

# Assemblage et mapping de données NGS, on en est où ? peut-on s'en passer ?

Pierre Peterlongo

Journée de la plateforme Genouest 26 Octobre 2010

# Séquençage – « intro classique »

## INTRODUCTION

A new generation of sequencing technologies is revolutionizing molecular biology [1–15]. They have dramatically lowered the costs per sequenced nucleotide and increased throughput by orders of magnitude. Illumina's Solexa and Applied Biosystems' SOLiD can generate gigabases of nucleotide sequence per week. However, a perceived limitation of these ultra-high-throughput technologies is their short read-lengths. The first incarnation of Illumina's Genome Analyzer (GA) typically generated sequence

# Quelles technologies ?

- Plus connues:
  - **454: Genome Sequencer**
  - **Illumina Solexa Genome Analyser**
  - **Applied Biosystem: SOLiD**



Peterlongo - Assembler... ou pas.



# Quelles technologies ?

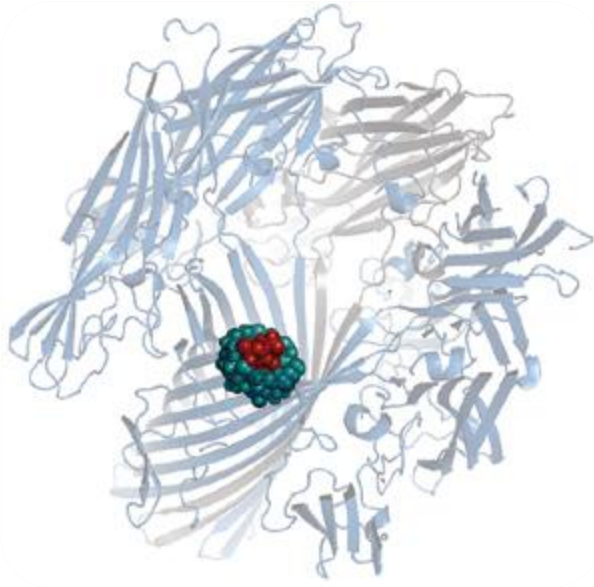
- Moins connues/utilisées:
  - Helicos: Heliscope
  - Polonator



# NNGS / Third Generation

(1000 \$ = human genome, 0.0000005 \$/base)

- Oxford Nanopore - reads unlimited
- Pacific Bioscience - reads = 100.000



# Illumina vs. 454



## **Illumina: Genome Analyzer Iix**

- Reads: 108bp
- 240M reads / run
- 25Gbp / run
- 10 jours / run
  
- Taux erreur très bas
  - 70% de reads parfaits
- pas d'indel => bonne complémentarité avec 454
- Prix

## **454: GSFLX Titanium**

- Reads: 500bp
- 1M reads / run
- 0.5Gbp / run
- Run : 8h
  
- Taux d'erreur élevé dans les homopolymers
- Bons assemblages à partir de 20x

# Que faire des reads ?

## Premiers reflexes

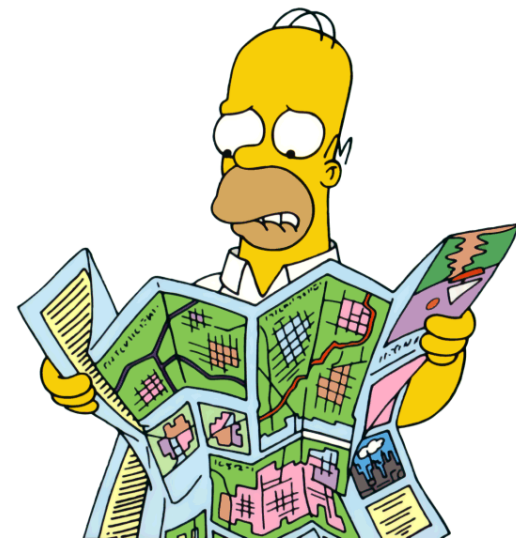
- Avec un génome de référence fiable :
  - Read mapping
- Sans génome de référence :
  - Assemblage *de novo*
- Si on dispose de génome(s) peu fiables :
  - Mapping + assemblage *de novo*

