LEIDEN UNIVERSITY MEDICAL CENTER

Bioinformatics at the LGTC

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Illumina platforms

Figure 1: HiSeq 2500.

Characteristics:

- High throughput (3 genomes).
- Paired end.
- High accuracy.
- Read length 2×125bp.
- Relatively long run time (6 days).
- Relatively expensive.

Introduction

Next generation sequencing data

```
1  @SGGPP:4:101
2  TTCGGGGGCTGGCAAATCCACTTCCGTGACACGCTACCATTCGCTGGTGGT
3  +
4  −’+4589,53330−0&07+03:54/2362−+.488587>@/25440++0(+
5  @SGGPP:4:102
6  CGGTAAACCACCCTGCTGACGGAACCCTAATGCGCCTGAAAGACAGCGTTC
7  +
8  34/−−0’+.000(.55:;:99(0(+2(22(0316;185;;0;:<<>=AA59
9  @SGGPP:4:106
10 TCGTTAACGACTTTGTTCGCCACCGCAACCGCCTGTTTCGGGTCACAGGCA
11 +
12 09875;5?<;?@A4?B:BBB<AA>CCC>C>BB0.−>=0488+3444:@5@<
13 @SGGPP:4:112
14 TTGATGAATATATTATTTCAGGGAATAATTATGACACCTTTAGAACGCATT
15 +
16 70<<@::5:<;==7;>>/79<:.:494.8(,,8:753/5@5??C>B???B7
```

Listing 1: A FastQ file.

Introduction

The best match to the reference genome

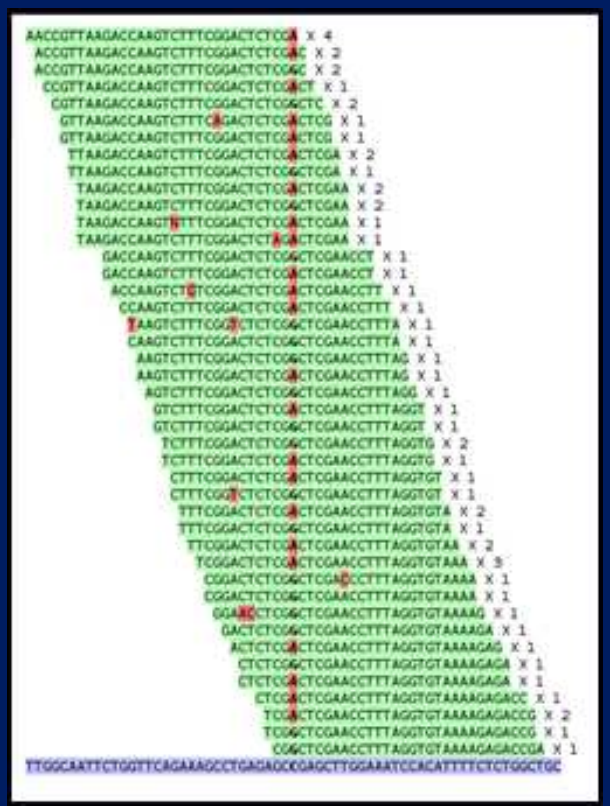

Figure 2: Visualisation of an alignment.

Very efficient.

- The reference genome needs to be *indexed*.
- Finding an alignment is as easy as looking up a word in a dictionary.

