

# Séquençage nouvelle génération et maladies infectieuses

# La genopole Institut Pasteur

- Quatre équipes convergeant sur le NGS (HiSEQ Illumina)

	Hier	Aujourd'hui
Génomique	Séquençage Sanger	NGS Illumina
Transcriptome et épigénome	Puces à ADN	NGS Illumina
Génotypage des pathogènes et santé publique	MLST Séquençage Sanger Genescan	NGS Illumina
Génotypage et génétique	Puces Infinium Illumina	Séquençage d'Exome

# Etude des maladies infectieuses côté hôte / côté microbe

- Identification de fonctions spécifiques (facteurs de virulence, résistance)
- Etude de la diversité et de l'évolution des organismes pathogènes
- Expression génétique et régulation
- Découverte et caractérisation de nouveaux agents pathogènes
- Etude des communautés microbiennes
- Réponse transcriptionnelle de l'hôte à l'infection
- Résistance aux agents pathogènes (miRNA)
- Génétique de la prédisposition aux maladies infectieuses

# Capacité des séquenceurs: possibilité / nécessité de multiplexer

- HiSEQ 2000: 90 M de *clusters* par ligne
  - 100 bases single reads: 9 Gb

## Génomique

- 2000 *Escherichia coli*
  - 4000 *Streptococcus agalactiae*
- ⇒ Multiplexé par 12: > 160 x de couverture pour *E. coli*

## Transcriptomique (30% ARNm: 30 M de lectures)

- 5000 gènes: en moyenne 6000 lectures / gène

# Logiciels d'analyse

- Recherches de SNPs : MAQ, BWA, Bowtie, CLC
- Assemblage de novo: Velvet, CLC
- Visualisation : Tablet, Artemis, Mapview

Recherche de gènes *d'intérêt*  
Résistance - Virulence

# Resistance aux antibiotiques

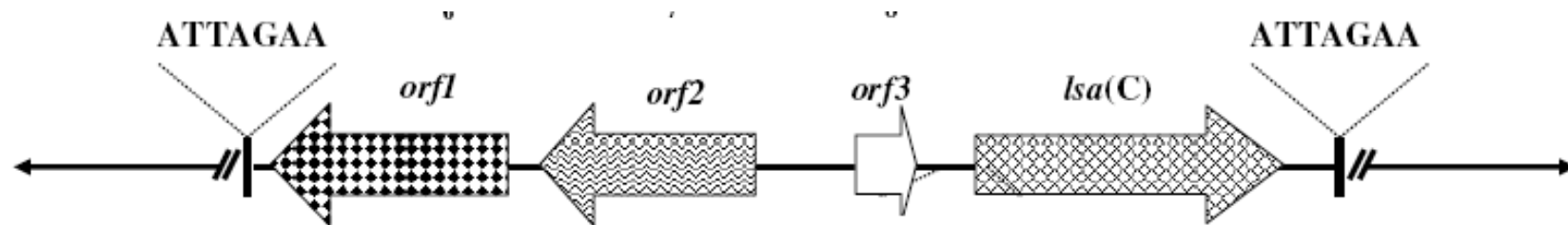
Q. Souches de *Streptococcus agalactiae* de Nouvelles Zélande avec un phénotype inhabituel de résistance aux lincosamides et aux streptogramines A

- Le mécanisme est inconnu.
- Possibilité de transfert horizontal de la résistance?

**Objectif:** identifier le ou les déterminants de résistance, identifier le support génétique (îlot génétique?).

# Méthodes - résultats

- Analyse de 18 isolats de différents sérotypes ayant les mêmes caractéristiques.
- Séquençage 454 d'un isolat représentatif
- Identification de gènes / îlots spécifiques à cette souche



- Lsa(C): transporteur ABC faiblement similaire à d'autres gènes Lsa
- Confère la résistance à la lincomycine et à la clindamycine
- Orf1: integrase
- Orf2: protéine de réplication
- Orf3: régulateur transcriptionnel



# Conclusions

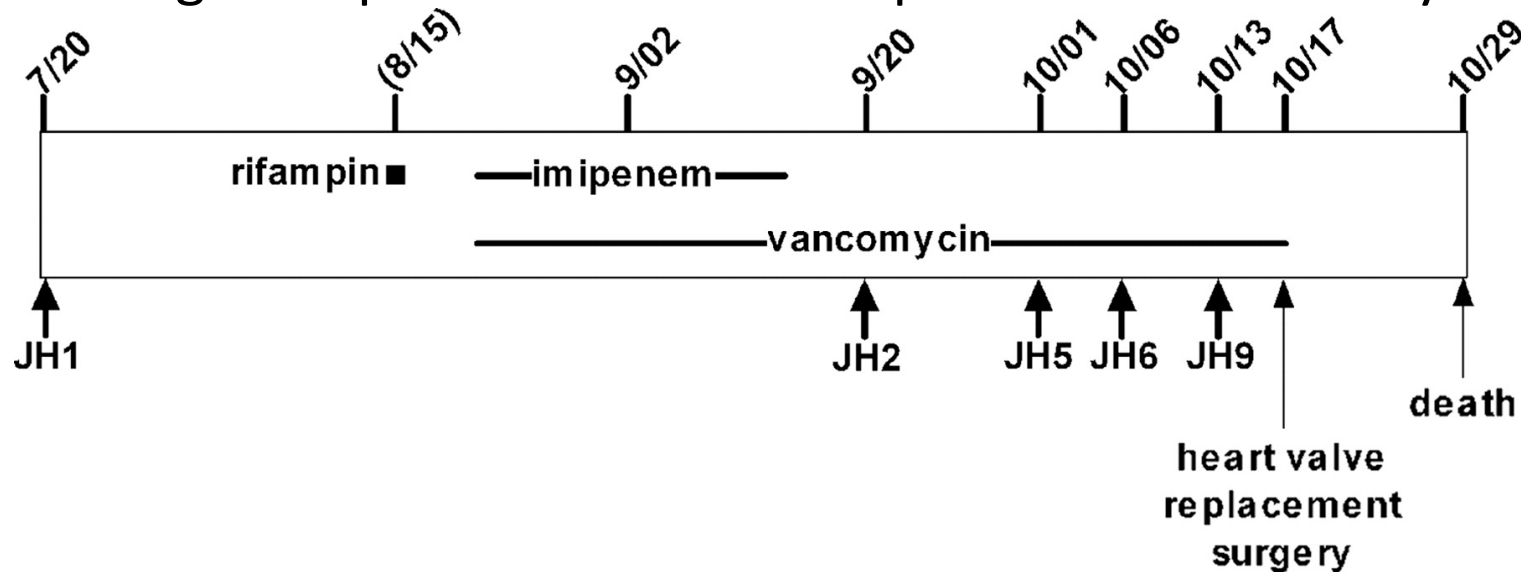
- Identification du gène conférant la résistance aux lincosamides
- Nouveau membre de la famille des transporteurs ABC Lsa
- Porté par un îlot génomique (mobile?)
- Le même élément a été retrouvé pour les 18 souches

# Recherche de mutations ponctuelles

## Etude de l'évolution

# Evolution de *S. aureus* au cours d'une infection (Mwangi et al. PNAS 2007)

- ⇒ Acquisition de la résistance (*in vivo*) au cours de l'infection
- ⇒ Séquençage d'isolats prélevés sur 3 mois
- ⇒ Base génétique de la faible susceptibilité à la vancomycine



Date (mois/jour) de prélèvement des souches analysées, traitements antibiotiques

⇒ Uniquement 35 mutations dans 31 gènes (25 substitutions – 6 synonymes, 1 insertion, 9 délétions)

# Emergence séquentielle des mutations

	Van	Rif	Oxa	Dap	1	2	3-6	7	8	9	10	11-18	19-33	34-35
<b>JH1</b>	1.0	0,012	0.75	0.01	+									?
<b>JH2</b>	4.0	16	25	0.05	-	+	+	+	+					?
<b>JH5</b>	6.0	16	0.75	0.05	-	+	+	+	+	+				?
<b>JH6</b>	8.0	16	1.5	1.0	-	+	+	+	+	-	+	+		?
<b>JH9</b>	8.0	16	0.75	1.0	-	+	+	+	+	+	+	+	+	?

