Functional genomics (I): Introduction and designing experiments

Melissa Burke [1], Laura Huerta [2]

- Gene Expression
- Beginner
- 1 hour

This is the first part of our functional genomics course. This course will give you an introduction to functional genomics. We will introduce you to different types of functional genomics studies and discuss best practices when designing your own experiments.

We will also explore some examples of how functional genomics is being applied to drug discovery and plant sciences.

Learning objectives:

- Describe some types of functional genomics studies
- Apply best practices when designing your own functional genomics experiments

Overview

This functional genomics course is divided into three parts, in which we will look at:

Functional genomics (I): Introduction and designing experiments (this part of the course)

- what functional genomics is;
- functional genomics case studies in plants and drug discovery;
- types of functional genomics experiments;
- designing a functional genomics experiment.

Functional genomics (II): Common technologies and data analysis methods [3]

- types of functional genomics technologies;
- data analysis.


- reproducibility and submitting your functional genomics data;
What is functional genomics?

**Functional genomics** is the study of how genes and intergenic regions of the genome contribute to different biological processes. A researcher in this field typically studies genes or regions on a “genome-wide” scale (i.e. all or multiple genes/regions at the same time), with the hope of narrowing them down to a list of candidate genes or regions to analyse in more detail.

The goal of functional genomics is to determine how the individual components of a biological system work together to produce a particular phenotype. Functional genomics focuses on the **dynamic expression of gene products in a specific context**, for example, at a specific developmental stage or during a disease. In functional genomics, we try to use our current knowledge of gene function to develop a model linking genotype to phenotype.

There are several specific functional genomics approaches depending on what we are focused on (Figure 1):

- DNA level (genomics and epigenomics);
- RNA level (transcriptomics);
- protein level (proteomics);
- metabolite level (metabolomics).
Figure 1 Functional genomics is the study of how the genome, transcripts (genes), proteins and metabolites work together to produce a particular phenotype.

Together, transcriptomics, proteomics and metabolomics describe the transcripts, proteins and metabolites of a biological system, and the integration of these data is expected to provide a complete model of the biological system under study.

In this course, we focus on DNA and RNA level approaches.

**Functional genomics case studies**

Examples of biological questions that can be tackled using functional genomics experiments are:

- Why do some cancer drugs only work effectively on a subset of patients with the disease?
- Why are some cultivars of rice more resistant to drought than others?
- What makes some individuals more susceptible to skin allergies?

On the next pages we will look at some examples of how functional genomics is being applied to drug discovery and plant sciences.

**Gramene: functional genomics in plants**

Functional genomics (in combination with other approaches) plays an important role in plant sciences. It is being
used in many diverse ways from informing selective breeding in tomatoes (1 [5]), producing alternative gluten-free grains (and beer) (2 [5]) and understanding the response of plants to climate change (3 [5]).

Translating findings from one species to another is an important aspect of plant functional genomics and it has just become easier with the development of the Gramene database.

The Gramene database [6] is an integrated resource for comparative genome and functional analysis in plants (Figure 2). The database provides agricultural researchers and plant breeders with valuable biological information on genomes and plant pathways of numerous crops and model species, thus enabling powerful comparisons across species.

Through its two main frameworks, genomes (collaboration with Ensembl Plants [7]) and pathways (The Plant Reactome [8]), Gramene provides online resources for visualising and comparing plant genomes and biological pathways (Figure 2).

As of April 2016, the current release of Gramene contains 39 reference genomes including: rice, maize, wheat, barley, soybean, Arabidopsis, Brassicas, poplar, medicago, tomato, potato, banana, cocoa, peach, grapevine, Amborella, spikemoss and algae. In addition to curated rice pathways, the Plant Reactome incorporates pathway projections to 58 plant species, including two wild peanut ancestors and the common bean.

Gramene genome browser

For genomes, the Gramene genome browser [9] allows you to explore plant genome features, functional ontologies, variation data and comparative phylogenomics [10]. For each of the 39 plant reference genomes
represented in the genomes section of Gramene you can find:

- protein-coding and non-coding genes, splice variants, cDNA and protein sequences in the browser’s gene annotation section;
- homologues, gene trees and whole genome alignments across multiple species in the comparative genomics section;
- sequence variants in the variation section.

The Gramene genome browser interacts with the Expression Atlas [12] to display expression data from curated expression studies in plants (Figure 3).

**Figure 3** Expression data of a rice prolamin gene across different tissues. Prolamins are rice seed storage proteins that accumulate in the endosperm during seed maturation, and serve as a source of carbon and nitrogen.

