LEIDEN UNIVERSITY MEDICAL CENTER

Bioinformatics practices

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

General overview: from sample to data.

Introduction

General overview: from sample to data.

First stage:

- Research question.
- Make a plan.
- Make agreements.

Introduction

General overview: from sample to data.

First stage:

- Research question.
- Make a plan.
- Make agreements.

Second stage:

- The lab work.

Introduction

General overview: from sample to data.

First stage:

- Research question.
- Make a plan.
- Make agreements.

Second stage:

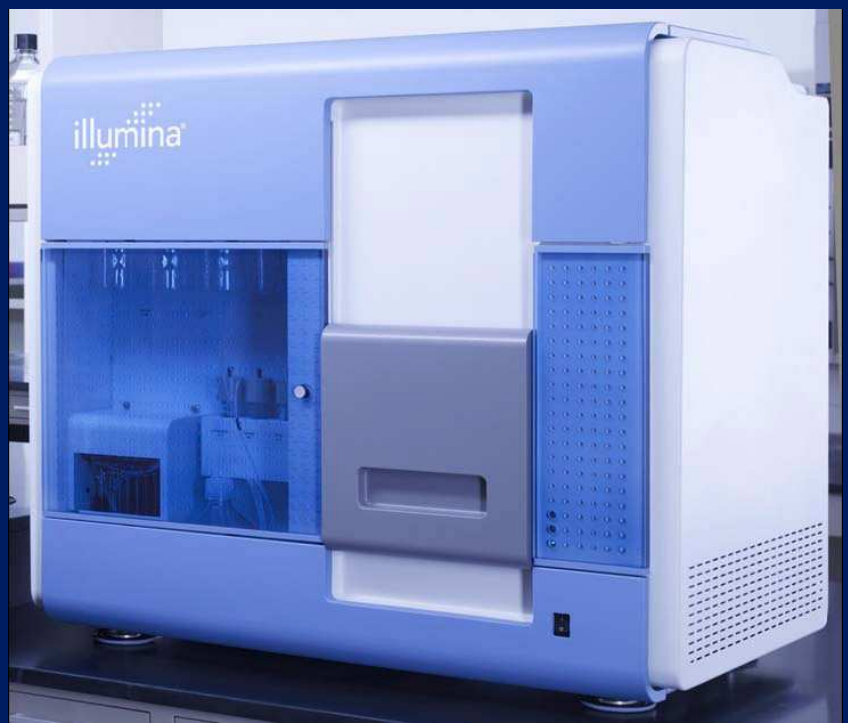
- The lab work.

Third stage (optional):

- Data analysis.

Introduction

The Illumina Genome Analyser II.



Strong points:

- Read length: $\pm 100 \times 2$.
- Runtime: 8 days.
- Output: ± 40 Giga bases.

Introduction

The Illumina HiSeq 2000.

Strong points:

- Read length: $\pm 150 \times 2$.
- Runtime: 8 days.
- Output: ± 150 Giga bases.

Introduction

The Roche / 454.

Strong points:

- Read length: ±500.
- Forensic science (STR markers).

Introduction

The Roche / 454.

Strong points:

- Read length: ±500.
- Forensic science (STR markers).

But:

- Difficulties with mono nucleotide stretches.
- Low yield.

Introduction

The Helicos.

Strong points:

- True single molecule sequencing.
- Forensic science (PCR free, GC-bias free).

Introduction

The Helicos.

Strong points:

- True single molecule sequencing.
- Forensic science (PCR free, GC-bias free).

But:

- Read length: ± 32.
- 5% “dark nucleotide” rate.

Introduction

The Ion Torrent.

Strong points:

- Read length: ±350.
- Runtime: 2 hours.
- Able to pool over 100 amplicons in one run.

Introduction

The Ion Torrent.

Strong points:

- Read length: ±350.
- Runtime: 2 hours.
- Able to pool over 100 amplicons in one run.

But:

- Difficulties with mono nucleotide stretches.
- Extremely low yield.

Introduction

The PacBio.

Strong points:

- Read length: up to 10 kilobases.
- True single molecule sequencing.
- Circular DNA sequencing.
- Strobe sequencing.

Introduction

The PacBio.

Strong points:

- Read length: up to 10 kilobases.
- True single molecule sequencing.
- Circular DNA sequencing.
- Strobe sequencing.

But:

- Extremely high error rate.
- Dark nucleotides.

First stage

The requirements and feasibility:

- Research question.
- Discuss the experiment.
- Discuss the technical possibilities.
- Budget.
- Discuss data analysis.

First stage

The requirements and feasibility:

- Research question.
- Discuss the experiment.
- Discuss the technical possibilities.
- Budget.
- Discuss data analysis.

If data analysis:

- Discuss analysis in general.
- Make sure that the boundaries are clear.
- Preliminary decision about who will do the work.

First stage

The requirements and feasibility:

- Research question.
- Discuss the experiment.
- Discuss the technical possibilities.
- Budget.
- Discuss data analysis.

If data analysis:

- Discuss analysis in general.
- Make sure that the boundaries are clear.
- Preliminary decision about who will do the work.

Make a tender offer.

Second stage

Planning the sequencing run.

- Samples arrive (guaranteed by customer).
- Data into LIMS.
  - Customer.
  - Date.
  - Type of experiment.

Second stage

Planning the sequencing run.

- Samples arrive (guaranteed by customer).
- Data into LIMS.
  - Customer.
  - Date.
  - Type of experiment.
- Sample preparation.
- Sequencing.
- Updating LIMS.
  - Flowcell id / lane number / barcode.
  - Type: paired end / single end.

