RNA-seq

Olivier Bouchez

Responsable NGS

GeT-PlaGe

genotoul

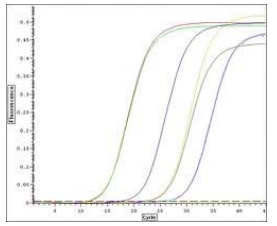
GeT Génome et Transcriptome

Qu’est-ce que le RNA-seq ?

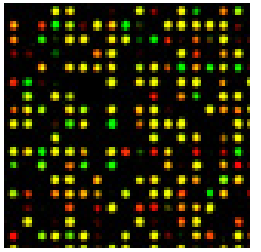
Introduction

Quelques applications les plus utilisées pour l'étude de l'expression des gènes et leurs différences :

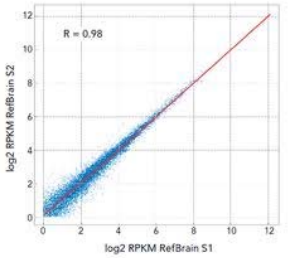
✓**PCR quantitative : on étudie un seul gène à la fois par réaction**
(cf présentation de Jean-José Maoret)



✓**Puces à ADN : on étudie des centaines/milliers de gènes à la fois, mais tous connus => on choisit lesquels étudier**
(cf présentation de Véronique Le Berre)



✓**RNAseq : on étudie tous les gènes à la fois sans à priori**



Introduction

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-séquencage : séquençage d'un fragment d'ADN et comparaison du résultat obtenu avec une séquence de référence connue**

- ✓ **Séquencage *de novo* : séquençage d'un génome pour lequel il n'existe pas de séquence de référence, détermination d'une séquence inconnue**

GeT Génome et Transcriptome

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**
- ✓ ….

BMC Genomics. 2013 Oct 5;14(1):683. [Epub ahead of print]

"A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids"

Davey MW, Gudimella R, Harikrishna JA, Sin LW, Khalid N, Keulemans J.

Abstract

BACKGROUND: Modern banana cultivars are primarily interspecific triploid hybrids of two species, Musa acuminata and Musa balbisiana, which respectively contribute the A- and B-genomes. The M. balbisiana genome has been associated with improved vigour and tolerance to biotic and abiotic stresses and is thus a target for Musa breeding programs. However, while a reference M. acuminata genome has recently been released (Nature 488:213--217, 2012), little sequence data is available for the corresponding B-genome.To address these problems we carried out Next Generation gDNA sequencing of the wild diploid M. balbisiana variety 'Pisang Klutuk Wulung' (PKW). Our strategy was to align PKW gDNA reads

PLoS One. 2013 Oct 1;8(10):e74183. doi: 10.1371/journal.pone.0074183.

High-throughput RNA sequencing of pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants.

Howard BE, Hu Q, Babaoglu AC, Chandra M, Borghi M, Tan X, He L, Winter-Sederoff H, Gassmann W, Veronese P, Heber S.

Department of Computer Science, North Carolina State University, Raleigh, North Carolina, United States of America.

Abstract

We report the results of a genome-wide analysis of transcription in Arabidopsis thaliana after treatment with Pseudomonas syringae pathovar tomato. Our time course RNA-Seq experiment uses over 500 million read pairs to provide a detailed characterization of the response to infection in both susceptible and resistant hosts. The set of observed differentially expressed genes is consistent with previous studies, confirming and extending existing findings about genes likely to play an important role in the defense response to Pseudomonas syringae. The high coverage of the Arabidopsis transcriptome resulted in the discovery of a surprisingly large number of alternative splicing (AS) events - more than 44% of multi-exon genes showed

Proc Natl Acad Sci U S A. 2013 Oct 28. [Epub ahead of print]

Differential expression of olfactory genes in the southern house mosquito and insights into unique odorant receptor gene isoforms.

Leal WS, Choo YM, Xu P, da Silva CS, Ueira-Vieira C.

Department of Molecular and Cellular Biology, University of California, Davis, CA 95616.

Abstract

The southern house mosquito, Culex quinquefasciatus, has one of the most acute and eclectic olfactory systems of all mosquito species hitherto studied. Here, we used Illumina sequencing to identify olfactory genes expressed predominantly in antenna, mosquito's main olfactory organ. Less

Front Genet. 2013 Aug 2;4:145. doi: 10.3389/fgene.2013.00145. eCollection 2013.

Mammalian miRNA curation through next-generation sequencing.

Brown M, Suryawanshi H, Hafner M, Farazi TA, Tuschl T.

Laboratory of RNA Molecular Biology, Howard Hughes Medical Institute, The Rockefeller University New York, NY, USA.

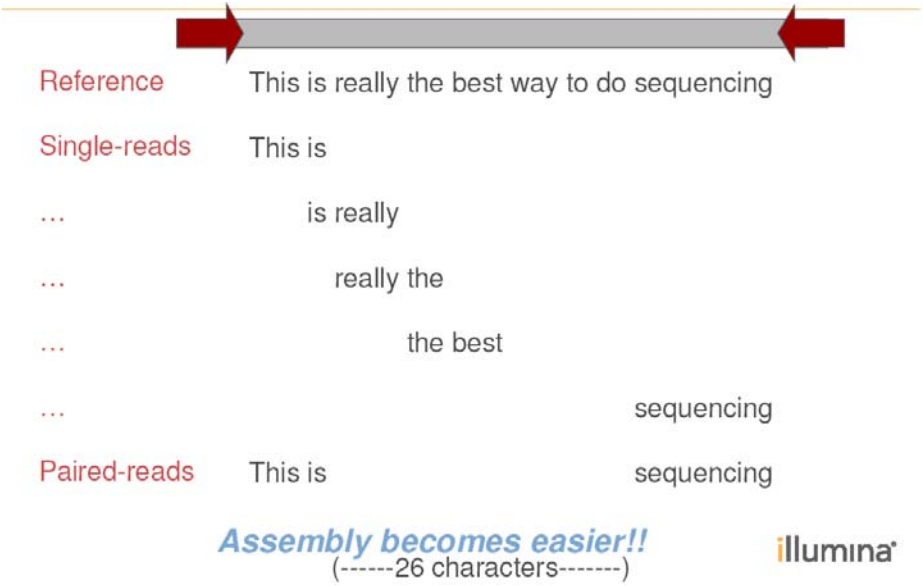
Abstract

Characteristic small RNA biogenesis processing patterns are used for the discovery of novel microRNAs (miRNAs) from next-generation sequencing data. Here, we highlight and discuss key criteria for mammalian - specifically human - miRNA database curation based on small RNA sequencing data. Sequence reads obtained from small RNA cDNA libraries are aligned to reference genomic regions, and miRNA genes are revealed by their

Introduction

- ✓ **Lecture single read (SR) et paired end (PE)**

- ✓ **PE : facilite l'alignement des séquences sur le génome de référence et/ou l'assemblage**



GeT Génome et Transcriptome

Exemples de séquenceurs NGS

Séquençage par synthèse
Longueur max: 2 x 300 pb
Débit max : >15 GB
Temps de run : 39 h

Illumina MiSeq

Séquençage par synthèse
1 flowcell, 8 lanes/ flow cell, Longueur : 2 x 150 pb
Débit : >700 GB/flowcell
Temps de run : 2,5 jours

Illumina HiSeq3000

Ion Torrent S5

Semiconductor sequencing
Longueur : 100 pb
Débit : 80 millions de séquences
Temps de run : quelques heures

Ion Torrent PGM

Semiconductor sequencing
Longueur : 200-400 pb
Débit : 5 millions de séquences
Temps de run : quelques heures

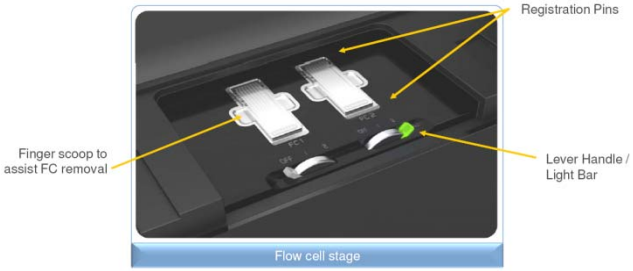
PacBio RSII

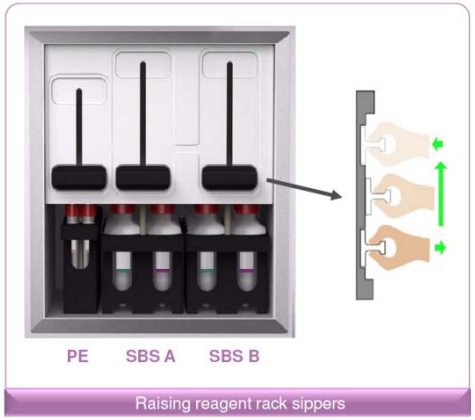
Single Molcule Real Time Sequencing
Longueur : >10 kb
Débit : 70000 séquences
Temps de run : 6 h

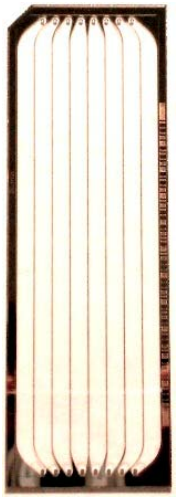
Présentation d'un séquenceur NGS : Illumina HiSeq3000

Spécifications :

