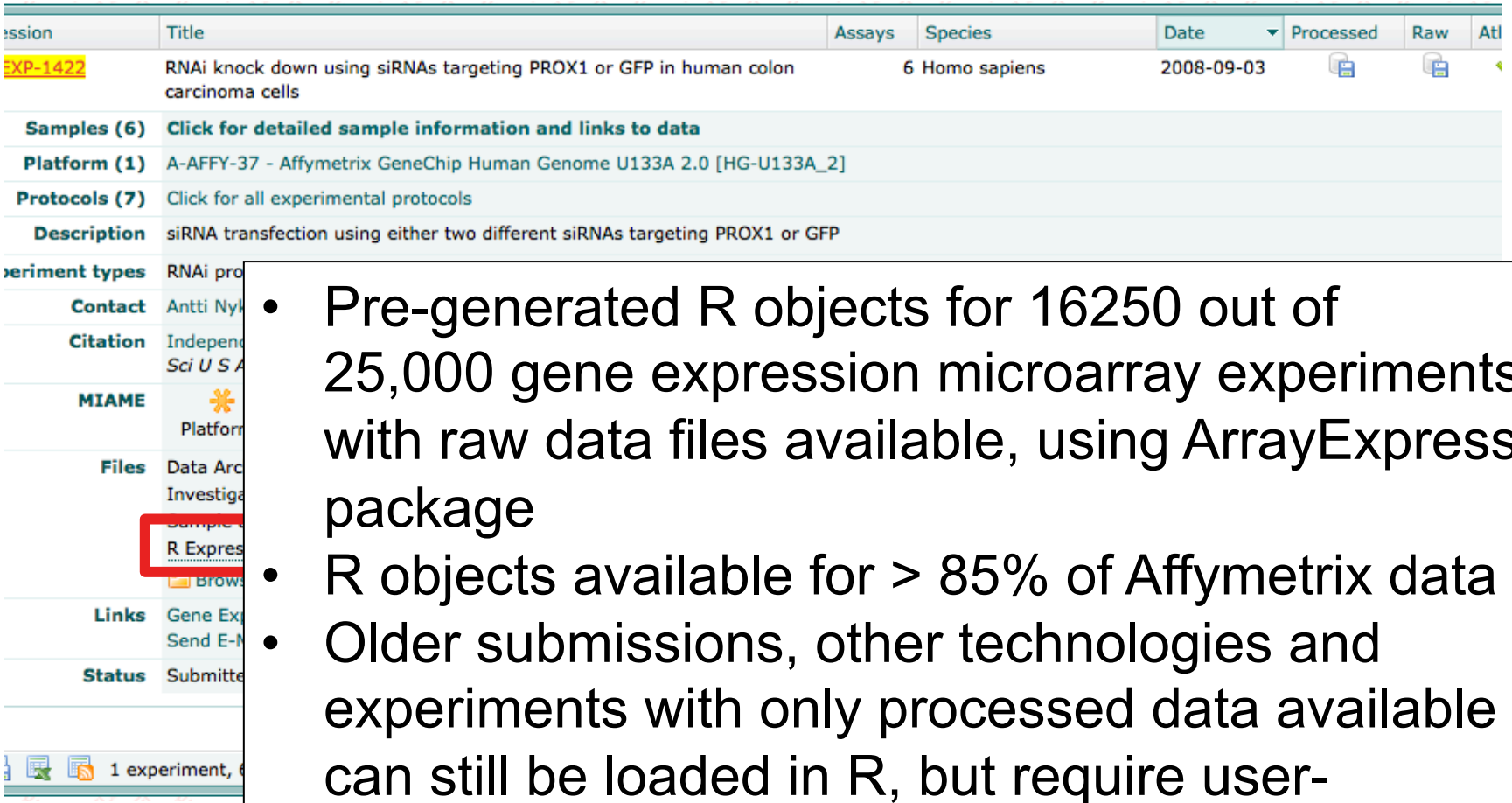
Samples (6)	**Click for detailed sample information and links to data**
Platform (1)	A-AFFY-37 - Affymetrix GeneChip Human Genome U133A 2.0 [HG-U133A_2]
Protocols (7)	Click for all experimental protocols
Description	siRNA transfection using either two different siRNAs targeting PROX1 or GFP
Experiment types	RNAi profiling by array, cellular modification, co-expression, in vitro
Contact	Antti Nykänen <antti.z.nykanen@helsinki.fi>
Citation	Independent filtering increases detection power for high-throughput experiments. Bourgon, Richard; Gentleman, Robert; Huber, Wolfgang. *Proc Natl Acad Sci U S A* , UKPMC 20460310
MIAME	Platforms – Protocols Factors Processed Raw
Files	Data Archives: E-MEXP-1422.processed.1.zip, E-MEXP-1422.raw.1.zip Investigation Description: E-MEXP-1422.idf.txt Sample and Data Relationship: E-MEXP-1422.sdrf.txt R ExpressionSet: E-MEXP-1422.eSet.r Browse all available files
Lin	Send E-MEXP-1422 data to GENOMESPACE
Status	Submitted on 27 August 2008, released on 3 September 2008, last updated on 27 March 2012

