



LEIDEN UNIVERSITY MEDICAL CENTER

***E. coli* plasmid and gene profiling using Next Generation Sequencing**

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics



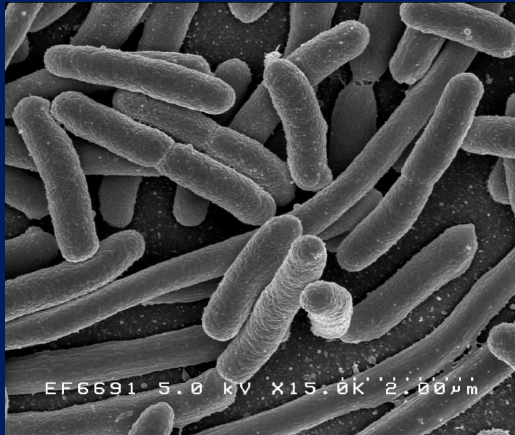
General overview

Figure 1: Escherichia coli.

Genome published in 1997.

- Genome size 4.6×10^6 basepairs.
- 4,288 genes in the assembly.
 - 86 tRNA genes.
- 2,584 operons in the assembly.
 - 7 rRNA operons.

Genome published in 1997.

- Genome size 4.6×10^6 basepairs.
- 4,288 genes in the assembly.
 - 86 tRNA genes.
- 2,584 operons in the assembly.
 - 7 rRNA operons.

However, per individual strain:

- Between 4,000 and 5,500 genes.
- 16,000 genes in total (pangenome).

Genome published in 1997.

- Genome size 4.6×10^6 basepairs.
- 4,288 genes in the assembly.
 - 86 tRNA genes.
- 2,584 operons in the assembly.
 - 7 rRNA operons.

However, per individual strain:

- Between 4,000 and 5,500 genes.
- 16,000 genes in total (pangenome).

Very diverse, only 20% of the genome is shared between all strains.

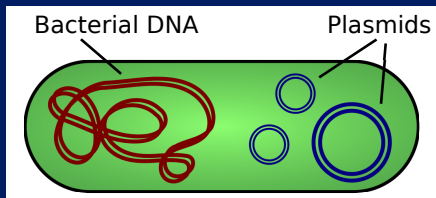
Plasmids

Figure 2: Schematic overview of a cell containing plasmids.

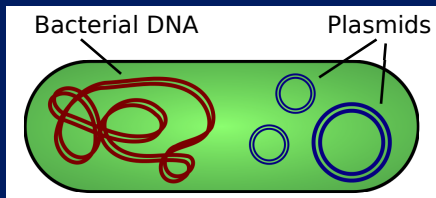
Plasmids

Figure 2: 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.

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Genes:

- Multi Locus Sequence Typing (MLST).
 - Uses household genes.
 - Fragments of 450 to 500 basepairs.

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Genes:

- Multi Locus Sequence Typing (MLST).
 - Uses household genes.
 - Fragments of 450 to 500 basepairs.
- Antibiotic resistance.
 - The gene may be known, the plasmid may not be.

Profiling

Plasmids:

- May carry antibiotic resistance genes.
- Not all of them are known.
- May be highly similar to other plasmids.

Genes:

- Multi Locus Sequence Typing (MLST).
 - Uses household genes.
 - Fragments of 450 to 500 basepairs.
- Antibiotic resistance.
 - The gene may be known, the plasmid may not be.
- Efflux pumps.
- ...

Antibiotic resistance testing

Figure 3: Classical antibiotic resistance test.

Goals

Clinical:

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

Goals

Clinical:

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

Technical limitations:

- The result must be delivered fast.

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.

Why Next Generation Sequencing?

We analyse *everything* in one go.

- The genome, all plasmids are sequenced.
- Known but also *unknown* DNA is sequenced.
- Data can be re-analysed.
 - Is gene X also in there?

Why Next Generation Sequencing?

We analyse *everything* in one go.

- The genome, all plasmids are sequenced.
- Known but also *unknown* DNA is sequenced.
- Data can be re-analysed.
 - Is gene X also in there?

We did a pilot on the HiSeq 2000.

- Successful.
- A bit slow (it takes two weeks for a HiSeq to finish).
- Way too much data per sample.
 - Over 200 times more data per sample than needed.
- Found a contamination (Streptococcus).

Sequencers: Ion Torrent



Figure 4: Ion torrent.

Characteristics:

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

Sequencers: Ion Torrent



Figure 4: Ion torrent.

Fast and inexpensive.