- ✓ **700 Gb**
- ✓ **Paired-end (2x150 bp) : ~600 millions de séquences/lane**







1 flowcell = une lame = 8 lignes

Worflow

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

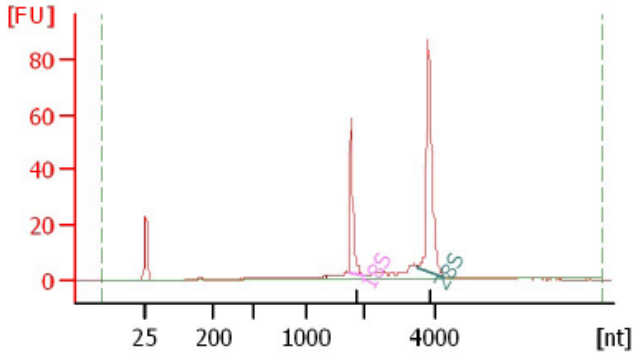
Contrôles qualité des ARN

- **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 Rétrotanscription**

- **Evaluation de l'état de dégradation de l'ARN par un dosage au BioAnalyzer (Agilent)**

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



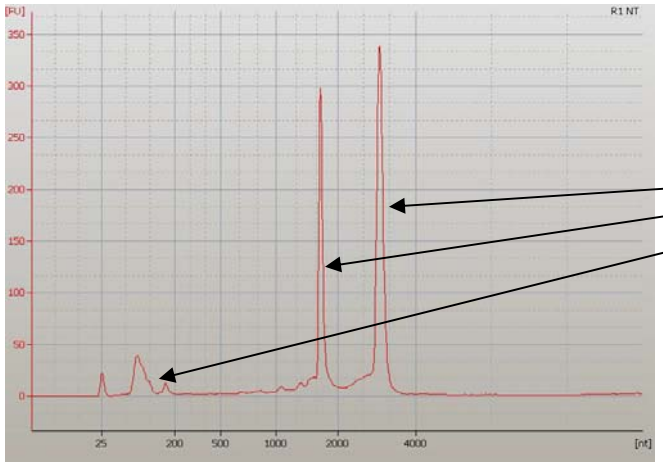
Overall Results for sample 6 : 9-H-

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

ARNs

La plupart des ARNs sont des ARN ribosomiques => 90 à 98 % des ARNs

Nécessité de les éliminer pour pouvoir étudier les ARN messagers

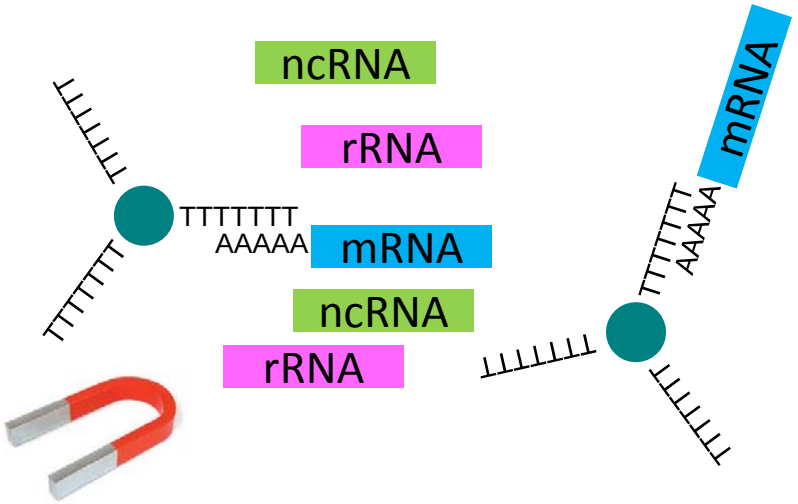


ARN ribosomiques

Elimination des ARN ribosomiques

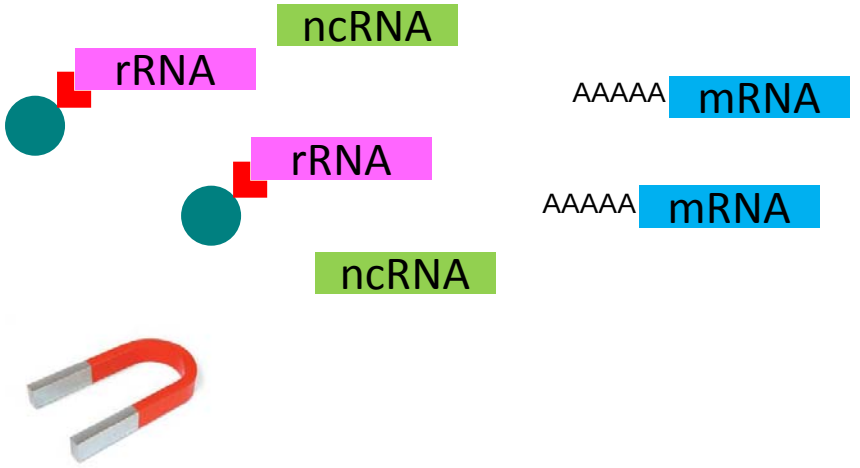
Exemples de protocoles permettant de sélectionner les ARN messagers :

Utilisation de billes magnétiques poly-T



✓Valable pour les eucaryotes, mais pas pour les bactéries (pas d'ARNm poly-A)
✓Très efficace
✓Sélection seulement des ARN poly-A

Utilisation de billes magnétiques permettant de sélectionner les ARN ribosomiques



✓Récupération des ARN messagers et des ARN non-codons
✓Moins efficace
✓Nécessité d'optimisation pour chaque espèce

Exemple de préparation des librairies

Kits Illumina TruSeq

1µg ARN total

↓

Purification et fragmentation des ARNm

↓

Synthèse du cDNA double brin

↓

Réparation des extrémités

↓

Adénylation des extrémités 3'

↓

Ligation des adaptateurs

↓

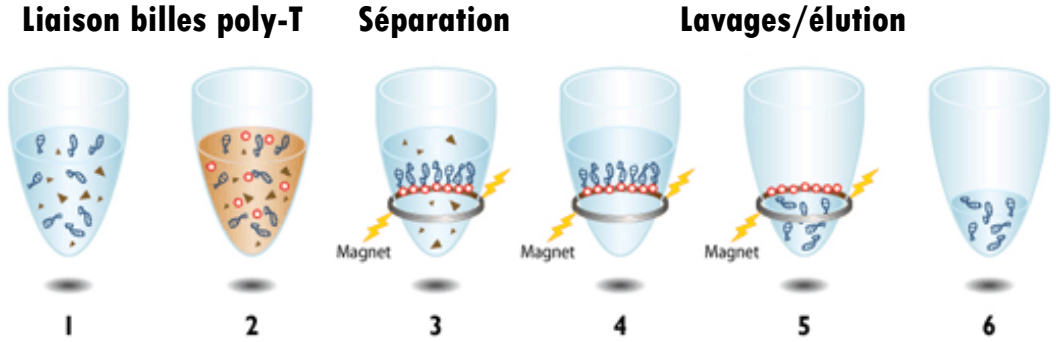
Amplification par PCR

↓

Vérification des librairies

Exemple de préparation des librairies

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



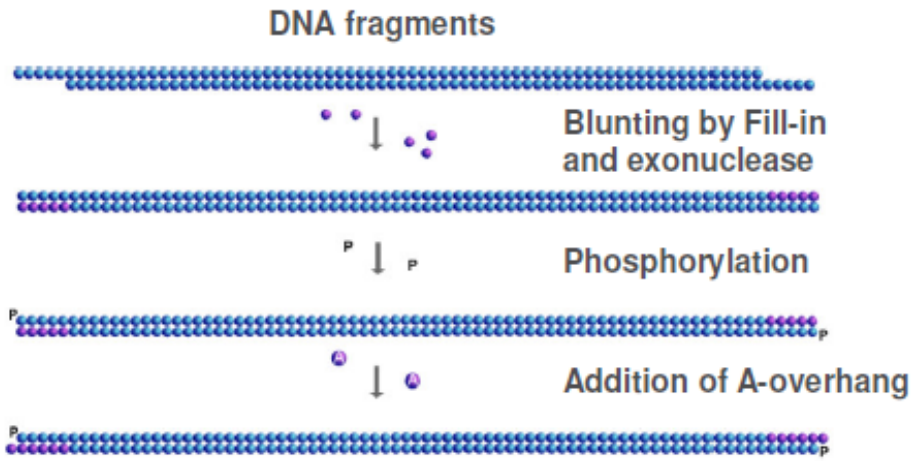
	Echantillon non traité	Echantillon traité			
		Pourcentage d'élimination des ARNr			
		90%	98%	99%	99,90%
ARNr	98%	83%	50%	33%	5%
Autres	2%	17%	50%	67%	95%

Kits Illumina TruSeq

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

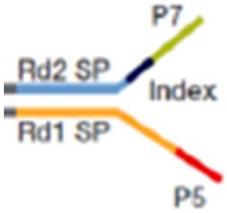
Exemple de préparation des librairies

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

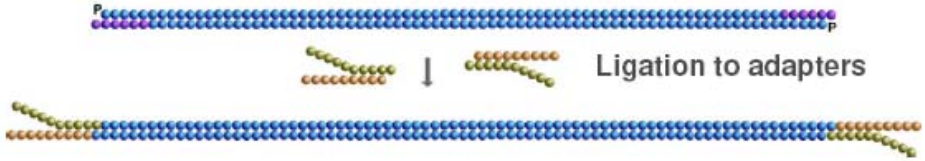


Exemple de préparation des librairies

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

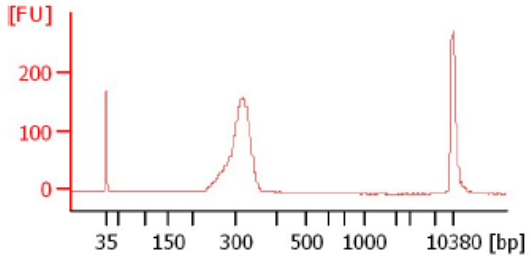


  - **P5 et P7: Fixation sur la flowcell**
  - **SP 1 et 2: primers de séquençage**
  - **Tag: index (6 bases): multiplexage par 24**



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

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



Informations importantes

✓**Important : le séquençage sera ori**

Curr Genomics. 2013 May;14(3):173-81. doi: 10.2174/1389202911314030003.

Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling.

Mills JD, Kawahara Y, Janitz M.

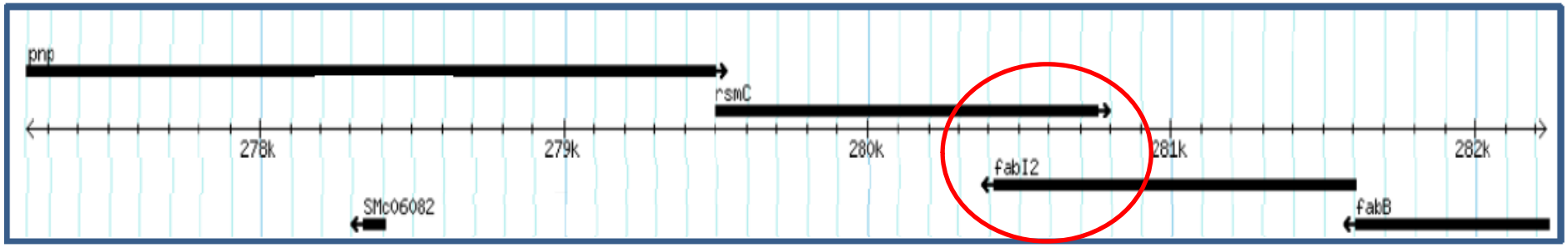
School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia.

Abstract

RNA-Seq is a recently developed sequencing technology, that through the analysis of cDNA allows for unique insights into the transcriptome of a cell. The data generated by RNA-Seq provides information on gene expression, alternative splicing events and the presence of non-coding RNAs. It has been realised non-coding RNAs are more then just artefacts of erroneous transcription and play vital regulatory roles at the genomic, transcriptional and translational level. Transcription of the DNA sense strand produces antisense transcripts. This is known as antisense transcription and often results in the production of non-coding RNAs that are complementary to their associated sense transcripts. Antisense tran-scription has been identified in bacteria, fungi, protozoa, plants, invertebrates and mammals. It seems that antisense tran-scriptional 'hot spots' are located around nucleosome-free regions such as those associated with promoters, indicating that it is likely that antisense transcripts carry out important regulatory functions. This underlines the importance of identifying the presence and understanding the function of these antisense non-coding RNAs. The information concerning strand ori-gin is often lost during conventional RNA-Seq; capturing this information would substantially increase the worth of any RNA-Seq experiment. By manipulating the input cDNA during the template preparation stage it is possible to retain this vital information. This forms the basis of strand-specific RNA-Seq. With an ability to unlock immense portions of new in-formation surrounding the transcriptome, this cutting edge technology may hold the key to developing a greater under-standing of the transcriptome.

Pourquoi est-ce important :

✓**Ex chez les bactéries, le génome est très dense en gènes et plusieurs gènes peuvent se chevaucher, il faut donc savoir quand on séquence à quel gène correspond la séquence obtenue**



✓**Identification de transcrits « reverses » impliqués dans la régulation de la traduction des ARN**

