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

Laura Huerta [1], Melissa Burke [2]

- Gene Expression
- Beginner
- 2 hours

This is the second part of our functional genomics course. This course will introduce you to common technologies in functional genomics studies, including microarrays and next generation sequencing (NGS), with a special focus on RNA-sequencing (RNA-seq). We will look at important considerations when designing your experiments, data analysis methods, and discuss when to use one technology over another.

We recommend that you look at Functional genomics (I): Introduction and designing experiments [3] before starting this course.

Learning objectives:

- Evaluate different functional genomics technologies and know when to use them
- List some of the considerations when designing your own functional genomics experiments using these technologies
- Describe the steps involved in data analysis of microarray and RNA-seq experiments

Introduction

Many different types of technologies can be used in functional genomics experiments ranging from real-time PCR [4] to high-throughput technologies such as microarrays and next generation sequencing [5].

Choosing the right technology

The type of technology that you choose to use will depend on the scale and intent of your experiment. When designing your experiment you will need to think about the advantages and disadvantages of each of the technologies and whether they are suited to the type of experiment that you are planning.

For example, have you thought about how many genes you want to analyse? Real-time PCR is more suited to the analysis of a small number of genes while microarrays and next generation sequencing technologies can analyse thousands of genes simultaneously. Next generation sequencing also has the added advantage in that it allows the discovery of new genes and transcripts.

In this section of our course we discuss the application of real-time PCR, microarrays and next-generation sequencing (with a focus on RNA-sequencing) to functional genomics experiments. We will touch on the advantages and disadvantages of each technology, common analysis methods and look at how these technologies can be applied to a variety of biological questions.

Real-time PCR
Real-time polymerase chain reaction (real-time PCR [4]) is commonly used to measure gene expression [6]. It is more sensitive than microarrays in detecting small changes in expression but requires more input RNA and is less adaptable to high-throughput studies (1 [7]). It is best suited for studies of small subsets of genes. Its one major shortcoming is that the sequence of the specific target gene of interest must be known (so you can design the PCR primers), hence real-time PCR can only be used for studying known genes.

**Real-time PCR steps**

The first step in a real-time PCR reaction is the conversion of RNA to complementary DNA (cDNA) - this process is known as reverse transcription (Figure 1). The next step uses fluorescent reporters and a PCR reaction to amplify and detect specific genes (Figure 1). Two types of fluorescent reporters are commonly used; these are SYBR green and Taqman probes.
Figure 1 Real-time PCR involves conversion of RNA to cDNA via reverse transcription, followed by several rounds of PCR to amplify and detect the genes of interest. The products can be detected in ‘real-time’ by using SYBR-green or Taqman probes.
SYBR green and Taqman probes

SYBR green is a dye that fluoresces only when bound to double stranded DNA (i.e the PCR product) (Figure 1).

Taqman probes are made of a gene-specific nucleic acid probe [8], joined to reporter and quencher [9] molecules. The probe binds to the DNA between the forward and reverse primer. While the reporter and quencher are bound to the probe, the quencher absorbs the fluorescence emitted by the reporter. During the extension phase of the PCR reaction the probe is degraded, releasing the reporter and allowing its fluorescence to be detected. The advantage of the Taqman method is that probes with different coloured reporters can be combined in multiplex assays [10].

For both SYBR green and Taqman methods, the amount of fluorescence in a sample is detected in ‘real-time’ and plotted against the cycle number (Figure 1). The amount of fluorescence is proportional to the amount of PCR product. The time point at which the fluorescence reaches a defined threshold is relative to the level of gene expression. The design of real-time PCR experiments requires prior knowledge of the gene sequence and careful consideration of the types of controls to include.

For a comprehensive discussion about real-time PCR, you can have a look at Qiagen: Critical Factors for Successful Real-time PCR [11].

Microarrays

Microarrays are a collection of DNA probes that are usually bound in defined positions to a solid surface, such as a glass slide, to which sample DNA fragments can be hybridised. The probes are generally oligonucleotides [12] that are ‘ink-jet printed’ onto slides (Agilent) or synthesised in situ (Affymetrix [13]). Labelled single-stranded DNA or antisense RNA [14] fragments from a sample of interest are hybridised to the DNA microarray [15] under high stringency conditions. The amount of hybridisation [16] detected for a specific probe [8] is proportional to the number of nucleic acid fragments in the sample.

One-colour or two-colour arrays?

