Transcriptomics
&
Applied Genomics (TAG)

David Hot

Anca Lucau
Ségolène Caboche
Ludovic Huot
Stéphanie Slupek
Yves Lemoine

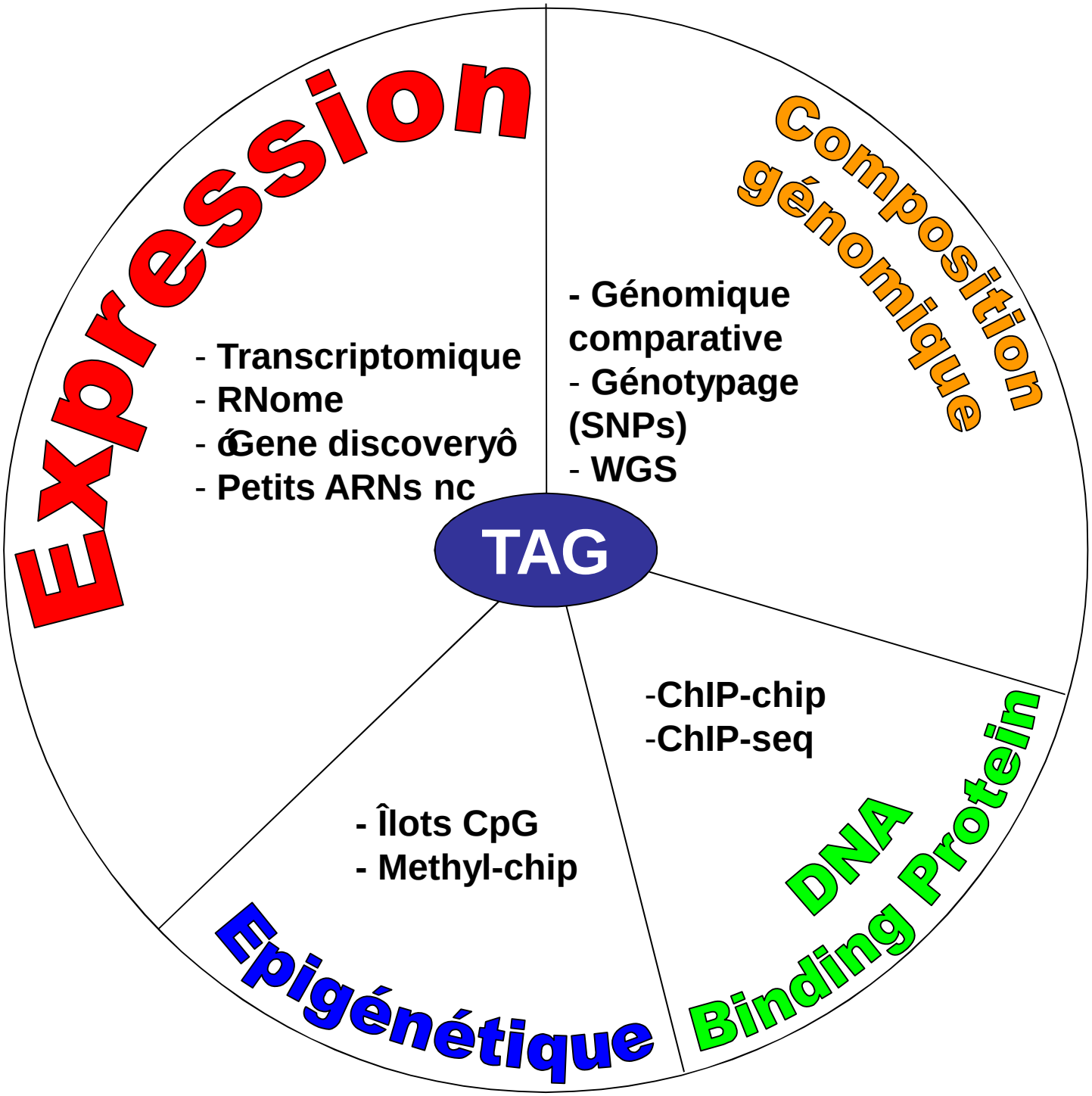
Léa Siegwald
Alexandre D'Halluin

Christophe Audebert
Gaël Even
Sophie Merlin
Alexandre Loywick

Institut Pasteur de Lille

CIIL

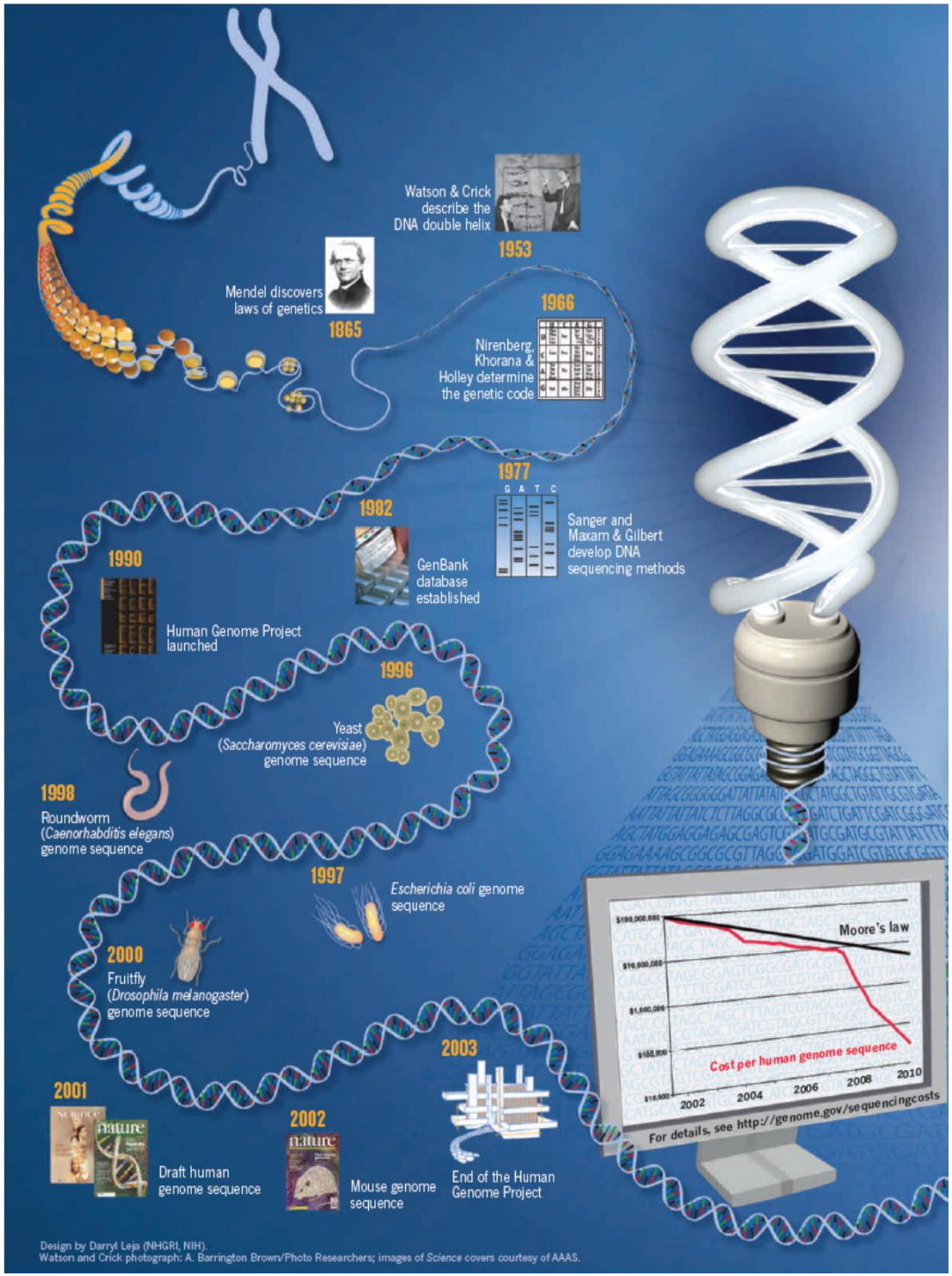
Pegase Biosciences



Institut Pasteur de Lille

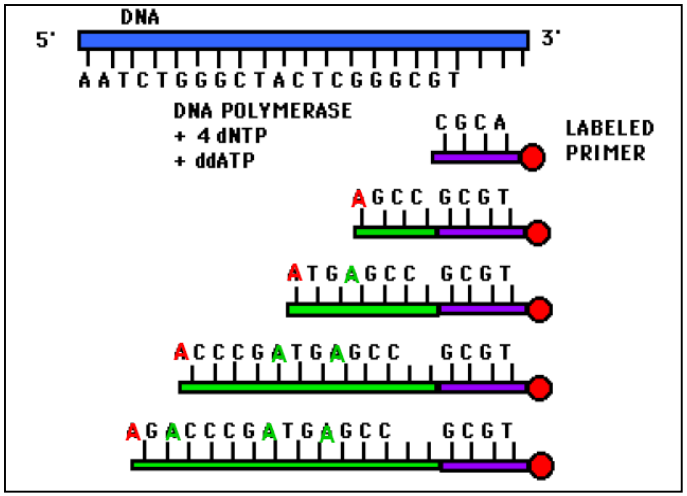
Fondation reconnue dôutilité publique

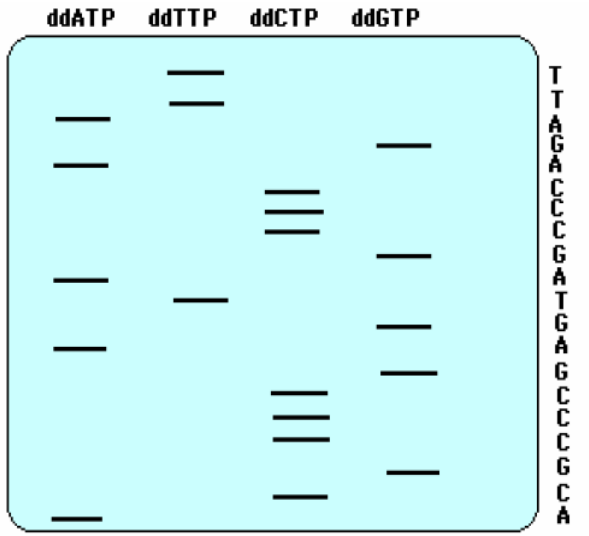
Institut Pasteur de Lille



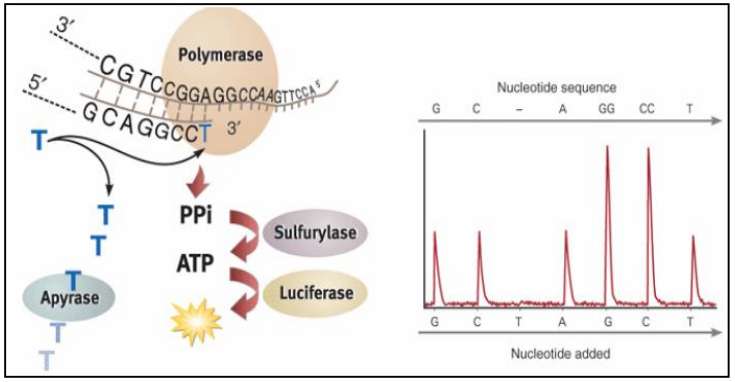
Nature 470, 204ï 213
(10 February 2011)

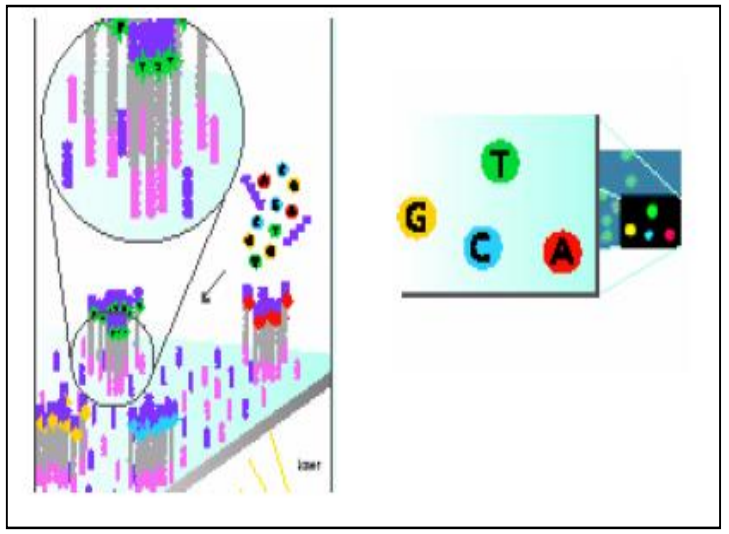
SANGER





High speed sequencing
High through-put sequencing
Next generation sequencing
Massively parallel sequencing
Séquençage parallélisé





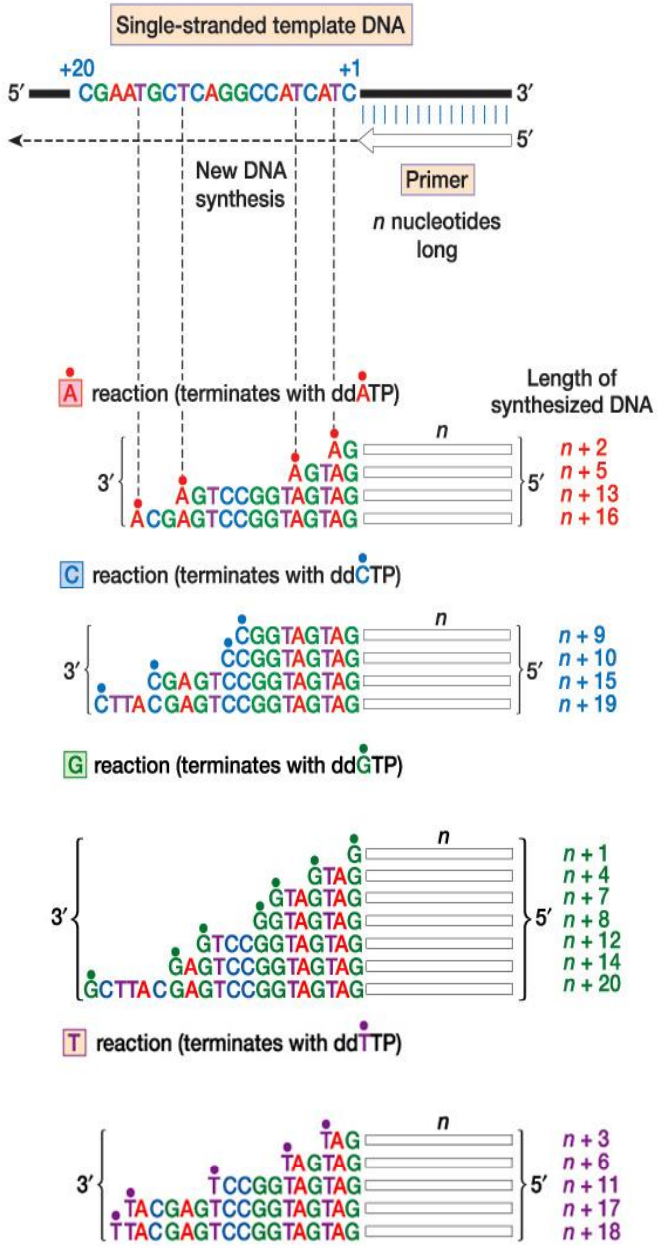


Figure 7-2 part 2 of 2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

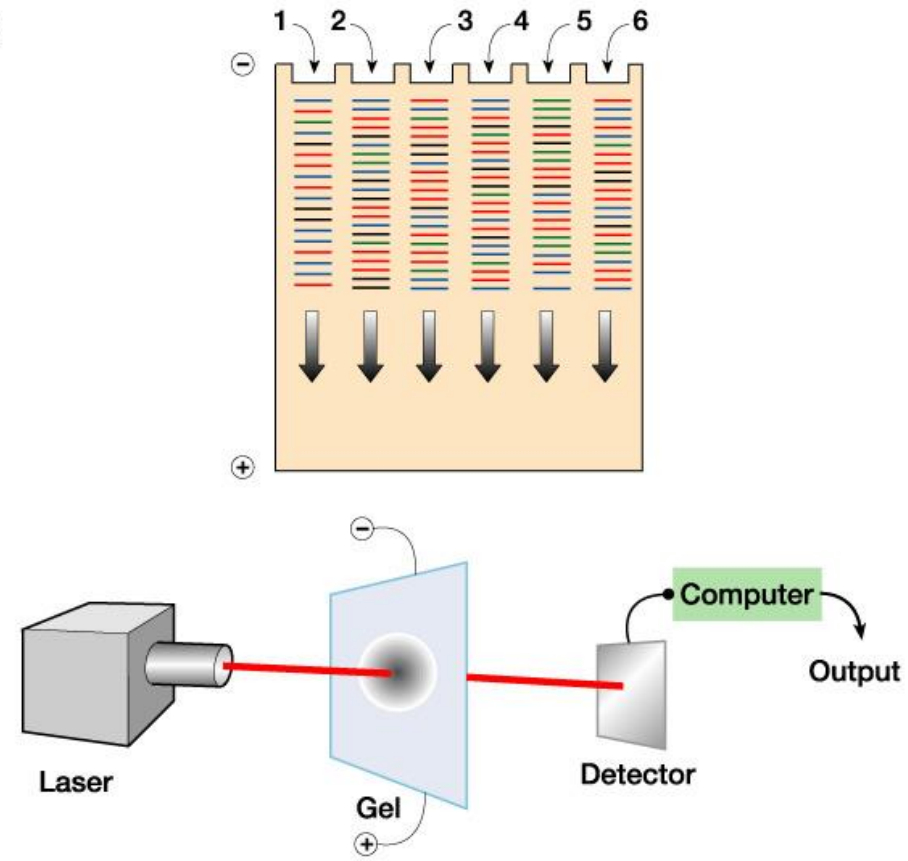
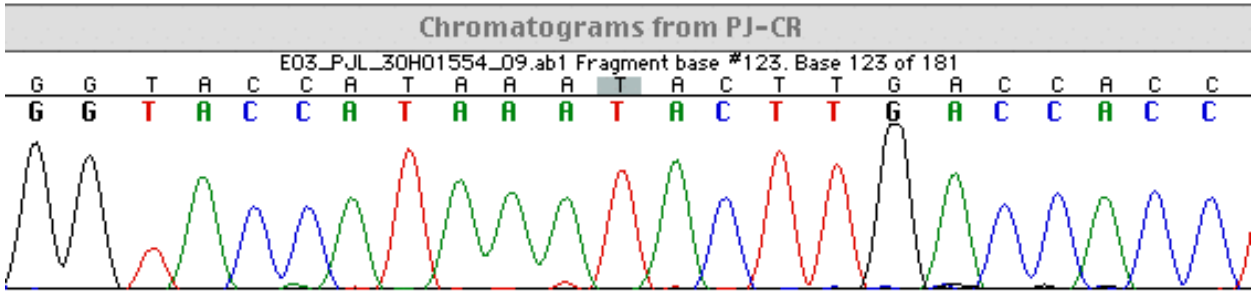
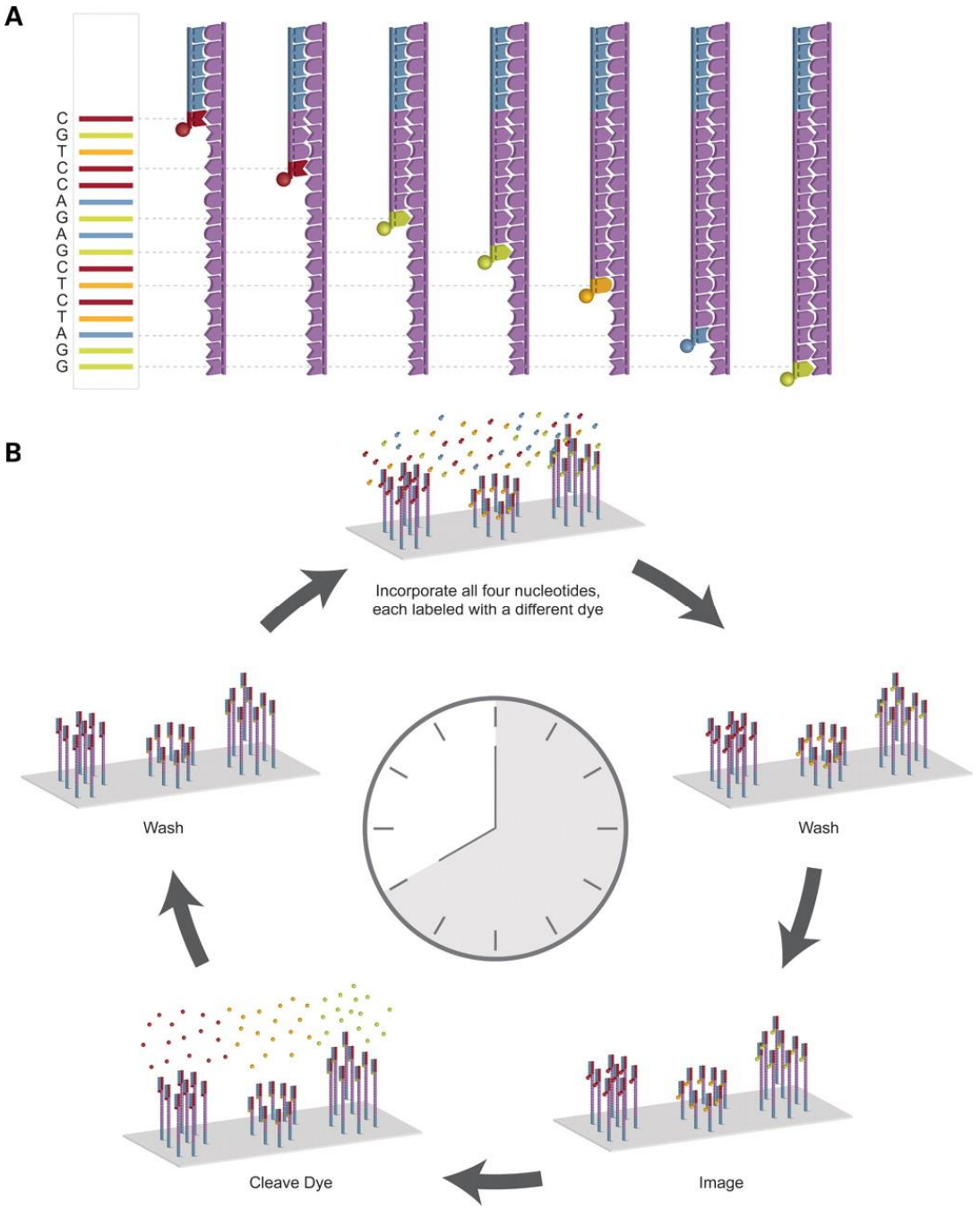


