LEIDEN UNIVERSITY MEDICAL CENTER

ICT and research at the LUMC

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

ICT and research

Division 5.

Introduction

ICT and research

Division 5.

- Center for Human and Clinical Genetics.

Introduction

ICT and research

Division 5.

- Center for Human and Clinical Genetics.
  - Human Genetics.

Introduction

ICT and research

Division 5.

- Center for Human and Clinical Genetics.
  - Human Genetics.
    - Leiden Genome Technology Center.

Introduction

ICT and research

Division 5.

- Center for Human and Clinical Genetics.
  - Human Genetics.
    - Leiden Genome Technology Center.

High throughput sequencing.

Illumina sequencing

DNA sequencing

Figure 1: HiSeq 2000.

Illumina sequencing

DNA sequencing

Figure 1: HiSeq 2000.

Figure 2: Flowcell.

Illumina sequencing

DNA sequencing

Figure 1: HiSeq 2000.

Figure 2: Flowcell.

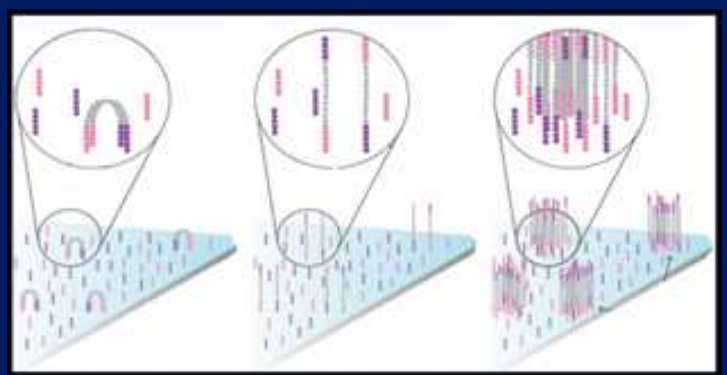

Figure 3: Amplification.