### Plant Reactome

For pathways, Plant Reactome [8] is a database of plant metabolic and regulatory pathways. Currently, it features Oryza sativa (English rice) as a reference species. Gene homology-based pathway projections for other species are developed by using Oryza sativa ssp. Japonica as a reference.

The database contains more than 240 rice pathways of which 200 were manually curated. In Plant Reactome you can search and browse pathways, as well as find gene expression data (Figure 4).
Drug discovery

The first step in the drug discovery pipeline is the characterisation of the disease process and identification of drug ('therapeutic') targets. Here, a 'target' is defined as a protein or messenger RNA which, when modified by a drug, favourably affects the outcome of a disease.

In traditional drug discovery pipelines (Figure 5), less than 10% of candidate targets turn out to be valid after the investment of millions of dollars and six to seven years of work [4] [5]. More than 50% of drugs fail at phase II clinical trials due to the lack of efficacy, i.e. the drug is not harmful to humans, binds to the intended target and has drug-like properties (e.g. a reasonable half-life in the human body), but it does not significantly affect the disease.
Figure 5 It is well known that de novo drug discovery and development is a 10-17 year process from idea to marketed drug. The probability of success is lower that 10%.

Using functional genomics to guide drug discovery

The pharmaceutical industry is keen on reducing the staggering rate of failure for targets in the drug discovery pipeline. One approach is to improve the identification and selection of potential targets, so drug development teams can focus on more hopeful candidate targets from the beginning. In this respect, results from many types of functional genomics experiments can provide information or evidence for the relationship between potential targets and their associated disease at the:

- DNA level (single nucleotide polymorphism and copy number variations, epigenetics);
- RNA level (gene expression microarrays and RNA-seq);
- and the protein level (DNA/RNA-protein interactions e.g. ChIP-seq).

Functional genomics in cancer studies

Using functional genomics to guide drug discovery has been particularly useful for cancer studies. For example the discovery that the gene HER2 is over-expressed in certain types of breast cancers led to the development of the drug Herceptin (The HER2 Journey). More recently, high throughput analyses and meta-analyses of data from breast cancer samples have uncovered many additional targets for which existing drugs can be re-purposed, significantly speeding up the process of drug discovery.

Open Targets

Traditionally, acquiring, analysing and integrating the huge amount of DNA, RNA and protein data to discover new target-disease relationships required advanced bioinformatics expertise. To make this process accessible to biologists and applicable to all aspects of human disease, a new collaborative initiative called Open Targets was set up by EMBL-EBI, Wellcome Trust Sanger Institute, GSK and Biogen (Figure 6).

Open Targets provides an open access and user friendly web-based tool for the identification of novel associations between targets and diseases. It integrates data from many sources to calculate a 'score' for each potential association between a target and a disease. These sources include:
- functional genomics databases (e.g. Expression Atlas);
- genome variation databases (GWAS [21], the Cancer Genome Atlas, the European Variation Archive);
- protein (UniProt [22]);
- chemical (ChEMBL [23]);
- pathway (Reactome [24]);
- and phenotypic resources as well as the text-mined results from published literature (Europe PMC [25]).

Figure 6 Open Targets provides a tool for discovering potential therapeutic targets and the links between these targets and human diseases.

The idea is that potential therapeutic targets for each disease are ranked based on this association 'score', i.e. based on the strength of the association with the disease and the likelihood that modifying these targets would have a beneficial outcome. You can also see how a particular target is linked with other diseases. This is all linked to publicly available experimental evidence which can be visualised and explored using the Open Targets platform.

Common study types in functional genomics

Functional genomics experiments measure changes in the DNA (genome and epigenome), RNA (transcriptome [26]), or interactions between DNA/RNA and proteins that influence the phenotype [27] of a sample. Common branches of functional genomics include (Figure 7):

- genotyping;
- transcription profiling;
- epigenetic profiling;
nucleic acid-protein interactions;
meta-analysis.

Figure 7 The molecules that can be analysed by functional genomics and the technologies used in the analysis.

Transcription profiling

This is one of the most popular study types, also known as 'expression profiling'. It involves the quantification of gene expression [16] of many genes in cells or tissue samples at the transcription (RNA) level. The quantification can be done by collecting biological samples and extracting RNA (in most cases, total RNA) following a treatment or at fixed time-points in a time-series, thereby creating 'snap-shots' of expression patterns.

For common reference genomes with well-annotated transcripts and genes (e.g. the human genome), a researcher can choose to focus on quantifying transcription of all or a subset of transcripts, genes, coding exons, non-coding RNA [28], and so forth.

