Functional genomics: An introduction to EMBL-EBI resources

Amy Tang [1]
Gene Expression
Beginner
0.5 hour

This course will introduce the field of functional genomics and provide an overview of the major resources for this field at EMBL-EBI.

Updated in August 2014.

For a more in depth look at functional genomics, we have just launched three new courses (May 2016) covering an introduction to functional genomics and designing experiments [2], common technologies and data analysis methods [3], and submitting your functional genomics data [4].

Learning objectives:

- Know what functional genomics is and the questions that can be addressed in this field of study
- List the main technologies used in functional genomics experiments
- Describe the main functional genomics resources at EMBL-EBI: ArrayExpress and Expression Atlas

What is functional genomics?

The aim of functional genomics [5] studies is to understand the complex relationship between genotype and phenotype on a global (genome-wide) scale. Studies investigate a range of processes such as transcription, translation [6] and epigenetic regulation, in an attempt to answer relevant biological questions such as:

- When and where are genes expressed?
- How do gene expression [7] levels differ in various cell types and states?
- What are the functional roles of different genes and in what cellular processes do they participate?
- How are genes regulated? Where are the active gene promoters in a particular cell type?
- How do genes and gene products interact?
- How does gene expression change in various diseases or following a treatment?
Figure 1. Functional genomics integrates information from various molecular methodologies to gain an understanding of how DNA sequence is translated into complex information in a cell (DNA → RNA → Proteins → biological process).

Experiments

Functional genomic experiments (Figure 1) typically utilise large-scale, high-throughput assays to measure and track many genes or proteins in parallel under different experimental or environmental conditions (e.g. with samples from patients and healthy individuals). This "genome-wide" approach allows the function of different parts of the genome to be discovered by combining information from genes, transcripts and proteins.

Technologies used in functional genomic studies

High-throughput technologies

Functional genomic studies frequently rely on high-throughput technologies such as microarrays and high-throughput sequencing (HTS). Examples of these include:

Microarrays

- **Expression-profiling** - used to measure the expression of thousands of genes at once, using oligonucleotide probes (usually ≤50 basepairs in length) designed from transcript cDNA or exon sequences across the genome.
- **Tiling microarrays** - often used for mapping transcription factor binding sites or locations of epigenetic marks (e.g. histone modifications). They use overlapping oligonucleotide probes (usually ≤50bp) covering several megabases of genomic sequences.

HTS

- **RNA sequencing (RNA-Seq)** - is used to sequence cDNA in order to get information about a sample's RNA content.
- **ChIP sequencing (ChIP-Seq)** - uses Chromatin ImmunoPrecipitation (ChIP) with DNA sequencing to identify protein-binding sites on DNA.

You can learn more about these applications in this review article (12) (1) (13)).

These studies produce massive amounts of data, e.g. expression values from tens of microarray chips, each with thousands of probes, or tens of millions of very short sequence reads from HTS
machines. These data can only help us gain insight into underlying biological processes, if they are carefully recorded and stored in databases, along with the experimental workflows employed and annotated detail for each sample. This will then allow the data to be queried, compared, analysed, interpreted and shared (2) [13] by the research community. Recently, the advent of more sophisticated applications of HTS technology, such as single-cell RNA-seq (sequencing one cell at a time instead of a population of cells) calls for even more meticulous documentation.

As a result, functional genomics [5] databases, such as ArrayExpress [14], were set up to provide easy access to functional genomics data which have been annotated sufficiently to capture the biology underpinning the experiments. Moreover, the databases distribute the annotated data in a structured and standardised format, facilitating information sharing and automated programmatic data-mining.

**Functional genomic resources at EMBL-EBI**

EMBL-EBI has two main functional genomic resources:


The relationship between the two databases and other software / resources is shown in Figure 2 (below).

![Figure 2. The relationship between the ArrayExpress and Expression Atlas databases.](image)

ArrayExpress and the Atlas are linked to other EMBL-EBI databases such as **ENA** [17], **Ensembl** [18] and **Uniprot** [19], which provide additional information on DNA and protein sequences.
and functions (Figure 2). Moreover, the two databases link out to data analysis software such as Bioconductor [20] (6 [13]) - a tool for the analysis and comprehension of high-throughput genomic data.

ArrayExpress

ArrayExpress [14] (3 [13]) is a database of functional genomics [5] data. Its stores microarray [21] and HTS data that are described and archived according to the community guidelines for microarray (MIAME [22]) and HTS (MINSEQE [23]).

The focus of the database is on individual experiments and it can be used to retrieve actual experiments as well as the information describing them. Very often, one experiment corresponds to one publication in a journal. In ArrayExpress, you can access the raw data [24] files, sample annotations or assays relating to an experiment of interest.

You can use ArrayExpress to:

- Search by keywords or experiment's properties (e.g. citation, transcriptomics [25] platform, species or sample annotation [26]) and identify experiments of interest;
- Download data associated with experiment(s), alongside its annotation, for your own analysis;
- Submit microarray or HTS data that you want to publish.

Submitting to ArrayExpress

Data generated by microarray-based and HTS-based platforms can be submitted directly to ArrayExpress [27] using our webform-based tool Annotare [28]. Submitting data to ArrayExpress is often a pre-requisite of publishing a functional genomics experiment, and allows the scientific community to perform searches on the latest functional genomics data.

Apart from direct submission, ArrayExpress also contains data imported via our in-house pipeline [29] from the Gene Expression Omnibus [30] (GEO) - a public repository of functional genomics data at the NCBI - to expand the number of studies that are available in ArrayExpress.

Expression Atlas

Expression Atlas [31] (the 'Atlas') (4,5 [13]) is a database of analysed gene expression [7] data that contains two components:

a. Baseline Atlas [32] - providing gene expression data for normal, untreated tissues or commonly used cell lines;

b. Differential Atlas [33] - allowing queries on genes that are up- or down- regulated in different experimental conditions.