Characteristics:

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

General overview

We screen for 130 known plasmids and 400 genes.

General overview

We screen for 130 known plasmids and 400 genes.

Output:

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

General overview

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.

General overview

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.

Alignment

```

ACCCTTANGACCAAGTCTTTCGGACTCTCGA X 4
ACCCTTANGACCAAGTCTTTCGGACTCTCGAC X 2
ACCCTTANGACCAAGTCTTTCGGACTCTCGAC X 2
CCCTTANGACCAAGTCTTTCGGACTCTCGACT X 1
CCTTANGACCAAGTCTTTCGGACTCTCGACTC X 2
GTTAANGACCAAGTCTTTCGACTCTCGACTCG X 1
GTTAANGACCAAGTCTTTCGGACTCTCGACTCG X 1
TTAANGACCAAGTCTTTCGGACTCTCGACTCGA X 2
TTAANGACCAAGTCTTTCGGACTCTCGACTCGA X 1
TANGACCAAGTCTTTCGGACTCTCGACTCGAA X 2
TANGACCAAGTCTTTCGGACTCTCGACTCGAA X 2
TANGACCAAGTCTTTCGGACTCTCGACTCGAA X 1
TANGACCAAGTCTTTCGGACTCTCGACTCGAA X 1
GACCAAGTCTTTCGGACTCTCGACTCGAACCT X 1
GACCAAGTCTTTCGGACTCTCGACTCGAACCT X 1
ACCAAGTCTTTCGGACTCTCGACTCGAACCT X 1
CCAAGTCTTTCGGACTCTCGACTCGAACCT X 1
TAAGTCTTTCGGACTCTCTCGACTCGAACCTTA X 1
CAAGTCTTTCGGACTCTCGACTCGAACCTTA X 1
AAGTCTTTCGGACTCTCGACTCGAACCTTAA X 1
AAGTCTTTCGGACTCTCGACTCGAACCTTAA X 1
AGTCTTTCGGACTCTCGACTCGAACCTTAA X 1
GTCTTTCGGACTCTCGACTCGAACCTTAA X 1
GTCTTTCGGACTCTCGACTCGAACCTTAA X 1
TCTTTCGGACTCTCGACTCGAACCTTAA X 2
TCTTTCGGACTCTCGACTCGAACCTTAA X 1
CTTTCGGACTCTCGACTCGAACCTTAA X 1
CTTTCGGACTCTCGACTCGAACCTTAA X 1
TTTTCGGACTCTCGACTCGAACCTTAA X 2
TTTTCGGACTCTCGACTCGAACCTTAA X 1
TTTTCGGACTCTCGACTCGAACCTTAA X 2
TCTTCGGACTCTCGACTCGAACCTTAA X 3
CGACTCTCGACTCGACTCGAACCTTAA X 1
CGACTCTCGACTCGAACCTTAA X 1
GACTCTCGACTCGAACCTTAA X 1
GACTCTCGACTCGAACCTTAA X 1
CTCTCGACTCGAACCTTAA X 1
CTCTCGACTCGAACCTTAA X 1
CTCGACTCGAACCTTAA X 1
CTCGACTCGAACCTTAA X 1
TCTCGACTCGAACCTTAA X 2
TCTCGACTCGAACCTTAA X 1
CGACTCTCGAACCTTAA X 1
CGACTCTCGAACCTTAA X 1
TTGGCAATTCTGTTTCGAGAAAGCTGAGACCGACTTGGAAATCCACTTTCTCTGGCTGC
  
```

Figure 5: Variant calling.

MLST

Pipeline:

- Map all reads to the genome.
- Make a consensus sequence.
- Select genes.

MLST

Pipeline:

- Map all reads to the genome.
- Make a consensus sequence.
- Select genes.

Tools:

- `tmap` for alignment.
- `samtools/bcftools` for building a consensus sequence.
- In house program to select a region.

Plasmid detection

Pipeline:

- Select all reads that do not map to the genome.
- Map these reads to each plasmid individually.
- Calculate the *horizontal coverage*.

Plasmid detection

Pipeline:

- Select all reads that do not map to the genome.
- Map these reads to each plasmid individually.
- Calculate the *horizontal coverage*.

Tools:

- **samtools** to extract unmapped reads.
- **tmap** for alignment.
- In house program to make a **wiggle** track.
- In house program to find *covered regions*.

