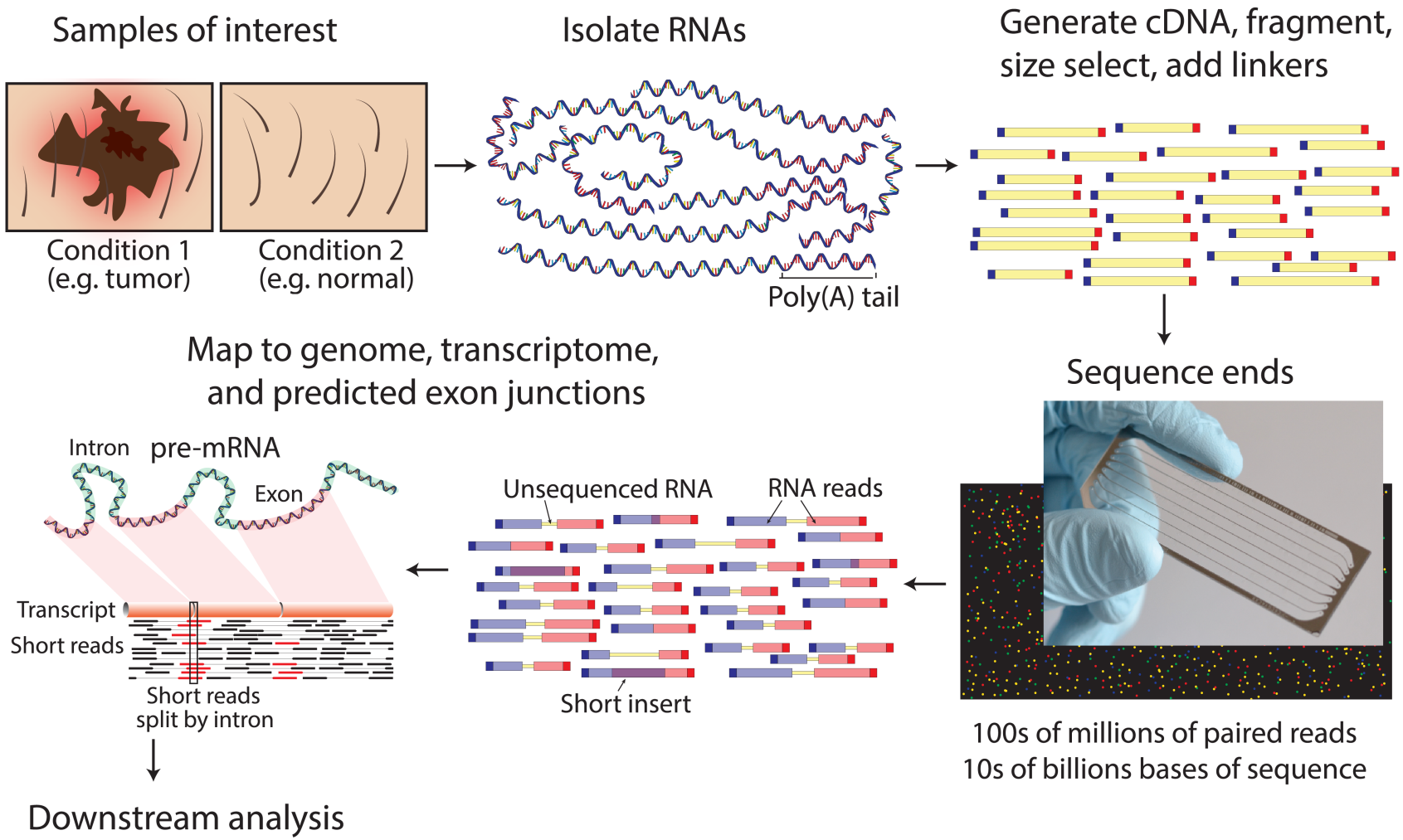
3-4 outils sur le traitement de données RNA-Seq