Introduction

Consistent deviations from the reference

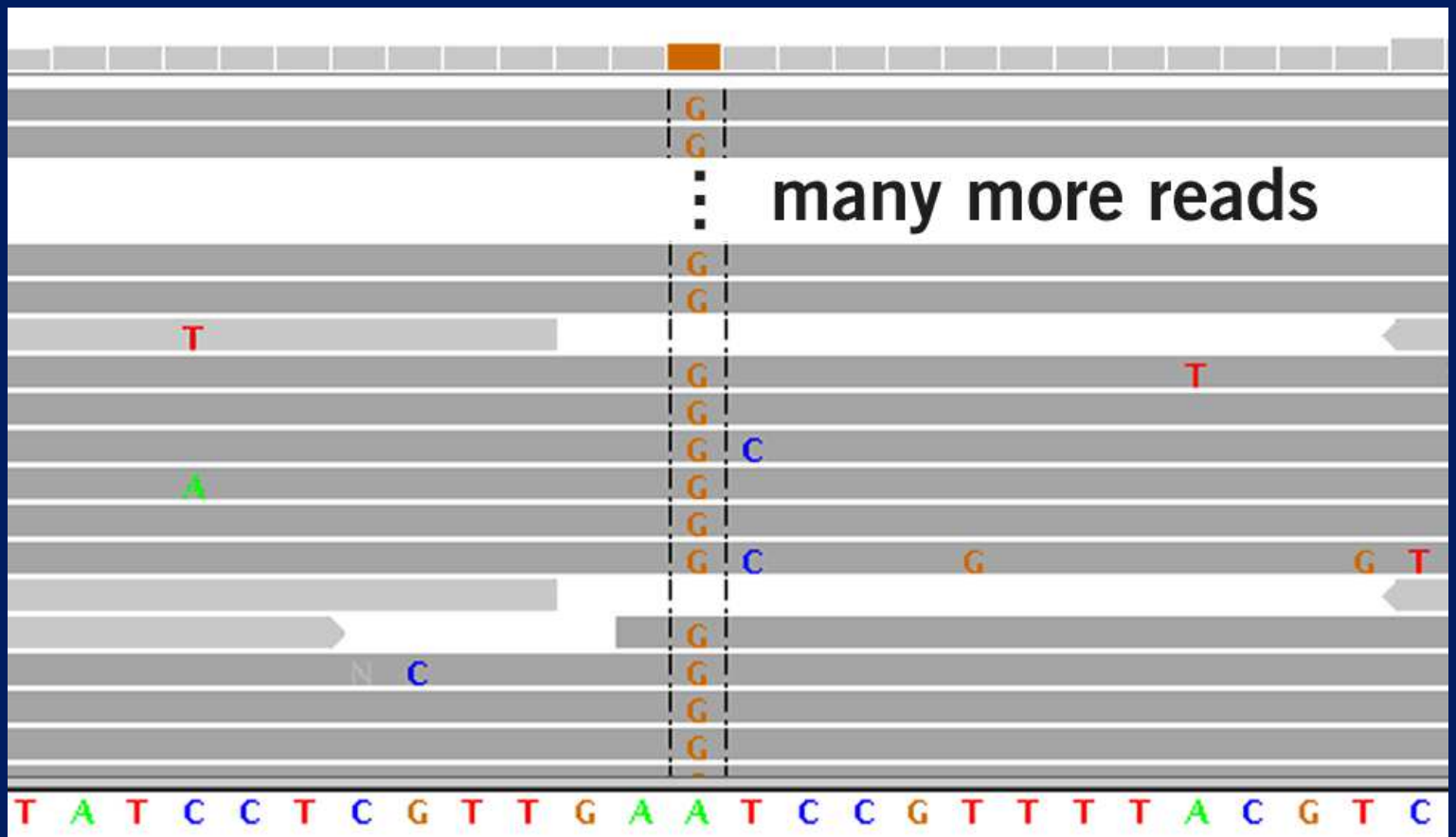

Figure 3: Result of an alignment.

Introduction

Resequencing

Exome:

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

Introduction

Resequencing

Exome:

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

Full genome:

- Analyse everything.

Introduction

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 1: Gain of using a cluster.