Figure 7-3 part 1 of 2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

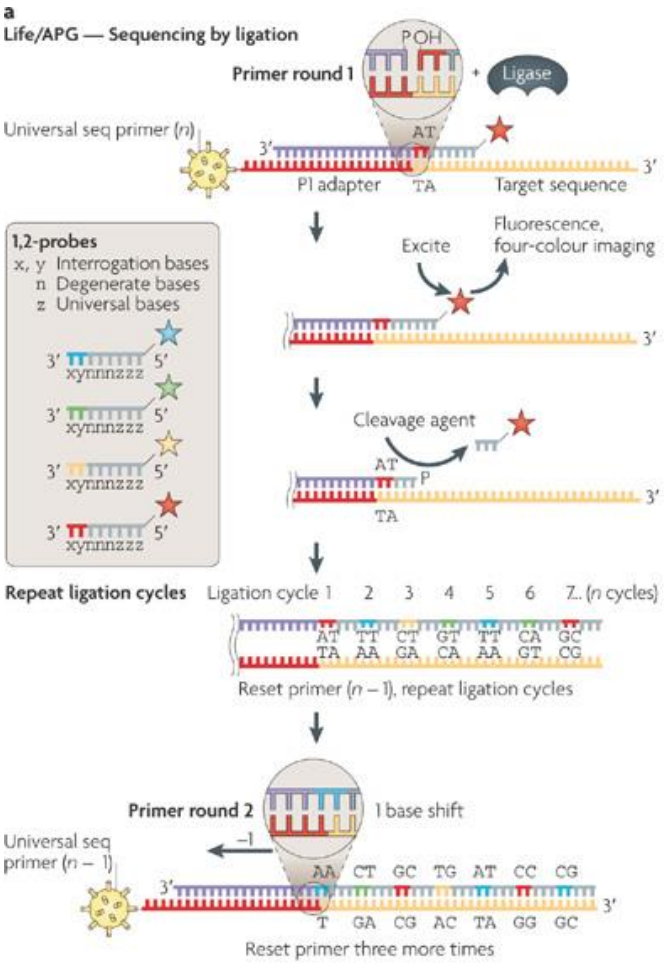


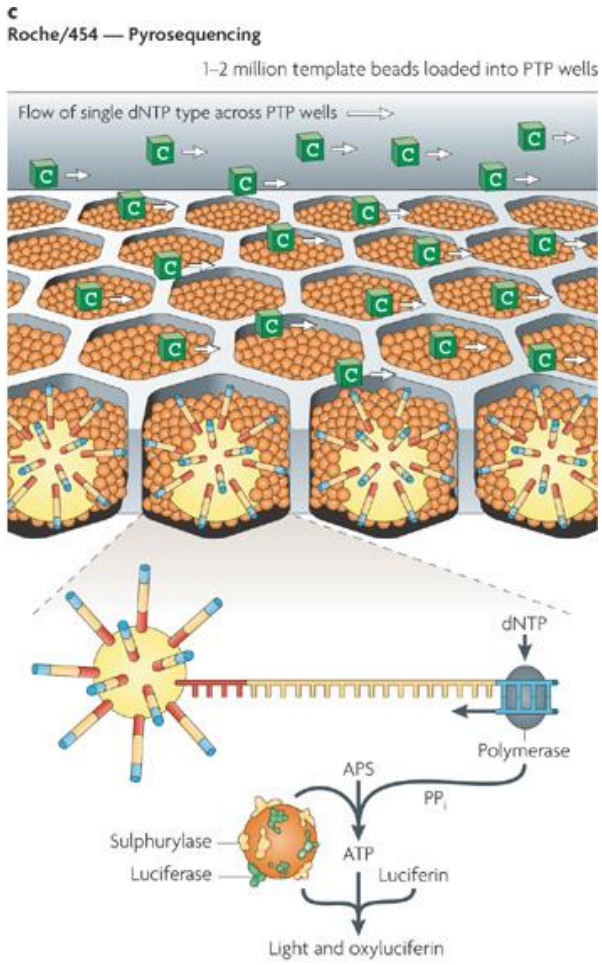


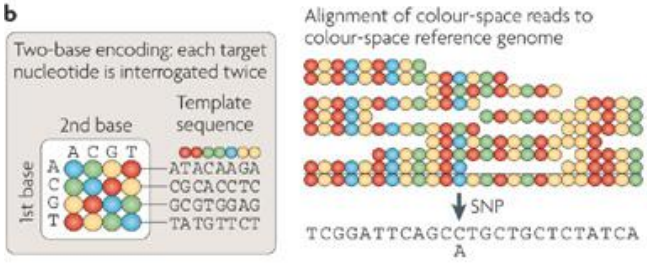
Fondation reconnue d'utilité publique

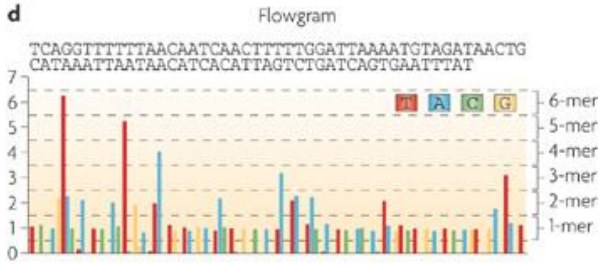
Schadt E E et al. Hum. Mol. Genet. 2010;19:R227-R240

Institut Pasteur de Lille





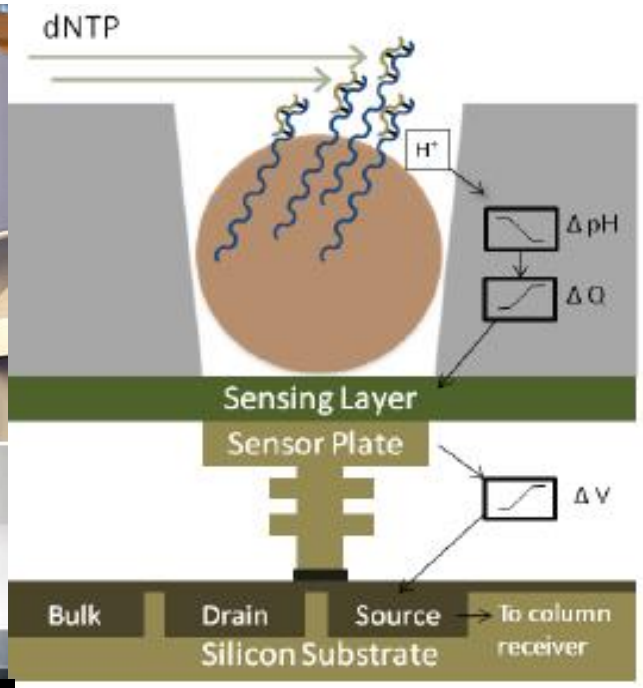




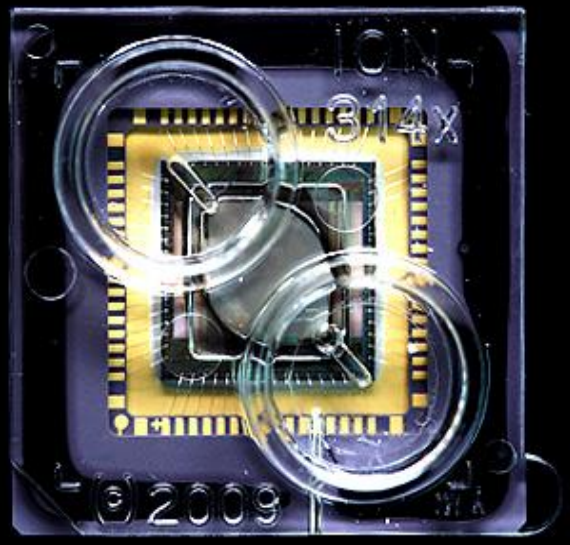
Nature Reviews | Genetics

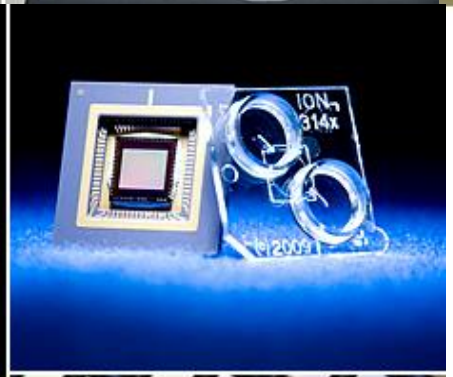
Personal Genome Machine® (PGM) Ion Torrent

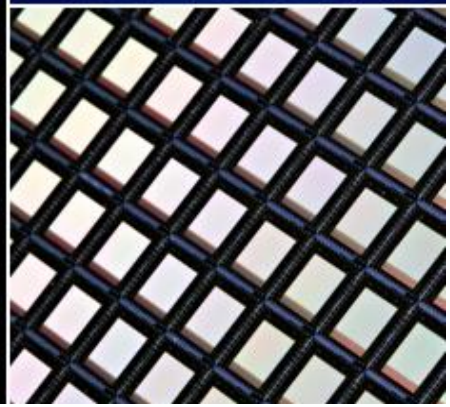
Ion Proton (Ion Torrent)



semiconducteurs







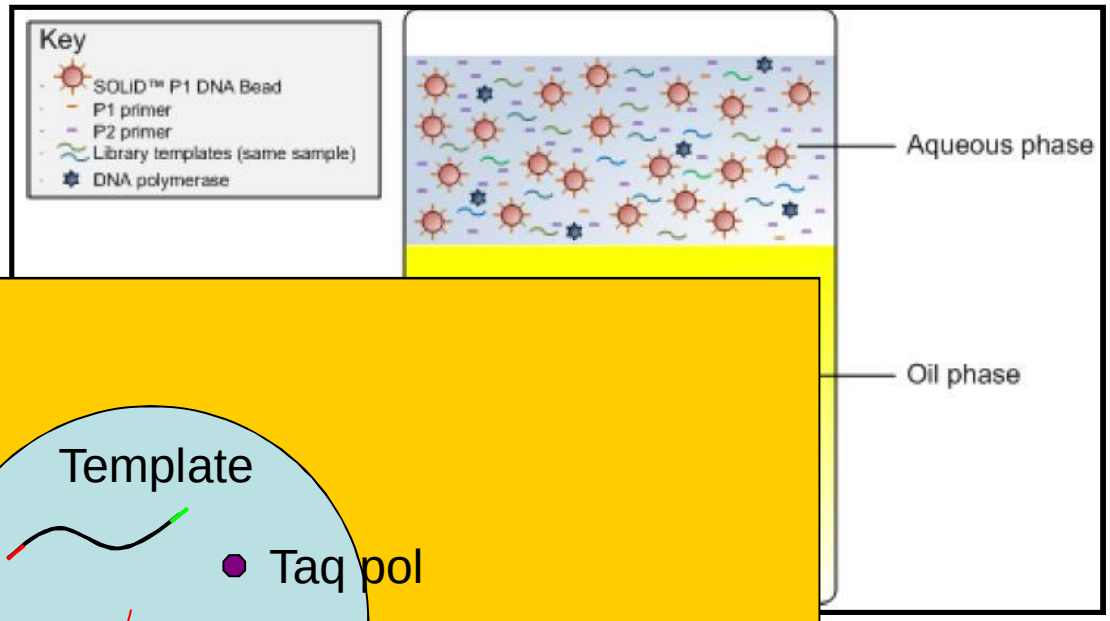
- pas de molécules fluorescentes
- run < 2 heures
- qualité données brutes : Q20
- R&D très active

ion torrent

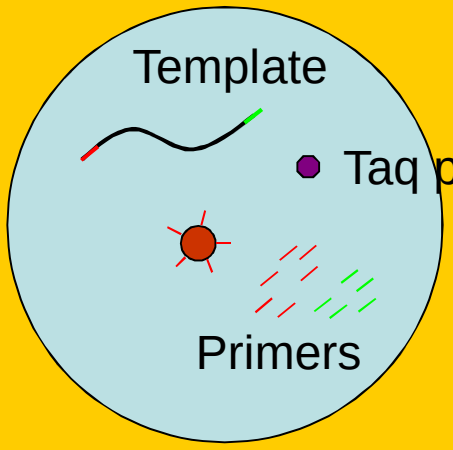
by life technologies™

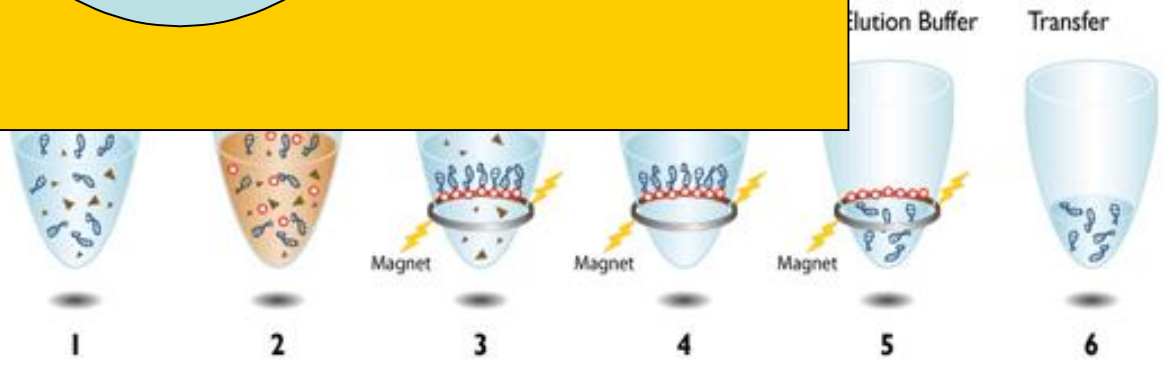
Préparation de la matrice

Emulsion



PCR





WorkFlow – PGM / Ion Torrent

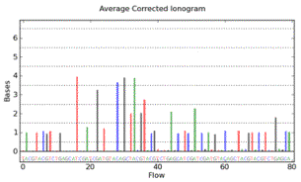
A **ADN**

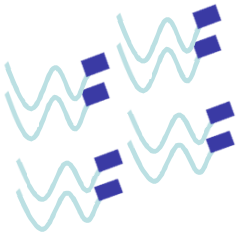
B **Préparation dûne librairie compatible**