Coverage

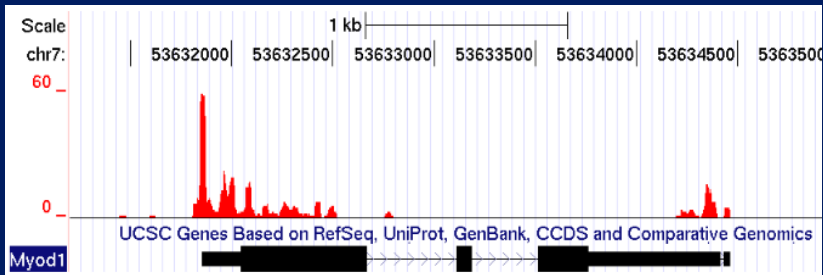


Figure 6: Coverage / depth histogram.

Coverage

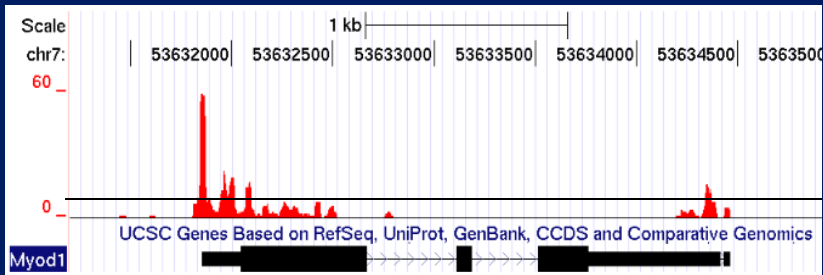


Figure 6: Coverage / depth histogram.

Coverage

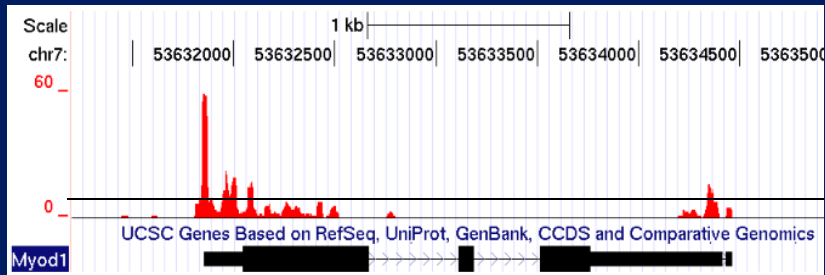


Figure 6: Coverage / depth histogram.



Figure 7: Coverage summary.

Antibiotic resistance genes detection

Pipeline:

- Select genes from the genome or plasmids.
- Calculate the non-N content of the consensus sequence.

Antibiotic resistance genes detection

Pipeline:

- Select genes from the genome or plasmids.
- Calculate the non-N content of the consensus sequence.

Tools:

- In house program to select a region.
- In house program to calculate the non-N percentage.

Technical issues

Between 66% and 80% of the reads map to the genome.

Technical issues

Between 66% and 80% of the reads map to the genome.

The other needs to be mapped to the 130 plasmids and 278 additional genes.

- Alignment is not much faster for small reference sequences.

Technical issues

Between 66% and 80% of the reads map to the genome.

The other needs to be mapped to the 130 plasmids and 278 additional genes.

- Alignment is not much faster for small reference sequences.

In total, the analysis would take around $\frac{130+278}{3} = 136$ times longer than the initial alignment.

Clusters

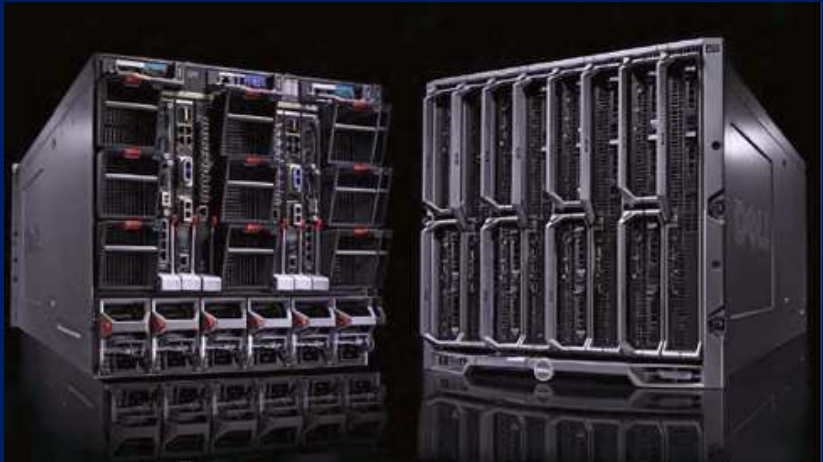


Figure 8: Dell M610 blade server

Automatic scheduling on a cluster

```
1  %.bam: %.sam
2      $(SAMTOOLS) view -bt $(call MAKEREF, $@) -o $@ $<
3
4  %.flagstat: %.bam
5      $(SAMTOOLS) flagstat $< > $@
```

Listing 1: Makefile snippet.

