EMBO Practical Course on Analysis of High-Throughput Sequencing Data

Gabry Rustici [1]

- Gene Expression
- Intermediate
- > 3 hours

Here you will find a collection of materials including videos of lectures from the 'EMBO Practical Course on Analysis of High-Throughput Sequencing Data' that took place at EMBL-EBI on 23-29 October 2011.

This course was aimed at advanced PhD students and post-doctoral researchers who are applying or planning to apply high throughput sequencing technologies and bioinformatics methods in their research.

The aim of this course was to familiarize the participants with advanced data analysis methodologies and provide hands-on training on the latest analytical approaches.

Learning objectives:

- Lectures give insight into how biological knowledge can be generated from RNA-Seq and ChIP-Seq experiments and illustrate different ways of analysing such data
- Practicals consist of computer exercises that will enable the participants to apply statistical methods to the analysis of RNA-Seq and ChIP-seq data, mostly using R and Bioconductor

How to take this course

This course makes use of lecture material from a seven-day course called 'EMBO Practical Course on Analysis of High-Throughput Sequencing' that took place at EMBL-EBI in October 2011.

If you see this video icon, you will be able to watch a video of the lecture, presented in Mediasite [2] format. (Note that these videos will only run on Google Chrome [3], Internet Explorer or Safari.)
Practical handouts and the code used in the lectures are available for you to download.

You can go through all of the lectures and exercises and study the full course or you can pick out lectures that are relevant to you. If studying the whole course, please note that it is not designed to be completed in a single day and is best carried out bit-by-bit over several days.

Written introductions are provided for each lecture to help you decide which ones to study.

**Course Overview**

The course agenda can be found [here](https://www.ebi.ac.uk/training/online).

Click here to watch a short overview video [6] of the course (4 minutes)

Please be aware that the practicals in this course were tested on R 2.14.0 (development version). Most exercises will still work with the current R release [7] but the results obtained might slightly differ or the code used might need to be modified.

**Introduction to R and Bioconductor**

In this lecture, Nicolas Delhomme, a bioinformatician from the Furlong Group at EMBL Heidelberg, provides an introduction to R [7] and Bioconductor [8], which is the software that will be used throughout the course to perform analysis of next generation sequencing [9] data, focusing on post-alignment analysis steps.

Click here to watch the video [10] (18 minutes)

Here he briefly describes the Bioconductor packages available for short read analysis. This is just a lightweight introduction and the packages are described more extensively in the next session.

**Bioconductor packages for short read analysis - An Overview**

In this lecture, Nicolas Delhomme describes the Bioconductor [8] packages available for short read analysis. He provides an introduction to the various applications of these packages, covers what are
the issues of the technologies that are commonly used in next generation sequencing [9] and walks users through a typical workflow in the analysis of next generation sequencing data.

Click here to watch the video [11] (60 minutes)

The workflow will cover:

- quality assessment of the data;
- alignment of short reads to the reference genome;
- pre-processing [12] of the data;
- working with the relevant annotation [13] for your data;
- determine a table of counts (for RNA-Seq [14]) or peaks (for ChIP-Seq);
- export and visualise the results.

The code used by Nicolas during this lecture can be found here [15].

In order to run this code you will need to install some of the following Bioconductor packages in your R session:

- RnaSeqTutorial [16]
- GenomicRanges [17]
- biomaRt [18]

You can access the materials and data used in this session here [19].

**Bioconductor packages for short read analysis - Details**

In this lecture, Nicolas Delhomme describes in detail the Bioconductor [8] packages that can be used for the analysis of short reads.

Click here to watch the video [20] (48 minutes)

Particular emphasis is put on the following packages:

- ShortRead [21] and Rsamtools [22] for file I/O, quality assessment, and high-level, general purpose data summary.
- IRanges [23] and GenomicRanges [17] for range-based (e.g., chromosomal regions) calculation, data manipulation, and general-purpose data representation.
- **BSgenome** [25] for accessing and manipulating curated whole-genome representations.
- **Biostrings** [26] for alignment, **pattern** [27] matching (e.g., primer removal), and data manipulation of large biological sequences or sets of sequences.
- **rtracklayer** [28] for import and export of tracks on the **UCSC genome browser** [29].
- **biomaRt** [18] for access to Biomart databases.

The code used by Nicolas during this lecture can be found [here] [30].

This lecture was followed by a practical; the handout for the practical can be found [here] [31], along with the code [32] and two additional packages developed by Nicolas for this course, **easyRNASeq** [33] and **RnaSeqTutorial** [34].

In order to follow this practical you will need to install some of the following Bioconductor packages in your R session.

- biomaRt
- Biostrings
- BSgenome
- BSgenome.Dmelanogaster.UCSC.dm3 [35]
- BSgenome.Hsapiens.UCSC.hg19 [36]
- EatonEtAlChIPseq [37]
- genomeIntervals [38]
- GenomicFeatures
- GenomicRanges
- hgu95av2probe [39]
- IRanges
- leeBamViews [40]
- Rsamtools
- rtracklayer
- ShortRead

Please refer to the practical handout [31] to identify the packages that you will need.