1: *blaR1* (regulation de *blaZ* ( $\beta$ -lactamase))

2: *vraR* regulator of cell wall synthesis (vancomycine résistance)

3-6: *rpoB* Rifampicin résistance

7: *rpoC*

9: SA1249 fonction inconnue

10: *agrC*, système de régulation (vancomycine et daptomycine résistance)

11: *yycH*, système de régulation essentiel (Vanco. et Daptomycine résistance)

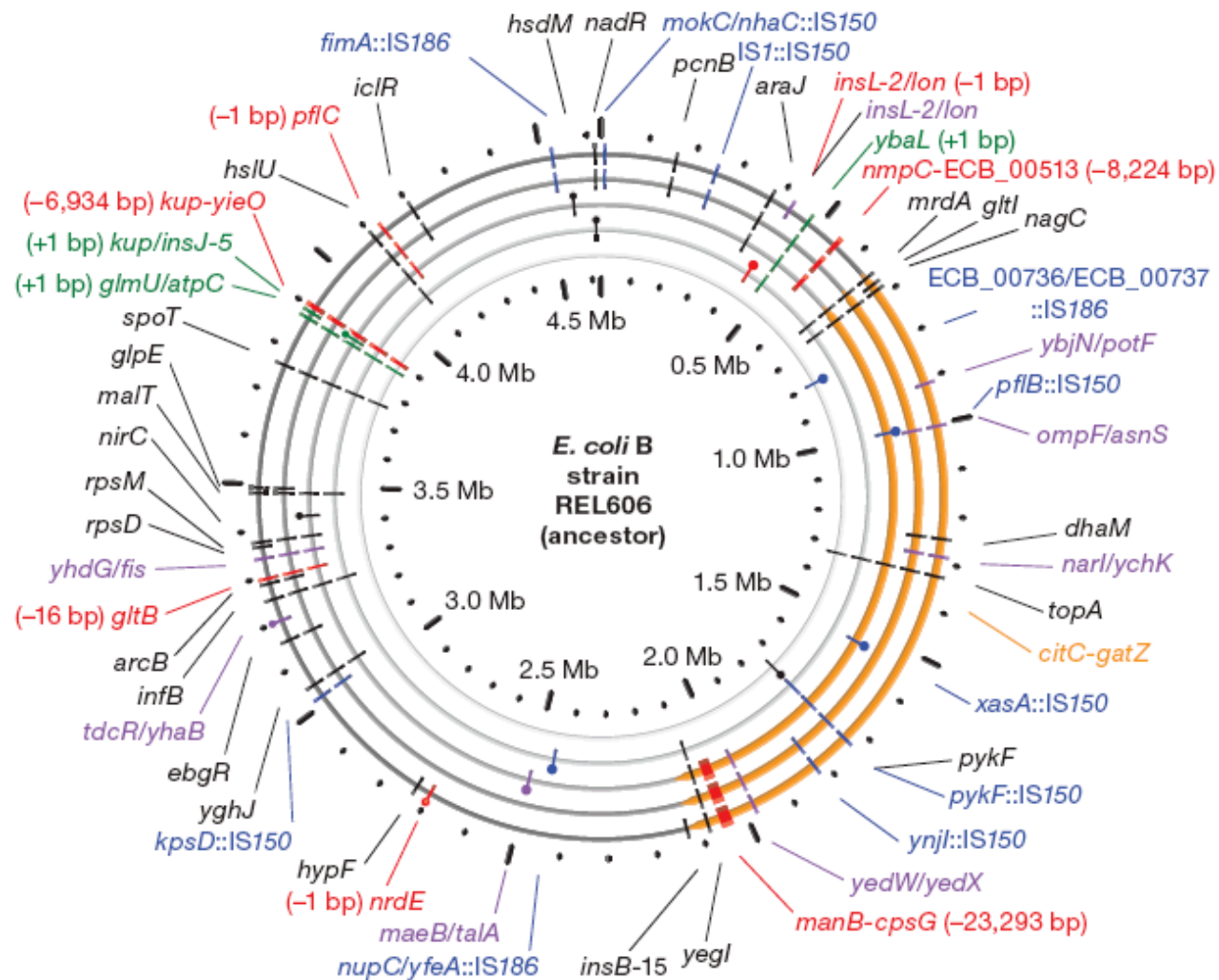
14: *isdE*, 15: *prsA* (frame shift)

## Evolution de *S. aureus in vivo*: conclusions

- Les souches bactériennes évoluent au cours de l'infection.
- Le séquençage permet de comprendre l'acquisition de la résistance et l'échappement aux défenses de l'hôte.
- La souche évoluée a un fitness réduit

# Identification de mutations – Expérience de Lenski

- Principe: évolution de souches d'*E. coli* dans des conditions suboptimales (MM, glucose 25mg/l) sur 40000 générations (20 ans)
  - Augmentation de la vitesse de croissance par accumulation de mutations
  - Identification de certaines mutations. Apparition de phénotype mutateur
- => Séquençage d'isolats après 2 000, 5 000, 10 000, 15 000, 20 000 et 40 000 générations
- => Après 20 000 générations: 29 SNPs (toutes non synonymes) et 16 indels



Clone sequenced from generation...

□ 2K	■ 5K	■ 10K	■ 15K	■ 20K
------	------	-------	-------	-------

Evolved mutations... ⤴ Off line of descent to 40K clone

SNP	■ Protein coding	DIP	■ Deletion	■ IS insertion
■ Intergenic	■ Insertion	■ Inversion		

# Conclusions

- L'adaptation est rapide - gain de fitness diminue au cours du temps
- L'évolution est constante sur 20 000 générations
- Mais pratiquement toutes les mutations apportent un bénéfice (non synonyme)
- Remplacement d'un clone par un clone portant une mutation sélectionnée

⇒ Séquençage d'un plus grand nombre d'isolats



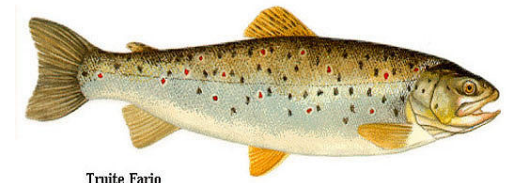
Etude évolutive et des populations  
Analyse génomique de collections de souches

# *Streptococcus agalactiae*

- Pathogène opportuniste
- Chez l'homme
  - Commensal des tractus digestif et urinaire
  - Cause principale d'infections néonatales



- Spectre large d'hôtes animaux
  - Mammites chez les bovins et les camélidés
  - Infection invasive chez les poissons



Truite Fario



## Séquençage de sept isolats

- 2-22 [ST261](#) *S. difficile* strain, tilapia Israel 1984
- SS1218 [ST261](#) frog, Louisiana
- SS1219 [ST260](#) frog, Brazil 1982
- 90-503 [ST260](#) hybrid striped bass, Louisiana (1990)
- 05-108A [ST260](#) Tilapia Honduras (2005)
  
- CF01173 [ST7](#) Trout farm England
- SS1014 [ST7](#) Fish USA

# Les souches ST7

Nb de SNP

comparaison avec  
les souches  
humaines A909 et  
H36B

	A909	H36B	CF01173 UK	SS1014 USA
A909		3835	<b>389</b>	3484
H36B			3625	<b>689</b>
SS1173				<b>3514</b>
SS1014				

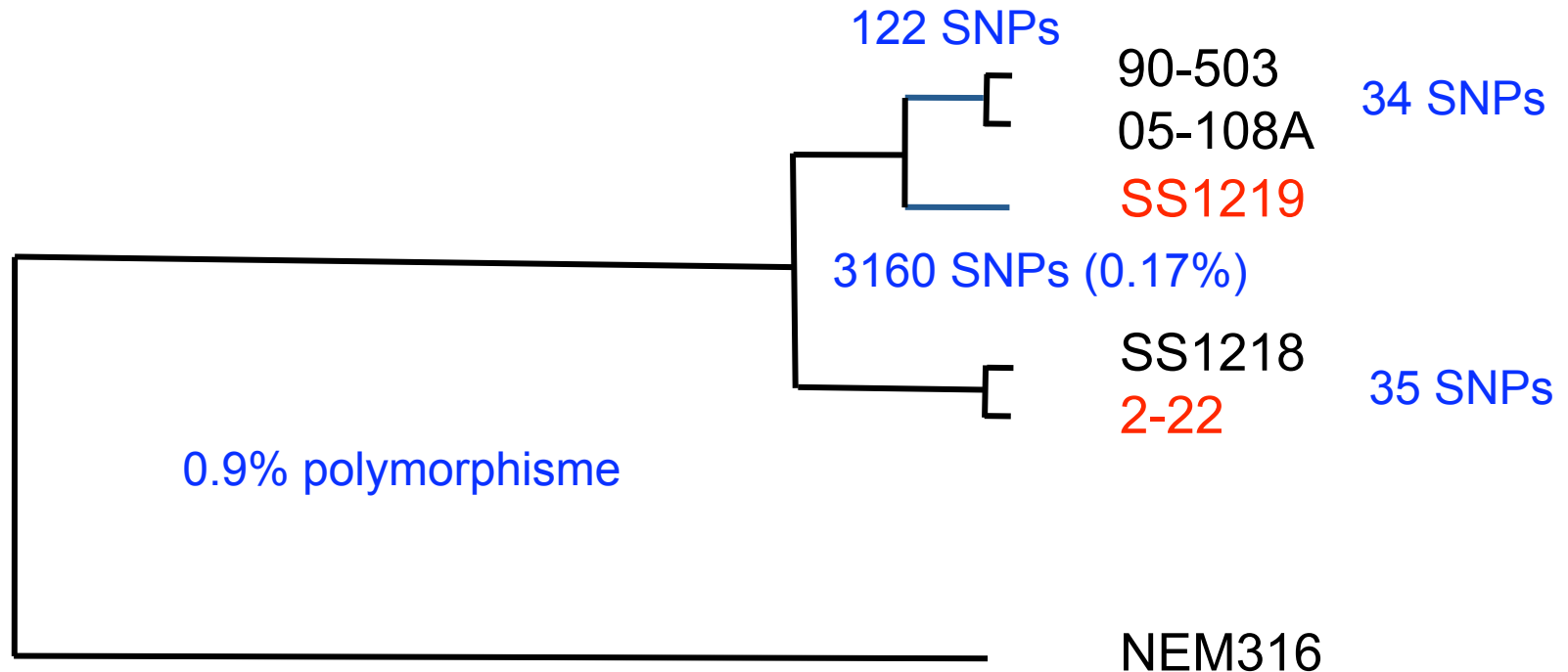
⇒ Les souches ST7 de poisson sont proches des souches humaines

⇒ Les deux souches se sont adaptées indépendamment à l'hôte poisson

# CC 260 un lignage spécifiquement pathogène de poisson

- Phénotype spécifique: thermosensible, capacité réduite à utiliser des sources de C., hydrophobicité de surface
- Très pathogène et transmissible: provoque des épidémies dans les élevages; DL50 = 100 bactéries
- *Streptococcus agalactiae* de sérotype 1b
- Isolée dans différentes parties du monde : Israël, USA, Brésil (eaux chaudes)

