



GET
Génome et
Transcriptome

RNA-seq

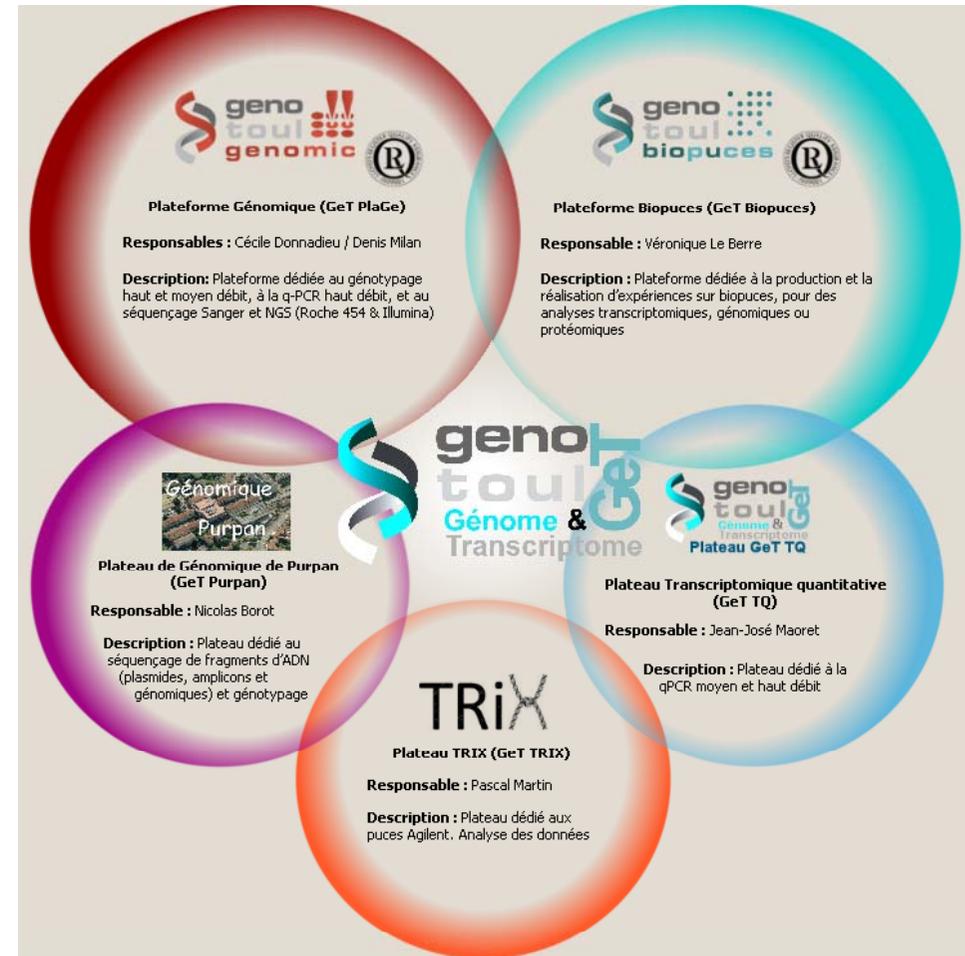
Olivier Bouchez
Nathalie Marsaud

Plateforme GeT : Génome et Transcriptome

Deux plateformes IBiSA et 3 plateaux techniques regroupés depuis 2010

**Responsable scientifique
Denis Milan**

- **Coordination des nouveaux investissements**
- **Interaction dans la mise en place de nouvelles technologies**
- **Partage de l'expertise de chacune des structures**
- **Animation d'une communauté plus large**



- **Certification ISO 9001:2008**



Expertise et mise à disposition d'une plateforme technologique en génomique :

- **Séquençage / Génotypage**
- **Analyse d'expression**
- **PCR temps réel**



Auzeville (INRA)



- **Un partenariat fort et historique avec la PF Bioinformatique**
- **Plateforme stratégique INRA : 147 k€ dotation 2011**
- **Gestion financière et RH des non titulaires par le SAIC de l'INP**

GeT-PlaGe : Ressources Humaines



David

Laure

Johanna

Clémence

- **17 personnes: 4 permanents INRA (dont Denis), 1 CDD TH, 3 permanents Instituts ou labo (100% GA, 100% CNRS LIPM, 30% INSERM), 9 CDD sur activité**
- **4 départs en 2011**
- **Renforcement de la partie administrative**
- **Renforcement de l'équipe technique**
- **Renforcement de l'équipe info/Bioinfo**
- **Partenariat étroit avec la plateforme Bioinfo**



Le partenariat avec la PF Bioinformatique

- L'équipe technique PlaGe prend le relais de la plateforme bioinfo sur l'analyse qualitative des données
- Le développement en collaboration du pipeline de traitement des données et du site NG6
- Des investissements de concert (CDD IBiSA 2008, CPER, France Génomique).

PROJETS RUNS DOWNLOAD

recherche OK

Projets > NG6 > ReadStats

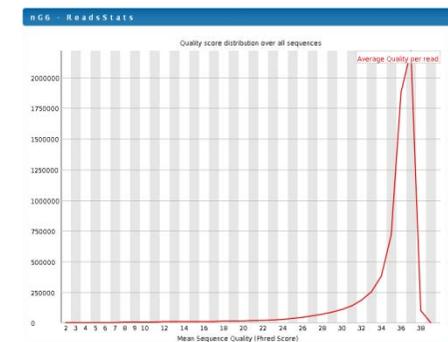
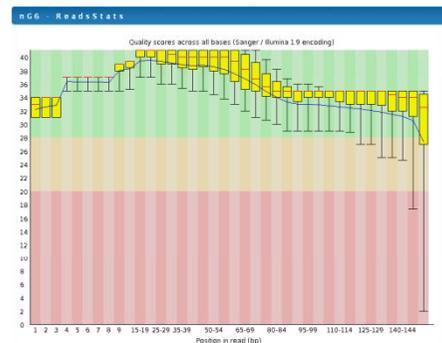
Analyse ReadStats :
 Statistiques sur les lectures et leurs qualités par mids.
 L'ensemble des résultats de cette analyse occupent 1.54 Mb d'espace disque.

Statistiques sur les lectures et sur leur qualité

Echantillons (n)	Statistiques par position				Statistiques par séquence						
	Qualité	GC%	Ns	Contenu	Nombre de séquence	Qualité	GC%	Longueur	Niveau de duplication	Errors	Séquences sur-représentées
001_6_1_seq (95)	FAIL	WARN	PASS	FAIL	32 708 900	PASS	35(PASS)	101(PASS)	WARN	WARN	PASS
001_6_2_seq (95)	FAIL	WARN	PASS	FAIL	32 708 900	PASS	34(WARN)	101(PASS)	WARN	WARN	PASS
002_6_1_seq (142)	FAIL	WARN	PASS	FAIL	29 594 529	PASS	35(WARN)	101(PASS)	FAIL	WARN	PASS
002_6_2_seq (142)	FAIL	WARN	PASS	FAIL	29 594 529	PASS	34(WARN)	101(PASS)	WARN	WARN	PASS
003_6_1_seq (118)	FAIL	WARN	PASS	FAIL	32 779 500	PASS	34(WARN)	101(PASS)	FAIL	WARN	PASS
003_6_2_seq (118)	FAIL	WARN	PASS	FAIL	32 779 500	PASS	34(PASS)	101(PASS)	FAIL	WARN	PASS

Aide Statistiques par position :

- Qualité : WARN = plus petit quartile d'une base inférieur à 10, ou si la médiane d'une base est inférieure à 25. FAIL = plus petit quartile d'une base inférieur à 5, ou si la médiane d'une base est inférieure à 20.
- GC% : WARN = GC% par position +/- 5% de la moyenne. FAIL = GC% par position +/- 10% de la moyenne.
- Ns : WARN = Ns% > 5%. FAIL = Ns% > 10%.



France Génomique

- **Les objectifs de France Génomique :**
 - **Un réseau national intégrant CNS/CNG & 7 PF IBiSA**
 - **Un portail unique pour les utilisateurs**
 - **Des développements (bio & bioinfo) et une veille mis en commun**
 - **Du matériel complété en fonction des besoins**
- **Pour GeT-PlaGe & PF Bioinfo sur les années 1&2 :**
 - **Un nouvel Hiseq 2000**
 - **Une contribution à l'automatisation de la préparation des librairies**
 - **1 CDD Biologiste + 2 CDD Bioinfo sur 2 ans + 1 CDD coordination bioinfo nationale**
 - **300 k€ d'infrastructure informatique**





GET
Génome et
Transcriptome

Qu'est-ce que le RNA-seq ?

