HOW TO CONDUCT A FULL GENOME PROJECT

Historical perspective
Why sequence a genome ?
How to sequence a genome ?

Gilbert GREUB

Institute of Microbiology
University of Lausanne

INTRODUCTION

- Historical perspective
- Why sequencing bacterial genomes ?
- Who is sequencing bacterial genomes?
- Which genome should be sequenced ?
- How to sequence bacterial genomes ?

Historical perspective

1977 Invention of dideoxy chain terminator sequencing
Sanger et al. PNAS 1977;74:5463

1979 Sequencing of a bacteriophage genome
5.3 kb (bacteriophage phiX174)
Sanger et al. Nature 1979;265:887

1981 First human mitochondrial genome sequence
Anderson et al. Nature 1979;290:457

1982 First use of shotgun sequencing
48 kb (bacteriophage lambda)
Sanger et al. J Mol Biol 1982;162:729

1986 Development of automated fluorescent sequencing
Smith et al. Nature 1986;321:674

1995 First complete genome sequences of bacteria
Haemophilus influenzae — Fleischmann et al. Science 1995;269:496
Mycoplasma genitalium — Fraser et al. Science 1995;270:397

Historical perspective

1996 Complete genomes of two species of a genus
Mycoplasma pneumoniae — Himmelreich et al. NAR 1996;24:4420

1997 First genome of *Escherichia coli*
Blattner et al. Science 1997;277:1453

1998 Genome of *M. tuberculosis*
Cole et al. Nature 1998;393:537

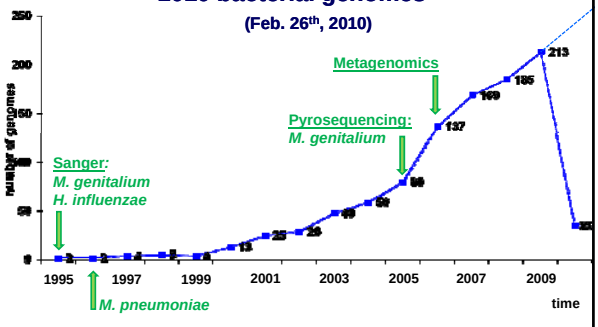
1999 Complete genomes of two isolates of a species
Helicobacter pylori — Alm et al. Nature 1999;397:176

2005 First genome done by pyrosequencing
Mycoplasma genitalium — Margulies et al. Nature 2005;437:376

2006 Bacterial metagenomic of the Sargasso sea
Venter et al. Science 2006;304:66
Metagenomic of human bowel microbiota
Gill et al. Science 2006;312:1355

Historical perspective

1010 bacterial genomes
(Feb. 26th, 2010)



http://www.genomesonline.org

INTRODUCTION

- Historical perspective
- Why sequencing bacterial genomes ?
- Who is sequencing bacterial genomes?
- Which genome should be sequenced ?
- How to sequence bacterial genomes ?

Why sequencing bacterial genomes ?

Considered as a « distraction of effort and of funding from hypothesis-driven research »
Pallen et al. Mol Mcrobiol 1999;32:907

⇩

Plateform for hypothesis generation

More efficient than the competitive duplication of effort to sequence the same gene cluster
(*icm/dot* system in *Legionella*)

Why sequencing bacterial genomes ?

Insight in the biology of model organisms
- *Escherichia coli* (Blattner et al. Science 1997;277:1453)
- *Bacillus subtilis* (Kunst et al. Nature 1997;390:249)

Insight in microbial evolution
- Evidences of reductive evolution
  - *Rickettsia prowazeckii* (Andersen et al. Nature 1998;396:133)
  - *Mycobacterium leprae* (Cole et al. Nature 2001; 409:1007)
- Unexpected level of horizontal transfer
  - 2nd genome of *Escherichia coli* (Hayashi et al. DNA Res 2001; 8:11)
- Amoebae as melting-pot for genes exchange
  - *Rickettsia belii* (Ogata et al. PLOS Genetic 2005)
  - Marseillevirus (Boyer et al. PNAS 2009)

Why sequencing bacterial genomes ?

Insight in the bacterial metabolism
- Lipopolysaccharide biosynthesis pathway
  - *Haemophilus influenzae* (Hood et al. Mol microbiol 1996;22:951)

Search for new vaccines candidates
- *Neisseria meningitidis* (Pizza et al. Science 2000;287:1816)

Search for new drugs/new drug targets
- ATP synthase of *M. tuberculosis* (Andries et al. Science 2005; 307:223)

Improve diagnostic tests
- Immunogenic proteins (Greub et al. PLOS One 2009; 4:8423)

Test the approach for larger genomic projects
- Human genome project (Venter et al. Science 2001; 291:1304)

Why sequencing bacterial genomes ?