# Read Mapping

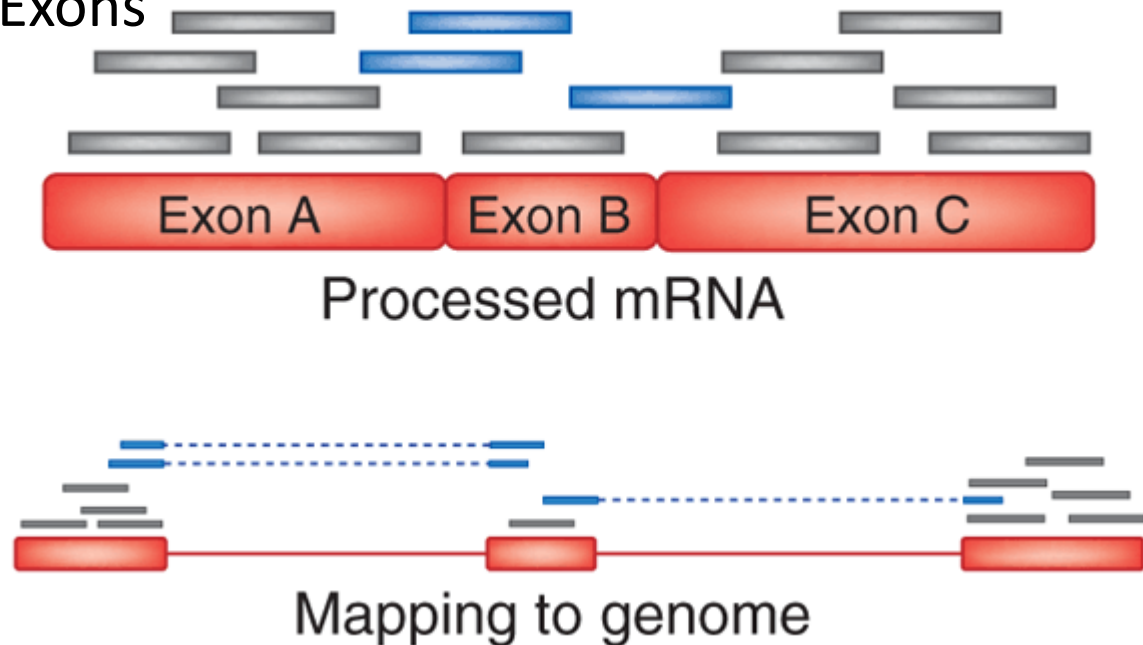
```
M.A.Q Viewer
ATAGGTTATAGCACAGGAAGAAGGAATAGGAGAAAAACAAGTATCTACATAGAACTTTTCAGTGTAAAAAATCCCAAAAACCGGTTGACAATTGCCAA
ATAGGtTATAGcaCagGcagaag AATAGGAGaAAAAACAAGTATCTACATAGAACTTT GTGTAaaaaATCCCAAAAACCGGTTGACAATTGtn A
ATAGGtTATAGCACAGGgaGaaGGcn AGGAGAAAAACAAGTATCTACATAGAACTTTTCAG GTAAAAAATCCCAAAAACCGGTTGACAATTGCcaA
ATAGGTTATAGCACAGGAAGAAGGAA GGAGAAAAACAAGtAtCTACATAGAActtTCaGt TAAAAAATCCCAAAAACCGGTTGACAATTGcCaC
ATAGGTTATAGCACAGGAAGAAGGAATAG AGAAAAaCAAAGTATCTACATAGAACTTTTCAGTGT AAAAATCCCAAAAACCGGTTGACAATTGCCAa
ATAGGTTATAGCACAGGAAGAAGGAATAGGAGA AAAACAAGTATCTACATAGAACTTTTCAGTGTAAAA AtCCCAAAAACCGGTGACAATTgcCAA
AtACGTTatAGCACAgGAaAaGgaATagGAcaa CAAAGTATCTACATAGAACTTTTCAGTGTAAAAaTC CAAAAACCGGTTGACAATTGCCAa
ATAGGTTATAGCACAGGAAGAAGGAtAGGAGaaa aAAGTATCTACATAGAaCTTtCAGTGTAAaAGTCC AAAaaaCCgGTTGACAATTGCCAa
AtAGGTTATAGCACAGGAAGAAGGAATAGGAGAAAA AAAGTATCTACATAGAActTTTCAGTGTAAAAAATCC AAAAAACCGGTTGACaaTTGCCaA
AT GGTtTATAGCACAGGAaGAAGGAtAGGAgAaaaaaac AAGTATCTACATAGAActTTTCAGTGTAAaAAATccc AAaACCGGTTGACaATTGCCAa
AT tTATAGCACAGGAAGAAGGAATAGGAGAAAAAACAA gAgCTaCaTAgAGGCTTTTCAGTGTAAAAaATCcCAAA aacCggTTGACAATTGCCAa
ATA tTATAGCACAGGAAGAAGGAATAGGAGAAAAAACAA TAtCTACATAGAActTTTCAGTGTAAAAAATCCCAAA AaCCgGTTGACAATTGCCAa
ATAGG TATAGCACaGGAAGAAGGAAaTAGGagAAaAAaCaAc ATCTACATAGAActTTTCAGTGTAAAAaATCCCAaaA aCCGgtTGACAAttGCCAa
ATAGG aATAGCACAGGAAGAAGGAATAGGAGAAAAaACAag TCTACATAGAActTTTCAGTGTAAAAAATcCcaAaAa CCGGTTGACAaATTGCCaa
ATAGGTT tAGCACAGGAAGAAGGAATAGGAGAAAAACAAGT CTACATAGAActTTTCAGTGTAAAAAATCCCAAAAaa CGGTTGAcAATTgCCaa
ATAGGTTA AgCACAGGAAGAAGGAATAGGAGAAAAACAAGTA CtAcAtAGAActTTTCAGTGTAAAAAATCCCAAAAAA CggTtGACAATTGCCAa
ATAGGTTA ACAGGAAGAAGGAATAGGAGAAAAACAAGTATCT CATAGAActTTTCAGTGTAAAAAATCCCAcAAaCCG ttGACAATTGCCAA
ATAGGTTA ACAGGAAGAAGGAATAGGAGAAAAACAAGTATCT CATAGAActTTTCAGTGTAAAAAATCCCAaAaAACCg tGtcaATtGcCAa
ATAGGcTat ACAGGAAGAAGGAATAGGAGAAtAAACAAGTATCT ATAGAActTTTCAGTGTAAAaAATCCCAAAAACCGg aGACAATTGCCAA
ATaGGTTaTt AGGAAGAAGGAATAGGAGAAAAACAAGTATCTAc tAGAActTTTCAGTGTAAAAAATCCCAAAAACCGGT ACAAtTGCCAA
ATAGGTTATAG GGAAGaaGGAATAGGAgaaAaAaCAaGtTCTcac AGAaCTTTTCAGTGTAAAAAATCCCAcAaaACcgGat gccaa
ATAGcTTATAGCACA GAaGAAGGAATAGGAGAAAAACAAGTATCTaCAT GAActTTTCAGTGTAAAAAATCCCAAAAACcGGtTG CCAa
ATAGGTTATAGCACAG AAGAAGGAATAGGAGAAAAACAAGTATCTACATA AActTTTCAGTGTAAAAAATCCCAAAaAaacCGGTTGt CCAa
ATAGGTTATAGCACAGG AGAAGGAATAGGAGAAAAACAAGTATCTACATAG tTTCAGTGTAAAAAATcCCAAAAACCGTTGAcAt CCAa
ATAGGTTATAGCACAGG AGGAATAGGAGAAAAACAAGTATCTacATAGaaG TTCAGTGTAAaAaATCCCAaAAAaCCGgtGAcGat
ATAGGTTATAGCACAGGAAG gGAATAGGAGAAAAACAAtTATctAcocTagCAnn TCAGTGTAAAAAAtCCCAAAAACCGGTTGACAATT
ATAGGTTATAGCACAGGcAGA ATAGGAGAAAAACAAGTATCTACATAGAActTTTC TGTAaaaaATCCCAAAAACCGGTTGACAaTTGcca
ATAGGTTATAGCACAGGAaGg TAGGAGAAAAACAAGTATCTACTTAGAACTTTTct GTAAAAAATCCCAAAAACCGGTTGACAATTGCcna
ATAGGTTATAGCACAGGAAGaAGGA TAaGAGAAAAaCAAAGTATCTACATAGAActTTTCA GTAAAAAATCCCAAAAACCGGTTGAcAATTGccaa
ATAGGTTATAGCACAGGAAGAAGGAATAGGA AAAAAACAAGTATCTAGAGGAGAActTTTCAGTgtAA atccCAAAAACCGGTTGACAATTGCCAA
aTagGtTATAGC aGAAGAAGGAATAGGAGAAAAACAAGTATCTAGAGGAGAActTTTCAGTGTAAAAAATCCCAAAaAaCcGgtt
ATAGGTTATAGCACAGGaaGaaGg TAGGAGAAAAACAAGTATCTACATAGAActTTTCA GTAAAAAATCCCAAAAACCGGTTGACAATTGCCAA
Read: 165181|+!s!xa_0011_8_0109_6382[11]IA 40
```



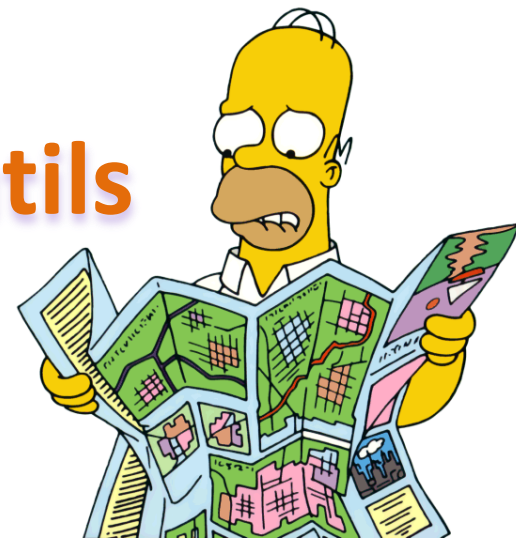
# Read Mapping



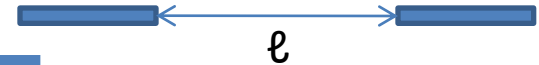
- Difficultés:
  - Reads mappant à plusieurs locus
  - RNASeq:
    - Introns Exons



# Read Mapping : Les outils



- Idée globale:
  - Indexation Reads/Génomme
  - Utilisation de graines ( $\approx$ Blast)



Outil	Algo	Long.	Gaps	Pairs	Qual.
Bfast	hash ref		X	X	
Bowtie	FM-index			X	X
BWA	FM-index	X	X	X	
MAQ	hash reads		X	X	X
Mosaik	hash ref	X	X	X	
Novoalign	hash ref		X	X	X
<b>GASSST</b>	hash ref	X	X		