A major design consideration in a microarray experiment is whether to measure the expression levels from each sample on separate microarrays (one-colour array) or to compare relative expression levels between a pair of samples on a single microarray (two-colour array) (Figure 2). The overall performance of one-colour and two-colour arrays is similar (2 [7]).
Figure 2 One-colour vs two-colour arrays.

**Figure 2 description:**
In two colour microarrays, two biological samples (experimental/test sample and control/reference sample) are labelled with different fluorescent dyes, usually Cyanine 3 (Cy3) and Cyanine 5 (Cy5). Equal amounts of labelled cDNA are then simultaneously hybridised to the same microarray chip. After this competitive hybridisation, the fluorescence measurements are made separately for each dye and represent the abundance of each gene in one sample (test sample, Cy5) relative to the other one (control sample, Cy3).

The hybridisation data are reported as a ratio of the Cy5/Cy3 fluorescent signals at each probe. By contrast, in one colour microarrays, each sample is labelled and hybridised to a separate microarray and we get an absolute value of fluorescence for each probe.

**Counteracting dye bias effects when using two-colour arrays**

One issue for two-colour arrays is related to dye bias effects introduced by the slightly different chemistry of the two dyes. It is important to control for this dye bias in the design of your experiment, for example by using a **dye swap design** or a reference design (Figure 3). In a dye swap design, the same pairs of samples (test and control) are compared twice with the dye assignment reversed in the second hybridisation. However, the most common design for two colour microarrays is the **reference design** in which each experimental sample is hybridised against a common reference sample.
A dye swap design involves hybridising the same two samples (test and control) twice. In one hybridisation, the control sample is labelled with green (Cy3) dye and the test sample with red (Cy5) dye. In the second hybridisation dye assignment is reversed. In a reference design, both test and control samples are labelled with one dye (usually Cy5), while a reference sample is labelled with the other dye (usually Cy3) and co-hybridised.

Performing replicates

Replicates are essential for reliably detecting differentially expressed genes in microarray experiments. Without replicates, no statistical analysis of the significance and reliability of the observed changes is possible; the typical result is an increased number of both false-positive and false-negative errors in detecting differentially expressed genes. There are two types of replicates: technical replicates and biological replicates (see the Experimental design and data analysis in RNA-seq experiments [17] section).

In Expression Atlas [18], a minimum acceptable number of biological sample replicates (three) is enforced to ensure sufficient statistical power to detect differential expression (3 [7]).

Analysis of microarray data

Microarrays can be used in many types of experiments including genotyping, epigenetics, translation profiling and gene expression profiling.

Gene expression profiling is by far the most common use of microarray technology. Both one and two colour microarrays can be used for this type of experiment. The process of analysing gene expression data is similar for both types of microarrays and involves (Figure 4):

1. feature extraction;
2. quality control;
3. normalisation;
4. differential expression analysis;
5. biological interpretation of the results;
6. submission of data to a public database.
Figure 4 Overview of the microarray data analysis pipeline.
We will now go through each of these steps in more detail.

**Feature extraction**

Feature extraction is the process of converting the scanned image of the microarray into quantifiable (computable) values and annotating it with the gene IDs, sample names and other useful information (Figure 5) (4 [7]).

![Feature extraction process](image.png)

**Figure 5** Feature extraction involves the conversion of the scanned microarray image to quantifiable values that are saved in binary (e.g. CEL) or text format.

This process is often performed using the software provided by the microarray manufacturer. The output of this process is raw (i.e. unprocessed) data files that can be in binary or text format (Table 1).

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Typical raw data format</th>
<th>How to open / Analysis software examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix</td>
<td>.CEL (binary)</td>
<td>R packages (affy, limma, oligo…)</td>
</tr>
<tr>
<td>Agilent</td>
<td>feature extraction file (tab-delimited text file per hybridisation)</td>
<td>Spreadsheet software (Excel, OpenOffice, etc.)</td>
</tr>
<tr>
<td>GenePix (scanner)</td>
<td>.gpr (tab-delimited text file per hybridisation)</td>
<td>Spreadsheet software (Excel, OpenOffice, etc.)</td>
</tr>
<tr>
<td>Illumina</td>
<td>.idat (binary)</td>
<td>R packages (e.g. illuminaio)</td>
</tr>
<tr>
<td></td>
<td>.txt (tab-delimited text matrix for all samples)</td>
<td>Spreadsheet software (Excel, OpenOffice, etc.)</td>
</tr>
<tr>
<td>Nimblegen</td>
<td>NimbleScan, .pair (tab-delimited text matrix for all samples)</td>
<td>Spreadsheet software (Excel, OpenOffice, etc.)</td>
</tr>
</tbody>
</table>

After the feature extraction process, the data can be analysed. Array manufacturers often provide software to open and analyse their raw data files. These programs may not always be available, may become obsolete after a few years, or may not be flexible enough for your needs. There are several free software tools that are suitable for the downstream processing of microarray files. Examples are the Galaxy platform [19], GenePattern [20], GeneSpring [21] (licence required) and the statistics software R [22].

