Next Generation Sequencing and pathogen identification

Suzanne Bastian, Mily Leblanc-Maridor, Olivier Plantard

Introduction : definitions, aims, history

Main sequencing methods and their evolution

NGS and bioinformatics

NGS : the entrance of biology in « big science »

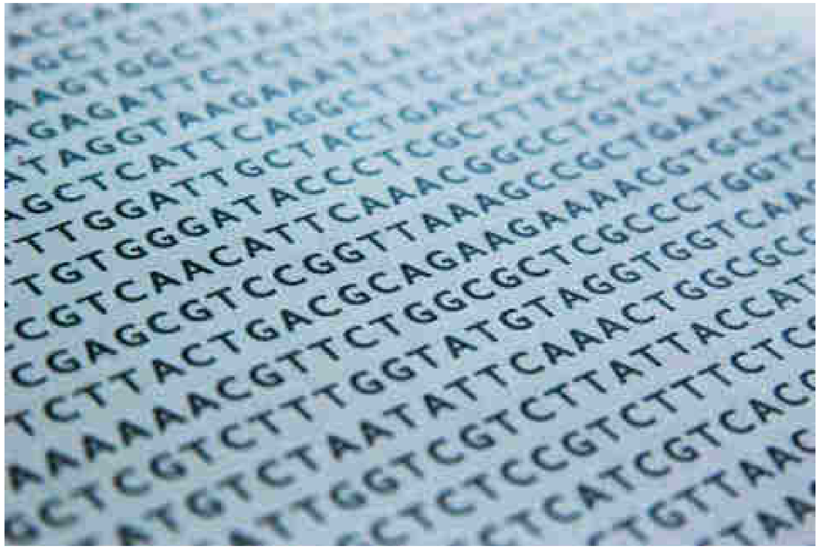
Séminaire Cœur de BioEpAR

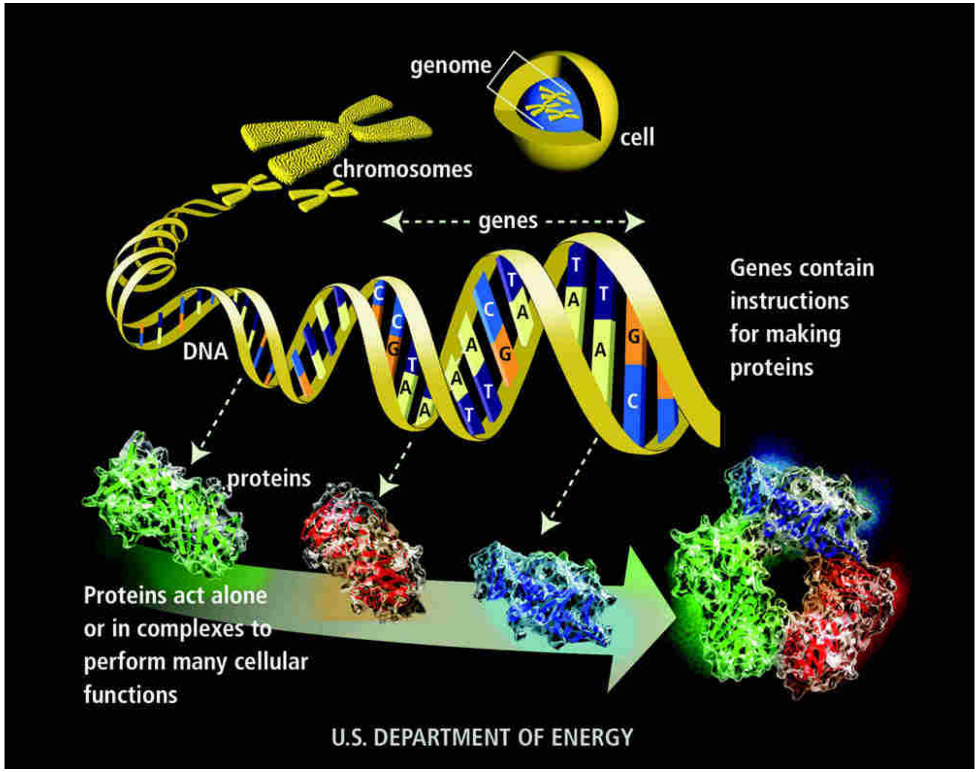
Jeudi 24 avril 2014

1) Introduction : definitions, aims, history

NGS = Next Generation Sequencing / **High Throughput** Sequencing
(robotisation, parallélisation)

Sequencing = détermination de l'ordre des quatres nucléotides (G,A,T,C) des molécules d'ADN, support de l'information génétique d'un organisme




genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

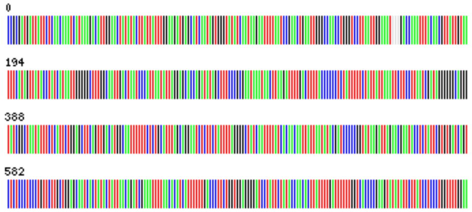
Séquençage = méthode de choix pour l'étude des **génomes** (contenu en ADN d'une cellule [exhaustif])
Génomique = comprendre comment fonctionne le génome
Le développement des NGS et l'essor de la génomique sont intimement liés.

Cette nouvelle accessibilité du génome rend possible son étude à différentes échelles
(espèce, population, cellule)

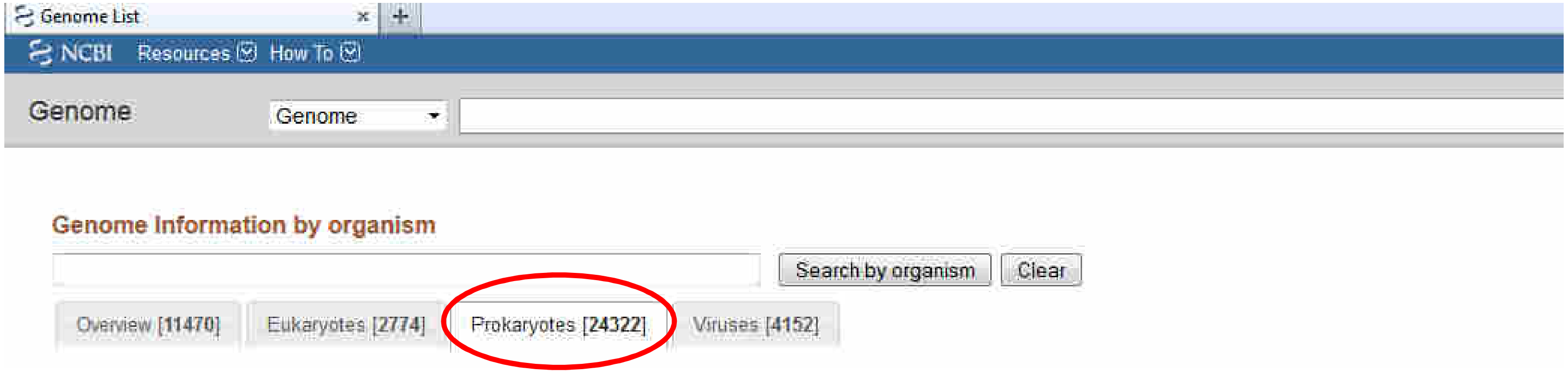
1) Introduction : definitions, aims, history

Génomique et identification de pathogènes :

➔ Métagénomique échantillons complexes
➔ Métabarcoding (code barre pour chaque espèce / base de données complète)



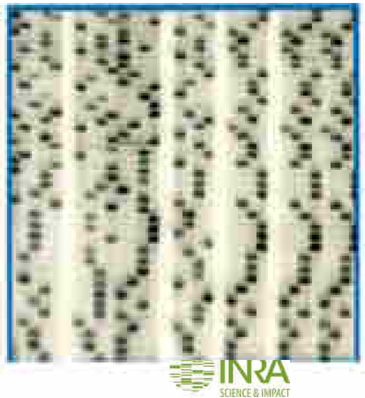
Nombre de génome bactériens connus > 24 000.



7300 espèces de bactéries connues à ce jour.
Estimation du nombre d'espèces : entre 5 et 10 millions

2) Main sequencing methods and their evolution

Pre-1992
"old fashioned
way"



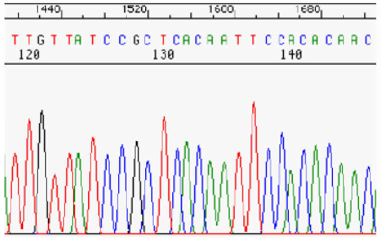
3000 nucleotides
per week

1992-1999
ABI 373/377

1999
ABI 3700

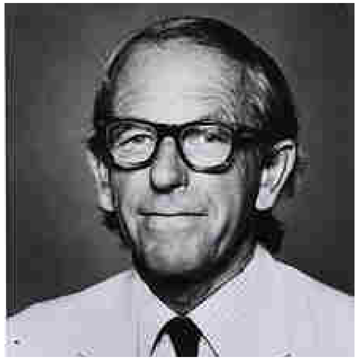
2003

ABI 3730XL



690 000 nucleotides
per day / machine

Séquençage « Sanger »

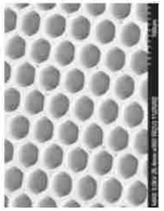


1977 : first genome sequenced
virus bactériophage ϕX174

Karen Staehling-Hampton

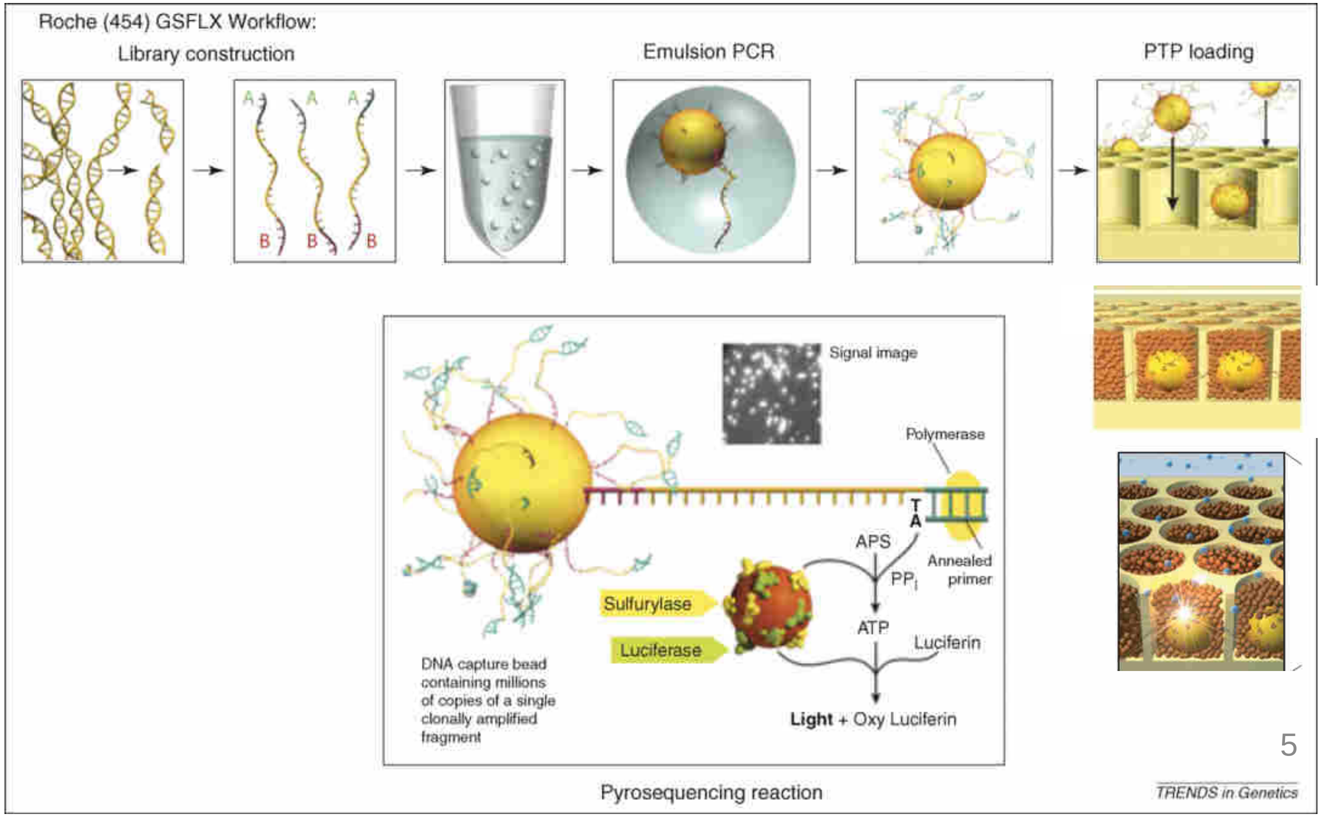
Le séquençage massivement parallèle

NGS 1.5 : Séquenceur Roche / 454: La révolution du pyroséquençage (October 2005)



Version Titanium:

- Lectures de ~ 400 bases
- 1 millions reads/run
- 400 Mb /jour

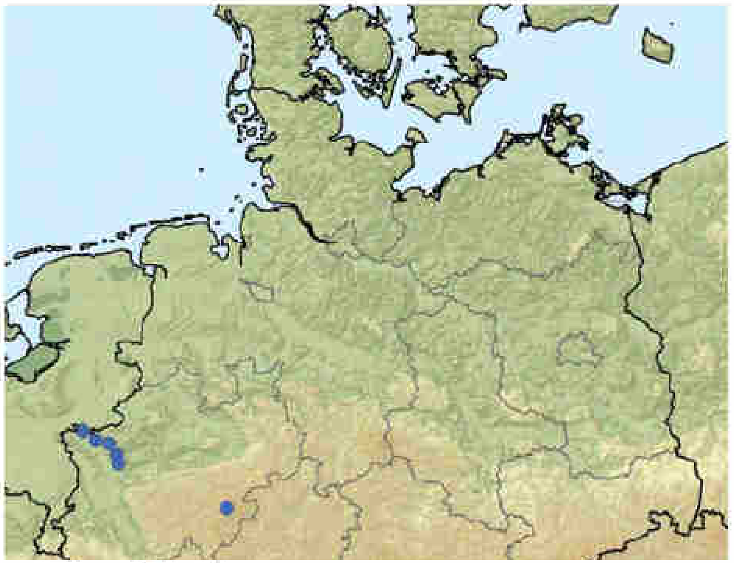


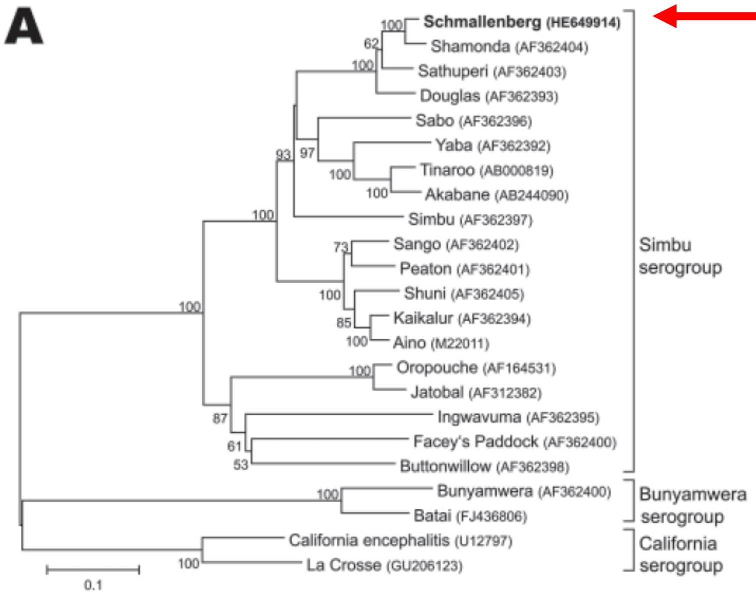
Emerging Infectious Diseases • www.cdc.gov/eid • Vol. 18, No. 3, March 2012