+ rapide +sensible sur  
gros reads et tx erreurs > 5%

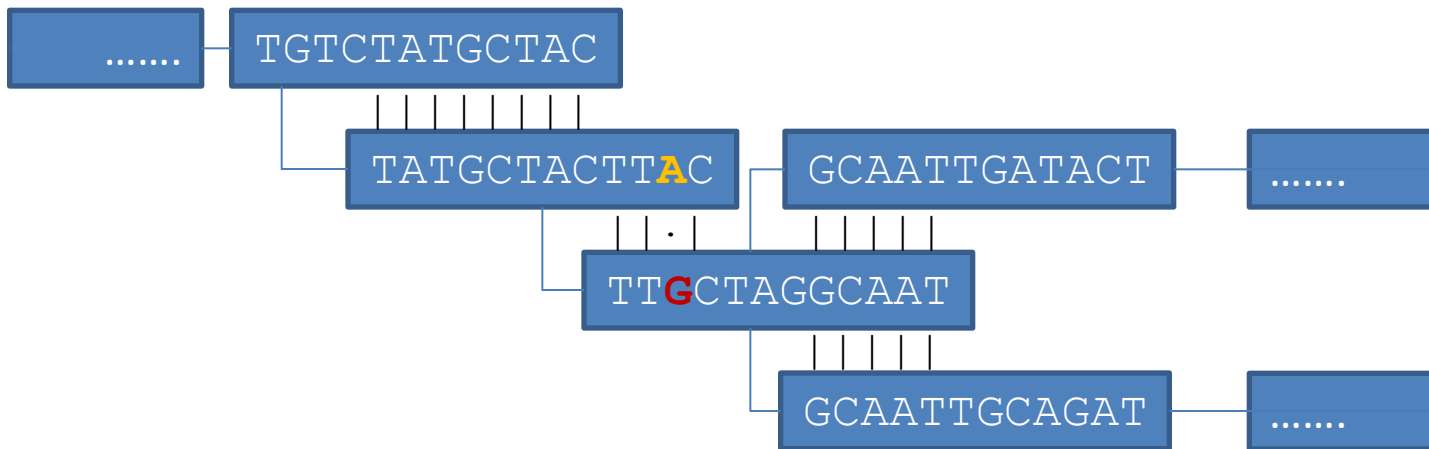
# Assemblage: Le problème

- Million/milliards de pièces
- Sans l'image
- Endroit, envers ???
- Erreurs de séquençage :
  - Les pièces ne s'assemblent pas parfaitement
- Couverture :
  - Certaines parties du puzzles absentes



# Stratégies d'assemblage

- Graphe des reads chevauchants (overlap graph)

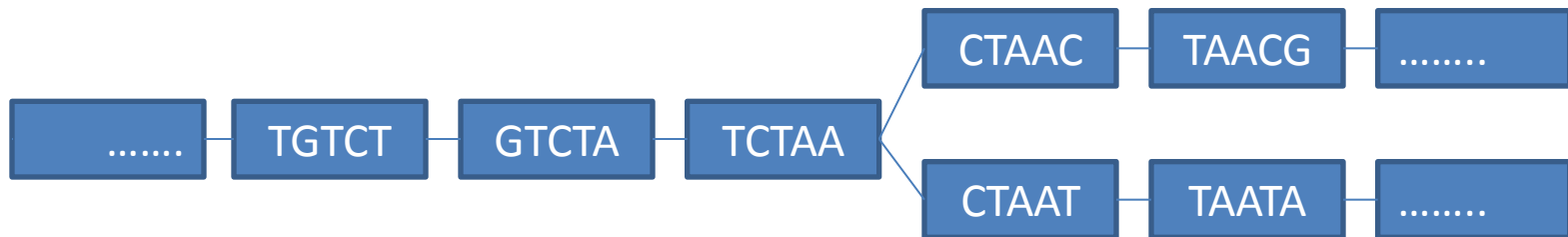


# Stratégies d'assemblage



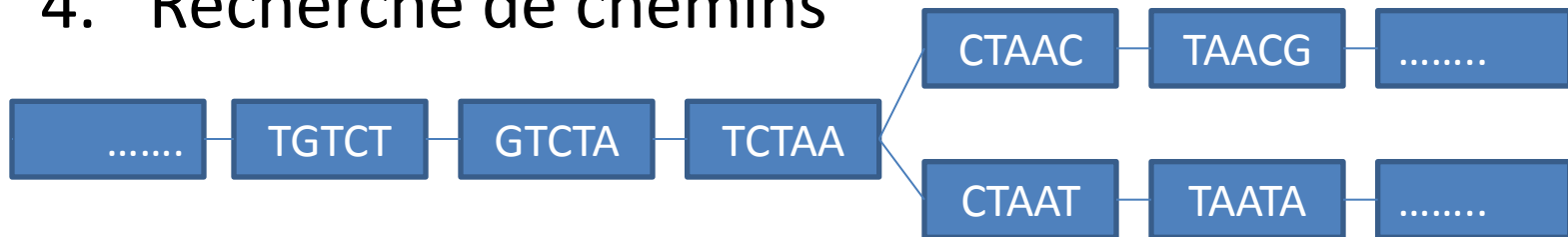
Nicolaas Govert de Bruijn

- Graphe de de-Bruijn
  - Reads  $\rightarrow$  k-mers
  - 1 Nœud = 1 k-mer
  - une arrête = 1 (k-1)-mer



# Stratégies d'assemblage

- Graphe de de-Bruijn / assemblage
  1. Nettoyage des reads
  2. Construction du graphe
  3. Nettoyage du graphe
  4. Recherche de chemins



- ...TGTCTAA
- CTAACG...
- CTAATA...

# Principaux outils d'assemblage

- Glouton

- SSAKE

- SHARCGS

- VCAKE



- Overlap

- Newbler (454)

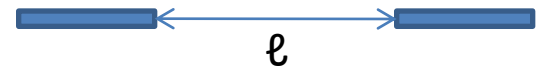
- CABOG (454)

- Edena (Illumina)

- Shorty (454+Illu)

# Principaux outils d'assemblage

- De-Bruijn
  - Euler (substitutions,  $\neq k$ , réduction graphe, paires)
  - **Velvet** (réduction graphe: topologie, couverture, paires)
  - ABySS (Parallèle: 3,5 Milliards de reads Illumina)
  - AllPaths (qualités des reads, larges génomes)
  - **SOAPnovo** (Overlap+DB, large génomes)
  - ...
  - Symbiose : Assembleur paires





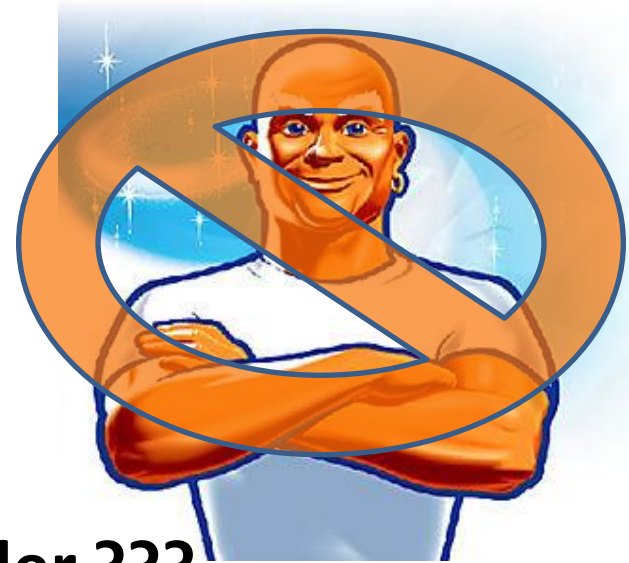
# Limitations des outils d'assemblage

- Gestion des erreurs de séquençage
  - Nettoyage statistique
- Gestion des répétitions
  - Divers heuristiques de résolutions