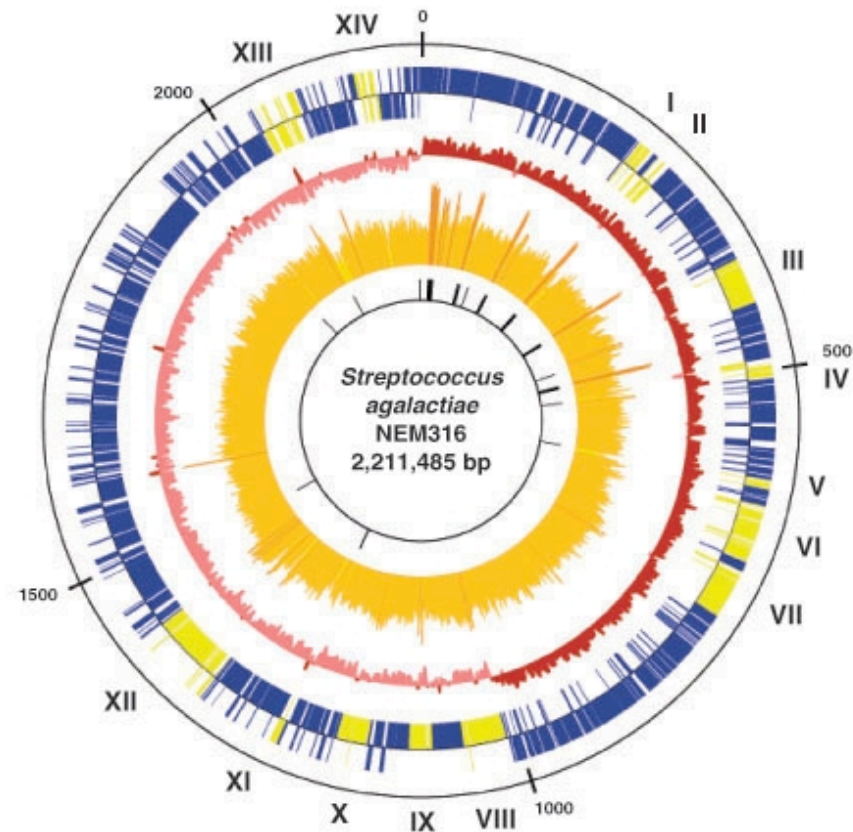
# Les souches isolées de poissons correspondent à un lignage spécifique



⇒ Expansion récente des groupes 2-22 – 90503

⇒ Ce lignage est ancien et distant des souches humaines

Taille: 2-22: 1834 kb - ss1219: 1845 kb

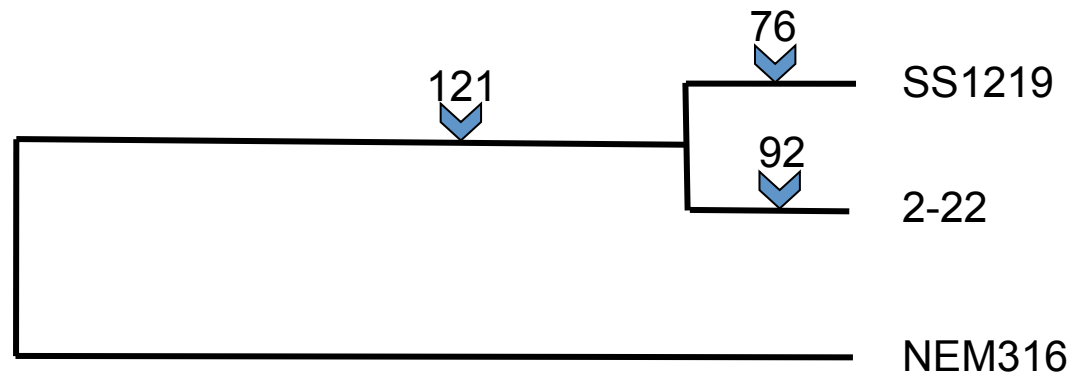


- ⇒ Perte de pratiquement tous les îlots
- ⇒ Nombreuses délétions et pseudogènes



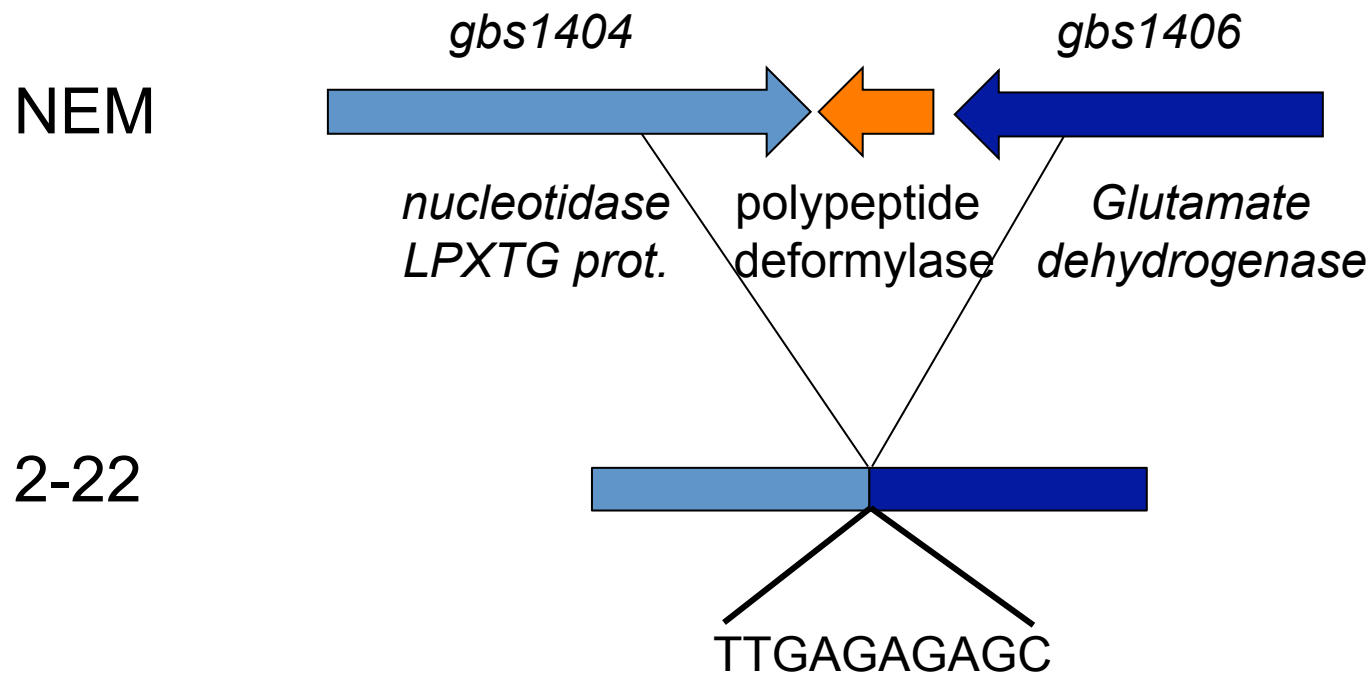
# Perte de gènes

- Correspond au génome core des souches humaines: c.a. 1800 gènes
- 213 pseudogènes dans la souche 2-22
- 197 pseudogènes dans la souche SS1219
- 121 gènes sont inactivés dans les deux souches



- ⇒ Touchent des fonctions métaboliques (source de C), l'énergétique, les composants de surface, les facteurs de virulence et des régulateurs.
- ⇒ Des gènes d'un même opéron peuvent être inactivés différemment dans les deux lignages.
- ⇒ Processus en cours

