EBI: Next Generation Sequencing Practical Course

Emily Perry

DNA & RNA
Gene Expression
Beginner
> 3 hours

Here you will find a collection of video lectures from the 'Hands on training in Next-Generation Sequencing' course that took place at EMBL-EBI on 12-14 March 2012, available for you to watch in your own time.

The workshop provides an introduction to Next Generation Sequencing (NGS) and addresses how to interpret the data and how the analysis is applied in Genomics.

Note: the videos in this course will only run on Google Chrome, Internet Explorer and Safari.

Learning objectives:

- Understand some principles behind NGS
- Know the challenges created by NGS
- Know how to submit and retrieve NGS data to and from databases
- Understand the uses of NGS data in: Whole genome assembly; Gene expression analysis; Genome annotation; Gene regulation analysis; Variation studies

How to take this course

This course makes use of lecture material from a three-day course called 'Hand on training in Next Generation Sequencing' that took place at EMBL-EBI in March 2013.

If you see this video icon, you will be able to watch a video of the lecture, presented in Mediasite format. (Note that these videos will only run on Google Chrome, Internet Explorer or Safari.)

The slides for each lecture are also available as pdf files for you to study. Lecture links will open in a new tab and pdfs will open in your default pdf viewer (eg. Adobe Acrobat or Apple Preview).
Some of the lectures have associated exercises. Links to online courses, workbooks and data files are provided where appropriate.

You can go through all of the lectures and exercises and study the full course (some six hours of video plus reading and exercises), 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. Note that some lectures form a series and you may find it difficult to follow later lectures in a series without first viewing the earlier ones - these lectures are grouped together and should be easy to identify.

What is Next-Generation DNA Sequencing

Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies including:

- Illumina (Solexa) sequencing
- Roche 454 sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing

These recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology.

The next few pages will give you an overview of how Illumina, 454 and Ion torrent sequencing work.

Illumina sequencing

In NGS, vast numbers of short reads are sequenced in a single stroke.

To do this, firstly the input sample must be cleaved into short sections. The length of these sections will depend on the particular sequencing machinery used.

In Illumina sequencing, 100-150bp reads are used. Somewhat longer fragments are ligated to generic adaptors and annealed to a slide using the adaptors. PCR is carried out to amplify each read, creating a spot with many copies of the same read. They are then separated into single strands to be sequenced.
The slide is flooded with nucleotides and DNA polymerase. These nucleotides are fluorescently labelled, with the colour corresponding to the base. They also have a terminator, so that only one base is added at a time.

An image is taken of the slide. In each read location, there will be a fluorescent signal indicating the base that has been added.
The slide is then prepared for the next cycle. The terminators are removed, allowing the next base to be added, and the fluorescent signal is removed, preventing the signal from contaminating the next image.

Computers are then used to detect the base at each site in each image and these are used to construct a sequence.

The process is repeated, adding one nucleotide at a time and imaging in between.
All of the sequence reads will be the same length, as the read length depends on the number of cycles carried out.

### 454 sequencing

Roche 454 sequencing can sequence much longer reads than Illumina. Like Illumina, it does this by sequencing multiple reads at once by reading optical signals as bases are added.

As in Illumina, the DNA or RNA is fragmented into shorter reads, in this case up to 1kb. Generic adaptors are added to the ends and these are annealed to beads, one DNA fragment per bead. The fragments are then amplified by PCR using adaptor-specific primers.

Each bead is then placed in a single well of a slide. So each well will contain a single bead, covered in many PCR copies of a single sequence. The wells also contain DNA polymerase and sequencing buffers.

The slide is flooded with one of the four NTP species. Where this nucleotide is next in the sequence, it is added to the sequence read. If that single base repeats, then more will be added. So if we flood with Guanine bases, and the next in a sequence is G, one G will be added, however if the next part of the sequence is GGGG, then four Gs will be added.
The addition of each nucleotide releases a light signal. These locations of signals are detected and used to determine which beads the nucleotides are added to.
This NTP mix is washed away. The next NTP mix is now added and the process repeated, cycling through the four NTPs.

