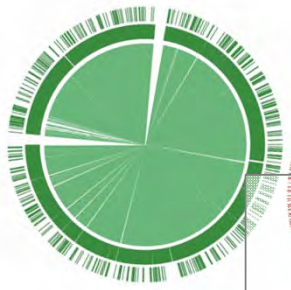


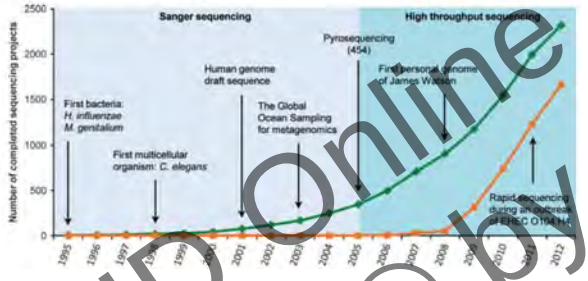
How to sequence a genome ?

Prof. Gilbert GREUB
Institute of Microbiology
University of Lausanne & University Hospital Center
Lausanne, Switzerland

INTRODUCTION

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

Historical perspective



Bertelli & Greub. Clin Microbiol Infect 2013

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

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

INTRODUCTION

- Historical perspective
⇨ **Why 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
  - 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

- Catalase
- T3SS
- Adhesins
- …

INTRODUCTION

- Historical perspective
- Why sequencing bacterial genomes ?
- Which genome should be sequenced ?

⇨ How to sequence bacterial genomes ?

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)
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 — 10-20% of effort
2. Sequencing (shotgun/pyrosequencing)
3. Assembly
4. Gap closure
5. Annotation
6. Exploit your data (wetlab)
7. Communicate your results

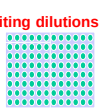
80-90% of time and effort

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*

Steps of a genome project

1. Choice of species, strain ⇨ culture: DNA — 80-90% of effort
2. ⇨ Sequencing (shotgun/pyrosequencing)
3. Gap closure
4. Annotation
5. Exploit your data (wetlab)
6. Communicate your results

10-20% of time and effort

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 FLX Titanium System

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

Solexa - illumina technology

HiSeq 2000

Throughput: up to 25 Gb per day
Reads per run : 1 billion
Reads length: 35 to 100 bp

Fast

Short reads

Accurate in homopolymers

Steps of a genome project

1. Choice of species, strain ⇨ culture: DNA — 10-20% of effort
2. Sequencing (shotgun/pyrosequencing)
3. ⇨ Assembly
4. Gap closure
5. Annotation
6. Exploit your data (wetlab)
7. Communicate your results

80-90% of time and effort

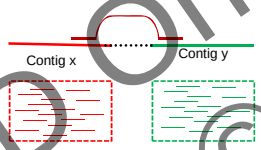
De novo genome assembly

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

Scaffold

Contigs

Reads



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

Steps of a genome project

1. Choice of species, strain ⇨ culture: DNA — 10-20% of effort
2. Sequencing (shotgun/pyrosequencing)
3. Assembly
4. ⇨ Gap closure
5. Annotation
6. Exploit your data (wetlab)
7. Communicate your results

80-90% of time and effort

SKIP THIS STEP
⇩
"DIRTY GENOME"

Greub et al. PLOS One 2009; 4:8423)
Bertelli & Greub. CMI 2013

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

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

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

Function annotation

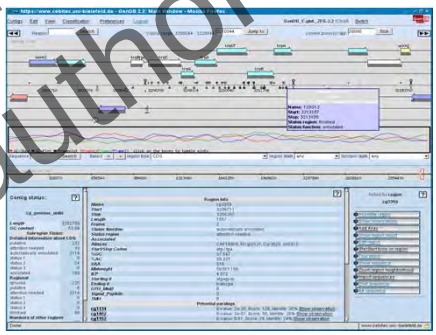
Bacterial genomes range from 1 - 13 mio basepairs
Approx. 1000 genes per 1 mio basepairs

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

Function annotation with GenDB



http://www.cebitec.uni-bielefeld.de

CDS, tRNA, rRNA, IS elements, oligos, mutations, operons, terminators, …

Automatic & manual annotation

Submission to BLAST against nr and SwissProt, Pfam, TIGRfam, Interpro, SignalP, TMHMM, ...

KEGG, COG, GO

Genome maintenance and re-annotation

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

Conclusions