# Mécanisme de délétion



⇒ Probablement RecA indépendant, mais résultant de glissement de la polymérase



# Réduction en cours de ces gènes

```
Query: 1 MNEIKCPHCCTAFAINSEYHQLLEQIRGDAFDKEVSRLEKERLILGEQAKNQLQEVVV 60 MNEIKCPHCCTAFAINSEYHQLLEQIRGDAFDKEVSRLEKERLILGEQAKNQLQEVVV 60
Sbjct: 43355 MNEIKCPHCCTAFAINSEYHQLLEQIRGDAFDKEVSRLEKERLILGEQAKNQLQEVVV 43176 MNEIKCPHCCTAFAINSEYHQLLEQIRGDAFDKEVSRLEKERLILGEQAKNQLQEVVV 574456

Query: 61 EKDKETIAKLQYKVKQFLIEKDNLLKDNEYQLAEQLNPKDMMLRDLENQIDRLRLEHENS 120 EKDKETIAKLQYKVKQFLIEKDNLLKDNEYQLAEQLNPKDMMLRDLENQIDRLRLEHENS 120
Sbjct: 43175 EKDKETIAKLQYKVKQFLIEKDNLLKDNEYQLAEQLNPKDMMLRDLENQIDRLRLEHENS 43050 EKDKETIAKLQYKVKQFLIEKDNLLKDNEYQLAEQLNPKDMMLRDLENQIDRLRLEHENS 574276

Query: 121 QEALTKVERERDAIQNLHIQEKEKDLALASVKSDYEVQLKAANEQVEFYKNFKAQQSTK 180 QEALTKVERERDAIQNLHIQEKEKDLALASVKSDYEVQLKAANEQVEFYKNFKAQQSTK 180
Sbjct: 43049 QEALTKVERERDAIQNLHIQEKEKDLALASVKSDYEVQLKAANEQVEFYKNFKAQQSTK 42876 QEALTKVERERDAIQNLHIQEKEKDLALASVKSDYEVQLKAANEQVEFYKNFKAQQSTK 574102

Query: 181 AVGESLEHYAETEFNKVRHLAFPNAYFEKDNTLSSRGSKGDFIYREKDNLEFLSIMFE 240 AVGESLEHYAETEFNKVRHLAFPNAYFEKDNTLSSRGSKGDFIYREKDNLEFLSIMFE 240
Sbjct: 42875 AVGESLEHYAETEFNKVRHLAFPNAYFEKDNTLSSRGSKGDFIYREKDNLEFLSIMFE 42696 AVGESLEHYAETEFNKVRHLAFPNAYFEKDNTLSSRGSKGDFIYREKDNLEFLSIMFE 573922

Query: 241 MKNESDDTIKHKHNEFFKELDKDRREKSCEYAVLVTMLEADNDYNTGIVDVSHKYPKM 300 MKNESDDTIKHKHNEFFKELDKDRREKSCEYAVLVTMLEADNDYNTGIVDVSHKYPKM 300
Sbjct: 42695 MKNESDDTIKHKHNEFFKELDKDRREKSCEYAVLVTMLEADNDYNTGIVDVSHKYPKM 42516 MKNESDDTIKHKHNEFFKELDKDRREKSCEYAVLVTMLEADNDYNTGIVDVSHKYPKM 573748

Query: 301 YVIRPQFFIQLIGILRNAALNTLKYQELALMKEQNIDITHFEEDLDIFKNAFAKNYNSA 360 YVIRPQFFIQLIGILRNAALNTLKYQELALMKEQNIDITHFEEDLDIFKNAFAKNYNSA 360
Sbjct: 42515 YVIRPQFFIQLIGILRNAALNTLKYQELALMKEQNIDITHFEEDLDIFKNAFAKNYNSA 42492 YVIRPQFFIQLIGILRNAALNTLKYQELALMKEQNIDITHFEEDLDIFKNAFAKNYNSA 573718

Query: 361 SKNFQKAIDEIDKSIKRMEAVKAALTTSENQRLANNKLDVSVKLLTRKNPTMKAKFDA 420 SKNFQKAIDEIDKSIKRMEAVKAALTTSENQRLANNKLDVSVKLLTRKNPTMKAKFDA 420
Sbjct: 42491 SKNFQKAIDEIDKSIKRMEAVKAALTTSENQRLANNKLDVSVKLLTRKNPTMKAKFDA 42312 SKNFQKAIDEIDKSIKRMEAVKAALTTSENQRLANNKLDVSVKLLTRKNPTMKAKFDA 573538

Query: 421 LKD 423 LKD 423
Sbjct: 42311 LKD 42303 LKD 573529
```

NEM316 /SS1219

NEM316 / 2-22

# Perte des répétitions dans les protéines répétées

## serin rich protein

2-22 NFLSESASTSASTSASMSASTSASTSASTSASTSASTSASTSSSSSVTSN  
NEM KLLSESASTSASTSA 310 x SX SASTSA-TSSSSSVTSN

## Fibrinogen binding protein A

2-22 NQSQGNV LERRQRDVENKSQV GQLIEKNPLFSK  
NEM NQSQGNV LERRQRDVENKSQV 17x GQLIEKNPLFSK

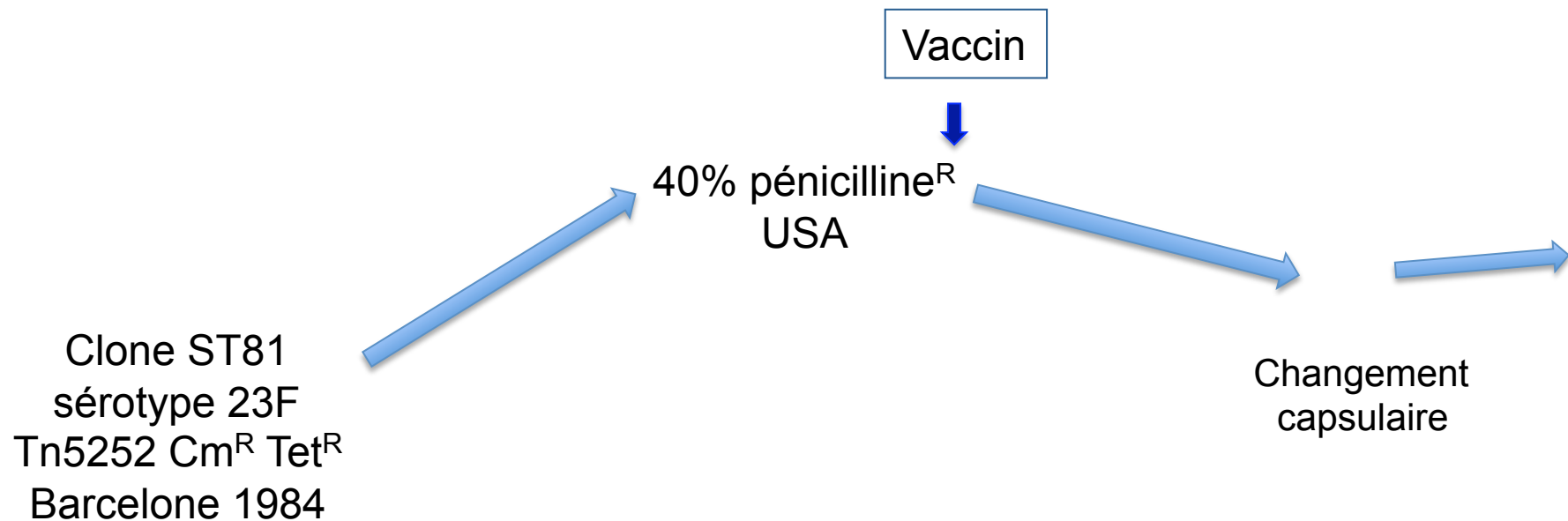
Deletion of one rDNA operon

# Rappelle l'évolution d'espèces hautement pathogènes pour l'homme

- Espèces dérivant d'espèces pathogènes opportunistes à large spectre d'hôte.
  - Très virulente, mais portage rare.
    - *Salmonella typhi* et *paratyphi* A de *S. typhimurium*
    - *Bordetella pertussis* et *parapertussis* de *B. bronchiseptica*
- ⇒ Pertes de fonctions: réduction génomique, pseudogènes et expansion d'IS
- ⇒ Associées à un *goulot d'étranglement* récent

# Evolution du clone de pneumocoque Spain23F-1

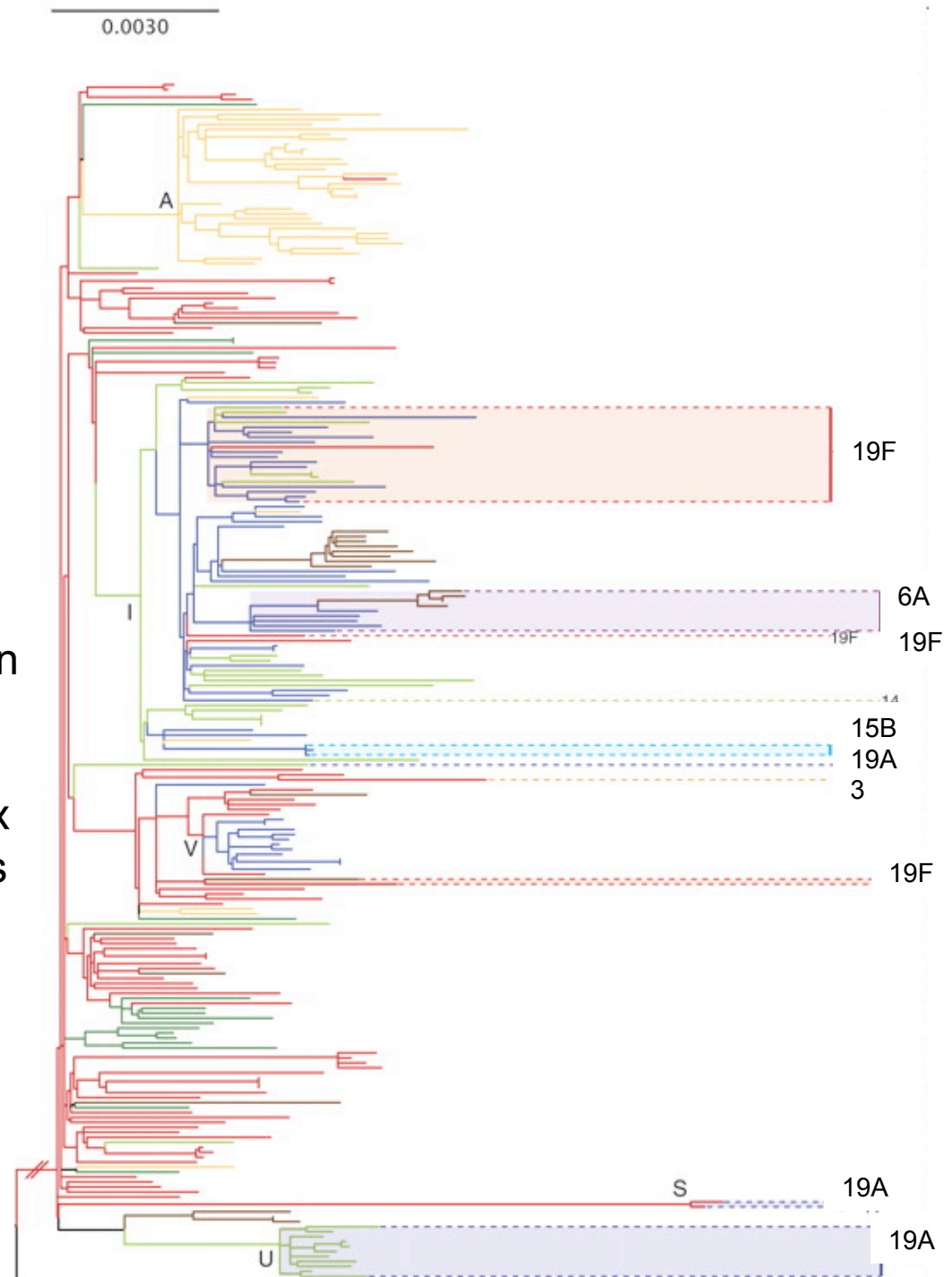
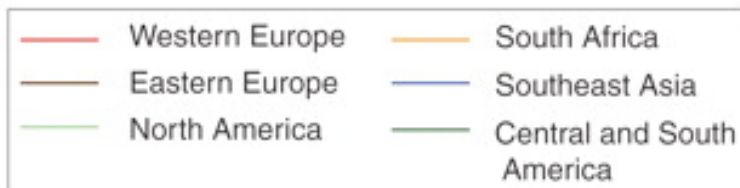
PMEN1 : Pneumococcal molecular epidemiology network clone 1