This kind of sequencing generates graphs for each sequence read, showing the signal density for
Each nucleotide wash. The sequence can then be determined computationally from the signal density in each wash.

All of the sequence reads we get from 454 will be different lengths, because different numbers of bases will be added with each cycle.

**Ion Torrent: Proton / PGM sequencing**

Unlike Illumina and 454, Ion torrent and Ion proton sequencing do not make use of optical signals. Instead, they exploit the fact that addition of a dNTP to a DNA polymer releases an $H^+$ ion.

As in other kinds of NGS, the input DNA or RNA is fragmented, this time ~200bp. Adaptors are added and one molecule is placed onto a bead. The molecules are amplified on the bead by emulsion PCR. Each bead is placed into a single well of a slide.

Like 454, the slide is flooded with a single species of dNTP, along with buffers and polymerase, one NTP at a time. The pH is detected in each of the wells, as each $H^+$ ion released will decrease the pH. The changes in pH allow us to determine if that base, and how many thereof, was added to the sequence read.
The pH change, if any, is used to determine how many bases (if any) were added with each cycle.

The dNTPs are washed away, and the process is repeated cycling through the different dNTP species.
Improvements on the previous technology

The four main advantages of NGS over classical Sanger sequencing are:

- speed
- cost
- sample size
- accuracy

NGS is significantly cheaper, quicker, needs significantly less DNA and is more accurate and reliable than Sanger sequencing. Let us look at this more closely. For Sanger sequencing, a large amount of template DNA is needed for each read. Several strands of template DNA are needed for each base being sequenced (i.e. for a 100bp sequence you'd need many hundreds of copies, for a 1000bp sequence you'd need many thousands of copies), as a strand that terminates on each base is needed to construct a full sequence. In NGS, a sequence can be obtained from a single strand. In both kinds of sequencing multiple staggered copies are taken for contig construction and sequence validation.

NGS is quicker than Sanger sequencing in two ways. Firstly, the chemical reaction may be combined with the signal detection in some versions of NGS, whereas in Sanger sequencing these are two separate processes. Secondly and more significantly, only one read (maximum ~1kb) can be taken at a time in Sanger sequencing, whereas NGS is massively parallel, allowing 300Gb of DNA to be read on a single run on a single chip.

The reduced time, manpower and reagents in NGS mean that the costs are much lower. The first
human genome sequence cost in the region of £300M. Using modern Sanger sequencing methods, aided by data from the known sequence, a full human genome would still cost £6M. Sequencing a human genome with Illumina today would cost only £6,000.

Repeats are intrinsic to NGS, as each read is amplified before sequencing, and because it relies on many short overlapping reads, so each section of DNA or RNA is sequenced multiple times. Also, because it is so much quicker and cheaper, it is possible to do more repeats than with Sanger sequencing. More repeats means greater coverage, which leads to a more accurate and reliable sequence, even if individual reads are less accurate for NGS.

Sanger sequencing can be used to give much longer sequence reads. However, the parallel nature of NGS means that longer reads can be constructed from many contiguous short reads.

**Key tasks in sequence analysis**

This is a talk by Aylwyn Scally from the Wellcome Trust Sanger Institute. This lecture is 52 minutes long.

He discusses:

- sequence formats and quality.
- alignment of Next-Gen sequences to reference sequences.
- identification and labelling of variants, such as SNPs and structural variations.
- sequence assembly.

For each of these topics, he covers some of the programmes available for carrying out these analyses, the important issues to consider and the kinds of file inputs and outputs involved.

[Click here to watch the video](52 minutes)

The sound quality is poor for the first few slides, please bear with us as it does improve. This is the only lecture affected.

[Click here to view the slides] [6]
This series of lectures was presented by the Sequence Read Archive (SRA) team at the EMBL-EBI. They detail the role of the SRA in the curation of raw NGS data, and how users can submit and browse data.