Example 1: Transcription profiling by high throughput sequencing of four rice cultivars with variation in growth response to phosphate starvation [29]

Example 2: Molecular analysis of brain neoplasia [30]

Genotyping

Genotyping studies are those which identify differences in the DNA sequence (i.e. genotype) of a sample. The genomic DNA samples are often obtained from two contrasting groups of samples, e.g. drought-resistant rice cultivars vs. drought-sensitive counterparts, with the aim of identifying differences in the genotype which may
explain the difference in phenotype [27].

Genotyping studies can be designed to identify DNA sequence differences at three levels:

- **Single nucleotide polymorphisms** (SNPs, pronounced ‘snips’): SNP analysis focuses on differences in the DNA sequence at the single nucleotide level.
  Example: **SNP 6.0 profiling of plasma DNA in breast cancer patients** [31]

- **Copy number variations** (CNVs): CNVs refer to an increase or decrease in the number of copies of a segment of DNA (e.g. a gene, or a locus-specific DNA repeat element). Each ‘copy’ can be as short as 50 bases or up to 100 kilobases.
  Example: **A genome-wide copy number variant study of suicidal behavior** [32]

- **Structural variations**: they are an order of magnitude larger than CNVs and often cover megabases of DNA, and can be caused by chromosomal rearrangement events.
  Example: **Transcription profiling by high throughput sequencing of different maize lines to discover and characterise ‘presence-absence variation’ in the maize genome** [33]

One common extension of genotyping studies in humans are **genome-wide association studies** (GWAS [34]). Samples from cases (e.g. rheumatoid arthritis patients) and controls (e.g. healthy individuals) are genotyped across specific sites in the genome, followed by statistical analysis to find SNPs which are significantly more prevalent in one group (e.g. the disease cases). Such SNPs may then suggest an association between the SNPs and disease susceptibility.

### Epigenetic profiling

Epigenetics is the study of how biochemical modifications or physical interaction of DNA/chromatin affect gene regulation [35] in a cell, where such modifications/interactions are not related to changes in the underlying DNA sequence.

At the DNA level, **methylation** [36] of **CpG dinucleotides** [37] (often located near gene promoters) can be detected by first converting unmethylated cytosines into **uracil** [38] using bisulfite, which allows methylated and unmethylated cytosines to be distinguished.

At the chromatin level, modifications of the tails of histone proteins (e.g. methylation, acetylation) can be mapped by ‘immunoprecipitation’, where chromatin and proteins are chemically cross-linked reversibly. The genomic DNA associated with the modification/protein of interest is then ‘pulled-down’ (precipitated) with specific antibodies raised against the modification/protein. After precipitation, the cross-linking is reversed to release the genomic DNA for further analysis.

**Example:** **Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single base resolution** [39]

### DNA/RNA-protein interactions
Transcription factors, ribosomes and other DNA/RNA-binding proteins can bind to nucleic acid sequences and influence the transcription and translation [40] of genes. The immunoprecipitation technique has also been applied to study protein binding sites on RNA.

Example: RIP-seq of *Saccharomyces cerevisiae*... [41]

**Meta-analysis**

Meta-analysis is a branch of functional genomics in which data from pre-existing experiments is combined to create statistically more powerful models of a biological process. This type of analyses has become popular as it allows the identification of subtle events that could not be detected in smaller studies. Functional genomics databases such as ArrayExpress [42] and Expression Atlas [43] play an important role in these studies as reliable, well annotated sources of functional genomics data.

Example: Genome-wide analysis of over 106,000 individuals identifies 9 neuroticism-associated loci [44]

**Designing functional genomics experiments**

Like all experiments, functional genomics experiments begin with a biological question (Figure 8). This can be hypothesis driven (e.g. infection with influenza triggers interferon signalling pathways) or discovery/curiosity driven (e.g. what is the effect of salt treatment on gene expression in *Arabidopsis* plants?).

![Figure 8 Design elements for functional genomics studies.](image)
Three important things to consider when designing a functional genomics experiment are:

- scale and intent;
- data analysis;
- reproducibility.

We will discuss these in more detail over the next few pages.

**Scale and intent of the experiment**

The scale of an experiment encompasses both the number of samples and the number of genes to be analysed (8 [5]).

The number of samples is often a trade off between:

- the number of replicates required to produce statistically robust results;
- the ease of obtaining the samples;
- the budget for the project.

The scale and intent of the experiment will influence the methods used in the study.

- **Real-time PCR** [45] analysis is well suited to studies that aim to analyse a small number of genes in a small number of samples.

- **Microarrays** are useful for measuring large numbers of genes (or the whole transcriptome [26]), but they have reduced sensitivity compared with PCR.

- **RNA-sequencing** (RNA-seq [17]) and **next generation sequencing** [46] (NGS) are best suited for in depth analyses and are useful for discovery based projects such as those that aim to identify new transcripts, study non-coding RNAs, map transcription start sites or characterise the precise location of epigenetic modifications.