⇒ Etude de l'évolution de ce clone au niveau génomique et à l'échelle mondiale  
(Croucher et al. Science 2011)

# Phylogéographie of PMEN1

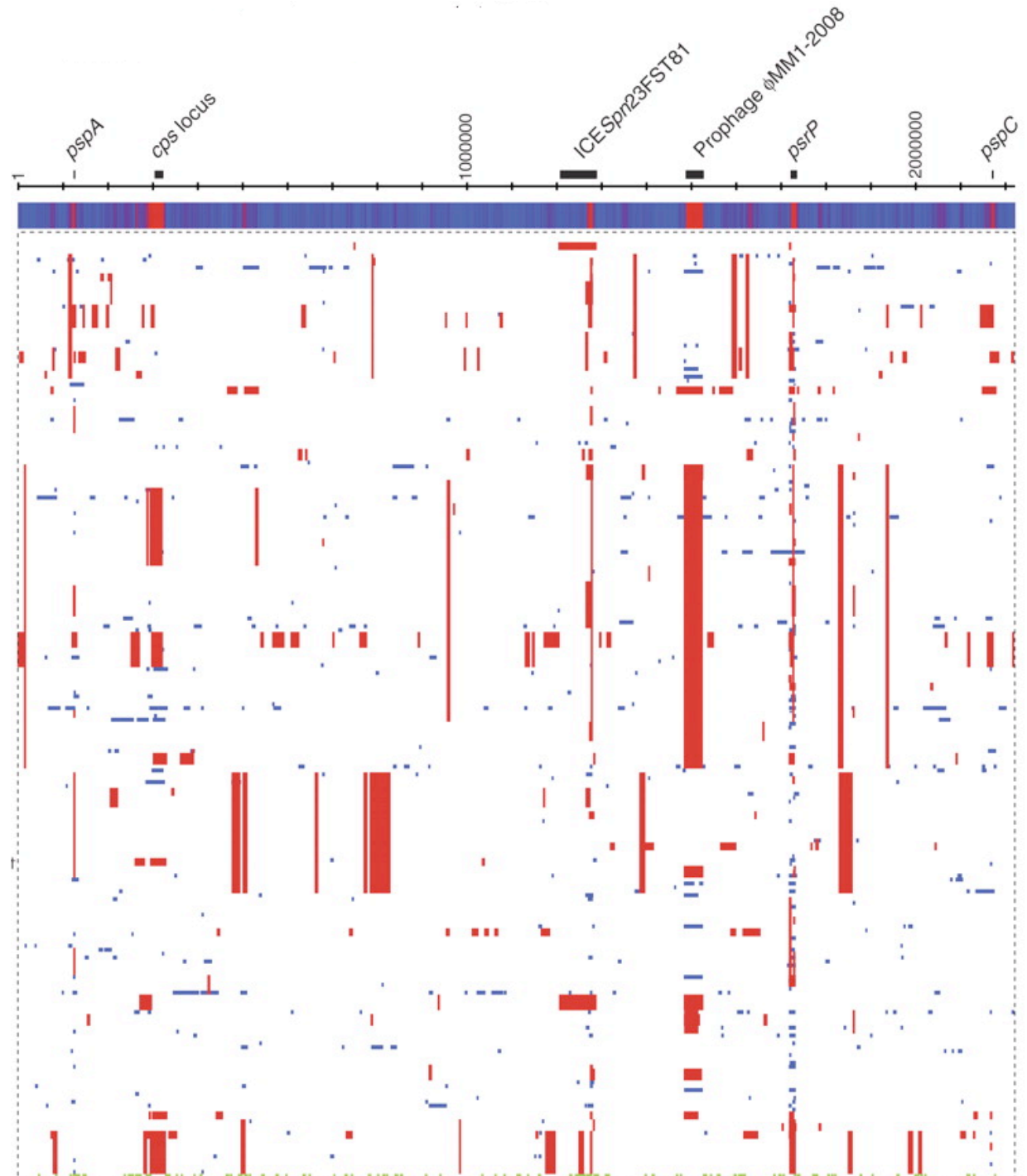
- 57736 site polymorphes
  - 50720 recombinaisons
  - 7016 mutations
- Taux de mutations: 3 mutations par génome et par an
- Origine: 1930 avant l'aire des antibiotiques
- Identification de clones locaux
- Identification de changements de type capsulaire





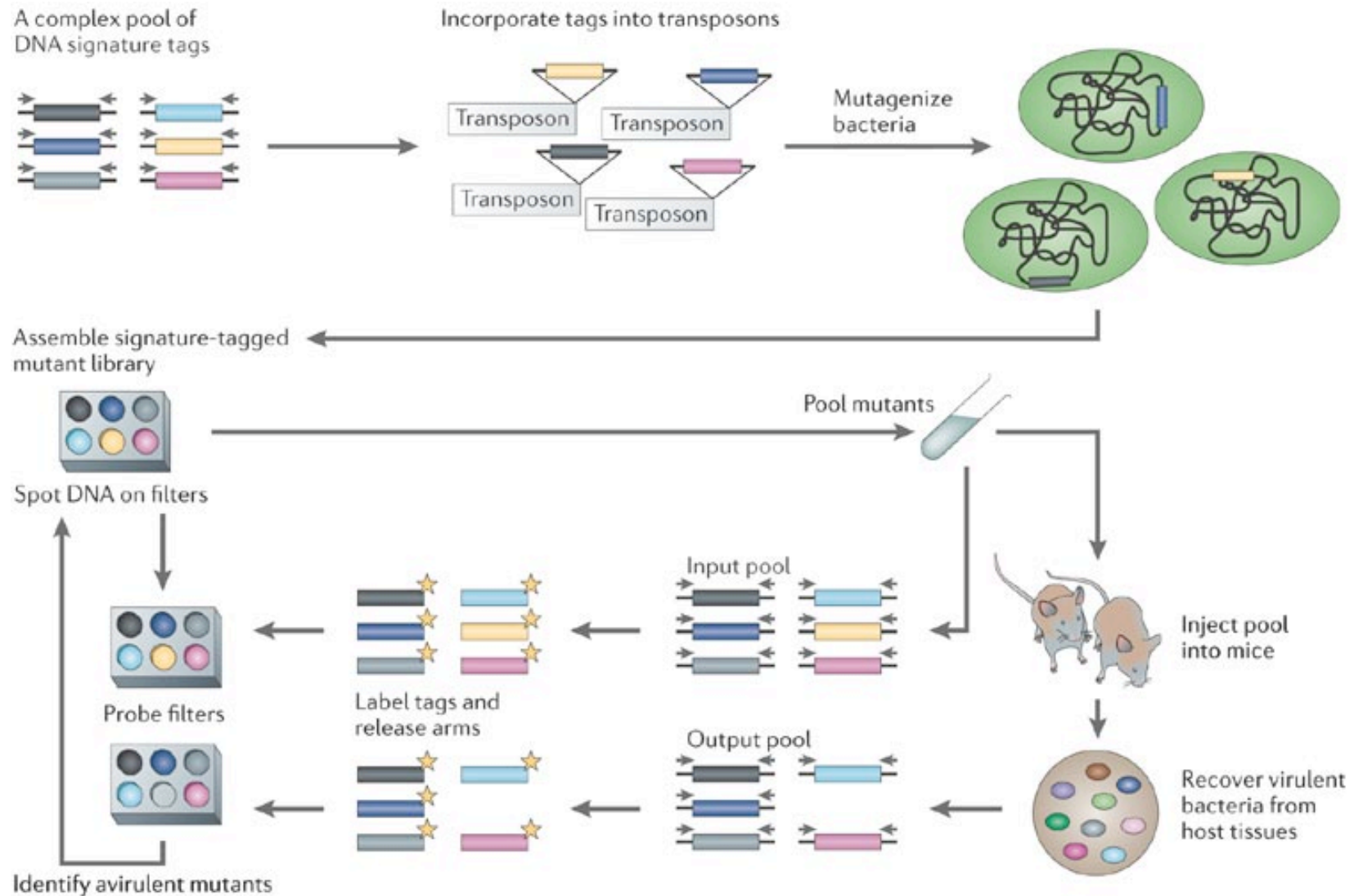
# Variations de séquence et recombinaison

- 74% du génome soumis à des recombinaisons.
- Taille des régions: 3 pb à 72 kb, moyenne 6.3kb
- Des régions préférentielles de recombinaison