Equipment

Clusters

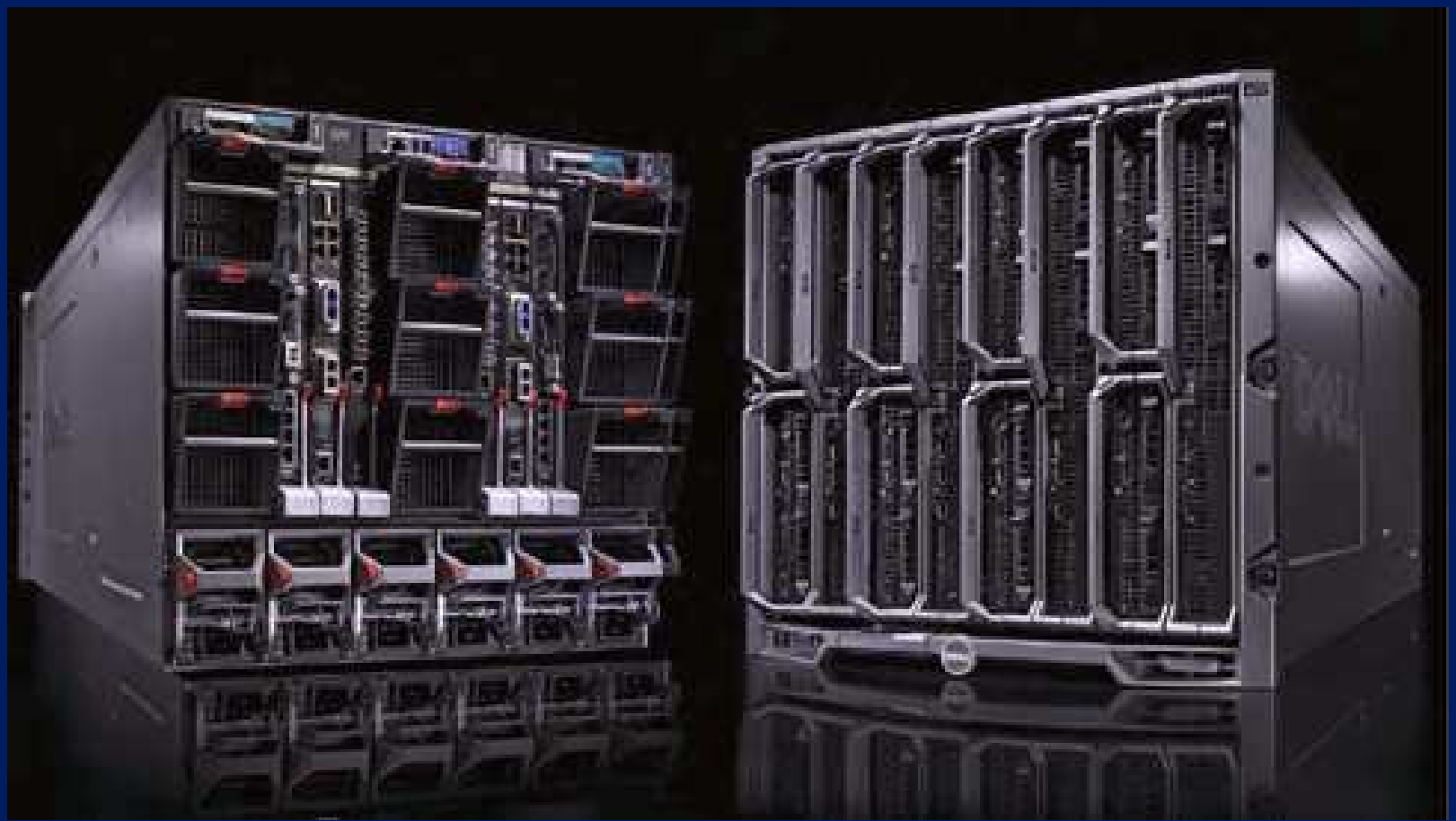

Figure 4: Dell M610 blade server

Equipment

Local cluster

Some figures:

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

Equipment

Local cluster

Some figures:

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

Funded by four departements:

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

Equipment

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 2: Usage of the storage.

