Functional genomics (III): Submitting your data and functional genomics databases

Melissa Burke [1], Laura Huerta [2]

- Gene Expression
- Beginner
- 1 hour

This is the third and final part of our functional genomics course. In this course, we will look at why, when, where to submit your functional genomics data.


Learning objectives:

- Describe the importance of sharing your functional genomics data
- Evaluate two functional genomics databases

Why should you share your data?

When we talk about data sharing, we are referring to providing the information needed to interpret, reuse and reproduce the experiment. The underlying idea is to make research data more available, citable, discoverable, interpretable, reusable and reproducible (Figure 1).

Providing access to data is crucial. As a consequence of the scientific method itself, where the validity of a conclusion depends on the ability to reproduce the underlying results, the publication of scientific data allows other scientists to replicate and then, importantly, extend knowledge in unanticipated directions. Sharing data and using it to build knowledge is what science is all about. By sharing your data, you can improve the reproducibility and visibility of your research, which increases your chances of being cited and forming new collaborations (1 [5]).
The idea of sharing functional genomics data is to make data more discoverable, accessible, interpretable, reusable and reproducible.

Sharing data also makes sense from an economic perspective and there is now a move towards open data to facilitate research and development in industry ([The Open Data Institute](https://www.opendatainstitute.org), 1-2 [5]).

**Journals and funding bodies often require you to submit your functional genomics data**

Guided by the principle of peer review, many funding bodies and journals now require the deposition of functional genomics data in a public database as a way of ensuring reproducibility and standardisation. Rather than mandating data sharing, [Scientific Data](https://www.nature.com/scientificdata) [7] has taken a different approach. This approach seeks to reward scientists for releasing their data and ensures data quality adheres to community standards.

A striking example of the importance of sharing data comes from the Ebola virus outbreak in 2014 and 2015: [Make outbreak research open access](https://www.nature.com/news/make-outbreak-research-open-access-1.15881) [8] (3 [5]).

**What kind of data should we share?**

The repeatability of published studies implies not only making raw data [9] (unprocessed data files obtained from the microarray [10] scanner or from the sequencing machine) available, but also providing detailed information and
protocols about the overall study and individual samples (known as metadata [11]).

What metadata are required?

Providing sufficient metadata is essential to understand the associated data and to make the data reusable and the results reproducible. Several 'minimum metadata reporting standards' have been created with that aim. These include MIAME [12] (Minimum Information About a Microarray Experiment) guidelines for microarray data and MINSEQE [13] (Minimum Information about Sequencing Experiments) for RNA-seq [14] data (Figure 2).

Both MIAME and MINSEQE emphasise the importance of providing not only the raw data but also the following:

- general information about the experiment (e.g. a summary of the experiment and its goals, contact information, and any associated publication);

- sample data relationships (e.g. which raw data file relates to which sample, which hybridisations are technical replicates and which are biological replicates);

- detailed sample annotation [15], including 'obvious' information such as the organism from which the samples were obtained, as well as highlighting the experimental factors and their values (e.g. listing 'compound' and 'dose' as factors in dose-response experiments, and specifying which compound was used and at what dose).

Curation of functional genomics experiments is essential to ensure that public datasets contain sufficient metadata so that the experiment can be understood without referring to an associated paper. Unfortunately, the metadata in public data sets often lack sufficient information and, even when extra information is found in an associated publication, it is often incomplete and in a non-standard format.

Tools to help annotate your experiments

There are annotation tools such as Annotare [16] that help you to annotate your experiments easily, thus improving metadata quality (see section ‘How to submit your data [17]’).

When to submit your data

When should you submit your data? As soon as you have the data! This will give the curators time to check through your submission and to work with you to ensure your submission complies with MIAME or MINSEQE guidelines. Both ArrayExpress [18] and GEO [19] databases give submitters full control over when to release the data to the public, so depositing early does not imply premature distribution of your unpublished data (Figure 3). Don’t wait until your manuscript is accepted, as curation can take longer than expected, and may result in a delay in the publication of your paper.
Figure 3 The recommended and not recommended approaches to submitting functional genomics data.
Public resources on functional genomics

Where to submit your data

Functional genomics data are challenging in terms of data transfer, storage, retrieval and maintenance, which is why large public access databases have been established for them.

Public databases facilitate the deposition and internet-based redistribution of data. They also provide considerable benefit to data generators who, with time, invariably want to free up computational storage space and guard against accidental data loss, especially for large data sets which are costly to (re)generate. New collaborations and new research opportunities are facilitated by data deposition because few large data sets are analysed to their full potential.

For some databases, a submitter can also be provided with a period of exclusive use of the data thereby allowing them to maximally benefit from their efforts (4 [5]).