➤ Quelques définitions

- ✓ **Séquençage** : déterminer la succession linéaire des bases A, C, G, T de l'ADN, la lecture de cette séquence permet d'étudier l'information biologique contenue par celle-ci
- ✓ **Séquençage Nouvelle Génération (NGS, Next Generation Sequencing)** : Séquençage à très haut débit, génération d'un très grand nombre de séquences simultanément
- ✓ **RNA-seq** : transcriptome sequencing. Informations sur les ARN via le séquençage de l'ADN complémentaire (cDNA)
- ✓ **Re-séquençage** : séquençage d'un fragment d'ADN et comparaison du résultat obtenu avec une séquence de référence connue
- ✓ **Séquençage *de novo*** : séquençage d'un génome pour lequel il n'existe pas de séquence de référence, détermination d'une séquence inconnue

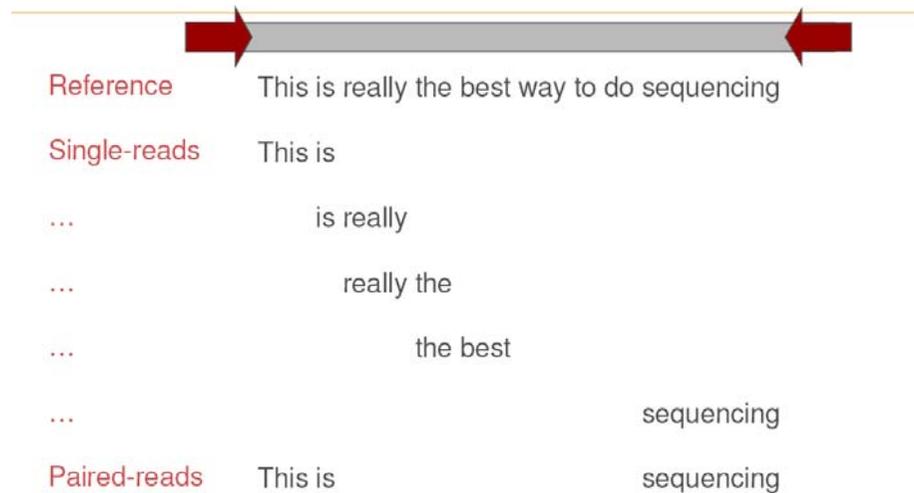
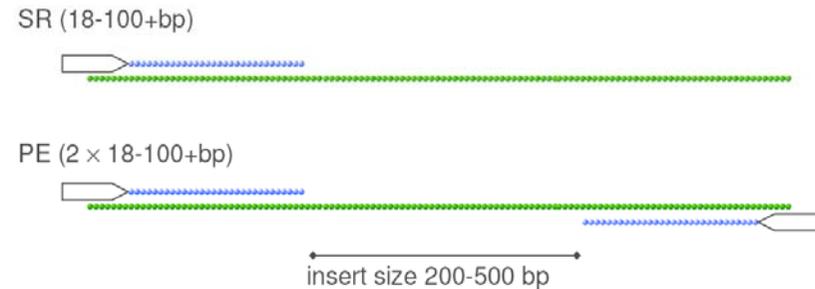
Pourquoi faire du RNA-seq ?

➤ **L'accès aux séquences des ARN permet de :**

- ✓ **Annoter un génome**
- ✓ **Réaliser un catalogue de gènes exprimés**
- ✓ **Identifier des nouveaux gènes**
- ✓ **Identifier des transcrits alternatifs**
- ✓ **Quantifier l'expression de gènes (comparaisons entre différentes conditions expérimentales)**
- ✓ **Identifier des petits ARNs**
- ✓ **Identifier des « starts » de transcription**
- ✓ **....**

Introduction

- ✓ **Lecture single read (SR) et paired end (PE)**
- ✓ **PE : facilite l'alignement des séquences sur le génome de référence et/ou l'assemblage**



Assembly becomes easier!!
 (-----26 characters-----)

illumina

Séquenceurs NGS PlaGe



Illumina HiSeq 2000 x 2

Séquençage par synthèse
2 flowcells
Longueur : 2 x 100 pb
Débit : 240-300 GB/flowcell
Temps de run : 15 jours
Multiplex 24



Illumina MiSeq

Séquençage par synthèse
Longueur : 2 x 150 pb
Débit : >1 GB
Temps de run : 27 h
Multiplex 24



GS FLX 454 XL+

Pyroséquençage
Longueur : 400-700 pb
Débit : 1 million de séquences, ~400-700 MB
Temps de run : 10h

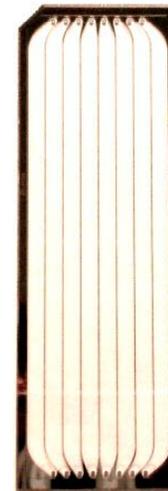
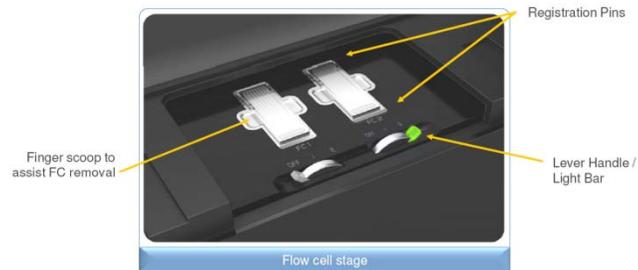
Spécificités

Machine	Roche GS FLX	HiSeq 2000 v3	MiSeq	PacBioRS	Ion Torrent	Life Tech	Oxford Nanopore
Type	PCR billes	PCR Cluster	PCR cluster	Simple molécule	PCR Billes	Simple Molécule	Simple Molécule
Taille Fragments	500 → 1000	100 → 150	150	1000	100-200	5 000 à 10 000	→ 100kb
Nb fragments / run	1 Million	6 Milliards	13 Millions	50 000	100 000	40 000	8 000
Mb / run (durée)	700 Mb (8h)	600 000 Mb (14j)	3 400 Mb (27h)	50 Mb (15 min)	10 Mb (2h)	40 Mb (2h)	10 000 Mb (24h)
Cout marginal	12 € / Mb	0.1 € / Mb	1 € / Mb	10-200 € / Mb	5-10 € / Mb	5 € / Mb	?
	2008	2010	2011	2012	2011	2012	2012

Présentation d'un séquenceur NGS : Illumina HiSeq2000

Spécifications :

- ✓ **600 Gb (2 flowcell de 8 lanes)**
- ✓ **Paired end (2x100 bp): ~360 millions de séquences/lane**



1 flowcell = une lame = 8 lignes



1 Library Preparation



Fragment DNA
Repair ends
Add A overhang
Ligate adapters
Purify

2 Cluster Generation



Hybridize to flow cell
Extend hybridized template
Perform bridge amplification
Prepare flow cell for sequencing



3 Sequencing



Perform sequencing
Generate base calls

4 Data Analysis



Images
Intensities
Reads
Alignments

Contrôles qualité des ARN

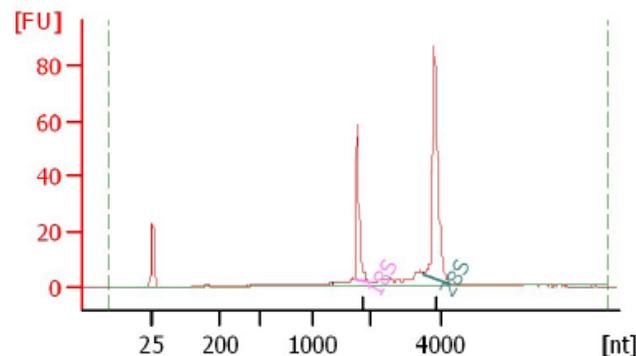
➤ Dosage au Nanodrop

- ✓ **Concentration de l'ARN en ng/μL (au moins 100 ng/μL)**
- ✓ **Ratio 260/280: contamination par les protéines; doit être > 1.8**
- ✓ **Ratio 260/230: contamination par les sels (résidus de l'extraction); doit être > 1.8**
 → Peut engendrer une diminution du rendement de la Rétrotranscription



➤ Evaluation de l'état de dégradation de l'ARN par un dosage au Bioanalyser (Agilent)

- ✓ **RIN (RNA Integrity Number) : compris entre 0 et 10 ; doit être > 8.5**
- ✓ **28S/18S : doit être > 1.8**

