

XXX Molecular atlas

Life scientists are increasingly using a range of different –omics techniques to study their subjects, integrating transcriptomics, proteomics and metabolomics data to achieve a systems-based understanding. Our ArrayExpress, Expression Atlas, PRIDE and MetaboLights resources are working towards offering a single, integrated view of –omics data.

In 2012 EMBL-EBI launched MetaboLights, a resource for metabolomics experiments and derived information. The Expression Atlas underwent a change, with Robert Petryszak stepping in to lead development of a prototype Baseline Expression Atlas, which uses sequencing-based expression data to report ‘absolute’ gene expression levels. Meanwhile, ArrayExpress accepted its millionth assay, and a number of substantial improvements were made to its user interface. The Proteomics Services Team now co-ordinates international data deposition strategies for mass spectrometry-based proteomics data through the Proteom-eXchange consortium.

ArrayExpress

The ArrayExpress Archive is a database of functional genomics experiments, including gene expression, from which you can query and download data collected to MIAME and MINSEQE standards. ArrayExpress is one of three international repositories recommended by many journals for holding microarray or RNAseq functional genomics data supporting publications.

<http://ebi.ac.uk/arrayexpress/>

Expression Atlas

The Expression Atlas is an added-value database for the reanalysis of gene expression data from EBI resources such as ArrayExpress. It shows which genes are expressed under different conditions, and how expression differs between conditions. The Expression Atlas currently holds RNA expression data from microarray or RNAseq experiments, and future development will include protein and metabolite expression data.

<http://ebi.ac.uk/gxa/>

PRIDE

The PRoteomics IDentifications database is a centralised, standards-compliant, public data repository for proteomics data. It includes protein and peptide identifications, post-translational modifications and supporting spectral evidence.

<http://ebi.ac.uk/pride/>

Metabolights

MetaboLights is a resource for Metabolomics experiments and derived information. It is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations.

<http://ebi.ac.uk/metabolights/>



Alvis Brazma

PhD in Computer Science, Moscow State University, 1987. Postdoctoral research at New Mexico State University, US.

At EMBL-EBI since 1997.

Summary of progress 2012



Ugis Sarkans

ArrayExpress

- Offered new ways of working with ArrayExpress data: R objects, integration with GenomeSpace, visualisation in Ensembl for sequencing-based data;
- Continued evolving the ArrayExpress user interface to provide a richer, more consistent service.



Helen Parkinson

ArrayExpress and Expression Atlas

- Produced EMBL-EBI's first RDF endpoint for gene expression data;
- Released Zooma, an automated annotation knowledgebase supporting ArrayExpress, the EBI BioSamples Database and the Expression Atlas;
- Welcomed the one millionth assay to ArrayExpress.



Henning Hermjakob

PRIDE

- Achieved full production status for the ProteomeXchange consortium, co-ordinating international data deposition strategies for mass spectrometry-based proteomics data;
- Handled over 80 submissions and more than 70 million new spectra;
- Released format-specific libraries (Griss et al., 2012; Reisinger et al., 2012);
- Released an updated deposition tool, PRIDE Converter (Côté et al., 2012);
- Released data analysis support, PRIDE Inspector (Wang et al., 2012);
- Implemented the capability to deposit large raw datasets in PRIDE.



Christoph Steinbeck

Metabolights

- Developed the MetaboLights database and archive, which was released in February 2012;
- Assumed co-ordination of COSMOS (European Commission co-ordination action) for metabolomics standards and data exchange.

Functional genomics

The Functional Genomics team provides bioinformatics services and conducts research in gene expression and high-throughput sequencing applications.

We participate in software development related to biomedical informatics and systems microscopy and are responsible for the Expression Atlas, which is growing to include proteomics and metabolomics data. Together with our Production and Development teams, we develop ArrayExpress, the archive of functional genomics data and the EBI BioSamples Database. We also contribute substantially to training in transcriptomics and the use of EMBL-EBI bioinformatics tools.

Our research efforts centre on developing new methods and algorithms for analysing gene expression data and integrating different types of data across multiple platforms. We are particularly interested in cancer genomics and transcript isoform usage, and collaborate closely with the Marioni group and others throughout EMBL.