Worflow

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

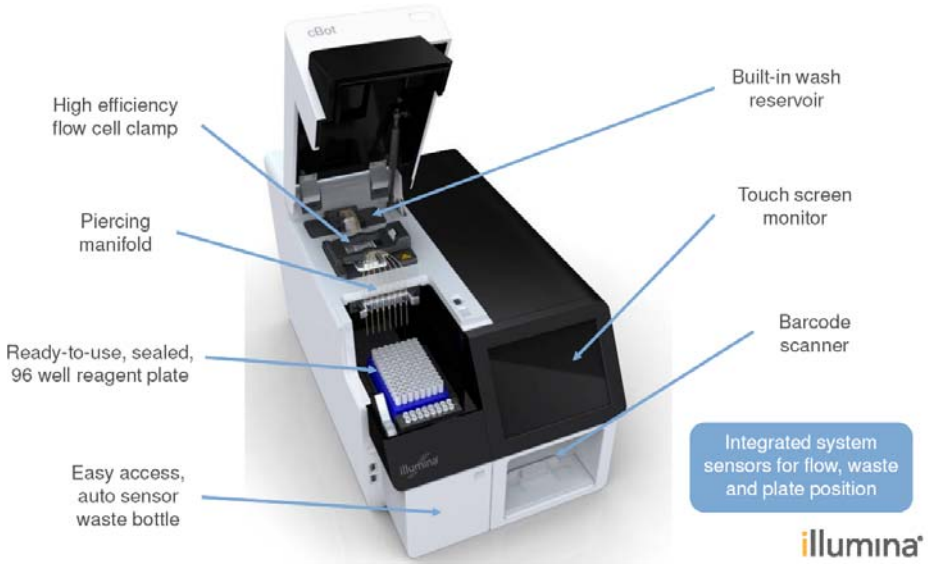
Perform sequencing
Generate base calls

4 Data Analysis

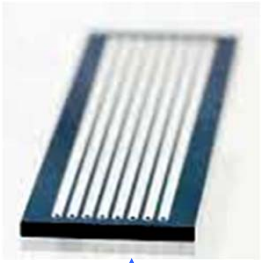
Images
Intensities
Reads
Alignments

Génération des clusters

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



Flowcell

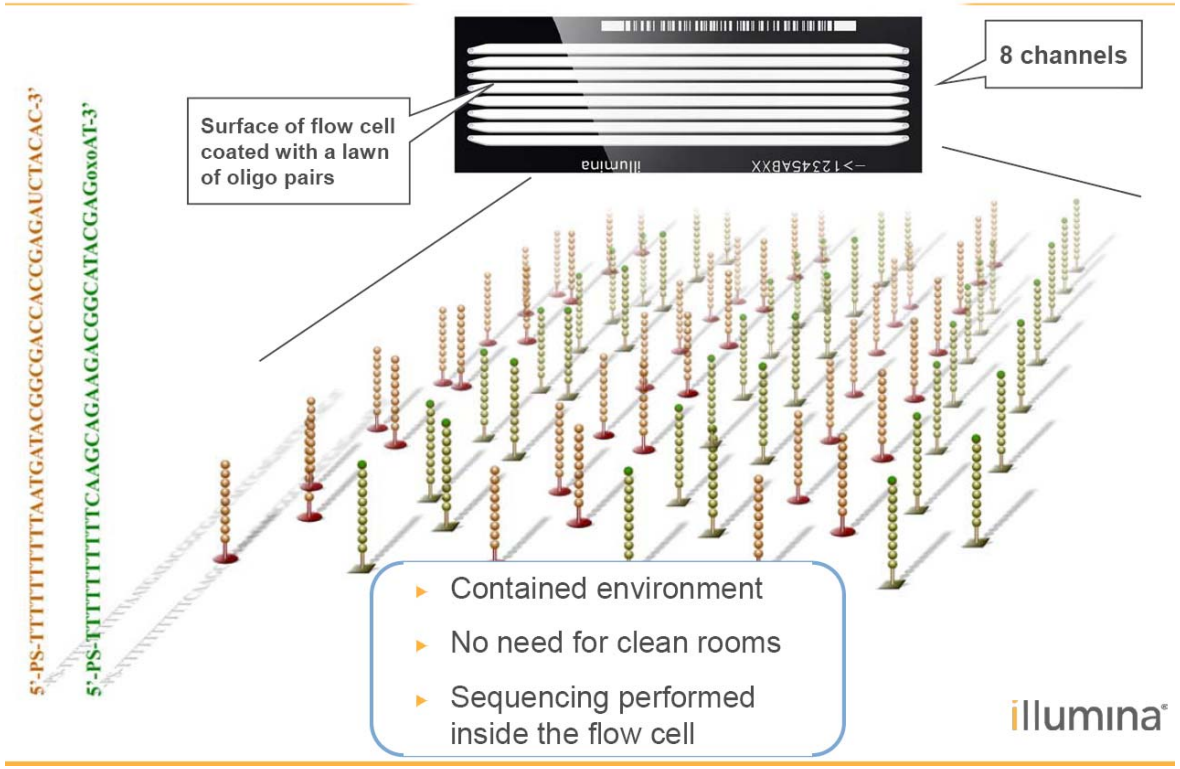


Librairies

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

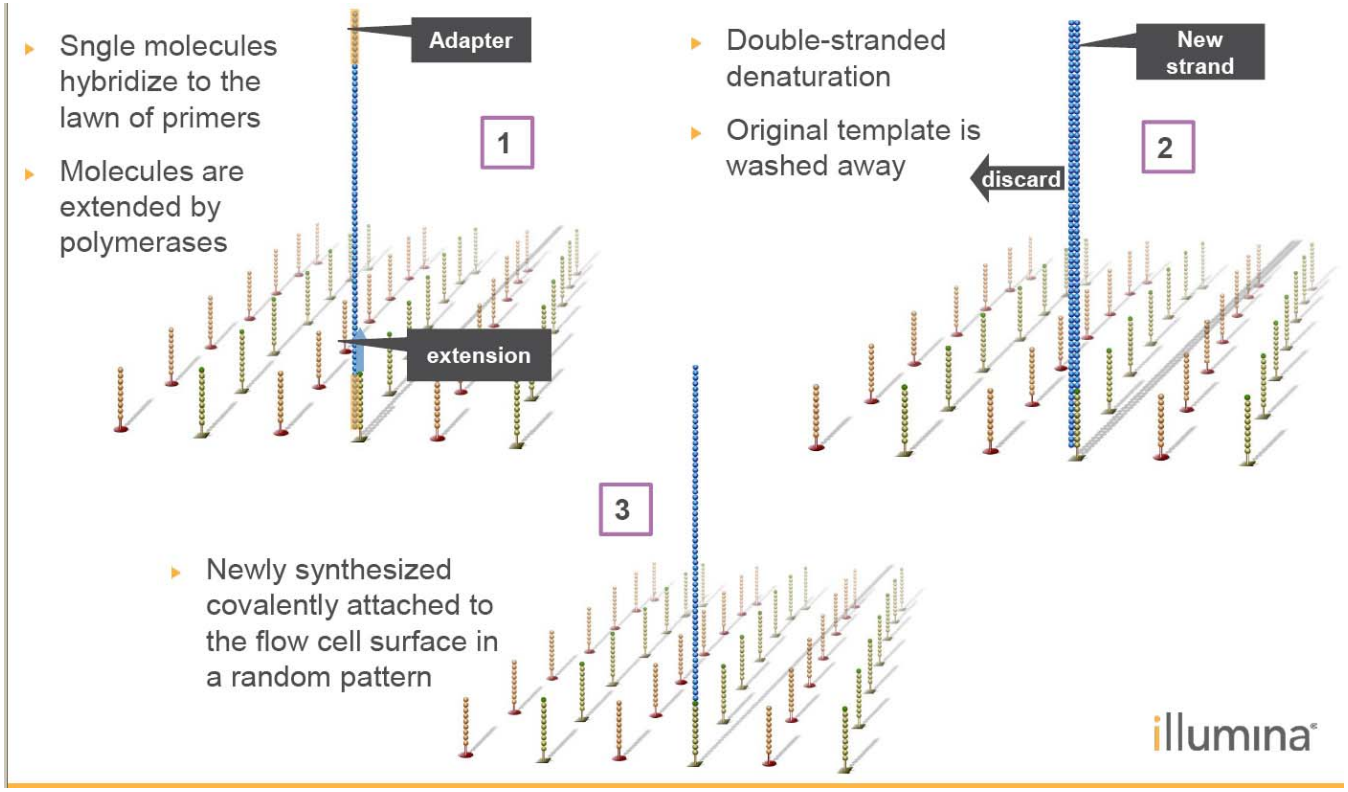
Génération des clusters

Design de la flowcell



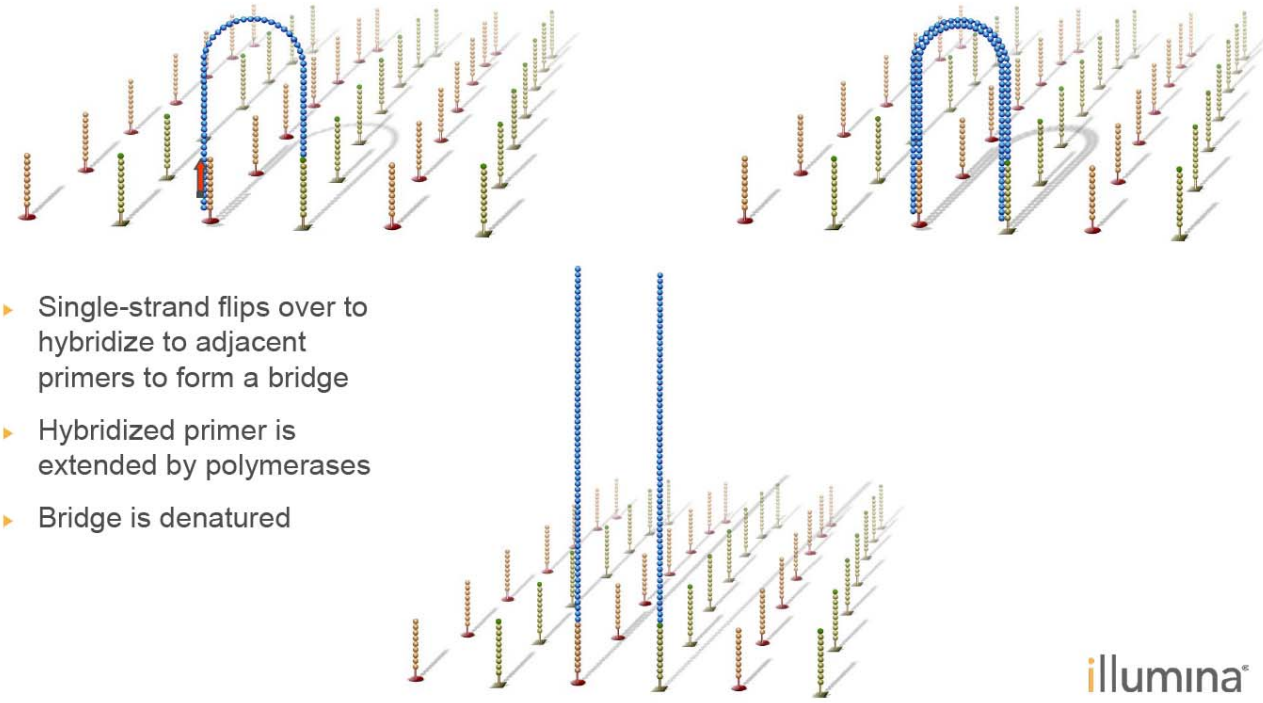
Génération des clusters

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



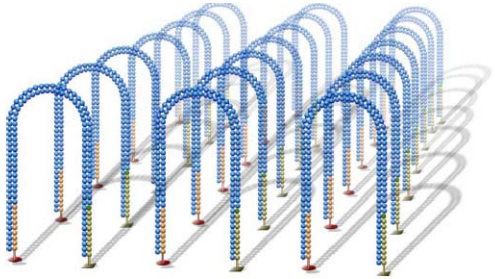
Génération des clusters

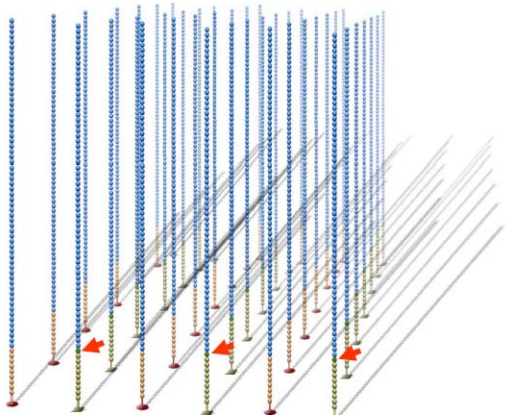
- ✓ **Formation d'un pont**
- ✓ **Synthèse du brin complémentaire**
- ✓ **Dénaturation**