C **Ion Xpress™ Préparation de la matrice**

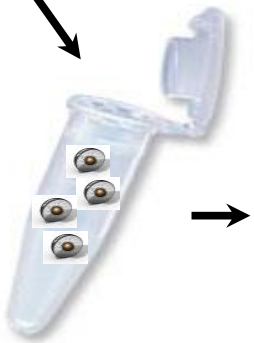
D **Sequençage**

E **Données compatibles FASTQ**

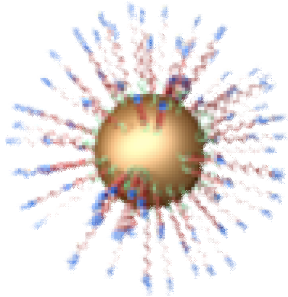


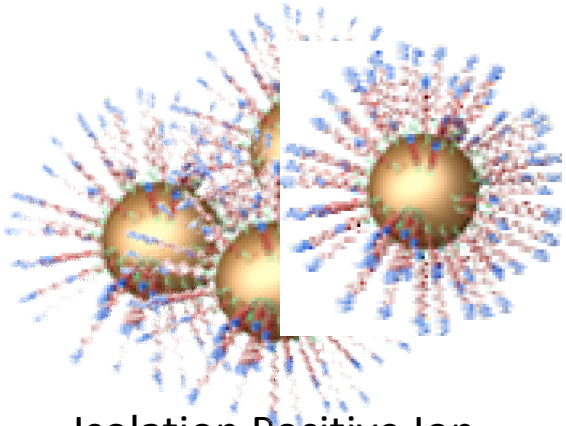


Préparation de la librairie



Amplification Clonale

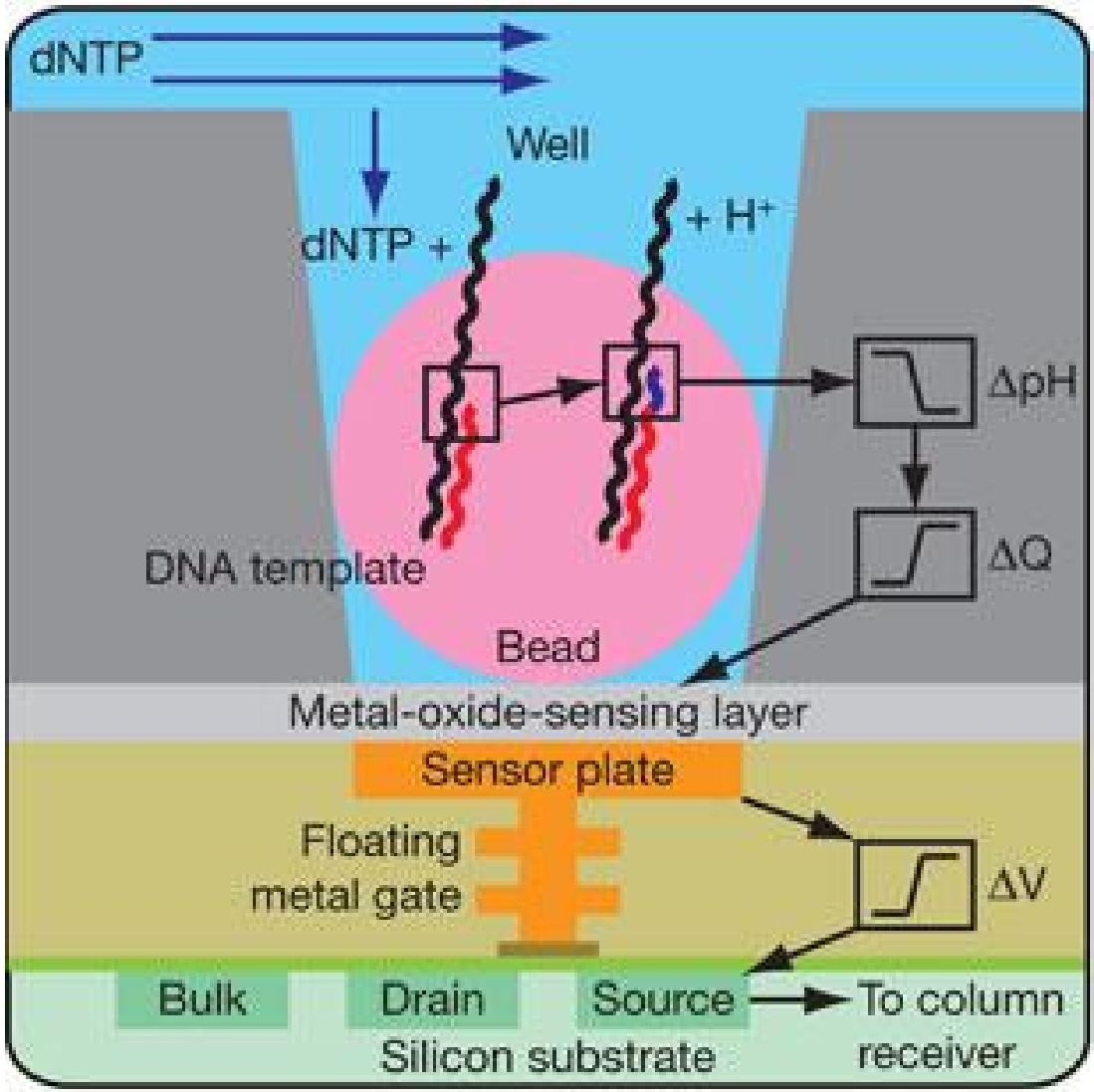




Isolation Positive Ion Sphere™ Particles

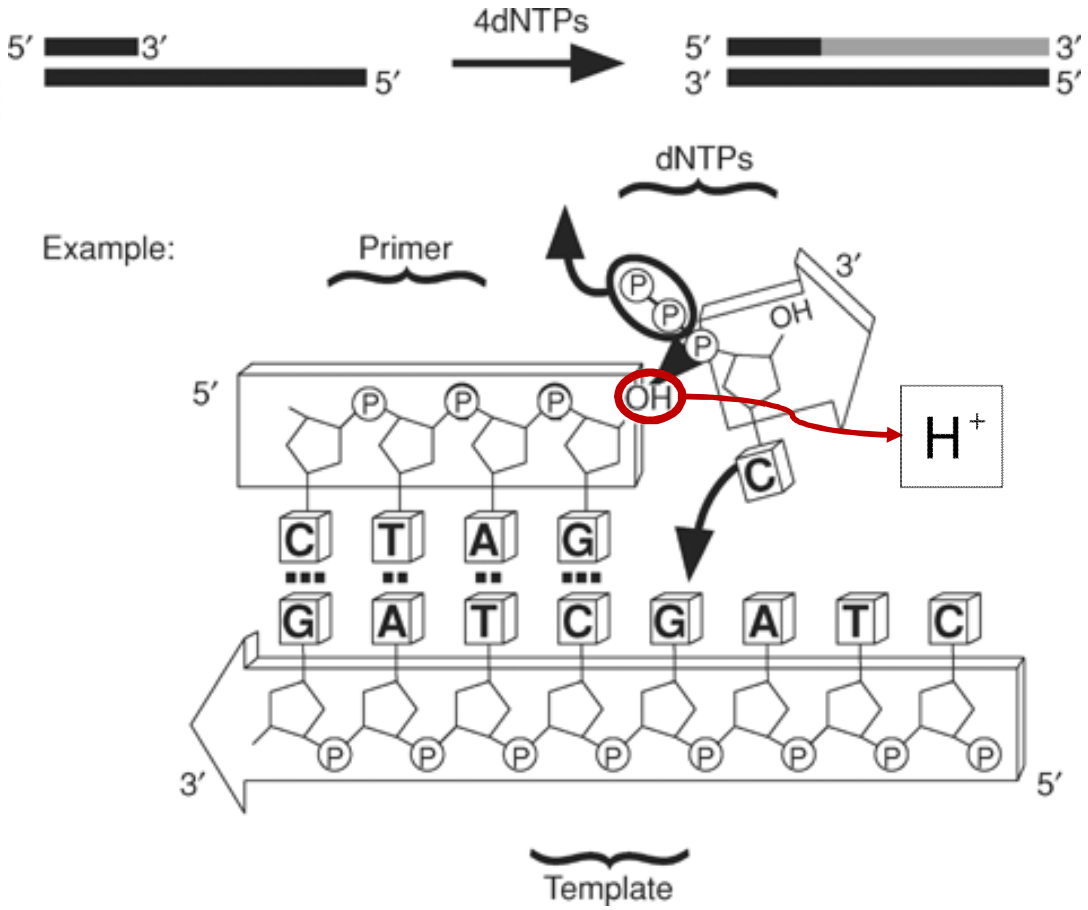
Chargement des puces et Sequençage

Fondation reconnue dûutilité publique

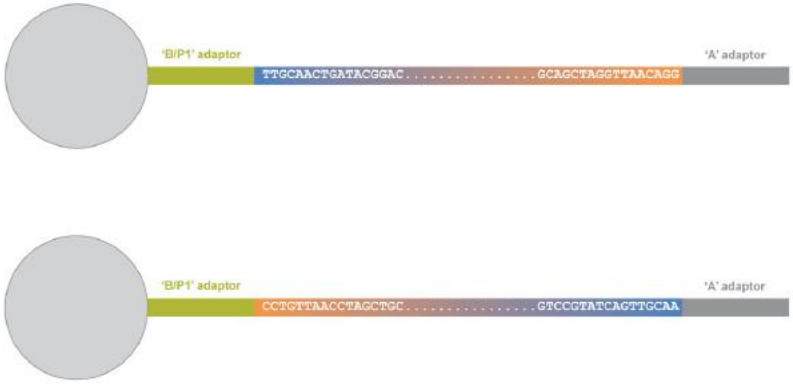


Fondation reconnue d'utilité publique

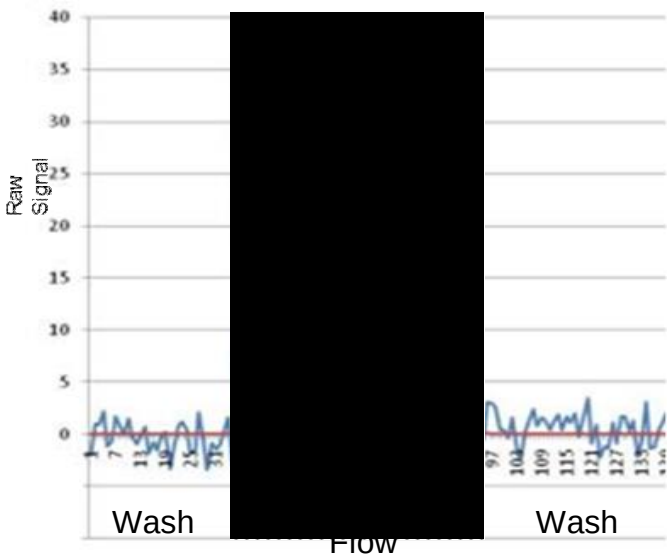
Polymérisation avec libération de proton



La donnée bruteé le ionogramme



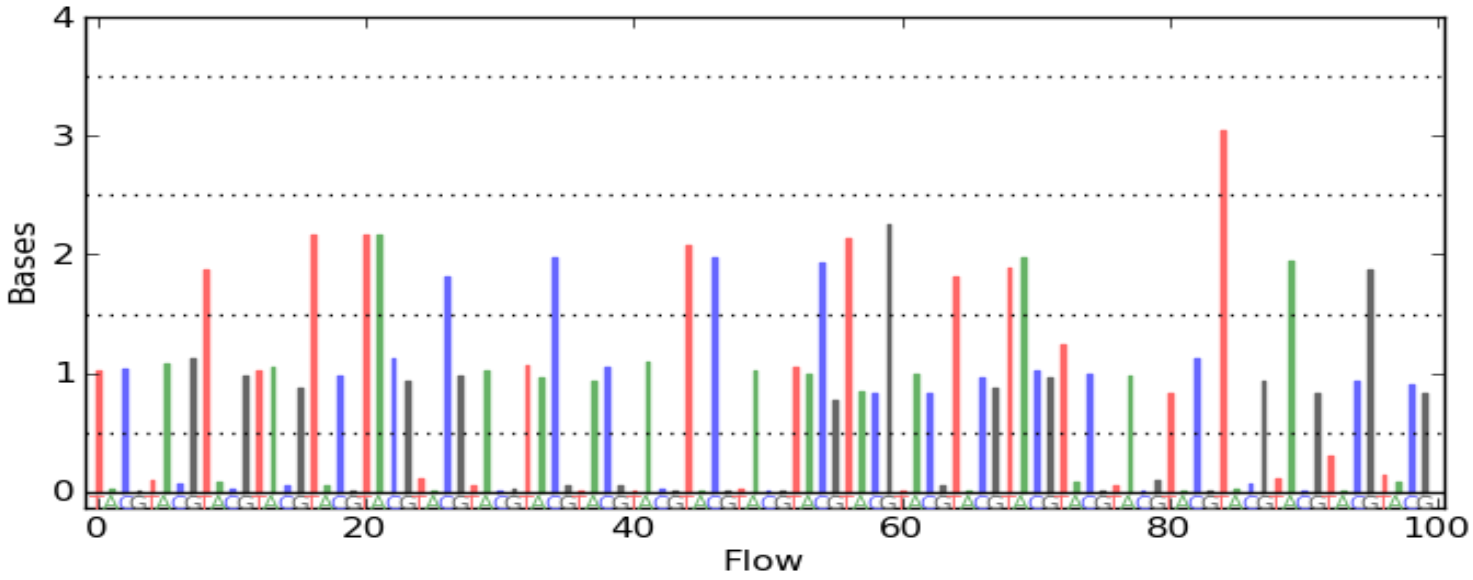
Single Well Incorporation Trace



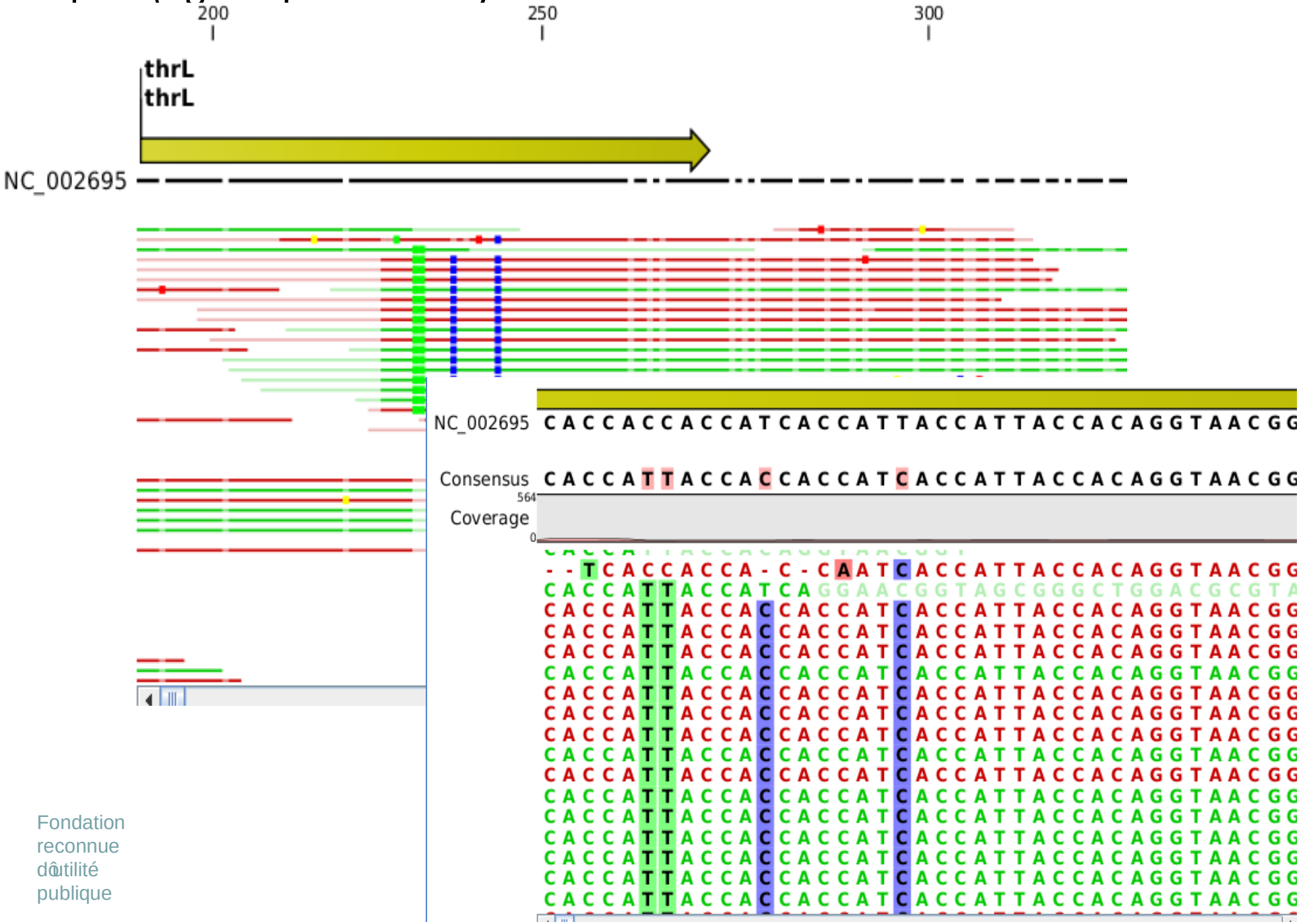
Frames (15 frames = 1 second)

Cycles : T→A→C→G

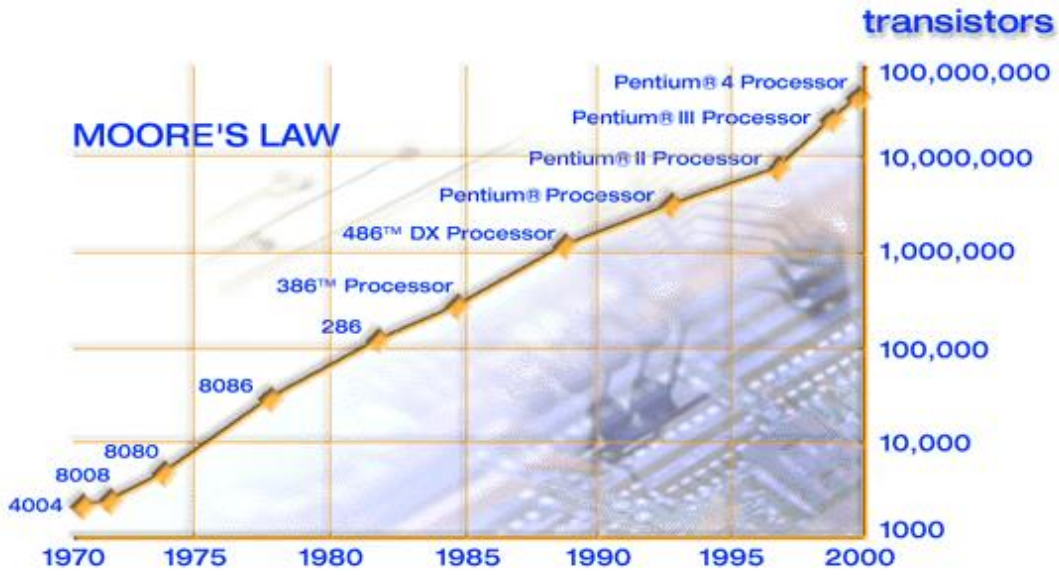
420x pour 200 nucl.
850x pour 400 nucl.

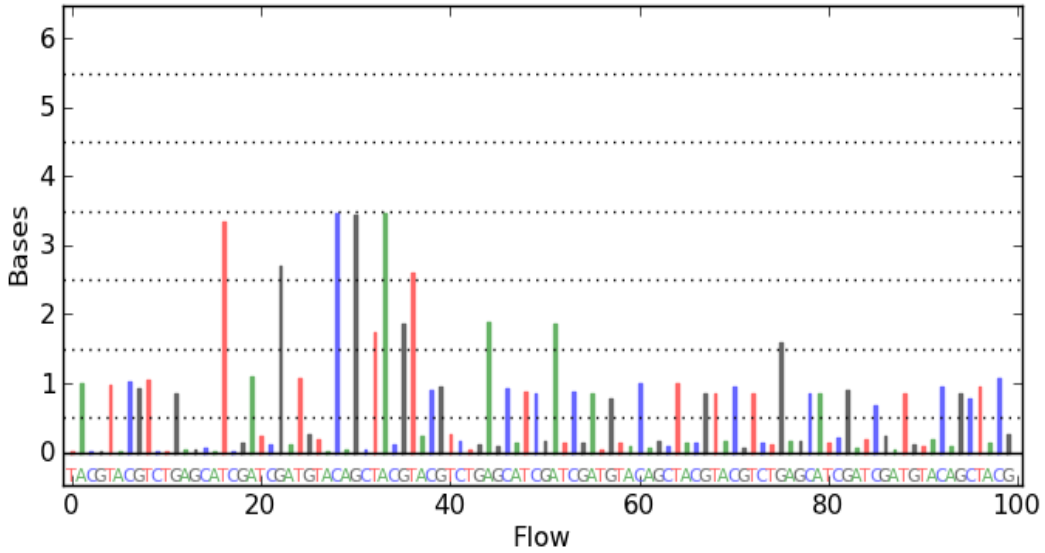


Séquençage et primo-analyse



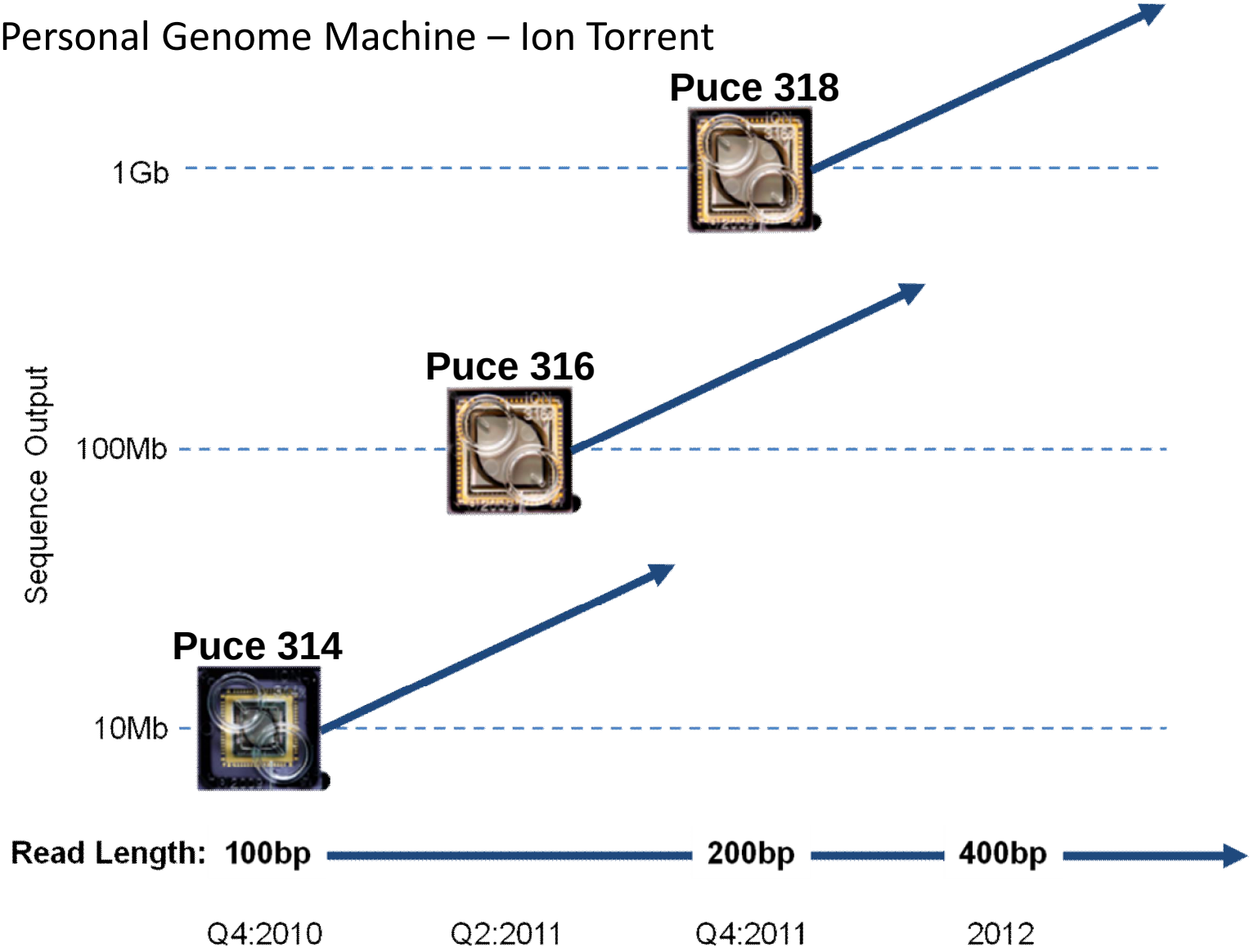
Principes généraux





Adapter le débit de séquences… à une problématique précise

Personal Genome Machine – Ion Torrent



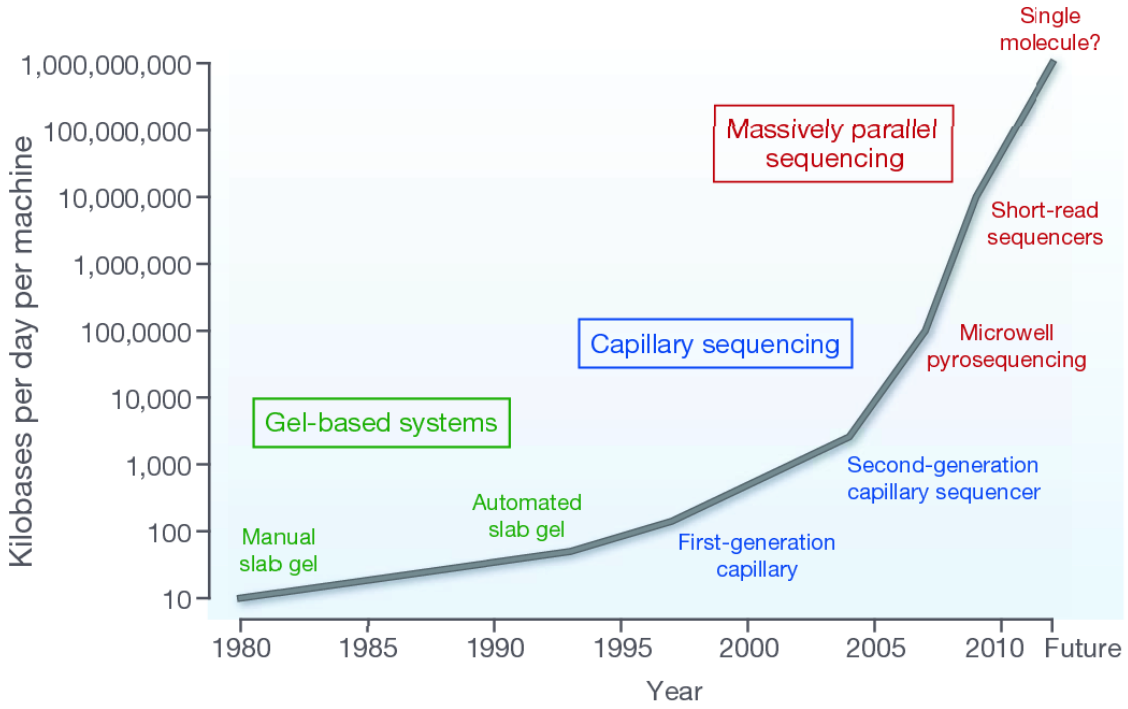


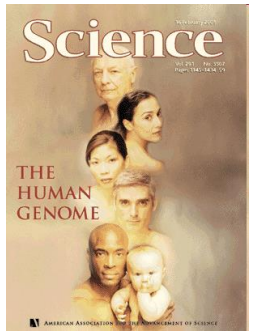
Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.