Identify virulence factors

> 50% genomes from pathogens

40 species > 1 genome (2 to 12)

Explain taxonomic skewness in favor of:
- **Proteobacteria**
- **Firmicutes (Staphylococcus)**

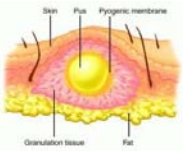
Staphylococcus aureus

Pigmented colonies

Gram positive coccis

Pyogenic bacteria

Epidermolysis



Genome sequencing of *S. aureus* allowed to identify:

VIRULENCE FACTORS

- Novel adhesins with a LPXTG motif (anchored in cell membrane)
- Exoenzymes encoding lipases, proteases
- Putative enterotoxin (homology to a diarrheal toxin of *Bacillus cereus*)
- Hemolysins
- Leukocidins

Reviewed in
Pallen et al. Bacterial pathogenomics 2007;5:120

- Bacteriophage encoding the Panton-Valentin leukocidin gene (*PV-luk*)

Baba et al. Lancet 2002;359:1819

Genome sequencing of *S. aureus* allowed to identify:

MOBILE GENETIC ELEMENTS

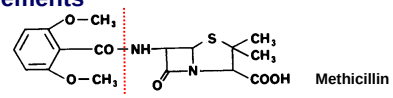
- Bacteriophages
- Staphylococcal cassette chromosomes (SCC)
- *S. aureus* pathogenicity islands
- Plasmids
- Transposons
- Conjugative elements
- Insertion sequences (IS)

Reviewed in
Pallen et al. Bacterial pathogenomics 2007;5:120

Genome sequencing of *S. aureus* allowed to conclude:

Many mobile genetic elements:

- Frequent horizontal transfer
- Virulence genes carried on the mobile genetic elements
- Resistance genes carried on the mobile genetic elements



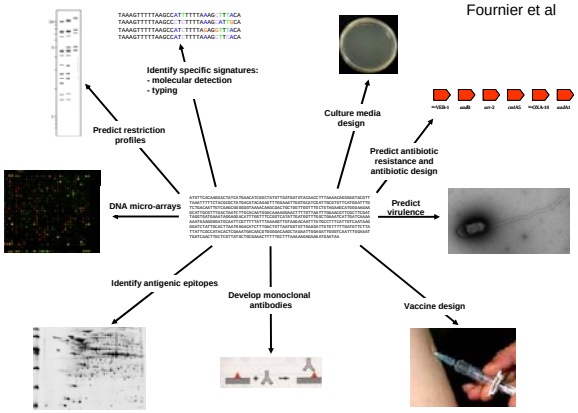
Methicillin

Ex: SCC elements: *mecA*, methicillin resistance (MRSA)
Plasmid pUSA03: resistance to macrolides/lincosamides
Plasmid pLW1043:vancomycine resistance transposon

Reviewed in
Pallen et al. Bacterial pathogenomics 2007;5:120

Why sequencing bacterial genomes ?

Fournier et al



INTRODUCTION

- Historical perspective
- Why sequencing bacterial genomes ?
- Who is sequencing bacterial genomes?
- Which genome should be sequenced ?
- How to sequence bacterial genomes ?

Who is sequencing genomes

USA	
The Institute for genomic research	www.tigr.org
DOE Joint genome institute	www.jgi-doe.gov
University of Washington	www.genome.washington.edu
University of Wisconsin	www.genome.wisc.edu
UK	
The Sanger Institute	www.sanger.ac.uk/projects
xBASE	http://xbase.bham.ac.uk/
France	
Genoscope	www.genosope.cns.fr
Unité des Rickettsies	www.ifr48.fr
Japan	
National biotechnology center	www.bio.nite.go.jp
RIKEN genomic center	www.gsc.riken.go.jp

Who is sequencing genomes

Before 2006 : mainly large genome sequencing center

Many groups worldwide since the availability high-throuput sequencing technologies

In Switzerland:

S. aureus (Lausanne/Geneva)
Chlamydia (Lausanne)
Mycoplasma (Bern/Lausanne)
Bartonella (Basel)
…

INTRODUCTION

- Historical perspective
- Why sequencing bacterial genomes ?
- Who is sequencing bacterial genomes?
- Which genome should be sequenced ?
- How to sequence bacterial genomes ?

Which genomes to sequence ?