Quality control

Quality control of microarray data begins with the visual inspection of the scanned microarray images to make sure that there are no obvious splotches, scratches or blank areas (4 [7]).

After feature extraction, the data analysis software packages can be used to make diagnostic plots (for example of background signal, average intensity values and percentage of genes above background) to help identify problematic arrays, reporters or samples (Figure 6).

![Quality control plots](image)

**Figure 6** Examples of quality control plots made when analysing differential expression data in Expression Atlas. Left to right: Array intensity distributions, PCA plot, density estimates. See the [Expression Atlas help pages](https://www.ebi.ac.uk/training/online) for an explanation of what each of these plots represents.

Normalisation

Normalisation [26] of microarray [15] data is used to control for technical variation between assays, while preserving the biological variation (4 [7]). There are many ways to normalise the data, and the methods used depend on:

- the type of array;
- the design of the experiment;
- assumptions made about the data (e.g. 'the majority of genes represented on the microarray are not expected to be differentially expressed [27] in the test group relative to controls');
- and the package being used to analyse the data.

For Expression Atlas [18], Affymetrix [13] microarray data is normalised using the 'Robust Multi-Array Average' (RMA [28]) method within the 'oligo' package.

Agilent microarray data is normalised using the 'limma' package: 'quantile normalisation' for one-colour microarray data; 'Loess normalisation' for two colour microarray data.

Find out more about normalisation methods

For a detailed discussion of normalisation methods see Grant et al. 2007 (4 [7]).
Differential expression analysis

The goal of differential expression analysis is to identify genes whose expression differs under different conditions. An important consideration for differential expression analysis is correction for multiple testing. This is a statistical phenomenon that occurs when thousands of comparisons (e.g. the comparison of expression of multiple genes in multiple conditions) are performed for a small number of samples (most microarray experiments have less than five biological replicates per condition). This leads to an increased chance of false positive results (4 [7]).

For Expression Atlas, the 'limma' package that is used to identify differentially expressed genes incorporates a method to correct for multiple testing. This method creates a log$_2$ fold change ratio between the test and control condition and an 'adjusted' p-value that rates the significance of the difference (5 [7]).

Biological interpretation of gene expression data

Many of the methods for visualising and interpreting microarray data can also be used for RNA-seq experiments. Some common ways of visualising and interpreting gene expression data are discussed later in this course [29].

Submission of data to a public repository

Once you have finished generating microarray data (e.g. raw data files and processed/normalised data files are ready), it is important that you submit the data files together with metadata to a public database such as ArrayExpress. This helps to ensure the reproducibility of your experiment and is now a requirement of many journals and funding bodies.

Learn more about submitting your data in part III of this functional genomics course [30].

Limitations of microarrays

Hybridisation-based approaches are high throughput and relatively inexpensive, but have several limitations which include (6 [7]):

- reliance upon existing knowledge about the genome sequence;
- high background levels owing to cross-hybridisation [31];
- limited dynamic range of detection owing to both background and saturation signals;
- comparing expression levels across different experiments is often difficult and can require complicated normalisation methods [32].

Next Generation Sequencing (NGS)

In contrast to microarray methods, sequence-based approaches directly determine the nucleic acid sequence of a given DNA or cDNA molecule.

The first major foray into DNA sequencing was the Human Genome Project. This project, which used first-generation sequencing, known as Sanger sequencing (the chain-termination method), took 13 years, cost $3 billion
and was completed in 2003.

Compared to conventional Sanger sequencing using capillary electrophoresis, the short read, massively parallel sequencing technique is a fundamentally different approach that revolutionised sequencing capabilities and launched the second-generation sequencing methods - or next-generation sequencing (NGS) - that provide orders of magnitude more data at much lower recurring cost.

There are several different types of next generation sequencing technologies but the sequencing by synthesis method (used by Illumina sequencing platforms) is by far the most popular.

These Youtube videos give a useful overview of how Sanger sequencing and the different next generation sequencing technologies work and when to use them.