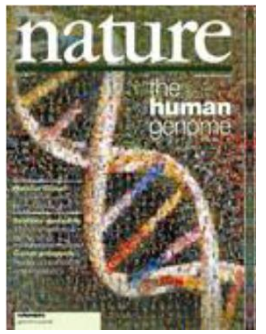
Temps de doublement = 16 mois

1998 0.3 10^6 bases/jour
2008 10^9 bases/jour

Séquençage du génome humain

Génome humain = 25000 gènes
3,4 milliard de bases
4000 livres de 250 pages





1990-2001

Séquençage du génome humain :
- Débuté en 1990 par Human Genome Project (HGP) et Celera
- Publié (ébauche) en 2001 et assemblé (99.99%) en 2003
- Coût = 3 milliards de $

Séquençage du génome humain

Fin 2012

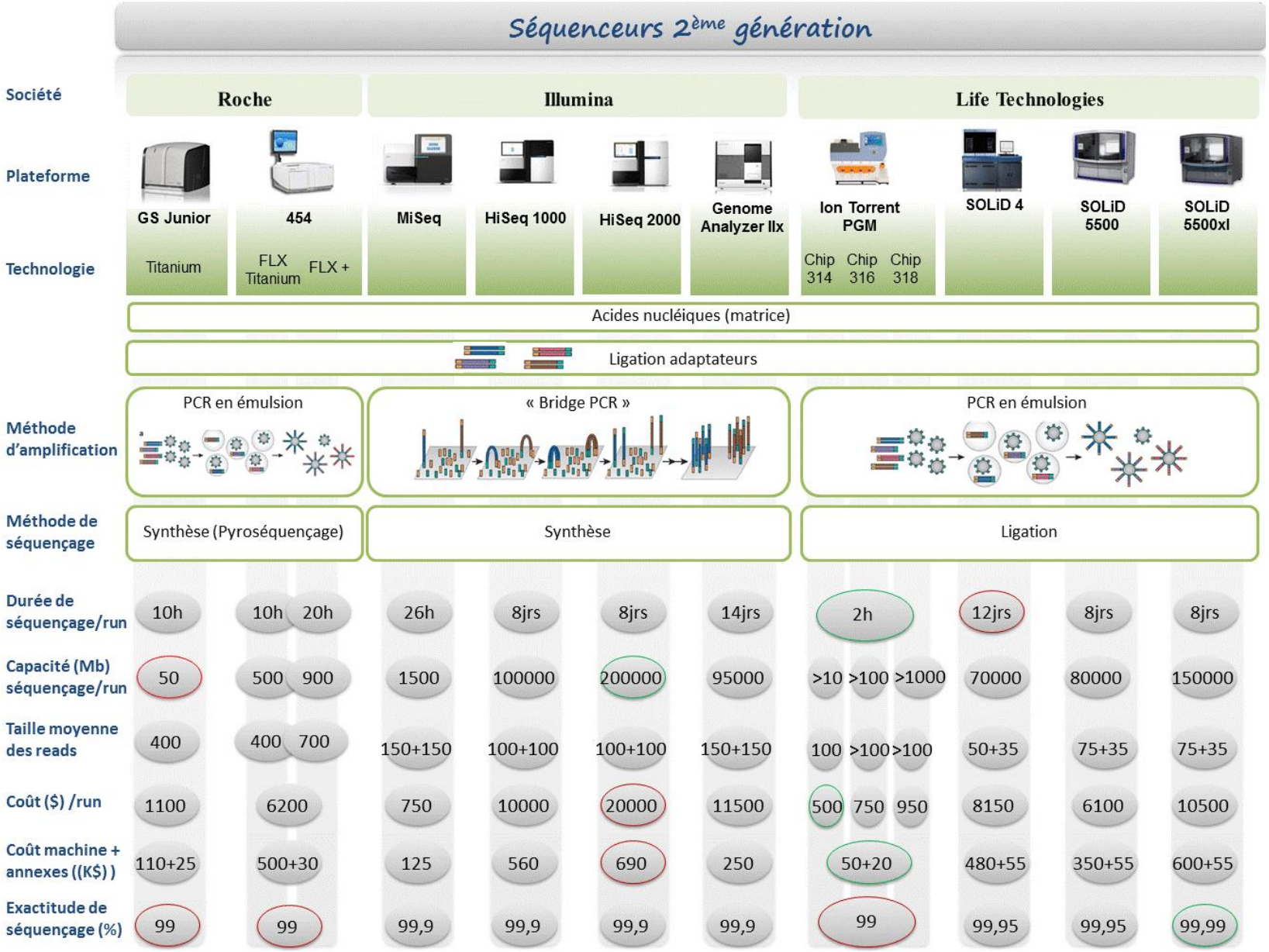
Séquençage du génome humain :

- IonProton (Life Technologies)
- Durée du run = 1 journée
- Coût = machine : 149000$; Consommables : 1000$

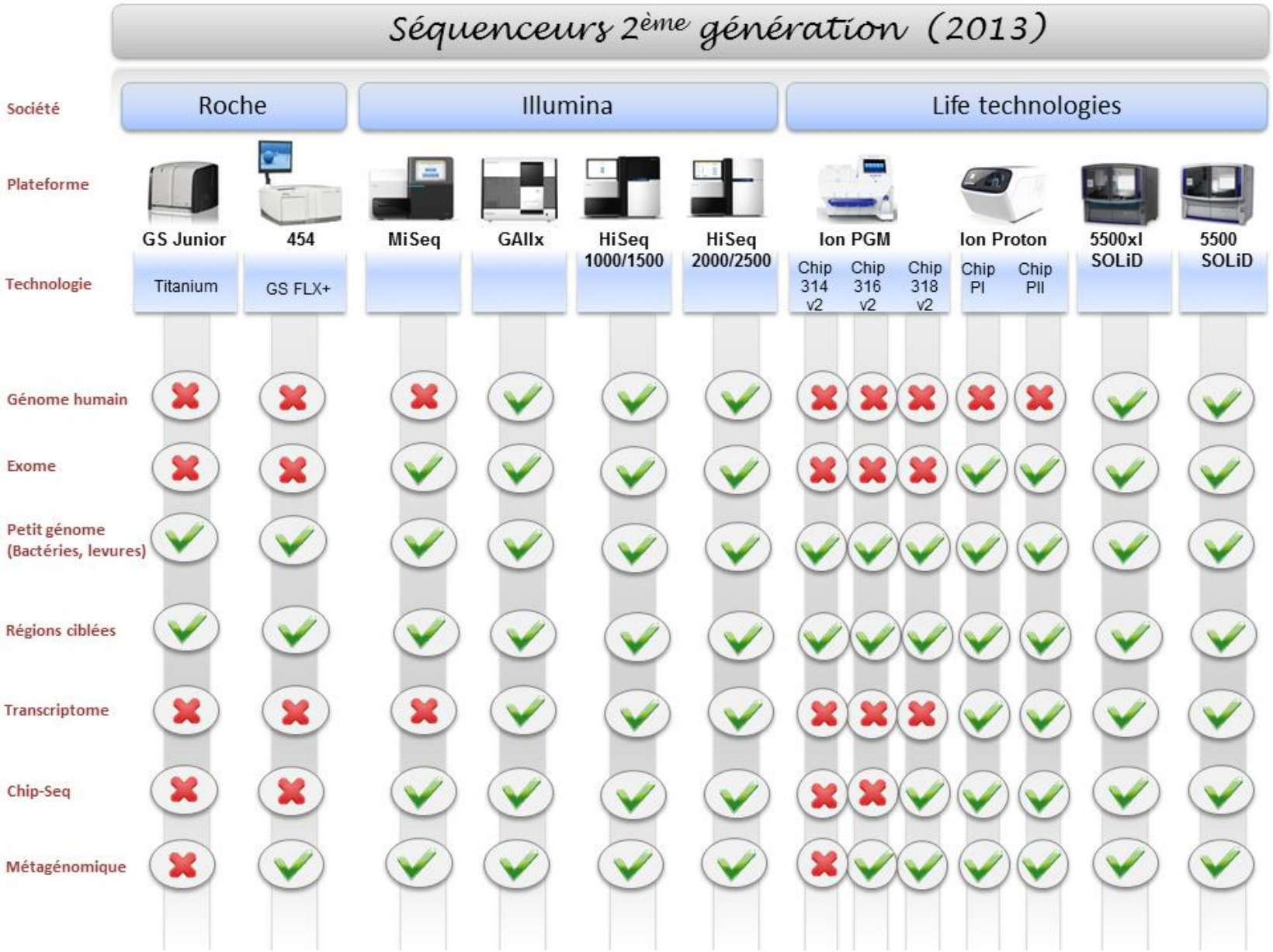
Chip PI

Institut Pasteur de Lille

Séquenceurs 2ème génération



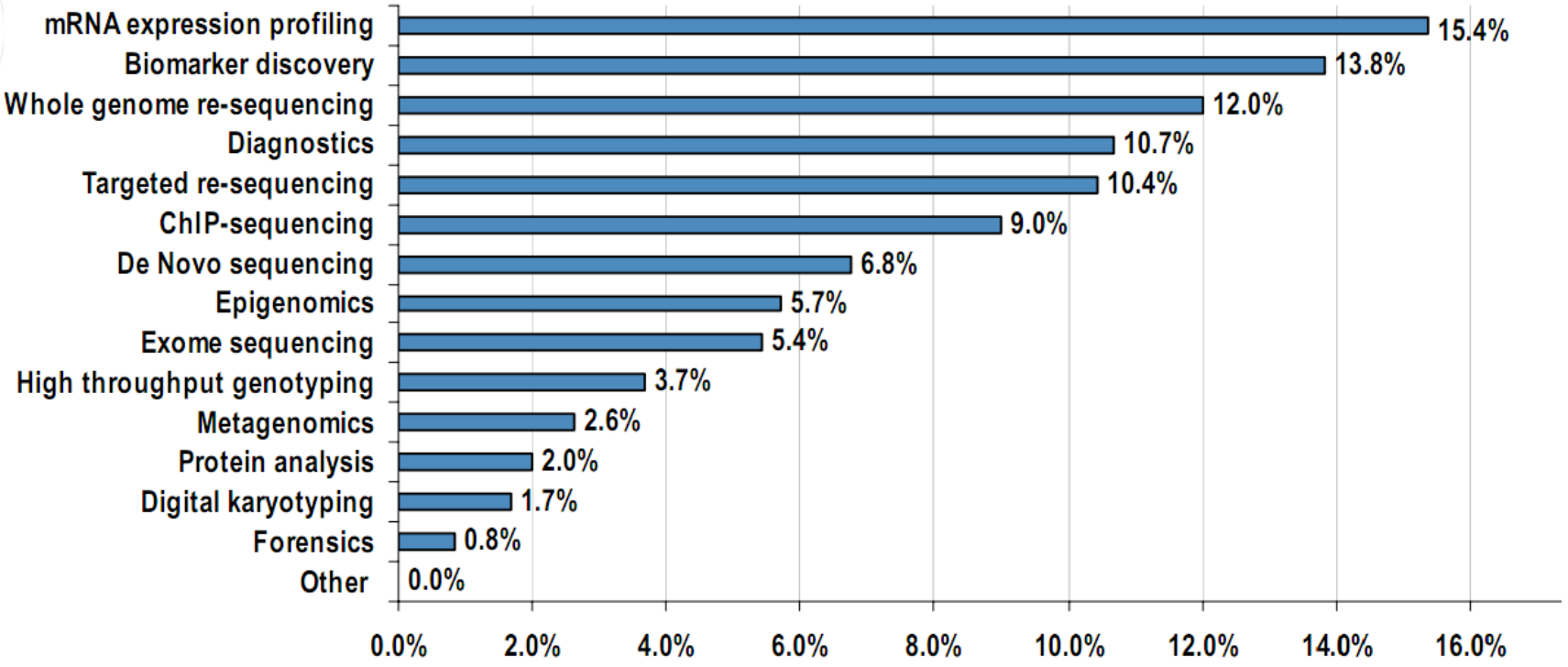
adaptation : Renaud Blervaque / biorigami.com



adaptation : Renaud Blervaque / biorigami.com

Une plateforme commune pour des champs d'applications couvrant tout le spectre des études transcriptomiques et génomiques

Pourcentage des différents champs d'applications sollicitant les plateformes de séquençage haut-débit (données 2009-2010)



Adapté du rapport J.P. Morgan : Next Gen Sequencing Survey (2010)

Re-séquençage

- Séquençage de souches d'*E. coli* isolées lors de l'épisode épidémique de l'été 2011 dans le Nord
- Entre le 6 juin et 1er juillet 2011 => 12 cas de syndrome hémolytique et urémique (SHU) liés à la consommation de steaks hachés (source InVS)
- Le mercredi 29 Juin : réception d'ADN génomique d'un isolat
  Procédure de validation => Proposition travail 316x (1ère mondiale)
- Le vendredi 1er juillet : séquençage du génome entier sur puce 316 (100Mb)
- Intérêts : - Faisabilité (3 j de manipes, 1 semaine d'analyse)
  - Souches multiples, 4 groupes différents
  - Identification région hypervariable => diag. différentiel testé, validé
  - Base pour le développement d'un pipeline d'analyse automatique

Institut Pasteur de Lille

Institut Pasteur de Lille

Centre d'Infection & d'Immunité de Lille

GENES DIFFUSION
Nous innovons, vous progressez.

GenoScreen

ars
Agence Régionale de Santé
Nord-Pas-de-Calais

Communiqué de presse

05/07/2011

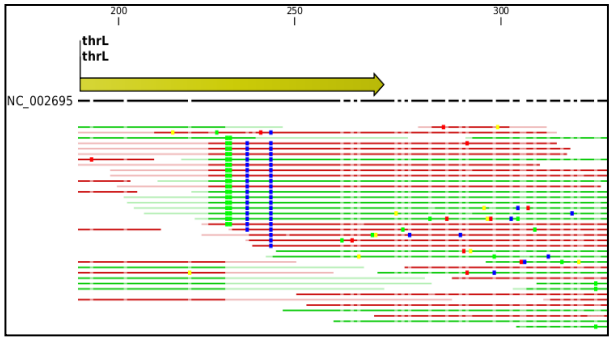
L'*Escherichia coli* séquencée en un temps record à l'Institut Pasteur de Lille

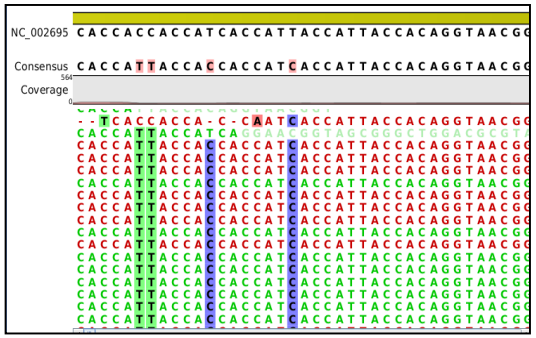
Une première mondiale : le patrimoine génétique du germe d'*Escherichia coli*, responsable de l'intoxication d'une dizaine de personnes dans le Nord de la France, a été déchiffré en moins de trois jours grâce à une nouvelle technologie de séquençage par semi-conducteurs (Ion Torrent, Life Technologies) et une collaboration efficace entre le Centre Hospitalier Régional Universitaire de Lille, l'Institut Pasteur de Lille et la société Gènes Diffusion.

Vendredi 1er juillet, soit moins de trois jours après la réception de la bactérie incriminée, les équipes de la plateforme de Transcriptomique et Génomique Appliquée de l'Institut Pasteur de Lille* et de la société Gènes Diffusion, qui travaillent en collaboration sur le site de l'Institut Pasteur de Lille, ont réalisé un séquençage complet du génome de la bactérie *E. coli*, isolée d'un des enfants hospitalisés au CHRU de Lille.

Comparaison aux souches répertoriées

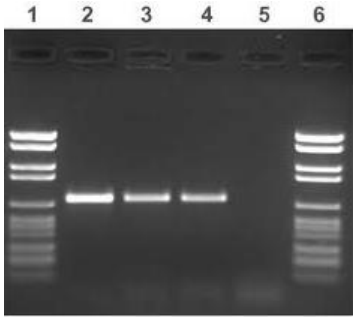
SNPs-Indels



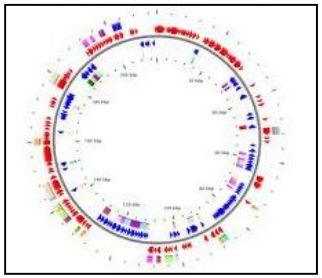


Région hypervariable

```
Query  1        GAATCGTCAGTTGCGGGTGGAAAAAATCCTGAATCGCTCCGGCATCCAGCCGTTGCACAG
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1404311  GAATCGTCAGTTGCGGGTGGAAAAAATTCTGAATCGCTCTGGCATCCAGCCATTGCACCG

Query  61       CAAATGCTCGTTTGCGAATTATCAGGTGCAGAACGACCGCCAGCGATACGCGTTGAGTCA
                |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1404371  CAAATGCTCGTTTTCGAATTACCAGGTGCAGAACGAAGGCAGCGATACGCGTTGAGTCA

Query  121      GGCGAAATCCATCGCTGATGAACTGACGACCGGGTGTACAAATTTTGCGTTCAGCGGAAA
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1404431  GGCGAAATCCATCGCTGATGAACTGATGACCGGGTGTACAAATTTTGCGTTCAGCGGAAA

Query  181      ACCTGGTACCGGGAAGAACCACTTAGCGGCAGCTATCGGGAATCGCCTGCTGAAAGACGG
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1404491  ACCTGGTACCGGGAAGAACCACTTAGCGGCAGCTATCGGGAATCGCCTGCTGAAAGACGG

Query  241      TCAGACAGTGATTGTGGTTACCGTGGCTGATGTTATGAGTGCCCTGCACGCCAGCTATGA
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1404551  TCAGACAGTGATTGTGGTTACCGTGGCTGATGTTATGAGTGCCCTGCACGCCAGCTATGA
```

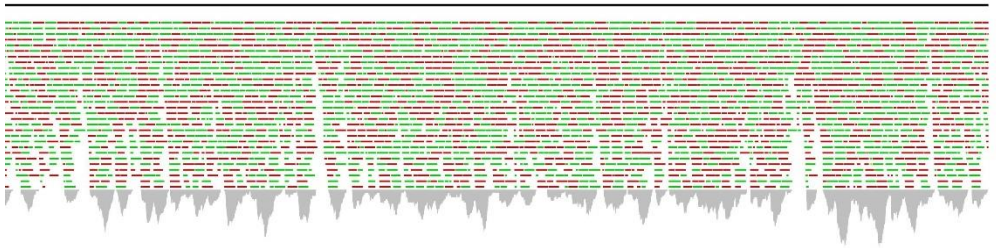


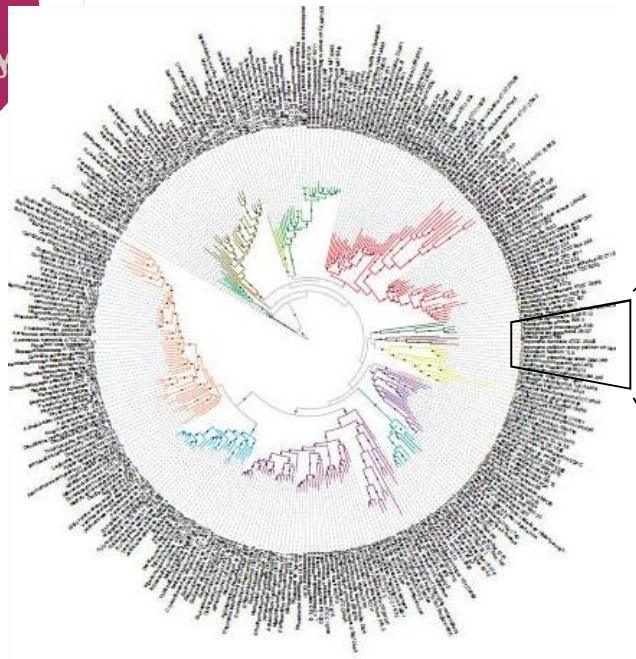
Identification des caractéristiques génomiques (/ souches voisines)