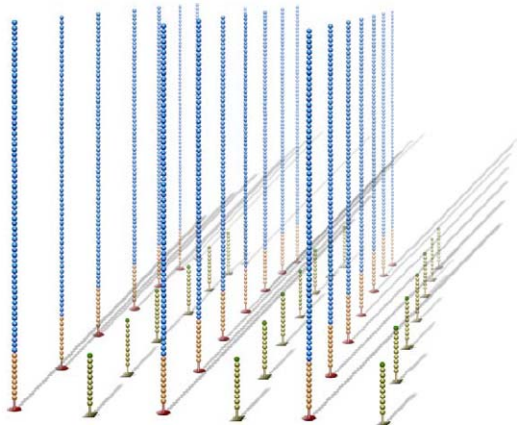


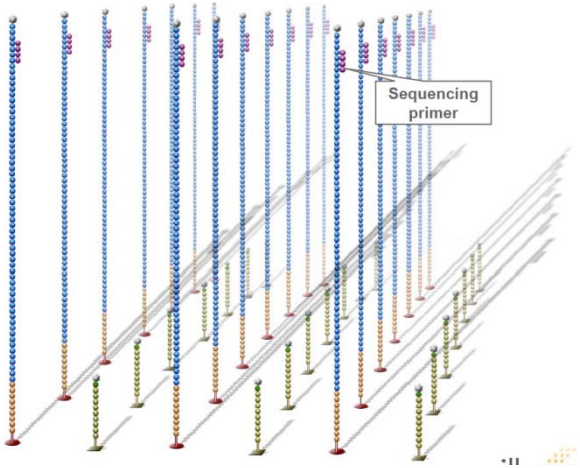
Génération des clusters

- 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**

Worflow

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

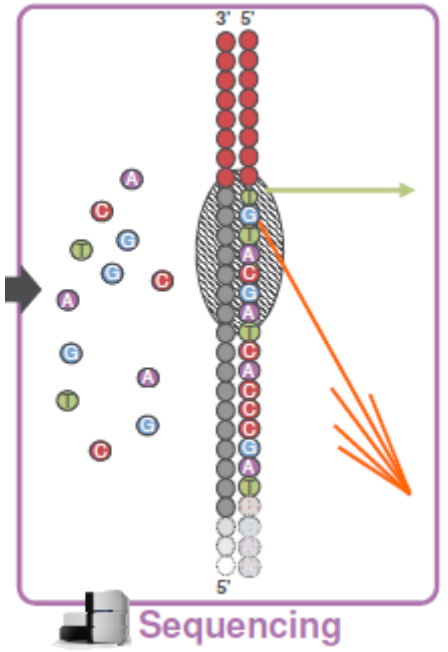
Perform sequencing
Generate base calls

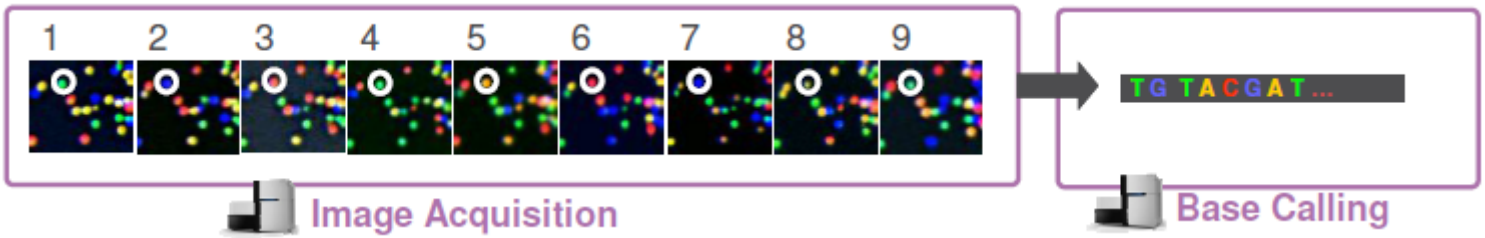
4 Data Analysis

Images
Intensities
Reads
Alignments

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





Worflow

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

Analyses qualitatives des données

Images

Base calling HCS/RTA

Fichiers Bcl

Séquences/qualité

CASAVA

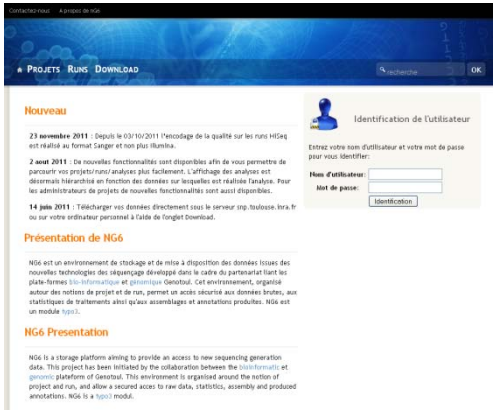
Démultiplexage

FastQ Illumina filter

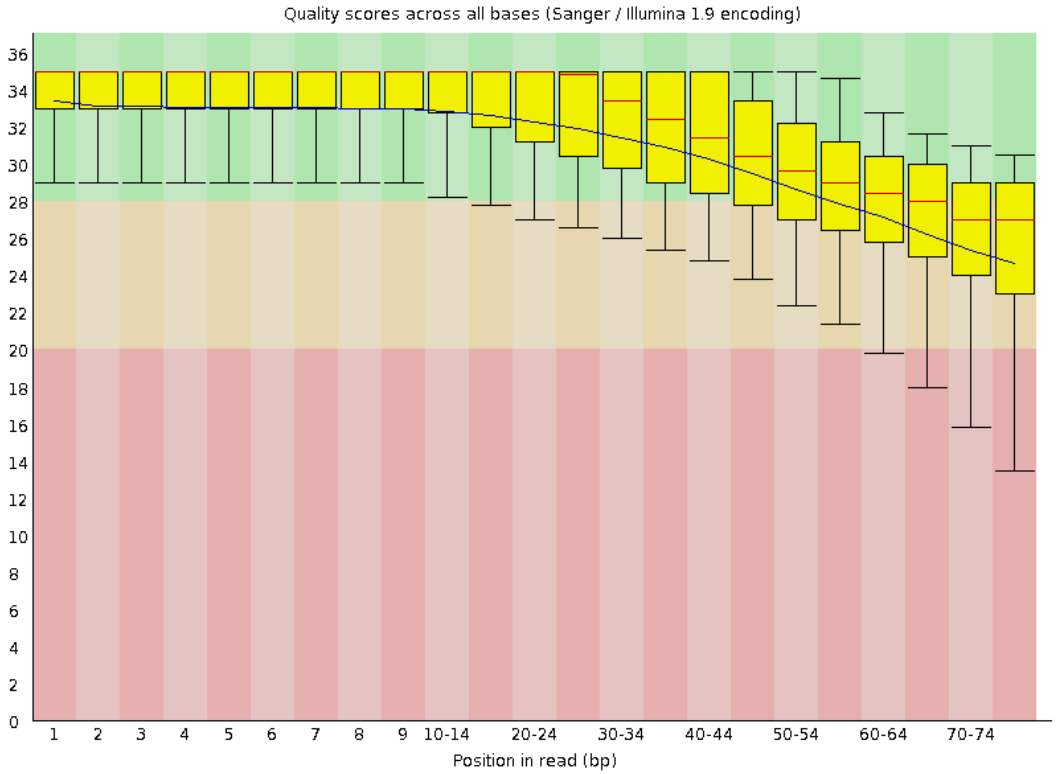
Séquences/qualité

FastQC

Affichage NG6



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

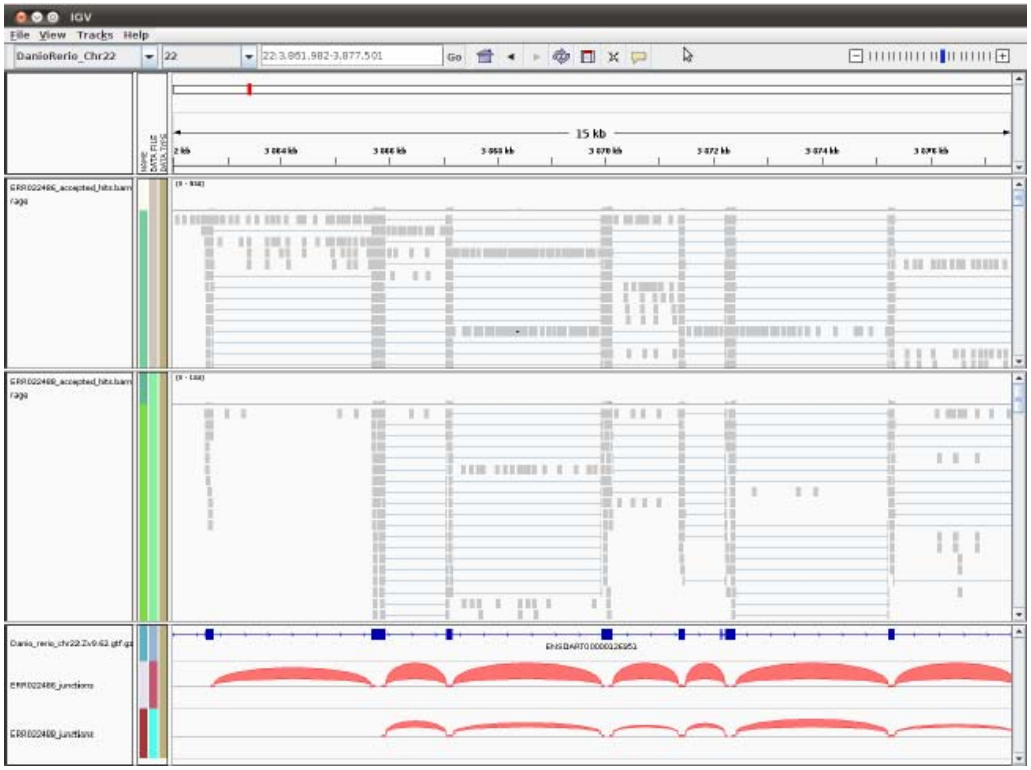
Critères de qualité

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

Analyses bioinformatiques et statistiques

Attention : pour toutes les analyses RNA-seq, cette partie est la face cachée de l'iceberg

Un exemple de ce que l'on peut faire avec ces données : alignement sur un génome de référence



Comptage du nombre de reads sur chaque gène, normalisation, comparaisons et interprétation

Analyses bioinformatiques et statistiques

- **Différentes approches d'alignement des séquences:**
  - ✓ **De novo : pas de génome de référence, transcriptome non disponible, très coûteux en terme de calculs, résultats très variables**
  - ✓ **Transcriptome de référence : la plupart sont incomplets**
  - ✓ **Génome de référence : 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 allèlique**

    → **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 <u>AVANT</u> 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 ?

Exemples de biais connus du RNA-seq

Préparation des librairies

- **Influence de la préparation des librairies**
  - ✓ **Synthèse du cDNA avec des randoms primers:**
    - o **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. Brenner[2] 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.

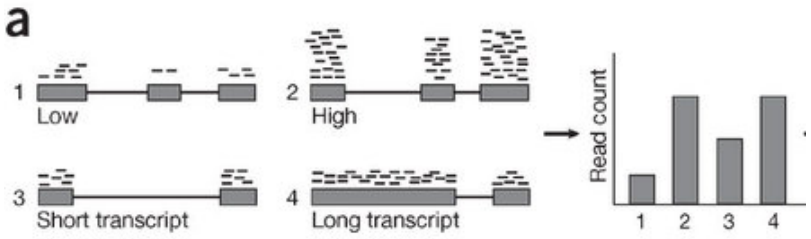
  - ✓ **Amplification par PCR**
    - o **Idée: supprimer totalement la PCR : actuellement, on réduit le nombre de cycles (10 cycles au lieu de 15)**

Longueur du transcrit

- **Influence de la longueur du transcrit:**
  - **Nombre total de reads d'un transcrit : Exemple de calcul utilisé : RPKM (Read Per Kb per Million reads mapped)**

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

  - **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)**



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

Liyan Gao[1,†], Zhide Fang[2,†], Kui Zhang[1], Degui Zhi[1] and Xiangqin Cui[1,*]

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

Associate Editor: Ivo Hofacker

Biol Direct. 2009 Apr 16;4:14.

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

Oshlack A, Wakefield MJ.

Abstract

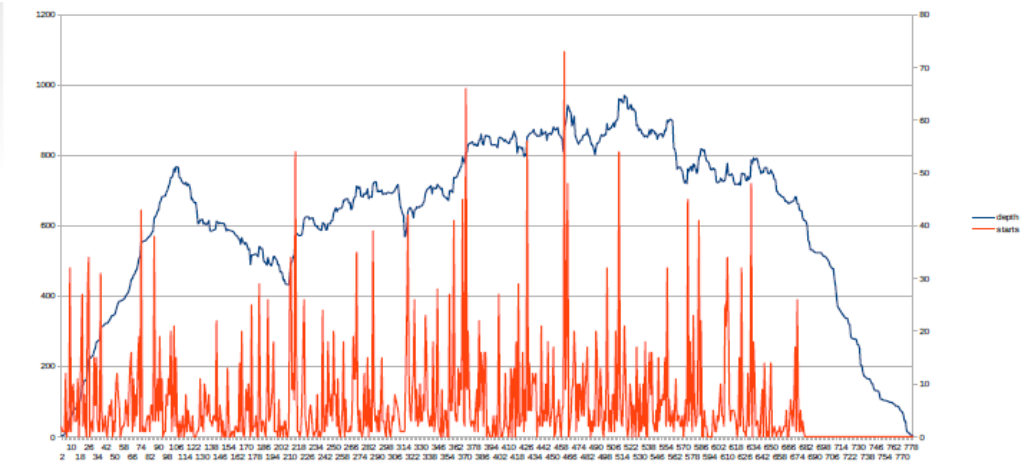
Background: Several recent studies have demonstrated the effectiveness of deep sequencing for transcriptome analysis (RNA-seq) in mammals. As RNA-seq becomes more affordable, whole genome transcriptional profiling is likely to become the platform of choice for species with good genomic sequences. As 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).