Data analysis

Pipelines

Figure 5: Scene from “Modern times”.

Example: E.coli data analysis

The E.coli

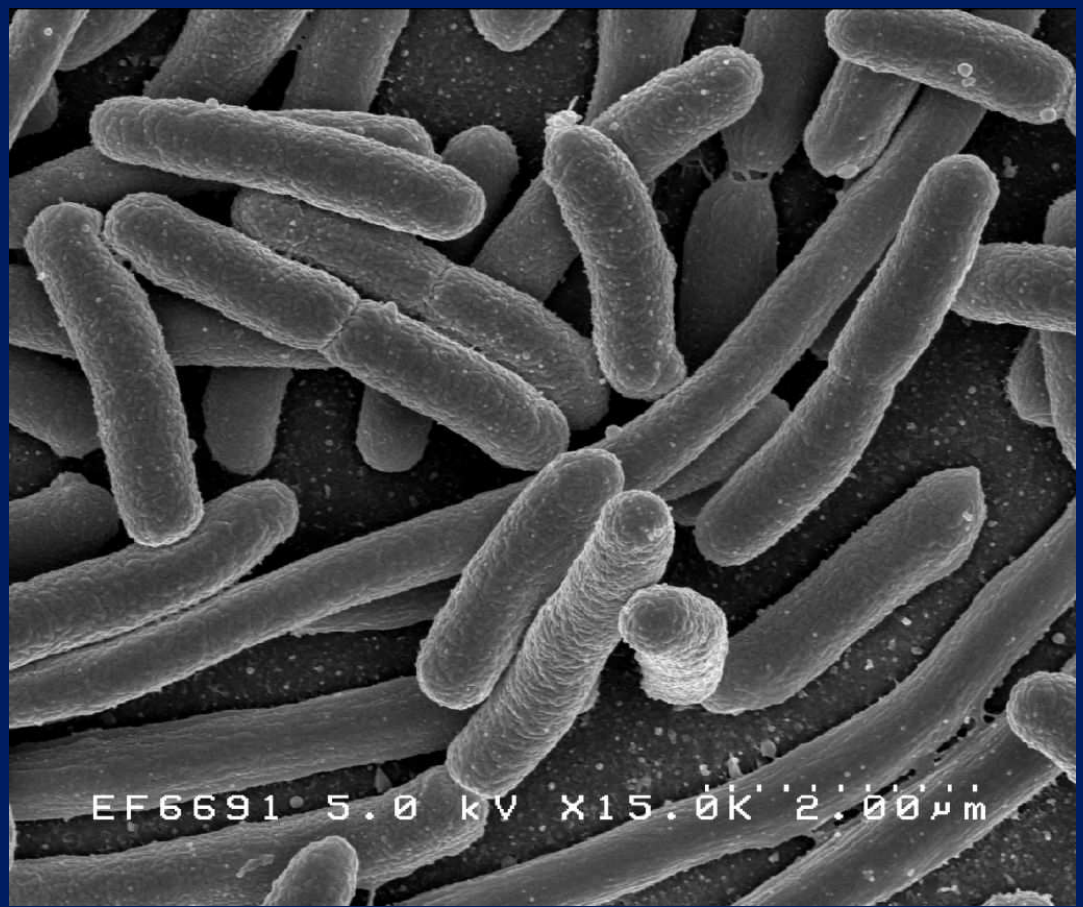

Figure 6: Escherichia coli.