Major achievements

The co-ordination of Expression Atlas development was taken over by Robert Petryszak, following the departure of Misha Kaphushesky in early 2012. The major focus was on planning and prototype development for the baseline Expression Atlas, which utilises high-throughput-sequencing-based expression data to report absolute (rather than relative) gene expression levels. We also started working on the close integration of gene-expression, protein-expression and metabolite-measurement data.

Major developments in the ArrayExpress Archive and the EBI BioSamples Database are described by Helen Parkinson and Ugis Sarkans (see also Rustici et al., 2013 and Gostev et al., 2012).

In 2012 we organised and participated in over 25 training events, including Bioinformatics Roadshows and on-site courses. These included the EMBO practical course on the analysis of high-throughput sequencing data, which was the most popular and oversubscribed training event in 2012.

Our team developed a prototype database for systems microscopy data and loaded seven datasets from project partners in the Systems Microscopy Network of Excellence, which was funded under the EU's Seventh Framework Programme (FP7).

As a part of our participation in the GEUVADIS project (funded by the FP7), we analysed mRNA and small RNA from lymphoblastoid cell lines of 465 individuals who participated in the 1000 Genomes Project. Our group led the

analysis of transcript isoform use and fusion gene discovery. By integrating RNA and DNA sequencing data, we were able to link gene expression and genetic variation, and to characterise mRNA and miRNA variation in several human populations. All of the data generated in the project are available through ArrayExpress.

The human transcriptome contains in excess of 100 000 different transcripts. We analysed transcript composition in 16 human tissues and five cell lines to show that, in a given condition, most protein coding genes have one major transcript expressed at significantly higher level than others, and that in human tissues the major transcripts contribute almost 85% to the total mRNA. We also found that the same major transcript is often expressed in many tissues. These observations can help prioritise candidate targets in proteomics research and help predict the functional impact of the detected changes in variation studies. Our findings, submitted for publication in 2012, point towards a lower degree of transcriptome complexity than recently estimated. Other research includes an exploration of the utility of gene expression data in the public domain (Rung and Brazma, 2012).

Angela Goncalves, a PhD student in the Functional Genomics group, gained her doctorate in 2012 and published some of her major findings in *Nature Genetics* and *Genome Research* (Goncalves, et al. 2012).



Alvis Brazma

PhD in Computer Science, Moscow State University, 1987. Postdoctoral research at New Mexico State University, US.

At EMBL-EBI since 1997.

Future plans

In 2012 we began the process of integrating Expression Atlas data with PRIDE proteomics data, and this process will continue in 2013. We will also undertake to integrate the metabolomics data in MetaboLights with the Expression Atlas. Our research will continue to focus on large-scale data integration and systems biology. We will develop methods for RNA-seq data analysis and processing, and apply these to address important biological questions, such as the role of alternative splicing and splicing mechanisms. Together with our colleagues at the International Cancer Genome Consortium, we will investigate the impact of cancer genomes on functional changes in cancer development and explore fusion genes and their role in cancer development.

References cited

Rustici, G., et al. (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res* 41(D1), D987-D990.

Kutter, C., et al. (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat Genet* 43, 948-955.

Selected publications

Rung, J. and Brazma, A. (2012) Reuse of public genome-wide gene expression data. *Nat Rev Genet* doi: 10.1038/nrg3394.

Fonseca, N.A., et al. (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169-3177.

Goncalves, A., et al. (2012) Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res* 22, 2376-2384.

Gostev, M., et al. (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res* 40 (D1), D64-D70. doi: 10.1093/nar/gkr937.

Kapushesky, M., et al. (2012) Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res* 40 (D1), D1077-D1081.