- [Next Generation Sequencing (NGS) - An Introduction](https://www.youtube.com/watch?v=kbhjJZlU-0M) [33]
- [When do I use Sanger Sequencing vs. NGS? - Seq It Out #7](https://www.youtube.com/watch?v=J5JqWQd1d-A) [34]

**Advantages of NGS**

NGS can be used to analyse DNA and RNA samples and is a popular tool in functional genomics. In contrast to microarray methods, NGS-based approaches have several advantages including:

- *a priori* knowledge of the genome or genomic features is not required;
- it offers single-nucleotide resolution, making it possible to detect related genes (or features), alternatively spliced transcripts, allelic gene variants and single nucleotide polymorphisms;
- higher dynamic range of signal;
- requires less DNA/RNA as input (nanograms of materials are sufficient);
- higher reproducibility.

If you want to learn more about NGS, you can take the [EBI Train online course: Next Generation Sequencing Practical Course](https://www.ebi.ac.uk/train/online/courses/next-generation-sequencing-practical-course) [35].

**RNA sequencing**

RNA sequencing (RNA-seq) is the application of next generation sequencing technologies to cDNA molecules. This is obtained by reverse transcription from RNA, in order to get information about the RNA content of a sample. Thus, RNA-seq is the set of experimental procedures that generates cDNA molecules derived from RNA molecules, followed by sequencing-library construction and massively parallel deep sequencing.

**Advantages of RNA-seq over hybridisation-based approaches**

RNA-seq provides several advantages over hybridisation-based approaches:

- RNA-seq has higher sensitivity for genes expressed either at low or very high level and higher dynamic range of expression levels over which transcripts can be detected (> 8000-fold range). It also has lower
technical variation and higher levels of reproducibility.

- RNA-seq is not limited by prior knowledge of the genome of the organism. Moreover, it can be performed in species for which genomes are not yet available, making RNA-seq particularly attractive for non-model organisms.

- RNA-seq gives unprecedented detail (to a single base resolution) about transcriptional features, such as novel transcribed regions, alternative splicing and allele-specific expression.

- Finally, whilst it is well known that microarrays are subject to cross-hybridisation [31] bias, RNA-seq is considered unbiased. However, several studies have observed a guanine-cytosine content bias in RNA-seq data and RNA-seq can suffer from mapping ambiguity for paralogous sequences.

**Applications of RNA-seq**

RNA-seq can be applied to a broad range of scientific questions such as:

- gene expression profiling between samples;
- study of alternative splicing events (differential inclusion/exclusion of exons in the processed RNA product after splicing of a precursor RNA segment) associated with diseases;
- identification of allele-specific expression, disease-associated single nucleotide polymorphisms (SNPs) and gene fusions to understand, e.g. disease causal variants in cancer.

Furthermore, single-cell RNA-seq has recently emerged as a way to study complex biological processes, cellular heterogeneity, and diversity, especially in the stem cell biology and neuroscience fields ([7,8](#)).

**Performing a RNA-seq experiment**

There is no common pipeline for all the different types of RNA-seq. Both the experimental design and the analysis procedures will be different depending on the organism being studied and the research goals. As the primary objective of many biological studies is a comparison of gene expression between samples, we outline a possible workflow for detecting differential expression using RNA-seq (Figure 7).
Functional genomics (II): Common technologies and data analysis methods

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

When we design an RNA-seq experiment we have to consider the following aspects (9 [7]):

- library construction;
- sequencing depth or library size;
- number of replicates.

Library preparation

Library preparation methods vary depending on the type of RNA being used, whether strand specificity is necessary and the type of reads.

One important aspect is choosing which population of RNA will be converted to a cDNA library. Using total RNA allows detection of non-coding as well as mRNAs, but may require additional enrichment steps (e.g. ribosomal RNA depletion) to allow the detection of transcripts of low abundance. PolyA⁺ RNA enrichment can be used to purify mRNAs and is useful in studies of eukaryotic organisms.

Figure 8 Paired-end sequencing and alignment. Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. Image courtesy of Illumina [36], Inc.

Another consideration is whether to generate a strand-specific library that retains the orientation of the original RNA transcript, which may be critical to identify antisense or non-coding RNA [37].

Furthermore, sequencing can involve single-end (SE) or paired-end (PE) reads. Paired-end sequencing means sequencing both ends of the cDNA fragments and aligning the forward and reverse reads as read pairs (Figure 8). Paired-end reads are preferable for de novo transcript discovery or isoforms expression analysis, as well as to characterise poorly annotated transcriptomes.