This introductory talk is by Rasko Leinonen from the European Nucleotide Archive (ENA), who host the SRA.

In this talk, you should learn about the aims of the SRA and gain an overview of the services they offer. You will learn about how to submit data to the SRA. This includes:

- Where to submit different kinds of NGS data.
- The format your data should be in.
- The metadata required for each sequencing project.
- How to set up a submission account.
- What happens to your submitted data.

Click here to watch the video [7] (30 minutes)

Click here to view the slides [8]

**SRA - Browsing and data retrieval**

The data submitted to the SRA is fully searchable, due to indexing from the metadata associated with it. This talk is by Marc Rossello from the ENA and lasts 17 minutes.

In his talk he gives an overview of how to search for different kinds of sequencing data. This includes searching via keywords and sequences. He also demonstrates how data is presented and stored by the SRA, and how to access it.
SRA - Webin interactive submission tool

This talk focuses on another method of submitting data to the SRA, the Webin interactive submission tool. This talk is by Arnaud Oisel from the ENA and is 11 minutes long.

The Webin tool is used by small-scale submitters, who only submit sequence data once a month or less, and to edit any data. This talk goes over the various stages of submitting and editing data via the Webin tool including:

- Setting up an account.
- Uploading files to the SRA.
- Adding metadata.
- SRA Accession numbers.
- Specifying a release date, so that your data can be released at the same time as your paper.
- Checking upload quality.
- Editing the associated metadata.
SRA - Submission process and REST XML submissions

This talk is by Rajesh Radhakrishnan from the ENA. This talk covers large scale data submission to the SRA.

In his talk Rajesh talks about:

- File submission types.
- Use of XML to add metadata objects.
- Validation of file submissions.
- Assignment of accession numbers.
SRA online tutorials

The ENA also have a number of tutorials available here.

Genome assembly

This series of talks is by Matthias Haimel who was working in Ensembl Genomes at the time, now at the University of Cambridge. These talks are intended to show how NGS data can be used for genome assembly. There are three talks, each with an exercise associated with them. It is intended that you view the talks and do the exercises in sequence.

Click here to download basic information on the exercises.
Part 1

In the first talk, Matthias introduces how we can assemble whole genomes from NGS data using the Velvet genome assembly programme. He discusses the different kinds of NGS reads that can be used. Following that is an explanation of how Velvet uses De Bruijn graphs to assemble genomes, including:

- What De Bruijn graphs are.
- How we can use them to assemble sequences.
- How we can resolve bubbles in De Bruijn graphs using Tour Bus.
- An example assembly process using De Bruijn graphs.
- The benefits and difficulties of using De Bruijn graphs.

This talk includes an introduction to the first practical, as well as questions and answers from the practical. If you do not wish to hear the questions and answers about the practical, pause the video at 28m30s.

Click here to watch the video [20] (30 minutes)

Click here to view the slides [21]

Assembly exercise 1

This exercise allows you to have a go at assembling contigs from single read NGS sequence data submitted to the SRA using Velvet.
Part 2

In this talk, Matthias looks at some of the algorithms used by Velvet, such as Pebble (used to join up the pairs of paired-end reads) and Rock band (for long read information). He also discusses how Velvet handles read quality and error correction.

Click here to watch the video [23] (10 minutes)

Click here to view the slides [24]

Assembly exercise 2

This exercise allows you to assemble contigs from paired-end reads. There is an investigation into the effect of data quality.

Click here to download the exercise pdf, [25]
Part 3

This last talk by Matthias is 10 minutes long.

In this talk he examines some of the Velvet extensions available:

- Columbus, which allows you to align reads to a reference sequence [26].
- Oases, for transcriptome [27] analysis.
- Metavelvet, to separate out and assemble a mixed pot of multiple genomes.

He also discusses some issues related to assembly quality, including:

- Read simulation.
- Testing of different methods in Assemblathon.
- Expected and observed coverage.