Small genomes

First complete genome sequences of bacteria
Mycoplasma genitalium Fraser et al. Science 1995;270:397

Complete genomes of two species of a genus
M. pneumoniae/genitalium Himmelreich et al. NAR 1996;24:4420

First genome done by pyrosequencing
Mycoplasma genitalium Margulies et al. Nature 2005;437:376

Historical strains

Haemophilus influenzae Fleischmann et al. Science 1995;269:496

↳ Discovery of restriction enzymes by Smith (1970)

Empirical choice → rationale choice

Which genomes to sequence ?

Small genomes (*Mycoplasma genitalium*)
Historical strains (*H. influenzae*)
Pathogenic strains (*S. aureus, M. tuberculosis, …*)
Model organisms (*E.coli*)
Special phenotype of a strain (MRSA)
Mutant strain (nitrosoguanidine mutants of *C. abortus*)
Type strain (ATCC/DSMZ/...)
No genetic tools available (Chlamydia, Rickettsia, Coxiella)
No in vitro growth (T. pallidum, M. leprae)
Area of expertise ↳ **Which species ?**
Which strain ?

Which genomes to sequence ?

Well characterized strain
with considerable body of prior knowledge
(may have lost its pathogenic potential)

versus

Fresh, minimally passaged clinical isolate
(may be genetically intractable)

↳ **How many passages ?**

INTRODUCTION

- Historical perspective
- Why sequencing bacterial genomes ?
- Who is sequencing bacterial genomes?
- Which genome should be sequenced ?
- How to sequence bacterial genomes ?

Steps of a genome project

1. Choice of species, strain ⇨ culture: DNA
2. Sequencing (shotgun/pyrosequencing)
3. Assembly
4. Gap closure
5. Annotation
6. Exploit your data (wetlab)
7. Communicate your results

10-20% of effort

80-90% of time and effort

↳ **Use of raw genome sequences: the « dirty genome » approach**

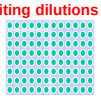
Greub et al. PLOS One 2009; 4:8423)

Culture

Grow enough bacterial cells to yield enough genomic DNA to create a shotgun library of *E.coli*

- **Biosafey issue with BSL3 pathogens**
- **Purification problems with strict intracellular bacteria**

Limiting dilutions



Gastrographine gradients
Sucrose barrier

- **No growth of some agents**
  - **17 months to get enough *T. whipplei***
  - **Growth in Armadillo of *Mycobacterium leprae***
  - **Growth in rabbit testis of *Treponema pallidum***

Sanger-shotgun sequencing:

Shotgun library in *E.coli*

Paired-end reads from:

- **10'000-80'000 small inserts: 1-5-6kb**
- **1'000-5'000 larger inserts: 10-40 kb for scaffolding**

⇩

Sanger dideoxy sequencing (Nobel Prize in 1980)

- **automated capillary sequencing**
- **reads of about 800 bp**

⇩

Coverage of about 10x

(amount of sequence obtained/estimated genome size)

Frangeul et al. Microbiology 1999

454 pyrosequencing

Genome Sequencer 20 System	Genome Sequencer FLX System	Genome Sequencer FLX Titanium System
		
Bp per run: 20 Mb Reads per run : 200'000 Read length : ~100 bp	Bp per run: 100 Mb Reads per run : 400'000 Reads length: 250 bp	Bp per run: 400 Mb Reads per run : >1'000'000 Reads length: 400 bp

Fast / high througput

«Long» reads (compared to Solexa)

Problems with homopolymers ?

Pyrosequencing versus Sanger

Pyrosequencing	Sanger
+	-
• Paired-end sequencing	• Linked-clones
-	+
• Faster • No clonal bias	• Slower • Clonal bias
• Problems with homopolymers • New – Few softwares	• No problems with homopolymers • Old – Largely tested – Freely available good softwares for genome assembly & finishing

Solexa - illumina technology

Genome Analyzer II System	HiSeq 2000
	
Throughput: 4 Gb per day Reads per run: 250 millions Reads length: 35 to 100 bp	Throughput: up to 25 Gb per day Reads per run : 1 billion Reads length: 35 to 100 bp

Fast

Short reads

Accurate in homopolymers

Our own experience in Lausanne

Parachlamydia ***Protochlamydia*** ***Waddlia***

2 runs

Tagged
&
Pooled

1 canal
1/3 canal per bacterium

1 run

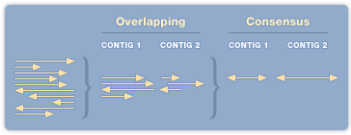
Conclusion

Correction of homopolymer errors
is possible with Solexa technology
ans is mainly useful :