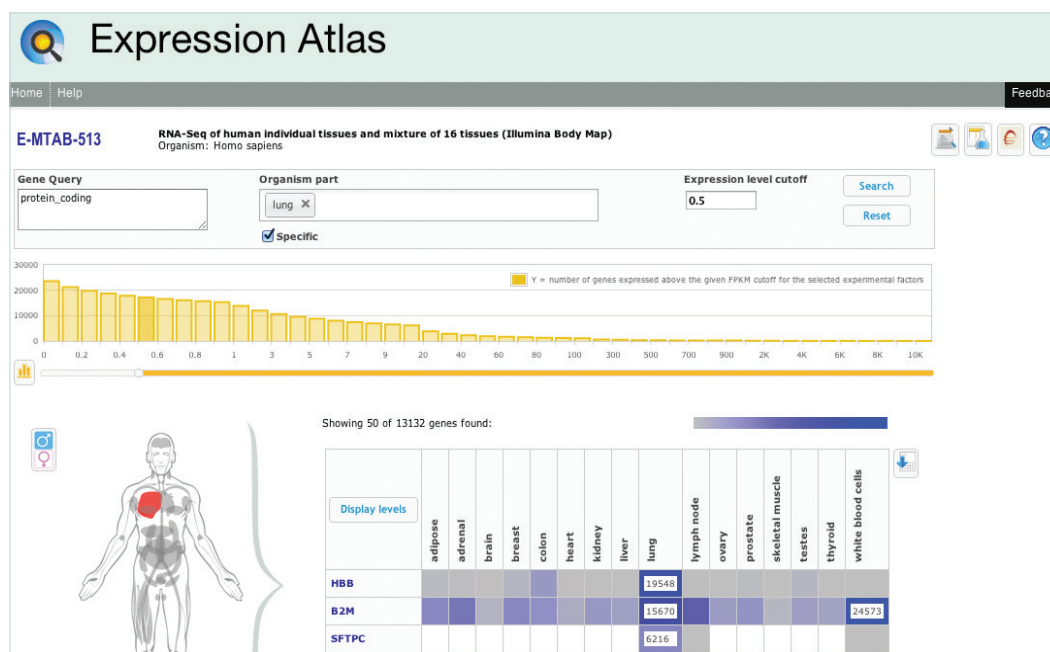


Figure. Prototype of the baseline Expression Atlas.

Functional genomics production

Our team manages data content, software delivery and usability for the ArrayExpress Archive, the Expression Atlas and the EBI BioSamples Database. The complex metadata in these resources represent experimental types, variables and sample attributes for which semantic mark-up in the form of ontologies is required. We develop ontologies and software for the annotation of complex biological data, including the Experimental Factor Ontology (EFO) for functional genomics annotation.

We are key collaborators in several projects funded by the European Commission, the US National Institutes of Health (NIH) and the US National Science Foundation (NSF). These range from Sybaris, which is investigating genome content and gene expression in fungal pathogens, to infrastructural projects such as BioMedBridges, which is developing integrative software to link different biomedical domains. As part of our collaboration with the National Center for BioOntology in the US, we support functional genomics data integration and analysis by developing tools for ontology manipulation and for the semantic web. With our partners in KOMP2 and the International Mouse Phenotyping Consortium, we manage, analyse and distribute complex phenotypic data from 20 000 knockout mouse lines. In the context of the InfraCOMP project, we promote mouse data integration in Europe.

Major achievements

We delivered a zoom-able, integrated browser for the National Human Genome Research Institute (NHGRI) genome-wide association study (GWAS) catalogue that uses semantic web technologies and EFO to support rich, dynamic queries. It allows users to query by trait, SNP (e.g., obesity, type II diabetes) and clinical measurement (e.g., blood glucose) to visualise associated regions of the genome. These in turn are integrated with Ensembl and the ontology term, synonyms and definitions.

As part of the EBI website redesign, our team is working to integrate our data more closely with other resources. By integrating the Expression Atlas dataset of under- and over-expressed genes with Ensembl, Reactome and Uniprot we have opened up possibilities for richer queries and data mining. This work represents EMBL-EBI's first production SPARQL endpoint.

EFO had two major releases in 2012: the first included numerous terms for the description of GWAS data and the second imported the Orphanet Ontology for genetic diseases. EFO supports Ensembl's annotation of phenotypes associated with variation.

A new NSF-funded collaboration with the Gramene plant database and the Ensembl Genomes team uses our team's curation and analysis expertise and the Atlas infrastructure to integrate gene expression datasets generated using Array and RNA-Seq technologies for the plant community. The results are generating new plant content for Atlas users.