Sequencing depth or library size

Sequencing depth or library size refers to the number of sequenced reads for a given sample. As the sample is sequenced to a deeper level, the reads are likely to cover a larger proportion of the genome/transcriptome [38], allowing more transcripts to be detected with more precise quantification. Optimal sequencing depth depends on the aims of the experiment and on the complexity of the target transcriptome.
Number of replicates

The number of replicates in a RNA-seq experiment depends on:

1. Technical variability in the RNA-procedures. Technical replicates are replicates where the biological material is the same but the technical steps used to measure gene expression are repeated (Figure 9). Technical reproducibility in RNA-seq is usually high at the level of sequencing but RNA-seq library preparation includes a number of steps (RNA fragmentation, cDNA synthesis, adapter ligation [39] and PCR amplification) that may introduce biases in the data.

2. Biological variability of the system under study. Biological replicates consist of different biological samples that are processed separately (Figure 9). Biological replicates are required if inference on the population is to be made, with three biological replicates being the minimum for any inferential analysis.

3. Desired statistical power, that is the capacity for detecting statistically significant differences in gene expression between experimental groups.

![Figure 9 Technical replicates and biological replicates in a RNA-seq experiment.](image)

Laboratory performance

**Library preparation**: a population of RNA (total or fractionated such as polyA⁺) is converted to a library of cDNA fragments with adapters (synthetic oligonucleotides [12] of a known sequence) attached to one or both ends. Adapter-ligated cDNA fragments are then amplified in preparation for sequencing.

**Sequencing and imaging**: to obtain the nucleic acid sequence from the amplified library, each molecule is
sequenced in a high throughput manner to obtain millions of short reads (sequences) from one end (single-end sequencing) or both ends (pair-end sequencing). Sequencing is usually performed by core facilities or external companies (using different platforms such as Illumina GA/HiSeq, Applied Biosystems SoLiD, Roche 454 Life Science). Although the platforms differ substantially in their chemistry and processing steps, the sequencing process usually generates millions of short (25-300 bp) reads with associated quality scores as FASTQ [40] files (Box 1).

**Box 1** FASTQ format.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

There are four line types in the FASTQ format:
1. A `@` title line which often holds just a record identifier.
2. Nucleotide sequence of the read.
3. A `+' line to signal the end of the sequence lines and the start of the quality string.
4. Quality score for each nucleotide of the read.

**Data analysis**

Once sequencing is complete, raw sequence data (FASTQ files) must undergo several analysis steps. The exact methods vary depending on the experiment but include:

- quality control;
- read alignment (with or without a reference genome);
- quantification of gene and transcript expression;
- and detection of differentially expressed genes (89 [7]).

We will go through each of these steps in more detail over the next few pages.
Quality control

This step involves pre-processing [41] the data to remove:

- adapter sequences (adapter trimming);
- low-quality reads;
- uncalled bases;
- and to filter out contaminants (sequences which don’t derive from the source organism). It is important to check that sequence quality is similar for all samples and discard outliers.

FastQC [42] is a popular tool to perform quality assessment. As a general rule, read quality decreases towards the 3’ end of reads, and if it becomes too low, bases should be removed to improve mappability [43].

Trimmomatic [44] can be used to remove PCR primers, adapter sequences and to trim the lower-scored bases and low quality N bases.

Read mapping or alignment

Once high-quality data are obtained from pre-processing [41], the next step is the read mapping or alignment. There are two main options depending on the availability of a genome sequence (Figure 10):

- When studying an organism with a reference genome, it is possible to infer which transcripts are expressed by mapping the reads to the reference genome (genome mapping) or transcriptome [38] (transcriptome mapping). Mapping reads to the genome requires no knowledge of the set of transcribed regions or the way in which exons are spliced together. This approach allows the discovery of new, unannotated transcripts.

- When working on an organism without a reference genome, reads need to be assembled first into longer contigs [45] (de novo assembly). These contigs can then be considered as the expressed transcriptome to which reads are re-mapped for quantification.
Alignment tools

There are many bioinformatics tools available to perform the alignment of short reads. One of the most popular RNA-seq mappers is TopHat [47], which aligns reads in two steps:

1. unspliced reads are mapped to locate exons (with Bowtie [48]);
2. unmapped reads are then split and aligned independently to identify exon [49] junctions [9] [7]).