1 experiment, 6 assays.

Link with GenomeSpace



Link with Bioconductor for microarray data

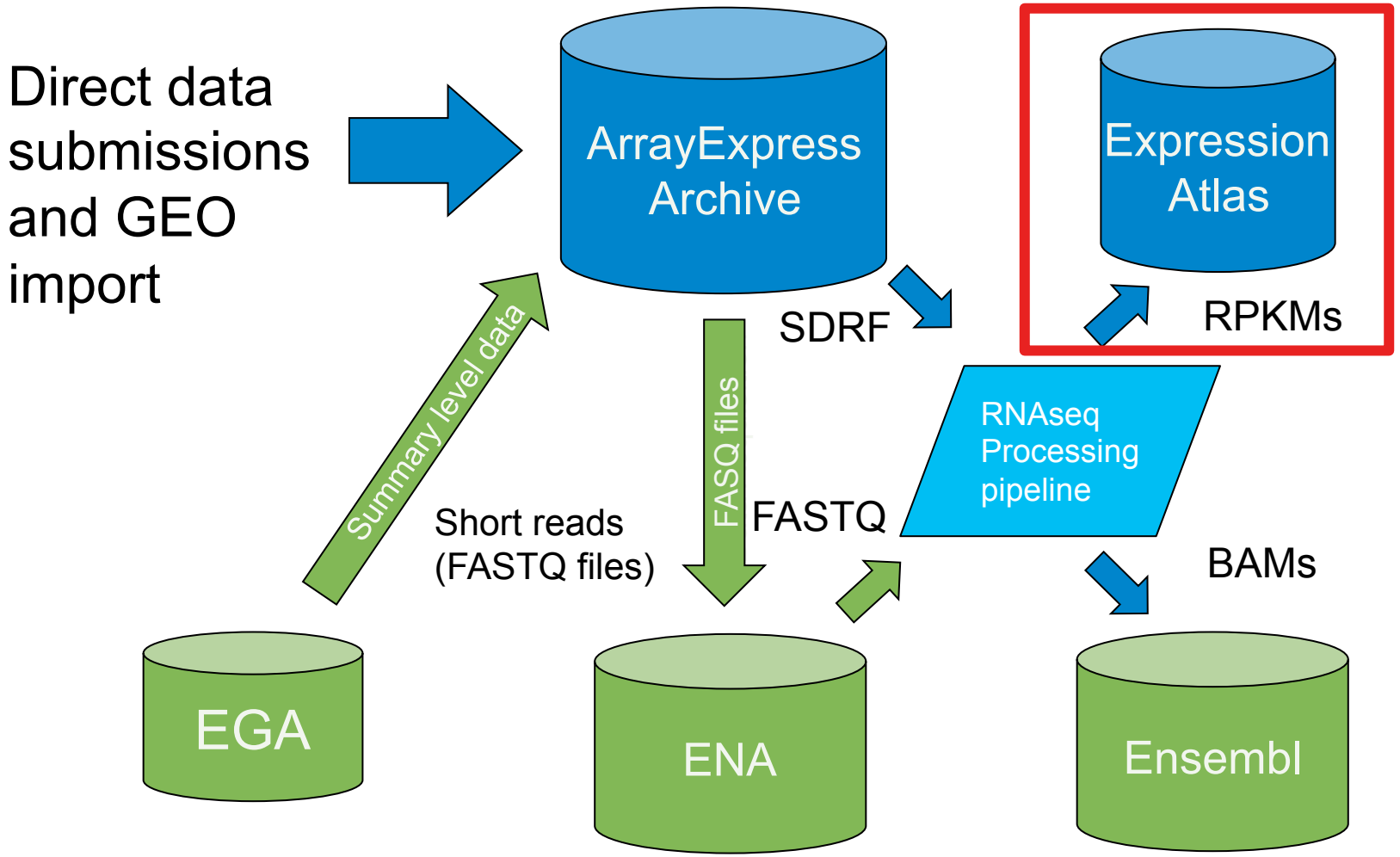


ession	Title	Assays	Species	Date	Processed	Raw	Atl
EXP-1422	RNAi knock down using siRNAs targeting PROX1 or GFP in human colon carcinoma cells	6	Homo sapiens	2008-09-03			
Samples (6)	**Click for detailed sample information and links to data**						
Platform (1)	A-AFFY-37 - Affymetrix GeneChip Human Genome U133A 2.0 [HG-U133A_2]						
Protocols (7)	Click for all experimental protocols						
Description	siRNA transfection using either two different siRNAs targeting PROX1 or GFP						
eriment types	RNAi pro						
Contact	Antti Nyk						
Citation	Indepen *Sci U S A*						
MIAME	Platfor						
Files	Data Arc Investiga R Expres Brows						
Links	Gene Ex Send E-M						
Status	Submitte						

1 experiment, (

- Pre-generated R objects for 16250 out of 25,000 gene expression microarray experiments with raw data files available, using ArrayExpress package
- R objects available for > 85% of Affymetrix data
- Older submissions, other technologies and experiments with only processed data available can still be loaded in R, but require user-specified settings

RNA-seq processing pipeline



Expression Atlas