Third stage

This is optional, mainly depending on the expertise of the client.

- A second discussion.
  - Details.
    - Which tools, how do they work.
    - Explain the report they will receive.
    - Answer any other questions.

Third stage

This is optional, mainly depending on the expertise of the client.

- A second discussion.
  - Details.
    - Which tools, how do they work.
    - Explain the report they will receive.
    - Answer any other questions.
  - What will be delivered.
  - When will the analysis be done.

Third stage

This is optional, mainly depending on the expertise of the client.

- A second discussion.
  - Details.
    - Which tools, how do they work.
    - Explain the report they will receive.
    - Answer any other questions.
  - What will be delivered.
  - When will the analysis be done.

Wait until we receive written confirmation before proceeding.

Third stage

Depending on the experiment:

Third stage

Depending on the experiment:

Targeted resequencing / full genome sequencing:

- GAPSS3 pipeline.

Third stage

Depending on the experiment:

Targeted resequencing / full genome sequencing:

- GAPSS3 pipeline.

SAGE/CAGE RNASeq:

- GAPSS2 pipeline.

Third stage

Depending on the experiment:

Targeted resequencing / full genome sequencing:

- GAPSS3 pipeline.

SAGE/CAGE RNASeq:

- GAPSS2 pipeline.

Specialised analysis:

- Hand work.

Third stage

Depending on the experiment:

Targeted resequencing / full genome sequencing:

- GAPSS3 pipeline.

SAGE/CAGE RNASeq:

- GAPSS2 pipeline.

Specialised analysis:

- Hand work.
    - But not less structured.

GAPSS3

Emphasis on quality control and information disclosure.

GAPSS3

Emphasis on quality control and information disclosure.

- FASTQC plots before and after data cleaning.
  - Sequence content.
  - GC content.
  - Length distribution.
  - ...

GAPSS3

Emphasis on quality control and information disclosure.

- FASTQC plots before and after data cleaning.
  - Sequence content.
  - GC content.
  - Length distribution.
  - ...
- QC after alignment.
  - Transition transversion rates.

GAPSS3

Emphasis on quality control and information disclosure.

- FASTQC plots before and after data cleaning.
  - Sequence content.
  - GC content.
  - Length distribution.
  - ...
- QC after alignment.
  - Transition transversion rates.
- QC after annotation.
  - dbSNP rate.

GAPSS3

Emphasis on quality control and information disclosure.

- FASTQC plots before and after data cleaning.
  - Sequence content.
  - GC content.
  - Length distribution.
  - ...
- QC after alignment.
  - Transition transversion rates.
- QC after annotation.
  - dbSNP rate.

In principle, we do not do any biological interpretation, but if we get a suggestion that helps clients guide them in their interpretation, we incorporate it.

GAPSS3

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).

GAPSS3

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.

GAPSS3

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.
- Database to keep track of the versions of all used tools and custom scripts.
  - If one or more tools are upgraded, the new versions are stored.
  - The version number of the pipeline is incremented.
  - The versions of all tools of all pipeline versions can be retrieved from this database.

GAPSS3

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.
- Database to keep track of the versions of all used tools and custom scripts.
  - If one or more tools are upgraded, the new versions are stored.
  - The version number of the pipeline is incremented.
  - The versions of all tools of all pipeline versions can be retrieved from this database.
- LaTeX documentation is automatically generated.
  - Compiled to pdf that can be handed over to the customer.

GAPSS3

Some implementation details.

- Framework in *bash*.
  - Stand alone scripts written in other languages (Perl, Python, . . .).
- *Sun grid engine* to submit jobs to our local cluster.
- Database to keep track of the versions of all used tools and custom scripts.
  - If one or more tools are upgraded, the new versions are stored.
  - The version number of the pipeline is incremented.
  - The versions of all tools of all pipeline versions can be retrieved from this database.
- LaTeX documentation is automatically generated.
  - Compiled to pdf that can be handed over to the customer.
- All individual commands are logged.

Custom analysis

When doing custom analysis, we are usually experimenting ourselves.

Custom analysis

When doing custom analysis, we are usually experimenting ourselves.

Make a log file of the analysis:

- Write down all tools used.
  - Their version number.
  - Where you got them.
  - What you installed to compile it.
  - ...

Custom analysis

When doing custom analysis, we are usually experimenting ourselves.

Make a log file of the analysis:

- Write down all tools used.
  - Their version number.
  - Where you got them.
  - What you installed to compile it.
  - ...
- If the analysis is only for one sample, log all commands.

Custom analysis

When doing custom analysis, we are usually experimenting ourselves.

Make a log file of the analysis:

- Write down all tools used.
  - Their version number.
  - Where you got them.
  - What you installed to compile it.
  - ...
- If the analysis is only for one sample, log all commands.
- Otherwise write a small script.
- Supply the log and the script with the output.

Custom analysis

When doing custom analysis, we are usually experimenting ourselves.

Make a log file of the analysis:

- Write down all tools used.
  - Their version number.
  - Where you got them.
  - What you installed to compile it.
  - ...
- If the analysis is only for one sample, log all commands.
- Otherwise write a small script.
- Supply the log and the script with the output.

Tip:

- Some terminals (e.g., xterm) have a log functionality.

Questions?

Acknowledgements:

Sophie Greve-Onderwater
Henk Buermans
Michiel van Galen
Yu-Ching Lai
Martijn Vermaat
Bradley ten Broeke
Michel Villerius
Johan den Dunnen