It is important to check the quality of the mapping process. The percentage of mapped reads is a global indicator of the overall sequencing accuracy and of the presence of contaminating DNA. Picard [50] can be used for quality control in mapping.

Either with a reference or de novo assembly, the complete reconstruction of transcriptomes using short reads is challenging. For example, short reads can sometimes align equally well to multiple locations (multi-mapped reads or multi-reads). Paired-end reads reduce the problem of multi-mapping, because a pair of reads must map within a certain distance of each other and in a certain order (Figure 8 [51]). Finally, long-read technologies, such as SMRT from Pacific Biosciences [52], provide reads that are long enough to sequence complete transcripts for most genes and are a promising alternative.

Quantification
The simplest approach to quantifying gene expression by RNA-seq is to count the number of reads that map (i.e. align) to each gene (read count) using programs such as HTSeq-count [53]. This gene-level quantification approach utilizes a gene transfer format (GTF) file containing gene models, with each model representing the structure of transcripts produced by a given gene.

Raw read counts are affected by factors such as transcript length (longer transcripts have higher read counts, at the same expression level) and total number of reads. Thus, if we want to compare expression levels between samples, we need to normalize the raw read counts. The measure RPKM (reads per kilobase of exon model per million reads) and its derivative FPKM (fragments per kilobase of exon model per million reads mapped) account for both gene length and library size effects.

Correcting for gene length is not necessary when comparing changes in gene expression within the same gene across samples. However, it is necessary for correctly ranking gene expression levels within the sample to account for the fact that longer genes accumulate more reads (at the same expression level).

### Differential gene expression analysis

Differential expression analysis means taking the normalized read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

### Methods for differential expression analysis

There are different methods for differential expression analysis such as edgeR [54] and DESeq [55] based on negative binomial (NB) distributions or baySeq [56] and EBSeq [57] which are Bayesian [58] approaches based on a negative binomial model [59]. It is important to consider the experimental design when choosing an analysis method. While some of the differential expression tools can only perform pair-wise comparison, others such as edgeR, limma-voom, DESeq and maSigPro can perform multiple comparisons.

In Figure 11, below, we outline the RNA-seq processing pipeline used to generate data for Expression Atlas.
Figure 11 RNA-seq processing pipeline used to generate gene expression data in Expression Atlas.

Figure 11 description:
Raw reads (FASTQ files) undergo quality assessment and filtering. The quality-filtered reads are aligned to the reference genome via TopHat2. The mapped reads are summarised and aggregated over genes via HTSeq. For baseline expression, the FPKMs are calculated from the raw counts by iRAP [60]. These are averaged for each set of technical replicates, and then quantile normalised within each set of biological replicates using limma [24].

Finally, they are averaged for all biological replicates (if any). For differential expression, genes expressed differentially between the test and the reference groups of each pairwise contrast are identified using DESeq2.

Biological interpretation of gene expression data

Many of the methods for visualising and interpreting gene expression data can be used for both microarray and RNA-seq experiment. Some of the most common methods are discussed below.
Heatmaps and clustering

A common method of visualising gene expression data is to display it as a heatmap (Figure 12). The heatmap may also be combined with clustering methods which group genes and/or samples together based on the similarity of their gene expression pattern. This can be useful for identifying genes that are commonly regulated, or biological signatures associated with a particular condition (e.g. a disease or an environmental condition) ([4] [7]).

In heat maps the data is displayed in a grid where each row represents a gene and each column represents a sample. The colour and intensity of the boxes is used to represent changes (not absolute values) of gene expression. In the example below, red represents up-regulated genes and green represents down-regulated genes. Black represents unchanged expression.

**Figure 12** An example of a heatmap in which genes have been grouped based on their pattern of gene expression.
Gene set enrichment analysis and pathway analysis

A common approach to interpreting gene expression data is gene set enrichment analysis based on the functional annotation of the differentially expressed genes (Figure 13). This is useful for finding out if the differentially expressed genes are associated with a certain biological process or molecular function.

The Gene Ontology [61], containing standardised annotation of gene products, is commonly used for this purpose. It works by comparing the frequency of individual annotations in the gene list (e.g. differentially expressed genes) with a reference list (usually all genes on the microarray [15] or in the genome). Enrichment of biological pathways supplied by KEGG, Ingenuity, Reactome [62] or WikiPathways can be performed in a similar way (11,12 [7]).

Popular tools for gene set enrichment and pathway analysis

- DAVID [63] (free online tool)
- GSEA [64] (free)
- Ingenuity [65] (licence required)
- Reactome [66] (free)

![Figure 13](image.png) An example of a pathway from Wikopathways.