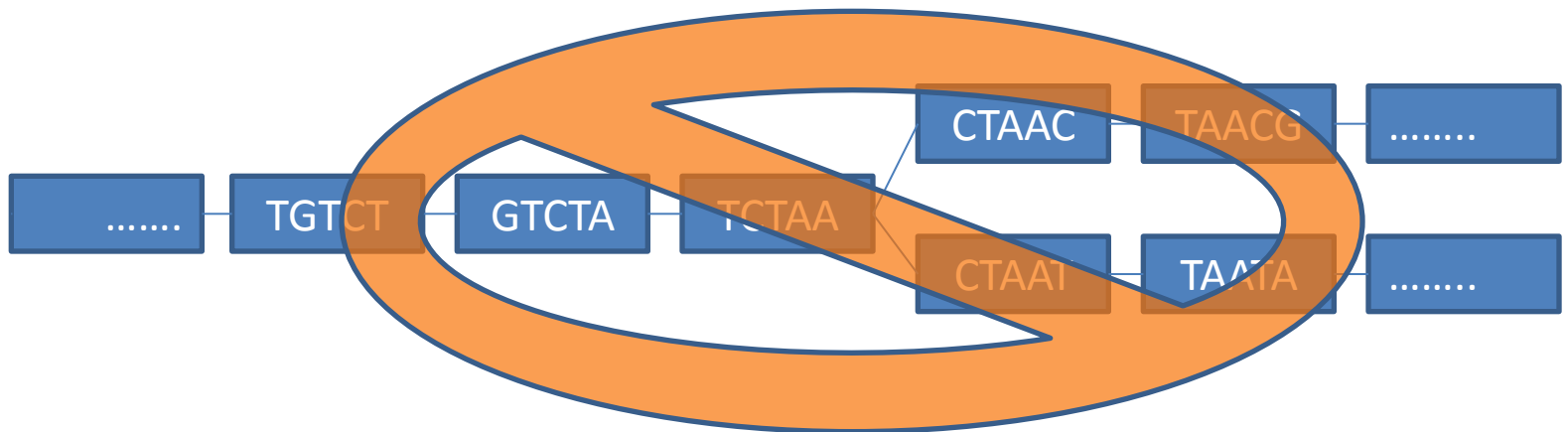


# Limitations des outils d'assemblage

- Gestion des erreurs de séquençage
  - Nettoyage statistique
- Gestion des répétitions
  - Divers heuristiques de résolutions



**Peut-on éviter d'assembler ???**



# Alcovna

## Algorithmes pour la Comparaison et la Visualisation de données Non Assemblées



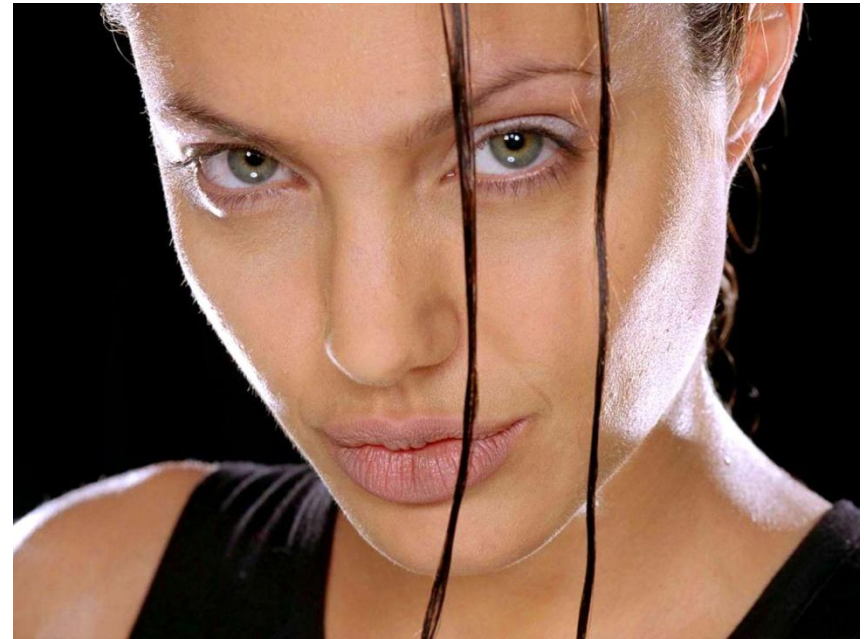
- Extraction d'informations dans les reads bruts:
  - **SNP : kisSnp**
  - Transcrits alternatifs / Splicing graphes
  - Éléments répétés
  - Recherche d'homologues

# Snp: Single Nucleotide polymorphism

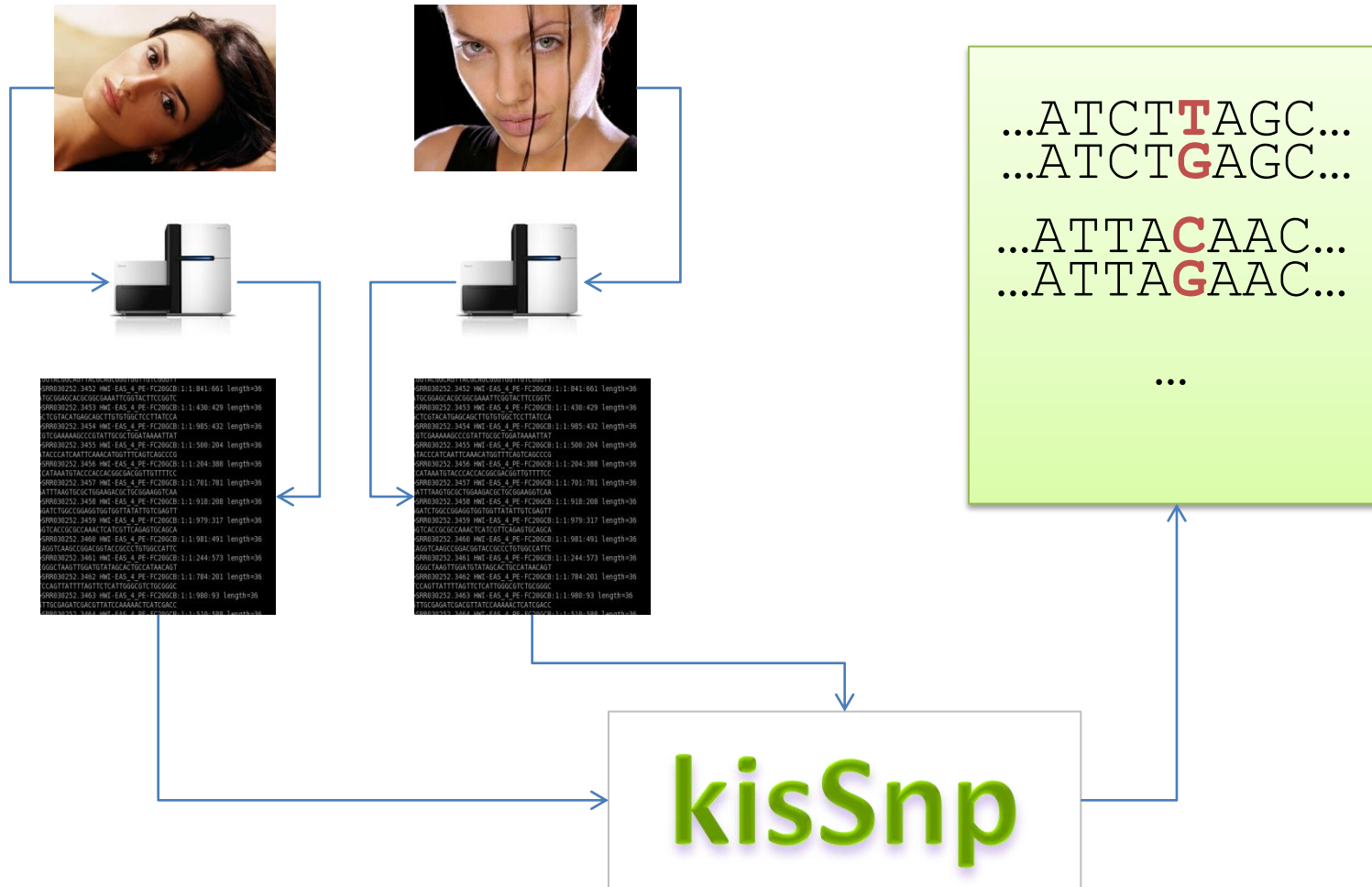
- Yeux marrons  
...ATCTTAGC...