Automatic scheduling on a cluster

```
1  %.bam: %.sam
2      $(SAMTOOLS) view -bt $(call MAKEREF, $@) -o $@ $<
3
4  %.flagstat: %.bam
5      $(SAMTOOLS) flagstat $< > $@
```

Listing 1: Makefile snippet.

To fully exploit a cluster, we use the *Make* language.

- Only describe dependencies.
- Implicit workflow.
- Error control.

Automatic scheduling on a cluster

```
1  %.bam: %.sam
2      $(SAMTOOLS) view -bt $(call MAKEREF, $@) -o $@ $<
3
4  %.flagstat: %.bam
5      $(SAMTOOLS) flagstat $< > $@
```

Listing 1: Makefile snippet.

To fully exploit a cluster, we use the *Make* language.

- Only describe dependencies.
- Implicit workflow.
- Error control.

The pipeline we made is only 122 lines long.

MLST

```
1 CAATGATGATCGACAGTATGGCTGTGCTCGATATCTTCATTCTTGCGGCT
2 AAAGCGGGCGGCGAACCACCACAAAGAATACCGGAACGAAGAAGATTGCCA
3 GTACCGTTGCGGTCACCATCCCGCCATTACACCGGTACCTACTGCGTTC
4 TGCGCGCCGGAACCAGCACCAGTACTGATAACCAGCGGCATAACGCCGAG
5 GATAAACGCCAGCGAGGTCATCAGGATCGGACGTAAACGCATCCGCACCG
6 CATCAAGCGTGCCTTCAATCAGACCTTTACCTTCTTTATCCATCAAGTCT
7 TTGGCGAATTGACGATAAGGATCGCGTTCTTCGCCGACAACCCAATGGT
8 TGTGAGCAGGCTACCTGGAAGTAAACGTCATTGGTCAGGCCACGGAAGG
```

Listing 2: Part of the consensus sequence of *acrB*.

MLST

```
1 CAATGATGATCGACAGTATGGCTGTGCTCGATATCTTCATTCTTGGCGGCT
2 AAAGCGGCGGCGGAACCACCACAAAGAATACCGGAACGAAGAAGATTGCCA
3 GTACCGTTGCGGTCACCATCCCGCCATTACACCGGTACCTACTGCGTTC
4 TGCGCGCCGGAACCAGCACCAGTACTGATAACCAGCGGCATAACGCCGAG
5 GATAAACGCCAGCGAGGTCATCAGGATCGGACGTAAACGCATCCGCACCG
6 CATCAAGCGTTCGCTTCAATCAGACCTTTACCTTCTTTATCCATCAAGTCT
7 TTGGCGAATTOGACGATAAGGATCGCGTTCTTCGCCGACAACCCAATGGT
8 TGTGAGCAGGCOCTACCTGGAAGTAAACGTCATTGGTCAGGCCACGGAAGG
```

Listing 2: Part of the consensus sequence of *acrB*.

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

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 1: Part of the plasmids Excel file.

Gene detection

Reference	Gene	Length	Cov	#4/#3
AB699171	CMY-87	959	90	0.093
AB715422	IMP-34	742	125	0.168
AB737978	ACT-16	1062	202	0.190
AB753456	IMP-42	739	417	0.564
AB753457	IMP-40	739	414	0.560
AB753458	IMP-41	731	364	0.497
AC_000091 . 1	accD	915	915	1.000
AC_000091 . 1	acrA	1194	1194	1.000
AC_000091 . 1	acrB	3150	3150	1.000

Table 2: Part of the genes Excel file.

Reusability

Plasmids and genes can be added easily.

Reusability

Plasmids and genes can be added easily.

Plasmids.

- Download a reference sequence.
- Index the reference sequence.
- Put the files in the right folder.

Reusability

Plasmids and genes can be added easily.

Plasmids.

- Download a reference sequence.
- Index the reference sequence.
- Put the files in the right folder.

Genes:

- Download a reference sequence.
- Find the gene in this reference sequence.
- Write down the coordinates of the gene.

This part is automated.



Acknowledgements:

Sunita Paltansing
Henk Buermans
Sandra Bernards
Johan den Dunnen