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

IRISA

Symbiose

INRIA

Biogenouest

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 Bio-systems' 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

Paszkiewicz and Studholme, *De novo* assembly of short sequence reads, Briefings in Bioinformatics, 11(5) 2010

Quelles technologies ?

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

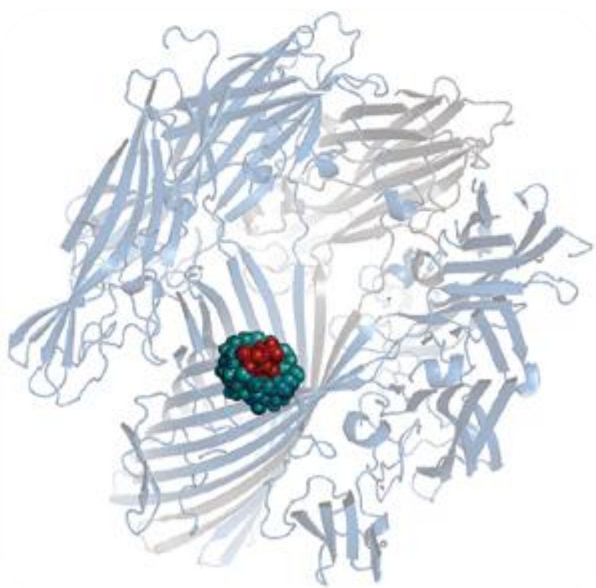
Quelles technologies ?

- 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

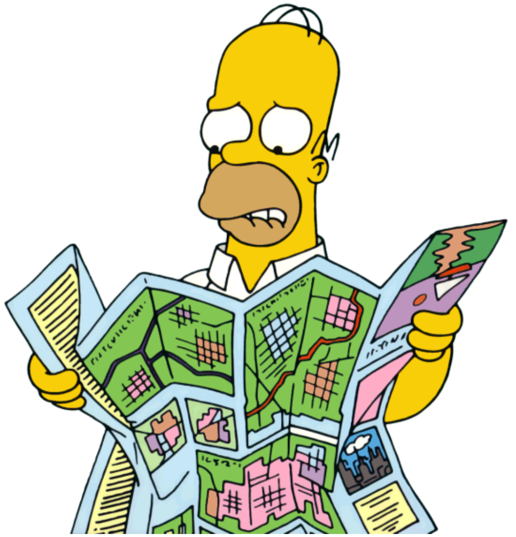
Que faire des reads ?
Premiers reflexes

- Avec un génome de référence fiable :
  - Read mapping
- Sans génome de référence :
  - Assemblage *de novo*

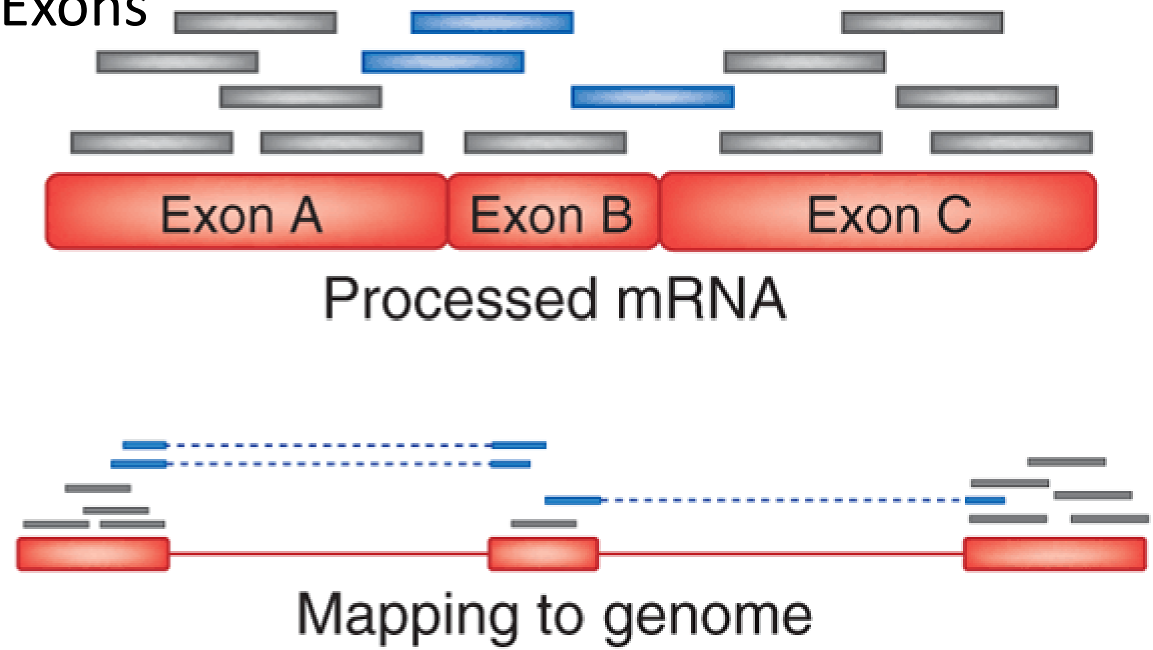
- Si on dispose de génome(s) peu fiables :
  - Mapping + assemblage *de novo*