- Yeux verts  
...ATCTGAGC...



# kisSnp: Entrée/Sortie



• Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. *Molecular Ecology*, 2010.

• Calling SNPs without a reference genome. *BMC Bioinformatics*, 11:130, 2010.

# kisSnp: Cœur de l'algo

## Cas le plus simple:

- SNP  
- Répétitions, ~~couverture variable~~
- Erreurs de séquençage



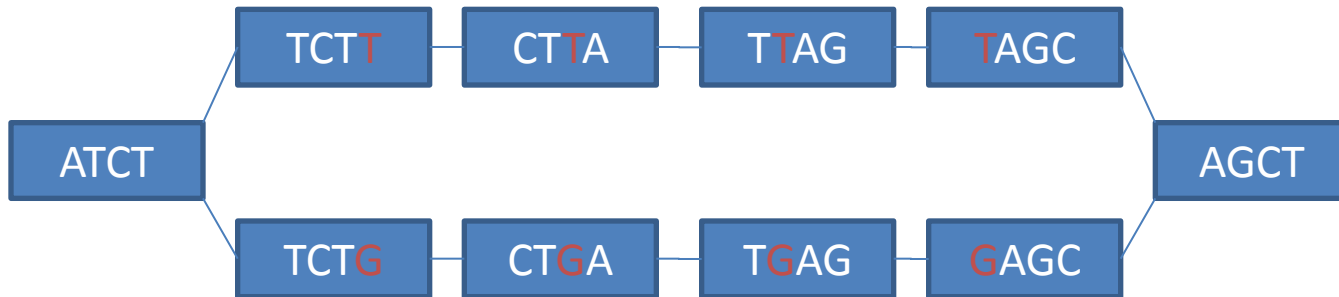
## Séquence:

- ...ATCTTAGC... 

# kisSnp: Cœur de l'algo

## Cas le plus simple:

- SNP  
- Répétitions, couverture variable
- Erreurs de séquençage



## Séquences:

• ...ATCT**T**AGCT...

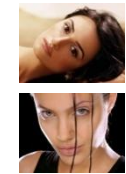
• ...ATCT**G**AGCT...



# kisSnp: Cœur de l'algo

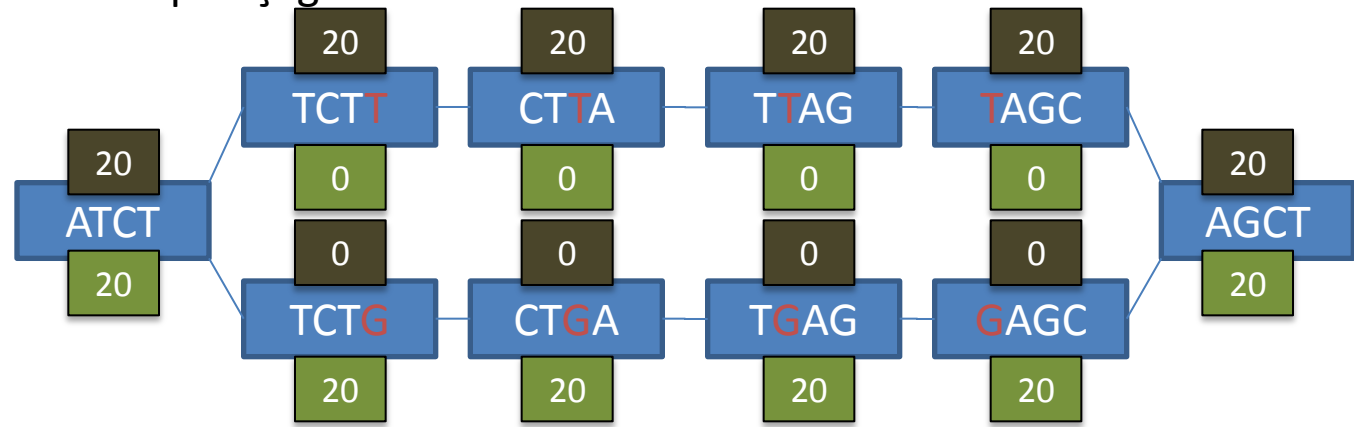
Cas le plus simple, avec comptage:

- SNP  
- Répétitions, ~~couverture variable~~
- Erreurs de séquençage



nombre de kmers

nombre de kmers



Séquences:

- ...ATCT**T**AGCT... 
- ...ATCT**G**AGCT... 



# kisSnp: Cœur de l'algo

## + répétitions, couverture variable

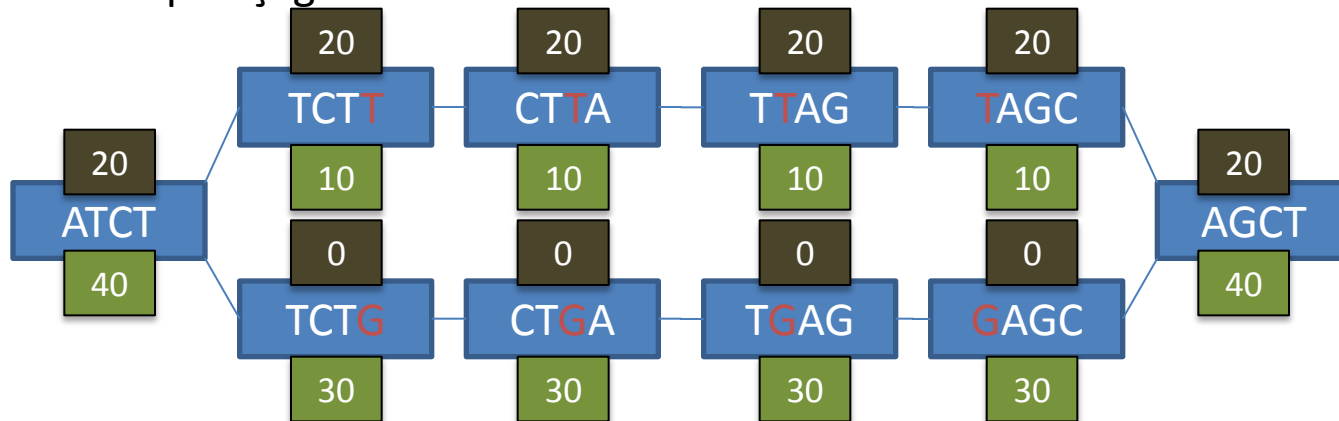
- SNP  
- Répétitions, couverture variable
- Erreurs de séquençage



nombre de kmers



nombre de kmers




## Séquences:

• ...ATCT**T**AGCT... 