Novel Orthobunyavirus in Cattle, Europe, 2011

Bernd Hoffmann,[1] Matthias Scheuch,[1] Dirk Höper, Ralf Jungblut, Mark Holsteg, Horst Schirrmeier, Michael Eschbaumer, Katja V. Goller, Kerstin Wernike, Melina Fischer, Angele Breithaupt, Thomas C. Mettenleiter, and Martin Beer

In 2011, an unidentified disease in cattle was reported in Germany and the Netherlands. Clinical signs included fever, decreased milk production, and diarrhea. Metagenomic analysis identified a novel orthobunyavirus, which subsequently was isolated from blood of affected animals. Surveillance was initiated to test malformed



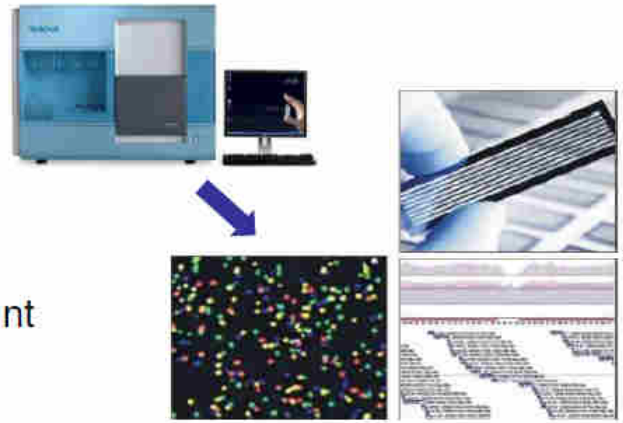


The Study

On a farm near the city of Schmallenberg (North Rhine-Westphalia, Germany; Figure 1), 3 blood samples obtained in October 2011 from dairy cows that had clinical signs at sampling (Table, BH 80/11) were pooled and analyzed by using metagenomics. We also investigated a blood sample from a healthy animal from a different farm (Table, BH 81/11). For metagenomic analysis, 4 sequencing libraries (Table) were prepared and sequenced by using the 454 Genome Sequencer FLX (Roche, Mannheim, Germany). Two libraries each were generated from DNA and RNA isolated from plasma samples (Table). By using a combination of BLAST (*1*) and sequence mapping with the 454 reference mapper application (version 2.6; Roche), reads were classified into different superkingdoms (Table). In addition to the anticipated high number of host

NGS 2.0 : Illumina/Solexa (february 2007)

four-color DNA sequencing-by-synthesis using reversible terminators with removable fluorescent dyes.



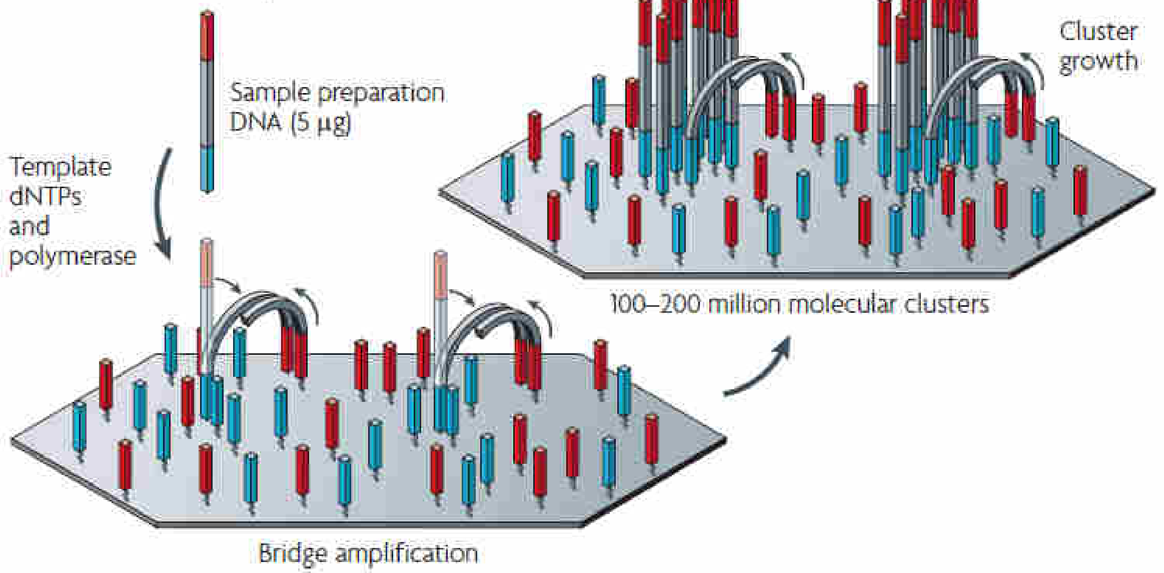
Version GAII

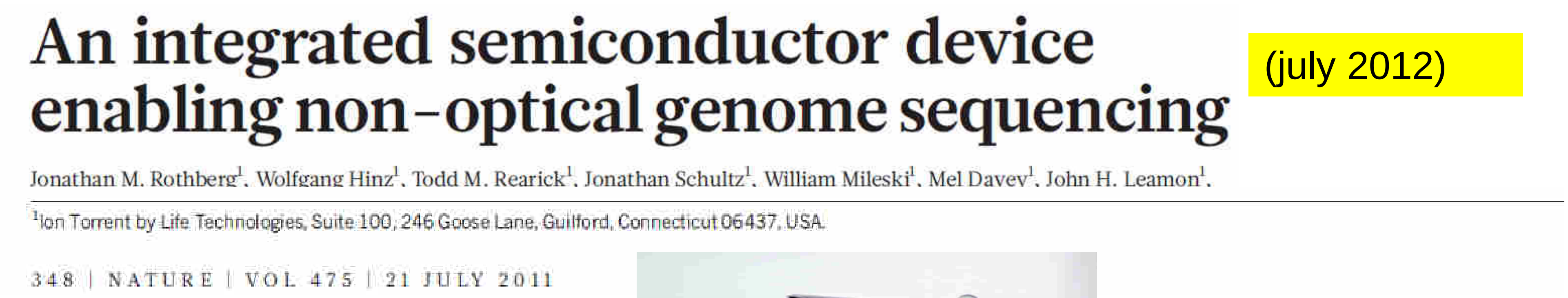
- Lectures de ~ 100 nt
- ~ 100 M lectures
- 10 Gb /run
- 1.5 Gb /jour

b Illumina/Solexa

Solid-phase amplification

One DNA molecule per cluster





An integrated semiconductor device enabling non-optical genome sequencing

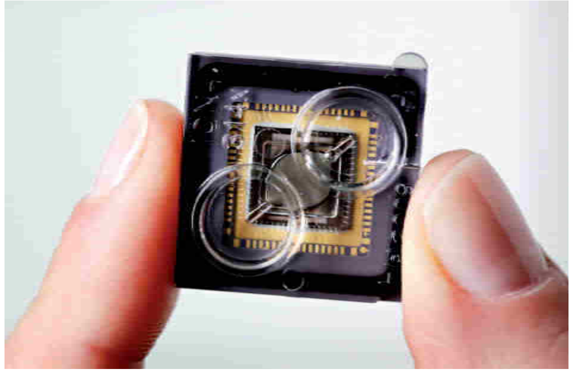
Jonathan M. Rothberg[1], Wolfgang Hinz[1], Todd M. Rearick[1], Jonathan Schultz[1], William Mileski[1], Mel Davey[1], John H. Leamon[1],

[1] Ion Torrent by Life Technologies, Suite 100, 246 Goose Lane, Guilford, Connecticut 06437, USA.

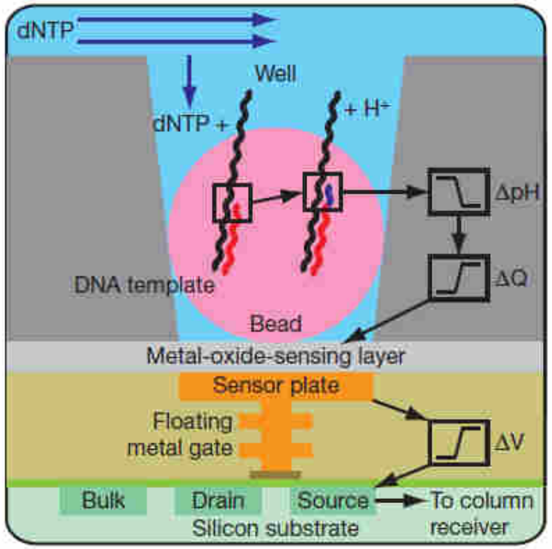
348 | NATURE | VOL 475 | 21 JULY 2011

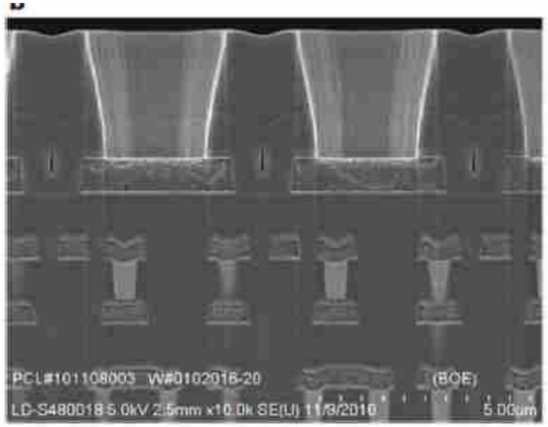
(july 2012)

49 500 $



99 $





PCL#101108003 W#0102016-20 (BOE)
LD-S480018 5.0kV 2.5mm x10.0k SE(U) 11/9/2010 5.00um

1,2 million de puits de 3,5 micrometres
1 run = 2 h = 25 millions de bases

OPEN ACCESS Freely available online

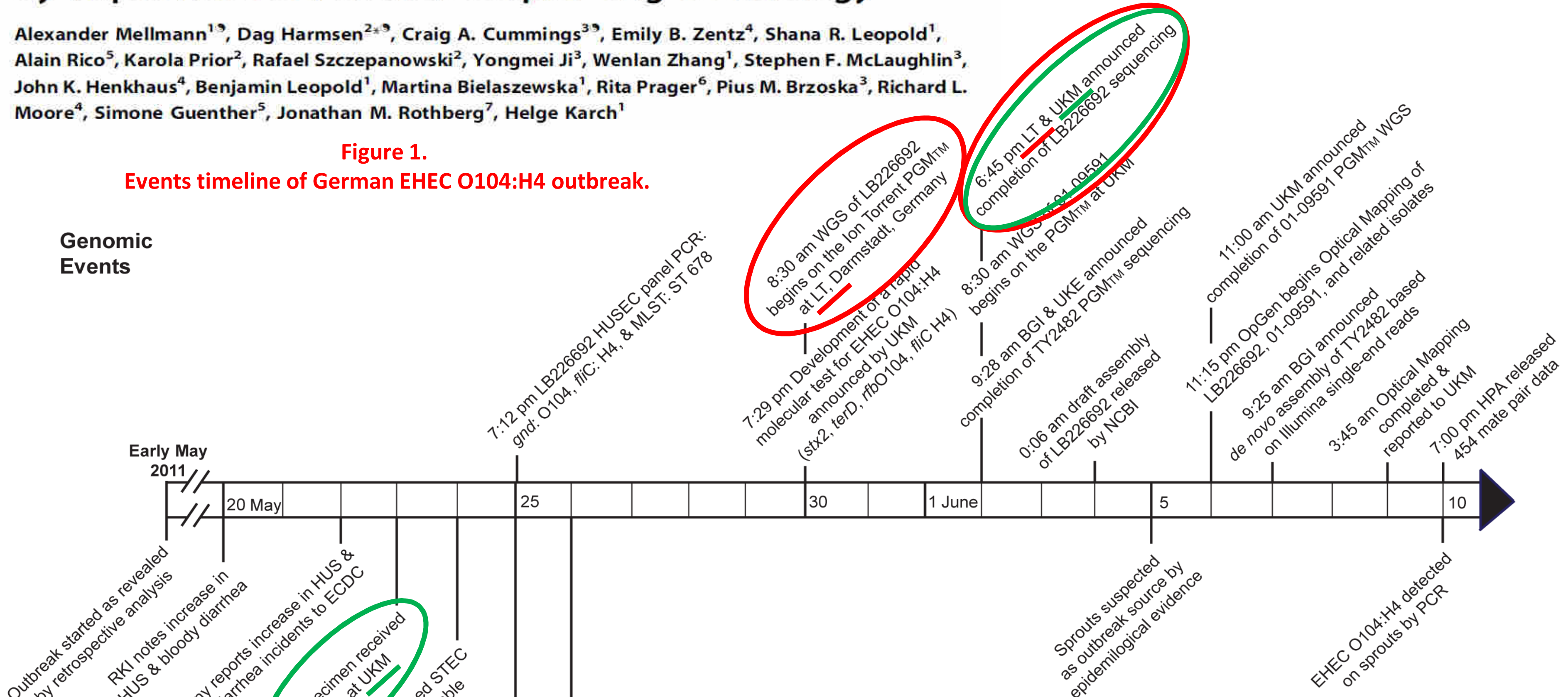
PLoS one

Prospective Genomic Characterization of the German Enterohemorrhagic *Escherichia coli* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology

Alexander Mellmann[1¶], Dag Harmsen[2*¶], Craig A. Cummings[3¶], Emily B. Zentz[4], Shana R. Leopold[1], Alain Rico[5], Karola Prior[2], Rafael Szczepanowski[2], Yongmei Ji[3], Wenlan Zhang[1], Stephen F. McLaughlin[3], John K. Henkhaus[4], Benjamin Leopold[1], Martina Bielaszewska[1], Rita Prager[6], Pius M. Brzoska[3], Richard L. Moore[4], Simone Guenther[5], Jonathan M. Rothberg[7], Helge Karch[1]

Séquençage en 62 heures

Figure 1.
Events timeline of German EHEC O104:H4 outbreak.