Read Mapping

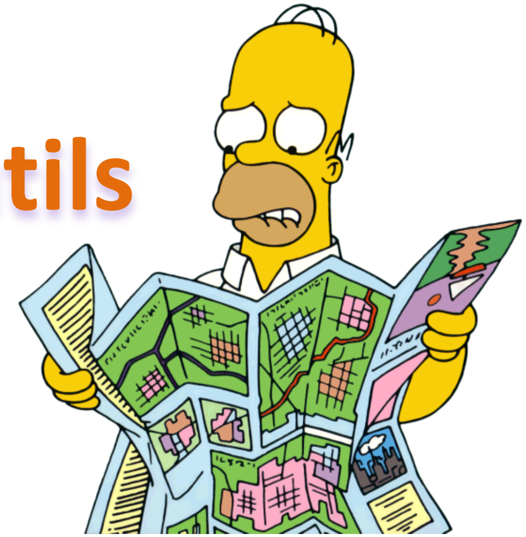
Read Mapping



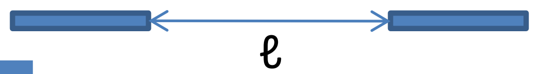
- Difficultés:
  - Reads mappant à plusieurs locus
  - RNASeq:
    - Introns Exons



Read Mapping : Les outils



- Idée globale:
  - Indexation Reads/Génome
  - Utilisation de graines (≈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
GASSST	hash ref	X	X		

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

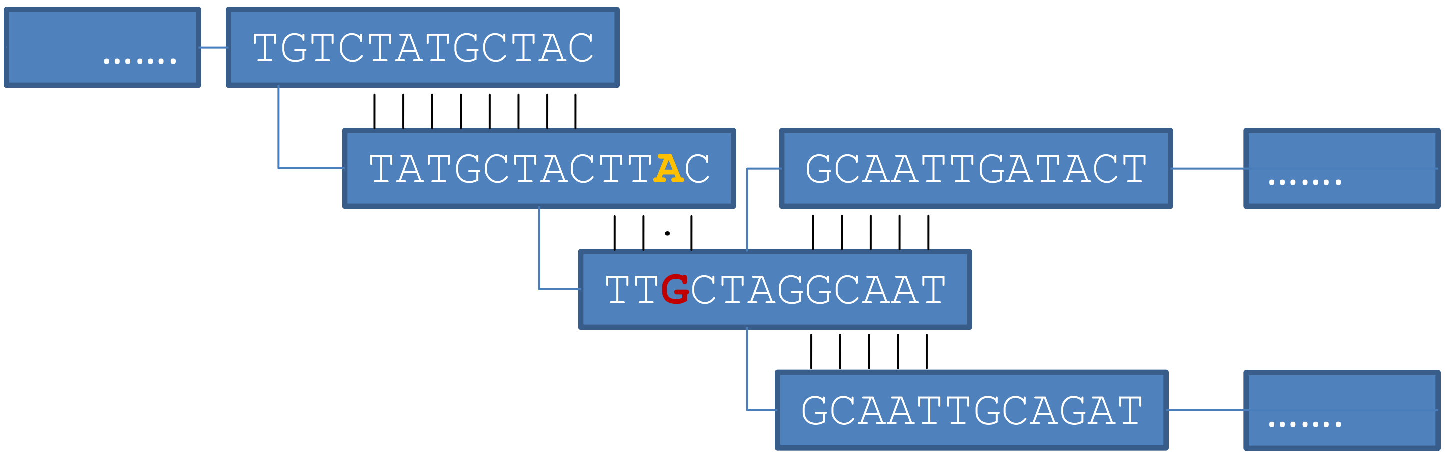
A survey of sequence alignment algorithms for next-generation sequencing, Li and Homer, Brief. Bioinfo., 11(5), 2010