SERPINE2 Homo sapiens

JSON XML

SERPINE2 is differentially expressed in 282 experiments: 135 organism parts: liver, bone marrow, ... (133 more);, 164 disease states: normal, glioblastoma, ... (162 more);, 11 cell types 2 cell types 62 cell types and 87 other conditions.

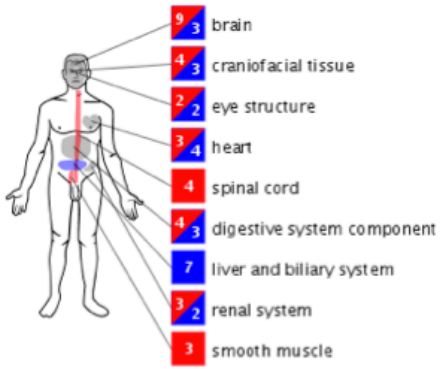
Synonyms	PI7, PN1, SERPINE2, nexin, GDN, PNI
Orthologs	SRPN6 (Anopheles gambiae) SRPN16 (Anopheles gambiae) SRPN5 (Anopheles gambiae) SRPN4 (Anopheles gambiae) srp-6 (Caenorhabditis elegans) srp-2 (Caenorhabditis elegans) srp-1 (Caenorhabditis elegans) NP_777094.1 (Bos taurus) ENSCING00000009089 (Ciona intestinalis) serpine2 (Danio rerio) SERPINE2 (Gallus gallus) Serpine2 (Mus musculus) GDN_RAT (Rattus norvegicus) srp-8 (Caenorhabditis elegans) NP_001023824.1 (Caenorhabditis elegans) Spn5 (Drosophila melanogaster) CG7219 (Drosophila melanogaster) CG6687 (Drosophila melanogaster) srp-3 (Caenorhabditis elegans) (Compare orthologs)
UniProt Accession	P07093, B4DIF2, B4DMR3, C9JN98, C9JRK5, C9JYC8, C9K031
Gene Ontology Term	cell differentiation, extracellular space, glycosaminoglycan binding, heparin binding, nervous system development...
InterPro Term	Protease inhibitor I4, serpin, Protease_inhib_I4_serpin
Search EB-eye	ENSG00000135919