Network analysis

Network analysis is complementary to pathway analysis and can be used to show how key components of different pathways interact. This can be useful for identifying regulatory events that influence multiple biological processes.
and pathways (11,12 [7]).

Genotyping, epigenetic and DNA/RNA-protein interaction methods

Genotyping

Variations in the DNA sequence such as copy number variations [67], insertions, deletions and SNPs are referred to as the genotype [68] of the sample and can be detected using (13,14 [7]):

1. **Comparative genome hybridisation**
   Both the test and a well-characterised reference sample are hybridised to the same microarray and the relative intensity between the two samples is used to infer copy number variation.

2. **Whole genome sequencing**
   The genome of the test sample is sequenced and computationally compared to a reference genome to identify many types of genetic variation.

3. **Exome sequencing**
   Uses sequence-specific probes in solution or fixed to microarrays to purify coding regions of the DNA. These regions are then amplified, sequenced and compared to a reference to identify variations in coding regions.

Epigenetic modifications

Epigenetic modifications are heritable chemical or physical changes in chromatin. There are two types of epigenetic modifications - DNA methylation [69] and histone modifications (15 [7]).

DNA methylation

DNA methylation involves the addition of a methyl or hydroxymethyl group to bases in the DNA sequence. The most commonly studied modification is the methylation of the C5 position on cysteine bases, or m5c (16,17 [7]).

DNA methylation can be measured using chromatin immunoprecipitation (ChIP) or bisulfite-based methods (Figure 14).

**ChIP-based methods**

ChIP-based methods use methylation-specific antibodies to purify methylated regions of the genome. The DNA is then analysed by microarray or NGS to identify these regions (16,17 [7]).
Figure 14 Methods for detecting DNA methylation. Immunoprecipitation and bisulfite-based methods can be used in combination with microarrays or next generation sequencing to identify regions of the genome that have been epigenetically modified.

Bisulphite-based methods

Bisulphite-based methods involve bisulphite treatment of the DNA sample, which converts unmethylated cytosine bases to a uracil base, while leaving methylated residues as a cytosine. The treated DNA is then analysed by microarray or NGS. This method can only detect types of methylation that are susceptible to bisulphite-induced changes (16,17 [7]).

For microarray analysis, the bisulphite-treated sample is competitively hybridised with an untreated sample. The signal intensity ratio between the samples is used to infer the level of methylation in specific regions of the genome.

For NGS, the treated and untreated samples are sequenced and compared to identify specific methylation sites within the genome. Bisulfite sequencing requires a well annotated genome.

Epigenetic modifications of RNA such as m6A methylation have recently been described and are thought to regulate RNA stability and localisation (18 [7]). These modifications can not be detected by bisulphite-based methods. They can be detected by MeRIP-Seq (methylated RNA immunoprecipitation sequencing), which combines immunoprecipitation of modified sequences with RNA-seq analysis.

Histone modifications

There are many types of histone modifications, including acetylation, methylation, ubiquitination, citrullination and...
phosphorylation of specific amino acids within the histone protein, usually towards the C-terminal ('tail') end of the protein. These modifications can both positively and negatively regulate gene expression by changing the way in which histones bind to DNA (19 [7]).

Histone modifications can be detected using a variety of techniques including mass spectrometry [70] and genomics approaches such as ChIP-chip [71] and ChIP-seq. The genomics approaches combine chromatin immunoprecipitation [72] (ChIP) of specific, modified histones and their associated DNA with microarray [15] (chip) or NGS (seq) of the DNA molecules to identify regions of the genome associated with these modifications (19 [7]) (Figure 15).

**Figure 15** Immunoprecipitation methods combined with microarrays or next generation sequencing can be used to
detect histone modifications.

**DNA/RNA-protein interactions**

DNA-protein and RNA-protein interactions can have a profound effect on gene expression and the spatial and temporal localisation of mRNA within a cell.

DNA-protein interactions include those between DNA and transcription factors or other regulatory proteins. RNA-protein interactions include those between RNA and the ribosome, and other RNA-binding proteins (20,21 [7]).

The analysis of DNA-protein and RNA-protein interactions uses similar methods which are based on:

1. immunoprecipitation of the nucleic acid-protein complex;
2. purification and analysis of the DNA/RNA by microarray or NGS to identify the regions of the genome or transcriptome that are interacting with a specific protein at a specific point in time (22 [7]) (Figure 16).