Primary databases (or archival databases) for functional genomics

Primary archives are repositories that store high-throughput functional genomics data sets generated using both microarray- and sequence-based technologies. Data sets are directly submitted into primary databases by researchers who are publishing their results in journals that require original data to be made freely available for review and analysis.

Functional genomics data are predominantly stored in one of two public databases:

- **ArrayExpress** [20] at EMBL-EBI (Figure 4);
- **GEO** [19] at NCBI.
Figure 4 A screenshot of an experiment in ArrayExpress showing the six essential components required for MIAME/MINSEQE compliance. The data should include: a description of the experiment, its goals, and the relevant biological background (1), a description of the samples and variables being tested (2), detailed information about the experimental and data processing methods (3), contact details for the authors and any details of any associated publications (4), the raw sequencing or microarray data (5) and the processed data (6).

Want to learn more about ArrayExpress?

See our free online courses on using and accessing data in ArrayExpress:

- ArrayExpress Quick tour [21]
- ArrayExpress: Discover functional genomics data quickly and easily [22]

How to submit your data

ArrayExpress uses a webform tool called 'Annotare' [23] for functional genomics data submissions (Figure 5). Annotare has features such as drop-down menus, term suggestions, and pop-up help windows to guide you through the submission process.
Getting help with your ArrayExpress submissions

As a general guide, provide as much information as possible, even if you think it’s not important or too ‘obvious’! If you get stuck, expert curators are always at hand to help with your submission via annotare [at] ebi.ac.uk (email). You can find many tips and tricks for a successful Annotare submission in the submission guide [16].

What happens after submission

If you submit to ArrayExpress, you will usually receive an automatically generated accession [24] number (e.g. E-MTAB-3533) within five minutes of submission, which can be quoted in your manuscript. Note that this doesn’t mean that the submission is finished!

A curator will then check through your submission to make sure that all of the necessary data is there and that it is properly annotated. Curators will email you if they have any questions, and may re-open the submission for further editing/corrections.

Once fully curated, the experiment will be loaded into ArrayExpress and the curators will send you an email to confirm that the submission is finished. It stays private until you decide to release it or the accession number is cited in a publication. You can change the public release date of the experiment at any time to match the peer review progress of your paper. You can add or edit publication details too.

Once the data is public it can be viewed on the ArrayExpress website (no login or password required), downloaded and may be selected for inclusion in the Expression Atlas [25].
Re-use of functional genomics data

The functional genomics data in archival databases can be downloaded and re-used provided that the database, accession number and original paper are cited.

The data can be used to:

- perform meta-analyses to answer a specific biological question;
- evaluate analysis methods;
- re-analysis for inclusion in value-added databases such as Expression Atlas.

Before reusing the data, it is important to:

- check that the data is available for all the samples;
- check that there is sufficient annotation and description of the samples and protocols so that you can understand the design of the experiment and how it was performed;
- review the design of the study.
  - Are there enough replicates?
  - What were the controls?
  - Does it address the biological questions you are asking?
- revise and re-annotate the probes/reference genomes;
- perform quality control analysis on the samples.

For a more detailed discussion of the re-use of public functional genomics data see Rung and Brazma. 2012 ([5]).

Secondary (value-added) databases

Secondary databases extract information from primary data to answer questions such as which genes are expressed under specific conditions and how gene expression differs between conditions. The added value comes from: data processing, additional annotation, mapping to standardised vocabularies or ontologies (such as Experimental Factor Ontology [26]) and analysis to extract gene expression profiles and other results from primary data.

Mapping metadata terms to controlled vocabularies is a core activity in curation, enabling efficient search of datasets by keywords, and is crucial for data sharing across research contexts.

Expression Atlas at EMBL-EBI (I)

One example of a value-added database is the Expression Atlas [27] (the Atlas). It contains pre-computed (analysed) gene expression values from RNA-sequencing and microarray expression experiments across different tissues, cell types, developmental stages and many other experimental conditions. It covers over 37 organisms including metazoans and plants.

Queries can be either in a:

- **baseline context**, e.g. find genes expressed in the macaque brain;
- **differential context**, e.g. find genes that are up- or down-regulated in response to auxin in Arabidopsis (Figure 6).
Expression Atlas consists of two components: differential and baseline expression. As of March 2016, Expression Atlas contains 2723 datasets, including 220 RNA-sequencing experiments.

All data in Expression Atlas are free to browse, download, reuse and are selected from the ArrayExpress archive. All datasets are manually curated to a high standard by in-house curators, using terms from the Experimental Factor Ontology (EFO) where possible, and processed using standardised analysis methods.