⊞ Show more properties

Experimental Factors

Organism part
studied in E-AFMX-5, E-GEOD-803, E-GEOD-1577, E-MEXP-1251, E-MEXP-433, ... (31 experiments)

1 3 Number of published studies where the gene is over/under expressed compared to the gene's overall mean expression level in the study.



show expression data for **all** values of this factor>>

Cell line
studied in E-GEOD-17482, E-GEOD-3860, E-GEOD-5258, E-GEOD-21750, E-MEXP-1006, ... (37 experiments)

Factor Value	U/D	Experiments
MCF-7	4	E-GEOD-20081, E-GEOD-5823, E-GEOD-3529, E-GEOD-19639
A549	3	E-GEOD-3202, E-GEOD-6013, E-MTAB-62
Mcf7	3	E-MTAB-62, E-MTAB-37, E-GEOD-5258
PC3	3	E-GEOD-5258, E-MEXP-2034, E-GEOD-14464
SW480	3	E-MEXP-1014, E-MTAB-37, E-GEOD-18005
BT474	1 2	E-MTAB-37, E-MTAB-62, E-MEXP-440

368 more value(s).
show this factor only>>

Expression Profiles

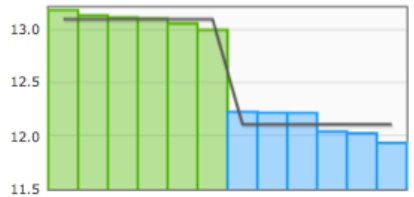
1 2 3 4 5 ... 56 57

282 experiments showing differential expression

E-MEXP-2241: Transcription profiling of human cells transfected with a miR-34a LNA inhibitor to investigate its role in oncogene-induced senescence

Experimental Factors

Compound Treatment | Genotype



ethanol ↑ 2.99 × 10 -10
4-hydroxytamoxifen ↓ 2.99 × 10 -10

Array Design: A-AFFY-44 Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133_Plus_2]