**Bioconductor packages for short read analysis - Quality Assessment**

In this lecture, Nicolas Delhomme covers quality assessment of next generation sequencing [9] data commenting on the different quality metrics available through the Bioconductor [8] package **ShortRead** [21].

[Click here to watch the video](41) (6 minutes)

The various plots that can be used to assess the quality of a sequencing run are presented.

**Introduction to ChIP-Seq data and analysis**
In this lecture, Bori Gerle, a postdoctoral fellow in the Luscombe Group at EMBL-EBI, gives an introduction to ChIP-seq data and its analysis.

Click here to watch the video [42] (37 minutes)

The lecture was followed by a practical; the handout [43] and the code [44] for this practical are provided. The dataset used for this practical is taken from Kasowski et al, (2010), Science 238(232-235). The dataset is available at GEO [45], accession [46] number: GSE19486.

In order to follow this practical you will need to install the following Bioconductor [8] packages in your R session:

- chipseq [47]
- BayesPeak [48]
- GenomicFeatures [24]
- ShortRead [21]
- rtracklayer [28]
- BSgenome [25]
- BSgenome.Hsapiens.UCSC.hg19 [36]
- seqLogo [49]
- DESeq [50]

In addition, you will need to install the Integrated Genome Browser (IGB [51]) to visualise the results of the practical.

Statistical Concepts and Methodologies for Data Analyses

In these two lectures (part 1 and part 2), Benilton Carvalho, from the Computational Biology and Statistics Group Department of Oncology at the University of Cambridge, talks about the statistical concepts and the methodologies involved in the analysis of next generation sequencing [9] data.

Click here to watch part 1 [52] (90 minutes)

Click here to watch part 2 [53] (93 minutes)

These lectures were followed by a short practical [54] that illustrated several of the concepts presented by Benilton in his lectures.

In order to follow this practical you will need to install the following Bioconductor [8] packages in your R session:
Differential Expression for RNA-Seq

In this lecture, Wolfgang Huber, Group Leader in the Genome Biology Unit at EMBL Heidelberg, gives an introduction to RNA-Seq data as well as the statistical and computational methods used to analyse such data.

Click here to watch the video [58] (78 minutes)

This lecture was followed by a practical that focused on the use of the Bioconductor [8] DESeq [50] package for the analysis of differential expression and RNA-seq style analysis. The practical is the vignette of the DESeq package and can be downloaded here [59], together with the code [60].

In order to follow this practical you will need to install the following Bioconductor packages in your R session:

- DESeq [50]
- pasilla [61]

Allele-specific expression and eQTL

In this lecture, John Marioni, Research Group Leader in Computational Biology at EMBL-EBI, talks about allele [62]-specific expression and eQTL. He provides biological motivation as well as examples of such applications.

Click here to watch the video [63] (66 minutes)

Summary

Various statistical packages are available in R and Bioconductor [8] to perform analysis of RNA-Seq [14] and ChIP-Seq data, focusing on the post-alignment analysis steps.

A typical workflow in the analysis of next generation sequencing [9] data includes:

- quality assessment of the data;
- alignment of short reads to the reference genome;
- pre-processing [12] of the data;
- working with the relevant annotation [13] for your data;
determining a table of counts (for RNA-Seq) or peaks (for ChIP-Seq) for downstream analysis;
visualising the data.

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.

Contributors

Gabry Rustici [1]

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.

Acknowledgements

We would like to express our gratitude to Nicolas Delhomme, Benilton Carvalho, Kathi Zarnack,
Borbala Gerle, John Marioni and Wolfgang Huber for contributing to this course and for all the hard
work they put into preparing the course materials. We also wish to thank Carl Bergman for filming
the lectures and Hugo [64] Neves, from the EMBL Photolab, for the post-processing of the lectures.

Financial support

This course is sponsored by EMBO [65].

Source URL: https://www.ebi.ac.uk/training/online/course/embo-practical-course-analysis-high-
throughput-seq

Links
[1] https://www.ebi.ac.uk/training/online/trainers/gabry
[52] http://medias01-web.embl.de/Mediasite/Play/5ed8ac0f2fef4e2e89d8accb87d8253d1d
[53] http://medias01-web.embl.de/Mediasite/Play/8dfc4c50907b4df6a9b8b8df9cd833721d
[54] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/user/18/private/statconceptsmethodspractical.pdf
[58] http://medias01-web.embl.de/Mediasite/Play/393c133b0c154632b2a3b5769b1608a41d
[62] https://www.ebi.ac.uk/training/online/glossary/allele
[63] http://medias01-web.embl.de/Mediasite/Play/7a15ffa352f64224a9cdd990c8bc64e01d
[64] https://www.ebi.ac.uk/training/online/glossary/hugo