![Figure 16](image)

**Figure 16** Immunoprecipitation methods combined with microarrays or sequencing can be used to identify genes or transcripts bound to specific proteins.

DNA–protein analyses are referred to as ChIP-chip or ChIP-seq, and require the crosslinking of DNA-protein with formaldehyde before the immunoprecipitation.

For RNA-protein analyses the cross-linking step is optional. RNA-immunoprecipitation (RIP)-chip/seq analyses can be performed with or without formaldehyde-based cross-linking before immunoprecipitation. Without cross-linking, this method is not well suited to detecting transient RNA-protein interactions and is prone to high background. Cross-linking and immunoprecipitation (CLIP)-chip/seq analyses use ultra violet cross-linking of RNA and proteins before immunoprecipitation, thus overcoming many of the limitations of RIP-based methods (20,22,23 [7]).
Summary

What kind of technologies can be used?

The type of technology used in a functional genomics experiment depends on the scale and intent of the experiment. Some commonly used technologies are real-time PCR, microarray analysis, next generation sequencing and RNA-seq.

How do you analyse functional genomics data?

The way functional genomics data is analysed varies depending on the technology used and the design and focus of the experiment. For gene expression studies, analysis involves quality control, data normalisation and statistical analysis to identify differentially expressed genes. Common tools for the visualisation and interpretation of gene expression data are hierarchical clustering (heatmaps), gene set enrichment analysis and pathway/network analysis. For high-throughput technologies such as RNA-seq data analysis will require some programmatic skills and bioinformatics expertise.

Quiz: Check your learning

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

Recommended online courses

- Functional genomics (I): Introduction and designing experiments [3]
- ArrayExpress: Discover functional genomics data quickly and easily [73]
- ArrayExpress: Quick tour [74]
- Expression Atlas: Quick tour [75]

Contact us

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


Contributors

Laura Huerta [1]
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.

Melissa Burke [2]
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.

Amy Tang
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 [100], 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 [101] and Expression Atlas [102] databases in the Gene Expression Team [103], promoting reproducible research by maintaining curation quality, and also by driving the development of the user-friendly ArrayExpress data submission tool Annotare [104]. 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 [105] GeneBuild team, curating gene models for key model organisms such as mouse and rat.

Source URL: https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods

Links
[1] https://www.ebi.ac.uk/training/online/trainers/lauhuema
[2] https://www.ebi.ac.uk/training/online/trainers/mburke
[4] https://www.ebi.ac.uk/training/online/glossary/real-time-pcr
[5] https://www.ebi.ac.uk/training/online/glossary/next-generation-sequencing
[6] https://www.ebi.ac.uk/training/online/glossary/gene-expression
[8] https://www.ebi.ac.uk/training/online/glossary/probe
[9] https://www.ebi.ac.uk/training/online/glossary/quencher
[10] https://www.ebi.ac.uk/training/online/glossary/multiplex-assays
[12] https://www.ebi.ac.uk/training/online/glossary/oligonucleotides
[13] https://www.ebi.ac.uk/training/online/glossary/affymetrix
[70] https://www.ebi.ac.uk/training/online/glossary/mass-spectrometry
[71] https://www.ebi.ac.uk/training/online/glossary/chip-chip
[72] https://www.ebi.ac.uk/training/online/glossary/chromatin-immunoprecipitation
[73] https://www.ebi.ac.uk/training/online/course/arrayexpress-discover-functional-genomics-data-qui
[74] https://www.ebi.ac.uk/training/online/course/arrayexpress-quick-tour-1
[75] https://www.ebi.ac.uk/training/online/course/arrayexpress-quick-tour-0
[76] http://europepmc.org/articles/PMC1435885
[77] http://europepmc.org/articles/PMC2920066
[78] http://europepmc.org/articles/PMC3964963
[80] http://europepmc.org/articles/PMC4402510
[81] http://europepmc.org/articles/PMC2949280
[83] http://europepmc.org/articles/PMC4648566
[88] http://europepmc.org/articles/PMC4108431
[97] http://europepmc.org/articles/PMC3624891
[99] https://www.ebi.ac.uk/training/online/trainers/amytang
[100] https://sites.google.com/site/brockdorfflab/
[101] http://www.ebi.ac.uk/arrayexpress
[102] http://www.ebi.ac.uk/gxa
[103] http://www.ebi.ac.uk/~rpetry/geteam/
[104] https://www.ebi.ac.uk/fg/annotare/