



TEPIC



> Next Generation Sequencing

Евгений Бахмет

Научные руководители:
Михаил Райко
Николай Вяххи

Цель — разработка курса Next Generation Sequencing на сайте stepic.org (до alignment) на базе соответствующей ВИКИ-КНИЖКИ



[Create account](#) [Log in](#)

Book [Discussion](#)

[Read](#)

[Latest draft](#)

[Edit](#)

[View history](#)

Next Generation Sequencing (NGS)

The latest reviewed version was checked on 24 June 2013. There are 4 pending changes awaiting review.

The Need for an Up-To-Date Synthesis of Next Generation Sequencing Know-How [\[edit\]](#)

Next generation sequencing (NGS) has become a commodity. With the commercialization of various affordable desktop sequencers, NGS will be of reach by more traditional wet-lab biologists. As seen in recent years, genome-wide scale computational analysis is increasingly being used as a backbone to foster novel discovery in biomedical research. However, as the quantities of sequence data increase exponentially, the analysis bottle-neck is yet to be solved.

The current sources for NGS informatics are extremely fragmented. A novice could read review articles in various journals, follow discussion threads on forums such as [BioStar](#)^[1] or [SEQanswers](#)^[2], or sign up for courses organized by various institutes. Finding a centralized synthesis is much more difficult. Books are available, but the development of the field is so fast that book chapters risk being obsoleted by the time they are even printed. Moreover, cost for a handful of authors to continually update their text would presumably take up a lot of their schedule.

Drawing from the obvious goodwill and community spirit displayed on discussion forums, and exploiting the collaborative tools made available by the Wikimedia foundation, we propose to initiate the editing of a collaborative WikiBook on NGS. Our plan is to collect a sufficient amount of text that people will be incentivized to contribute to it, essentially providing the same information as a forum but in a tidier form. Ultimately, our goal is to create a collective lab book that explains the key concepts and describes best practices in NGS.

TARGET AUDIENCE [\[edit\]](#)

This set of dynamic materials are designed for the bench biologists (advanced PhD students and early career postdoctoral researchers with no or basic bioinformatics experience and demonstrate interest in NGS data analysis). Advanced materials might be added as the community contributes and the needs and trends in the field develop. The flexibility of online material should allow the reader to ignore details in a first read, yet have immediate access to the details they need. However, the overall structure and style should be in priority designed for the non-bioinformatician reader.

Some chapter comes with practical exercise so readers may get themselves familiar with the steps.

- [Main Page](#)
- [Help](#)
- [Browse](#)
- [Cookbook](#)
- [Wikijunior](#)
- [Featured books](#)
- [Recent changes](#)
- [Donations](#)
- [Random book](#)
- ▼ [Community](#)
 - [Reading room](#)
 - [Community portal](#)
 - [Bulletin Board](#)
 - [Help out!](#)
 - [Policies and guidelines](#)
 - [Contact us](#)
- [Tools](#)
- ▼ [Languages](#)

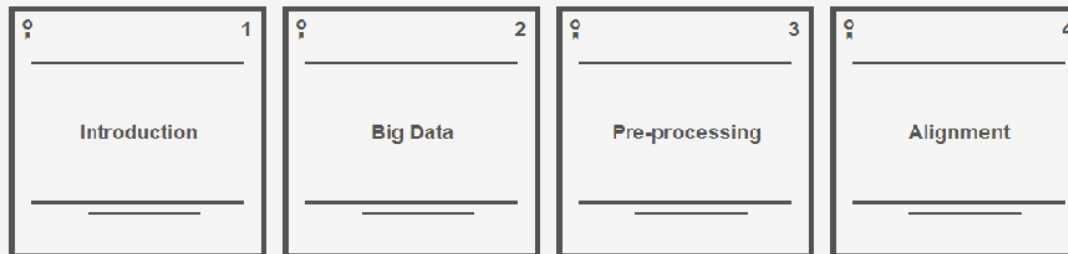
Based on [NGS WikiBook](#). Text is available under the [Creative Commons BY SA License](#).

Next generation sequencing (NGS) has become a commodity. With the commercialization of various affordable desktop sequencers, NGS will be of reach by more traditional wet-lab biologists. As seen in recent years, genome-wide scale computational analysis is increasingly being used as a backbone to foster novel discovery in biomedical research. However, as the quantities of sequence data increase exponentially, the analysis bottle-neck is yet to be solved.