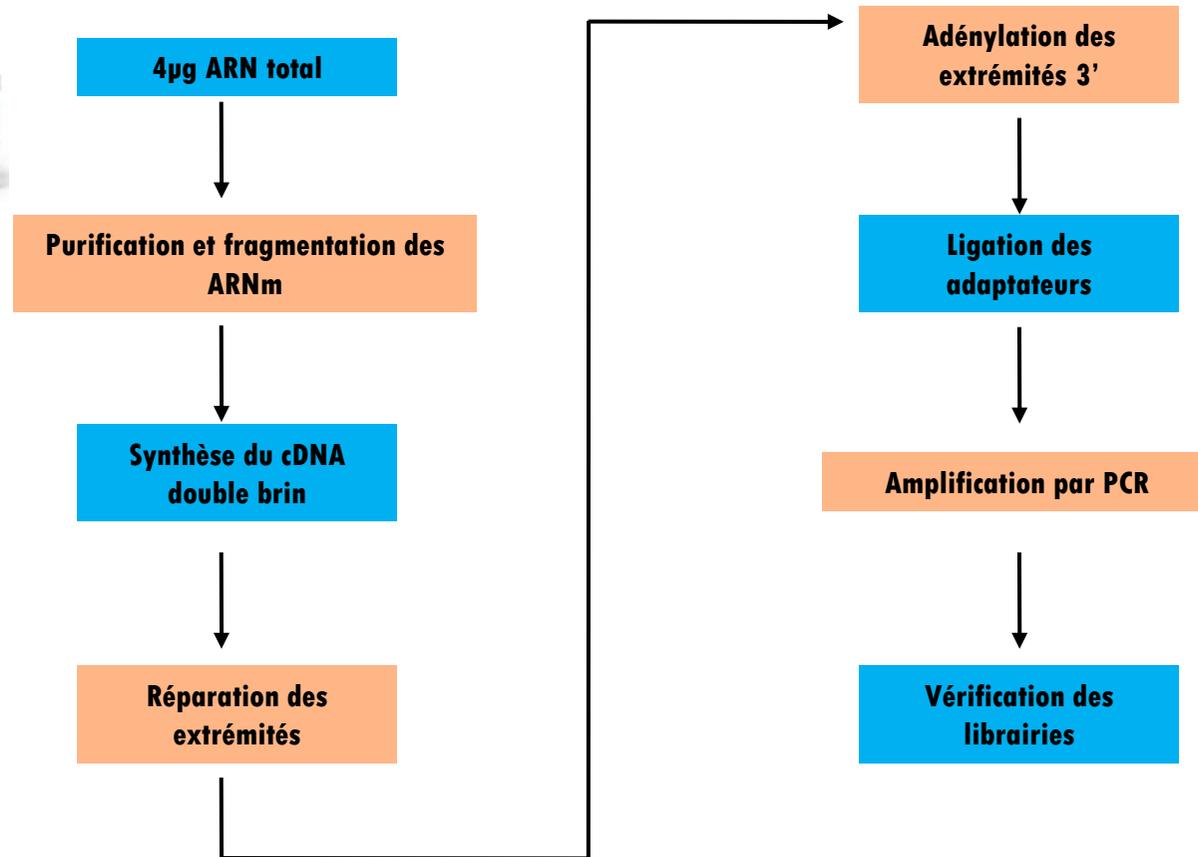


Overall Results for sample 6 : 9-H-

RNA Area:	223,9
RNA Concentration:	71 ng/μl
rRNA Ratio [28s / 18s]:	2,2
RNA Integrity Number (RIN):	9.9 (B.02.07)

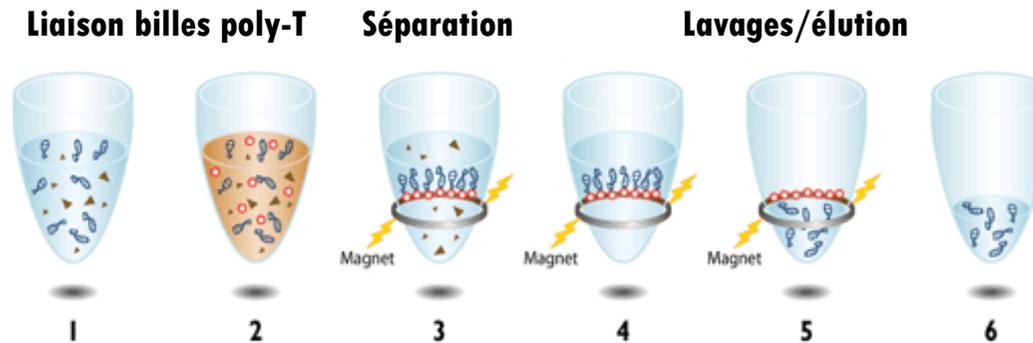
Préparation des librairies

Kits Illumina TruSeq



Préparation des librairies

- ✓ **Purification des ARNm sur billes magnétiques poly-T (sauf pour les bactéries)**



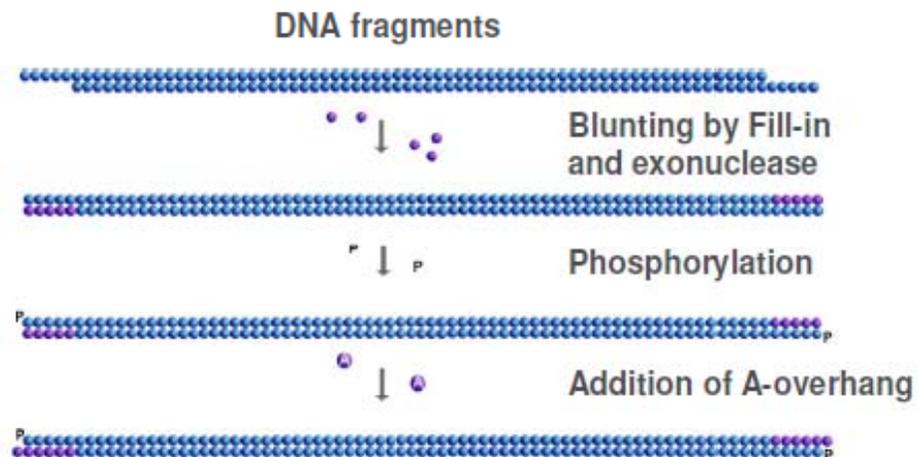
	Echantillon non traité	Echantillon traité			
		Pourcentage d'élimination des ARNr			
		90%	98%	99%	99,90%
ARNr	98%	83%	50%	33%	5%
Autres	2%	17%	50%	67%	95%

Kits Illumina TruSeq

- ✓ **Fragmentation chimique de l'ARN (tampon alcalin)**

Préparation des librairies

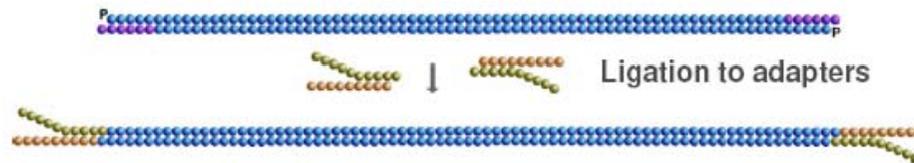
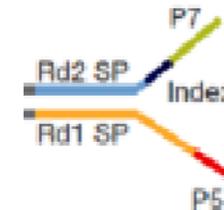
- ✓ **Rétrotranscription de l'ARN en cDNA double brin**
- ✓ **Réparation et phosphorylation des extrémités pour avoir des bouts francs**
- ✓ **Adénylation en 3'**



Préparation des librairies

✓ **Ligation des adaptateurs (contiennent l'index)**

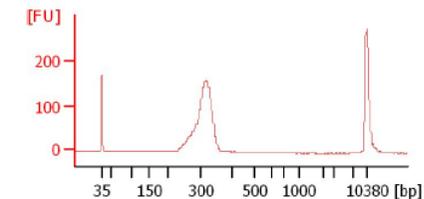
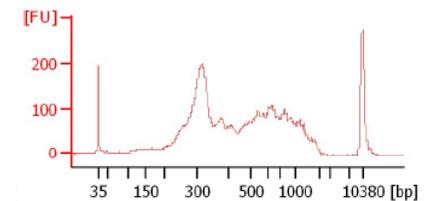
- **P5 et P7: Fixation sur la flowcell**
- **SP 1 et 2: primers de séquençage**
- **Tag: index (6 bases): multiplexage par 24**



✓ **Enrichissement par PCR des fragments+adaptateurs (Primers spécifiques de P5 et P7)**

✓ **Purification sur gel**

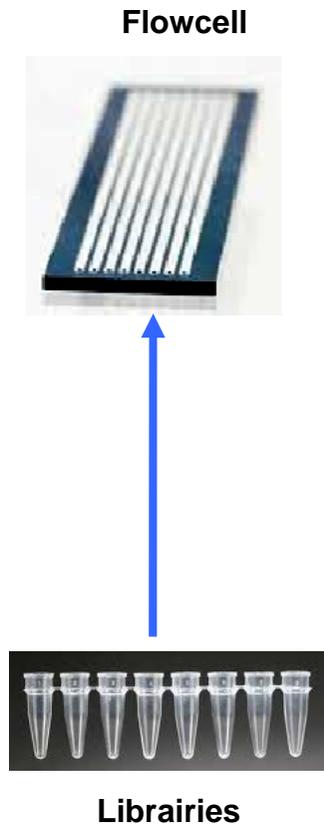
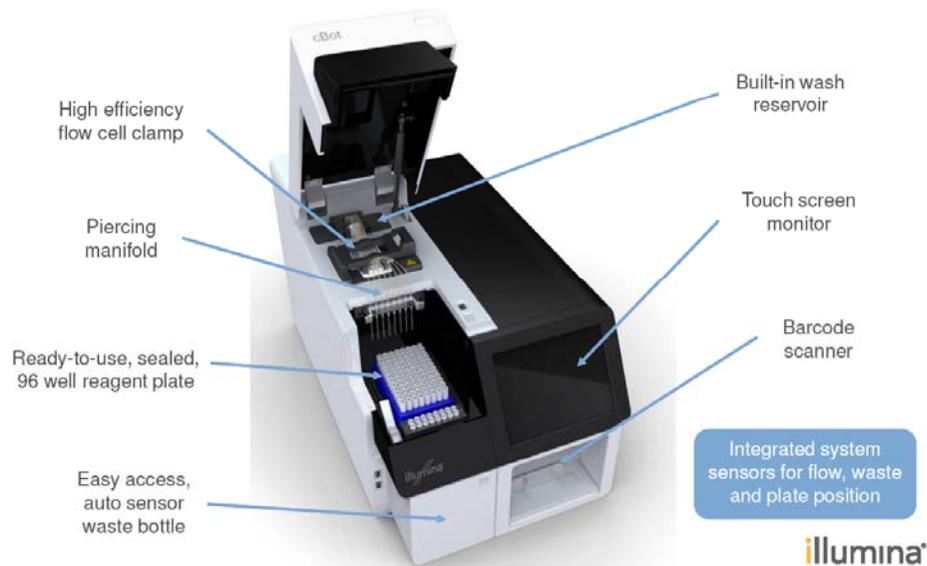
✓ **Contrôles qualités des librairies : Bioanalyser, qPCR**