Genomic Events

7:12 pm LB226692 HUSEC panel PCR: *gnd*: O104, *fliC*: H4, & MLST: ST 678

8:30 am WGS of LB226692 begins on the Ion Torrent PGMTM at LT, Darmstadt, Germany

7:29 pm Development of a rapid molecular test for EHEC O104:H4 announced by UKM (*stx2*, *terD*, *rfb*O104, *fliC* H4)

6:45 pm LT & UKM announced completion of LB226692 sequencing

8:30 am WGS of 01-09591 begins on the PGMTM at UKM

9:28 am BGI & UKE announced completion of TY2482 PGMTM sequencing

0:06 am draft assembly of LB226692 released by NCBI

11:00 am UKM announced completion of 01-09591 PGMTM WGS

11:15 pm OpGen begins Optical Mapping of LB226692, 01-09591, and related isolates

9:25 am BGI announced *de novo* assembly of TY2482 based on Illumina single-end reads

3:45 am Optical Mapping completed & reported to UKM

7:00 pm HPA released 454 mate pair data

Early May 2011

20 May

25

30

1 June

5

10

Outbreak started as revealed by retrospective analysis

RKI notes increase in HUS & bloody diarrhea

Germany reports increase in HUS & bloody diarrhea incidents to ECDC

Stool specimen received at the at UKM

LB226692 diagnosed STEC pure culture available

ECDC releases warning

RKI & BfR warn against consuming cucumbers, leaf lettuce and tomatoes

Spanish cucumbers suspected as outbreak source by PCR

Sprouts suspected as outbreak source by epidemiological evidence

EHEC O104:H4 detected on sprouts by PCR

Epidemiologic Events

Dealing with Data

PERSPECTIVE

On the Future of Genomic Data

Scott D. Kahn

Science 2011

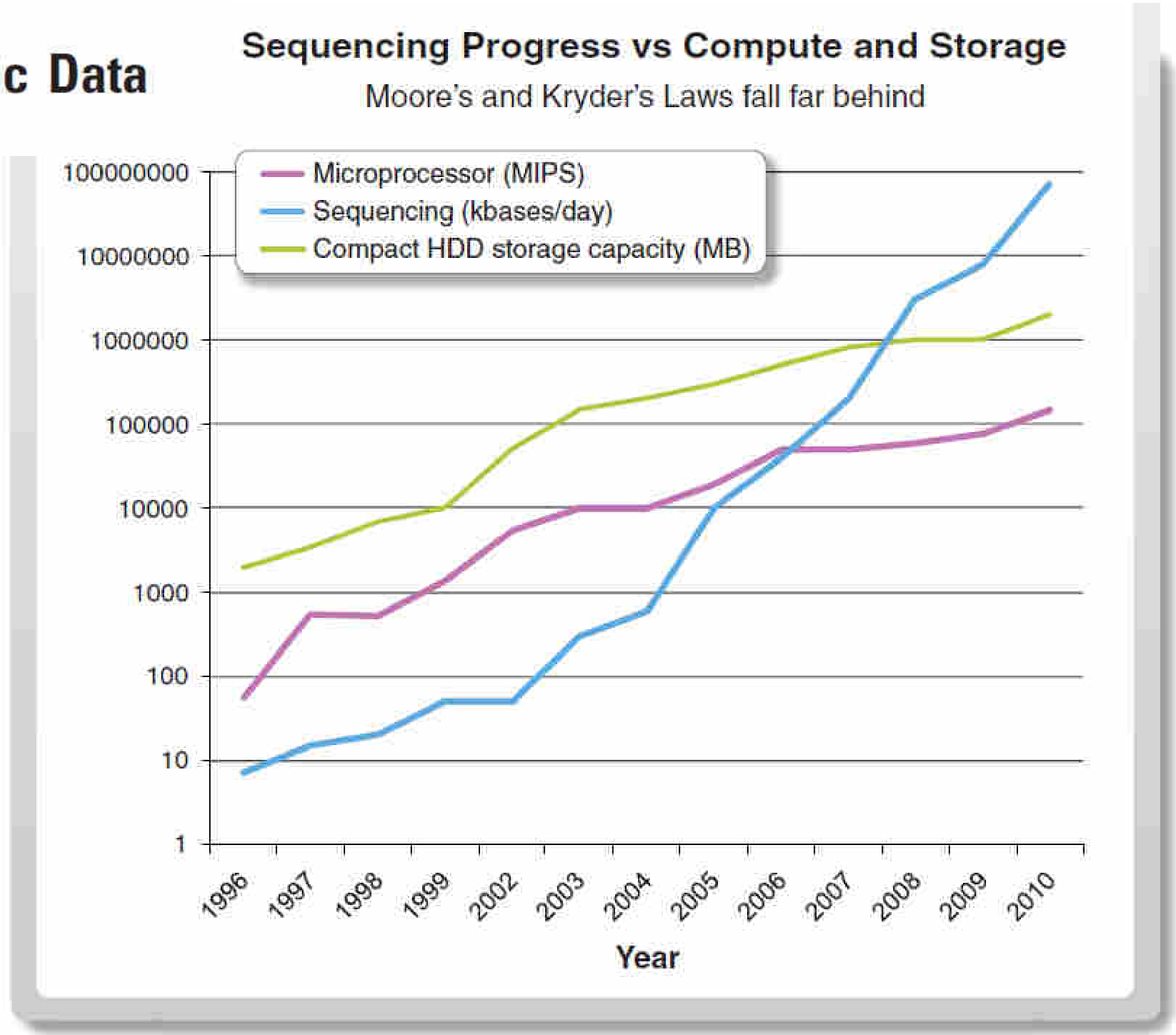
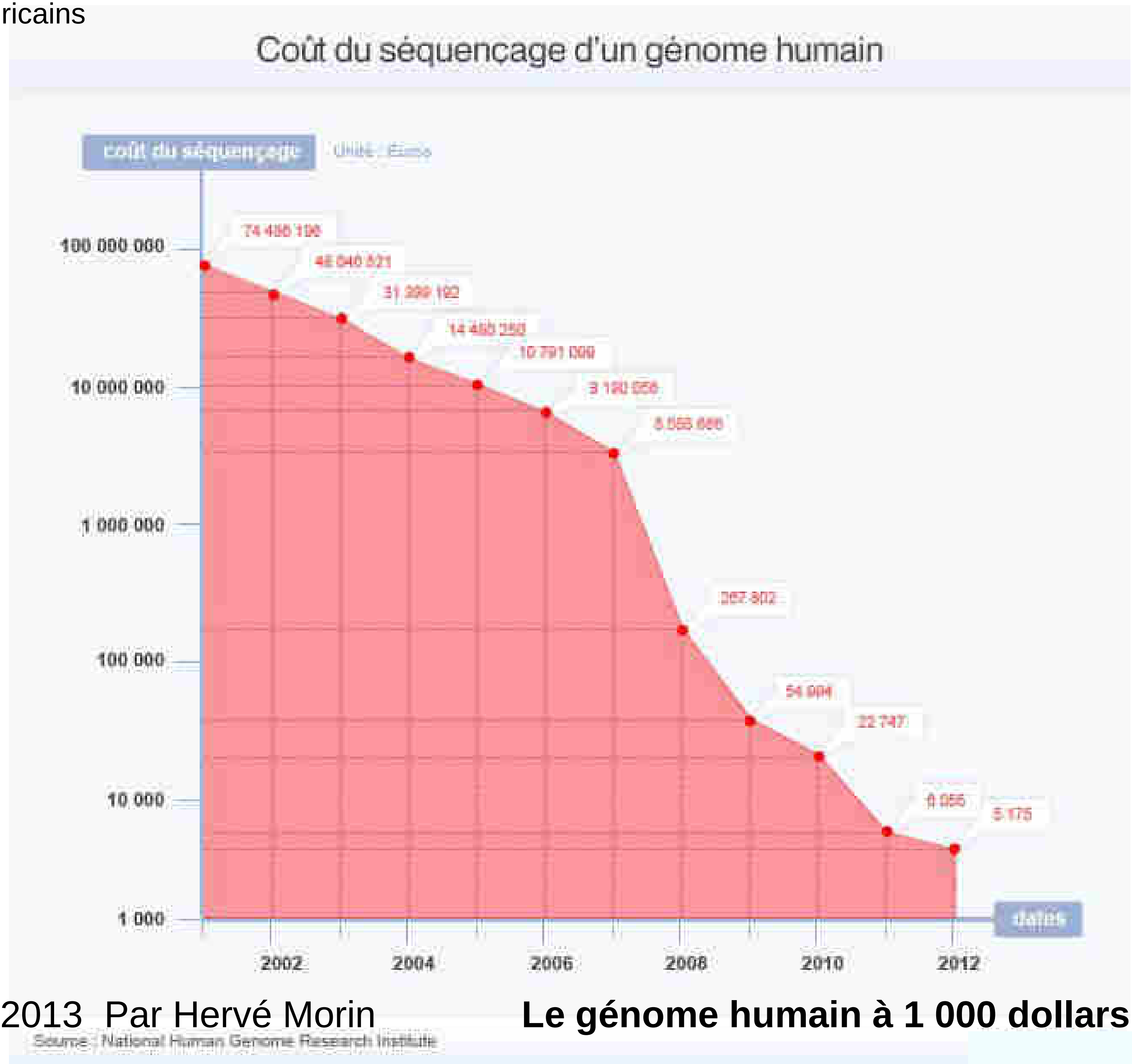


Fig. 1. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

Le programme HUMAN GENOME aura coûté, sur quinze ans, environ 2,7 milliards de dollars (2 milliards d'€) aux contribuables américains

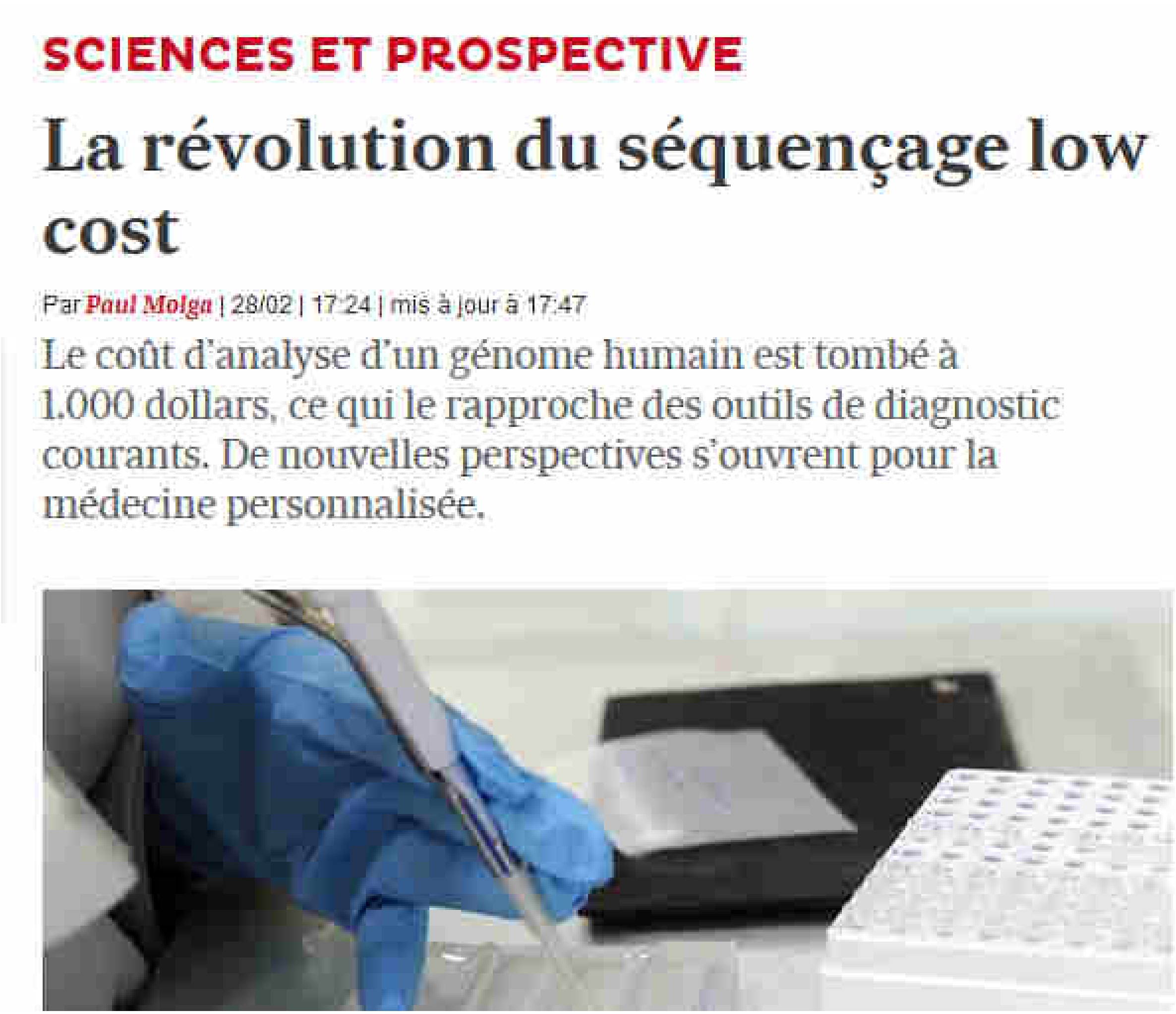

Coût du séquençage d'un génome humain

Source : National Human Genome Research Institute

LE MONDE | 01.01.2013 Par Hervé Morin

Le génome humain à 1 000 dollars

LesEchos.fr



SCIENCES ET PROSPECTIVE

La révolution du séquençage low cost

Par Paul Molga | 28/02 | 17:24 | mis à jour à 17:47

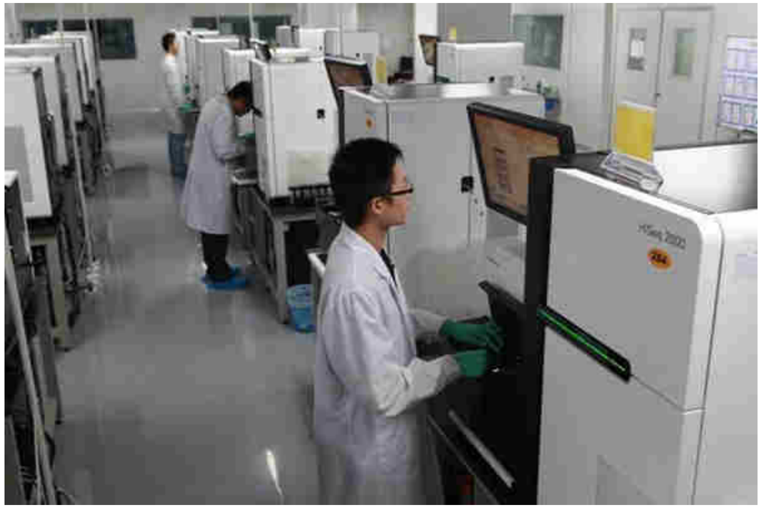
Le coût d'analyse d'un génome humain est tombé à 1.000 dollars, ce qui le rapproche des outils de diagnostic courants. De nouvelles perspectives s'ouvrent pour la médecine personnalisée.