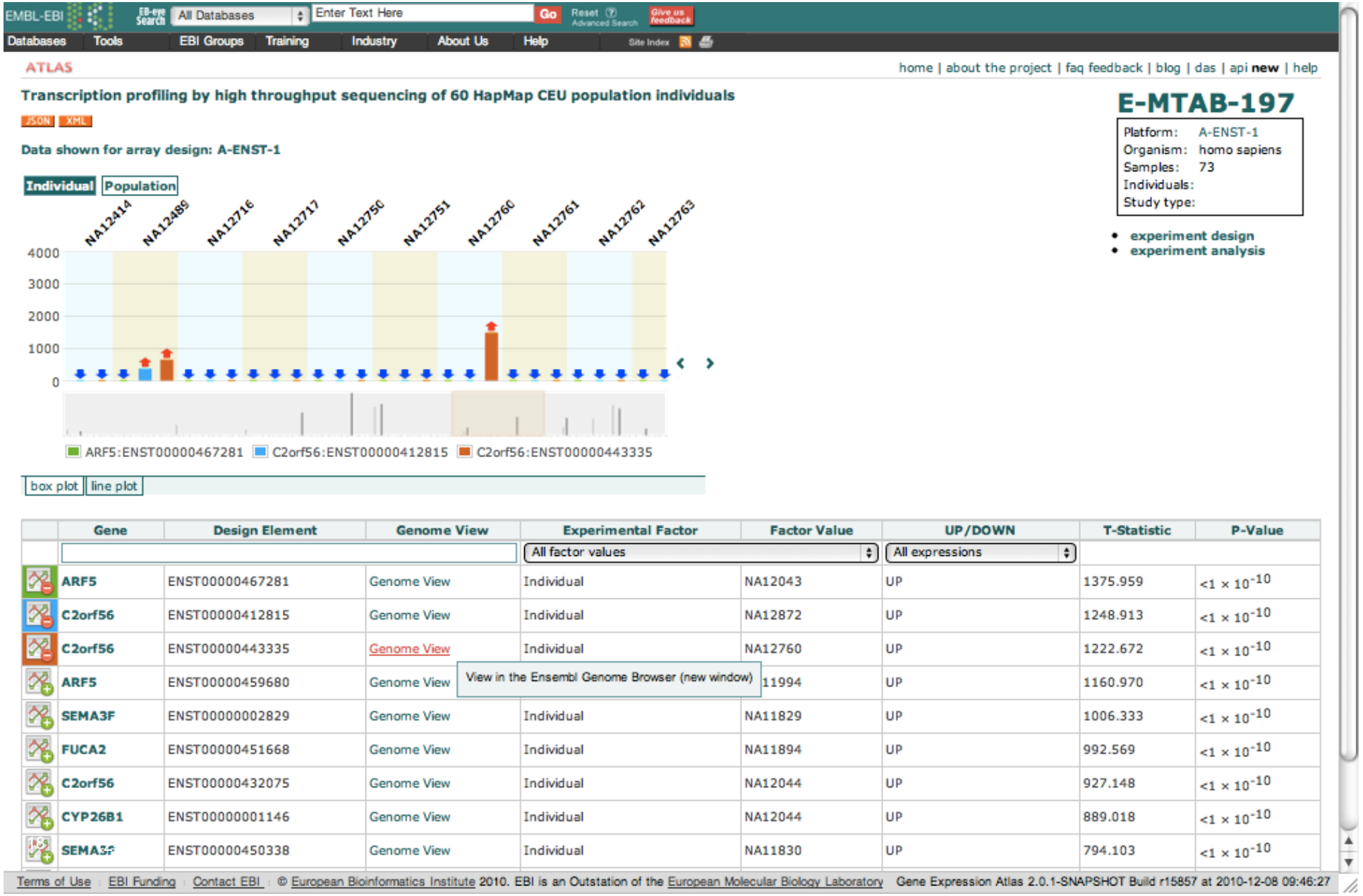
Show expression profile / experiment details

E-MEXP-2590: Transcription profiling of human colon cancer cell line SW620 treated wit tetradecylthioacetic acid (TTA) or control (NaOH).