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**

Questions

- ✓ **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 € €)**

- ✓ **Quelle est la longueur idéale pour les reads?**
  - **Plus les reads sont longs, plus l'alignement se fait facilement; 2X150 bp actuellement**

- ✓ **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**

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]

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

Questions

- ✓ **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

- ✓ **Quelle technique de séquençage utiliser?**
  - **Actuellement, technique la mieux adaptée (et la plus utilisée) : Illumina HiSeq (profondeur et longueur des séquences)**
  - **Pas besoin de réplicats techniques, mais absolument besoin de réplicats biologiques**

- ✓ **Séquençage orienté ou non-orienté?**
  - **Orienté : possibilité de détecter des starts de transcription, des nouveaux gènes, le brin codant du gène…**

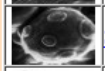
    MAIS : pas une couverture totale du transcriptome
  - **Non-orienté : couverture plus globale, grande efficacité**

    MAIS : séquençage seulement des ARN polyA+ (Kits Illumina)

- ✓ **Coûts : exemple de 10 échantillons séquencés sur 1 ligne en 2x150 pb (30 millions de couples de séquences par échantillon) : ~4000 € soit ~400 €/échantillon**

Questions

✓**Quelle profondeur en fonction de la taille du génome?**

Organisme	Position systématique	Nombre de chromosomes par jeu haploïde	Taille du génome (en Mb)	Nombre de gènes
 Escherichia coli	Bactérie	circulaire unique	4,7	4 288
 Saccharomyces cerevisiae	Ascomycète	16	14	6 200
 Caenorhabditis elegans	Nématode	6	100	19 100
 Arabidopsis thaliana	Angiosperme	5	130	25 000
 Drosophila melanogaster	Insecte	4	170	15 000
 Mus musculus	Mammifère	20	3 000	30 000
 Homo sapiens sapiens	Mammifère	23	3 000	30 000

Pas de loi en fonction de la taille du génome, mais plutôt en fonction du nombre de gènes, et encore…

http://www.inrp.fr/

GeT
Génome et
Transcriptome

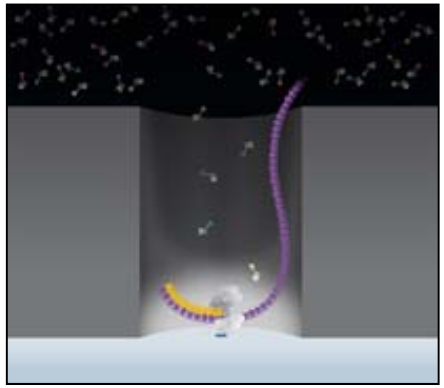
Iso-Seq : une autre utilisation du RNAseq

Autre utilisation du RNAseq : identifier des transcrits alternatifs

+ Séquençage molécule unique
+ Pas d'étape de PCR
+ Lectures longues (

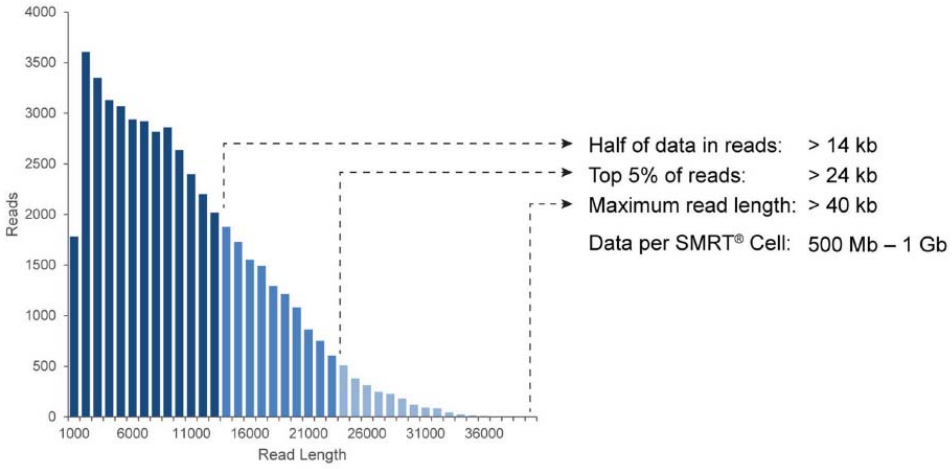
- Beaucoup d'erreur pour l'instant (>10%)