FORBES LIFE: THE ULTIMATE YACHTING MACHINE
JANUARY 17 • 2011 EDITION
Forbes
THE NEXT $100 BILLION TECHNOLOGY BUSINESS
JONATHAN ROTHBERG
HIS PERSONAL GENOME MACHINE COULD CHANGE YOUR LIFE

Dès lors, l’arrivée de la génomique impose la mise en place de nouveaux types de laboratoires faisant évoluer la biologie d’un stade artisanal à un niveau beaucoup plus automatisé, quasi industriel.

Les grands instruments de la biologie moléculaire, prémices de la médecine de demain **Pierre Tambourin**

Sanger Institute, UK



Beijing Genomics Institute, China

BGI recevra 1.5 milliard de $ de “fonds collaboratifs” sur les 10 prochaines années de la China Development Bank

Shenzen, 500 bioinformaticiens

Changement profond des métiers en biologie :

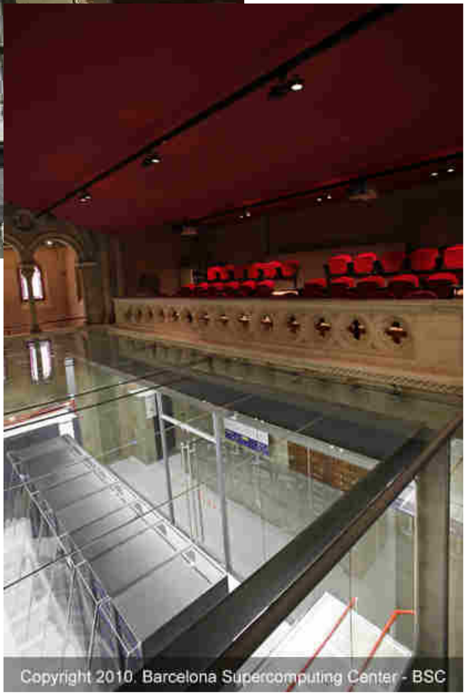
↘ ratio entre « biologie humide / biologie sèche »

Marenostrum, le supercalculateur le + puissant d'Europe

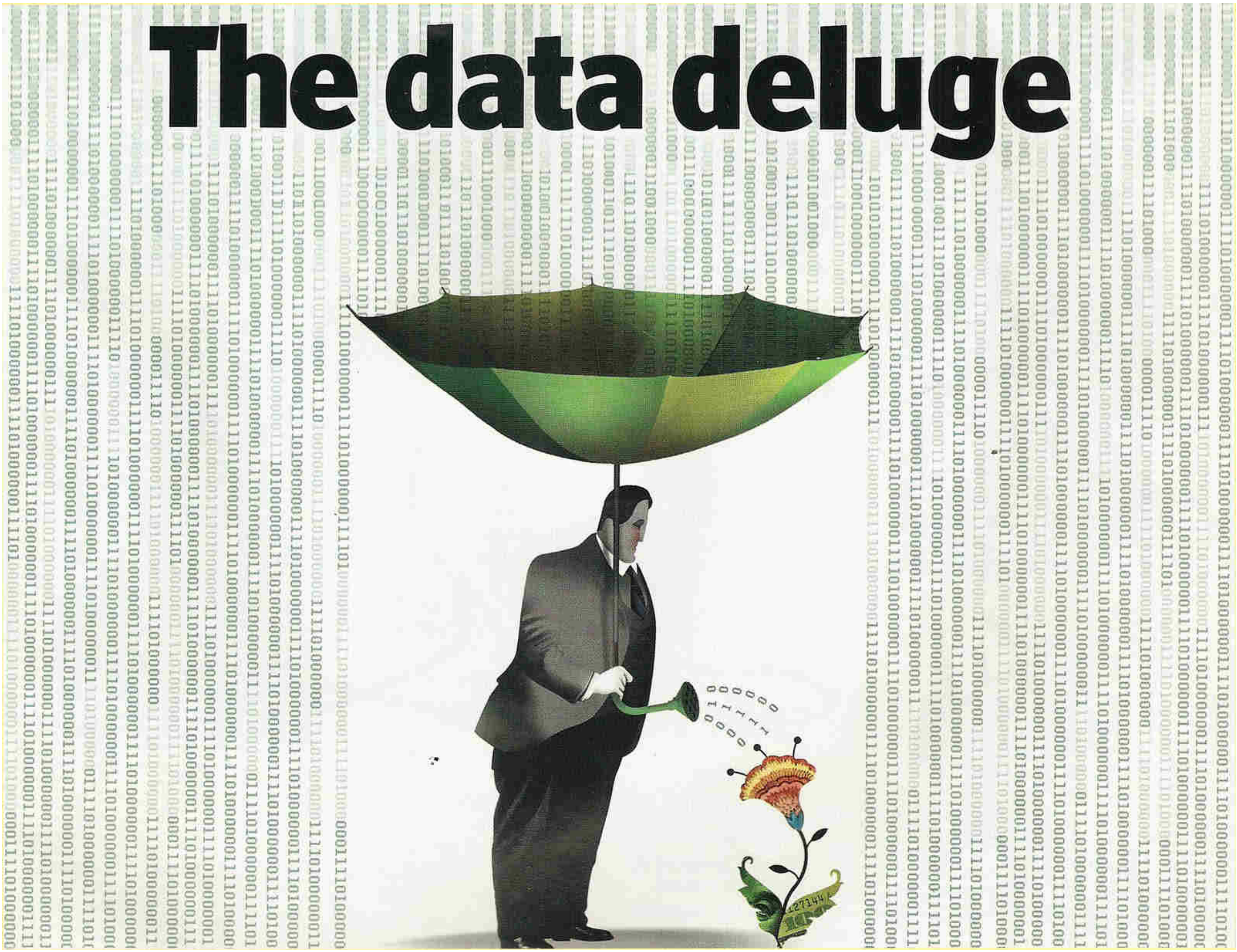
Copyright 2010. Barcelona Supercomputi

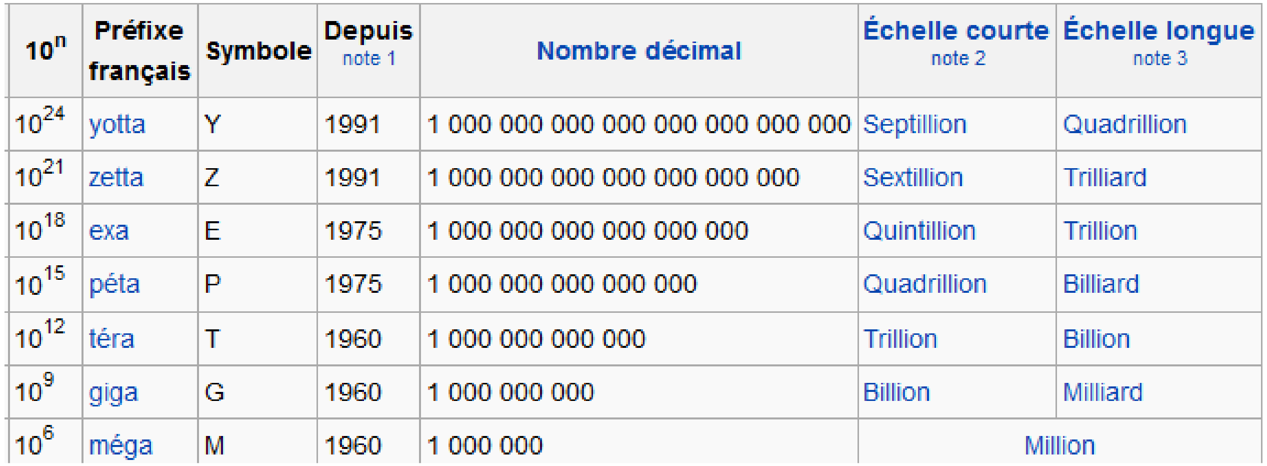
Copyright 2006, Barcelona Supercomputing Center - BSC

Chapelle Terro Girona, Barcelone



Copyright 2010. Barcelona Supercomputing Center - BSC

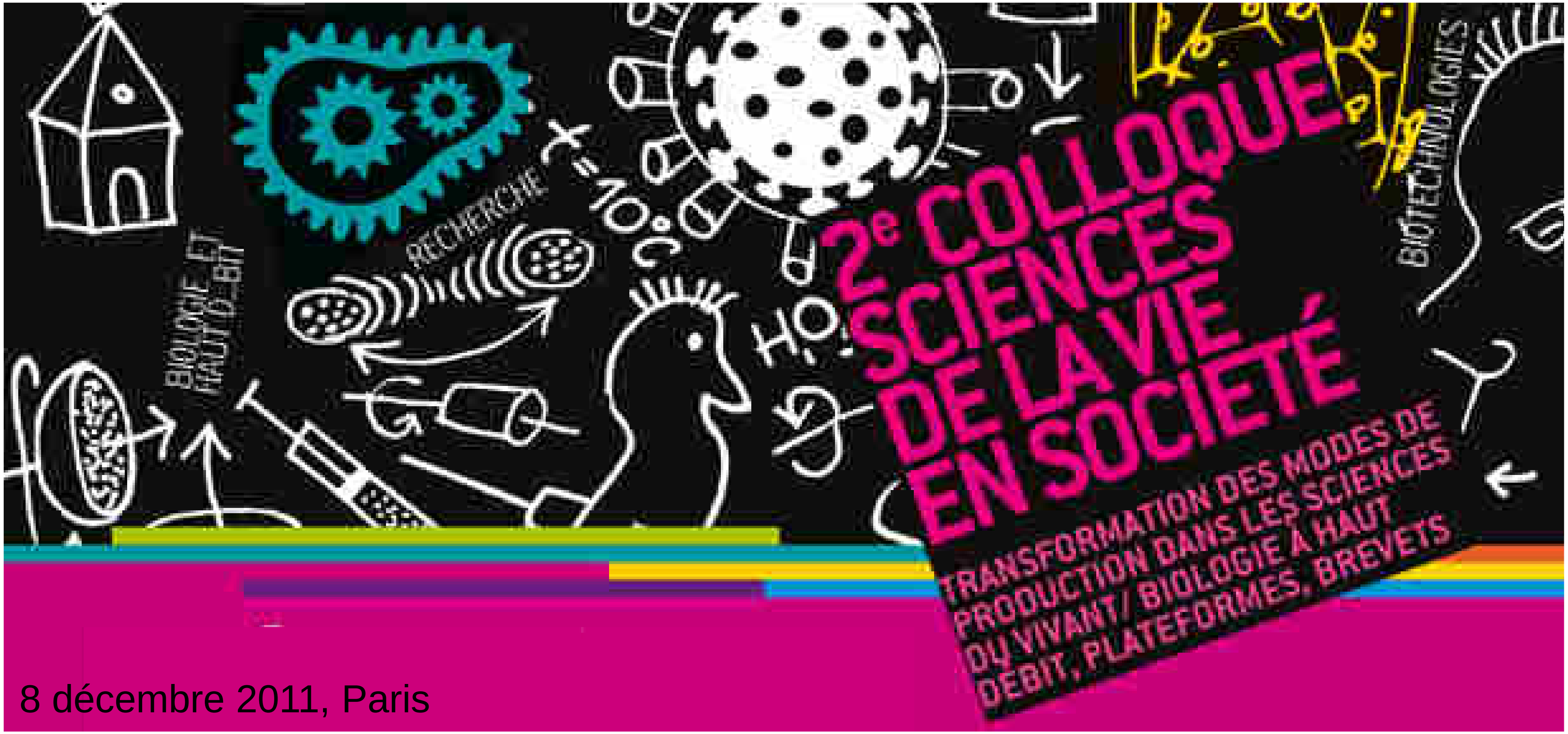




10ⁿ	Préfixe français	Symbole	Depuis note 1	Nombre décimal	Échelle courte note 2	Échelle longue note 3
10^{24}	yotta	Y	1991	1 000 000 000 000 000 000 000 000	Septillion	Quadrillion
10^{21}	zetta	Z	1991	1 000 000 000 000 000 000 000	Sextillion	Trilliard
10^{18}	exa	E	1975	1 000 000 000 000 000 000	Quintillion	Trillion
10^{15}	péta	P	1975	1 000 000 000 000 000	Quadrillion	Billiard
10^{12}	téra	T	1960	1 000 000 000 000	Trillion	Billion
10^{9}	giga	G	1960	1 000 000 000	Billion	Milliard
10^{6}	méga	M	1960	1 000 000	Million	

Data Center INRA Toulouse:

Mémoire vive 2 Teraoctets
400 Téraoctets de données
32 baies, 1500 serveurs


2e COLLOQUE
SCIENCES
DE LA VIE
EN SOCIÉTÉ
TRANSFORMATION DES MODES DE PRODUCTION DANS LES SCIENCES DU VIVANT/ BIOLOGIE À HAUT DÉBIT, PLATEFORMES, BREVETS
RECHERCHE
BIOLOGIE ET HAUT D_BIT
BIOTECHNOLOGIES
t=10°C
HO

8 décembre 2011, Paris

Biologie à haut débit et organisation de la recherche – une nouvelle économie des données ?