Etude fonctionnelle  
Signature Tagged Mutagenesis

# Signature Tagged Mutagenesis



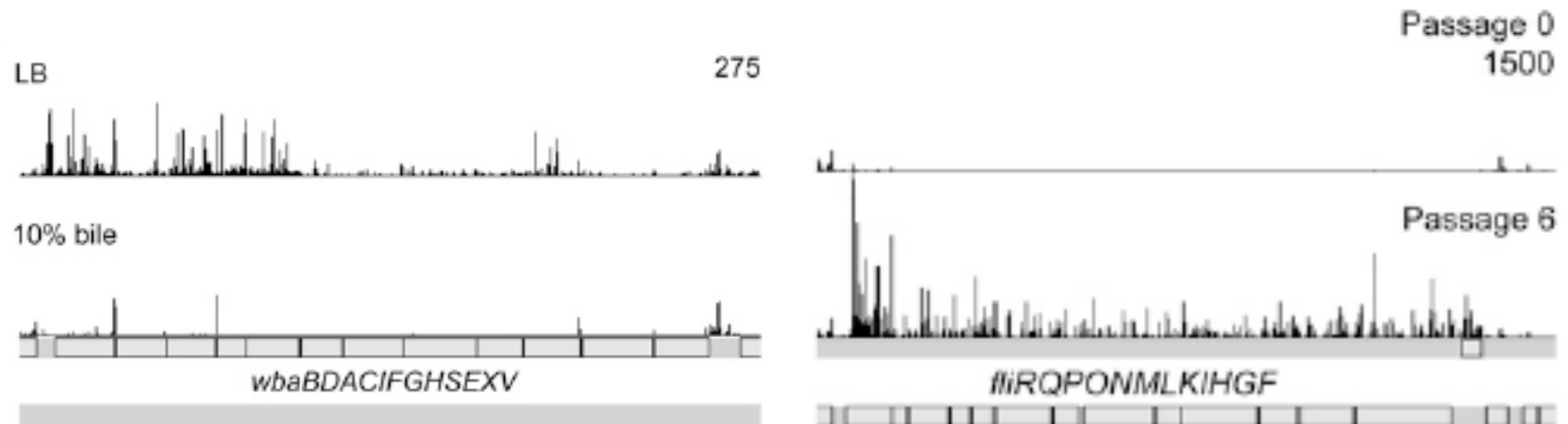
# Caractérisation de collections de mutants d'insertion STM haut débit

- Collection de mutants de *Salmonella* d'insertion simple (sans tag)
  - Création d'une banque Illumina classique
  - Etape d'amplification avec un oligo dans le transposon et un oligo Illumina
  - Séquençage : 6 millions de séquences
    - Filtre des séquences d'insertion du transposon
    - Identification du site d'insertion
  - Analyse statistique
  - Recherche de gènes essentiels
- ⇒ Application à *Salmonella* en LB ou en présence de sels biliaires

# Recherche de gènes essentiels de *Salmonella thyphi* en LB ou en présence de sels biliaries

(Langridge GC et al. Genome Res 2009)

- Cartographie de 370 000 transposons
- 356 gènes essentiels, 169 gènes contribuent au fitness en présence de sels biliaries



Le LPS devient essentiel en présence de bile

Inactivation du flagelle favorisée dans des conditions de laboratoire

# Conclusions

- Capacité de caractériser simultanément un très grand nombre de clones
- Identification de clones sélectionnés positivement et négativement
- Application dans des contextes multiples
- Risque: évolution rapide de la composition de la collection
- Application à des collections de mutants ponctuels (après mutagénèse chimique)

# Etude de l'ARN

Transcriptome

Cartographie des +1

Analyse des promoteurs

Découverte d'ARNs non codants

# Mapping the transcripts + operon structure

nature

Vol 464 | 11 March 2010 | doi:10.1038/nature08756

## ARTICLES

---

### **The primary transcriptome of the major human pathogen *Helicobacter pylori***

Cynthia M. Sharma<sup>1</sup>, Steve Hoffmann<sup>2</sup>, Fabien Darfeuille<sup>3,4</sup>, Jérémy Reignier<sup>3,4</sup>, Sven Findeiß<sup>2</sup>, Alexandra Sittka<sup>1</sup>, Sandrine Chabas<sup>3,4</sup>, Kristin Reiche<sup>5</sup>, Jörg Hackermüller<sup>5</sup>, Richard Reinhardt<sup>6</sup>, Peter F. Stadler<sup>2,5,7,8,9</sup> & Jörg Vogel<sup>1,10</sup>

Idée reçues pour *H. Pylori*  
- pas d'ARN nc  
- pas d'organisation en opéron

differential approach (dRNA-seq) selective for the 5' end of primary transcripts, allows to establish a genome-wide map of *H. pylori* transcriptional start sites and operons

- Establishment of a genome wide map of the *H. pylori* transcriptional start sites
- Description of operonic organization for most genes
- Genome wide antisense transcription
- Identification of internal promoter on + or – strands
- Identification of 60 small non coding RNA and few targets
- Characterization of *H. pylori* promoters

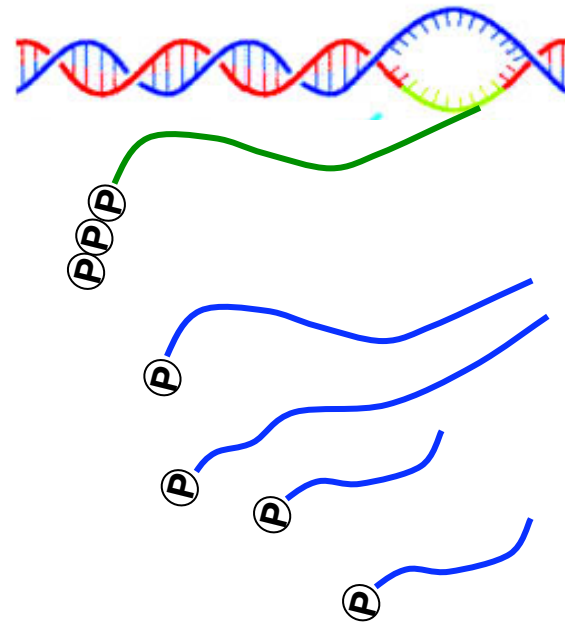


# Cartographie des ARNs chez *S. agalactiae*

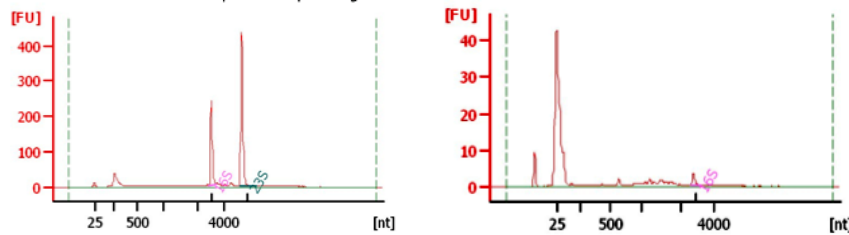
- Identification des +1: caractérisation des promoteurs et de l'organisation en opérons
- détection d'ARNs non annotés

Transcrits primaires

Produits de maturation  
ou de dégradation



# Protocol de séquençage des extrémités 5' des ARN bactériens



**Elimination des ARNr (partielle):  
Oligocapture avec le kit  
MicrobExpress**

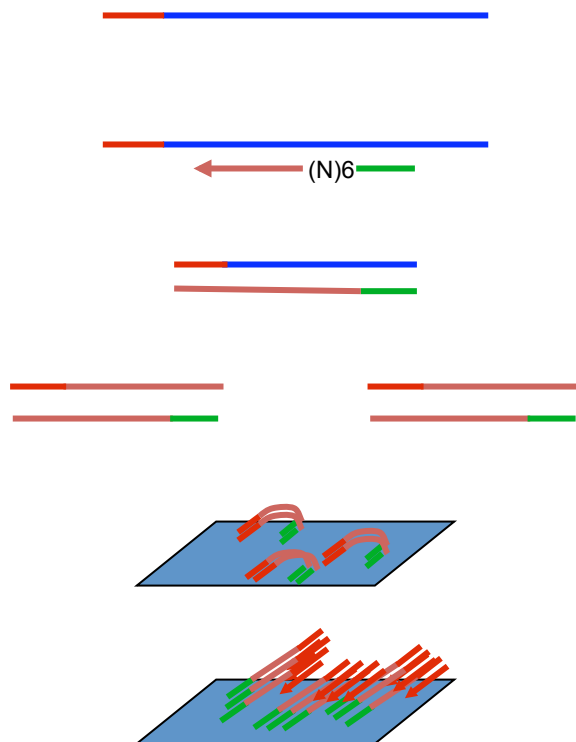
**Ligation d'un adaptateur 5':  
seulement possible sur les  
extrémités 5' mono P**