Génération des clusters

➤ **cBot**

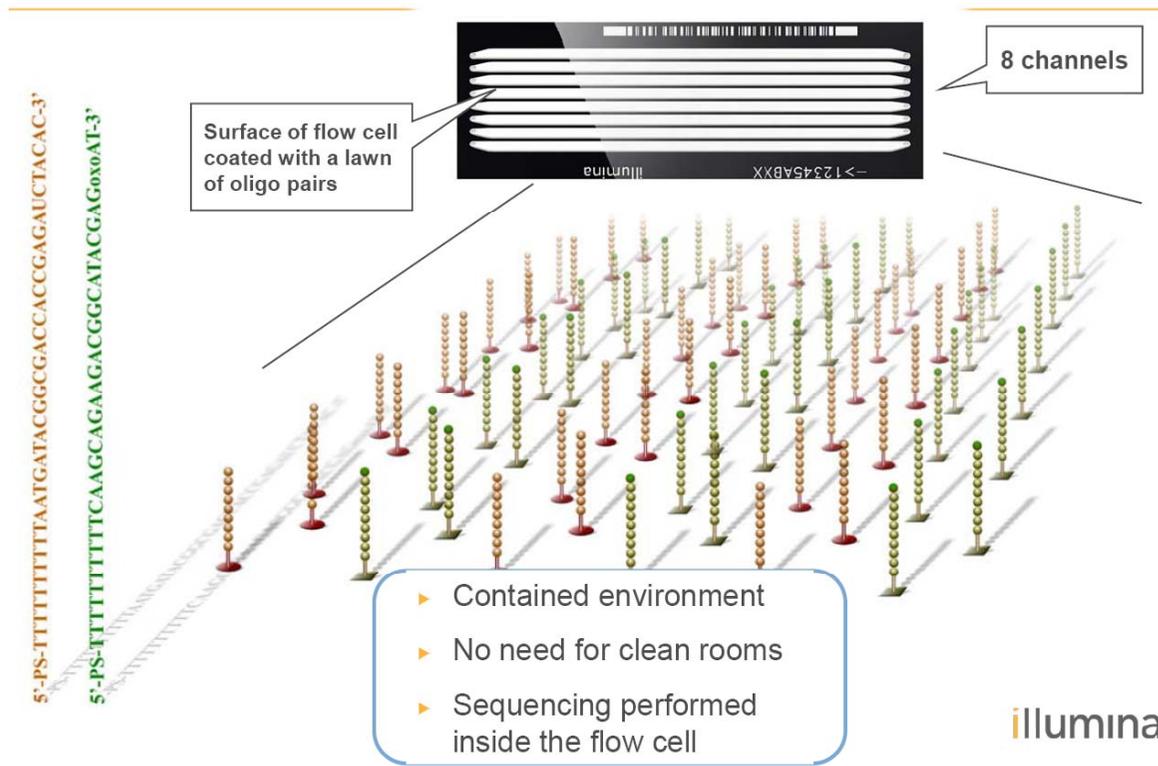
- ✓ **Transfert de la librairie sur la flowcell**
- ✓ **Amplification de la librairie : formation des clusters**



➤ **Génération de clusters : environ 5h**

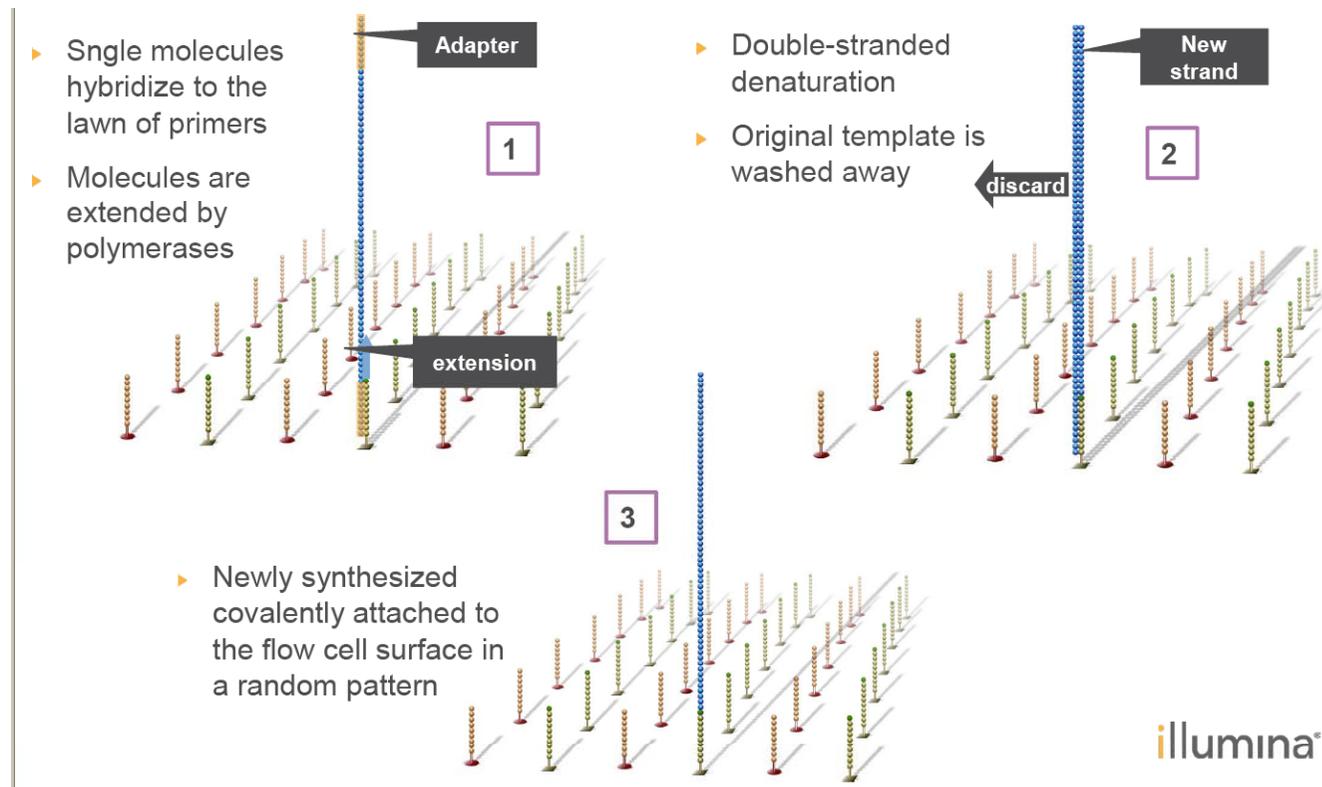
Génération des clusters

Design de la flowcell



Génération des clusters

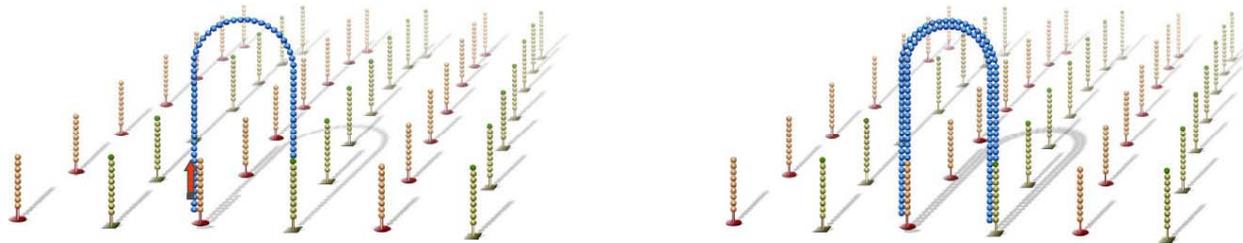
- ✓ **Hybridation des bibliothèques grâce aux adaptateurs à l'intérieur de la flowcell**
- ✓ **Synthèse du brin complémentaire**
- ✓ **Dénaturation**