• ...ATCT**G**AGCT... 

# kisSnp: Cœur de l'algo

## + erreurs de séquençage

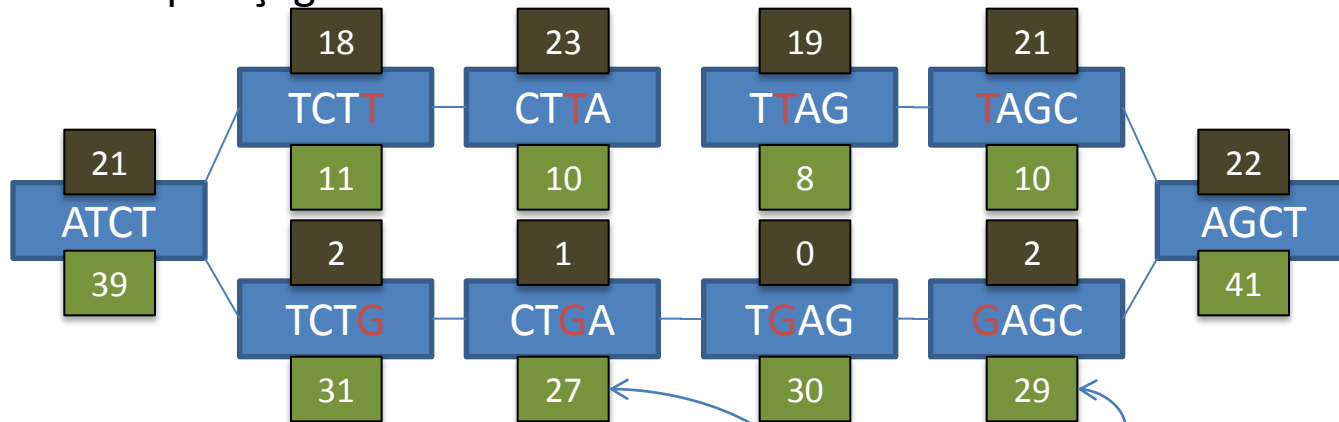
- SNP  
- Répétitions, couverture variable
- Erreurs de séquençage



nombre de kmers



nombre de kmers



## Séquences:

• ...ATCT**T**AGCT...



• ...ATCT**G**AGCT...



$\delta$

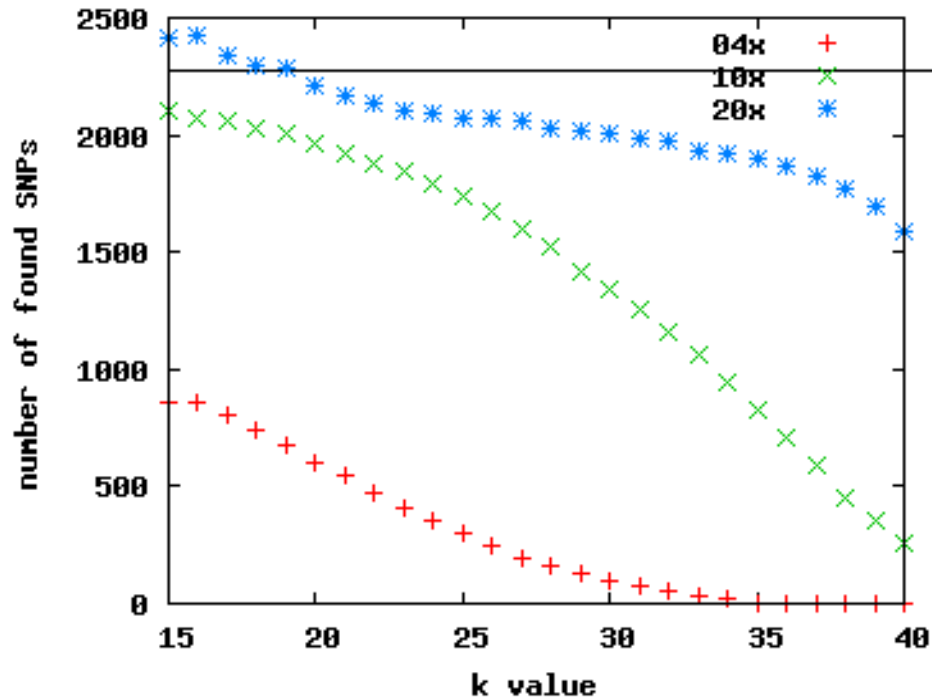
# Quelques resultats

- Sur données réelles et simulées
- 2 paramètres  $k$  et  $\delta$
- Qualité des données
  - couverture
  - erreurs
  - répétitions

Données  
simulées

# Apperçu global

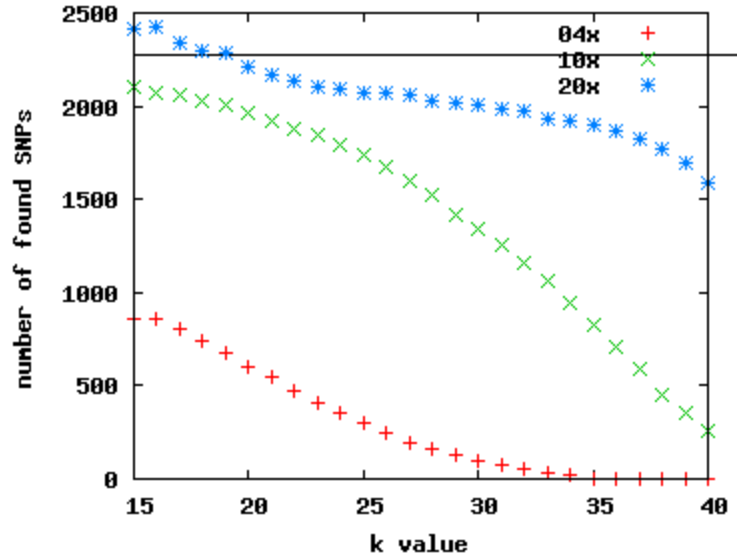
MC58 sequence (22 Mb)



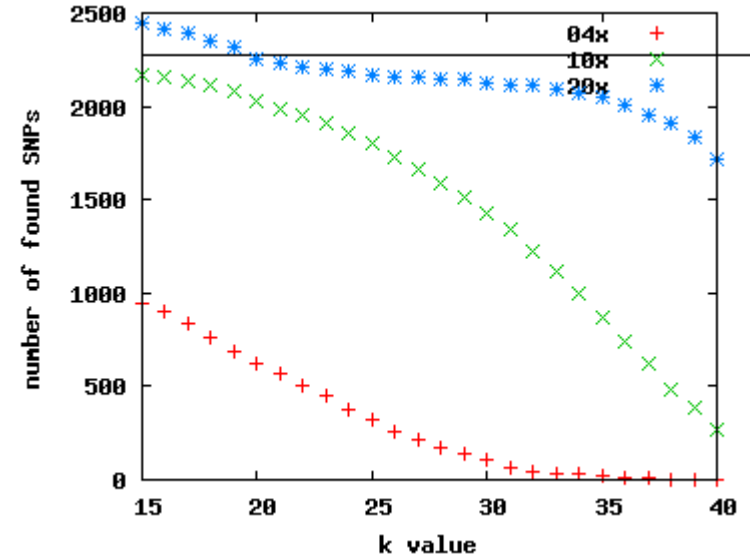
- avec  $k$  « assez gros ( $\geq 20$ ) » : aucun faux positifs
- jeux de paramètres robuste (ici  $\delta = 20$ )