Matthias Zytnicki, Christine Gaspin, Ignacio González

MIAT INRA

Plan

Le (s)RNA-Seq

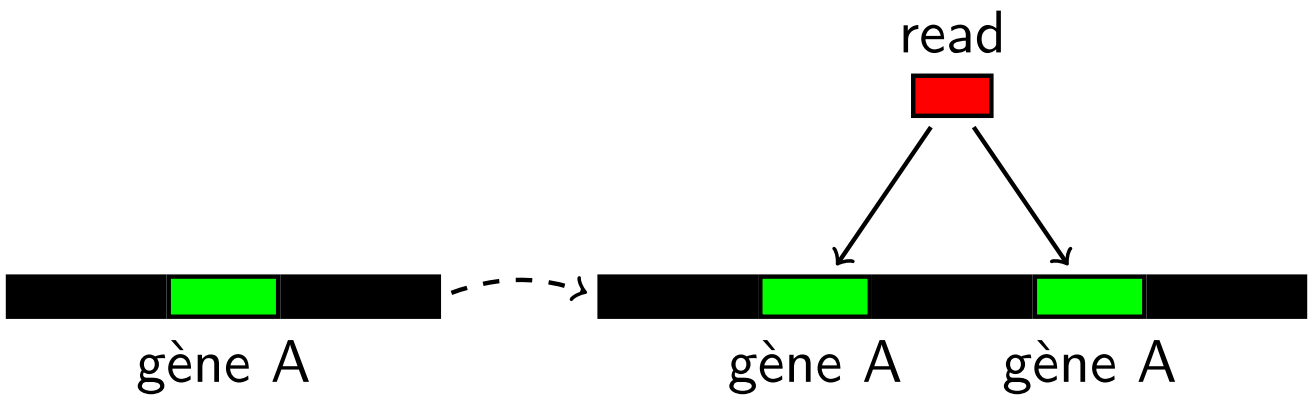


Griffith et al., PLoS Computation Biology, 2015

Le (s)RNA-Seq

- *Séquençage* des régions transcrites.
- *Alignement* des lectures sur le génome/transcriptome.
- *Quantification* des "gènes" en comptant le nombre de lectures.

$\Rightarrow$ Que se passe-t-il si un gène est dupliqué ?



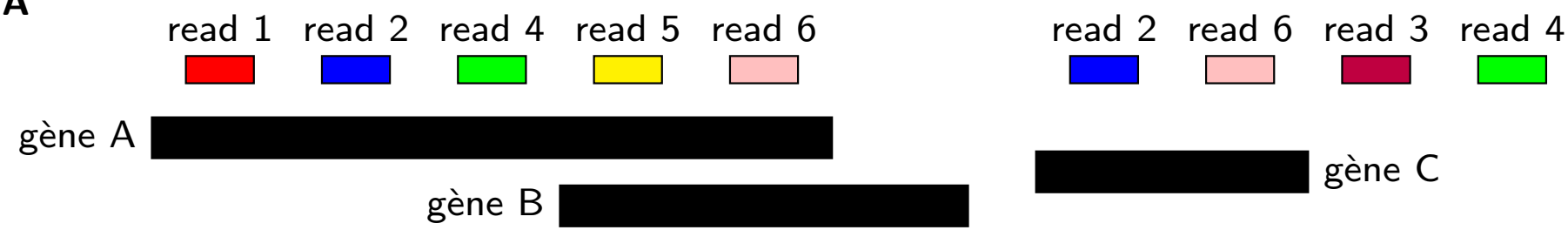
Quantification dans le cas de lectures multi-mappées

Différentes méthodes

- ne pas compter ces lectures,
- choisir un *hit* au hasard,
- pondérer chaque *hit*,
- utiliser une méthode d'estimation d'estimation de l'expression.

Pipe-line

A



B

read 1	A
read 2	A–C
read 3	
read 4	A
read 5	A–B
read 6	A–B–C

C

A	2
A–B	1
A–C	1
A–B–C	1

Les données

RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder

N Akula, J Barb, X Jiang, J R Wendland, K H Choi, S K Sen, L Hou, D T W Chen, G Laje, K Johnson, B K Lipska, J E Kleinman, H Corrada-Bravo, S Detera-Wadleigh, P J Munson & F J McMahon

Molecular Psychiatry (2014) **19**, 1179–1185 (2014)
doi:10.1038/mp.2013.170
Download Citation

Received: 28 March 2013
Revised: 24 October 2013
Accepted: 29 October 2013
Published online: 07 January 2014

6 réplicats par condition, séquencés en HiSeq. 2 x 700M–1M̄ lectures.

Résultats

Gènes non fusionnés

	htseq-count	featureCounts	mmquant
gènes diff. exp.	734	835	763
temps	4–5h	8–11m	21–29m

Gènes fusionnés

- 5–6% des lectures
- 256 gènes fusionnés différentiellement exprimés (516 gènes en tout)
- 33 gènes fusionnés avec une p-valeur $< 1\%$, incluant *ADK*, *GTF2I*, *hnRNP-A1*, *HTRA2*, *PKD1* et *RERE*.

Conclusion

- Un remplaçant de featureCounts (supporte les lectures pairées, et diverses options de chevauchement).
- En plus : le multi-mapping.
- En moins : le multi-threading (plusieurs threads par fichier).
- 3 façons de l'utiliser :
  - Téléchargement : https://bitbucket.org/mzytnicki/multi-mapping-counter
  - BioConductor (devel) : package Rmmquant
  - Galaxy Tool Shed

Introduction

Le sRNA-Seq

Il séquence tous les petits ARNs :

- miRNAs, siRNA, piRNAs, tRNAs, etc.
- produits de dégradation ;
- transcrits cryptiques.

Objectif

- Possibilité 1 : Dans le contexte différentiel, trouver les gènes de petits ARN différentiellement exprimés. Voir mmquant !
- Possibilité 2 : Établir le répertoire des petits ARNs, avec leur expression respective.
  - Méthode 1 : *Blaster* les lectures contre des banques de petits ARNs.
  - Méthode 2 : Aligner les lectures sur le génome, et comparer avec une annotation.