Our team's three MSc students worked with the Literature Services team to evaluate ontologies and literature versus ArrayExpress database content. They also evaluated the use of ontologies in R to perform phenotype-driven analyses and compared RNA-Seq analysis methodologies with array-based technologies for the same samples. All of the students received distinctions for their work.

Future plans

In 2013 we will increase our use of semantic web technologies in production as we drive automatic annotation and support curation using RDF triple stores. We will also explore the evolution and repair of RDF knowledgebases in the new DIACHRON project. Our team plans to release the



Helen Parkinson

PhD Genetics, 1997. Research Associate in Genetics, University of Leicester 1997-2000.

At EMBL-EBI since 2000.



Figure. SNPs at 10-8. P value associated with Type II Diabetes, shown in orange against all SNPs from the GWAS Catalog, 2012. www.ebi.ac.uk/fgpt/gwas.

BioRDF-R package, which will enable enrichment analysis of RDF data. We will host MSc students exploring human-mouse data integration. We will contribute standards and service registries to the BioMedBridges project and will work towards semantic interoperability.

Selected publications

Rustici, G., et al. (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res* 41 (D1), D987-D990.

Jupp, S., Parkinson, H. and Malone, J. (2012) Semantic Web Atlas: putting gene expression data into biological context. *SWAT4LS*.

Functional genomics development

Our team develops software for ArrayExpress, a core EMBL-EBI resource, and the EBI BioSamples Database, which centralises biological sample data.

Together with the Functional Genomics Production team, we build and maintain data management tools, user interfaces, programmatic interfaces, annotation and data submission systems for functional genomics resources. We also collaborate on a number of European 'multi-omics' projects in a data management capacity.

Major achievements

In 2012 we concentrated on data integration and quality. We implemented a link to GenomeSpace, a data-integration platform, making it easy for GenomeSpace users to retrieve and process data from ArrayExpress. To help users visualise sequencing-based ArrayExpress data, we began generating BAM files and linking them through a genome browser interface. We also enhanced the Bioconductor package for importing ArrayExpress data, and pre-processed a significant portion of ArrayExpress data; the resulting R objects being available to use on the ftp server.

We made several incremental improvements to the ArrayExpress user interface, including better visualisation of sequencing experiments. We invested significant effort aligning ArrayExpress with the look and feel of the new EMBL-EBI website, which will be launched in 2013. We also reviewed options for replacing MIAMExpress.

In 2012 we worked with other resources to ensure the EBI BioSamples Database provides useful, dependable sample data management services. We resolved some scalability problems, which made it possible to manage tens of millions of samples. In the process, we redeveloped the user interface, the programmatic interface and the backend layer. Because the user interface has the same underlying software architecture and components as ArrayExpress, it inherited functionality such as ontological query expansion.

The team participated in two large European projects: diXa and EU-AIMS. diXa is providing a sustainable data management infrastructure for the development of non-animal toxicity tests. Our role is to facilitate the reuse of ArrayExpress, PRIDE and MetaboLights and to build a portal linking studies in these resources. We are working on a solution that can be used for similar projects or for data that do not have a proper home because of the novelty of the technology used.

EU-AIMS is a five-year project centred on autism-spectrum disorder studies that began in 2012. Our role is to integrate genomics and imaging data.

Future plans

The next major milestone for ArrayExpress is replacing the aging MIAMExpress tool with Annotare, the next-generation data-submission and curation tool. Annotare will be adjusted for sequencing-based data submissions. The updated ArrayExpress interface will launch with the new EMBL-EBI website in early 2013, and will feature improvements such as batch queries and better representation of very large experiments.

The new relational backend of the BioSamples database will be deployed in early 2013, along with an improved set of programmatic access interfaces. This will allow BioSamples to be used as the primary storage facility for biological sample information for EMBL-EBI resources. The first customer will be the European Nucleotide Archive. We will also enable restricted access to pre-publication samples and work on a data-exchange mechanism with NCBI in the US.