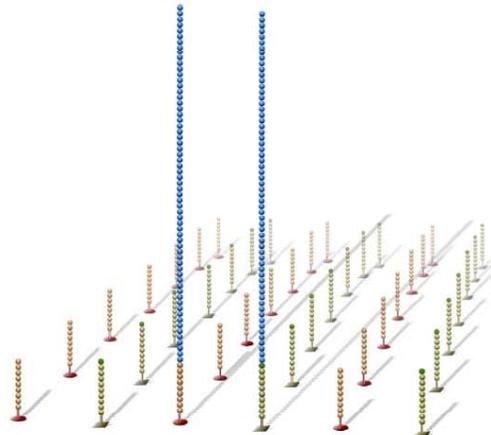
illumina®

Génération des clusters

- ✓ **Formation d'un pont**
- ✓ **Synthèse du brin complémentaire**
- ✓ **Dénaturation**



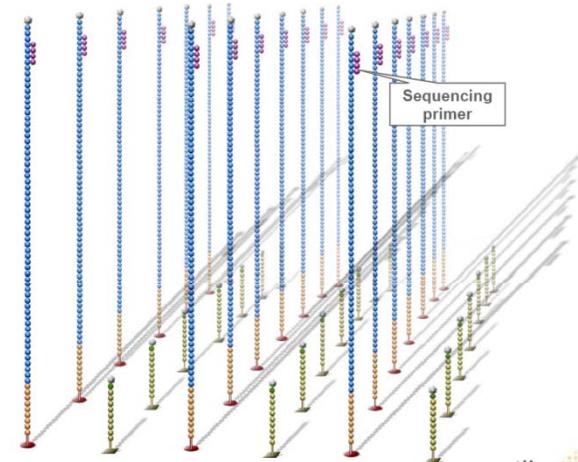
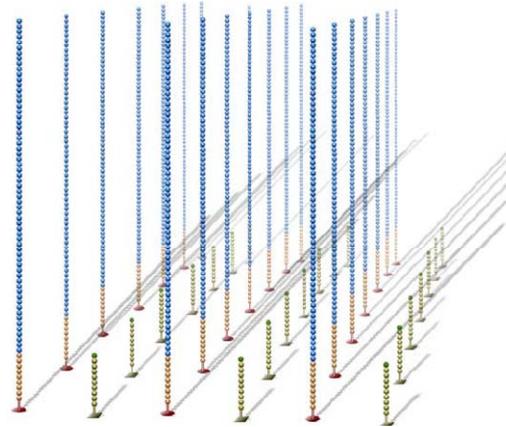
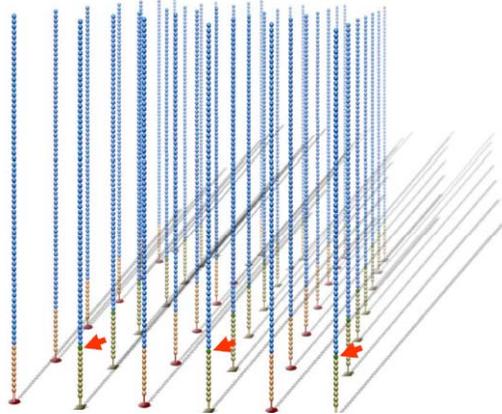
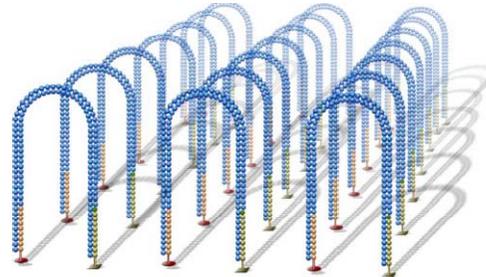
- ▶ Single-strand flips over to hybridize to adjacent primers to form a bridge
- ▶ Hybridized primer is extended by polymerases
- ▶ Bridge is denatured



illumina®

Génération des clusters

- ▶ Bridge amplification cycle repeated until multiple bridges are formed
- ▶ Bridges denaturation
- ▶ Reverse strands cleaved and washed away

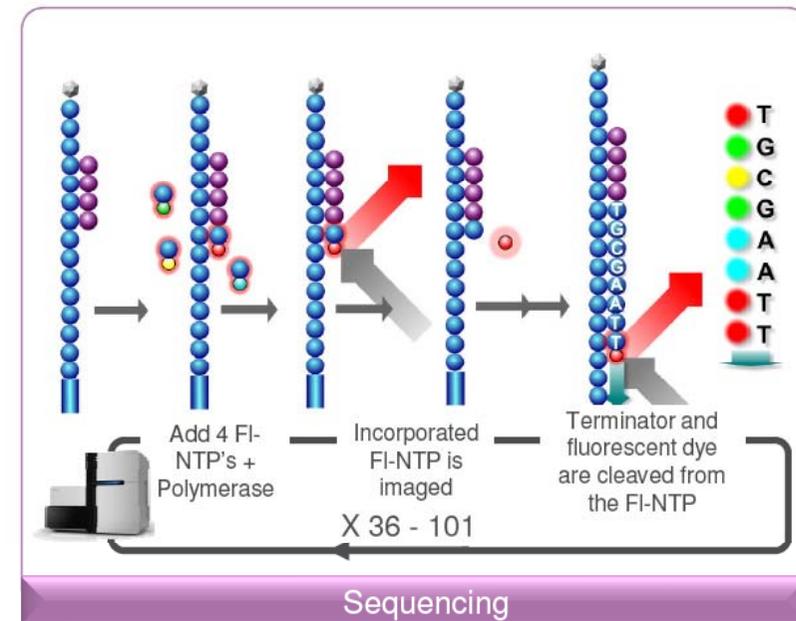


➤ **Formation de 750 000-850 000 clusters / mm²**

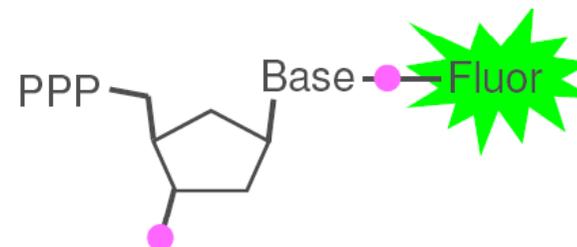
Principe du séquençage par synthèse

➤ Déroulement d'un cycle:

- ✓ **Incorporation des 4 nucléotides fluorescents en même temps**
- ✓ **Acquisition de l'image (50 min/cycle)**
- ✓ **Temps de run : 2 x100 bp : environ 15 jours**
- ✓ **Déprotection du nucléotide incorporé**



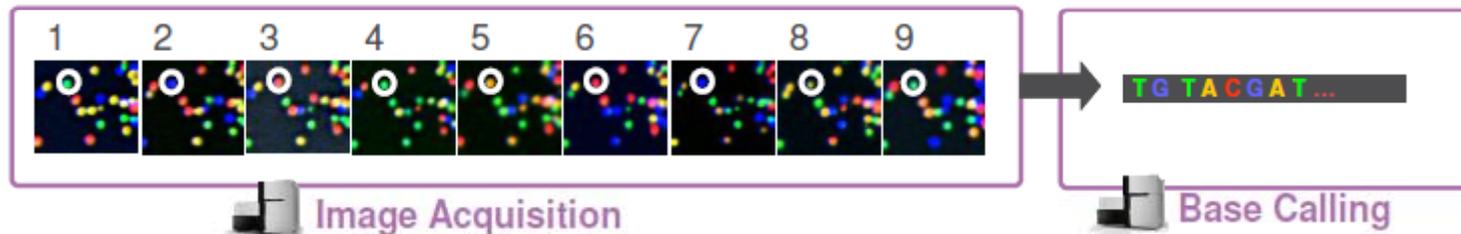
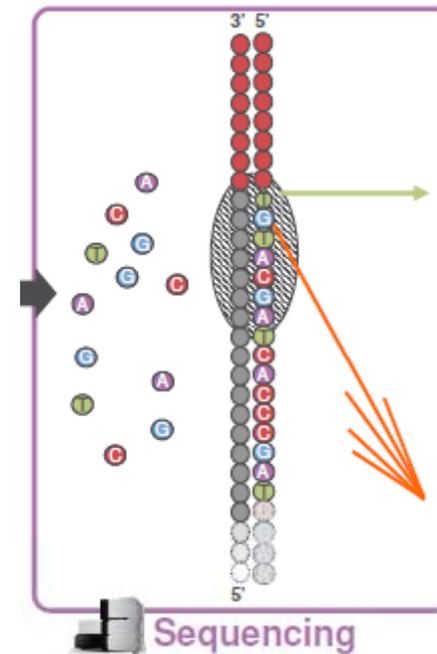
Rq: 1 seul jeu de caméras pour 2 flowcell



Principe du séquençage par synthèse

➤ Base calling:

- ✓ A chaque cycle, une base est incorporée
→ Détection de l'émission de fluorescence
- ✓ Chaque cluster est caractérisé par une position (X ; Y)
- ✓ A chaque cycle : une couleur détectée = une base
- ✓ **Base calling** = correspondance entre la fluorescence et la base pour un cluster