Introduction

Le sRNA-Seq

Il séquence tous les petits ARNs :

- miRNAs, siRNA, piRNAs, tRNAs, etc.
- produits de dégradation ;
- transcrits cryptiques.

Problème

Beaucoup de petits ARN sont répétés et des annotations sont ambiguës.

Objectif

- Possibilité 1 : Dans le contexte différentiel, trouver les gènes de petits ARN différentiellement exprimés. Voir mmquant !
- Possibilité 2 : Établir le répertoire des petits ARNs, avec leur expression respective.
  - Méthode 1 : *Blaster* les lectures contre des banques de petits ARNs.
  - Méthode 2 : Aligner les lectures sur le génome, et comparer avec une annotation.

Pipe-line

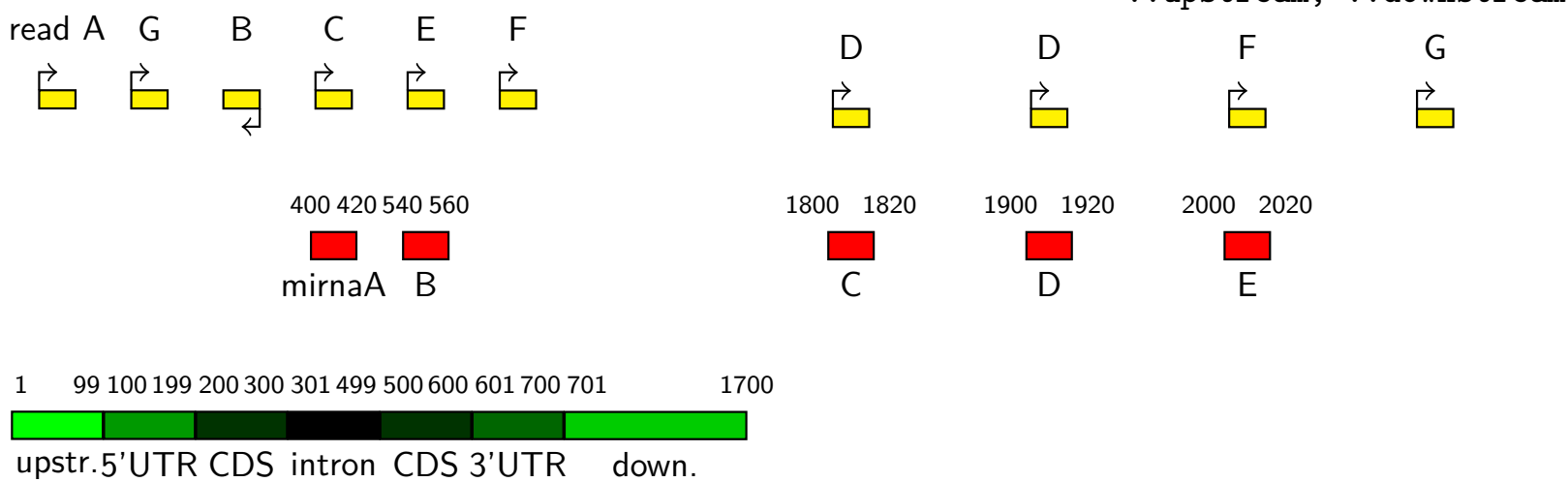
A

```
1   .   exon     100   300   .   +   .   gene_id "geneA"
1   .   exon     500   700   .   +   .   gene_id "geneA"
1   .   CDS      200   600   .   +   .   gene_id "geneA"
1   .   miRNA    400   420   .   +   .   gene_id "mirnaA"
1   .   miRNA    540   560   .   +   .   gene_id "mirnaB"
1   .   miRNA   1800  1820   .   +   .   gene_id "mirnaC"
1   .   miRNA   1900  1920   .   +   .   gene_id "mirnaD"
1   .   miRNA   2000  2020   .   +   .   gene_id "mirnaE"
```

B

```
Introns:
.:gene
Vicinity:
.:gene
Order:
.:CDS +, .:5'UTR +, \
     .:3'UTR +, .:miRNA
.:intron
.:gene -
.:upstream, .:downstream
```

C



D

A	upstream
B	gene (−)
C	miRNA
D	miRNA
E	CDS (+)—miRNA
F	3'UTR (+)—miRNA
G	5'UTR (+)