Les grands programmes de séquençage des génomes ont marqué l'entrée de la biologie dans le domaine de la « **big science** »

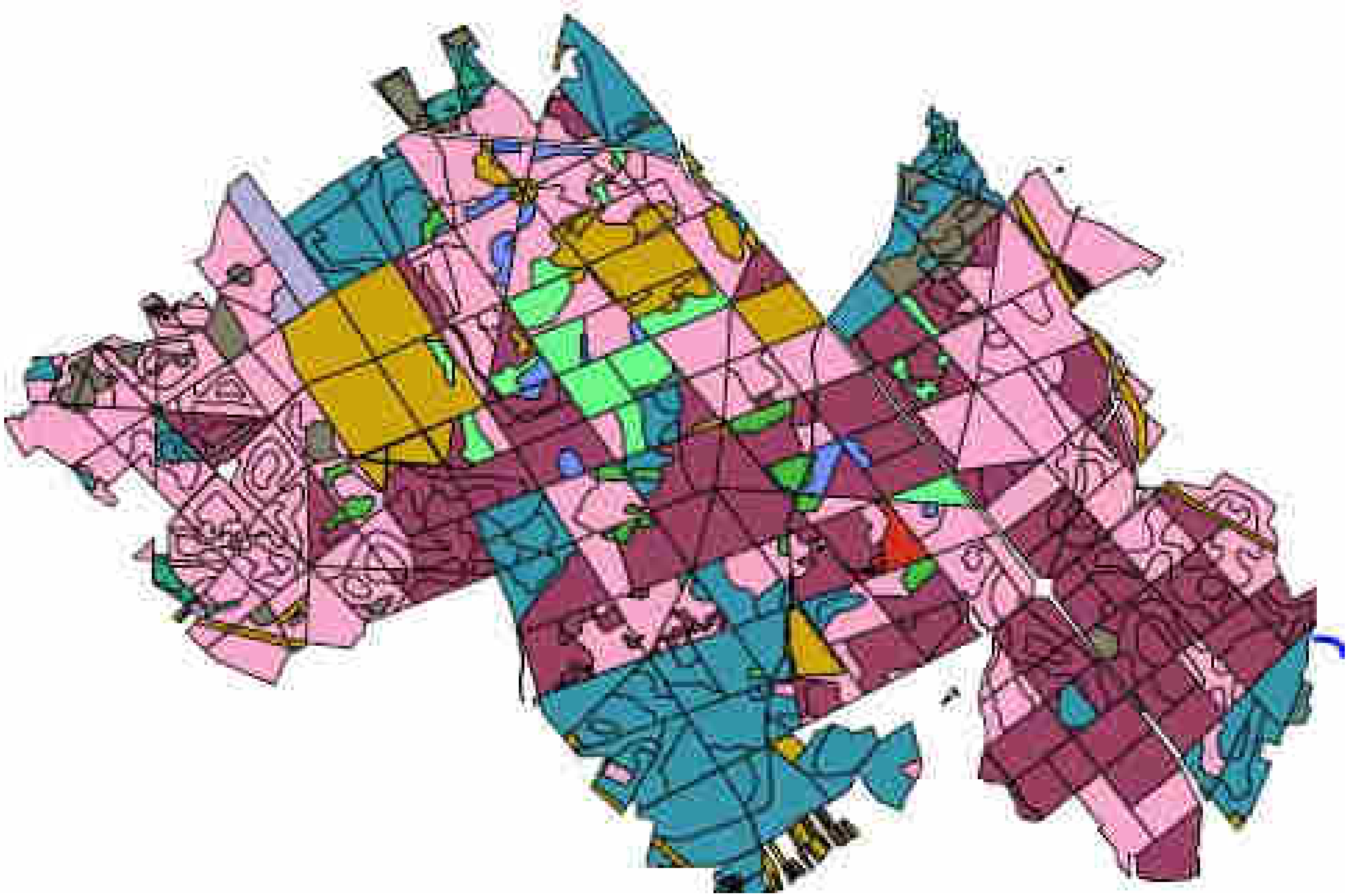
CERN

La révolution conceptuelle de la biologie à grande échelle

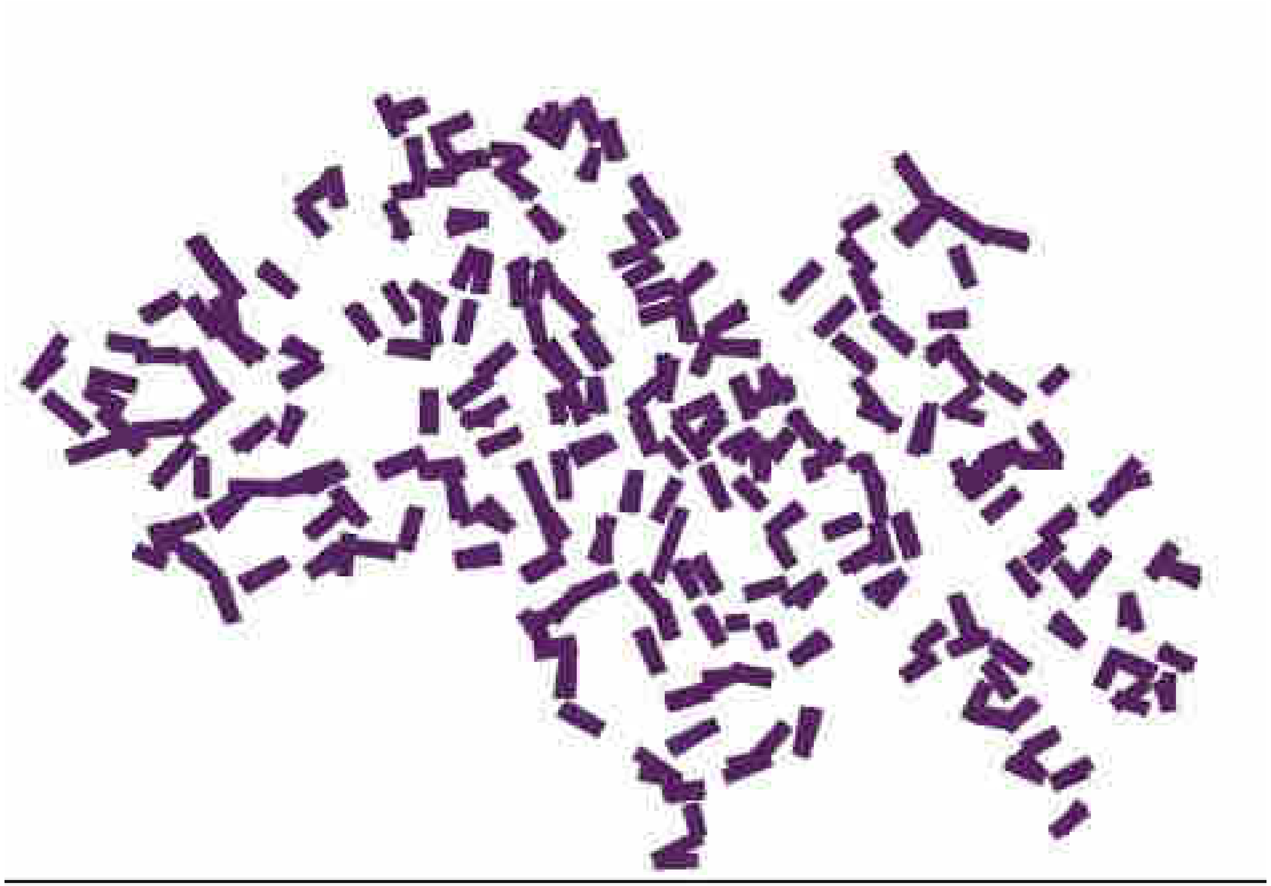
Ces grands instruments s’inscrivent pourtant dans un processus qui n’est pas simplement de doter la recherche de moyens financiers protégés. C’est aussi une manière de participer à une évolution qualitative très forte qui peut aboutir à des révolutions conceptuelles et médicales dont on imagine encore difficilement aujourd’hui les conséquences pour demain.

Metabarcoding for community studies

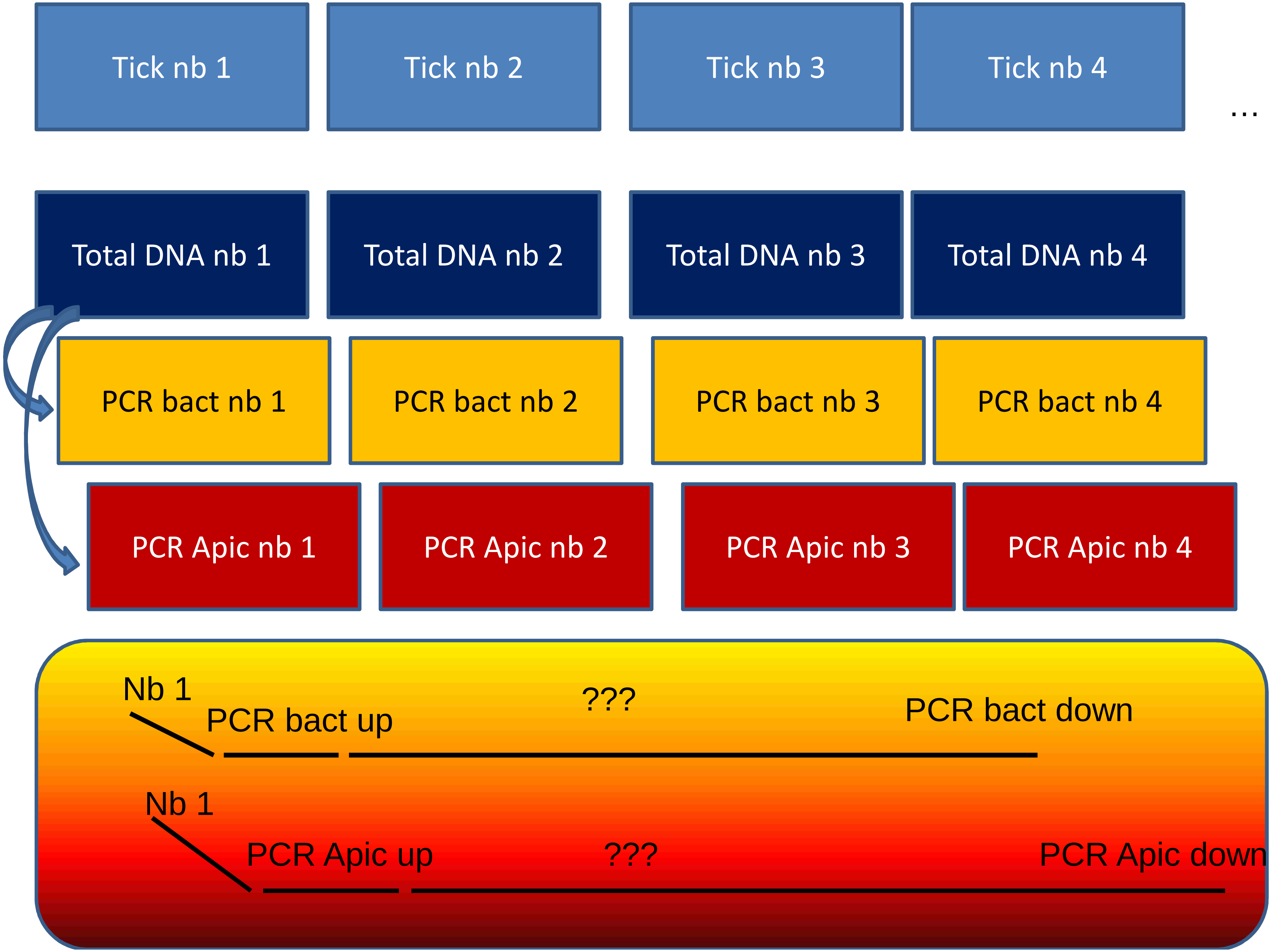
- Ixomic project : “NGS of *Ixodes ricinus* (PhD Elsa Quillery) and its **microbiome (Coll. BioEpAR-EpiA)”**
  - (INRA AIP Bioressources jan 2011 – jan 2013 seminar Paris feb 2013)
- Which pathogens are co-circulating in questing ticks of a suburban forest ?
- How diverse are the
  - Bacteria (EpiA)
  - Protozoans of the phylum Apicomplexa (BioEpAR)
- At one site ? vs. environmental factors ?



Senart forest (77) by dominating tree species in foresting sectors (IFN data)



20.000 (!) questing ticks collected by EpiA in Senart forest May 2011
Subset of
-190 adults
-190 nymphs
-190 larvae
At random among these ticks



Tick nb 1 Tick nb 2 Tick nb 3 Tick nb 4 …

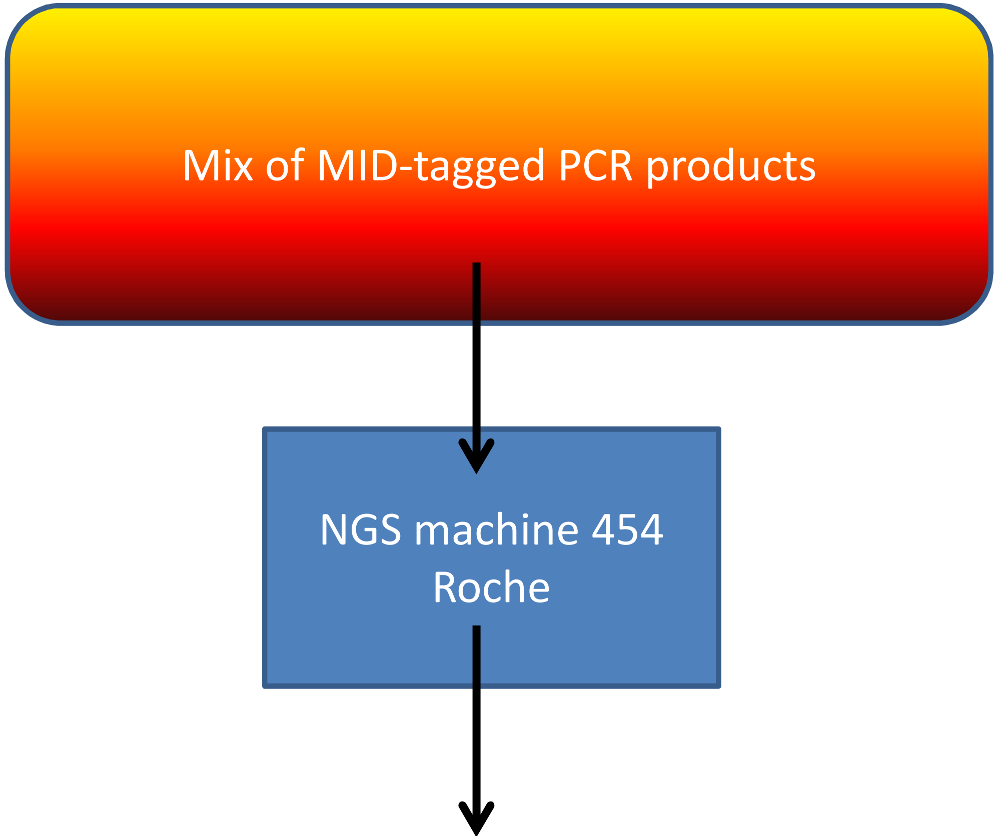
Total DNA nb 1 Total DNA nb 2 Total DNA nb 3 Total DNA nb 4