Déjà disponible



PacBio RSII

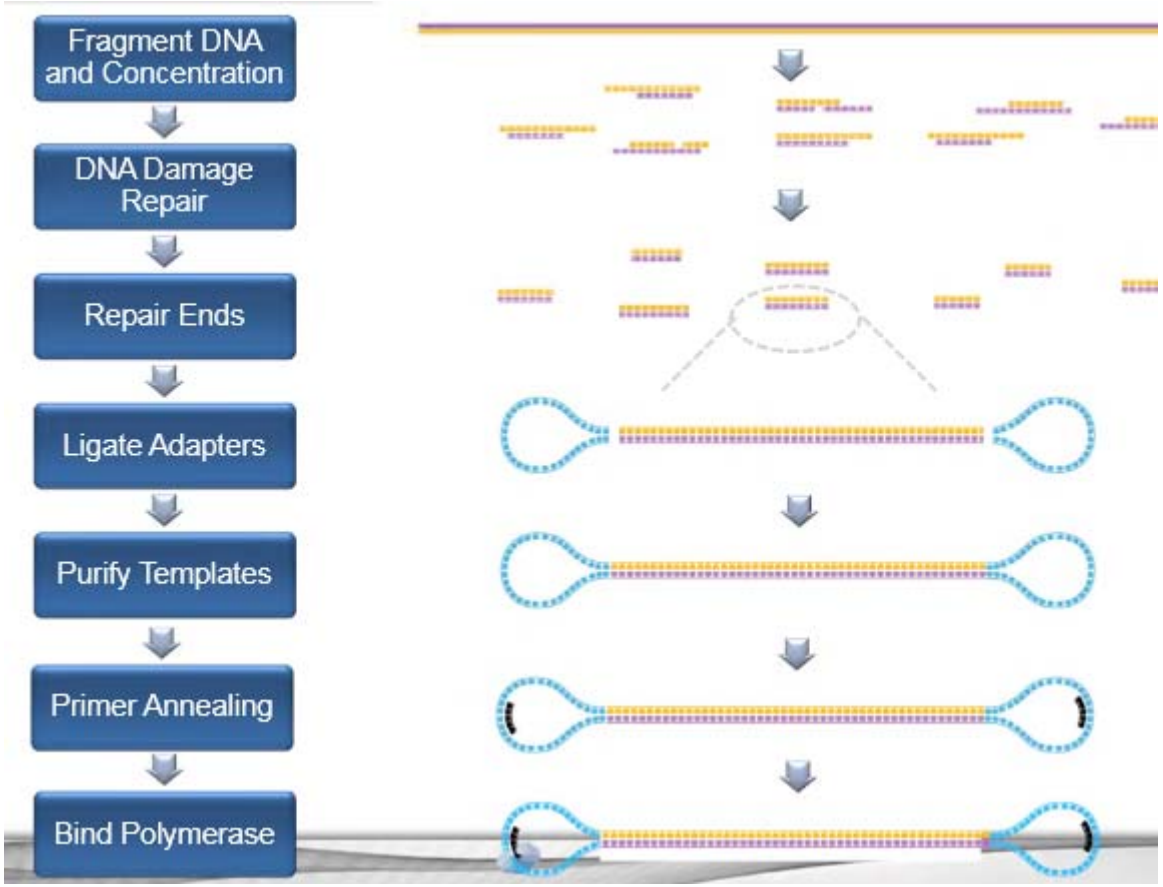
New : SEQUEL



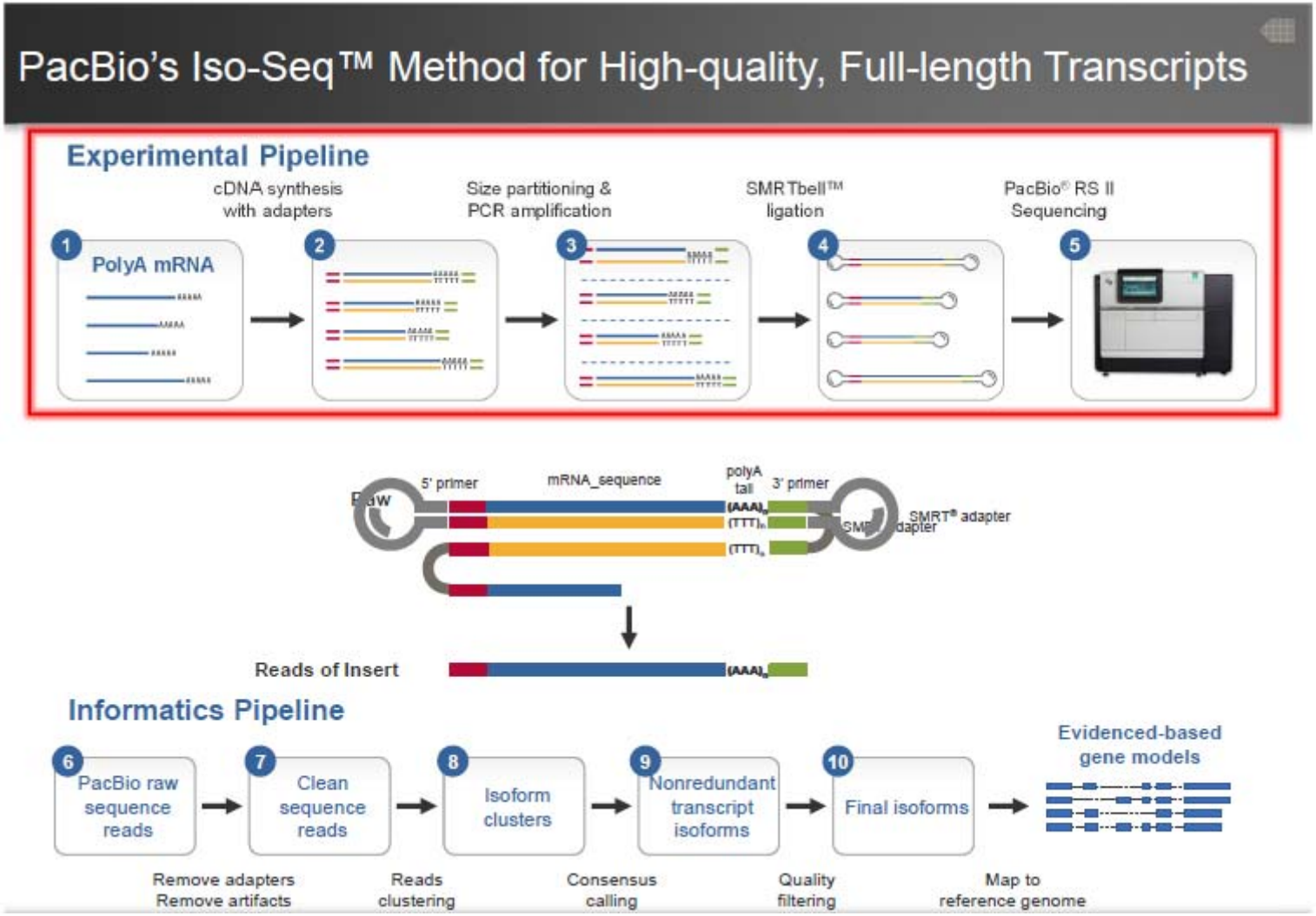
P6-C4, 4-hr movie, 20-kb BluePippin™ size-selected *E. coli* library (1 SMRT Cell)

http://www.pacificbiosciences.com

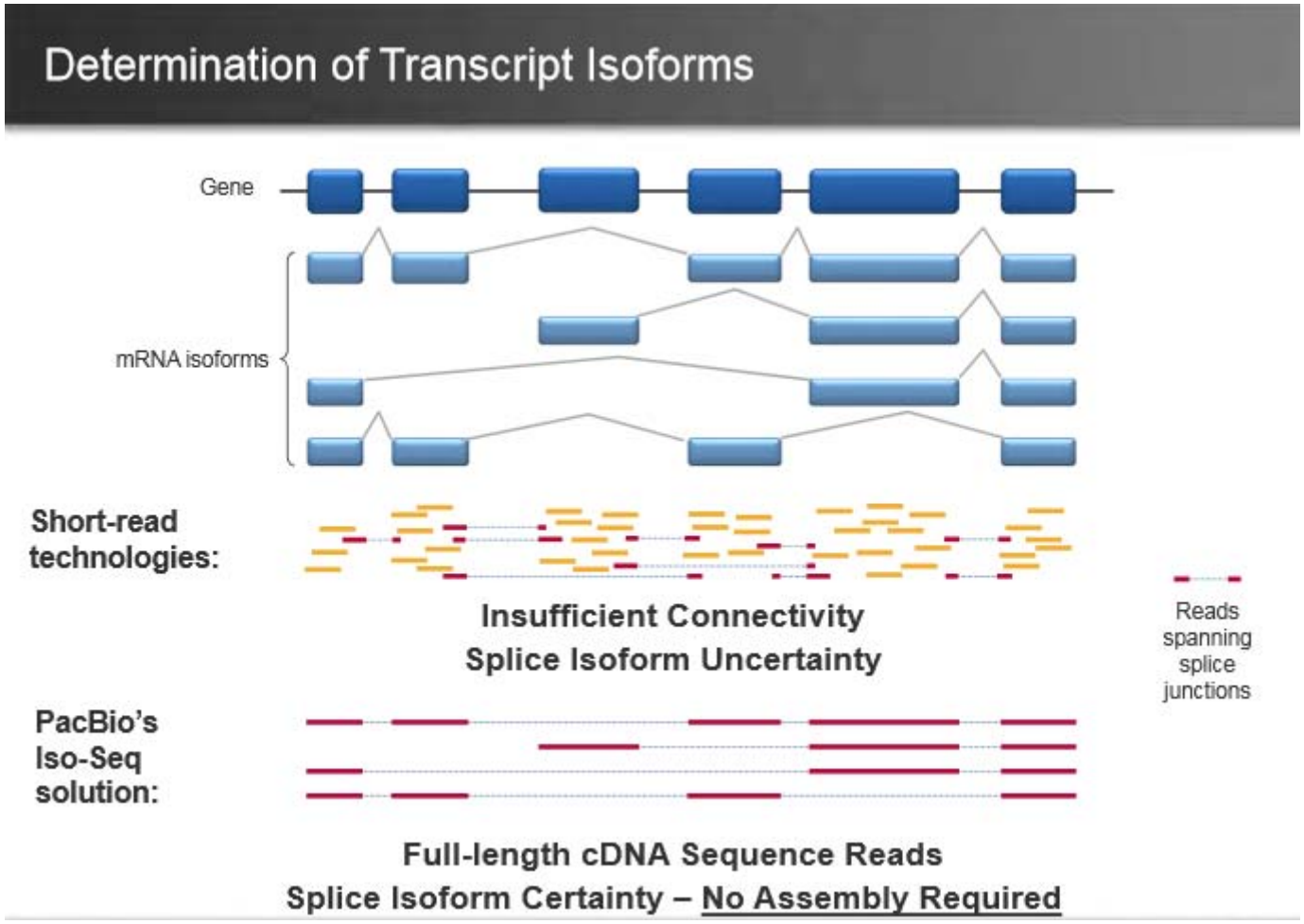
Pacific BioSciences Iso-Seq



Pacific BioSciences Iso-Seq



Pacific BioSciences Iso-Seq

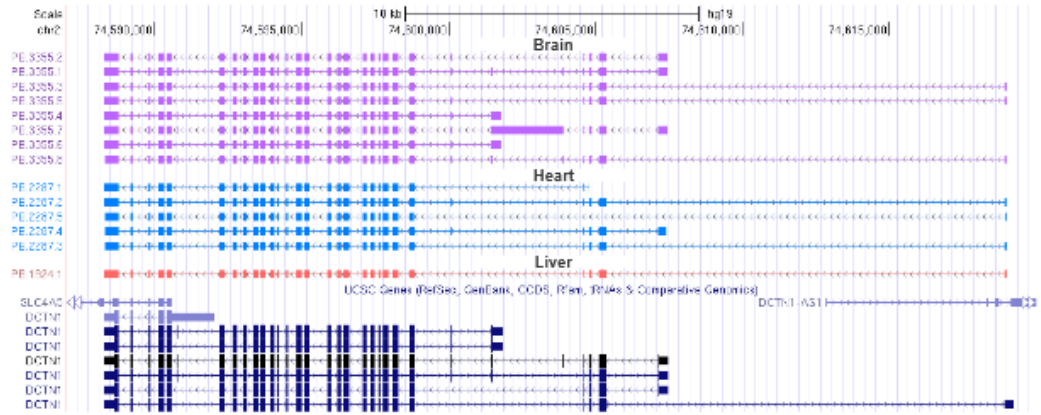


Pacific BioSciences Iso-Seq

PacBio Sequencing of Iso-Seq Libraries From 3 Human Tissues

Tissue	Size Fractions Sequenced				Number of Isoforms	Number of Genes	Transcript Lengths
	1-2 kb	2-3 kb	3-6 kb	5-10 kb			
Brain	■	■	■	■	10289	6356	418 – 8823 nt
Heart	■	■	■	■	6896	4351	467 – 8528 nt
Liver	■	■	■		6124	3497	419 – 4754 nt