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

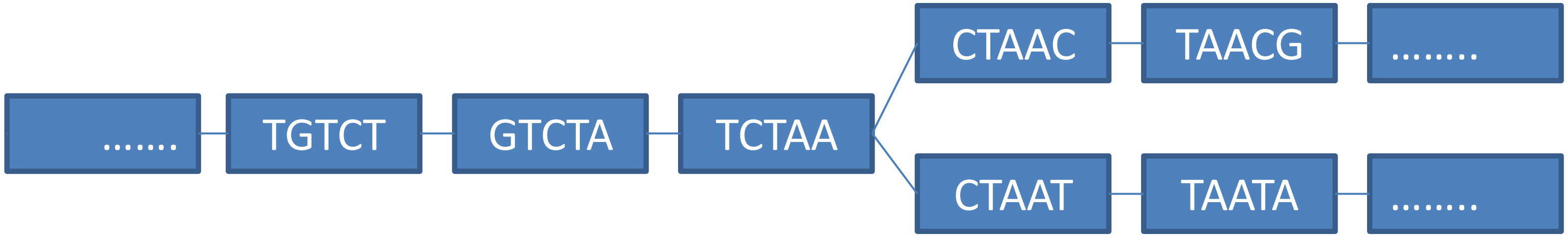
- Graphe des reads chevauchants (overlap graph)



Stratégies d'assemblage

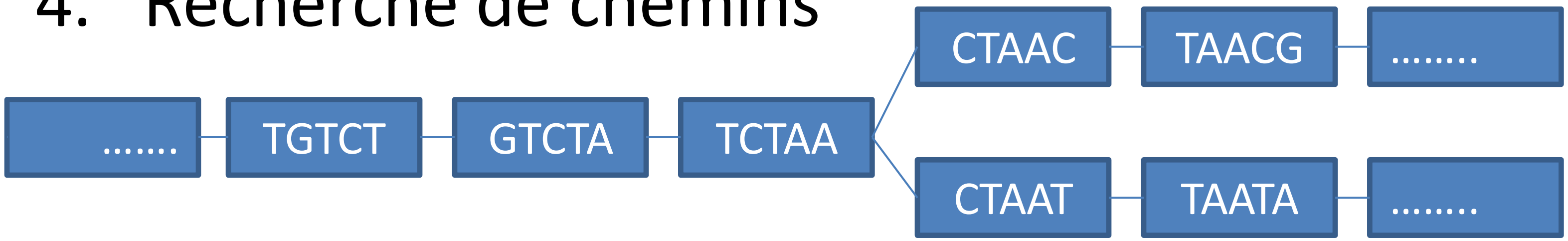

Nicolaas Govert de Bruijn

- Graphe de de-Bruijn
  - Reads -> 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

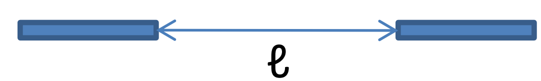
- Glouton
  - SSAKE
  - SHARCGS
  - VCAKE

- Overlap
  - Newbler (454)
  - CABOG (454)
  - Edena (Illumina)
  - Shorty (454+Illu)

Assembly algorithms for next-generation sequencing data, Miller, Genomics, 95(6), 2010

Principaux outils d'assemblage

- De-Bruijn
  - Euler (substitutions, ≠ 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