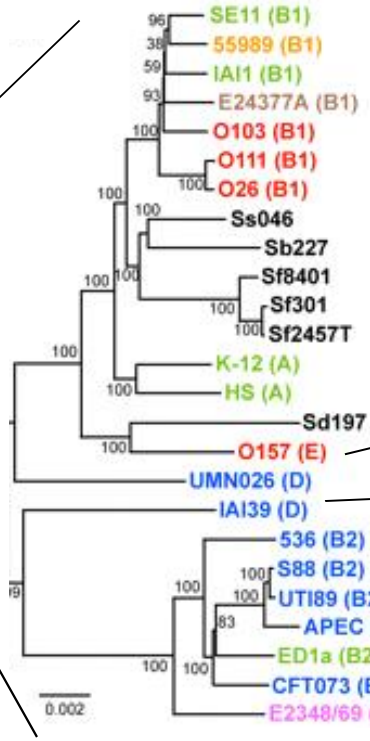


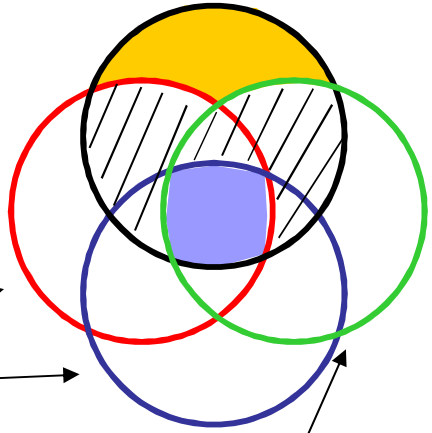
Fondation reconnue d'utilité publique

Principe dôanalyse dôune nouvelle séquence



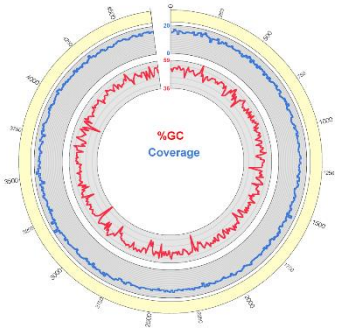


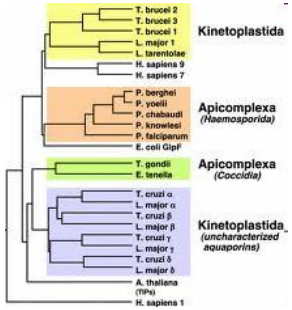




Objectifs du pipeline d'analyse

1 / Identification de la souche





2/ Replacer la souche dans un contexte phylogénétique

3/ Identifier les variations (SNPs, DIP, HTG, Délétions)

4/ Caractérisation automatique des éléments nouveaux (HTG)

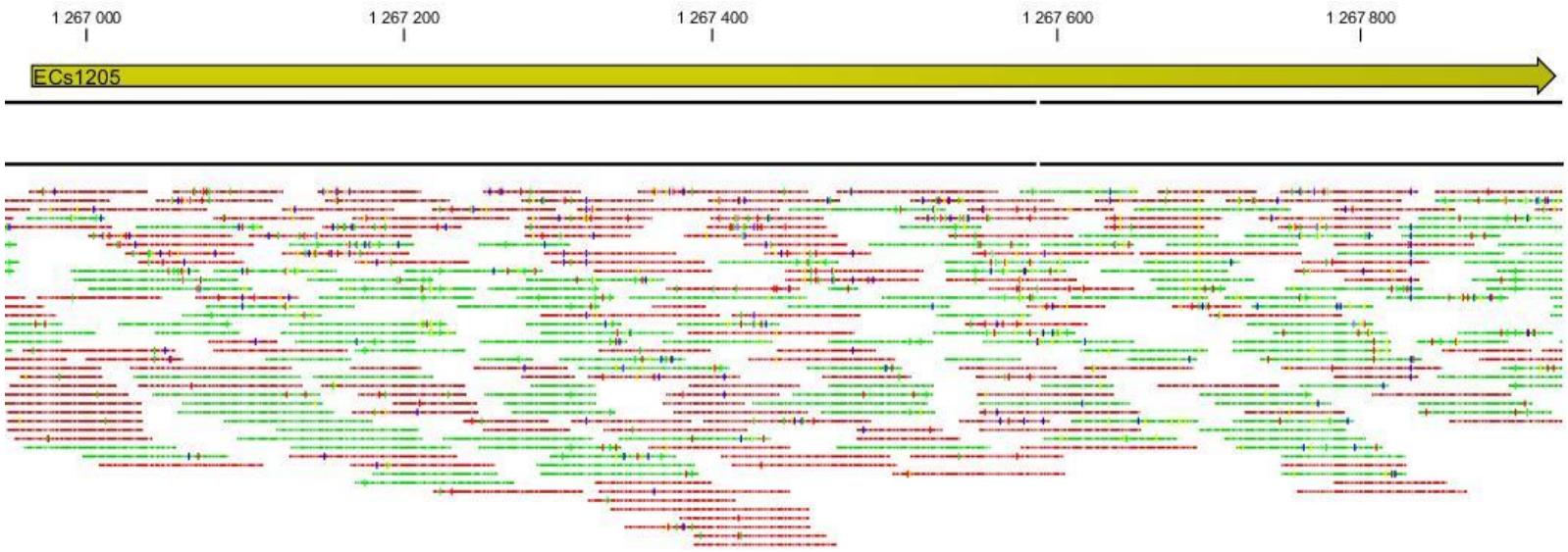
5/ Assemblage des éléments nouveaux (HTG)

Alertes en fonction du champ d'expertise

6/ Design automatique de test (oligos pour PCR diag.)

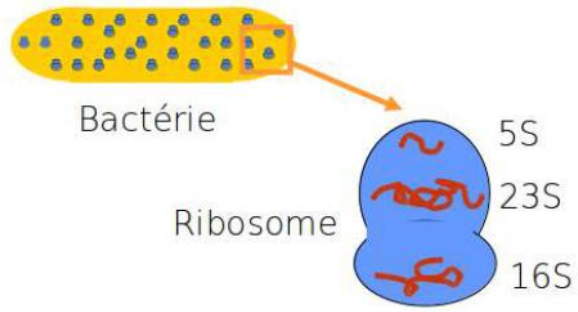
Processus auto-enrichi

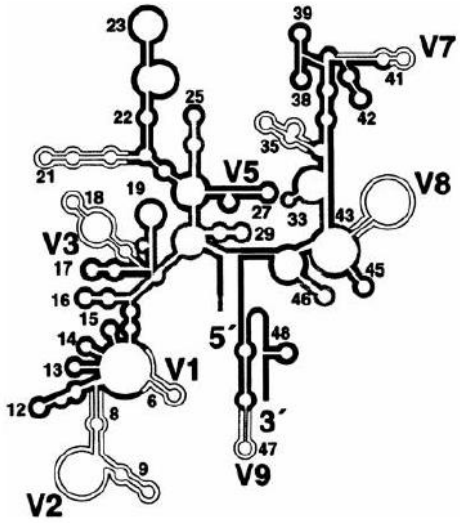
GENOME	Taille (bp)	Couverture (% taille totale)	Profondeur	% reads alignés	Nbr. mutations ponctuelles	Mut. ponctuelles intra CDS	Nbr. mutations codantes	Nbr. DIP	Nbr. DIP dans les CDS	Taux de mutations normalisés
CB9615	5386352	96.5	16.6	90.9	12908	10616	3706	437	231	0.0026
O111	5371077	87.3	14.6	79.9	78990	70597	13158	1521	519	0.0172
sakai	5498450	95.4	16.3	91.2	6627	5612	2352	423	224	0.0013
TW14359	5528136	95.9	16.3	91.9	7254	6021	2433	425	238	0.0014

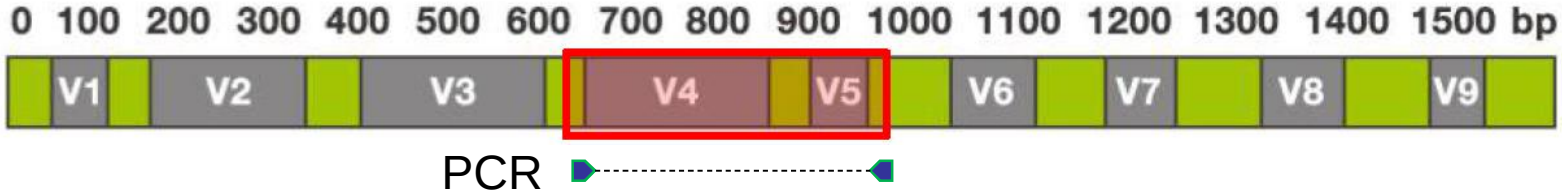


La métagénomique

Principe : Etablir le catalogue des microorganismes présents (microbiote) dans un environnement données (microbiome) via une approche de séquençage à haut débit

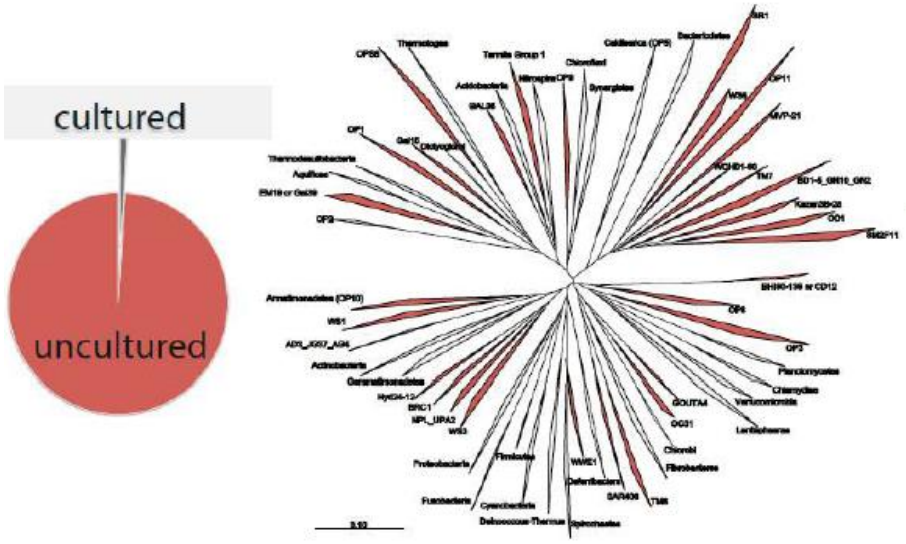




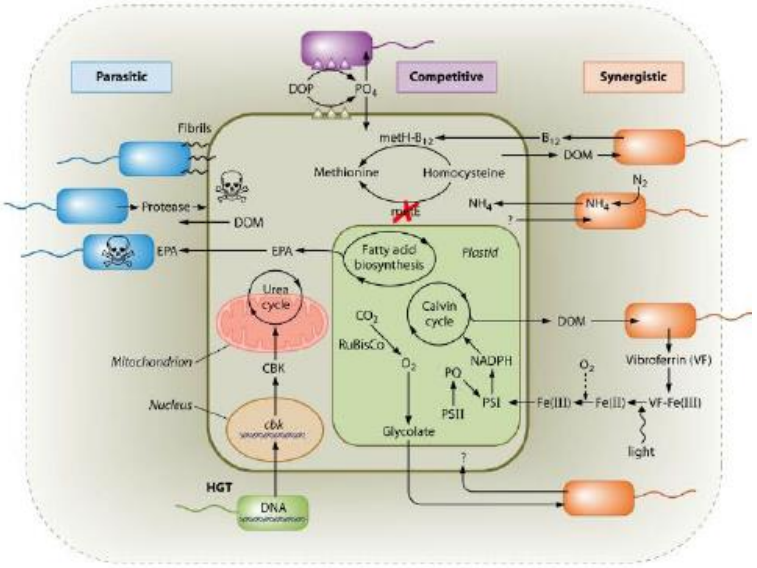


La métagénomique

Accéder aux génomes non cultivables



Qu'est-ce qui s'exprime dans mon échantillon ?

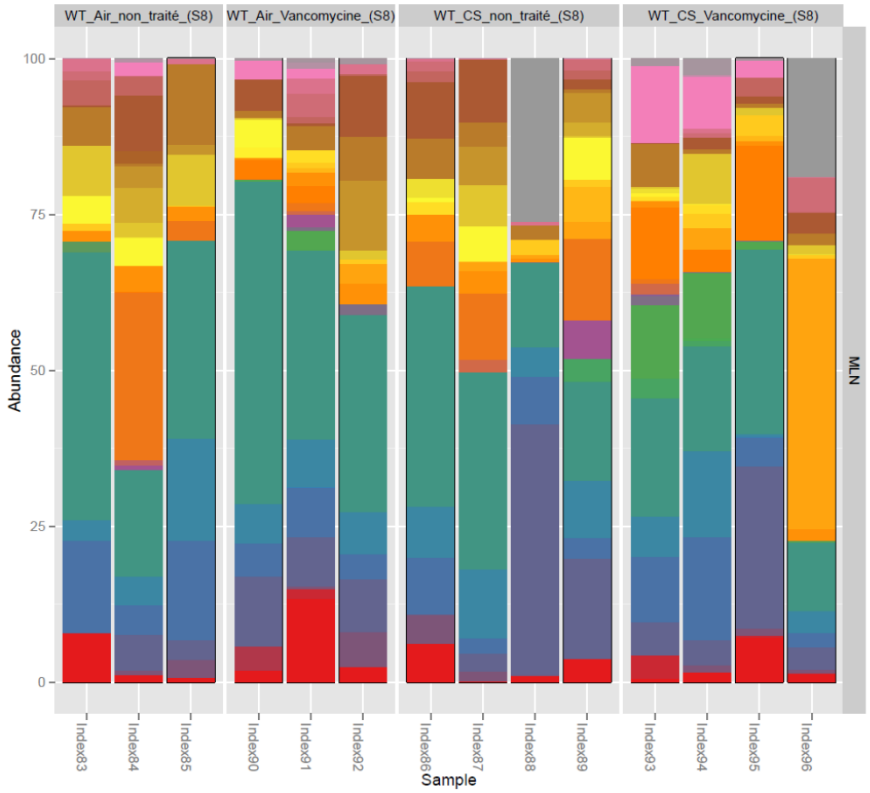


Microbiote Humain
>1.000.000 genes

Gènes Humain
~23.000 genes

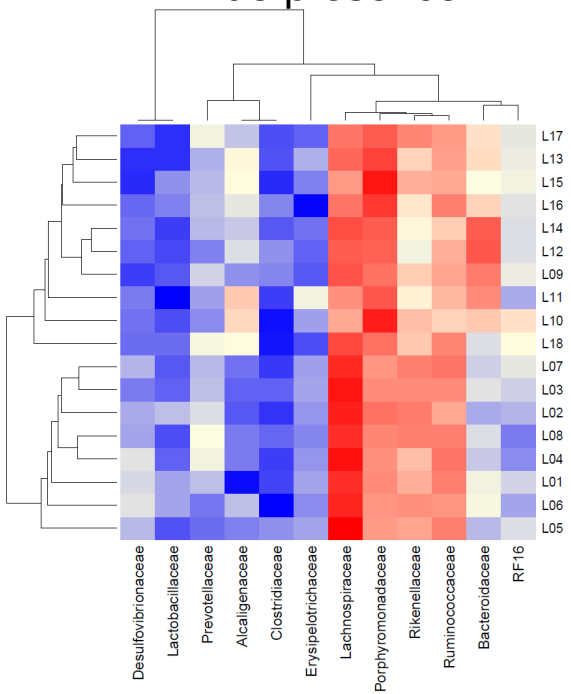
Fondation reconnue d'utilité publique

La métagénomique ï Analyse Biostatistique



Histogramme de diversité

Heatmap de corrélation de présence



Institut Pasteur de Lille

INTRODUCTION

Since 2008, researchers with the European consortium MetaHIT have been analyzing the collected genomes of the microorganisms present in our intestine : the microbiota.

RESEARCH

Little understood until now, the intestinal microbiota interests researchers as an avenue of inquiry to explain the evolution of chronic diseases.

FINDINGS

The MetaHIT consortium published two major findings in the scientific journal Nature : an established catalog of bacterial genes in the intestine; and the discovery of enterotypes.

PERSPECTIVES

MetaHIT opens avenues for further efforts in the field of human microbiome research : early detection of chronic diseases, personalized medicine and more healthful food.

MetaHIT — Metagenomics of the Human Intestinal Tract — European research project

Budget

22 million euros

The 4 year program was financed in large part by the European Union under the FP7 (*7th Framework Programme*).

Laboratories

8 countries
14 research & industrial

institutions are involved in the consortium, with more than 50 researchers and cooperation between Europe and China.

The microbiota

The microbiota is an ecosystem composed of billions of bacteria that make up a veritable "organ." Within 24 hours of birth, these bacteria colonize our digestive tract to form our intestinal microbiota (2kg for adults). MetaHIT focuses on the digestive tract since it is where the largest and most diversified bacterial community lives in our body.

Observations

chronic diseases
infectious diseases

Observations made in the past 50 years cannot be solely explained by variations of our genome.

Research themes

Nutrition. Better knowledge of the intestinal microbiota of individuals will enable the nutritional needs to adapt to everyones specific nutrient needs.

Medicine. With the study of the microbiota and the established catalogue of genes, we can have an unprecedented overview of the microbiota in healthy individuals and in patients. With the discovery of enterotypes we can imagine the upcoming development of new diagnostic or even prognostic tools for human health.

DEFINITION

***Enterotypes**

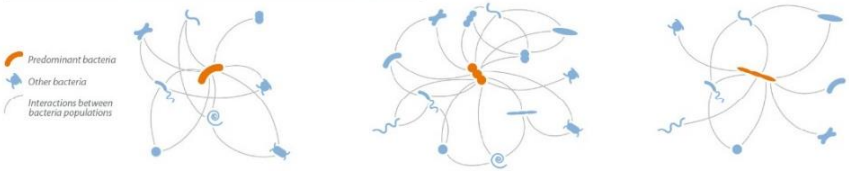
There are three in the world's population, each characterized by a predominant bacterial population.

Genome sequencing

3,3 million genes

The gut bacterial gene catalog, which can be compared to a *molecular scanner*, was established by metagenomic high throughtput sequencing and allows the observation of the human gut microbiome.

Discovery of the 3 enterotypes*



a. Bacteroides — b. Prevotella — c. Ruminococcus

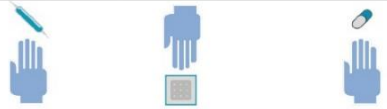
Chronic diseases

Disturbances in the microbiota can be early warning signs for certain diseases like Crohn's disease or diabetes.

Nutritional impact

If it is possible to reveal early warning signs of obesity, one can imagine nutritional intervention and diet advice being used to reestablish a healthy microbiota. The possibility of intervening directly in the flora, in the case of disturbance to the intestinal ecosystem, could also be envisioned.