In 2013 we will deploy the first version of the diXa data warehouse, which will serve as a prototype 'BioStudy' database for semi-structured assay data. For the EU-AIMS project we will deploy the first version of the data-management infrastructure and will begin integrating relevant genomics data with MRI and fMRI imaging data.



Ugis Sarkans

PhD in Computer Science, University of Latvia,
1998. Postdoctoral research at the University of
Wales, Aberystwyth, 2000.

At EMBL-EBI since 2000.

EBL > BioSamples

Search BioSample Database at EBI [clear](#)

leukemia Sort by: [Relevance](#)

Examples: [leukemia](#), [ArrayExpress AND "Mus musculus"](#)

1171 groups found, displaying groups 1 to 50. Pages: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) .. [24](#)

Accession	Title	Samples
SAMEG22636	Transcription profiling of bone marrow samples from human acute myeloid leukemia patients	43
SAMEG81358	hgu133a integrated dataset	5896
SAMEG24345	Transcription profiling of human haematological neoplasias from blood samples and cell lines	89
SAMEG23113	Exon array data from mouse APL tumors	15
SAMEG19131	Transcription profiling of human pediatric acute myeloid leukemia samples	58
SAMEG127805	Transcription profiling of human pre and pro B cells from patients with childhood acute lymphoblastic leukemia	4
SAMEG24245	Transcription profiling of infants (<1 year of age) diagnosed with Acute Lymphoblastic Leukemia (ALL) characterised by genomic translocation.	73
SAMEG34514	Expression data from Down syndrome and non-Down syndrome pediatric acute lymphoblastic leukemia cases	49
SAMEG98065	Epigenetic reprogramming in relapsed childhood ALL	14
SAMEG20664	Transcription profiling of human normal bone marrow CD34+ cells, promyelocytes, and neutrophils and PR9 cell line PML-RARA induction time series	106
SAMEG29066	H3K9acetyl ChIP on chip for integrative epigenomic study in leukemia samples	16
SAMEG13019	Genotyping of relapsed childhood high hyperdiploid acute lymphoblastic leukemia reveals the presence of preleukemic ancestral clones and the secondary nature of microdeletions and RTK-RAS mutations	22
SAMEG32325	miRNA profiling of human acute myeloid leukaemia Kasumi-1 cell line CD34+38- compartment.	1
SAMEG35559	aCGH data from mouse APL tumors	32
SAMEG84689	Gene expression profiling of Acute Myeloid Leukemia (AML) samples	6
SAMEG14586	Transcription profiling of human AML cell line OCI/AML2 treated with valproic acid and all trans retinoic acid	4
SAMEG25986	Transcription profiling of human T-cell acute lymphoblastic leukemia , T-cell lymphoblastic lymphoma and B-precursor acute lymphoblastic leukemia samples	29
SAMEG71318	In vitro prednisolone resistance signature in MLL-rearranged infant ALL	43
SAMEG84674	Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding (Illumina expression)	10
SAMEG97424	Glucocorticoid Resistance Gene Profiling in T-ALL Cell Lines	15
SAMEG23851	Transcription profiling by array of human bone marrow and peripheral blood CD34+ myeloblasts in acute myeloid leukaemia	10
SAMEG13412	Transcription profiling of acute monocytic leukemia (AML-M5) patients with DNMT3A mutations	10

Figure 1. The BioSamples Database in 2012.

ArrayExpress

Examples: [E-MEXP-31](#), [cancer](#), [p53](#), [Geuvadis](#) [Advanced](#)

[Home](#) [Experiments](#) [Arrays](#) [Help](#) [About ArrayExpress](#) [Login](#)

ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to [MIAME](#) and [MINSEQE](#) standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Data Content

Updated today at 16:30

- 36066 experiments
- 1042850 assays
- 9.35 TB of archived data

Latest News

Jan 2013 - [Tools for mapping high-throughput sequencing data](#) topped the list of 'most-read articles' in Bioinformatics last month. Written by N.A. Fonseca, J. Rung, A. Brazma and J.C. Marioni at EMBL-EBI, the paper discusses the results from a survey which focuses on classifying

Figure 2. The ArrayExpress Database in 2012.