RNAseq/NGS is more flexible than **microarray** [47] analysis because it is not restricted by prior knowledge of genetic sequence. It can also detect transcripts over a higher dynamic range. However, it is still relatively expensive and computationally and statistically intensive (9 [5]).

**Other considerations**

For each type of experiment there will be further considerations including:

- the technology (Taqman vs Sybr green PCR; one vs two colour microarrays, paired-end vs single-end sequencing);
- the platform (Illumina [48] vs Affymetrix [49] vs Agilent etc);
- the methods that are used to prepare the samples.
Data analysis

Although data analysis only happens after the wet-lab experimental procedures are performed, it is a very important aspect of the design of all functional genomics experiments and should be considered before embarking on the wet-lab procedures. This is especially true for RNA-seq experiments (8 [5], 9 [5], RNAseqlopedia: Experimental Design [50]).

Some important things to consider while planning your experiment are:

- How are you going to analyse your data?
- Is special software required? Is it free or does it need a licence?
- Is extra information needed? For example, a reference genome for aligning RNA-seq reads?
- Is bioinformatics expertise required? For example, if you’re expecting thousands of raw data files, each being several gigabytes in size, you will need some programmatic skills to manage them (open, copy, move, etc.).
- What are the hardware (storage and computational processing) requirements?
- What are the most appropriate statistical tests and has the experiment been designed so that you can use them?

Methods for analysing functional genomics data are discussed in the second part of this course ‘Functional genomics (II): Common technologies and data analysis methods [3]’.

Reproducibility and transparency

Reproducibility and transparency (or the lack of) is a hot topic at the moment, with many articles pointing out the limitations of published scientific papers (The Academy of Medical Sciences, UK [51]) (Figure 9). By ‘reproducible’, we mean both in a ‘mechanical’ sense (e.g. starting from a given raw data file, following an exact set of analysis procedures to obtain an identical processed data file) and in a ‘biological’ sense (e.g. replicating the conclusion of a study by sourcing comparable samples and following similar wet-/dry-lab protocols).
Some steps that you can take to ensure the reproducibility and transparency of your experiment are:

- ensure that the design of your experiment includes appropriate controls and replicates;
- use appropriate statistical tests and seek advice where necessary;
- keep back-up copies of your raw data (request them from the core facility if needed);
- document all procedures and parameters ('Materials and Methods') with as much detail as possible;
- learn about and follow the guidelines describing the minimum information required for the publication of real-time PCR (MIQE [52]), microarray (MIAME [53]) and RNA-seq experiments (MINSEQE [54]);
- submit your (well annotated) experiment to a public database for functional genomics experiments such as ArrayExpress or GEO.

**Summary**

**What is functional genomics?**

Functional genomics is the study of how genes, intergenic regions of the genome, proteins and metabolites work together to produce a particular phenotype. There are several specific functional genomics approaches depending on what you are focused on:
DNA level (genomics and epigenomics);
RNA level (transcriptomics);
protein level (proteomics);
and metabolite level (metabolomics).

Integration of data from these approaches is expected to provide a complete model of the biological system under study.

How to design an experiment
When designing an experiment it is important to consider the goals and scale of the experiment, how you will handle and analyse the data and the steps you need to take to ensure the reproducibility of the experiment.

For example:

- How many samples (and replicates) are needed to ensure statistically robust results?
- Which technology is most appropriate to the design and goals of the experiment?
- How will the data be analysed and stored?
- What extra information (metadata) is needed to fully describe the experiment and be compliant with the MIAME/MINSEQE guidelines for reproducibility?

What next?
Why not take our quiz on the next page and test your knowledge from this course? After that, we recommend taking a look at parts II and III of our functional genomics course. These cover:

- Part II: Common technologies and data analysis methods [3]

Quiz: Functional Genomics (I): Introduction and designing experiments

Questions: 5
Attempts allowed: Unlimited
Available: Always
Pass rate: 80 %
Backwards navigation: Allowed

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.

Learn more
Links and resources

- ArrayExpress [55]
- Expression Atlas [56]
- RNAseqlopedia [57]
- MIAME [53]/MINSEQ [54] guidelines
- EBI Gene Expression Team page [58]

Recommended online courses

- Functional genomics (II): Common technologies and data analysis [3]
- ArrayExpress: Discover functional genomics data quickly and easily [59]
- ArrayExpress: Quick tour [60]
- Expression Atlas: Quick tour [61]

Contact us

- mburke [at] ebi.ac.uk (Melissa Burke)
- lauhuema [at] ebi.ac.uk (Laura Huema)
- arrayexpress [at] ebi.ac.uk (ArrayExpress)

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 [2]
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.