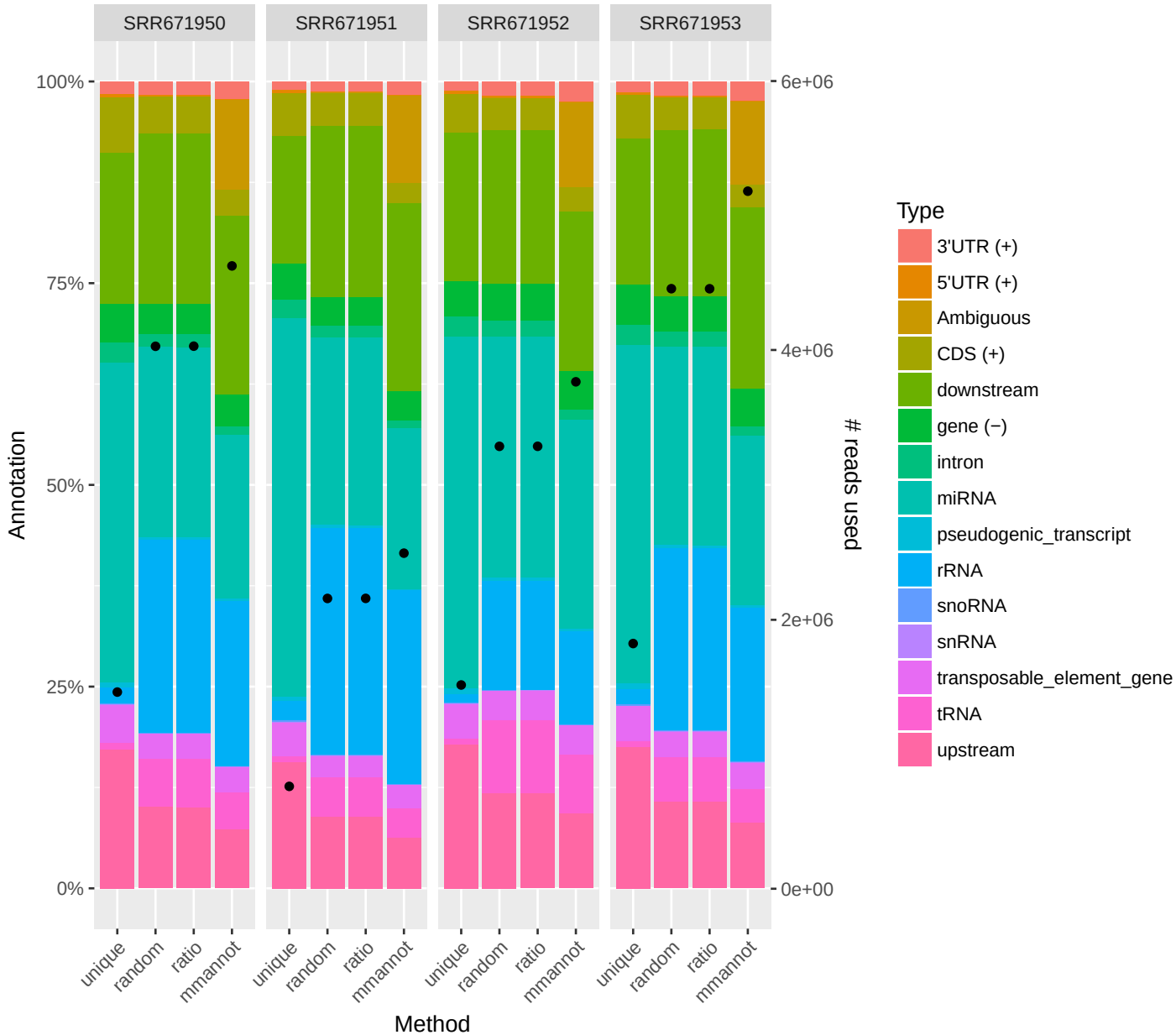
E

CDS (+)—miRNA	1
3'UTR (+)—miRNA	1
5'UTR (+)	1
miRNA	2
gene (−)	1
upstream	1

F

geneA (−)	1
geneA :3'UTR (+)—mirnaE	1
geneA :5'UTR (+)	1
geneA :CDS (+)—mirnaB	1
geneA :upstream	1
mirnaA	1
mirnaC—mirnaD	1

Résultats *A. thaliana*



Étude des classes

- La catégorie la plus représentée parmi les régions transcrites est miRNA — gène $(-)$.
- On retrouve des associations classiques :
  - *miR156*/*miR157* et *SPL*,
  - *miR163* et *PXMT1*,
  - *miR171* et *ATHAM*,
  - *miR400* et *PPR1*,
  - *miR403* et *Ago2*,
  - *miR824* et *AGL*.

Conclusion

- À publier. . .
- À télécharger : https://github.com/mzytnicki/mmannot

Plan

L'expression différentielle pour les nuls

Ce qu'on a :

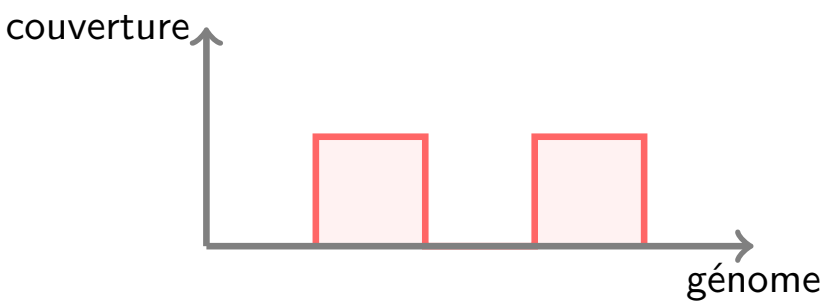
- une annotation
- 2 conditions
- plusieurs réplicats
- des millions de lectures par condition