Searching the Expression Atlas

You can search the Atlas by gene attributes (gene symbols, Ensembl [28] gene ID or keywords), as well as by sample attributes and experimental factors, taking advantage of ontology-driven query expansion. For example, searching for the disease ‘leukemia’ will automatically return expression data from samples of ‘leukemia’ itself and from its subtypes and closely related diseases (‘chronic lymphocytic leukemia’, ‘promyelocytic leukemia,’ etc), since the relationship between “leukemia” and related diseases is defined in the EFO.

Expression Atlas at EMBL-EBI (II)

An example of how differentially expressed genes are displayed in Expression Atlas is shown below in Figure 7.

The gene expression values are displayed as a heatmap where rows represent genes and columns are the experimental comparisons. In the header of the heatmap table, you will see the results of the gene set enrichment analysis (see the Gene set enrichment section [29] of the Functional genomics II course).
Functional genomics (III): Submitting your data and functional genomics databases

Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Figure 7 Example differential expression page with help annotations. Apart from the heatmap, gene set overlap summaries are shown. Each of the three plots shows a maximum of top 10 GO terms, InterPro domains and Reactome pathways ‘enriched’ in the comparison ‘tumor’ vs ‘paratumoral tissue’ in CTNNB1 mutation positive hepatocellular carcinoma (see Transcriptomic characterization of Hepatocellular Carcinoma with CTNNB1 mutation [30])

Gene set enrichment analysis is performed using the Piano package from Bioconductor [31]. For each comparison, the overlap between the set of differentially expressed genes and terms from GO, InterPro, and Reactome is performed, using Fisher’s exact test with multiple testing correction (FDR < 0.1).

Find out more about Expression Atlas

If you would like to learn more about how to use Expression Atlas, please take a look at the free online course: Expression Atlas: Quick tour [32].

Summary

Functional genomics databases

Primary functional genomics databases (e.g. ArrayExpress, GEO) are those that store and facilitate the sharing of functional genomics data. Secondary (value-added) functional genomics databases (e.g. Expression Atlas) aim to
enable the visualisation and interpretation of data stored in the primary databases.

**Submitting your own data**

When submitting your data you should start as soon as you have the data and provide as much information as possible!

Many funding bodies and journals now require researchers to submit their data to a public database before a paper can be accepted.

ArrayExpress uses a webform called Annotare for data submission. Annotare guides researchers through the submission process and helps them to ensure that they are providing enough information to be compliant with the MIAME and MINSEQE guidelines.

**Learn more**

**Links and resources**

- [ArrayExpress](#) [18]
- [Expression Atlas](#) [33]
- [RNaseqlopedia](#) [34]
- MIAME [12]/MINSEQE [13] guidelines
- [EBI Gene Expression Team page](#) [35]

**Recommended online courses**

- [Functional genomics (I): Introduction and designing experiments](#) [3]
- [Functional genomics (II): Common technologies and data analysis](#) [4]
- [ArrayExpress: Discover functional genomics data quickly and easily](#) [22]
- [ArrayExpress: Quick tour](#) [21]
- [Expression Atlas: Quick tour](#) [32]

**Contact us**

- mburke [at] ebi.ac.uk (Melissa Burke)
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

Melissa Burke [1]

EMBL-EBI
Scientific Training Officer (e-learning): Training Team

Melissa is the Scientific Training Officer (e-learning) for the Training Team at the EMBL-EBI. She joined the Training Team in July 2016 after having worked as a Scientific Curator for ArrayExpress/Expression Atlas at the EMBL-EBI. She has a PhD in Molecular Parasitology and has worked internationally as a postdoctoral researcher specialising in the functional genomics of infectious diseases.
Laura Huerta

EMBL-EBI
Senior Scientific Curator - Petryszak team: Gene Expression

Laura works as a Senior Scientific Curator to improve the content of functional genomics data in Expression Atlas. She is involved in developing and implementing metadata standards with a particular interest on data integration through ontology annotation. She manages training activities, delivering courses on functional genomics resources at EMBL-EBI and worldwide. She also interacts with software developers to improve user experience of Expression Atlas website. Laura joined EMBL-EBI in 2015 after receiving her PhD in Molecular Biology from the Polytechnic University of Valencia, focused on studying plant development regulated by hormones through the generation and analysis of transcriptomics data.

Source URL: https://www.ebi.ac.uk/training/online/course/functional-genomics-iii-submitting-your-data-and-f
[28] https://www.ebi.ac.uk/training/online/glossary/ensembl
[32] https://www.ebi.ac.uk/training/online/course/expression-atlas-quick-tour-0
[33] https://www.ebi.ac.uk/gxa/
[34] http://maseq.uoregon.edu/
[38] http://europepmc.org/abstract/MED/25719649
[40] http://europepmc.org/abstract/MED/23269463