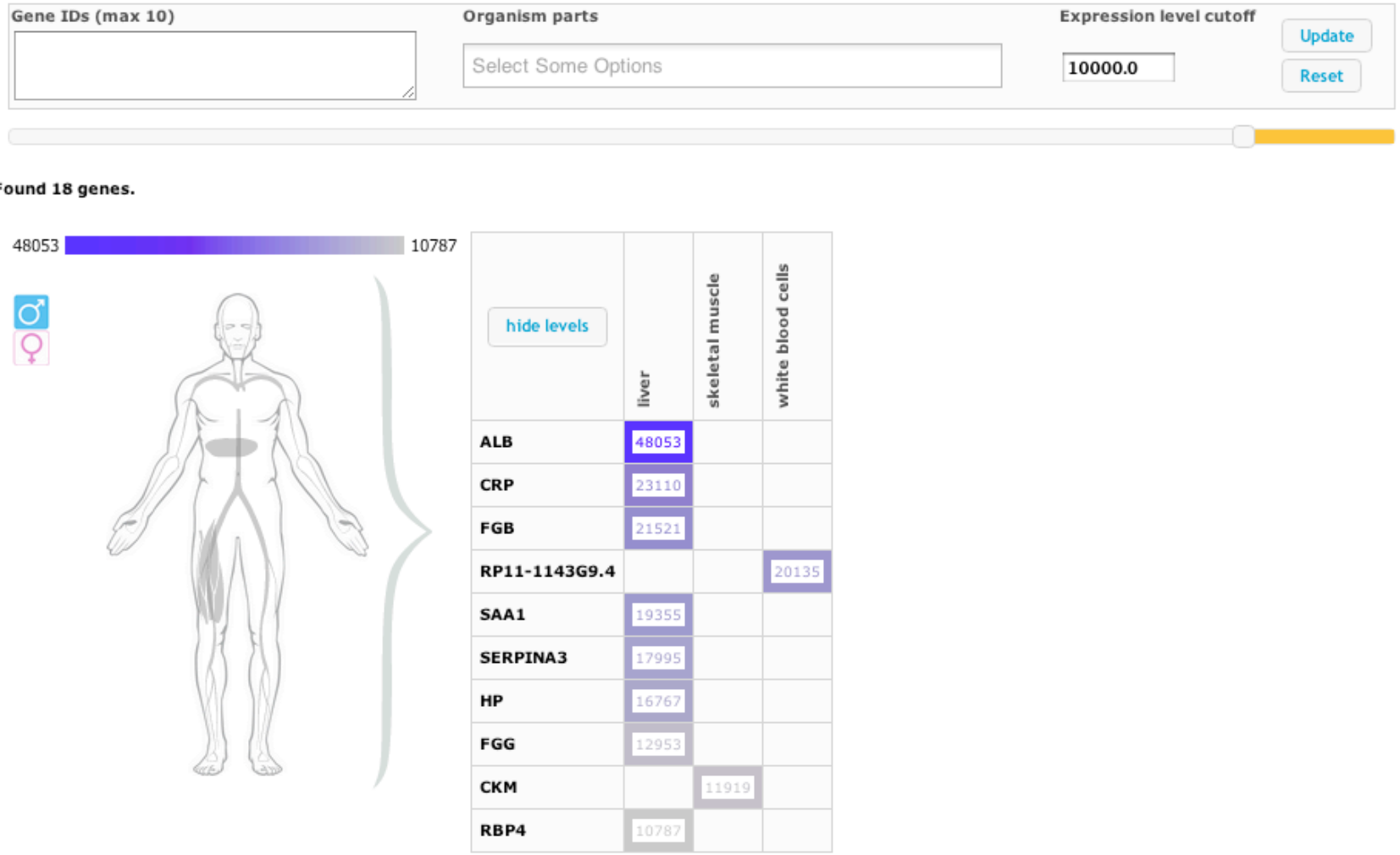
Experimental Factors

Compound Treatment | Dose

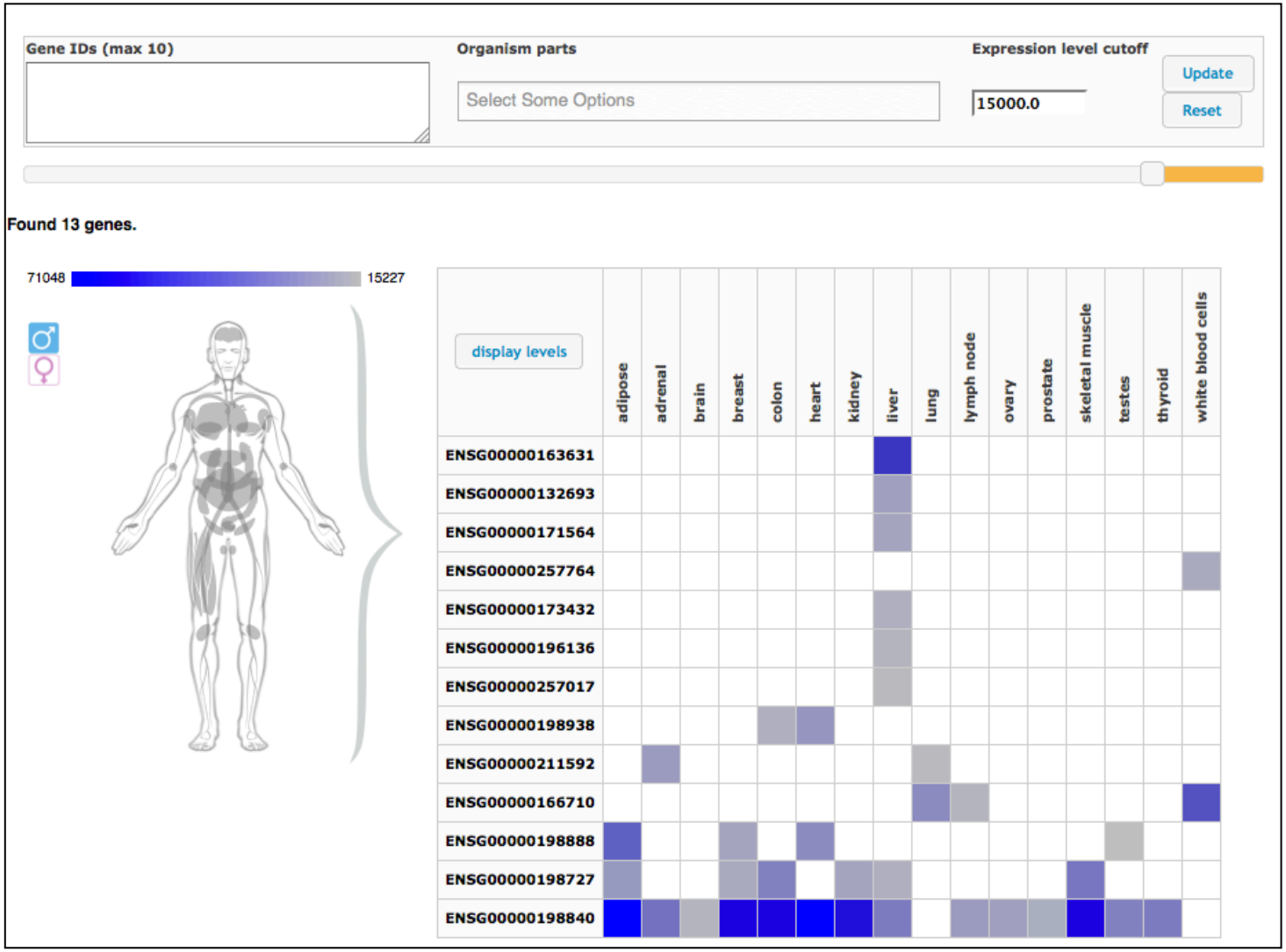
RNA-seq data in Atlas



New baseline Atlas prototype



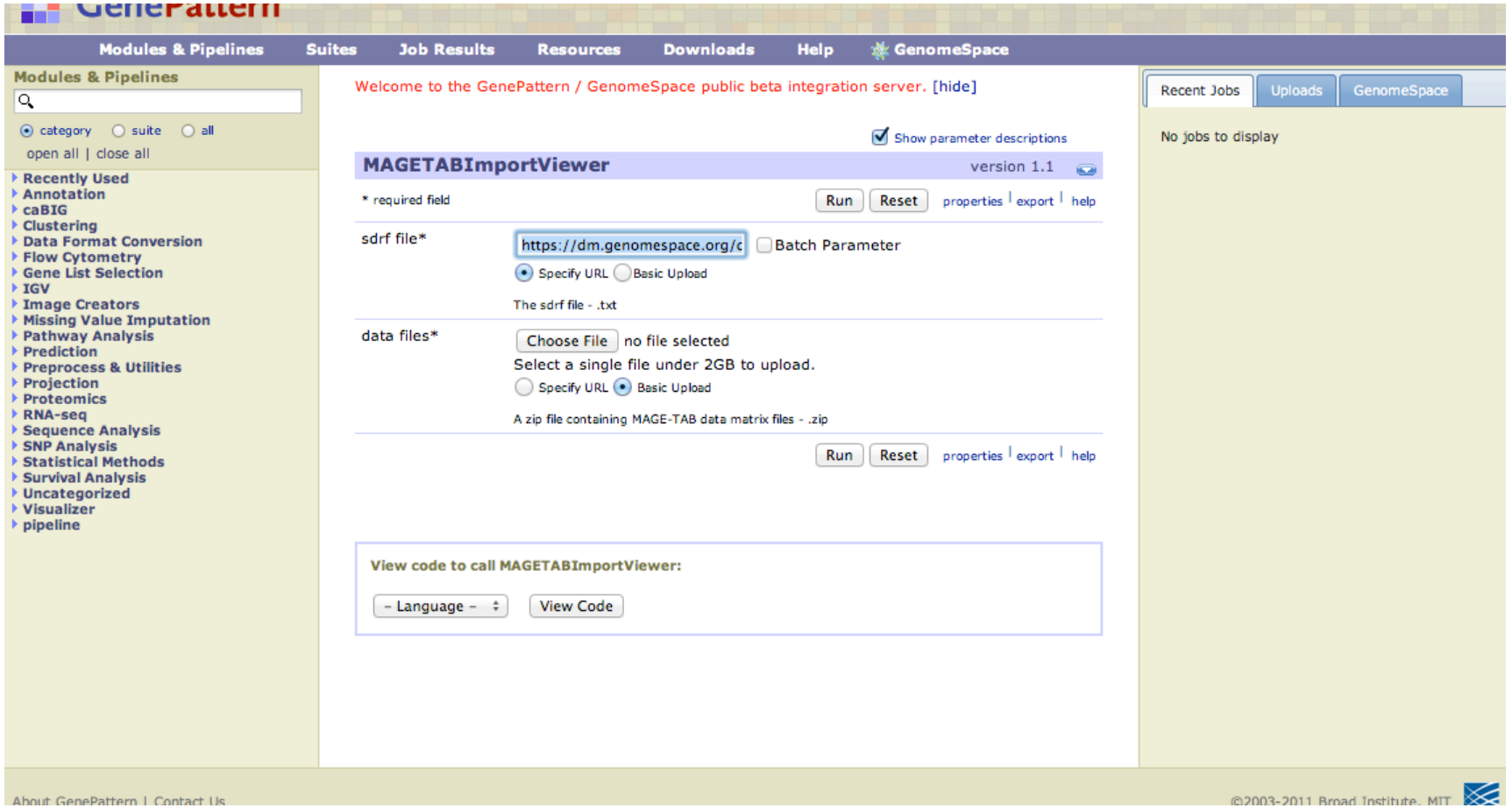
New baseline Atlas prototype



EMBL-EBI

How to improve our current pipeline?

- We are working towards understanding the effects of analyzing RNA-seq data using different tools
- Survey of different short read mapping methods: Fonseca at al, Tools for mapping high-throughput sequencing data, Bioinfomatics (2012)