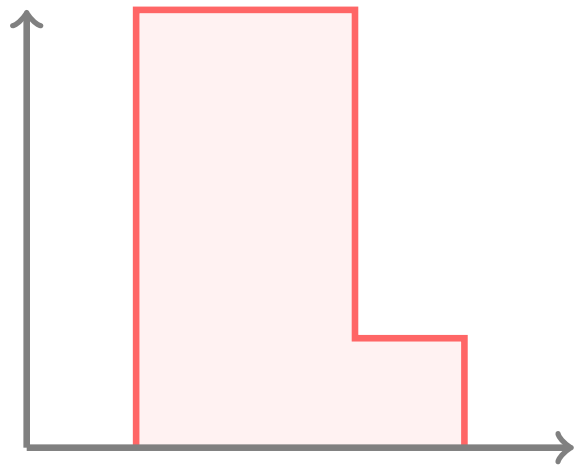
Ce qu'on aimerait avoir :

- En sRNA-Seq, l'annotation est souvent méconnue.
- Le but est de trouver les régions différentielles... sans annotation.

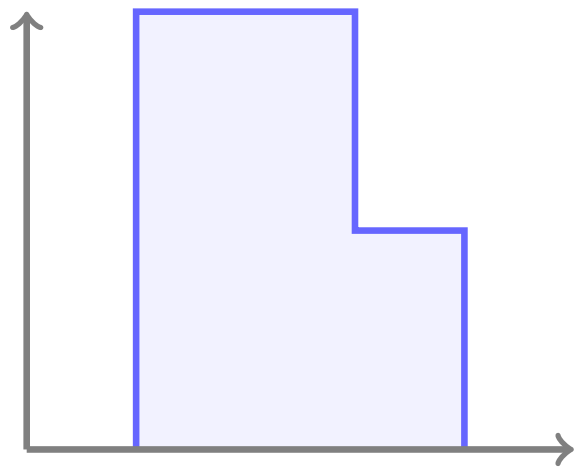
Particularités du sRNA-Seq

Les profiles peuvent être flous. . .