Unlike ArrayExpress [15] which focuses on experiments, the Atlas [34] focuses on genes and biological conditions, allowing you to ask biological questions such as:
• What genes are expressed in normal human liver? [35]
• What genes are expressed across a panel of ENCODE cell lines [36]?
• What genes are up- or down-regulated in drought and salt tolerance (DST) mutant Japanese rice plants vs wild type controls [37]?
• What genes are up- or down-regulated in Gata3 knockout mice vs wild type controls [38]?

How the Atlas is produced

The Atlas is composed of a sub-set of data sets from ArrayExpress, namely those on expression profiling, which are manually curated and then analysed in-house by a standard statistical pipeline [39] (Figure 2 [40]). The manual curation [41] step ensures only well-annotated data sets from well-designed experiments are included in the Atlas. For example, for an experiment to be considered for the differential atlas, it must have at least three biological replicates for each condition for proper downstream statistical analysis, and the intent of the experiment must be clear. Various quality-control metrics are also implemented during statistical analysis to discard sub-standard data, e.g. microarray [21] data with lots of background noise.

Summary

• Functional genomics attempts to understand the complex relationship between genotype and phenotype on a global scale using high-throughput technologies, such as microarrays and high-throughput sequencing.

• The functional genomics resources at EMBL-EBI consist of ArrayExpress [15] and the Gene Expression Atlas [42].

• ArrayExpress is a database of functional genomics data supported by scientific publications.

• The Expression Atlas is a database providing information about gene expression in different biological conditions (for example in different cell types, organisms parts, diseases, compound treatments and genotypes).

• The difference between the two databases is that ArrayExpress is built around experiments (containing information on data files, sample annotation and others), whereas the Atlas is built around genes and biological conditions and is used to visualise changes in gene expression associated with different biological conditions.

Using an analogy in cooking, you can think of ArrayExpress as a place where you can source all raw cooking "ingredients" (raw data files with associated sample annotation), and you are free to go away and create your own dishes or make a meal out of them.

If you do not fancy cooking, or don't have the time/skills/tools to do so, the Atlas is like a "canteen" where pre-defined "dishes" (analysed gene expression data) are already prepared from the finest "ingredients" for your consumption.

Quiz: Functional genomics quiz
Learn more

Recommended online courses

- Functional genomics (I): Introduction and designing experiments [2]
- Functional genomics (II): Common technologies and data analysis methods [3]
- ArrayExpress: Discover functional genomics data quickly and easily [43]
- ArrayExpress: Quick tour [44]
- Expression Atlas: Quick tour [45]

Find out more

- ArrayExpress help [46] and Expression Atlas help [47] provide an up-to-date guide on making the most out of the two databases.
- Ensembl [48] is a genome databases for vertebrates and other eukaryotic species.
- Uniprot [19] is a resource for protein sequence and functional information.
- GEO [30] is a public functional genomics data repository at the NCBI.
- Interpro [49] is a database used for the classification and automatic annotation of proteins and genomes.
- Gene Ontology [50] is a project which provides an ontology of defined terms representing gene product properties.
- Europe PubMed Central (PMC) [51] is a literature resource providing access to all PubMed abstracts, Europe PMC full text articles, patent abstracts and many more.

Books


Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to
improve our courses and thus enhance your learning experience.

References


Contributors

Amy Tang [1]
EMBL-EBI
Functional Genomics Curation and Training Project Leader

Amy studied Human Genetics at the University of Nottingham for her bachelor degree. She did her PhD at the Brockdorff's lab [58], then at MRC Clinical Sciences Centre / Imperial College London, working on the epigenetic control of X-chromosome inactivation in mouse early development. The project was split 50:50 into wet-lab and bioinformatics, this is when she first got interested in bioinformatics. After a short postdoc in the same lab, she returned to study at Imperial for a MSc in
Bioinformatics and Systems Biology, which helped her move laterally from wet-lab to bioinformatics.

Amy now leads the curation efforts for ArrayExpress [59] and Expression Atlas [60] databases in the Gene Expression Team [61], promoting reproducible research by maintaining curation quality, and also by driving the development of the user-friendly ArrayExpress data submission tool Annotare [62]. To allow users make the most out of the databases and to educate researchers on best practices for reproducible research, she also manages all training activities for the Team, delivering most of the hands-on courses on functional genomics resources at EMBL-EBI and worldwide. Through listening to Annotare submitters’ needs and gathering feedback at training courses, she also acts as users' advocate and works with software developers to improve user experience of ArrayExpress and Expression Atlas websites.

Before joining EMBL-EBI, Amy was a senior bioinformatician in the Ensembl [48] GeneBuild team, curating gene models for key model organisms such as mouse and rat.

---

Gabry Rustici [63]

EMBL-EBI
Training coordinator, Functional Genomics

Gabriella Rustici was previously the Research and Training Coordinator in the Functional Genomics Group at the European Bioinformatics Institute in Cambridge, UK. She earned her PhD in Genetics from Cambridge University in 2004, working on transcription profiling of the fission yeast cell cycle, and has since been involved in the analysis of functional genomics data.

---

All course materials in Train online are free cultural works licensed under a Creative Commons Attribution-ShareAlike 4.0 International license

**Source URL:** http://www.ebi.ac.uk/training/online/course/functional-genomics-introduction-embl-ebi-resource-1

**Links**
[1] http://www.ebi.ac.uk/training/online/trainers/amytang
[61] http://www.ebi.ac.uk/~rpetry/geteam/
[62] https://www.ebi.ac.uk/fg/annotare/
[63] http://www.ebi.ac.uk/training/online/trainers/gabry