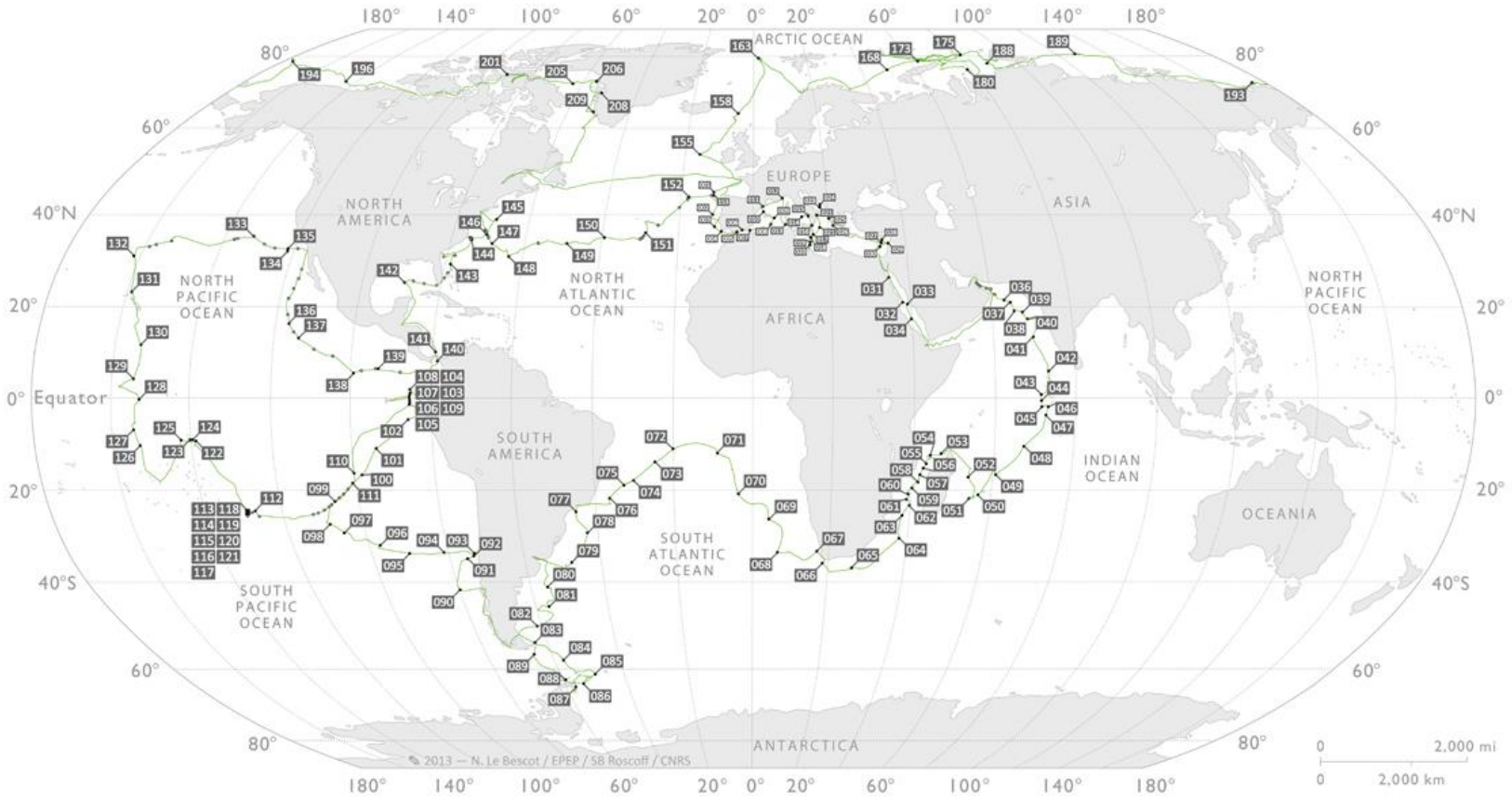
Personalized medicine



Classification by enterotype will help in the development of diagnostic tools able to reveal cases where a planned treatment would not be effective, and to adapt it accordingly.

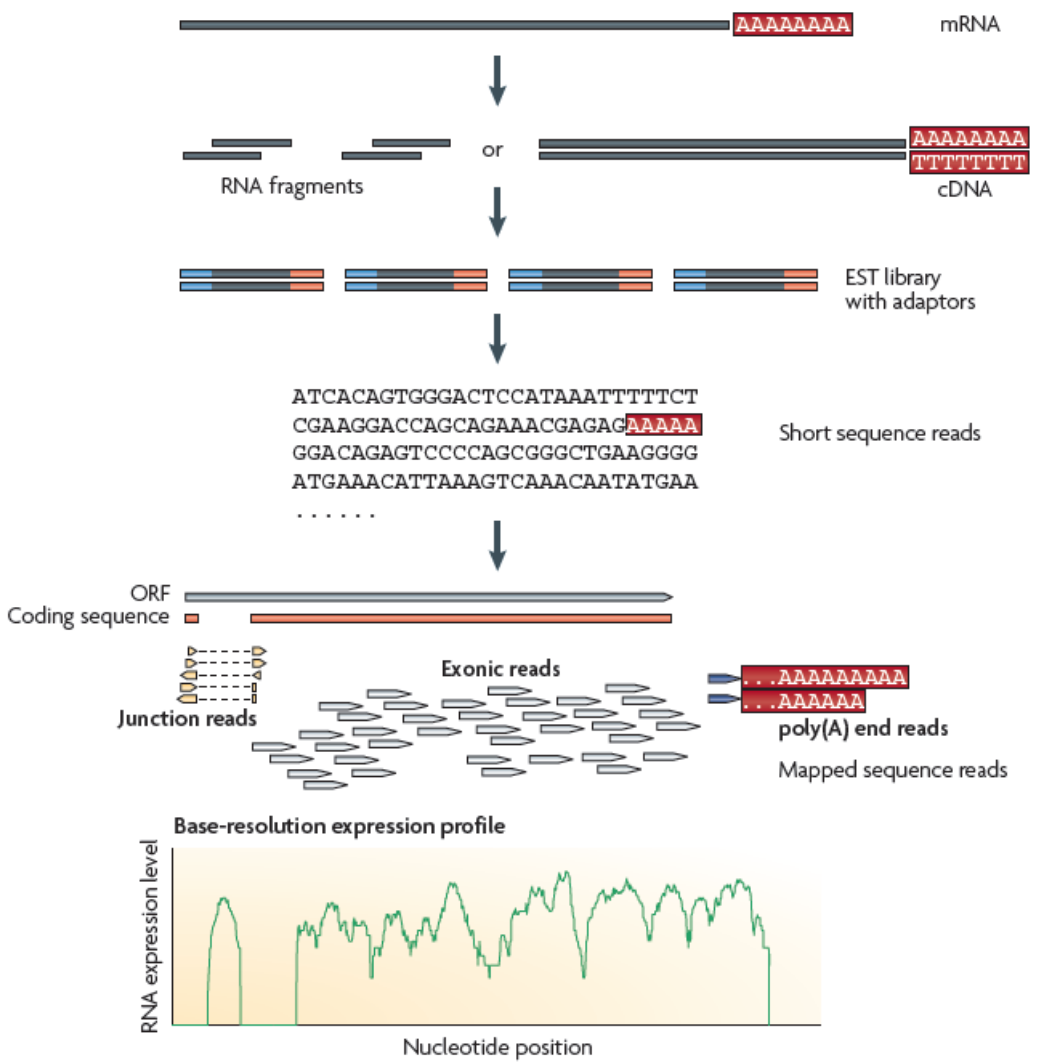
Tara Oceans

Institut Pasteur de Lille



Fondation reconnue d'utilité publique

RNA-seq


AAAAAAAA
mRNA
or
RNA fragments
AAAAAAAA
TTTTTTTT
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
.....
Short sequence reads
ORF
Coding sequence
Exonic reads
Junction reads
. . .AAAAAAAAA
. . .AAAAAA
poly(A) end reads
Mapped sequence reads
Base-resolution expression profile
RNA expression level
Nucleotide position

Mapping and quantifying mammalian transcriptomes by RNA-Seq

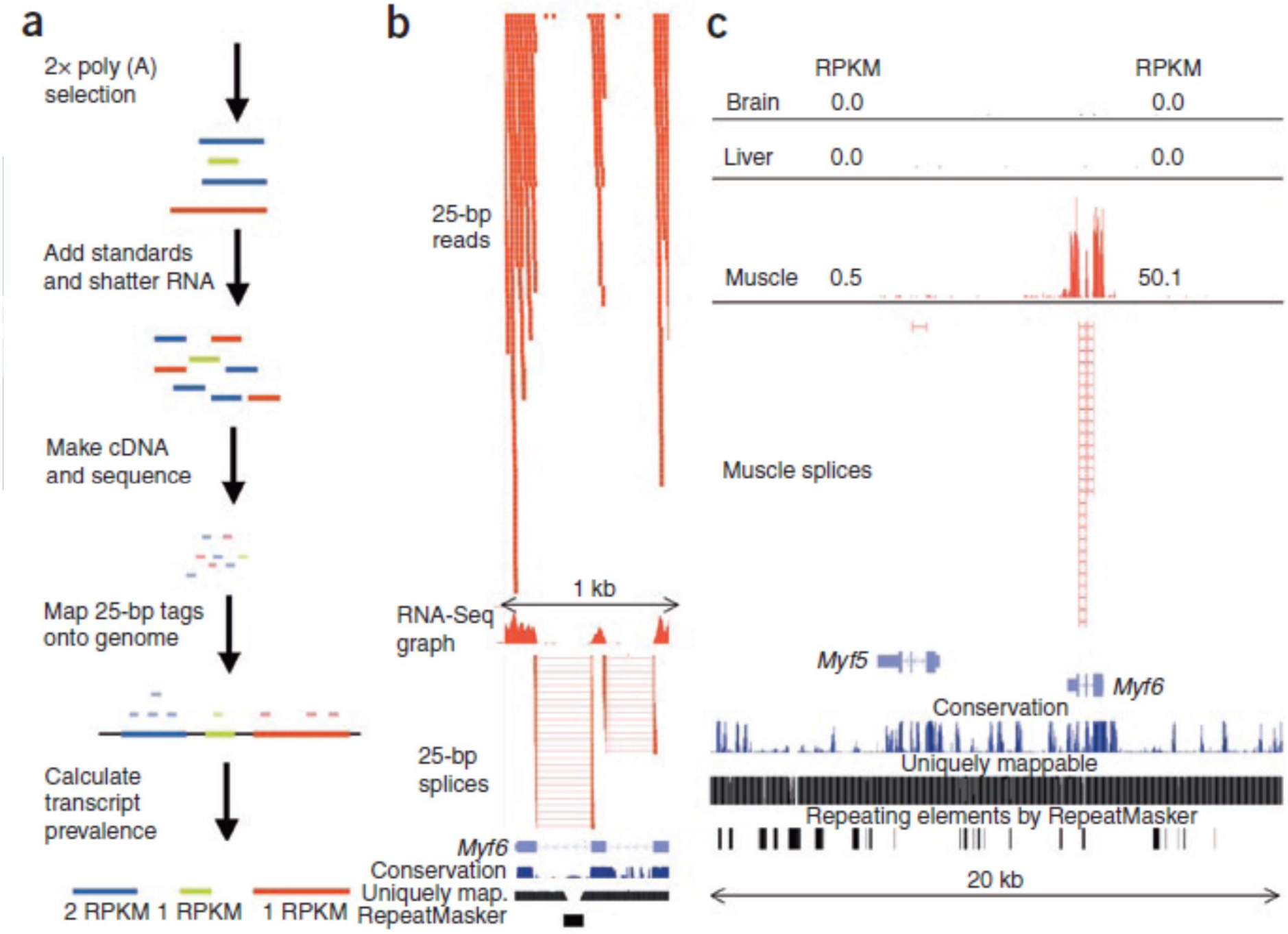
Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

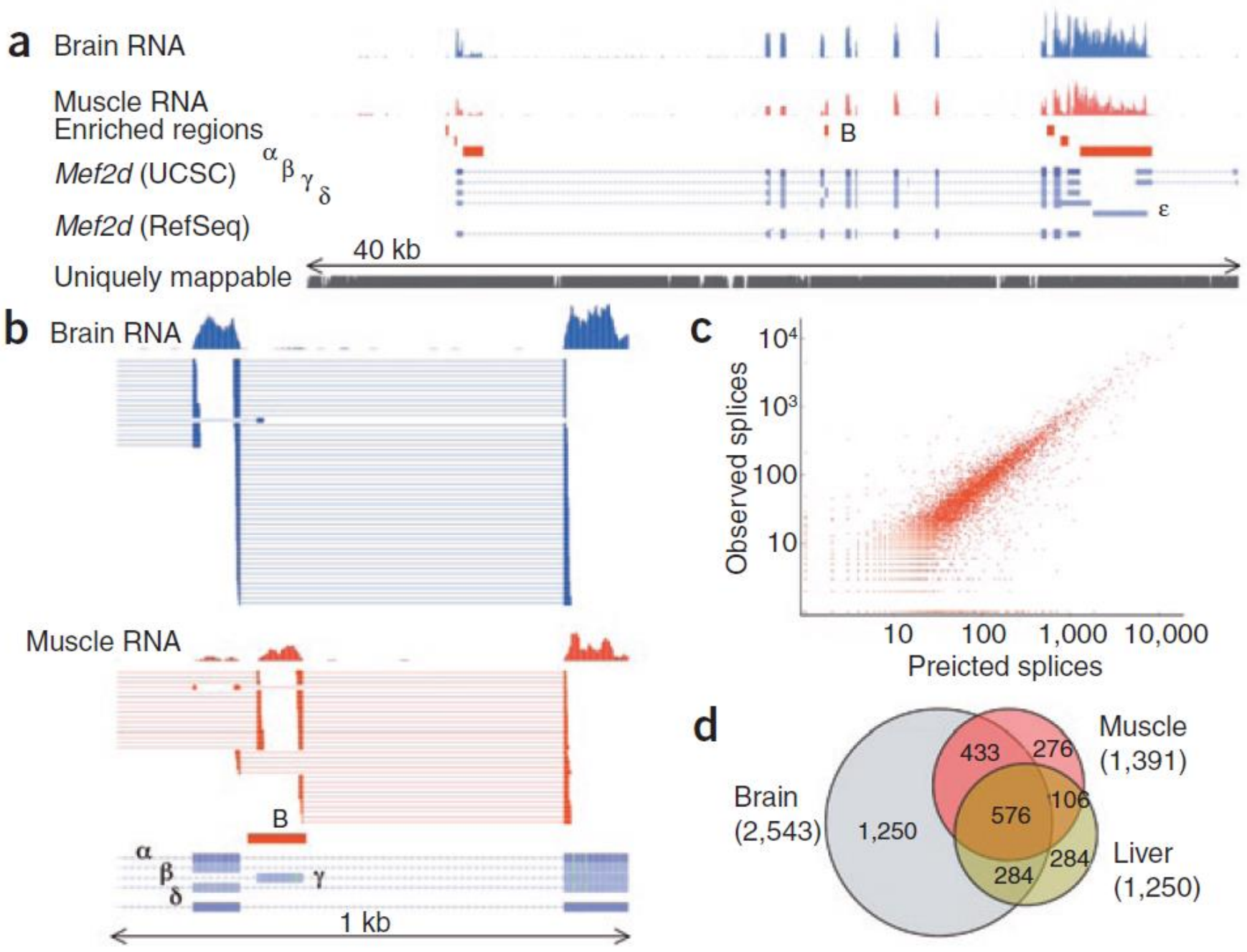
We have mapped and quantified mouse transcriptomes by deeply sequencing them and recording how frequently each gene is represented in the sequence sample (RNA-Seq). This provides a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. We report reference measurements composed of 41–52 million mapped 25-base-pair reads for poly(A)-selected RNA from adult mouse brain, liver and skeletal muscle tissues. We used RNA standards to quantify transcript prevalence and to test the linear range of transcript detection, which spanned five orders of magnitude. Although >90% of uniquely mapped reads fell within known exons, the remaining data suggest new and revised gene models, including changed or additional promoters, exons and 3′ untranscribed regions, as well as new candidate microRNA precursors. RNA splice events, which are not readily measured by standard gene expression microarray or serial analysis of gene expression methods, were detected directly by mapping splice-crossing sequence reads. We observed 1.45×10^5 distinct splices, and alternative splices were prominent, with 3,500 different genes expressing one or more alternate internal splices.

approaches to large-scale RNA analysis are serial analysis of gene expression (SAGE)[4,5] and related methods such as massively parallel signature sequencing (MPSS)[6], which use DNA sequencing of previously cloned tags 17–25 base pairs (bp) from terminal 3′ (or 5′) sequence tags. These sequence tags are then identified by informatic mapping to mRNA reference databases or, for longer tag lengths, to the source genome. A strength of SAGE and SAGE-like methods is that they produce digital counts of transcript abundance, in contrast to the analog-style signals obtained from fluorescent dye–based microarrays. However, SAGE-family assays provide no information about splice isoforms or new gene discovery, and fully comprehensive measurements of lower-abundance-class RNAs have not been achieved owing to cost and technology constraints. Expressed sequence tag (EST) sequencing of cloned cDNAs has long been the core method for reference transcript discovery[7–9]. It has both qualitative and quantitative limitations, imposed partly by historic sequencing capacity and cost issues, and more crucially by bacterial cloning constraints that affect which sequences are represented and how sequence-complete each clone is. Recently, dense whole-genome tiling microarrays have been developed and applied to transcriptomes for measur-

[1]Division of Biology, MC 156-29, California Institute of Technology, Pasadena, California 91125, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to B.W. (woldb@caltech.edu).

RECEIVED 2 MAY; ACCEPTED 27 MAY; PUBLISHED ONLINE 30 MAY 2008; DOI:10.1038/NMETH.1226


a
2× poly (A) selection
Add standards and shatter RNA
Make cDNA and sequence
Map 25-bp tags onto genome
Calculate transcript prevalence
2 RPKM
1 RPKM
1 RPKM
b
25-bp reads
1 kb
RNA-Seq graph
25-bp splices
Myf6
Conservation
Uniquely map.
RepeatMasker
c
RPKM
RPKM
Brain
0.0
0.0
Liver
0.0
0.0
Muscle
0.5
50.1
Muscle splices
Myf5
Myf6
Conservation
Uniquely mappable
Repeating elements by RepeatMasker
20 kb


a
Brain RNA
Muscle RNA
Enriched regions
Mef2d (UCSC)
Mef2d (RefSeq)
Uniquely mappable
α β γ δ
B
ε
40 kb
b
Brain RNA
Muscle RNA
B
α
β
δ
γ
1 kb
c
Observed splices
Preicted splices
10⁴
10³
100
10
10
100
1,000
10,000
d
Brain
(2,543)
Muscle
(1,391)
Liver
(1,250)
1,250
433
276
576
106
284
284

mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang[1,3], Catalin Barbacioru[2,3], Yangzhou Wang[2], Ellen Nordman[2], Clarence Lee[2], Nanlan Xu[2], Xiaohui Wang[2], John Bodeau[2], Brian B Tuch[2], Asim Siddiqui[2], Kaiqin Lao[2] & M Azim Surani[1]

Next-generation sequencing technology is a powerful tool for transcriptome analysis. However, under certain conditions, only a small amount of material is available, which requires more sensitive techniques that can preferably be used at the single-cell level. Here we describe a single-cell digital gene expression profiling assay. Using our mRNA-Seq assay with only a single mouse blastomere, we detected the expression of 75% (5,270) more genes than microarray techniques and identified 1,753 previously unknown splice junctions called by at least 5 reads. Moreover, 8–19% of the genes with multiple known transcript isoforms expressed at least two isoforms in the same blastomere or oocyte, which unambiguously demonstrated the complexity of the transcript variants at whole-genome scale in individual cells. Finally, for *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ (*Eif2c2*$^{-/-}$) oocytes, we found that 1,696 and 1,553 genes, respectively, were abnormally upregulated compared to wild-type controls, with 619 genes in common.

function[14,15]. Therefore, a more sensitive mRNA-Seq assay, ideally an assay capable of working at single cell resolution, is needed to meaningfully study crucial developmental processes and stem cell biology.

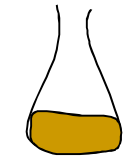
Here we modified a widely used single-cell whole-transcriptome amplification method to generate cDNAs as long as 3 kilobases (kb) efficiently and without bias[16,17]. With Applied Biosystems' next-generation sequencing SOLiD system, we found that it is feasible to get digital gene expression profiles at single-cell resolution. Using our mRNA-Seq assay with only a single mouse blastomere, we detected expression of 5,270 more genes than microarrays using hundreds of blastomeres. Using only a single blastomere, we also identified 1,753 previously unknown splice junctions, which have never been detected by microarrays at single-cell resolution. We found that hundreds of genes expressed two or more transcript variants in the same cell. We also found that in *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ mature oocytes, 1,696 and 1,553 genes, respectively, were abnormally upregulated, and 1,571 and 1,121 genes, respectively

[1]Wellcome Trust–Cancer Research UK Gurdon Institute of Cancer and Developmental Biology, University of Cambridge, Cambridge, UK. [2]Molecular Cell Biology Division, Applied Biosystems, Foster City, California, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to M.A.S. (as10021@mole.bio.cam.ac.uk) or K.L. (laokq@appliedbiosystems.com).