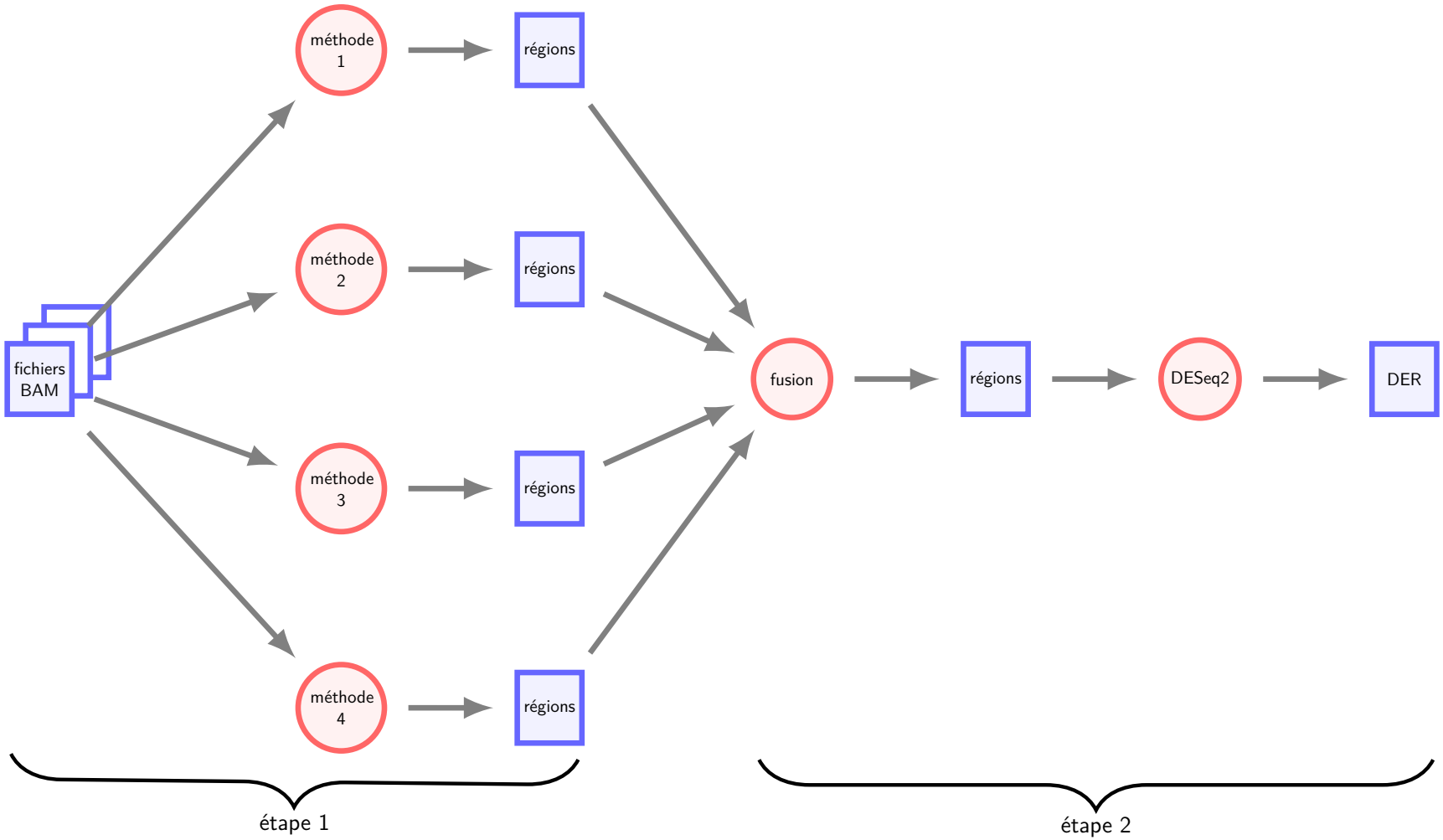




VS

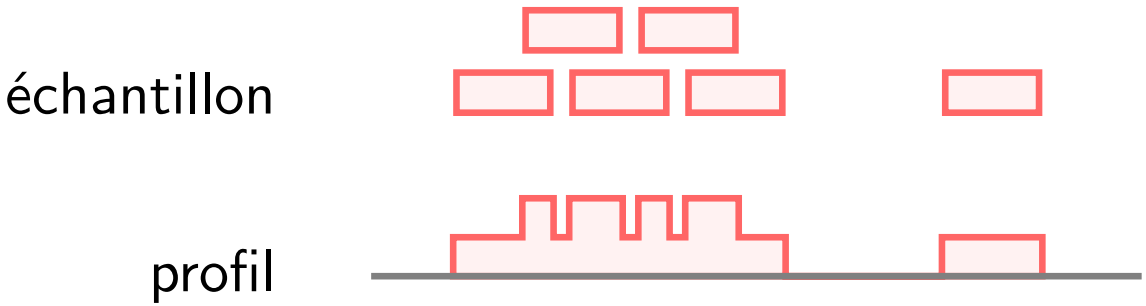


Pipe-line


méthode 1
régions
méthode 2
régions
fichiers BAM
méthode 3
régions
fusion
régions
DESeq2
DER
méthode 4
régions
étape 1
étape 2

Pré-processing

- Calculer la couverture



- Supprimer les régions peu exprimées.
- Normaliser les données (utilisant la méthode edgeR).

Méthode 1 : HMM

But

Agréger les différences

Étape 1 : Calculer une p-value pour chaque nucléotide

chrom.	pos.	échantillon 1	échantillon 2	...	p-value
chr 1	1	0	0		1
	2	0	0		1
	...	...	...		...
	100	300	310		0.1
	101	302	312		0.1
	103	303	313		0.1
	...	...	...		...
chr 2	1	0	0		1
	...	...	...		...