Example: E.coli data analysis

Plasmids

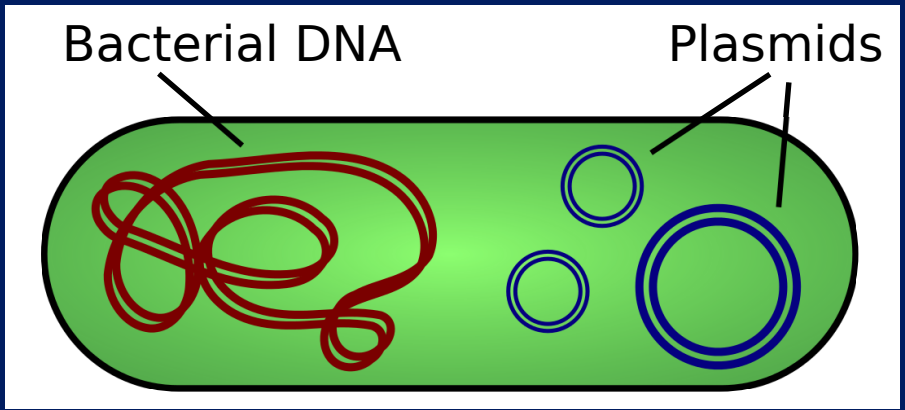

Figure 7: Schematic overview of a cell containing plasmids.

Example: E.coli data analysis

Plasmids

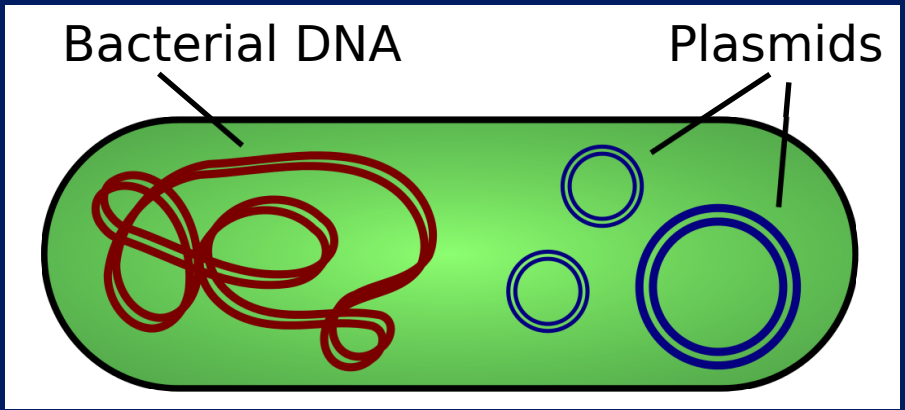

Figure 7: Schematic overview of a cell containing plasmids.

Plasmids are small DNA molecules.

- Separate and independent from the chromosome.
- Can be transferred to other species.
- Size between 1×10^3 and 1×10^6 basepairs.
- Copy number between 1 and 1,000.
- Variable between strains and individuals.

Example: E.coli data analysis

Antibiotic resistance testing

Figure 8: Classical antibiotic resistance test.

Example: E.coli data analysis

Goals

Clinical:

- Strain identification (MLST).
- Antibiotic resistance testing.
- Identifying efflux pumps.
- Find other important genes.

Example: E.coli data analysis

Goals

Clinical:

- Strain identification (MLST).
- Antibiotic resistance testing.
- Identifying efflux pumps.
- Find other important genes.

Technical limitations:

- The result must be delivered fast.

Example: E.coli data analysis

Goals

Clinical:

- Strain identification (MLST).
- Antibiotic resistance testing.
- Identifying efflux pumps.
- Find other important genes.

Technical limitations:

- The result must be delivered fast.

Next Generation Sequencing.

Example: E.coli data analysis

Sequencers: Ion Torrent

Figure 9: Ion torrent.

Characteristics:

- 3 hours per run.
- 1 day sampleprep, 1 day emulsion PCR.
- 4×10^6 reads.
- Read length $\pm$300bp.
- 2 *E. coli* per run.

Example: E.coli data analysis

Sequencers: Ion Torrent

Figure 9: Ion torrent.

Characteristics:

- 3 hours per run.
- 1 day sampleprep, 1 day emulsion PCR.
- 4×10^6 reads.
- Read length $\pm$300bp.
- 2 *E. coli* per run.

Fast and inexpensive.