PCR bact nb 1 PCR bact nb 2 PCR bact nb 3 PCR bact nb 4

PCR Apic nb 1 PCR Apic nb 2 PCR Apic nb 3 PCR Apic nb 4

Nb 1 PCR bact up ??? PCR bact down

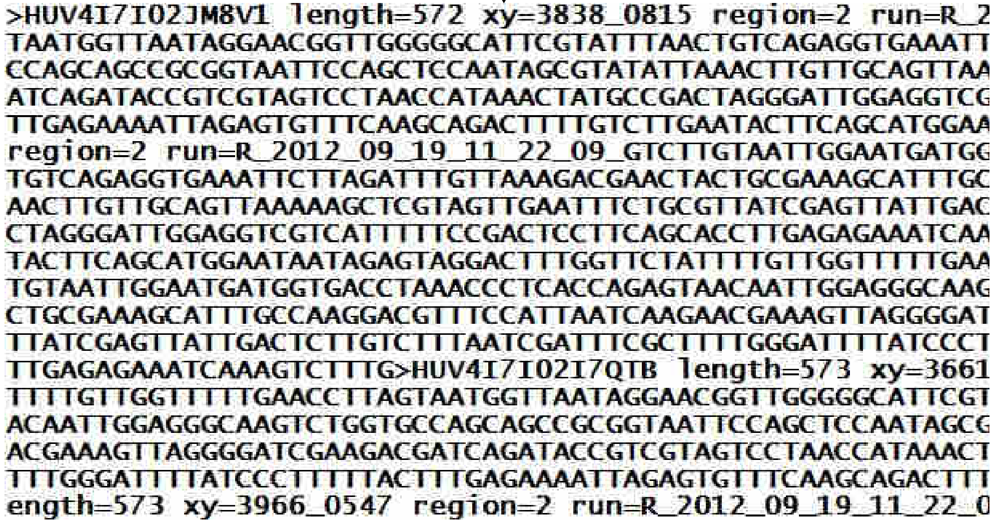
Nb 1 PCR Apic up ??? PCR Apic down



Mix of MID-tagged PCR products

NGS machine 454 Roche

8 Gb of data



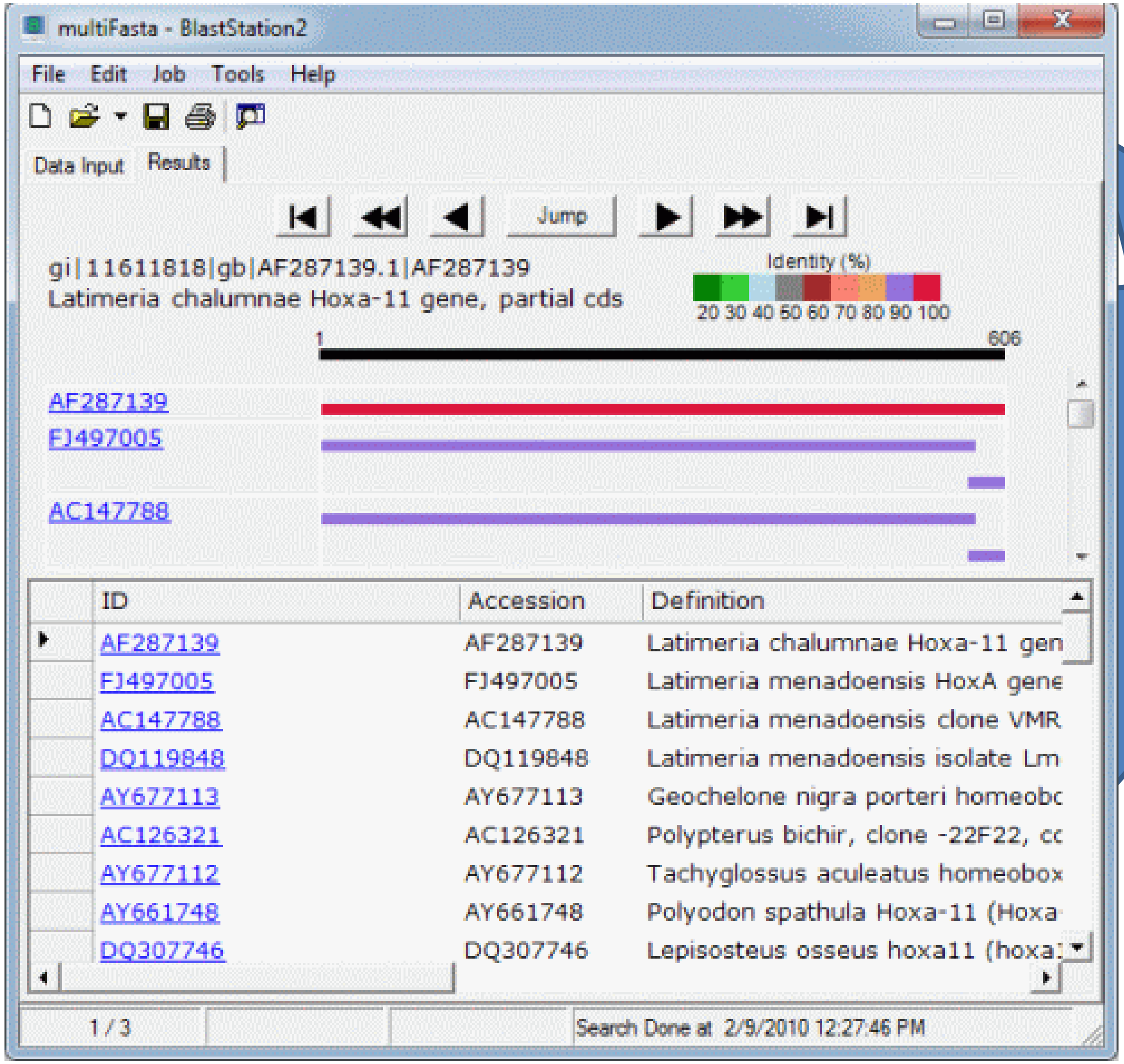
```
>HUV4I7I02JM8V1 length=572 xy=3838_0815 region=2 run=R_2
TAATGGTTAATAGGAACGGTTGGGGGCATTCGTATTTAACTGTCAGAGGTGAAATT
CCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAACTTGTTGCAGTTAA
ATCAGATACCGTCGTAGTCCTAACCATAAACTATGCCGACTAGGGATTGGAGGTCG
TTGAGAAAATTAGAGTGTTTCAAGCAGACTTTTGTCTTGAATACTTCAGCATGGAA
region=2 run=R_2012_09_19_11_22_09_GTCTTGTAATTGGAATGATGG
TGTCAGAGGTGAAATTCTTAGATTTGTTAAAGACGAACTACTGCGAAAGCATTTGC
AACTTGTTGCAGTTAAAAAGCTCGTAGTTGAATTTCTGCGTTATCGAGTTATTGAC
CTAGGGATTGGAGGTCGTCATTTTTCCGACTCCTTCAGCACCTTGAGAGAAATCAA
TACTTCAGCATGGAATAATAGAGTAGGACTTTGGTTCTATTTTGTTGGTTTTTGAA
TGTAATTGGAATGATGGTGACCTAAACCCTCACCAGAGTAACAATTGGAGGGCAAG
CTGCGAAAGCATTTGCCAAGGACGTTTCCATTAATCAAGAACGAAAGTTAGGGGAT
TTATCGAGTTATTGACTCTTGTCTTTAATCGATTTCGCTTTTGGGATTTTATCCCT
TTGAGAGAAATCAAAGTCTTTG>HUV4I7I02I7QTB length=573 xy=3661
TTTTGTTGGTTTTTGAACCTTAGTAATGGTTAATAGGAACGGTTGGGGGCATTCGT
ACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCG
ACGAAAGTTAGGGGATCGAAGACGATCAGATACCGTCGTAGTCCTAACCATAAACT
TTTGGGATTTTATCCCTTTTTACTTTGAGAAAATTAGAGTGTTTCAAGCAGACTTT
ength=573 xy=3966_0547 region=2 run=R_2012_09_19_11_22_0
```

Processing 8 Gb of data

- EpiA XB (bacteria)
- Sort
  - By length (~400 bp)
  - By quality
  - By PCR primers
- Bioinformatics e.g. Grep function “select lines with word ATTGTATC”

- BioEpAR SB (Piroplasmids)
- Sort
  - By length (~560 bp)
  - By quality
  - By PCR primers
- Galaxy platform user-friendly interface “Select” = grep “select lines with word TTATCGTATCA”

Metabarcoding = assign a barcode sequence to a species



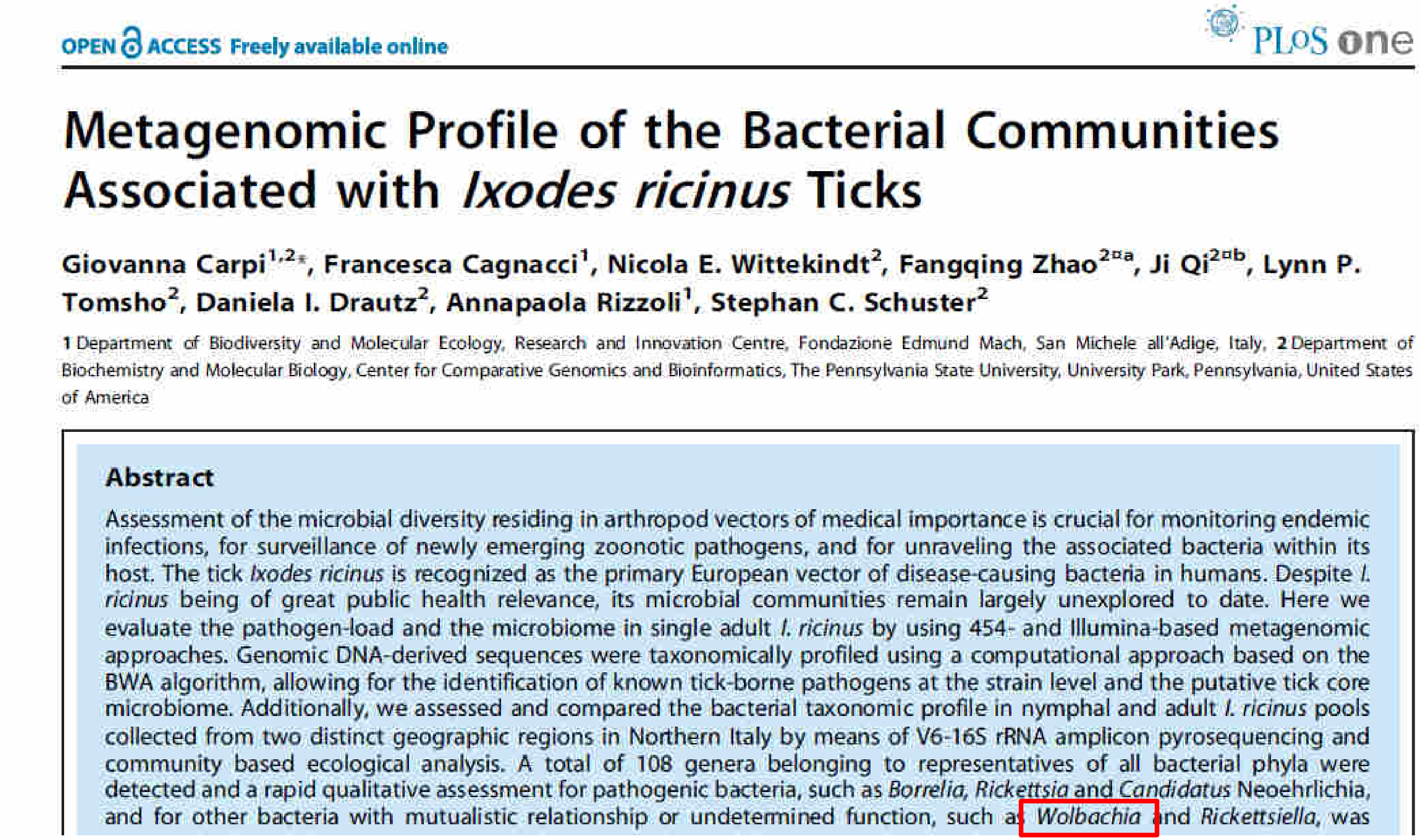
gi|11611818|gb|AF287139.1|AF287139
Latimeria chalumnae Hoxa-11 gene, partial cds

ID	Accession	Definition
AF287139	AF287139	Latimeria chalumnae Hoxa-11 gen
FJ497005	FJ497005	Latimeria menadoensis HoxA gene
AC147788	AC147788	Latimeria menadoensis clone VMR
DQ119848	DQ119848	Latimeria menadoensis isolate Lm
AY677113	AY677113	Geochelone nigra porteri homeobc
AC126321	AC126321	Polypterus bichir, clone -22F22, cc
AY677112	AY677112	Tachyglossus aculeatus homeobox
AY661748	AY661748	Polyodon spathula Hoxa-11 (Hoxa
DQ307746	DQ307746	Lepisosteus osseus hoxa11 (hoxa1

1 / 3 Search Done at 2/9/2010 12:27:46 PM

Basic Local Alignment Search Tool

BLAST



OPEN ACCESS Freely available online

PLoS one

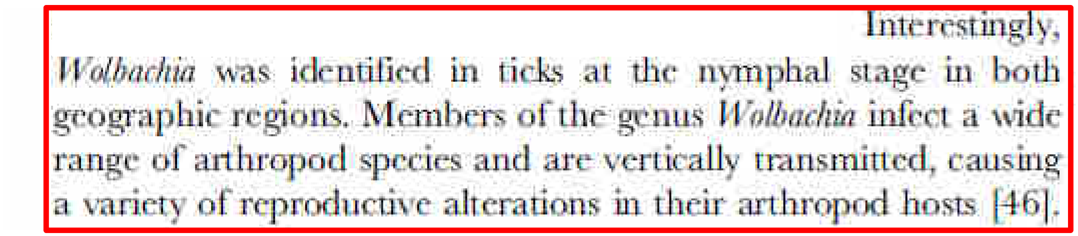
Metagenomic Profile of the Bacterial Communities Associated with *Ixodes ricinus* Ticks

Giovanna Carpi[1,2]*, Francesca Cagnacci[1], Nicola E. Wittekindt[2], Fangqing Zhao[2¤a], Ji Qi[2¤b], Lynn P. Tomsho[2], Daniela I. Drautz[2], Annapaola Rizzoli[1], Stephan C. Schuster[2]

1 Department of Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy, 2 Department of Biochemistry and Molecular Biology, Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania, United States of America

Abstract

Assessment of the microbial diversity residing in arthropod vectors of medical importance is crucial for monitoring endemic infections, for surveillance of newly emerging zoonotic pathogens, and for unraveling the associated bacteria within its host. The tick *Ixodes ricinus* is recognized as the primary European vector of disease-causing bacteria in humans. Despite *I. ricinus* being of great public health relevance, its microbial communities remain largely unexplored to date. Here we evaluate the pathogen-load and the microbiome in single adult *I. ricinus* by using 454- and Illumina-based metagenomic approaches. Genomic DNA-derived sequences were taxonomically profiled using a computational approach based on the BWA algorithm, allowing for the identification of known tick-borne pathogens at the strain level and the putative tick core microbiome. Additionally, we assessed and compared the bacterial taxonomic profile in nymphal and adult *I. ricinus* pools collected from two distinct geographic regions in Northern Italy by means of V6-16S rRNA amplicon pyrosequencing and community based ecological analysis. A total of 108 genera belonging to representatives of all bacterial phyla were detected and a rapid qualitative assessment for pathogenic bacteria, such as *Borrelia, Rickettsia* and *Candidatus* Neoehrlichia, and for other bacteria with mutualistic relationship or undetermined function, such as *Wolbachia* and *Rickettsiella*, was



Interestingly, *Wolbachia* was identified in ticks at the nymphal stage in both geographic regions. Members of the genus *Wolbachia* infect a wide range of arthropod species and are vertically transmitted, causing a variety of reproductive alterations in their arthropod hosts [46].