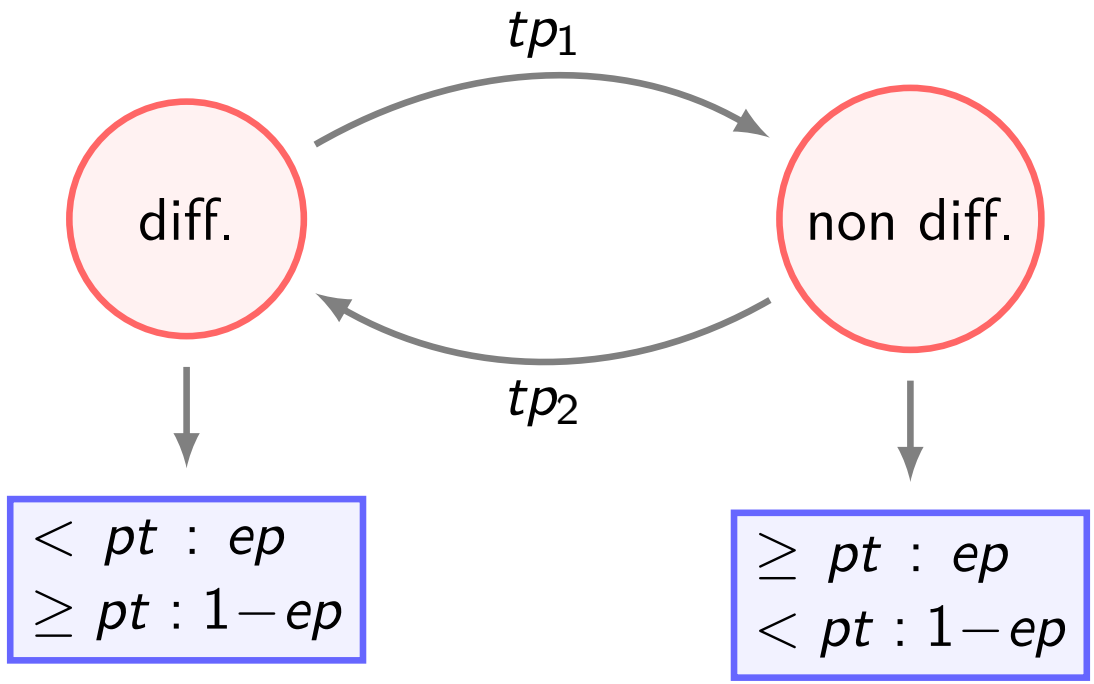
Méthode 1 : HMM

But

Agréger les différences

Étape 1 : Calculer une p-value pour chaque nucléotide

Étape 2 : Utiliser le HMM sur les p-valeurs



Méthode 1 : HMM

But

Agréger les différences

Étape 1 : Calculer une p-value pour chaque nucléotide

Étape 2 : Utiliser le HMM sur les p-valeurs

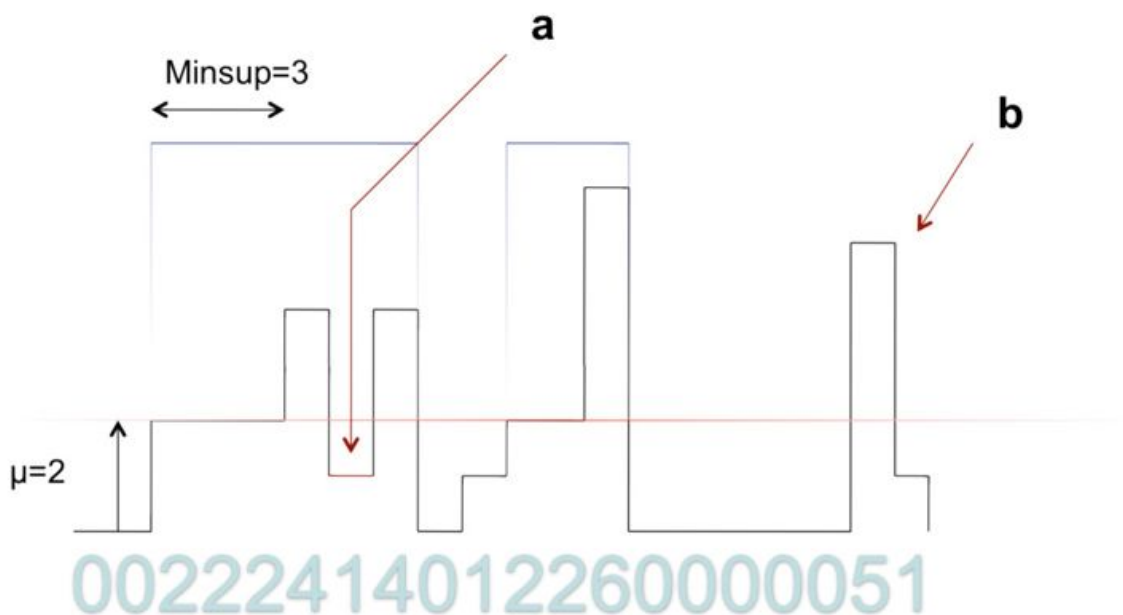
Étape 3 : Garder les régions de l'état « diff. »

Méthode 2 : Régions irréductibles

But

Zoomer sur les régions différentielles.

- Calculer la couverture moyenne dans chaque condition.
- Calculer le log-fold-change.
- Calculer les régions irréductibles.



A. Leśniewska & M.Okoniewski, BMC Bioinformatics, 2011

Protocole de comparaison

Autres approches

- derfinder
- ShortStack
- Utiliser d'autres annotations (gènes, miRNAs, snoRNAs, tRNA, etc.), et suivre la voie classique.

Données

- *A. thaliana*, *D. melanogaster*, et l'humain (avec des miRNAs validés).
- Données simulées sur l'humain.

Où on en est

On change la recette, tout le benchmark est à refaire. . . .

Sur l'humain

Annotation	total	srnadiff	derfinder	ShortStack
gènes	771	331	54	186
miRNAs	240	204	33	170
piRNAs	14	12	4	11
snoRNAs	42	31	1	29
tRNAs	92	72	19	61

Annotation	srnadiff	derfinder	ShortStack
srnadiff	1391	146	645
derfinder	146	199	103
ShortStack	541	93	617

Plan

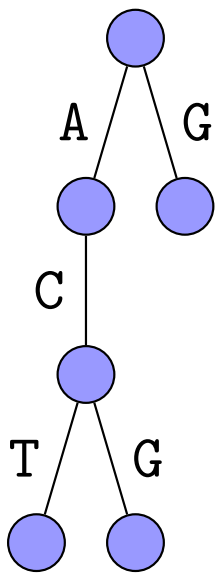
Idée principale

- lectures courtes ($\approx$ 20-30 pb)
- très répétées

$\Rightarrow$ on peut donc représenter efficacement les lectures de façon compacte