Illumina sequencing

Raw data analysis

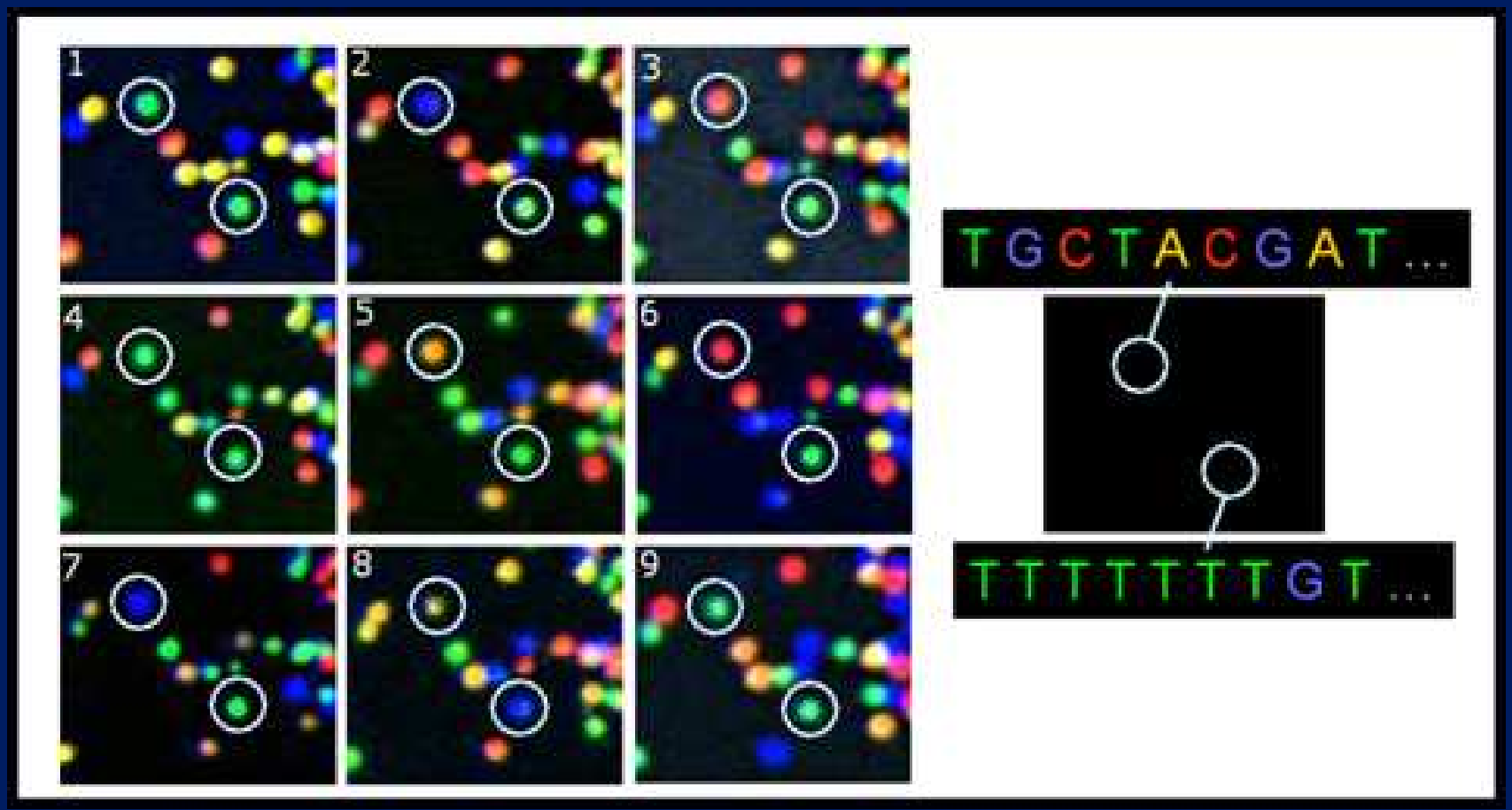

Figure 4: Base calling.

Illumina sequencing

Result of base calling

```
1  @FCC0ADRACXX:2:1101:1684:1875#GCGGAACT/1
2  AGGGGATGCAACCTCGAGGAGGAAAGGAACGAAAGAGGAAGGGAG...
3  +
4  BS\ceeeeggggggiihfhighihiiiihiihhhihfhighiiiii...
5  @FCC0ADRACXX:2:1101:1714:1885#GCGGAACT/1
6  ATTTTCTCAATGTCTACCACTGCGGGTAACACTTTGTGTTTCCCA...
7  +
8  bbbeeeeeggggggiihiiiiiiiiiighiiiiiiiiggfghiihi...
```

Listing 1: FastQ file.

These files are around 1T big.

Illumina sequencing

Some figures

Runtime 2 weeks (5-10 full genomes).

Produces 6T of pictures per flowcell, 500G base calling.

Needs continuous connection to the storage, otherwise it stalls.

Illumina sequencing

Some figures

Runtime 2 weeks (5-10 full genomes).

Produces 6T of pictures per flowcell, 500G base calling.

Needs continuous connection to the storage, otherwise it stalls.

Basecalling:

24	time in hours for base calling per flowcell
$12\times$	number of cores
$2\times$	number of flowcells
576	CPU hours per flowcell

On a desktop machine this would take 48 days.

PacBio sequencing

Pacific Biosciences Single Molecule, Real-Time

Figure 5: PacBio RS.