Informations importantes

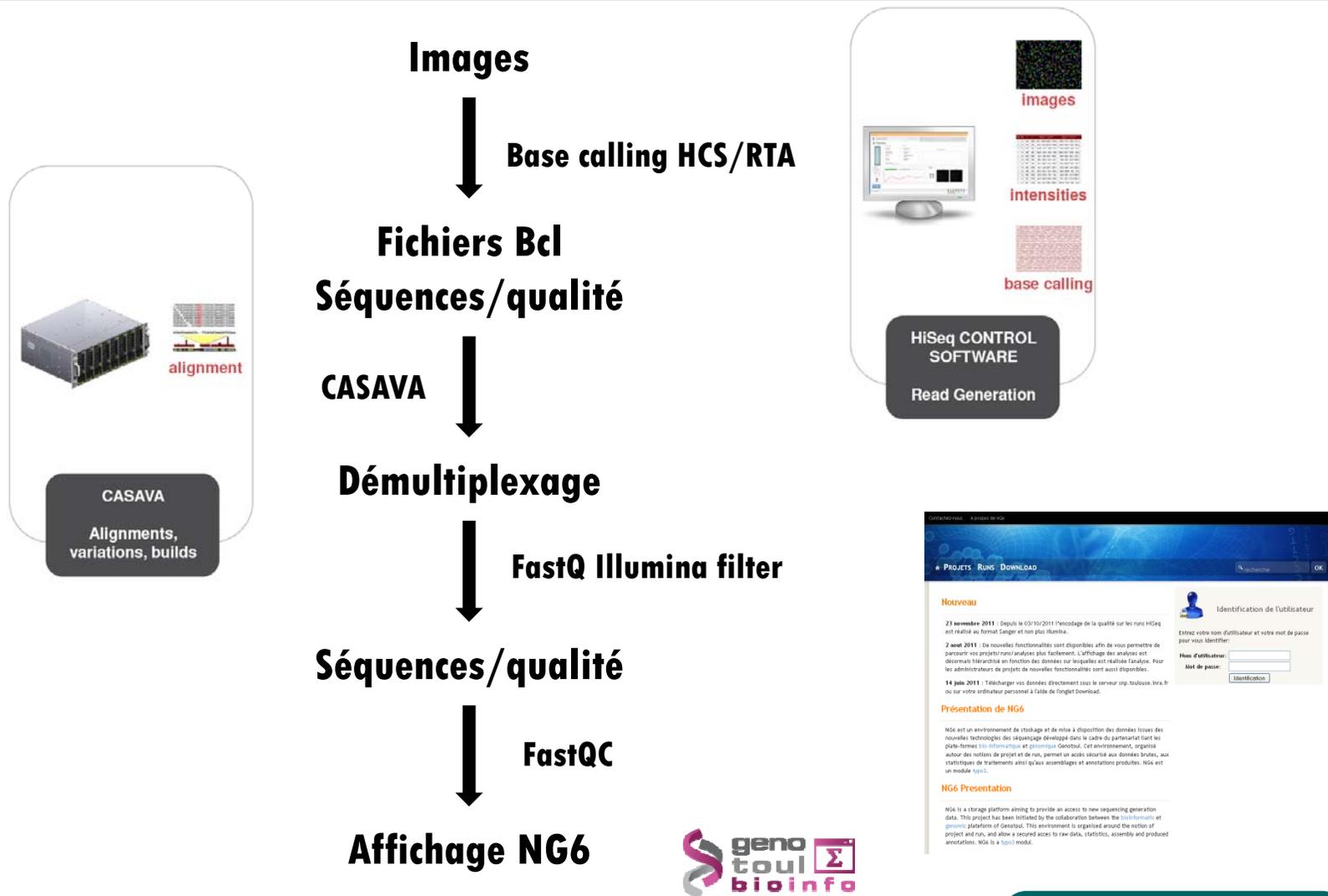
Kits Illumina TruSeq : le séquençage n'est pas orienté

- ✓ **Séquençage du brin codant ou du brin non codant**
- ✓ **RT réalisée avec des random primers**

Il existe d'autres kits permettant un séquençage orienté

- ✓ **Principe de ligation directe orientée d'un adaptateur sur l'ARN**
- ✓ **Détection des « starts » de transcription**
- ✓ **Identification du brin codant**
- ✓ **Possibilité de séquencer des petits ARN**

Analyses qualitatives des données



Analyses qualitatives des données : NG6

NG6 : un site de mise à disposition des données analysées : <http://ng6.toulouse.inra.fr/>

Contactez-nous A propos de nG6

PROJETS RUNS DOWNLOAD

recherche OK

Liste des projets auxquels vous avez accès :

Vous avez accès à 3 projets.
L'ensemble des données brutes et des résultats d'analyses occupent **31.97 Gb** d'espace disque pour l'ensemble des projets.

Afficher 10 éléments par page Rechercher :

Nom du projet	Description
BIOGECO Eucalyptus	Biogeco project for Eucalyptus species
BIOGECO Quercus	Biogeco project for Oak species
Demonstration	Projet public de démonstration

Affiche les éléments de 1 à 3 sur un total de 3 enregistrements

Analyses qualitatives des données : NG6

Runs > PhIX validation > ReadsStats

Analyse ReadsStats : Statistiques sur les lectures et leurs qualités.

Statistiques sur les lectures et sur leur qualité

Echantillons	Statistiques par position				Statistiques par séquence						
	Qualité	GC%	Ns	Contenue	Nombre de séquence	Qualité	GC%	Longueur	Niveau de duplication	K-mers	Séquences surreprésentées
s_1_1_qseq(1)	PASS	PASS	PASS	PASS	95456798	PASS	45(WARN)	101(PASS)	FAIL	WARN	PASS
s_1_2_qseq(2)	FAIL	PASS	PASS	PASS	95456798	PASS	44(WARN)	101(PASS)	FAIL	WARN	PASS

Aide Statistiques par position :

- Qualité : WARN = plus petit quartile d'une base inférieur à 10, ou si la mediane d'une base est inférieur à 25. FAIL = plus petit quartile d'une base inférieur à 5, ou si la mediane d'une base est inférieur à 20.
- GC% : WARN = GC% par position +/- 5% de la moyenne. FAIL = GC% par position +/- 10% de la moyenne.
- Ns : WARN = Ns% > 5%. FAIL = Ns% > 10%.
- Contenue : WARN = différence entre A et T ou G et C > 10% sur une position. FAIL = différence entre A et T ou G et C > 20% sur une position.

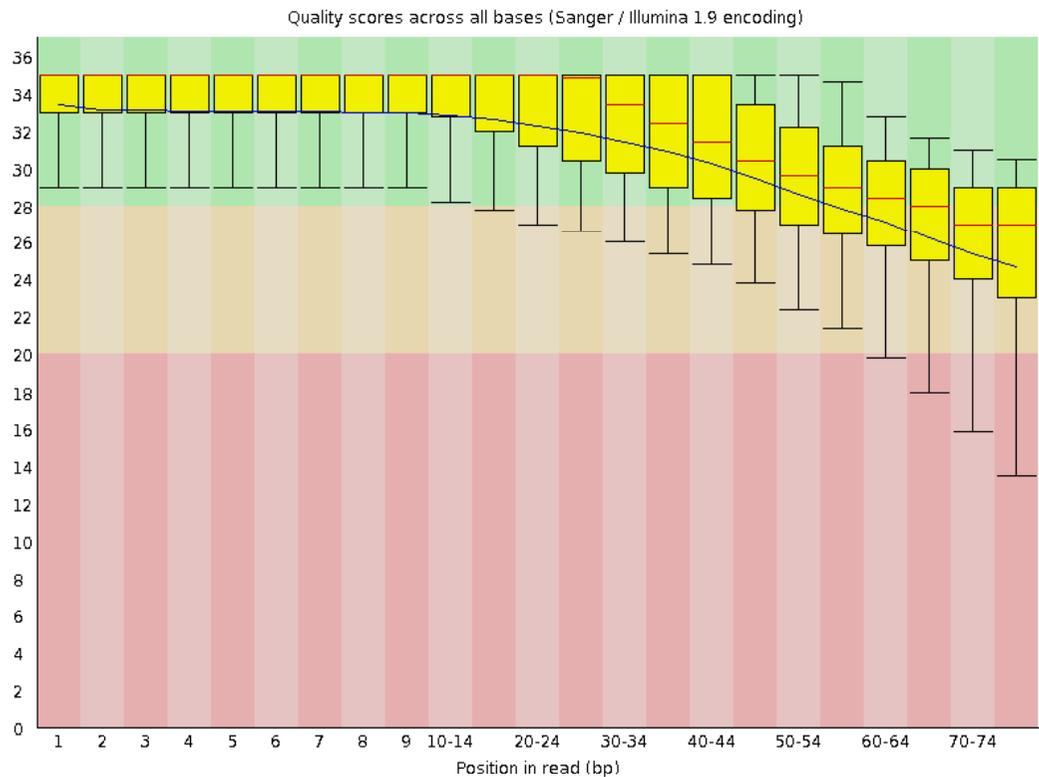
Aide Statistiques par séquence :

- Qualité : WARN = la qualité la plus observée < 27 (taux d'erreur de 0.2%). FAIL = la qualité la plus observée < 20 (taux d'erreur de 1%).
- GC% : WARN = plus de 15% des séquences ont un GC% différent de la distribution normale. FAIL = plus de 30% des séquences ont un GC% différent de la distribution normale.
- Longueur : WARN = toutes les séquences n'ont pas la même longueur. FAIL = une séquence a une longueur de 0pb.
- Niveau de duplication : WARN = plus de 20% des séquences sont non uniques. FAIL = plus de 50% des séquences sont non uniques.
- Séquences surreprésentées : WARN = une séquence représente plus de 0.1% du total. FAIL = une séquence représente plus de 1% du total.