**Synthèse d'ADNc par random priming: ajout de  
l'adaptateur 3'**

**Fractionnement par la taille (nécessaire pour le  
séquençage Illumina)**

**Amplification PCR (nombre de cycles limité)**

**Séquençage Illumina à partir du primer 5'**



# Convertir les extrémités 5'ppp en 5'p

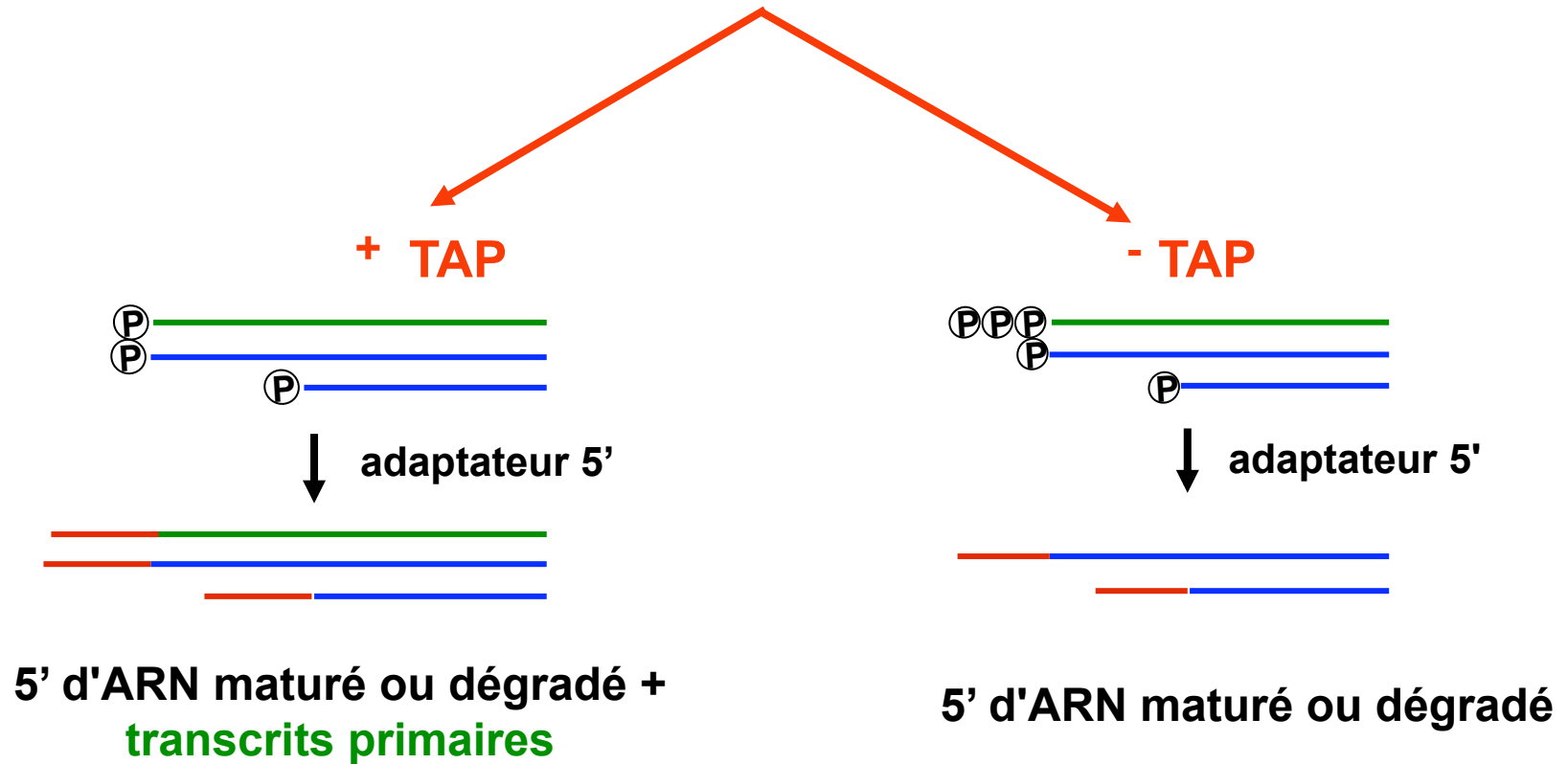
**Obtention des séquences des ARNm natif (Transcription Start Sites)**

**Comment?**

Traitement à la phosphatase acide de tabac (TAP)



# Séquençage des extrémités 5'



**+TAP** spécifique : cartographie des débuts de transcription  
+ et -TAP: caractérisation des produits de dégradation ou de maturation

# Résultats du séquençage Illumina

## +TAP

12 098 478 lectures

95 % sont cartographiées  
sur le génome

55 % = rRNAs et tRNAs

## -TAP

10 432 525 lectures

95 % sont cartographiées  
sur le génome

65 % = rRNAs et tRNAs



# Identification des débuts de transcription (TSS) par une analyse statistique

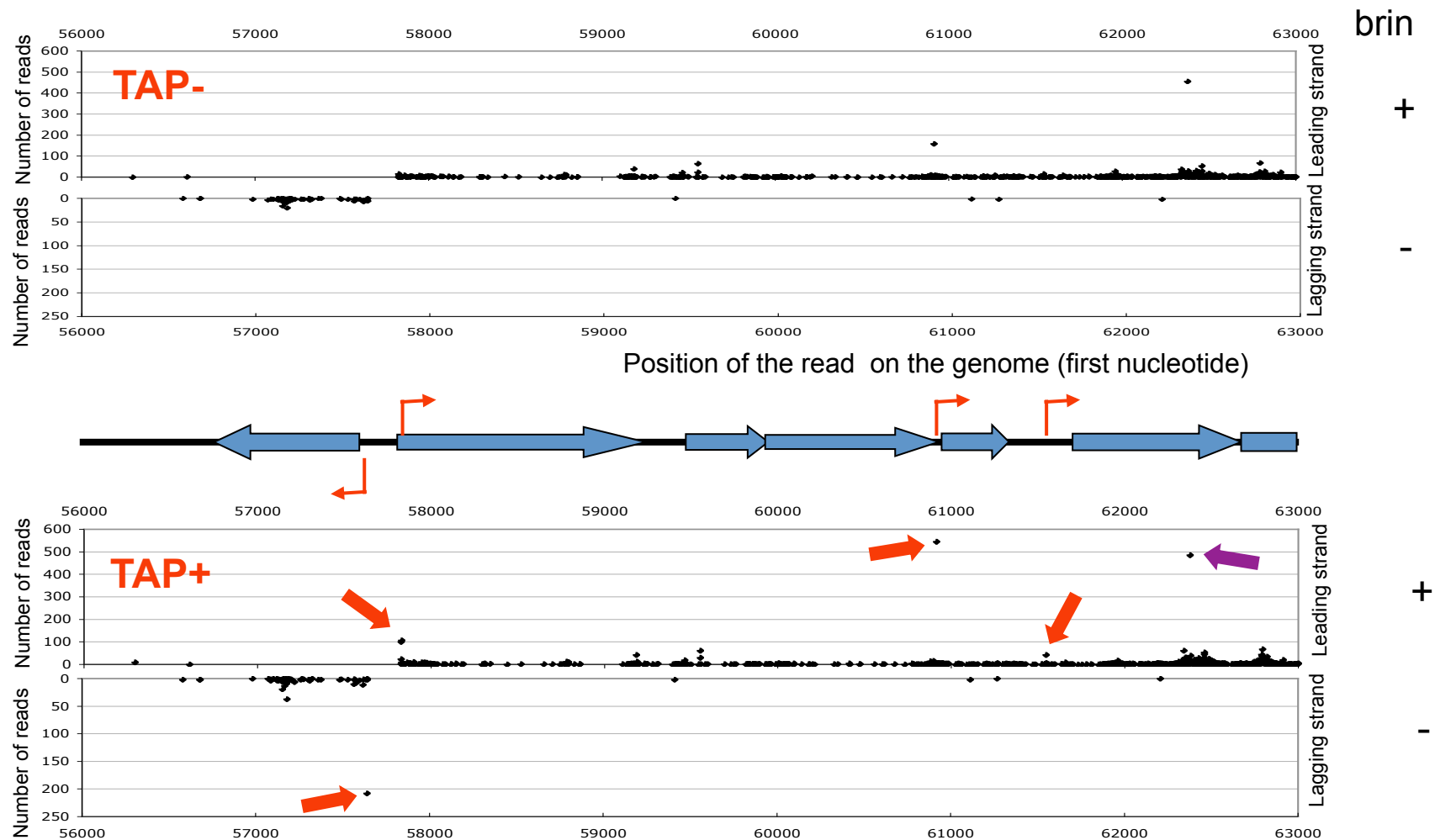
- Alignement des lectures sur la séquence génomique et quantification.
- Analyse statistique basée sur la différence du nombre de lectures + ou – traitement TAP

# Cartographie de 25 promoteurs sur les 70 premiers kb du génome

pos 5'	ATG	gene end		-37_'-27	-15_'-5	-TAP	+TAP
144 +	176	1537	gbs0001	TGTGGAAAAC	TGGTACAATA	7	114
2895 +	2898	3779	gbs0003	TAGTTTTCTT	GCTATAATGA	2	97
4523 +	4551	5666	gbs0006	TACCGCTTTT	TGATATAATA	211	1274
10080 +	10110	10382	gbs0009	TCAGATATTT	GCTAAAATAG	92	489
13111 +				GATAGTTTTA	GATTAGAATT	250	6047
13405 +	13444	13986	gbs0014	TATTTGAGAA	ATGATATAAT	113	811
13663 +	14009	15985	gbs0015	AATTGATTTT	AGCTATCATG	8	82
28945 +	28975	30318	gbs0016	AATATAATTT	GGTAAAATAT	7107	44603
30418 +	30442	31410	gbs0017	GGATTGAAAA	GCGTTAAAAT	83	833
31494 +	31518	32693	gbs0018	ACTGGACACA	TGTATAATA	10	68
33049 +	33507	34385	gbs0020	ATTATTTATT	ACGTTAGAAT	7	35
34437 +	34463	35455	gbs0021	TTGTGAAAAA	TGCTATAATA	114	4261
35802 +	35829	36533	gbs0023	TGTTGTATTA	TGTTAGAATA	46	231
43858 +	43858	44610	gbs0028	ATAAAGAGAA	TGTTATAATA	4	110
43860 +	44630	46177	gbs0029	AAAGAGAAAA	TTATAATAAG	1	23
46226 +	46370	47269	gbs0030	TTAGGTTTTTC	AGGTATACTA	15	602
47384 +	47416	48720	gbs0031	AATTTTATAG	TGATATTATA	54	196
57831 +	57888	59150	gbs0042	AAACTTAAAA	GTTATAAATT	8	107
60913 +	60992	61249	gbs0045	CTAGCGCTCA	TGGTAAACTA	158	545
64221 +	64256	65158	gbs0048	CATACTTAAA	TGATAATAATA	4	29
65447 +	65447	66445	gbs0049	TTCTTAACTT	TGTTATAATA	72	1592
66582 +	66597	67034	gbs0050	AACGTATTAT	TCTGTTATAAT	46	765
69420 +	69465	72107	gbs0053	CTATTTACAA	TGCTATAATA	188	1171