Example: E.coli data analysis

General overview of the pipeline

We screen for 130 known plasmids and 400 genes.

Example: E.coli data analysis

General overview of the pipeline

We screen for 130 known plasmids and 400 genes.

Output:

- MLST.
- List of plasmids.
    - Otherwise, similar plasmids.
- List of genes of interest.

Example: E.coli data analysis

General overview of the pipeline

We screen for 130 known plasmids and 400 genes.

Output:

- MLST.
- List of plasmids.
    - Otherwise, similar plasmids.
- List of genes of interest.

For the MLST, we need a *consensus sequence*.

- As opposed to a list of variants, which we normally use.

Example: E.coli data analysis

General overview of the pipeline

We screen for 130 known plasmids and 400 genes.

Output:

- MLST.
- List of plasmids.
    - Otherwise, similar plasmids.
- List of genes of interest.

For the MLST, we need a *consensus sequence*.

- As opposed to a list of variants, which we normally use.

For the list of plasmids and genes, we want a list we can open in Excel.

Example: E.coli data analysis

Results: MLST

```
1  CAATGATGATCGACAGTATGGCTGTGCTCGATATCTTCATTCTTGCGGCT
2  AAAGCGGCGGCGAACCACCACAAAGAATACCGGAACGAAGAAGATTGCCA
3  GTACCGTTGCGGTCACCATCCCGCCCATTACACCGGTACCTACTGCGTTC
4  TGCGCGCCGGAACCAGCACCAGTACTGATAACCAGCGGCATAACGCCGAG
5  GATAAACGCCAGCGAGGTCATCAGGATCGGACGTAAACGCATCCGCACCG
6  CATCAAGCGTCGCTTCAATCAGACCTTTACCTTCTTTATCCATCAAGTCT
7  TTGGCGAATTCGACGATAAGGATCGCGTTCTTCGCCGACAACCCAATGGT
8  TGTGAGCAGGCCTACCTGGAAGTAAACGTCATTGGTCAGGCCACGGAAGG
```

Listing 2: Part of the consensus sequence of acrB.

Example: E.coli data analysis

Results: MLST

```
1  CAATGATGATCGACAGTATGGCTGTGCTCGATATCTTCATTCTTGCGGCT
2  AAAGCGGCGGCGAACCACCACAAAGAATACCGGAACGAAGAAGATTGCCA
3  GTACCGTTGCGGTCACCATCCCGCCCATTACACCGGTACCTACTGCGTTC
4  TGCGCGCCGGAACCAGCACCAGTACTGATAACCAGCGGCATAACGCCGAG
5  GATAAACGCCAGCGAGGTCATCAGGATCGGACGTAAACGCATCCGCACCG
6  CATCAAGCGTCGCTTCAATCAGACCTTTACCTTCTTTATCCATCAAGTCT
7  TTGGCGAATTCGACGATAAGGATCGCGTTCTTCGCCGACAACCCAATGGT
8  TGTGAGCAGGCCTACCTGGAAGTAAACGTCATTGGTCAGGCCACGGAAGG
```

Listing 2: Part of the consensus sequence of acrB.

These sequences can be analysed directly by existing MLST classification software.

Example: E.coli data analysis

Results: Plasmid detection

Plasmid	Size	Reads	#3/#2	Cov	#5/#2
NC_001537	3895	18728	4.808	1418	0.364
NC_002119	9957	6130	0.615	789	0.079
NC_002127	3306	11749	3.553	1068	0.323
NC_002128	92721	11824	0.127	35783	0.385
NC_002142	68817	8163	0.118	15938	0.231
NC_002145	1549	46141	29.787	1549	1.000
NC_002487	5847	11669	1.995	1735	0.296
NC_002525	75582	420	0.005	1325	0.017
NC_004429	6349	961	0.151	1858	0.292

Table 3: Part of the plasmids Excel file.

Questions?

Sophie Greve-Onderwater
Henk Buermans
Johan den Dunnen