- This will be followed by comparison of different quantitation methods run on top of different mappers, on a selected datasets


GenePattern
Modules & Pipelines
Suites
Job Results
Resources
Downloads
Help
GenomeSpace
Modules & Pipelines
category
suite
all
open all | close all
Recently Used
Annotation
caBIG
Clustering
Data Format Conversion
Flow Cytometry
Gene List Selection
IGV
Image Creators
Missing Value Imputation
Pathway Analysis
Prediction
Preprocess & Utilities
Projection
Proteomics
RNA-seq
Sequence Analysis
SNP Analysis
Statistical Methods
Survival Analysis
Uncategorized
Visualizer
pipeline
Welcome to the GenePattern / GenomeSpace public beta integration server. [hide]
Show parameter descriptions
MAGETABImportViewer
version 1.1
* required field
Run
Reset
properties | export | help
sdrf file*
https://dm.genomespace.org/c
Batch Parameter
Specify URL
Basic Upload
The sdrf file - .txt
data files*
Choose File
no file selected
Select a single file under 2GB to upload.
Specify URL
Basic Upload
A zip file containing MAGE-TAB data matrix files - .zip
Run
Reset
properties | export | help
View code to call MAGETABImportViewer:
– Language –
View Code
Recent Jobs
Uploads
GenomeSpace
No jobs to display
About GenePattern | Contact Us
©2003-2011 Broad Institute, MIT