Click here to watch the video [28] (10 minutes)

Click here to view the slides [29]

Assembly exercise 3

This exercise covers contig assembly with long and variable length reads. It allows users to compare the parameters needed for long reads and paired-end reads.

Click here to download the exercise pdf. [30]
Exercise answers

You can download the answers to the exercises here. [31]

Variation: The 1000 Genomes Project

This talk is by Laura Clarke from the 1000 genomes project at the EMBL-EBI.

In this talk, Laura discusses the 1000 genomes project, which aims to document all human variation through NGS sequencing of the genomes of 1000s of individuals from different global populations. In the first half of the talk she covers:

- The strategy used in this project and the rationale for this strategy.
- The progress of the project.
- The pipelines involved to integrate data from a variety of sources.

In the second half of the talk, she discusses how users can access 1000 genomes data, including:

- The 1000 genomes FTP site [32].
- The 1000 genomes browser [33].
- The variant effect predictor [34].
- The variation pattern finder [34].
- Data slicing [34].
- The VCF to PED converter [34], for visualisation of haplotype data.
- The Perl API [35].

Click here to watch the video [36] (48 minutes)
ChIP-seq analysis

This talk by Remco Loos from the EMBL-EBI looks at ChIP-seq and lasts 35 minutes.

In his talk, Remco explains the principles behind ChIP (Chromatin ImmunoPrecipitation) and how NGS has improved ChIP technology but created challenges in terms of data quantity. He discusses the various stages and issues of ChIP-seq, including:

- Experimental design.
- Mapping software.
- Peak calling, and non-peak based analysis.
- Analysis of enriched areas.
- Viewing ChIP-seq data in genome browsers.
ChIP-Seq practical

This exercise allows you to do some basic analysis of ChIP-seq data. You will align ChIP-seq data to the mouse genome using Bowtie, then find immuno-enriched areas using the peak caller MACS. You will visualise the data in a genome browser and perform annotation and motif analysis on the predicted binding regions. This task takes 1.5 to 2 hours.

Click here to download the files and worksheet.

Before you download this folder, be aware that it is 118.5MB

You will need to install the following programmes and commands on your computer in order to carry out this practical:

- Bowtie
- MACS
- Samtools
- IGV
- bedGraphToBigWig
- Bedtools

In the NGS workshop, the files were pre-installed on the course computers in specific file locations. The commands in the worksheet direct to those file locations: you will need to alter the commands to go to the locations where you have saved your files.

RNA sequencing

RNAseq uses NGS to sequence extracts of RNA from cells or tissues. This can be used for transcriptome analysis and gene annotation. This next series of lectures will take you through these uses and there is a practical exercise for you to have a go at.

RNA-Seq in the Zebrafish gene build

This talk is by John Collins from the Wellcome Trust Sanger Institute. He explains how RNA-Seq data...
were incorporated into the zebrafish Ensembl gene build. This is an example of creating a gene-build combining RNA-Seq with other data, improving its quality.

He describes how RNA libraries were created from adult and embryonic zebrafish tissues and prepped for Illumina sequencing. He describes RNA-Seq was used to define 3' ends by enriching for polyA tails and trimming, thus improving 3' end prediction.


This research was supported by the Wellcome Trust Sanger Institute grant number 098051

**RNA-Seq Ensembl gene build**

This lecture is by Simon White of Ensembl.

In it, he explains how we can use RNA-Seq data to annotate genomes, including building complete gene sets from scratch, adding novel genes to existing annotations, improving and validating existing annotations and analysing the UTRs of genes. He explains the methods involved in RNA-Seq genome annotation.

He discusses the impact that this has had on the annotation of already annotated genomes, using
the example of zebrafish, and on creating annotation for new genomes, in this case the Tasmanian devil.

Click here to watch the video [51] (17 minutes)

Click here to view the slides [52]

RNA-Seq analysis - Transcriptome assembly and differential expression

In this talk, Myrto Kostadima from the EMBL-EBI discusses how we can use RNA-Seq data to analyse gene expression.