# Et les répétitions ?

MC58 sequence



MC58 sequence shuffle

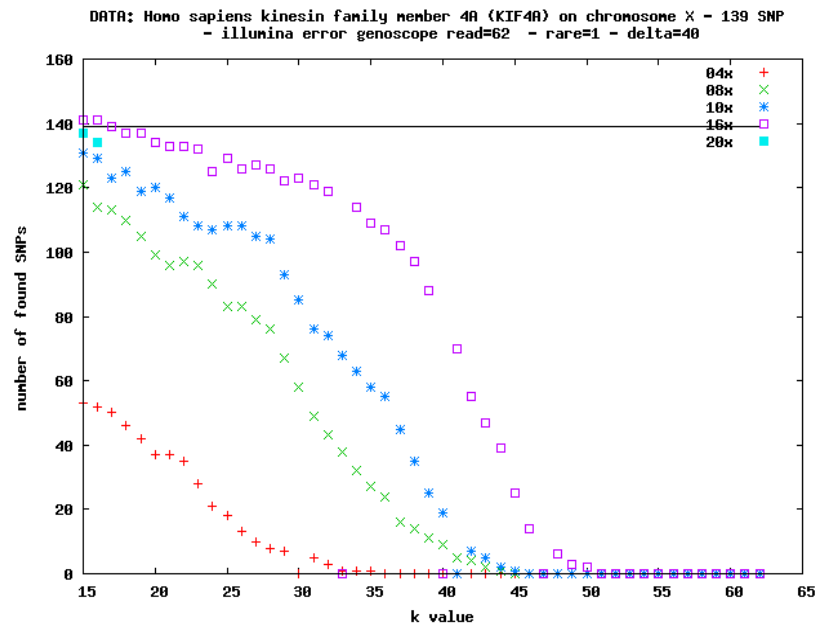


- Peu sensible aux répétitions

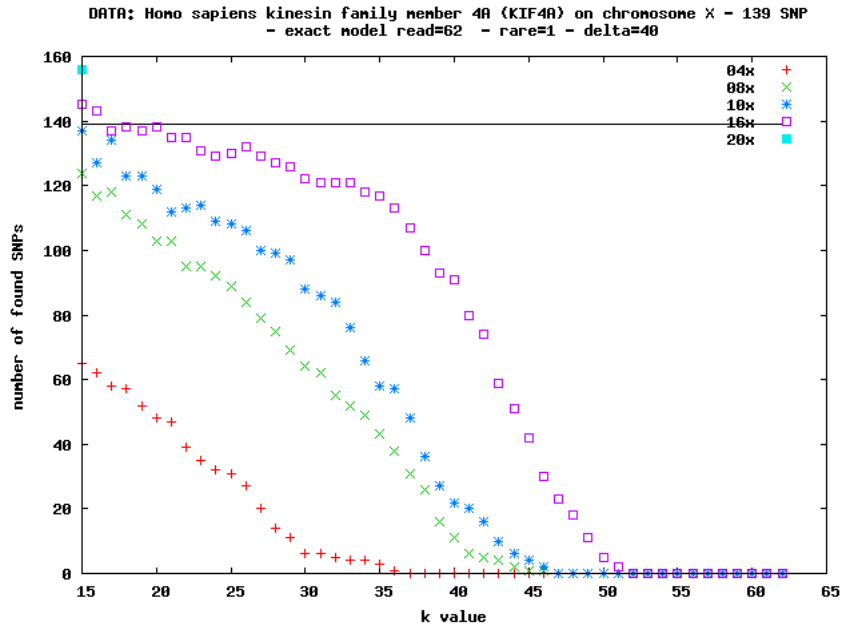
Données  
simulées

# Et les erreurs de séquençage ?

Avec erreurs (Illumina)



Sans erreurs



- Peu sensible aux erreurs de séquençage type illumina (substitutions)

Données  
réelles

"Genome evolution and adaptation in a long-term experiment with *Escherichia coli*", *Nature* 2009

## Expérimentations sur E. Coli

Etude originale : Comparaisons de générations

- 0 vs. 20K: 28 SNPs trouvés par mapping

KisSnp sur 0 vs. 20K reads:

- qq minutes, qq Gb Ram
- 27 des 28 SNPs
- 42 add. structures de SNP, non détectées précédemment



# Conclusion

## Bons points:

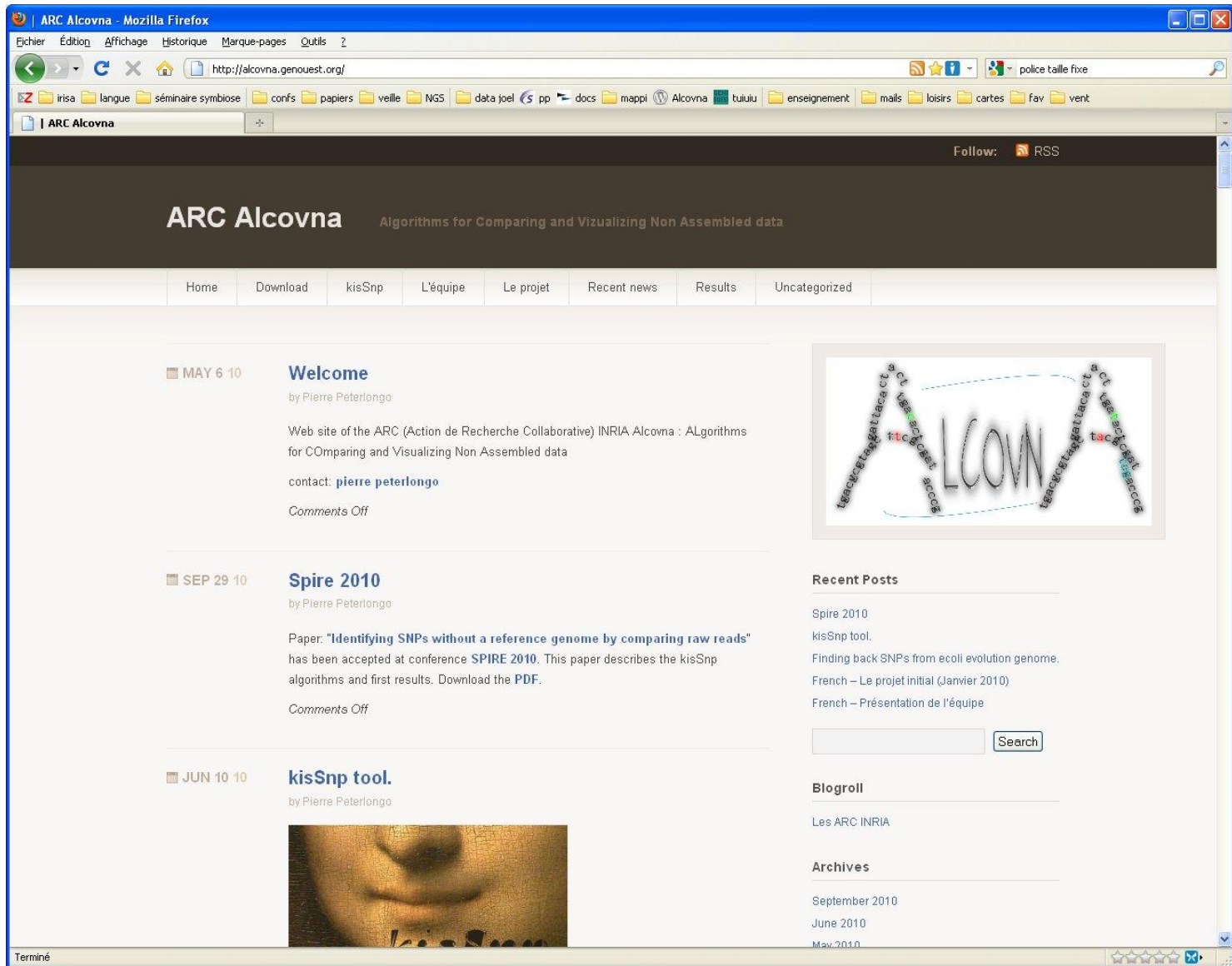
- Pas d'assemblage
- Pas de référence
- Pas de nettoyage
- Rapide et léger
- SNP:
  - Uniquement 2 paramètres
  - Peu sensible erreurs/repet.