- when sequencing with GS20
- « low » coverage with 454 reads
- depend on GC content, read quality, …

De novo genome assembly

Puzzle
- Millions of pieces
- Malformed pieces
- Often missing pieces
- Lots of identical pieces (blue sky)



Assembly by mapping

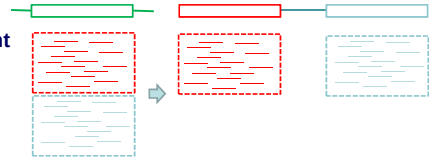
Know the final picture
Need a reference genome
Comparative genomic projects

Analysis of:
- single nucleotide polymorphisms (SNPs)
- deletions
- insertions
- inversions

difficult to detect

Genome assembly

Sequences issue
- Repeats
- GC content



Different software according to sequencing instrument:
- 454 sequences with Newbler
- Illumina sequences with ABySS/Velvet/…

⇩

First do each assembly with the appropriate software
and then mix the assemblies

Genome assembly

Quality of assembly
- Number of contigs
- N50 (length of contig that contain 50% of bases)
- Minimum (smaller contig size)
- Maximum (larger contig size)
- Consensus size (genome size)

Stringent criteria may improve assembly quality
by reducing the pairing possibilities

Genome assembly

Example

Assemblage	Nb of large contigs*	Nb of bases	Nb of contigs	Nb of bases	N50	Overlap Length	Overlap Identity
KNic454o1	101	2970836	232	2992519	81472	50	90
KNic454o2	98	2969410	229	2992100	87058	45	90
KNic454o3	94	2967805	242	2992303	91286	40	90
KNic454o4	94	2972966	209	2992305	87066	35	90
KNic454o5	104	2971832	303	3002554	91440	50	95
KNic454o6	93	2971911	281	3001360	97920	45	95
KNic454o7	99	2972218	278	2999600	91379	40	95
KNic454o8	98	2970140	273	2997839	91387	35	95

* larger than 500 bp

Gap closure

The strategies

Test combination of primers by PCR + sequencing (Multiplex/ long-range PCRs)

Reduce the complexity by:
- remove very small contigs initially (generally containing repeated sequences)
- mapping with a related genome (if available)
- use of cosmids / fosmids library

Genome walking

Re-assembly (improved softwares)

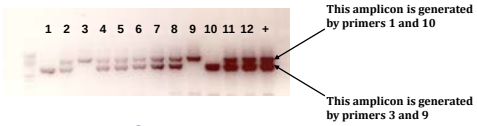
Gap closure

Test a variety of combinations

Multiplex PCRs

Primers divided in pools of 4 primers and PCRs performed using two pools of primers

When a PCR is positive, then eliminate each primer of the two pools, one at a time

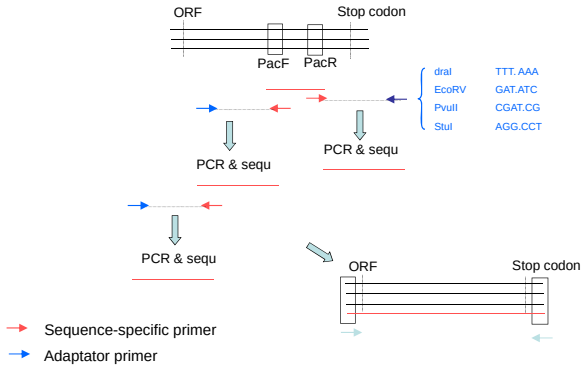


Long-range PCRs

To overcome difficult regions (repeats)

Gap closure

Genome walking



Gap closure

Example

93 contigs > 500bp

Avoid integrases, rhs, transposases, …

57 contigs > 3000bp

Gap closure

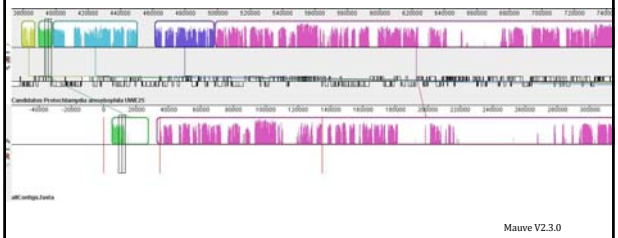
with 57 contigs:
1653 combinations should be individually challenged