RECEIVED 2 DECEMBER 2008; ACCEPTED 2 MARCH 2009; PUBLISHED ONLINE 6 APRIL 2009; CORRECTED ONLINE 19 APRIL 2009 (DETAILS ONLINE); DOI:10.1038/NMETH.1315

Differential RNA-seq : dRNA-seq

Transcription Starting Sites (TSS)



totRNA

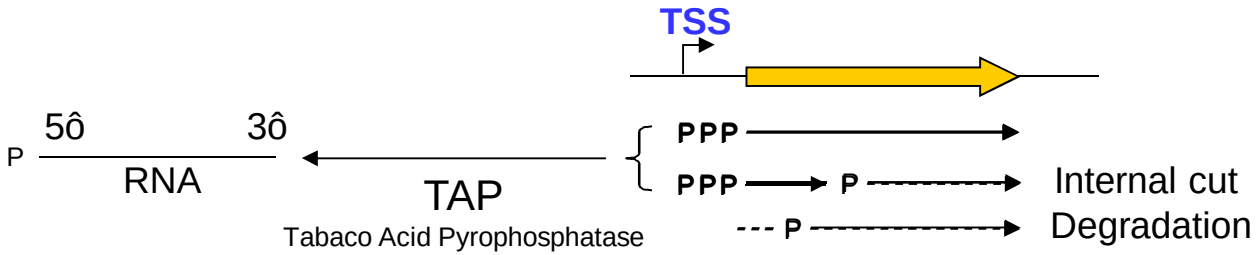
→ Discard of small RNAs (<200 nucl.) → Library (-)

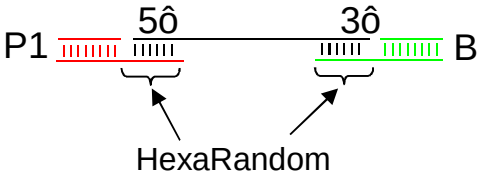
(RNA Clean XP, Agencourt)

↓

Digestion of non primary RNAs with

Terminator-5ô-phosphate-dependent exonuclease (Library (+))

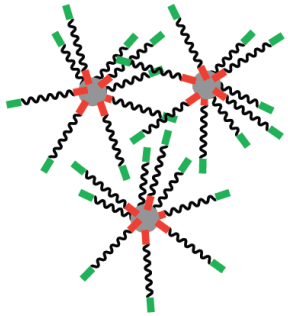




Ligation

Denaturation

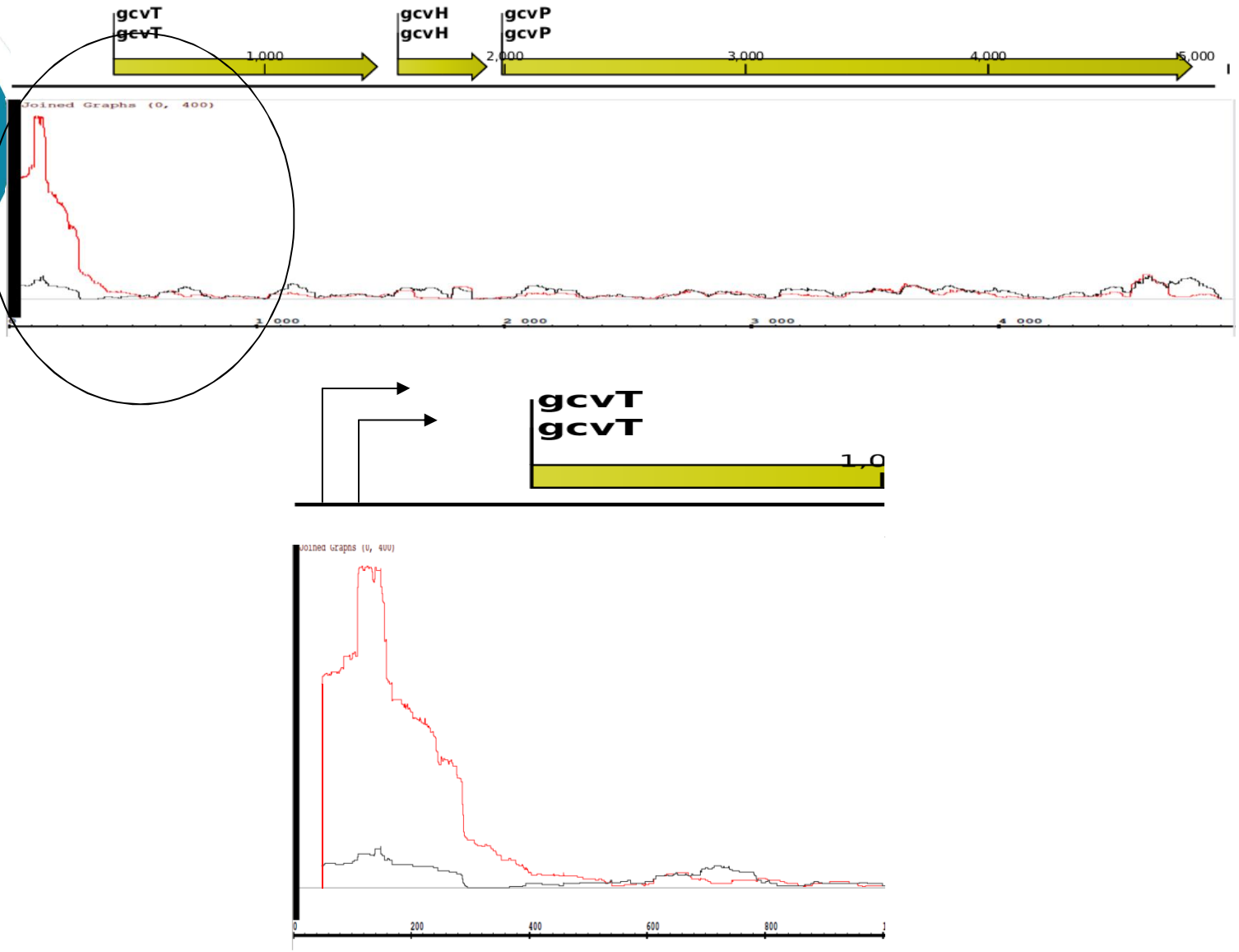
RT-PCR



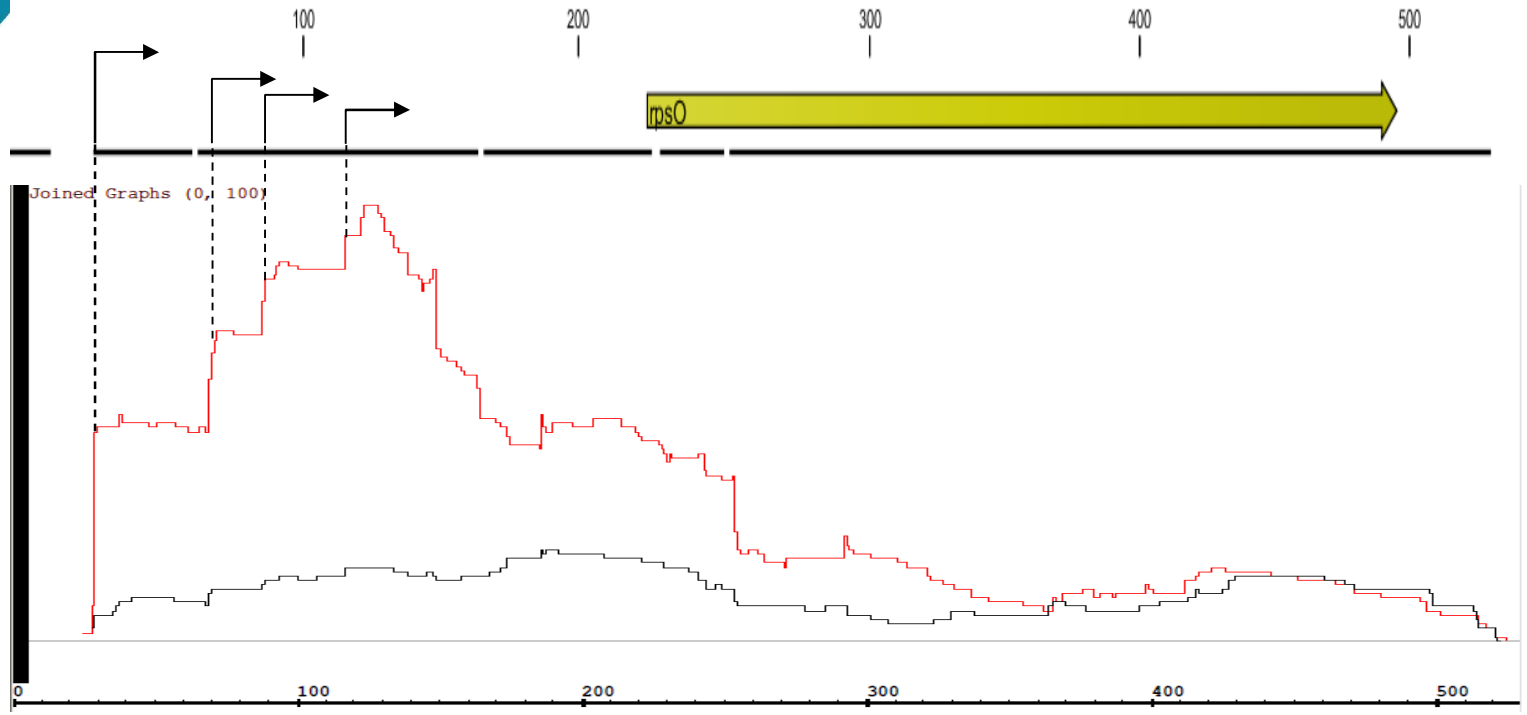
emPCR

PGM
Ion Torrent

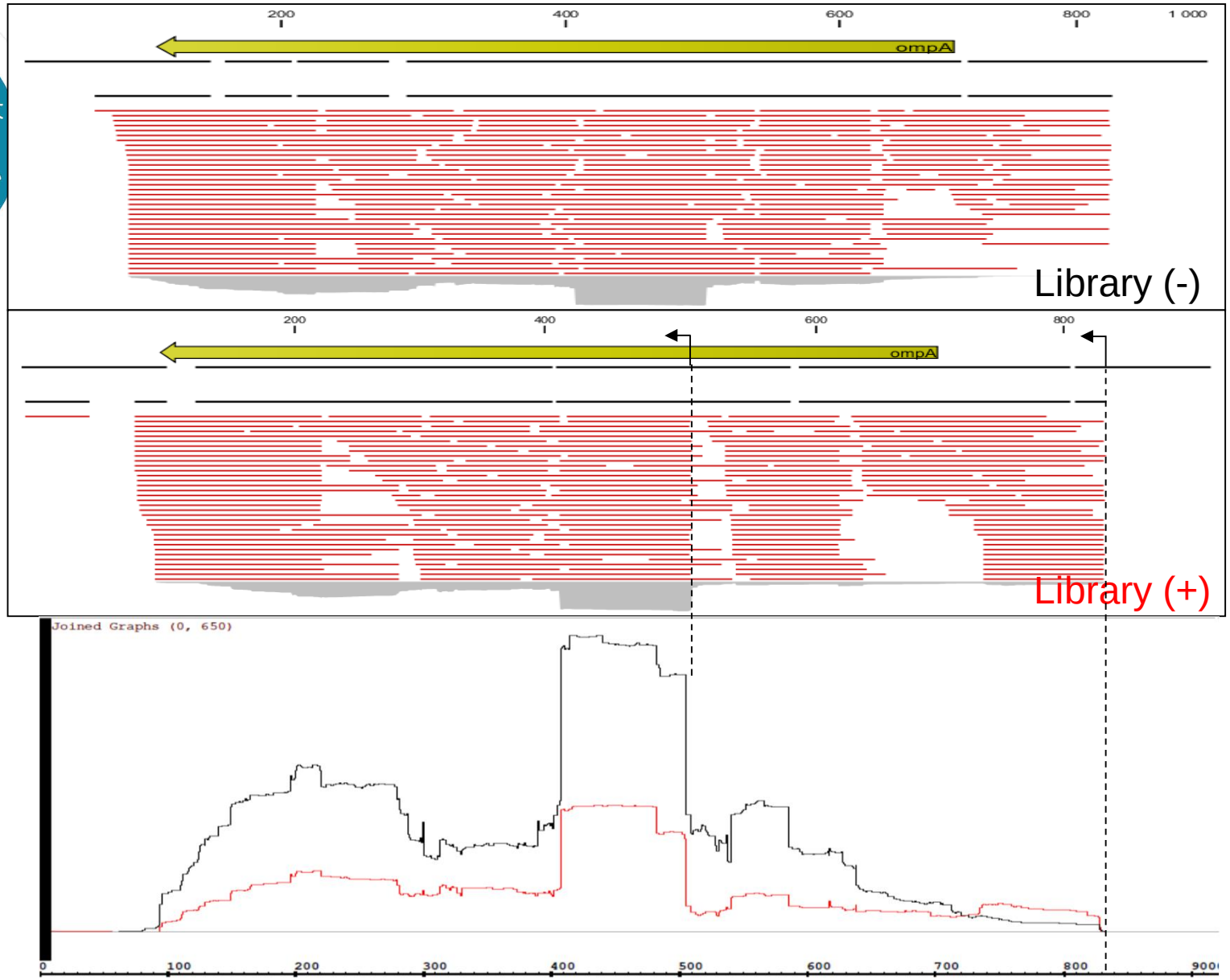
Primary TSS - Secondary TSS



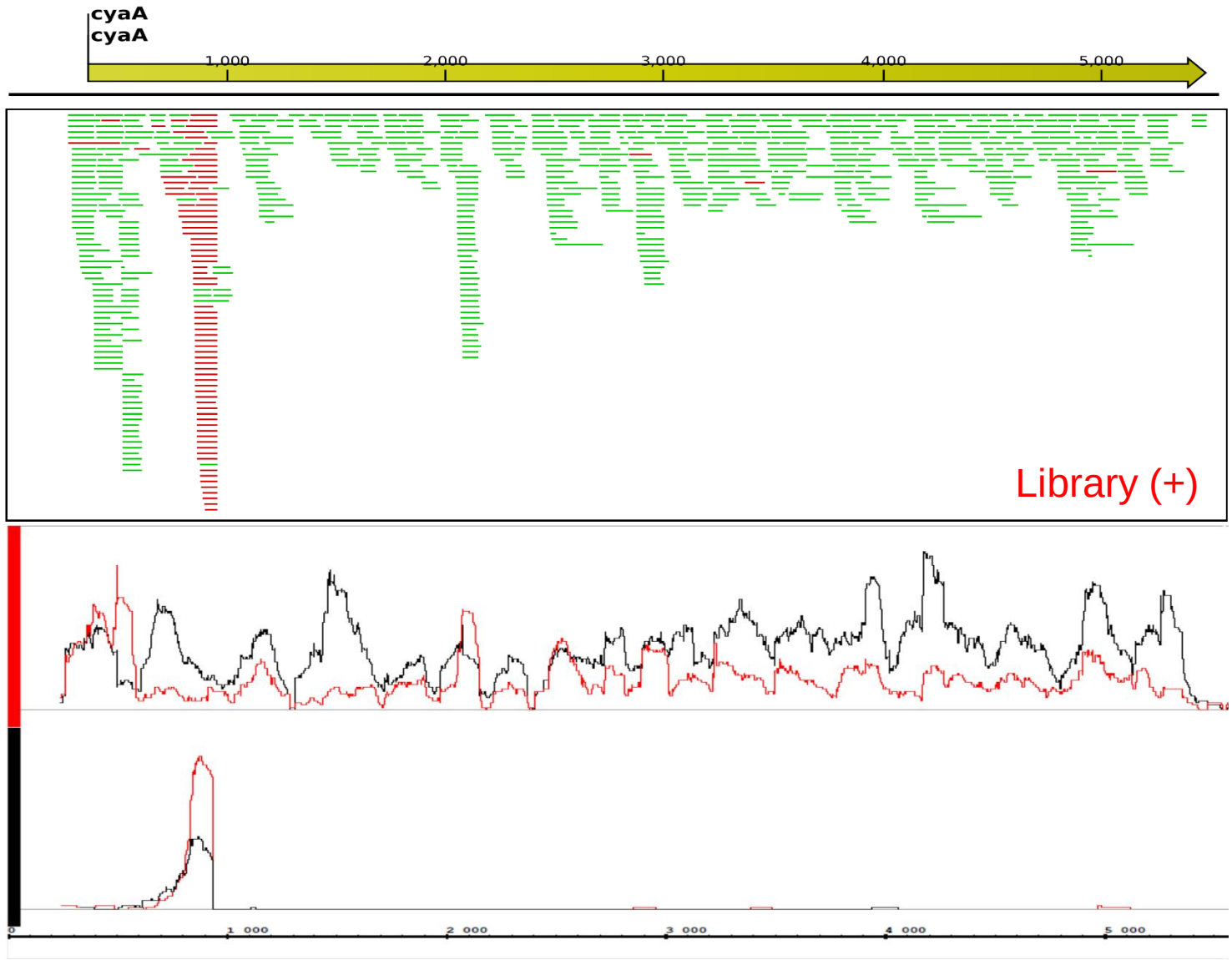
Primary TSS - Secondary TSS



Internal TSS



Antisens TSS



Next generation sequencing
ChIP-seq

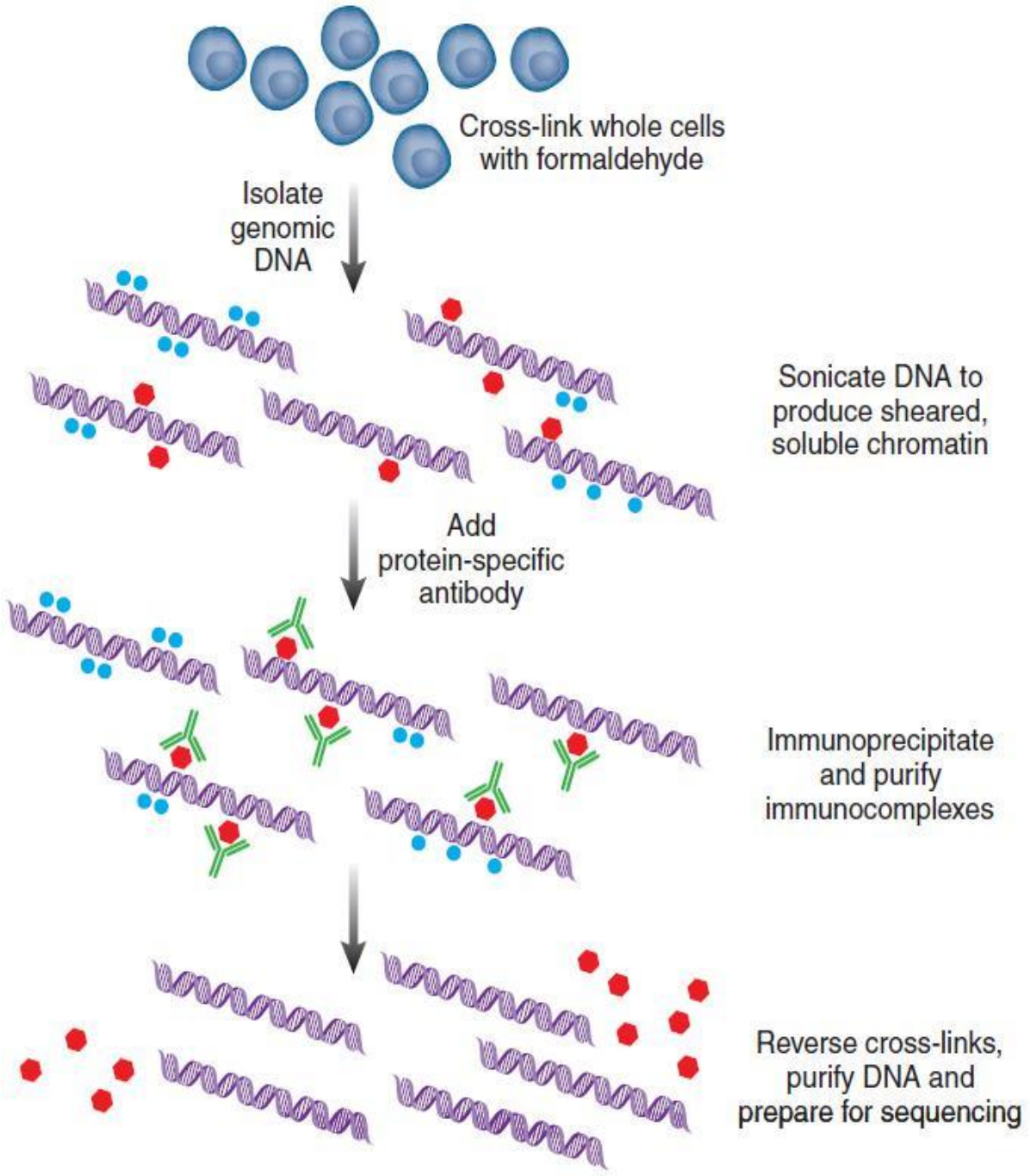
Science
AAAS

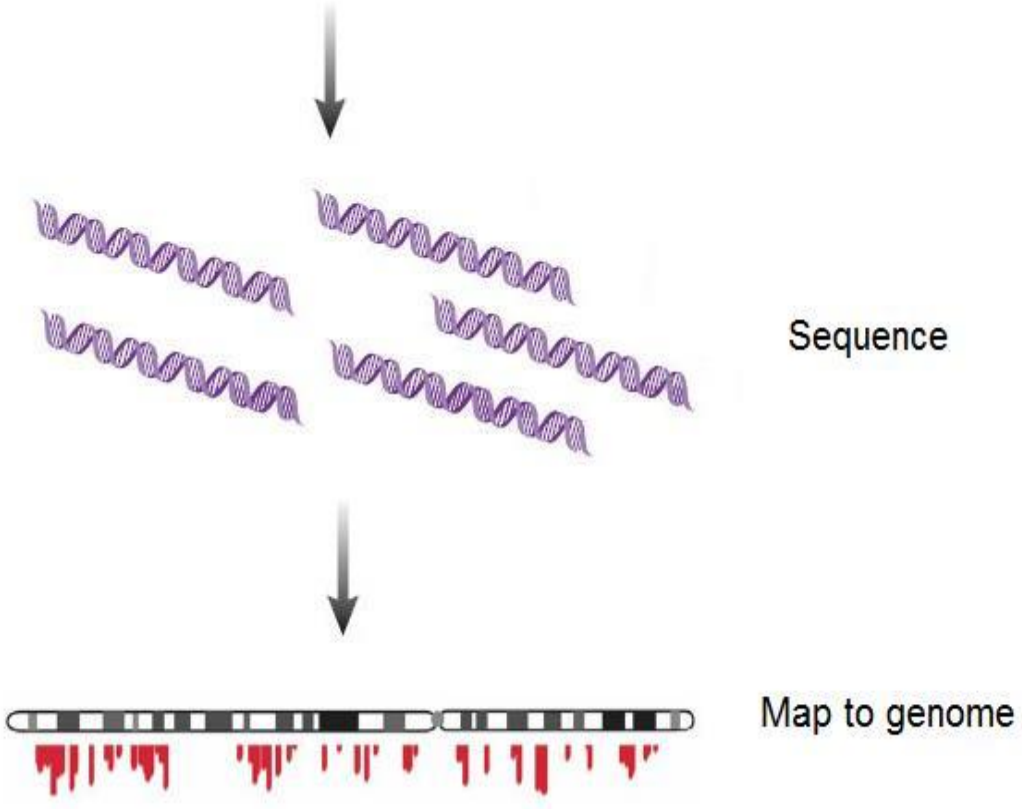
Genome-Wide Mapping of in Vivo Protein-DNA Interactions
David S. Johnson, *et al.*
Science **316**, 1497 (2007);
DOI: 10.1126/science.1141319

Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson,[1]* Ali Mortazavi,[2]* Richard M. Myers,[1]† Barbara Wold[2,3]†

In vivo protein-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these protein-DNA interactions comprehensively across entire mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIPSeq) based on direct ultrahigh-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element–1 silencing transcription factor) to 1946 locations in the human genome. The data display sharp resolution of binding position [±50 base pairs (bp)], which facilitated our finding motifs and allowed us to identify noncanonical NRSF-binding motifs. These ChIPSeq data also have high sensitivity and specificity [ROC (receiver operator characteristic) area ≥ 0.96] and statistical confidence ($P < 10^{-4}$), properties that were important for inferring new candidate interactions. These include key transcription factors in the gene network that regulates pancreatic islet cell development.


Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Add protein-specific antibody
Immunoprecipitate and purify immunocomplexes
Reverse cross-links, purify DNA and prepare for sequencing

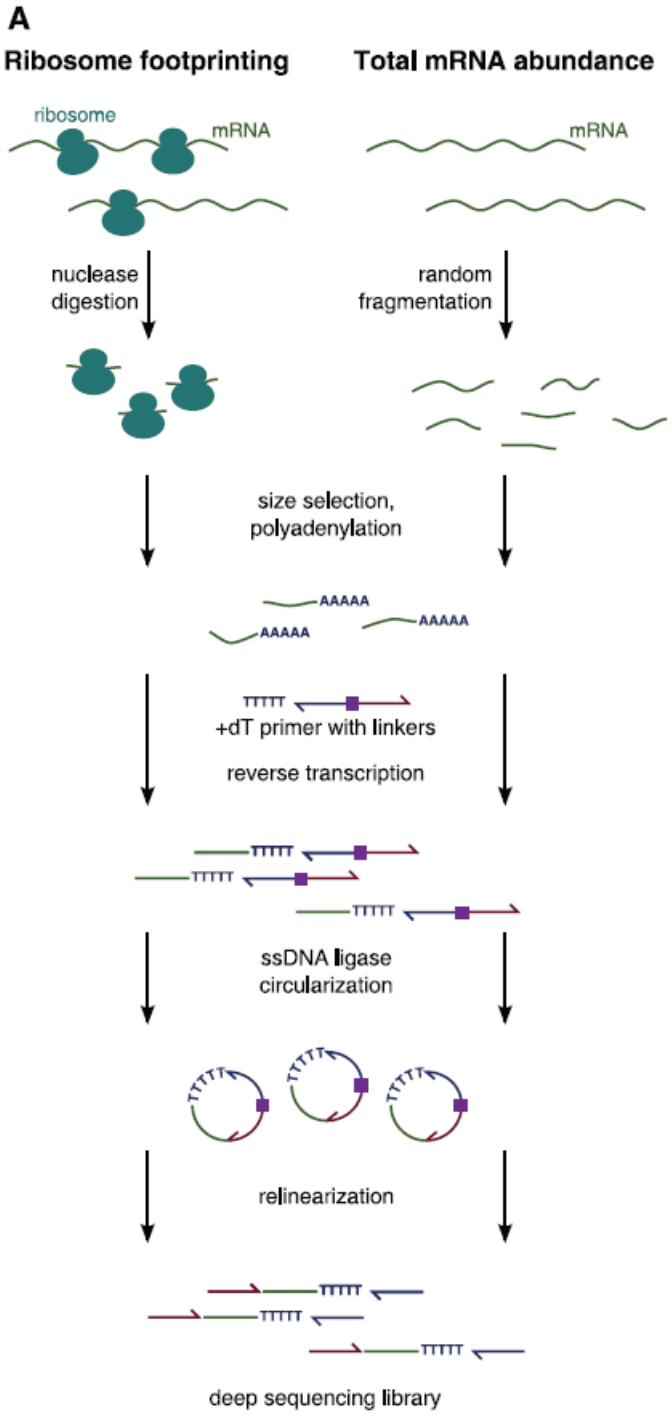

Sequence
Map to genome

Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling

Nicholas T. Ingolia,* Sina Ghaemmaghami,† John R. S. Newman, Jonathan S. Weissman

Techniques for systematically monitoring protein translation have lagged far behind methods for measuring messenger RNA (mRNA) levels. Here, we present a ribosome-profiling strategy that is based on the deep sequencing of ribosome-protected mRNA fragments and enables genome-wide investigation of translation with subcodon resolution. We used this technique to monitor translation in budding yeast under both rich and starvation conditions. These studies defined the protein sequences being translated and found extensive translational control in both determining absolute protein abundance and responding to environmental stress. We also observed distinct phases during translation that involve a large decrease in ribosome density going from early to late peptide elongation as well as widespread regulated initiation at non–adenine-uracil-guanine (AUG) codons. Ribosome profiling is readily adaptable to other organisms, making high-precision investigation of protein translation experimentally accessible.

Fig. 1. Quantifying mRNA abundance and ribosome footprints by means of deep sequencing. (**A**) Schematic of the protocol for converting ribosome footprints or randomly fragmented mRNA into a deep-sequencing library. (**B**) Internal reproducibility of mRNA-abundance measurements. CDSs were conceptually divided as shown, and the mRNA counts on the two regions are plotted. The error estimate is based on the χ^2 statistic.



Institut Pasteur de Lille

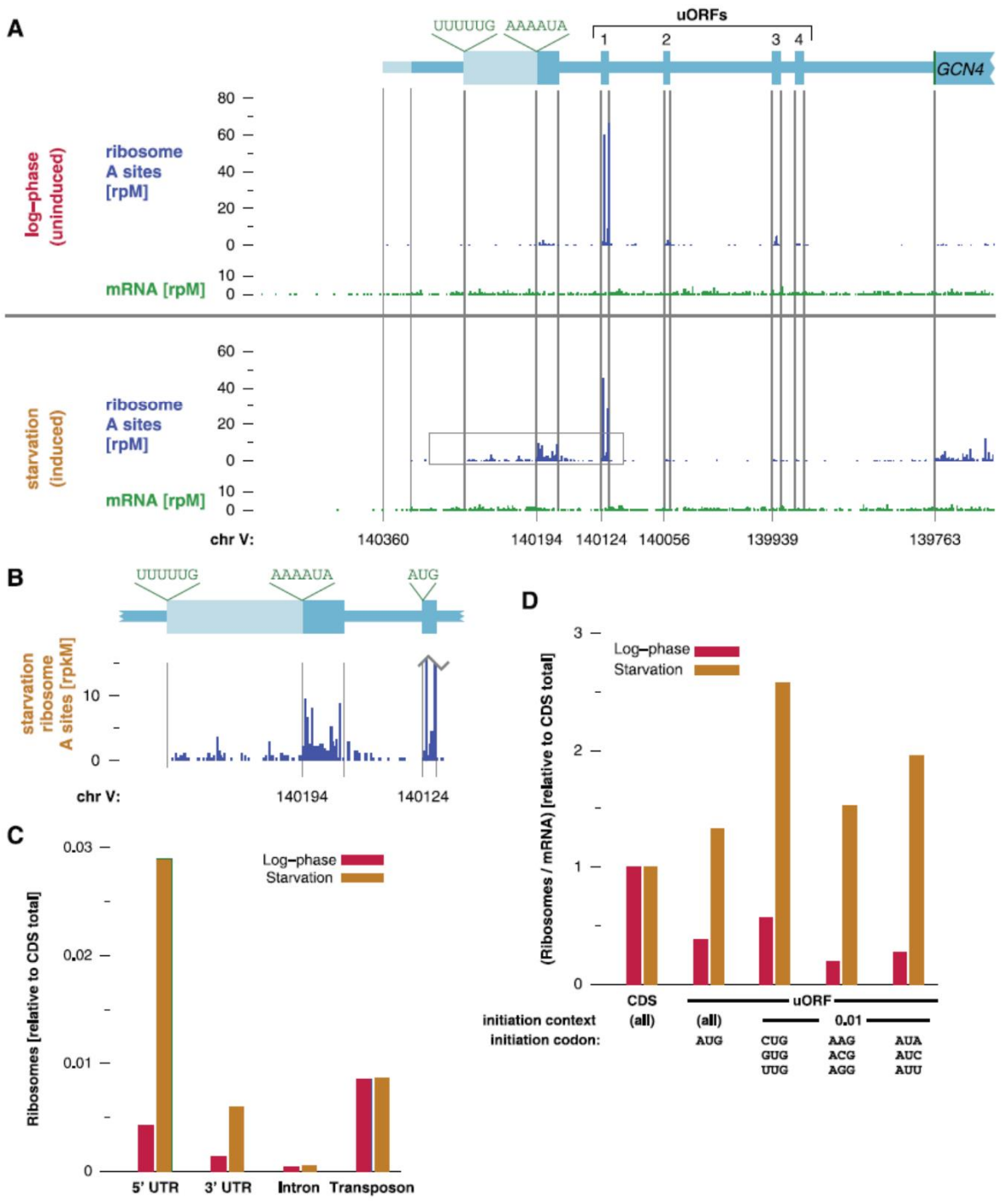


Fig. 5. Changes in 5′UTR translation during starvation. (**A**) Ribosome and mRNA densities in the *GCN4* 5′UTR in repressive and inducing conditions. The four known uORFs are indicated along with the proposed initiation sites for upstream translation. (**B**) Non-AUG uORF upstream of *GCN4*. Shown is an enlargement of the gray boxed area in (A). (**C**) Ribosome occupancy of noncoding sequences. The number of ribosome footprints mapping to different classes of regions is shown relative to the number of CDS reads. (**D**) Aggregate translational efficiency of uORFs (*14*).

Institut Pasteur de Lille

PNAS

Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore

David Stoddart, Andrew J. Heron, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley[1]

Department of Chemistry, University of Oxford, Oxford OX1 3TA, United Kingdom

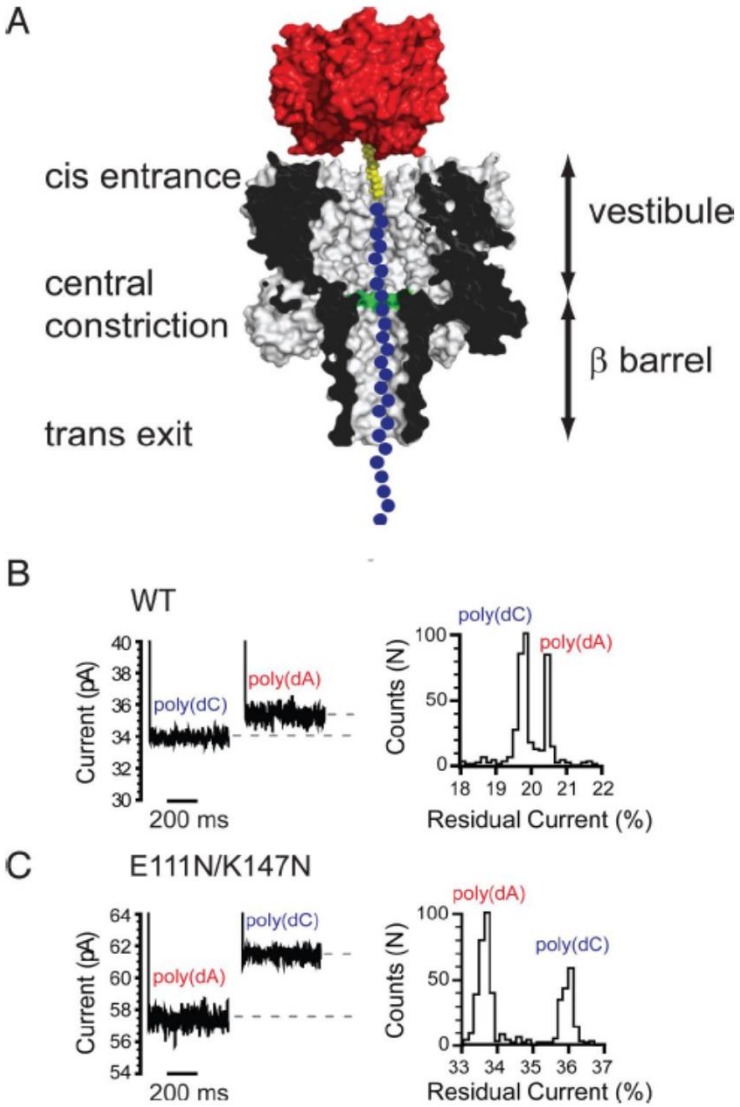
Edited by Daniel Branton, Harvard University, Cambridge, MA, and approved March 11, 2009 (received for review January 30, 2009)

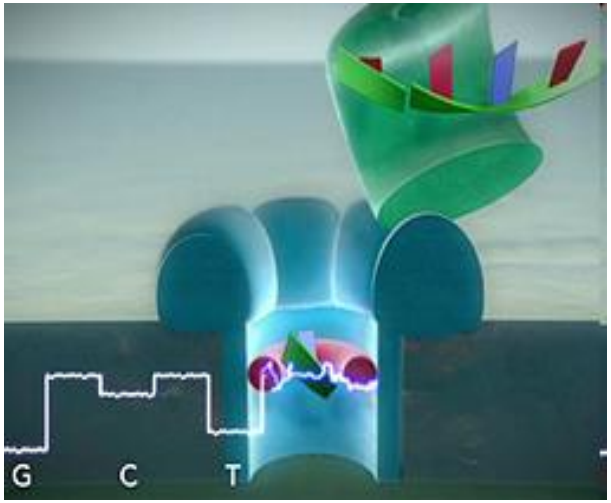
The sequencing of individual DNA strands with nanopores is under investigation as a rapid, low-cost platform in which bases are identified in order as the DNA strand is transported through a pore under an electrical potential. Although the preparation of solid-state nanopores is improving, biological nanopores, such as α-hemolysin (αHL), are advantageous because they can be precisely manipulated by genetic modification. Here, we show that the transmembrane β-barrel of an engineered αHL pore contains 3 recognition sites that can be used to identify all 4 DNA bases in an immobilized single-stranded DNA molecule, whether they are located in an otherwise homopolymeric DNA strand or in a heteropolymeric strand. The additional steps required to enable nanopore DNA sequencing are outlined.

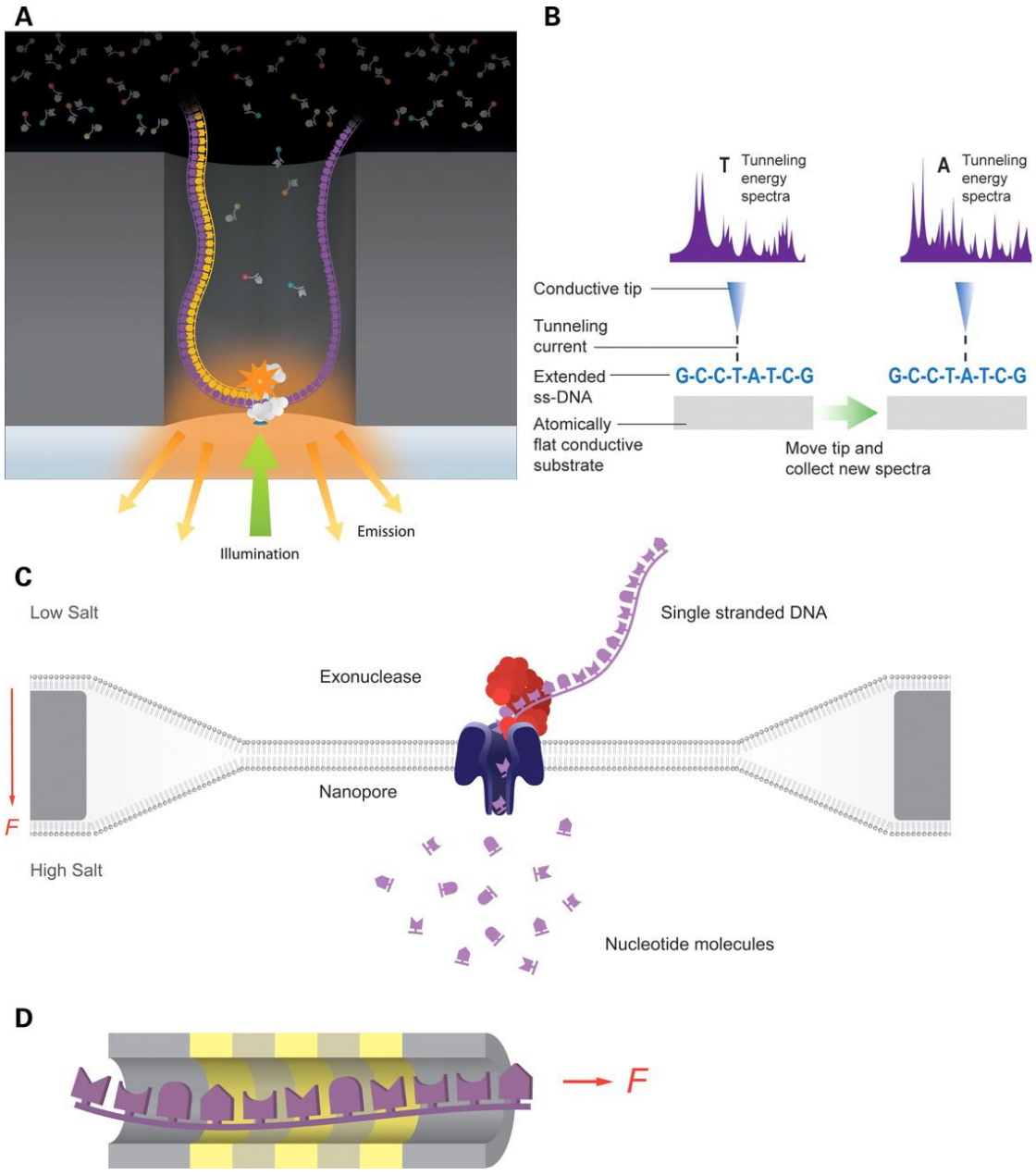
α-hemolysin | DNA sequencing | genomics | protein engineering | protein pore

Wang and colleagues first showed that the direction in which DNA enters the αHL pore (5′ or 3′ threading) affects the extent of current block (15), an observation supported by the data of Mathé et al. (16), and recent studies have shown that the rate of capture of DNA by protein pores is enhanced when the interior surfaces bear a net positive charge (17, 18). Under the high applied potentials required for threading, freely moving DNA is translocated through the wild-type (WT) αHL pore too quickly for bases to be identified, unless the bases are modified with bulky groups (19). In a step toward the management of this problem, the Ghadiri group have shown that ssDNA can be ratcheted one base at a time through the αHL pore by the action of a DNA polymerase (20). In the present work, we return to the problem of base identification and show that all 4 DNA bases can be distinguished in both homopolymeric and heteropolymeric immobilized DNA strands.

PHYSICAL SCIENCES


A
cis entrance
central
constriction
trans exit
vestibule
β barrel
B
WT
Current (pA)
poly(dC)
poly(dA)
200 ms
Counts (N)
Residual Current (%)
C
E111N/K147N
Current (pA)
poly(dA)
poly(dC)
200 ms
Counts (N)
Residual Current (%)


G
C
T



Human Molecular Genetics

Merci

TAG
Yves Lemoine
David Hot
Ségolène Caboche
Ludovic Huot
Renaud Blervaque
Stéphanie Slupek

Gènes Diffusion
Christophe Audebert
(c.audebert@genesdiffusion.com)
Gaël Even
Sophie Merlin

Adresse :
Institut Pasteur de Lille
1 rue prof. Calmette, LILLE
david.hot@pasteur-lille.fr
Tél. : 03 20 87 72 09

http://www.biorigami.com/

http://www.ciil.fr

Fondation reconnue d'utilité publique