PacBio sequencing

Base calling

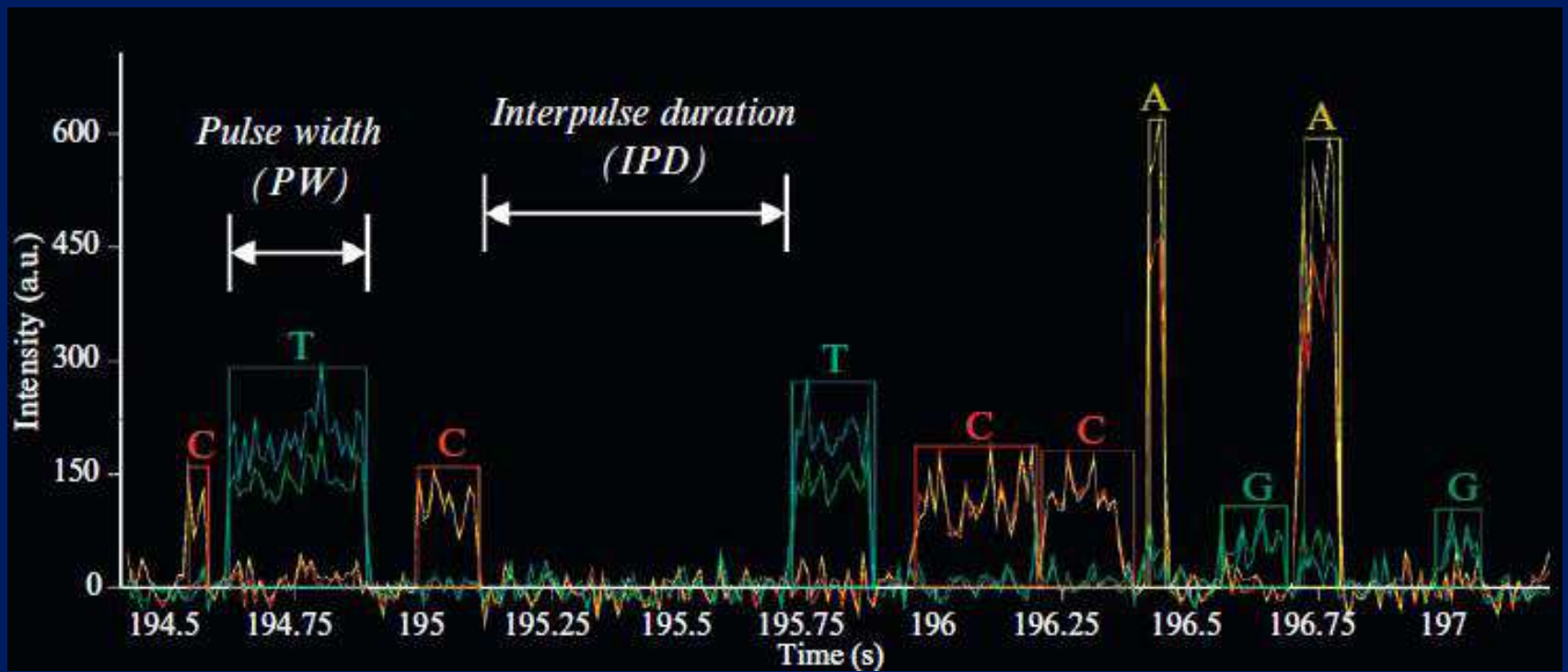

Figure 6: PacBio trace.

PacBio sequencing

Some figures

Runtime 90 minutes.

PacBio sequencing

Some figures

Runtime 90 minutes.

Per cell:

- 3G/sec.
  - 4 cameras.
  - 72 fps.
  - 10M pictures.
- 16T of raw data.
- 4G after base calling.

PacBio sequencing

Some figures

Runtime 90 minutes.

Per cell:

- 3G/sec.
  - 4 cameras.
  - 72 fps.
  - 10M pictures.
- 16T of raw data.
- 4G after base calling.

There are 16 cells in this machine:

- 256T per run.
- 64G after base calling.