## Futur

- SNPs intra individu
- SNPs  $n$  individus
  - + phylogénie
- Avancées Alcovna
  - ARN
  - Homologies



infos et téléchargement: <http://alcovna.genouest.org/>



The screenshot shows a Mozilla Firefox browser window displaying the website <http://alcovna.genouest.org/>. The browser's address bar and menu bar are visible at the top. The website header features the title "ARC Alcovna" and the subtitle "Algorithms for Comparing and Visualizing Non Assembled data". A navigation menu includes links for Home, Download, kisSnp, L'équipe, Le projet, Recent news, Results, and Uncategorized. The main content area displays three blog posts:

- MAY 6 10**: **Welcome** by Pierre Peterlongo. The post describes the website of the ARC (Action de Recherche Collaborative) INRIA Alcovna, which focuses on algorithms for comparing and visualizing non-assembled data. Contact information for Pierre Peterlongo is provided, and comments are turned off.
- SEP 29 10**: **Spire 2010** by Pierre Peterlongo. The post announces that a paper titled "Identifying SNPs without a reference genome by comparing raw reads" has been accepted at the SPIRE 2010 conference. It describes the kisSnp algorithms and first results, and provides a link to download the PDF. Comments are turned off.
- JUN 10 10**: **kisSnp tool.** by Pierre Peterlongo. The post features a partial image of a person's face with the text "kisSnp" overlaid.

On the right side of the page, there is a "Recent Posts" section listing "Spire 2010", "kisSnp tool.", "Finding back SNPs from ecoli evolution genome.", "French – Le projet initial (Janvier 2010)", and "French – Présentation de l'équipe". Below this is a search box with a "Search" button. A "Blogroll" section lists "Les ARC INRIA". An "Archives" section shows dates from "September 2010" to "May 2010". The browser's status bar at the bottom left shows "Terminé".

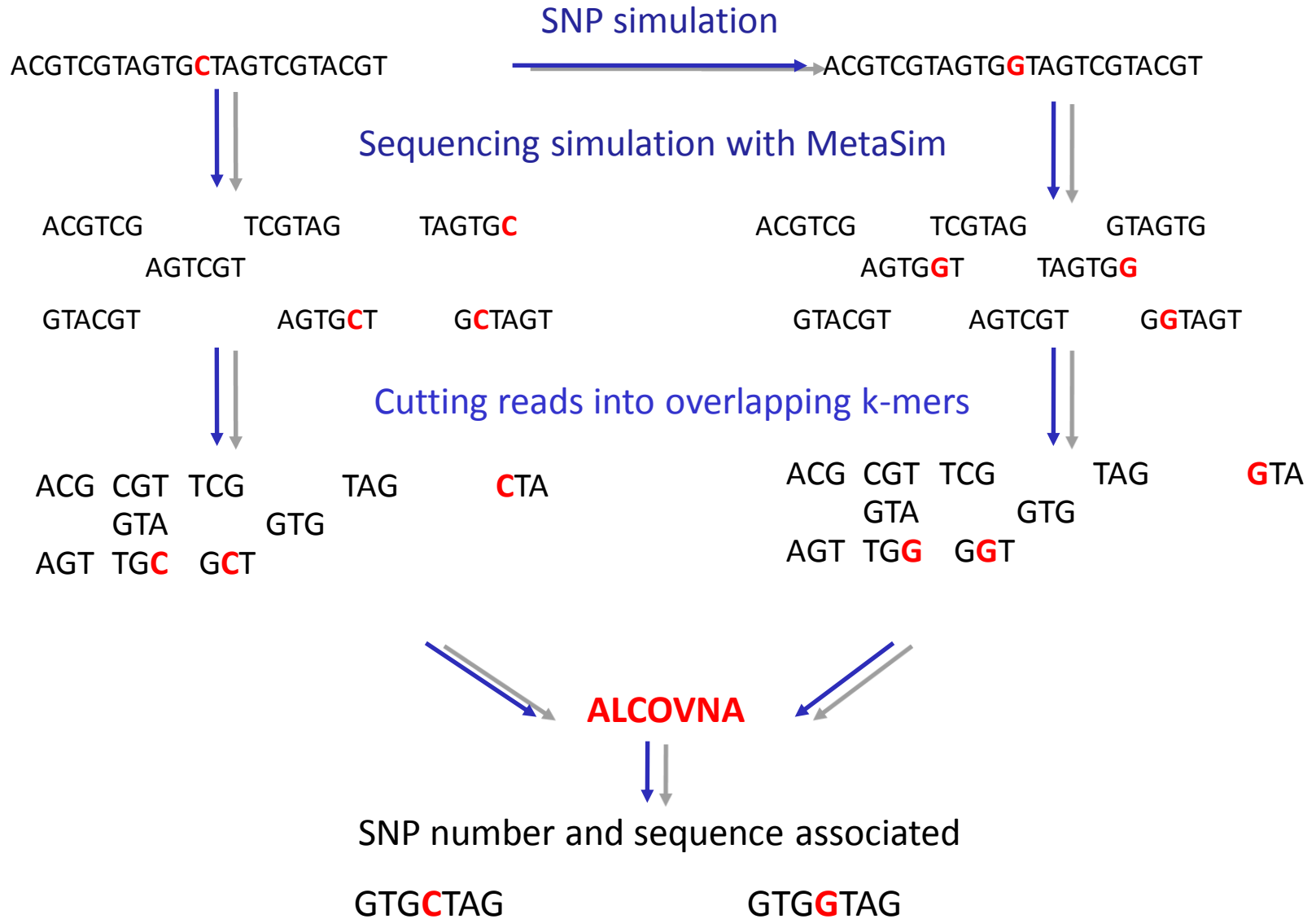
Peterlongo - Assembler... ou pas.

# Groupe NGS - IRISA

- Discussion  $\approx$  mensuelle
- Informel
- Biblio = Journal club ?
- Zone d'interactions développeurs / utilisateurs
- Soyez les bienvenus: pierre.peterlongo@inria.fr



# Experiment data



# NNGS / Third Generation

Features	Second-generation sequencers		
	454-FLX	Solexa	SOLiD
Read-length (bp)	240–400	35	35
Cost/human genome (US\$)	1 000 000	60 000	60 000
Run time (h/Gb)	75	56	42
Ease of use	Difficult	Difficult	Difficult

## Third-generation sequencers (single molecule-SBS)

Helicos tSMS	PacBio SMRT	Nanopore and modified forms	ZS Genetics TEM
30	100 000	Potentially unlimited?	Potentially unlimited?
70 000	Low	Low	Low
~12	<1	>20	~14
Easy	Easy	Easy	Easy