She describes the Cufflinks tool which is a spliced aligner. It can be used to identify different splice variants from RNA-Seq data, and determine their relative frequencies. She discusses how we can use Cufflinks to determine differential gene expression between tissues, individuals or disease states.

Click here to watch the video [53] (12 minutes)
RNA-Seq exercise

This exercise was set by Remco Loos and Myrto Kostadima from the EMBL-EBI. The exercise involves using a sample RNA-Seq dataset from zebrafish to perform a number of basic tasks:

- Aligning RNA-Seq data to the genome.
- Transcriptome reconstruction.
- Comparison of gene expression between two different conditions.

The exercise materials are available for your to work through at your leisure.

Click here to download the worksheet and files.

Before you download this folder, be aware that it is more than 650MB.

You will need to install the following programmes and commands on your computer in order to carry out this practical:

- [Samtools](#)
- [IGV](#)
- [bedGraphToBigWig](#)
- [Bedtools](#)
- [Tophat](#)
- [Cufflinks](#)

In the NGS workshop, the files were pre-installed on the course computers in specific file locations. The commands in the worksheet direct to those file locations: you will need to alter the commands to go to the locations where you have saved your files.
ENCODE consortium data

The ENCODE [58] (Encyclopaedia of DNA Elements) project attempts to identify all of the functional elements in the human genome. This talk by Steven Wilder from ENCODE examines the aims and progress of the ENCODE project. It covers:

- The strategy of the project and the rationale for that.
- The kinds of data included and from what sources.
- How to access the data in various ways.
- Reproducability of the data published by the project.
- The analysis pipeline.
- Biological validation of ENCODE data.

Click here to watch the video [59] (43 minutes)

Click here to view the slides [60]

Click here to visit the Nature ENCODE explorer [61]
Summary

Next-Generation Sequencing has changed the way we carry out molecular biology and genomic studies. It has allowed us to sequence and annotate genomes at a much faster rate. It has allowed us to study variation, expression and DNA binding at a genome-wide level.

As a result of this course, you should have some idea of the potential of NGS in different fields and the tools you can use for these analyses.

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

This course was put together using talks and exercises by:

- Laura Clarke
- John Collins
- Matthias Haimel
- Myrto Kostadima
- Rasko Leinonen
- Remco Loos
- Arnaud Oisel
- Rajesh Radhakrishnan
- Marc Rossello
- Aylwyn Scally
- Steven Wilder
- Simon White

We would like to thank the organisers of the March 2012 course for all their hard work, including Giulietta Spudich, Bronwen Aken, Alison Barker and Guy Cochrane.

We also wish to thank Doros Panayi for filming the lectures and Hugo Neves, from the EMBL Photolab, for the post-processing of the lectures.

The online course was compiled by Emily Pritchard, with input from Giulietta Spudich, Harold Swerdlow, Cath Brookshank, Mindi Sehra, Mark Adams and the speakers.
Additional reading

The following reviews on NGS might be of interest:

- Treffler R., Deckert V., Recent advances in single-molecule sequencing, [67] Current Opinion in Biotechnology, 21 (1) 4-11
- Guigó R. et al., 2006. EGASP: the human ENCODE Genome Annotation Assessment Project [68]. Genome Biology, 7, S2.1-31
- Engström P.G. et al., 2013. Systematic evaluation of spliced alignment programs for RNA-seq data, [69] Nature Methods, 10 (12), 1185-1191
- Steijger T. et al., 2013. Assessment of transcript reconstruction methods for RNA-seq, [70] Nature Methods, 10 (12), 1177-1184

The EMBL-EBI holds regular face-to-face Next Generation Sequencing Practical courses, as well as our more technical Analysis of High-Throughput Sequencing Data courses. Keep an eye on our training website [74] to see when the next course might be.

Develop your knowledge further with our online Analysis of High-Throughput Sequencing Data [75] course.

Source URL: https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course

Links
[1] https://www.ebi.ac.uk/training/online/trainers/emily