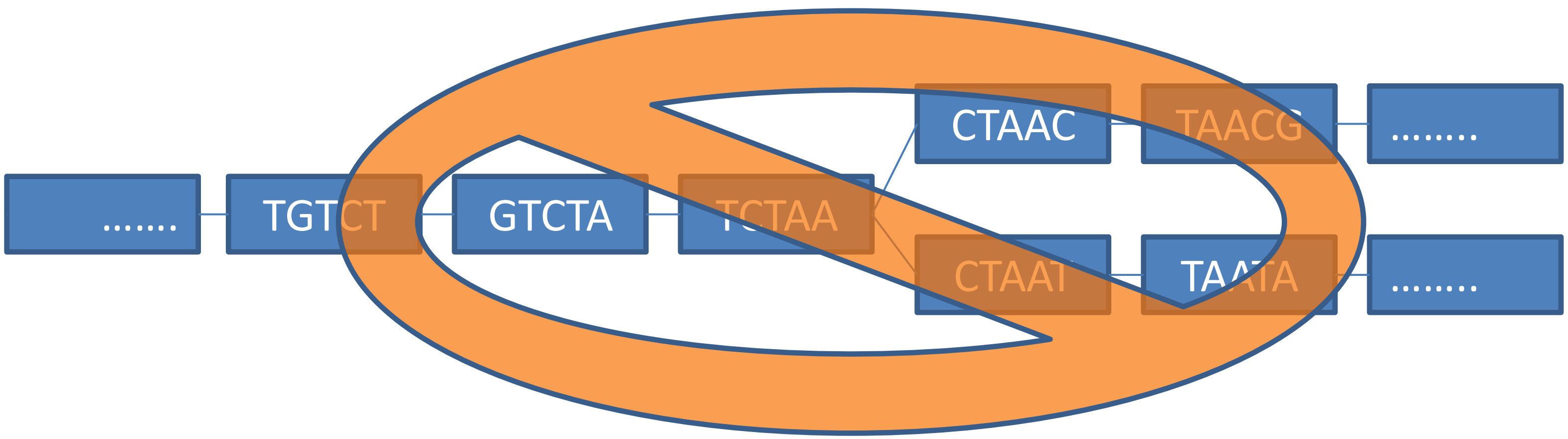
Limitations des outils d'assemblage

- 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
  - Eléments répétés
  - Recherche d’homologues