Secondary data analysis

Alignment

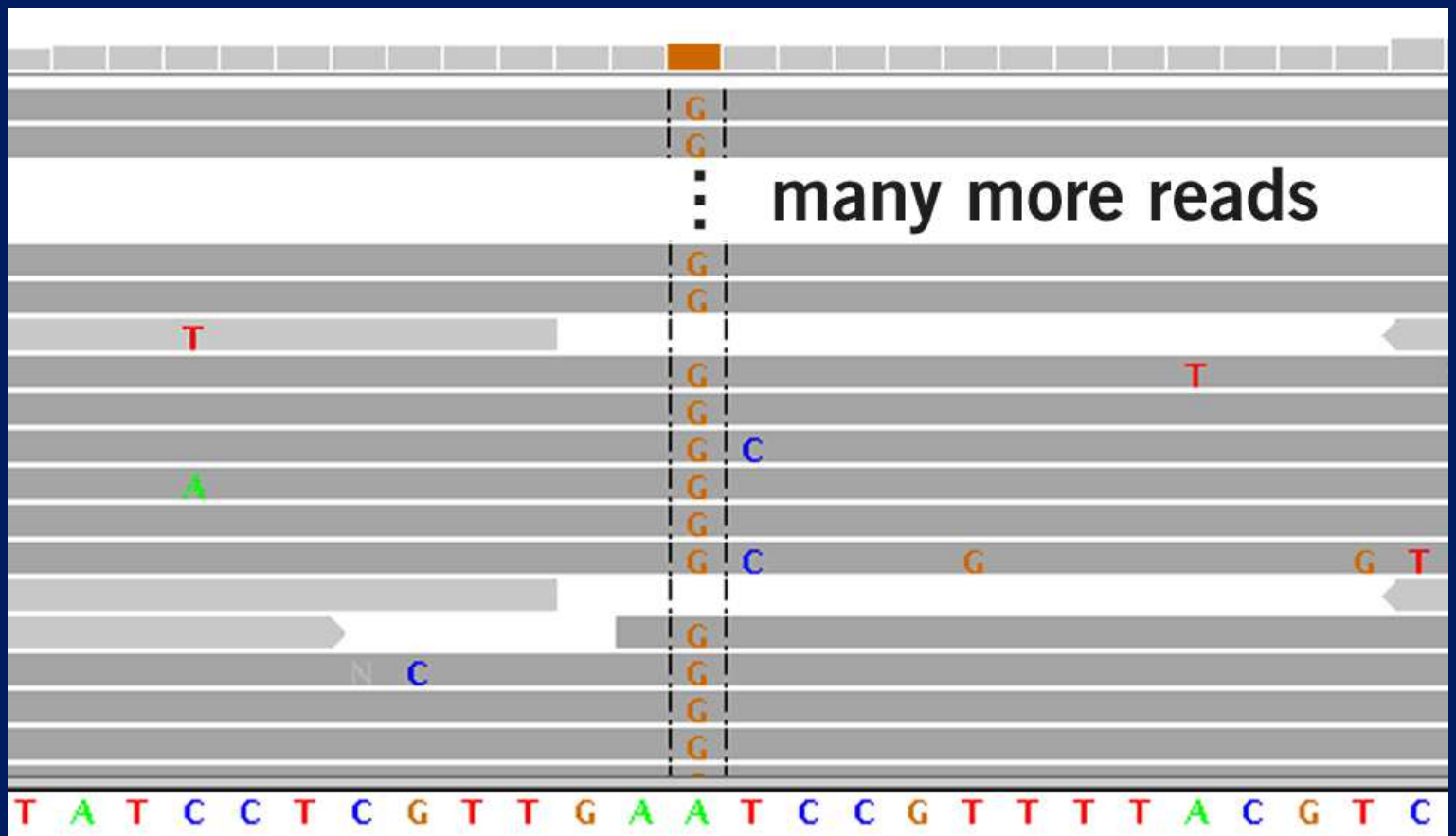

Figure 7: Result of an alignment.

Facilities

Clusters

Figure 8: Dell M610 blade server

Facilities

Local cluster

Some figures:

- 29 nodes.
- 368 cores.
- 94 users.

Facilities

Local cluster

Some figures:

- 29 nodes.
- 368 cores.
- 94 users.

Funded by four departements:

- Molecular Epidemiology.
- Clinical Genetics.
- Human Genetics.
- Parasitology.

Facilities

Storage

Funded by the same four departements.

Share	Size (TB)	Used
MolEpi	65	51
KG	27	13
HumGen	37	18
BMS	15	0
LGTC	105	89
SASC	5	0
GoNL	140	105
UCSC-bam	1	1
total	520	372

Table 1: Usage of the storage.

Common analyses

Resequencing

Exome:

- Only look at the genes.
- Will not detect everything.

Common analyses

Resequencing

Exome:

- Only look at the genes.
- Will not detect everything.

Full genome:

- Analyse everything.

Common analyses

Resequencing

Exome:

- Only look at the genes.
- Will not detect everything.

Full genome:

- Analyse everything.

type	desktop	cluster
exome	4 days	5 hours
genome	one year	3 days

Table 2: Gain of using a cluster.

Common analyses

Large projects

Genome of the Netherlands:

- 750 full genomes.

Common analyses

Large projects

Genome of the Netherlands:

- 750 full genomes.

Centre for Genome Diagnostics:

- 10 full genomes.

Common analyses

Large projects

Genome of the Netherlands:

- 750 full genomes.

Centre for Genome Diagnostics:

- 10 full genomes.

Geuvadis:

- QC for 667 samples done in two days.

These types of analysis would be impossible without the cluster.

Common analyses

Future

Sequencing in diagnostics (figures from Rotterdam):

- 100 exomes per week.
- 100×2G.

Common analyses

Future

Sequencing in diagnostics (figures from Rotterdam):

- 100 exomes per week.
- 100×2G.

We will switch to full genome sequencing soon:

- $\pm$40G per genome.
- 4T of data per week.

Common analyses

Future

Sequencing in diagnostics (figures from Rotterdam):

- 100 exomes per week.
- 100×2G.

We will switch to full genome sequencing soon:

- $\pm$40G per genome.
- 4T of data per week.

Needs to be more reliable.

Questions?

Acknowledgements:

Michiel van Galen
Martijn Vermaat
Michel Villerius
Johan den Dunnen