Fichiers résultats

[fastqc.tar.gz](#)

Analyses qualitatives des données : NG6



Q10 : 1 base sur 10 est incorrecte

Q20 : 1 base sur 100 est incorrecte

Ligne rouge : médiane (doit être > 20)

Very good quality calls

Reasonable quality

Poor quality

Critères de qualité

➤ Critères de qualité

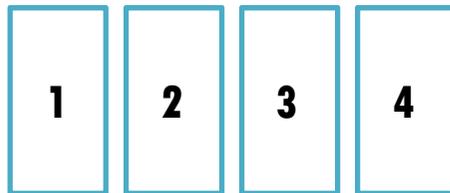
- ✓ **Le nombre de reads produites correspondant au nombre attendu**
- ✓ **Pas de contamination**
- ✓ **Longueur des reads correcte (100pb)**
- ✓ **Bonne qualité, mais ce n'est pas un critère rédhibitoire**
- ✓ **Bon alignement (re-séquençage avec génome de référence « propre ») : peu de reads non alignées**

Que faire ensuite ?

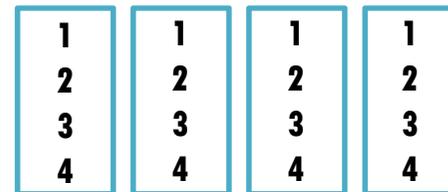
- **Différentes approches d'alignement des séquences:**
 - ✓ **De novo** : pas de génome de référence, transcriptome non disponible, très coûteux en terme calculs, résultats très variables
 - ✓ **Transcriptome de référence (en cours de développement)** : la plupart sont incomplets
 - ✓ **Génome de référence (en cours de développement)** : le plus utilisé, permet l'alignement de reads sur des parties non annotées, nécessite un « spliced aligner » pour eucaryotes
 - Etude à différents niveaux : gènes, transcripts, spécificité allélique
 - Découvertes de nouveaux transcripts, nouveaux isoformes, nouvelles structures de gènes (fusion)
- **IMPORTANT** : Discuter de la question biologique et du plan expérimental avec des Bioinformaticiens et Biostatisticiens **AVANT** de mettre en place l'expérience

Pourquoi ?

✓ **Multiplexage des échantillons possible sur différentes lignes**



**4 librairies séquençées sur 4
lignes, 1 librairie par ligne**



**4 librairies séquençées sur 4
lignes, 4 librairies par ligne**

Mêmes informations ?



GET
Génome et
Transcriptome

Exemples de biais connus du RNA-seq

Préparation des librairies

➤ Influence de la préparation des librairies

✓ Synthèse du cDNA avec des randoms primers:

○ la couverture du transcript n'est pas réellement aléatoire

- Spécificité de séquence de la polymérase?
- Réparation des extrémités?
- %GC ?

Published online 14 April 2010

*Nucleic Acids Research, 2010, Vol. 38, No. 12 e131
doi:10.1093/nar/gkq224*

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen^{1,*}, Steven E. Brenner² and Sandrine Dudoit^{1,3}

ABSTRACT

Generation of cDNA using random hexamer priming induces biases in the nucleotide composition at the beginning of transcriptome sequencing reads from the Illumina Genome Analyzer. The bias is independent of organism and laboratory and impacts the uniformity of the reads along the transcriptome. We provide a read count reweighting scheme, based on the nucleotide frequencies of the reads, that mitigates the impact of the bias.

✓ Amplification par PCR

- #### ○ Idée: supprimer totalement la PCR : actuellement, on réduit le nombre de cycles (10 cycles au lieu de 15)

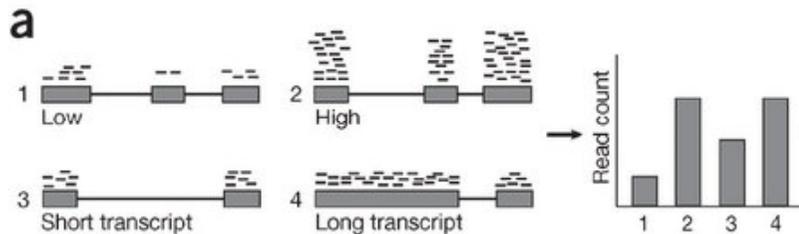
Longueur du transcript

➤ Influence de la longueur du transcript:

- Nombre total de reads d'un transcript : Exemple de calcul utilisé : RPKM (Read Per Kb per Million reads mapped)

$$\frac{\text{reads alignés sur gène d'intérêt}}{\text{nombre total de read alignés} \times \text{longueur du transcript (kb)}}$$

- A un même niveau d'expression, les transcrits longs auront plus de reads que les transcrits courts, donc l'expression différentielle des longs transcrits sera plus facilement identifiée (dépend de la profondeur de séquençage)



[Biol Direct](#). 2009 Apr 16;4:14.

Transcript length bias in RNA-seq data confounds systems biology.

Oshlack A, Wakefield MJ.

Abstract

Background: Several recent studies have demonstrated the effectiveness of deep sequencing for transcriptome analysis (RNA-seq) in mammals. As RNA-seq becomes more affordable, whole genome transcriptional profiling is likely to become the platform of choice for species with good genomic sequences. As yet, a rigorous analysis methodology has not been developed and we are still in the stages of exploring the features of the data.

Results: We investigated the effect of transcript length bias in RNA-seq data using three different published data sets. For standard analyses using aggregated tag counts for each gene, the ability to call differentially expressed genes between samples is strongly associated with the length of the transcript.

Conclusion: Transcript length bias for calling differentially expressed genes is a general feature of current protocols for RNA-seq technology. This has implications for the ranking of differentially expressed genes, and in particular may introduce bias in gene set testing for pathway analysis and other multi-gene systems biology analyses.

Reviewers: This article was reviewed by Rohan Williams (nominated by Gavin Huttley), Nicole Cloonan (nominated by Mark Ragan) and James Bullard (nominated by Sandrine Dudoit).

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 5 2011, pages 662–669
doi:10.1093/bioinformatics/btr005

Gene expression

Advance Access publication January 19, 2011

Length bias correction for RNA-seq data in gene set analyses

Liyao Gao^{1,†}, Zhidong Fang^{2,†}, Kui Zhang¹, Degui Zhi¹ and Xiangqin Cui^{1,*}

¹Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294 and ²Biostatistics Program, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA

Associate Editor: Ivo Hofacker

Autres biais

- **Couverture non aléatoire le long du transcrit**
- **Biais selon une spécificité de position et de séquence :**

Robert et al. Genome Biology, 2011, 12:R22

- **Biais selon l'aligneur utilisé**
- **Alignement multiple de certaines reads**

✓ **Combien de réplicats?**

- **Le maximum, à discuter avec les biostatisticiens....**

✓ **Combien de reads pour chaque échantillon?**

- **Entre 30M et 100M (dépendant de l'étude, (catalogue ou quantification?), et des € €)**

RNA-sequence analysis of human B-cells

Jonathan M. Toung,¹ Michael Morley,² Mingyao Li,³ and Vivian G. Cheung^{2,4,5,6}

¹Genomics and Computational Biology Program, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; ²The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA; ³Department of Biostatistics and Department of Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; ⁴Department of Pediatrics and Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; ⁵Howard Hughes Medical Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

✓ **Quelle est la longueur idéale pour les reads?**

- **Plus les reads sont longs, plus l'alignement se fait facilement; 2X100 bp actuellement mais passage à 2X150 bp bientôt**

✓ **Single read ou paired-end?**

- **Paired end facilite l'alignement des séquences, permet de détecter plus facilement les insertions/délétions et les jonctions entre les exons**

✓ **Déplétion des ARN ribosomaux?**

- ✓ **Séquençage des ARNs poly-A : forte sensibilité et forte précision pour l'étude du profil d'expression
MAIS : Pas de vision complète du génome**
- ✓ **Séquençage des ARN totaux déplétés en ARN ribosomaux : visibilité des ARNs non-codants et non-polyA
MAIS : Il faut augmenter la profondeur pour avoir une bonne sensibilité de détection**



GET
Génome et
Transcriptome

Comparaisons Microarrays/NGS

Malone and Oliver *BMC Biology* 2011, 9:34
<http://www.biomedcentral.com/1741-7007/9/34>


 BMC Biology
 Incorporating Journal of Biology

REVIEW Open Access

Microarrays, deep sequencing and the true measure of the transcriptome

John H Malone* and Brian Oliver

British Yeast Group Meeting 2008

Next-generation sequencing: applications beyond genomes

Samuel Marguerat¹, Brian T. Wilhelm² and Jürg Bähler^{1,3}

Cancer Research UK, Fission Yeast Functional Genomics Group, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, U.K.