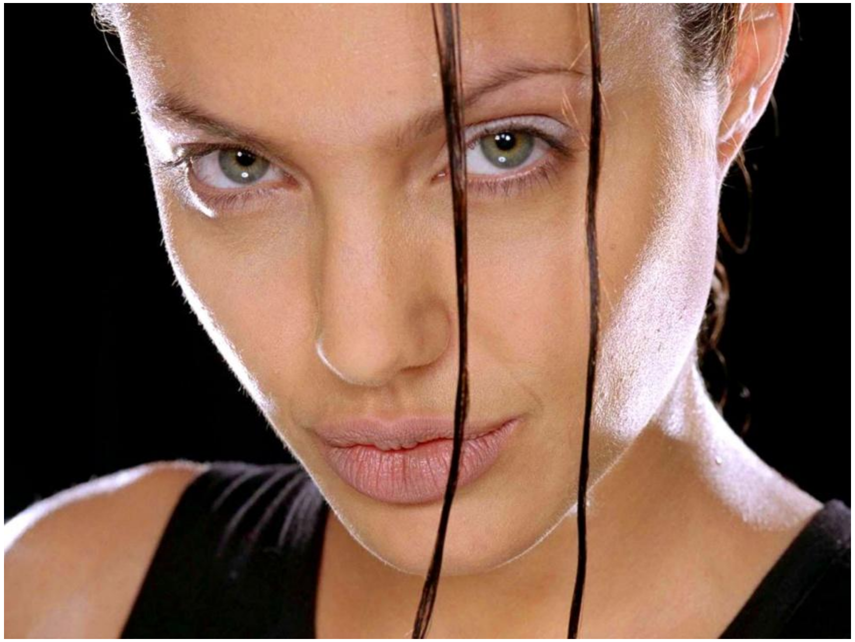
Snp: Single Nucleotide polymorphism

- Yeux marrons

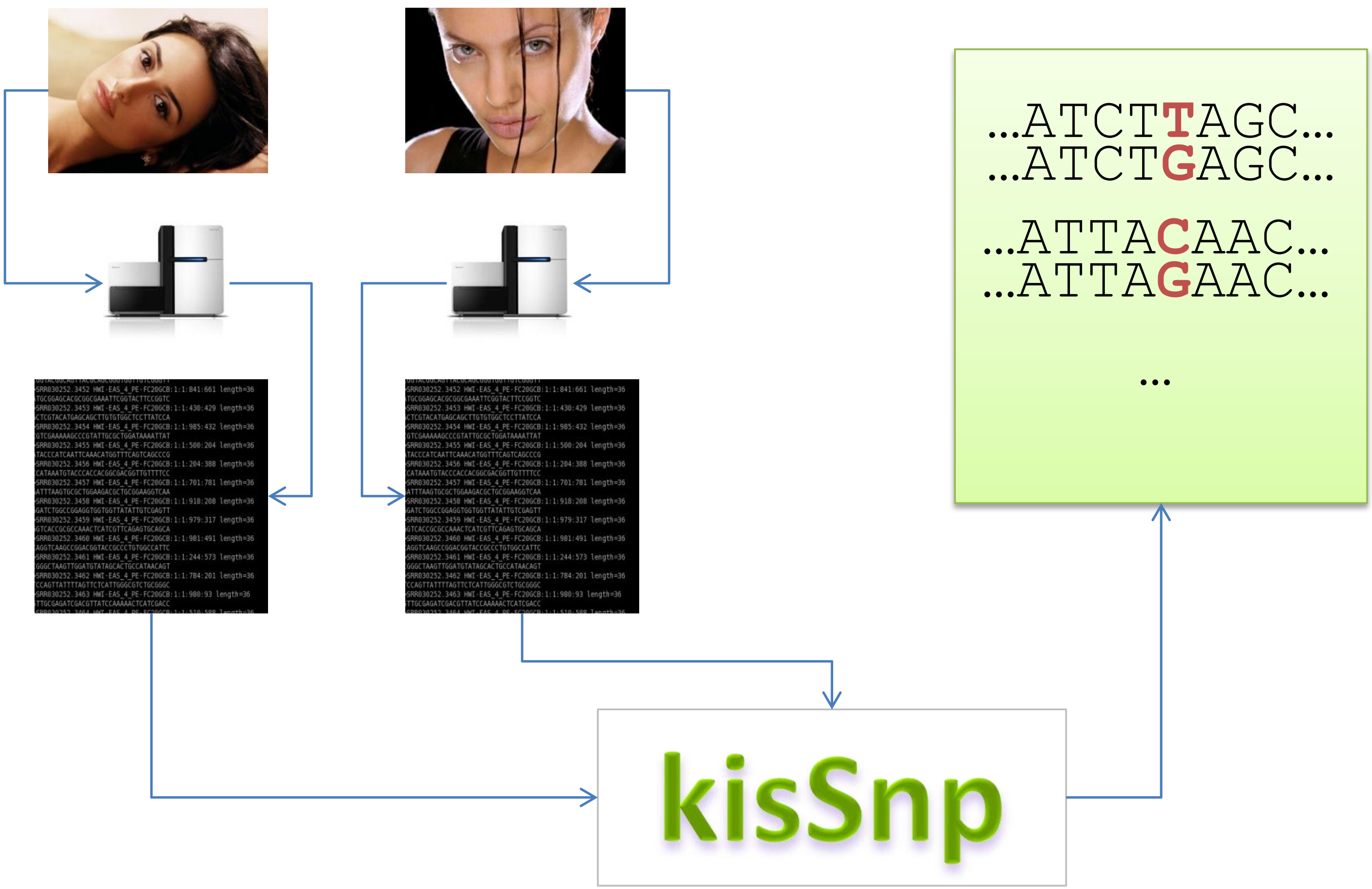
  …ATCT**T**AGC…

- Yeux verts

  …ATCT**G**AGC…



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~~

ATCT — TCTT — CTTA — TTAG — TAGC — AGCT

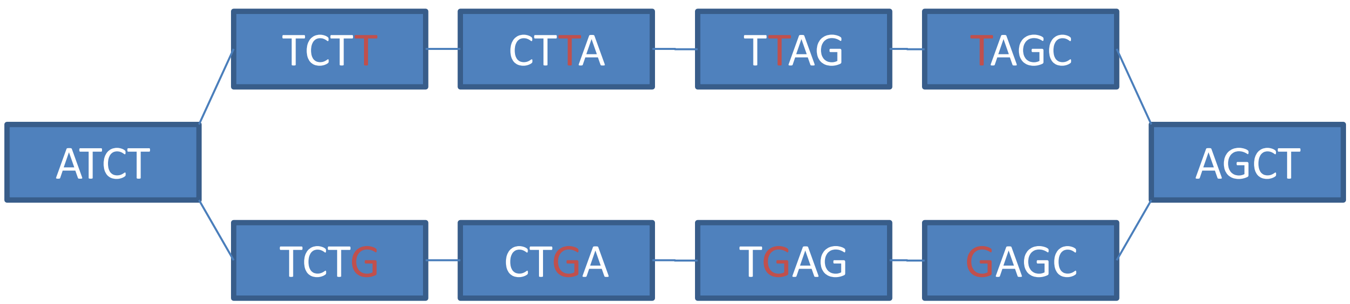
Séquence:

- …ATCTTAGC… 