# Exemple typique de distribution des lectures



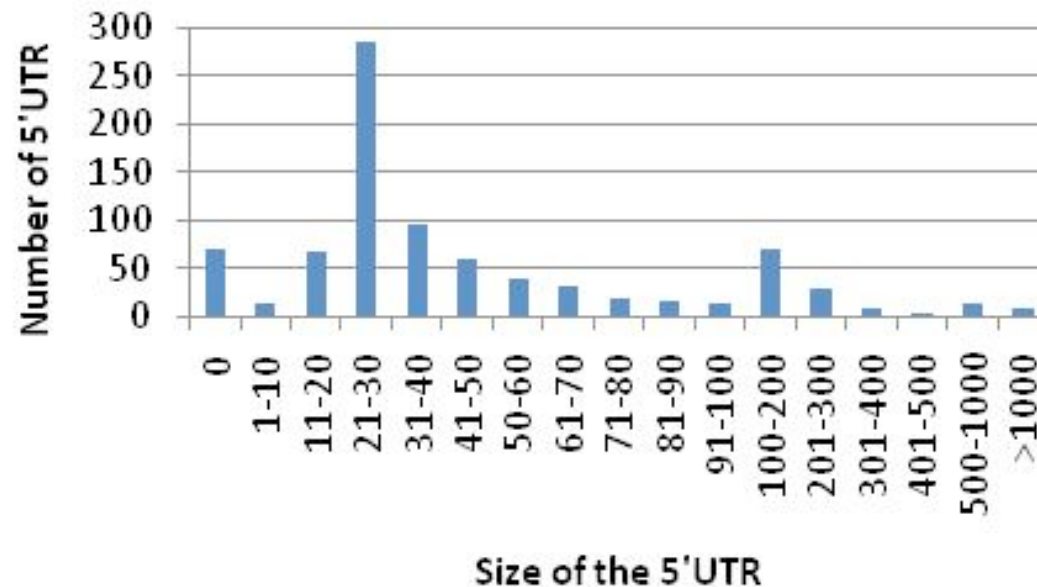
- ◆ TAP- permet de cartographier les régions transcrites
- ◆ La comparaison TAP-/TAP+ permet de déterminer les sites d'initiation de la transcription

# Analyse des débuts de transcription

- L'analyse statistique permet d'identifier 1837 débuts de transcription
- Prédiction de 1010 promoteurs
  - Débuts multiples de transcription pour un promoteur
  - 990 promoteurs avec une boîte -10: **TATAAT**
  - Très peu quantitatif: gène avec le plus grand nombre de lectures:
    - *gbs1286*: TAP+ 167 182; TAP- 681
    - *gbs0291*: TAP+: 108 845, TAP-: 356

# Localisation des promoteurs

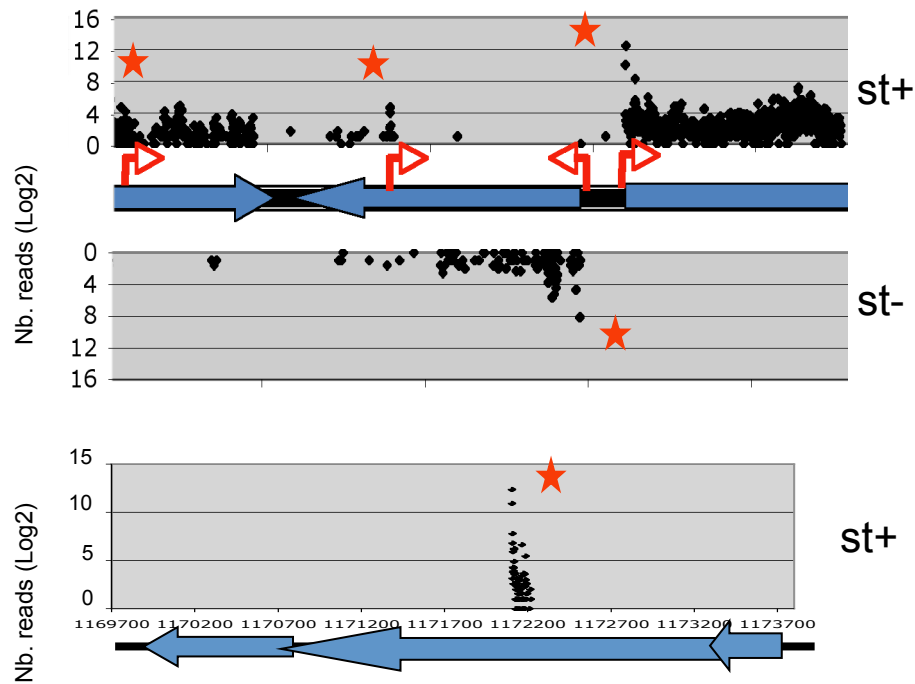
85% des promoteurs sont en amont de gènes (opérons)



15% des promoteurs sont dans des régions codantes: sens, anti sens et sRNA

# Catalogue de l'ensemble des transcrits chez *S. agalactiae*

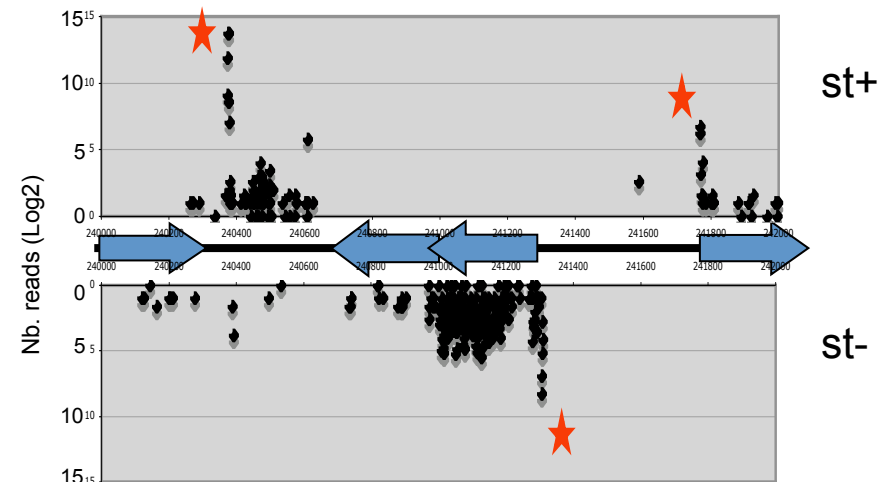
## Transcrit anti sens et sRNA



Distribution des lectures TAP+

★ TSS: statistiquement différent entre TAP+ et TAP-

## sRNAs intergénique



Log2 scale:	
0	↔ 1
4	↔ 16
9	↔ 256
12	↔ 4096

# Conclusions

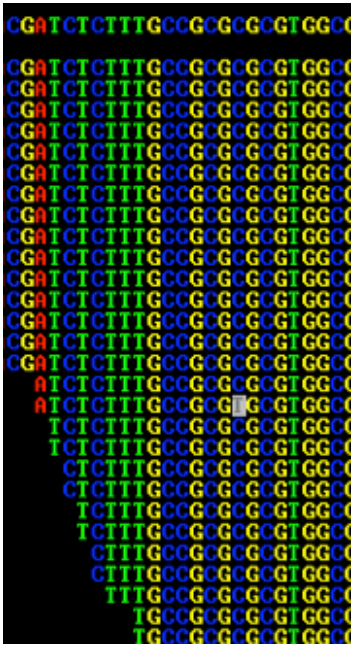
- Outil extrêmement puissant pour analyser les différents aspects de la transcription à l'échelle du génome
- Analyse *in vitro* / *in vivo* pour identifier des transcrits spécifiques
- Différentes variantes en fonction des questions posées
- Utilisation quantitative: remplacement des puces à ADN

# Etude métagénomique

- Analyse des flores complexes: étude du microbiome humain dans des conditions normales ou pathologiques
- Recherche d'agents pathogènes dans un fluide biologique
  - Normalement stérile : liquide céphalo-rachidien, sang etc.
  - En présence d'une flore normale : selle, gorges, peau, lavage bronchoalvéolaire
  - Analyse ADN et ARN
- Impact sur la santé de l'association de microorganismes

# Conclusions

- Impact majeur du NGS dans l'étude des maladies infectieuses:
  - Etudes fondamentales en microbiologie
  - Etude de la virulence
  - Génétique des populations et évolution
  - Réponse de l'hôte
  
  - Santé publique: épidémiologie moléculaire, maladies nosocomiales ...
  
  - Développement de médicaments (?)



# Journée Génomique et Transcriptomique Microbienne CIS Institut Pasteur 3 Octobre 2011