Mapping

Gap closure

Mapping

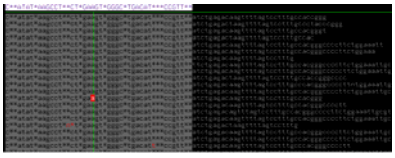
Alignment with related genomes if available may guide the first PCRs

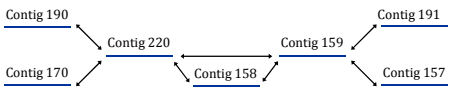


Gap closure

Look behind contigs

… the similarity between reads at contig ends allowing to draw a contig map





Gap closure

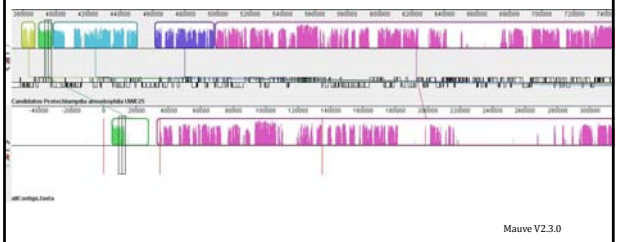
Major difficulties

- Ribosomal operons: 4 almost identical copies
- Rhs sequences: long repeated sequences (around 10 Kb) responsible for genome rearrangements and encoding internal multiple repetitions
- Repeated elements: transposases, integrases

Gap closure

Mapping

Alignment with related genomes if available may guide the first PCRs

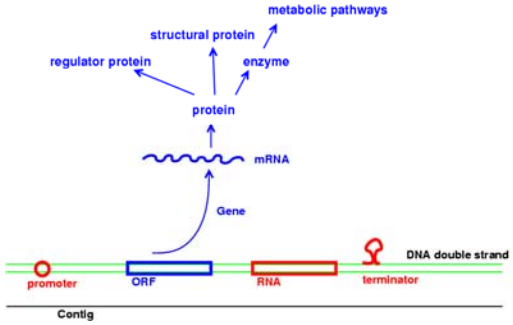


Annotation

- Only start when completed genome sequence (or almost complete)
- Repetitive sequences
- RNA: tRNAs / rRNAs / small regulatory RNAs
- Recently acquired DNA
- Leading strand / lagging strand } GC skew
- Origin of replication
- Identification and annotation of open-reading frames (ORFs)

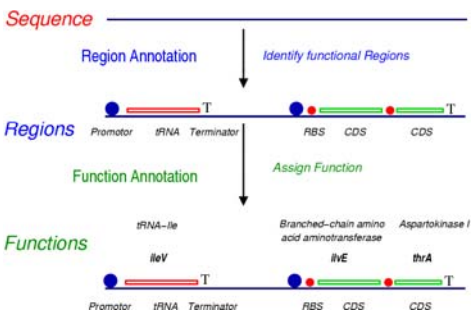
Annotation

From DNA sequence to molecular function



Annotation

The process of genome annotation



Regions annotation

Prediction of protein-coding sequences

Open-reading frames (ORFs)

⇩

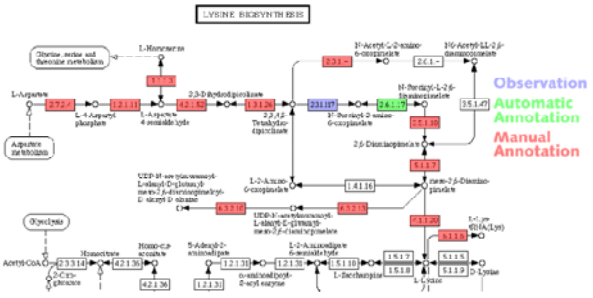
Protein-coding sequences (CDSs)

Predicted using Markov model
i.e. Glimmer
(Gene locator and Interpolated Markov Modeler)

1. Training set of trusted CDSs (long ORFs > 500 bp)
2. Use the model to predict CDSs

Annotation

GenDB – KEGG browser



Annotation

Bacterial genomes range from 1 - 13 mio basepairs
Approx. 1000 genes per 1 mio basepairs

1-10 tools used for gene prediction (per genome)
10-100 tools used for function prediction (per gene)

100 tools for about 2'000 different genes for a 3Mb genomes !
=> 80'000 tools are run

=> Extensive need for automation
=> Need for visualization

Steps of a genome project

1. **Choice of species, strain ⇨ culture: DNA**
2. **Sequencing (shotgun/pyrosequencing)**
3. **Assembly**
4. **Gap closure**
5. **Annotation**
6. **Exploit your data (wetlab)**
7. **Communicate your results**

10-20% of effort

80-90% of time and effort