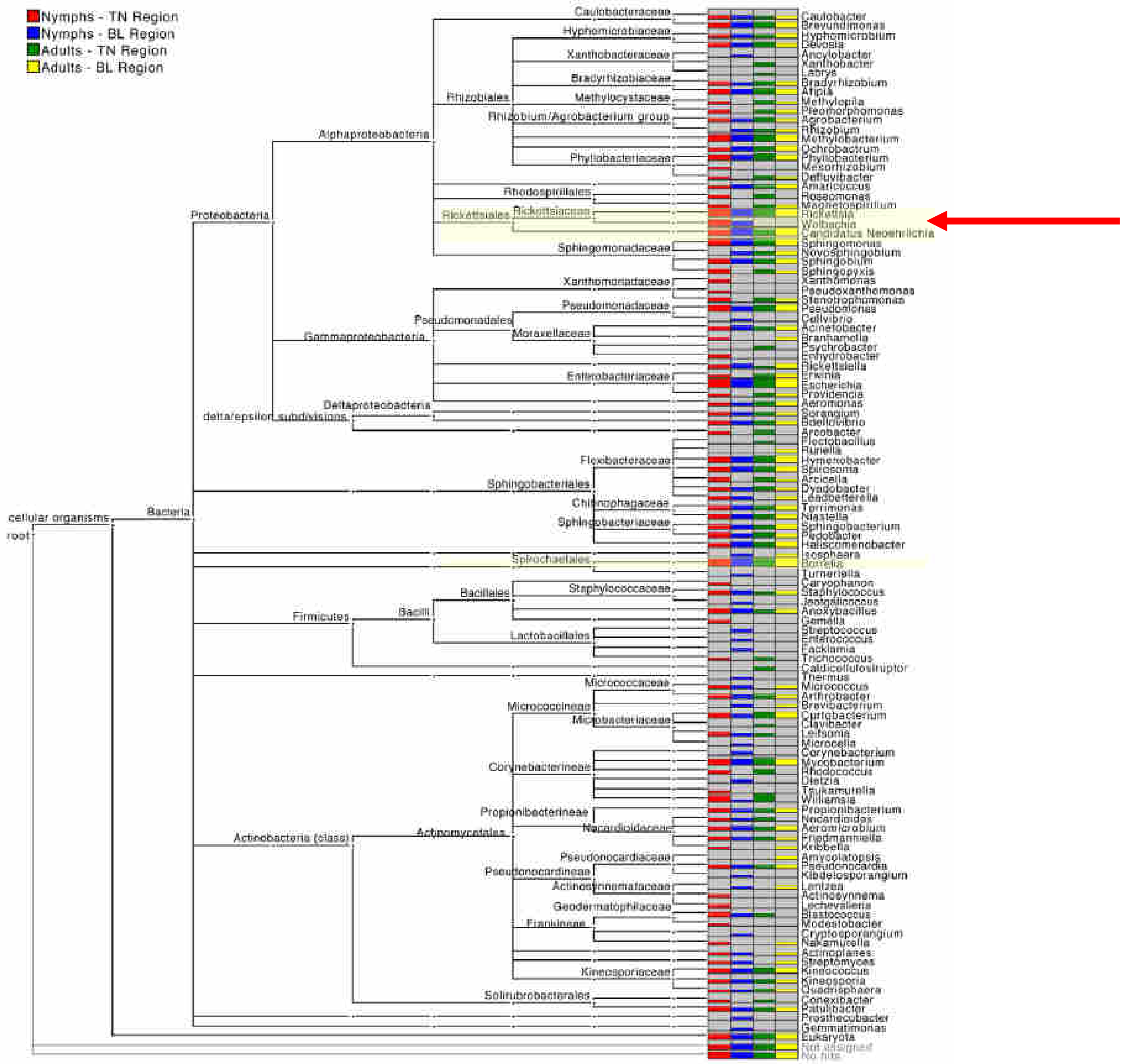
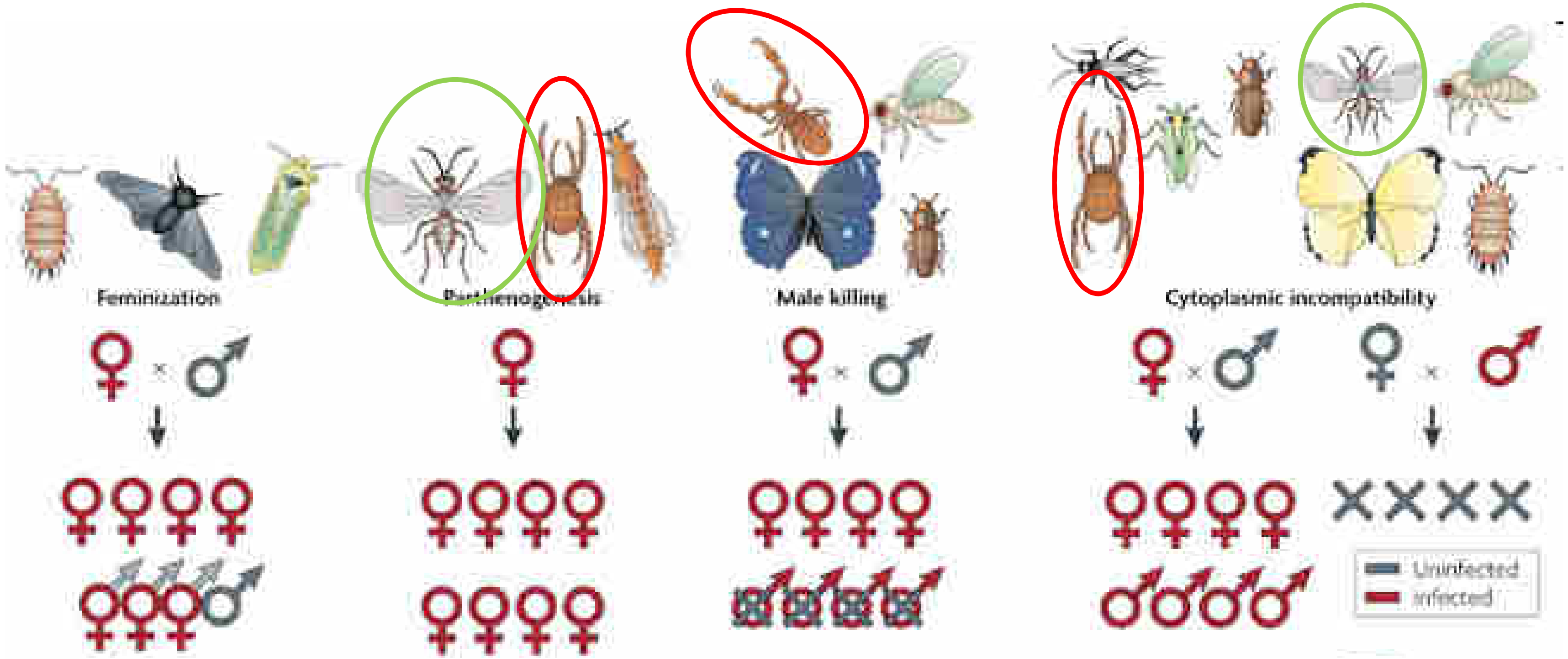


Figure 3. MEGAN comparison of bacterial taxonomic profiles of four tick pools (reflecting different life stages and geographic provenience) based on the V6 amplicon 16S rRNA gene sequence reads analyzed against the Ribosomal Database. The height of the bars corresponds to the number of hits for each genus. Highlighted in light yellow are bacteria genera recognized as tick-borne pathogens or tick endosymbionts.

Conséquences de *Wolbachia* sur son hôte

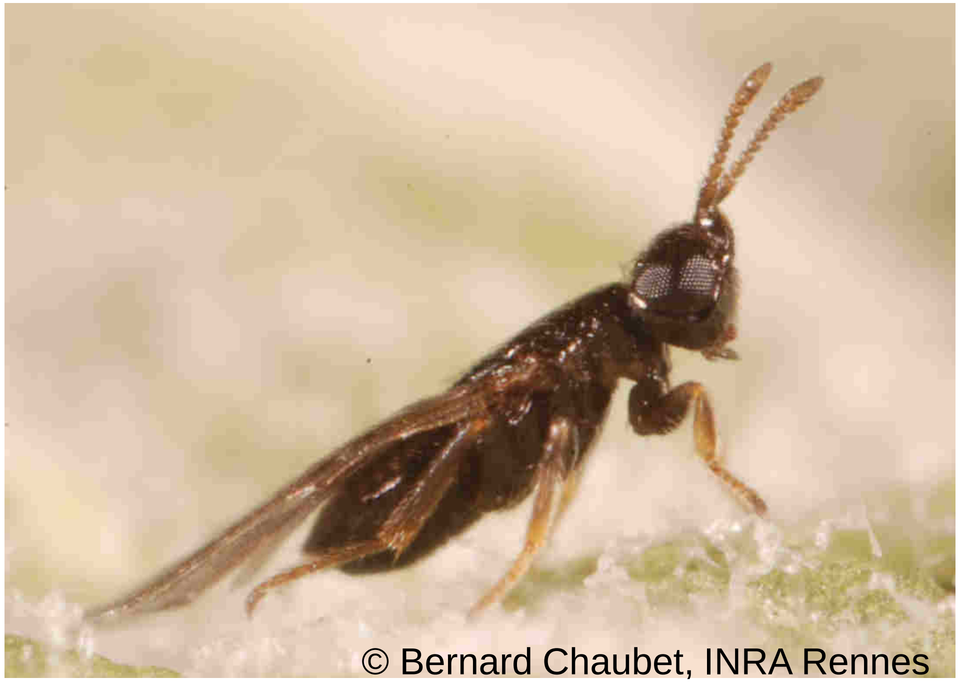
Quelles conséquences chez *I. ricinus* ?



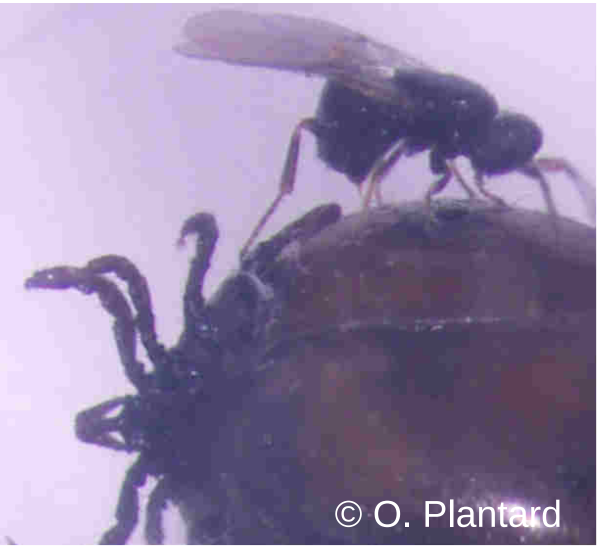
Les tiques hébergent de nombreux micro-organismes…
mais aussi des animaux de plus grosse taille :

¤ des nématodes

¤ des insectes parasitoïdes



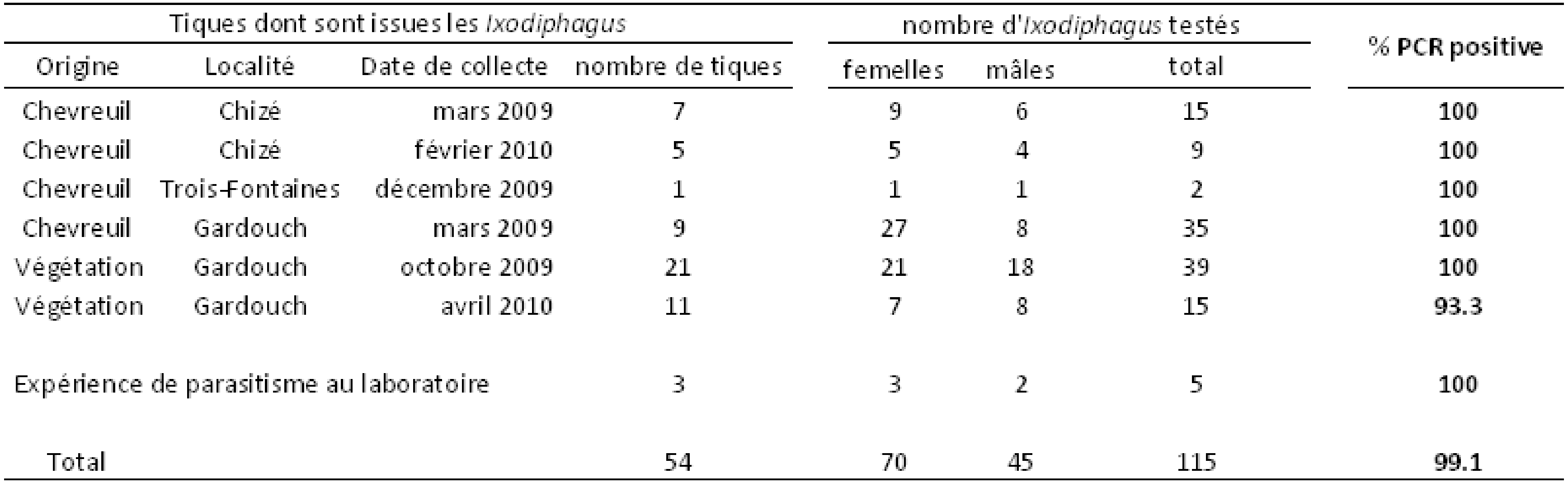
© Bernard Chaubet, INRA Rennes



© O. Plantard

Recherche de *Wolbachia* par PCR dans des *Ixodiphagus*

Amorces PCR définies dans le gène *Wsp* (excluant l'amplification d'*Anaplasma* ou d'*Ehrlichia*).