New Methods in Cardiovascular Biology

Deep mRNA Sequencing for In Vivo Functional Analysis of Cardiac Transcriptional Regulators
 Application to $G\alpha_q$

Scot J. Matkovich, Yan Zhang, Derek J. Van Booven, Gerald W. Dorn II

Microarray

- Pas de détection d'épissage alternatifs des transcrits sauf pour l'humain
- Matériel moins coûteux que le séquenceur

RNAseq

- A comparer avec les Tiling array (sondes sur toute la longueur du génome)
- Séquençage du cDNA double brin donc perte de l'information brin spécifique (RNA directionnel)
- Hétérogénéité dans le recouvrement de certaines zones d'expression quand on fait du séquençage
- Etude des transcrits faiblement exprimés

Conclusion : les données issues de microarrays et séquençage sont globalement similaires, pour les transcrits présents en grande quantité

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni, Christopher E. Mason, Shrikant M. Mane, et al.

Genome Res. 2008 18: 1509-1517 originally published online June 11, 2008
Access the most recent version at doi: [10.1101/gr.079558.108](https://doi.org/10.1101/gr.079558.108)

○ **Résultats:** comparaison du nombre de gènes différentiellement exprimés dans les 2 technologies:

- Parmi les gènes retrouvés dans une ou l'autre des technos, en qPCR:
 - 4/5 confirmés pour NGS (le 5^{ème} est un faux-positif)
 - 2/6 confirmés pour microarray

Donc, cela favorise les résultats NGS

○ **Conclusion:**

- Séquençage: très reproductible, identification de 40% de gènes différentiellement exprimés en plus par rapport aux microarrays
- Puce U133: les sondes ne couvrent qu'une petite partie du gène, donc pas de découverte de nouveaux transcrits ou d'épissages alternatifs

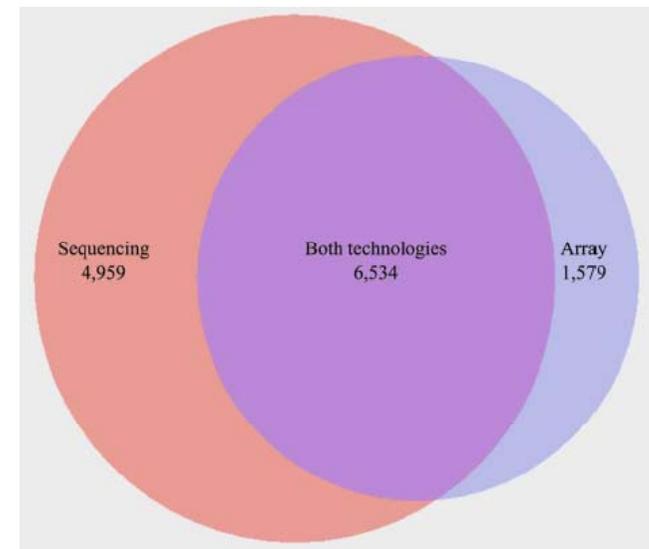


Figure 5. A Venn diagram summarizing the overlap between genes called as differentially expressed from the (left circle) sequence data and from the (right circle) array. The number of genes called by both technologies is indicated by the overlap between the two circles.

A Comparison of Next Generation Sequencing and Microarrays for Whole Transcriptome Expression Profiling

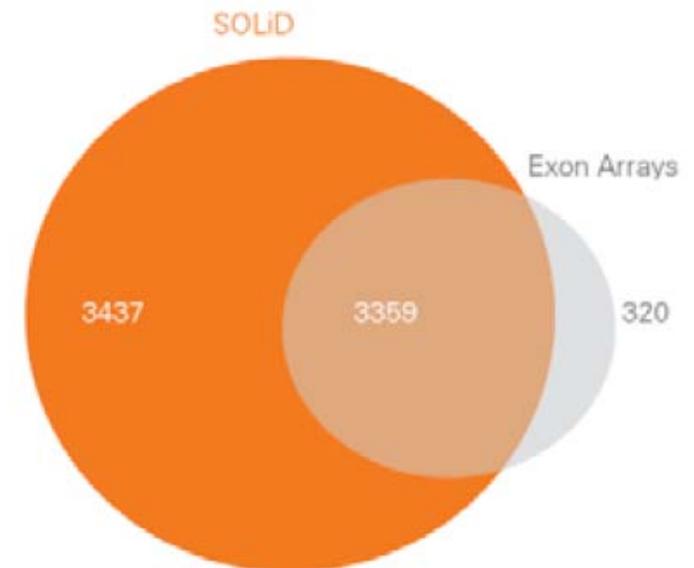
 applied biosystems

Penn Whitley, Andrew Lemire, Joel Brockman, Sheila Heater, Jeff Schageman, Jian Gu, Kristi Lea, Luming Qu, Charmaine San Jose, Natalie Hernandez Kelli Bramlett, Diane Ilsley and Robert Setterquist, Life Technologies/Ambion R&D, 2130 Woodward, Austin, TX, USA, 78744

○ Résultats:

- **Bonne concordance entre NGS et microarrays**
- **MAIS: NGS est plus précis et plus sensible → permet de faire plus de découvertes**
- **Microarrays sont assez bon marché et facile à utiliser; obtention des résultats assez rapidement**
- **Ces 2 techniques sont complémentaires**

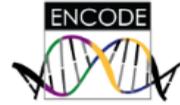
Figure 2. Concordance of Differentially Expressed Transcripts



Bilan de la comparaison

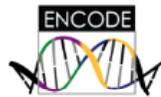
	Microarrays	Séquençage
Génome de référence	OUI Pour dessiner les oligonucléotides	NON OUI pour l'alignement
Technique	Hybridation avec les sondes	Accès direct à la séquence
Système de détection	Lecture de la fluorescence par un scanner: problème de détection dans les faibles et les fortes intensités	Précision de la lecture à une base donc étude de SNP possible
Coût	+	+++
Reproductibilité	+++	+++
Obtention des résultats	Rapide environ 4 jours	Assez long environ 3 semaines
Quantité de données générées	+	+++
Outils d'analyses statistiques	Bien définis	Encore en développement
Analyses	Etude de gènes différentiellement exprimés	Épissage alternatif, découverte de nouveaux exons, quantification de transcrits

Pour plus d'informations...



Encyclopedia of DNA Elements

<http://encodeproject.org/ENCODE/protocols/dataStandards/>



Guidelines for Experiments

Current Guidelines:

[ChIP-seq, ChIP-chip, DNase-seq, FAIRE-seq and DNase Standards v2.0 \(July 2011\)](#)

[RNA Standards v1.0 \(May 2011\)](#)

[RIP Standards v2.0 \(Jan 2012\)](#)

Infos sur :

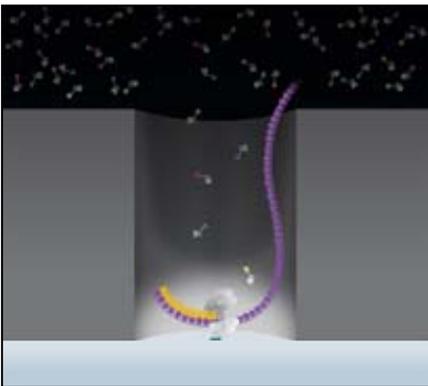
- ✓ Réplicats
- ✓ Nature du matériel
- ✓ Stats
- ✓ Etc...



GET
Génome et
Transcriptome

NGS 3^{ème} génération... l'avenir?

Pacific BioSciences



PacBio RS

- + Séquençage molécule unique
- + Pas d'étape de PCR
- + Taille des séquences

- Beaucoup d'erreur pour l'instant (>10%)

Déjà disponible

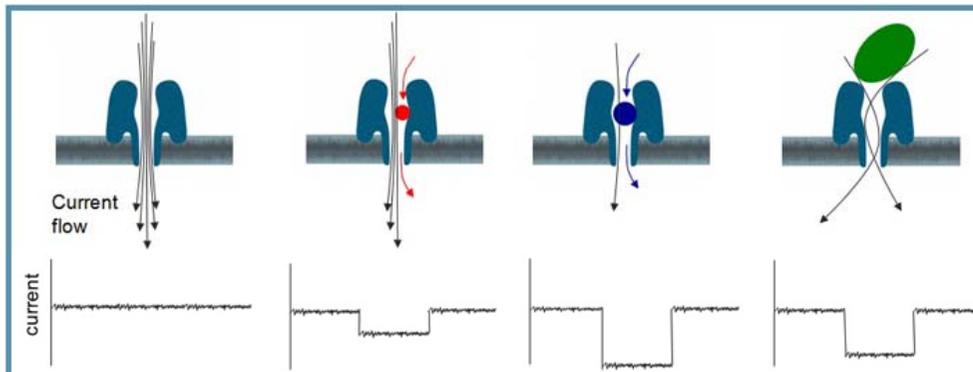
<http://www.pacificbiosciences.com>

Oxford Nanopore

MinION



GridION



- + Séquençage molécule unique
- + Pas d'étape de PCR
- + Taille des séquences

- Beaucoup d'erreur pour l'instant (>5%)

? Prix...

? Débit

? Date de sortie

<http://www.nanoporetech.com>



GET
Génome et
Transcriptome

Merci !