Exemple

`ACT`, `ACG`, `G` devient



Note

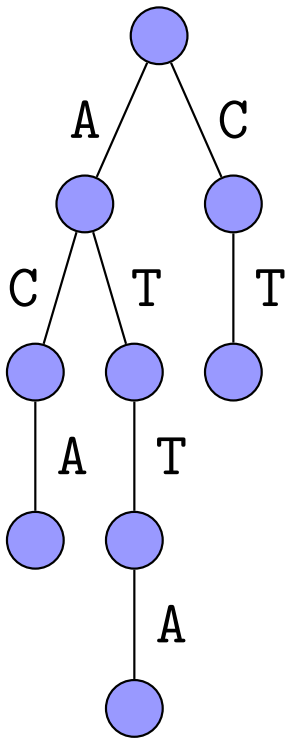
Le génome est également représenté sous forme d'arbre des suffixes (en l'occurrence : un tableau de suffixe).

Idée

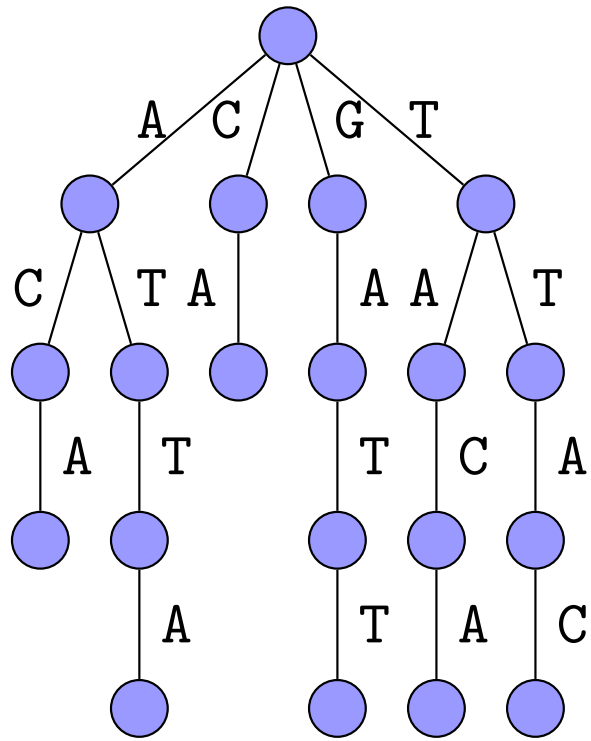
Mapper des lectures revient à comparer deux arbres.

Mapper les lectures sans erreur

Lectures



Arbre des suffixes

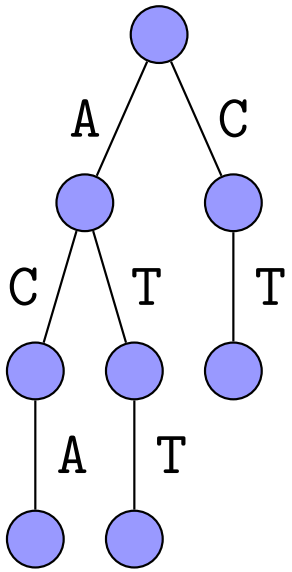


But

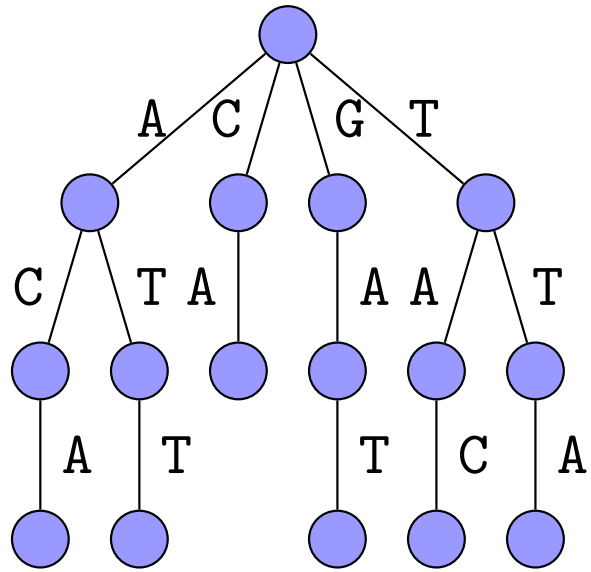
Trouver l'intersection des arbres.

Mapper les lectures avec erreurs

Lectures



Arbre des suffixes



Algo

- Pour chaque nœud de l'arbre des lectures, on calcule les nœuds de l'arbre du génome correspondants (modulo n erreurs).
- On procède récursivement.

Résultats (15M lectures)

Temps en minutes

outil	toutes lectures	réduites
`bwa aln`	16	6
`bwa aln -N`	$\approx$ 160	$\approx$ 60
`bwa mem`	5	1,5
`bwa mem -a`	6	1,8
`srnaMapper`	7	—

Mémoire en Mo

outil	toutes lectures	réduites
`bwa aln`	220	218
`bwa mem`	490	482
`srnaMapper`	4 135	—

Conclusion définitive

Beaucoup de choses à finir...