Tiques dont sont issues les *Ixodiphagus*				nombre d'*Ixodiphagus* testés			% PCR positive
Origine	Localité	Date de collecte	nombre de tiques	femelles	mâles	total	
Chevreuil	Chizé	mars 2009	7	9	6	15	100
Chevreuil	Chizé	février 2010	5	5	4	9	100
Chevreuil	Trois-Fontaines	décembre 2009	1	1	1	2	100
Chevreuil	Gardouch	mars 2009	9	27	8	35	100
Végétation	Gardouch	octobre 2009	21	21	18	39	100
Végétation	Gardouch	avril 2010	11	7	8	15	93.3
Expérience de parasitisme au laboratoire			3	3	2	5	100
Total			54	70	45	115	99.1

¤ La quasi-totalité des *Ixodiphagus hookeri* portent des *Wolbachia*

¤ Il y a bien transmission verticale de *Wolbachia* chez *Ixodiphagus hookeri* (les œufs contiennent la bactérie)

¤ Séquence du gène *Wsp* = 100% d'identité (500 pb) avec une séquence de *Wolbachia* amplifié chez d'autres insectes (dont des parasitoïdes chalcidiens)

➔La présence de *Wolbachia* dans des tiques est lié au parasitisme par *Ixodiphagus hookeri*

Plantard et al. PLOS One 2012

Processing 8 Gb of data

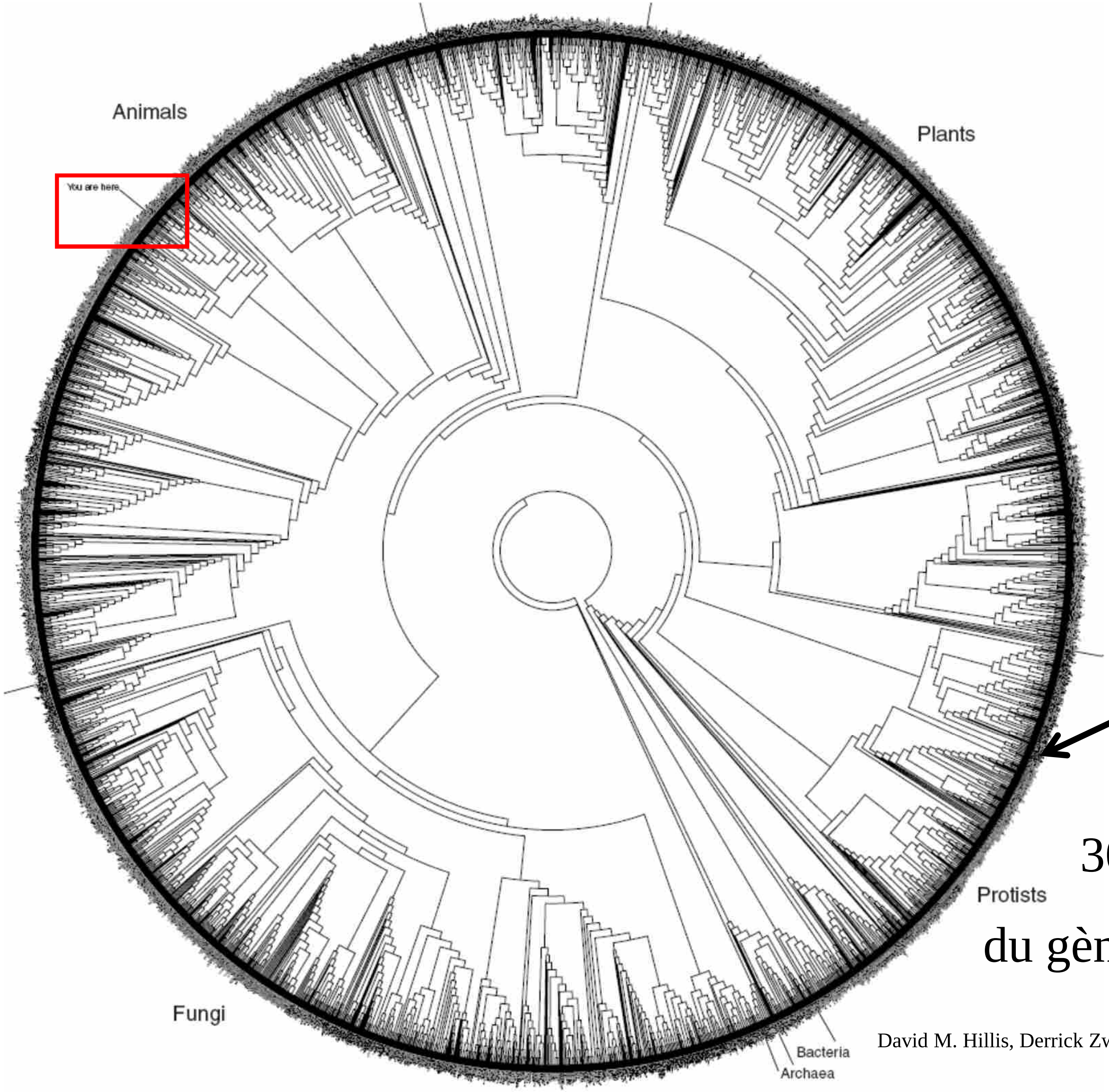
- EpiA XB (bacteria)
- Sort
  - By length (~400 bp)
  - By quality
  - By PCR primers
- Bioinformatics e.g. Grep function “select lines with word ATTGTATC”

- BioEpAR SB (Piroplasmids)
- Sort
  - By length (~560 bp)
  - By quality
  - By PCR primers
- Galaxy platform user-friendly interface “Select” = grep “select lines with word TTATCGTATCA”

What about the Piroplasmids (Apic – sequences) ?

- Information from End-point PCR studies :

  expect at least 4 species of Babesia, possibly more



Apicomplexes

3000 séquences

du gène ribosomique 18S

David M. Hillis, Derrick Zwickl, and Robin Gutell, University of Texas.

Data cleanup

- Ask a colleague (Cl. RISPE) expert in molecular evolution

  Modified BLAST algorithm to eliminate single errors
- Use a one-for-all OTU cleanup software
- Or…

- Try again with a different technology !

Background:
Mortality and cancer incidence in NZ Meat Workers

McLean *et al.* OEM 2004

- **Significant excess mortality from lung cancer**
- **Effect related to exposure to biological material contained in animal urine, faeces and blood**
- **Effect related to employment duration in selected biological exposure categories**

Aims of this study:

Multidisciplinary approach to identify potential causes of the increased cancer risk in meat-workers

Environmental monitoring to assess exposure to:

- Protein levels as a proxy for chronic antigenic stimulation.
- Urine, blood and faecal markers.
- **Specific pathogens with known carcinogenic properties in meat workers**
- The mutagenicity of whole bioaerosols *in vitro*.
- **Bacterial and viral pathogens using next-generation sequencing**

Aims of this study:

Biological monitoring to assess :

- **Serum antibody titres against specific pathogens** as a long term measure of exposure.
- The presence of specific pathogens in the airways as a **biomarker of exposure** in one of the target organs

Epidemiological methods to determine :

- Average exposure levels and variation between exposure groups to develop reliable exposure models for the agents measured
- To update and reanalyse the existing New Zealand meat workers cohort using these refined exposures estimates/informations

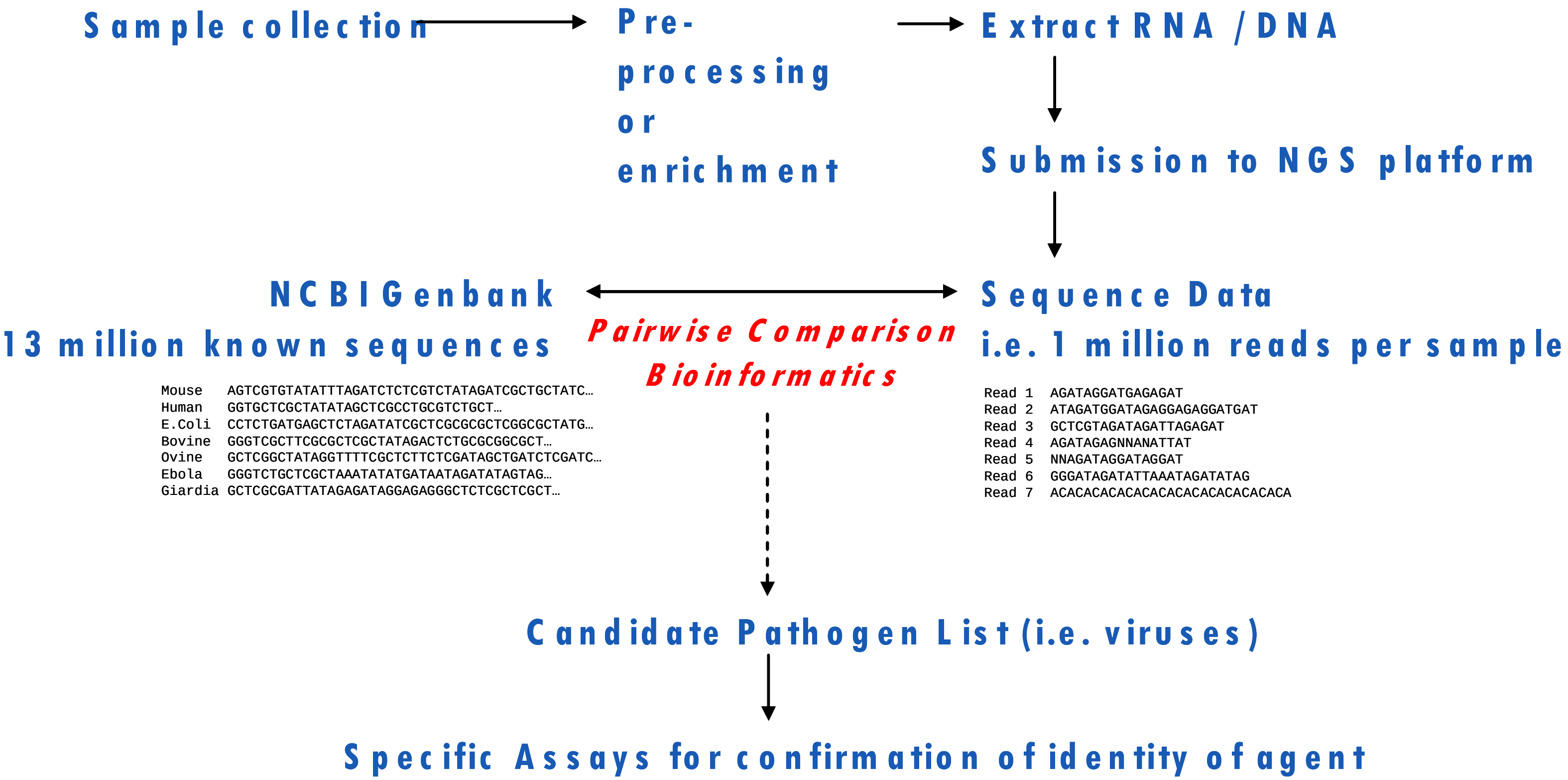
Experimental methods to confirm the biological reality of our results

BROHNZ Building Research in Occupational Health in New Zealand

Methods - Results:

Bioaerosol samples : Pathogen discovery

- Personal air samples
- Bulk air samples = environmental samples
- using deep-sequencing (NGS)



Sample collection

Pre-processing or enrichment

Extract RNA / DNA

Submission to NGS platform

Sequence Data
i.e. 1 million reads per sample

```
Read 1  AGATAGGATGAGAGAT
Read 2  ATAGATGGATAGAGGAGAGGATGAT
Read 3  GCTCGTAGATAGATTAGAGAT
Read 4  AGATAGAGNNANATTAT
Read 5  NNAGATAGGATAGGAT
Read 6  GGGATAGATATTAAATAGATATAG
Read 7  ACACACACACACACACACACACACACA
```

NCBI Genbank
13 million known sequences

```
Mouse   AGTCGTGTATATTTAGATCTCTCGTCTATAGATCGCTGCTATC…
Human   GGTGCTCGCTATATAGCTCGCCTGCGTCTGCT…
E.Coli  CCTCTGATGAGCTCTAGATATCGCTCGCGCGCTCGGCGCTATG…
Bovine  GGGTCGCTTCGCGCTCGCTATAGACTCTGCGCGGCGCT…
Ovine   GCTCGGCTATAGGTTTTCGCTCTTCTCGATAGCTGATCTCGATC…
Ebola   GGGTCTGCTCGCTAAATATATGATAATAGATATAGTAG…
Giardia GCTCGCGATTATAGAGATAGGAGAGGGCTCTCGCTCGCT…
```

Pairwise Comparison
Bioinformatics

Candidate Pathogen List (i.e. viruses)

Specific Assays for confirmation of identity of agent

Results:

Bioaerosol samples : Pathogen discovery

NGS First result : 454FLX on personal air sample

- **Technical problem on the NGS platform for analysis**
  - quantity, data lost, time to receive the first results…
- **Results : too many data – quality?**

Results:

Bioaerosol samples : Pathogen discovery

Hiseq on bovine pool of personal air samples

- Tow low quantity : pool of personal air samples
- **Identification of human papillomavirus**
- Validation of the extraction method
- **Huge amount of data, complexity (whole genome)**

Miseq on bulk air samples being analysed

- Validation of the extraction method (↗ quality + quantity)
- **Identification of bovine papillomavirus and coronavirus**
- **Identification of porcine adenovirus ?**
- **Huge amount of data : trouble for the bacteria analysis**

Results:

Bioaerosol samples : Pathogen discovery

Metagenomic approach

To reduce the amount of data and

To simplify the complexity of the data

↓

Metabarcoding approach

16s DNA analysis - Miseq on bulk air samples

- Analysis of the bacteria diversity
- Analysis in relation with different work task or environment

Actual projects or Future projects using NGS

Project RESPICARE (S. Assié) – Collab. G. Meyer et J.L. Guérin

- Antimicrobials and infectious respiratory diseases : integrated actions for drug reduction
- WP1 : Broad detection and study of the evolution of respiratory infectious agents

Thesis A. Rieux (C. Chartier) – Collab. Anses Niort

- Cryptosporidium (Molecular characterization ?) - Sanger
- 18S rRNA amplification + séquençage

Study of pig's microbiota by NGS (M. Leblanc-Maridor C. Belloc)

Diversity of the intestinal flora

Variations along a production cycle

Influences : pathogens? *Campylobacter*? *Salmonella*?

Collaborations envisagées (Anses Ploufragan, IFIP, Institut Pasteur…)

Take home messages

- Metagenomics can be used for metabarcoding studies
  - Markers at species-level
  - Need for reference sequences (Barcode of Life project) Systematics and Taxonomy
  - Assemblage studies on microbial community profiles
  - Quantitative approach possible (relative abundances)