Full-Length Non-Redundant Transcript Sequences

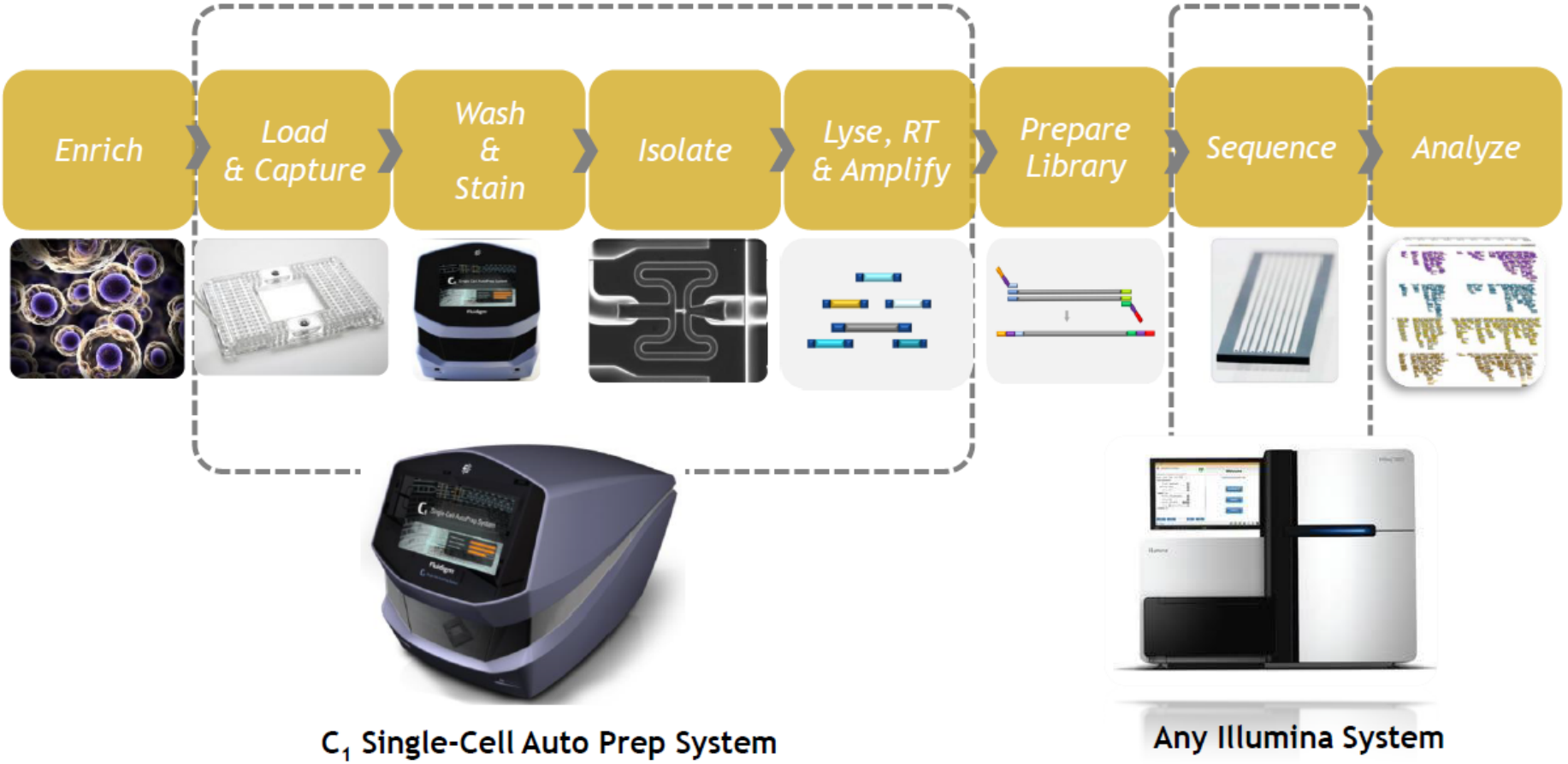


genotoul

GeT
Génome et
Transcriptome

Single Cell RNAseq

Fluidigm C1 Single Cell RNA seq



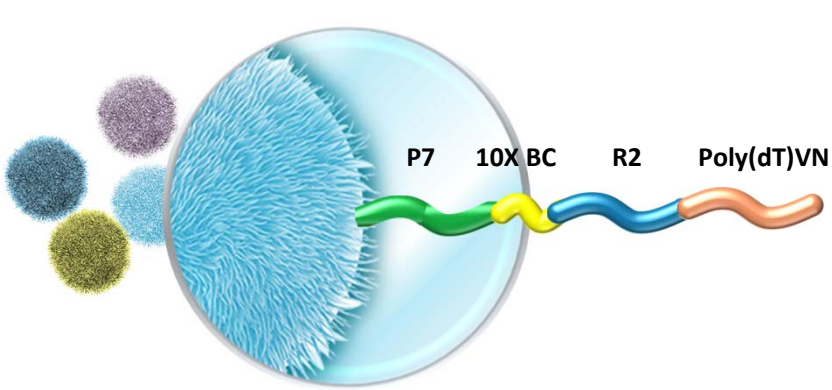
C_1 Single-Cell Auto Prep System

Any Illumina System

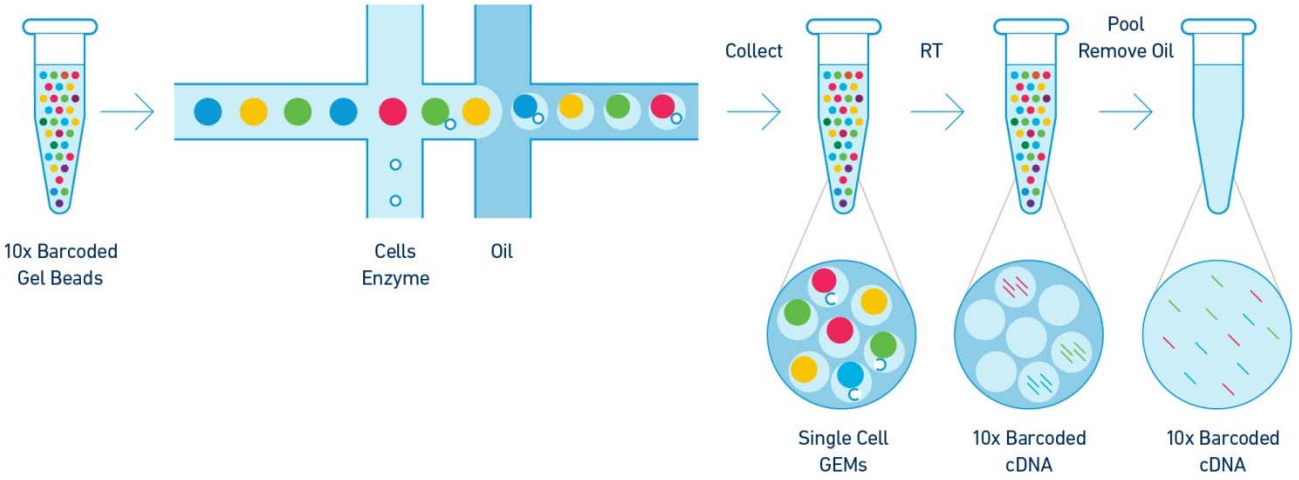
Fluidigm C1 Single Cell RNA seq

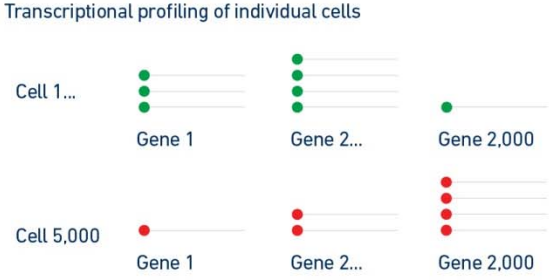
10XGenomics Chromium Single Cell RNAseq

Gel Bead Scaffold → Functional Oligo with Barcode → High-diversity Library



10XGenomics Chromium Single Cell RNAseq





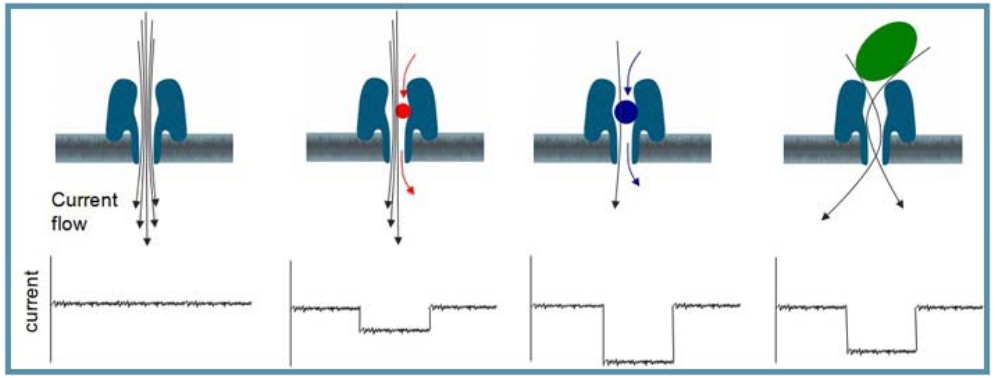
genotoul

GeT Génome et Transcriptome

Le futur?

Oxford Nanopore

MinION



PromethION

+ Séquençage molécule unique
+ Pas d'étape de PCR
+ Taille des séquences
+ Séquençage des ARN direct?

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

http://www.nanoporetech.com

genotoul

GeT
Génome et
Transcriptome

Questions?

GeT Génome et Transcriptome

Compar

PLoS One. 2013 Aug 20;8(8):e71462. doi: 10.1371/journal.pone.0071462.

Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data.

Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y.

Center for Quantitative Sciences, Vanderbilt University, Nashville, Tennessee, United States of America.

Abstract

RNAseq and microarray methods are frequently used to measure gene expression level. While similar in purpose, there are fundamental differences between the two technologies. Here, we present the largest comparative study between microarray and RNAseq methods to date using The Cancer Genome Atlas (TCGA) data. We found high correlations between expression data obtained from the Affymetrix one-channel microarray and RNAseq (Spearman correlations coefficients of ~0.8). We also observed that the low abundance genes had poorer correlations between microarray and RNAseq data than high abundance genes. As expected, due to measurement and normalization differences, Agilent two-channel microarray and RNAseq data were poorly correlated (Spearman correlations coefficients of only ~0.2). By examining the differentially expressed genes between tumor and normal samples we observed reasonable concordance in directionality between Agilent two-channel microarray and RNAseq data, although a small group of genes were found to have expression changes reported in opposite directions using these two technologies. Overall, RNAseq produces comparable results to microarray technologies in term of expression profiling. The RNAseq normalization methods RPKM and RSEM produce similar results on the gene level and reasonably concordant results on the exon level. Longer exons tended to have better concordance between the two normalization methods than shorter exons.

	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**