Target Audience

This set of dynamic materials are designed for the bench biologists (advanced PhD students and early career postdoctoral researchers with no or basic bioinformatics experience and demonstrate interest in NGS data analysis). Advanced materials might be added as the community contributes and the needs and trends in the field develop. The flexibility of online material should allow the reader to ignore details in a first read, yet have immediate access to the details they need. However, the overall structure and style should be in priority designed for the non-bioinformatician reader.

Some chapter comes with practical exercise so readers may get themselves familiar with the steps.



[Edit Course](#) [Add Lesson](#) ▼



FASTQ files – discussion of the various quality encodings

FASTQ files extend FASTA files in that they provide both sequence and quality. A FASTQ file thus typically consists of four lines.

1. A line starting with @ containing the sequence identifier
2. the actual sequence
3. a line starting with + after which the sequence identifier is optional
4. a line with quality values which are encoded in ASCII space

As such the 2nd and 4th line must have the same length One such entry is given below showing one sequence "ATGTCT"...

```
@HWI-ST999:102:D1N6AACXX:1:1101:1235:1936 1:N:0:  
ATGTCTCCTGGACCCCTCTGTGCCCAAGCTCCTCATGCATCCT  
+  
19DAADDDF9B9AGF=FGIEHCCD9DG=1E9?HHCf@HHG??B
```


FASTQ files extend FASTA files in that they provide both sequence and quality. A FASTQ file thus typically consists of four lines.

1. A line starting with @ containing the sequence identifier
2. the actual sequence
3. a line starting with + after which the sequence identifier is optional
4. a line with quality values which are encoded in ASCII spaceFASTQ files extend FASTA files in that they provide both sequence and quality. A FASTQ file thus typically consists of four lines.

Next Generation Sequencing

1. Introduction
2. Big Data
- 3. Pre-processing**
4. Alignment



Stepic  Full screen


Searching

Let's try to find some data.

At the <http://galaxyproject.org/> select "Use Galaxy" -> "Get Data" -> "EBI SRA" -> "SRR922404". The last one is an accession number of one reads array.

What animal or plant does this accession belong to?

- Escherichia coli
- Drosophila melanogaster
- Homo sapiens



[Solve Again](#)

Next Generation Sequencing

1. Introduction
2. Big Data
- 3. Pre-processing**
4. Alignment



Stepic



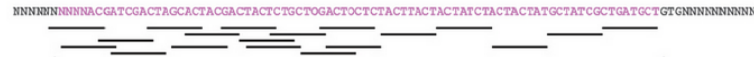
Full screen

Introduction

Alignment, also called mapping, of reads is an essential step in re-sequencing. Having sequenced an organism of a species before, and having constructed a reference sequence, re-sequencing more organisms of the same species allows us to see the genetic differences to the reference sequence, and, by extension, to each other. Alignments of data from these re-sequenced organisms is a relatively simple method of detecting variation in samples. There are certain instances (such as new genes in the sequenced sample that are not found in the existing reference sequence) that can not be detected by alignment alone; however, while other approaches, such as de novo assembly, are potentially more powerful, they are also much harder or, for some organisms, impossible to achieve with current sequencing methods.

Next-generation sequencing generally produces short reads or short read pairs, meaning short sequences of <~200 bases (as compared to long reads by Sanger sequencing, which cover ~1000 bases). To compare the DNA of the sequenced sample to its reference sequence, we need to find the corresponding part of that sequence for each read in our sequencing data. This is called aligning or mapping the reads against the reference sequence. Once this is done, we can look for variation (e.g. SNPs) within the sample.

Sequencing Data Aligned to Reference Genome



© 2010 Illumina, Inc.

Next Generation Sequencing

1. Introduction
2. Big Data
3. Pre-processing
- 4. Alignment**



Stepic
Full screen


Workflow (continuation)

Mapping

Now, we can continue to our data from "Pre-processing", and first fo all, we have to map our reads on the reference genome. For this step, at the left pannel select "NGS: Mapping" -> "Map with Bowtie for Illumina". Then, in the new window, choose the reference "Escherichia coli (str. K-12 substr. MG1655): eschColi_K12", select your last data at the "FASTQ file" field and click "Execute"

Look at the right pannel and check - what format now has your data?

SAM
 BAM
 FASTAQ



Solve Again

Next Generation Sequencing

1. Introduction
2. Big Data
3. Pre-processing

4. Alignment



Результат: курс дописан до раздела «alignment», внедрен workflow для упрощения понимания задач

Спасибо за понимание