kisSnp: Cœur de l'algo

Cas le plus simple:

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



Séquences:

- …ATCT**T**AGCT… 
- …ATCT**G**AGCT… 

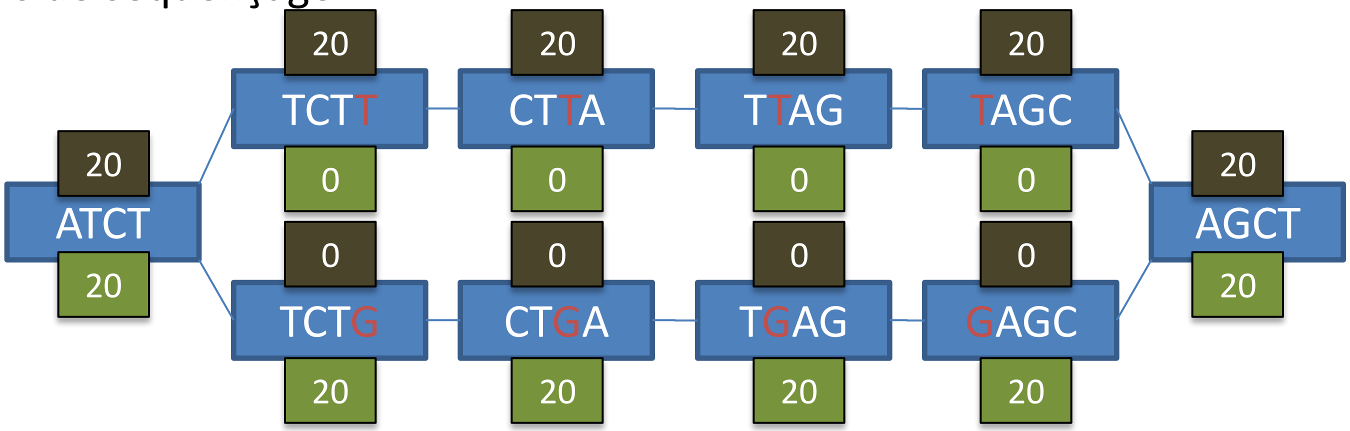
kisSnp: Cœur de l'algo

nombre de kmers

nombre de kmers

Cas le plus simple, avec comptage:

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



Séquences:

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

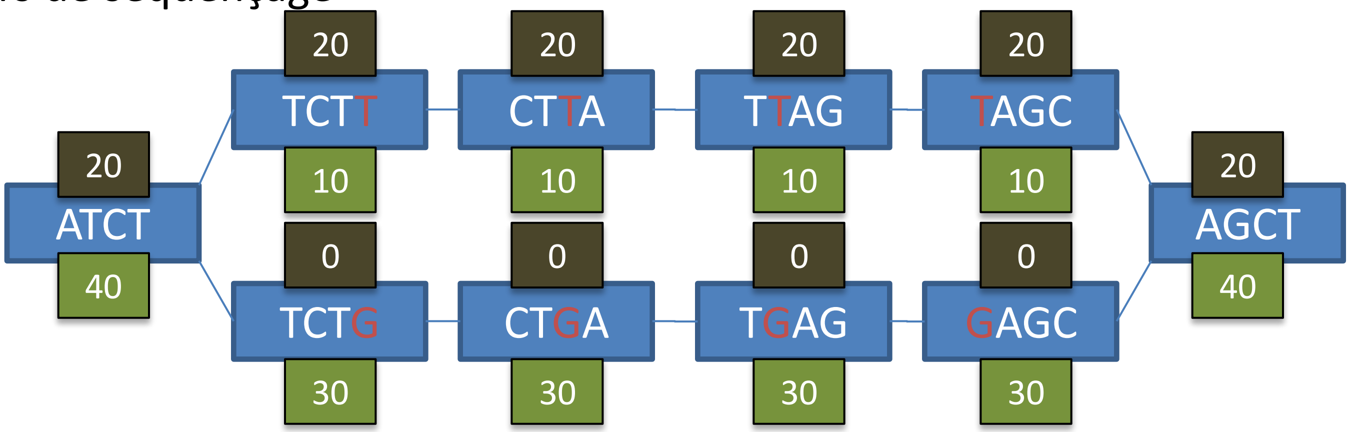
kisSnp: Cœur de l'algo

 nombre de kmers

 nombre de kmers

+ répétitions, couverture variable

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



Séquences:

- …ATCT**T**AGCT… 
- …ATCT**G**AGCT… 

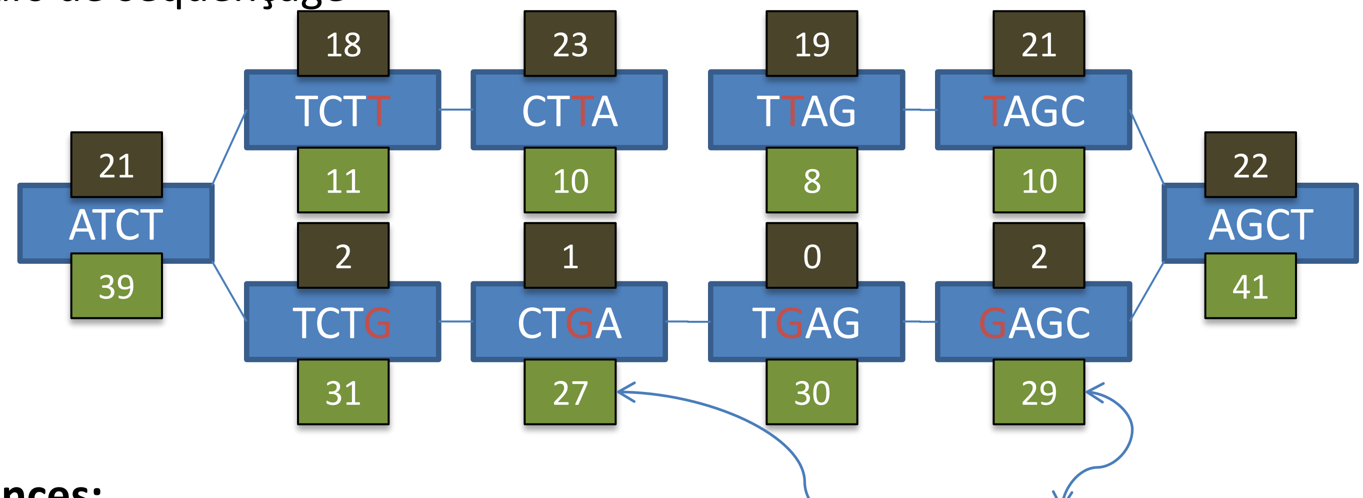
kisSnp: Cœur de l'algo


nombre de kmers


nombre de kmers

+ erreurs de séquençage

- SNP  
- Répétitions, couverture variable
- Erreurs de séquençage



Séquences:

- …ATCT**T**AGCT… 
- …ATCT**G**AGCT… 

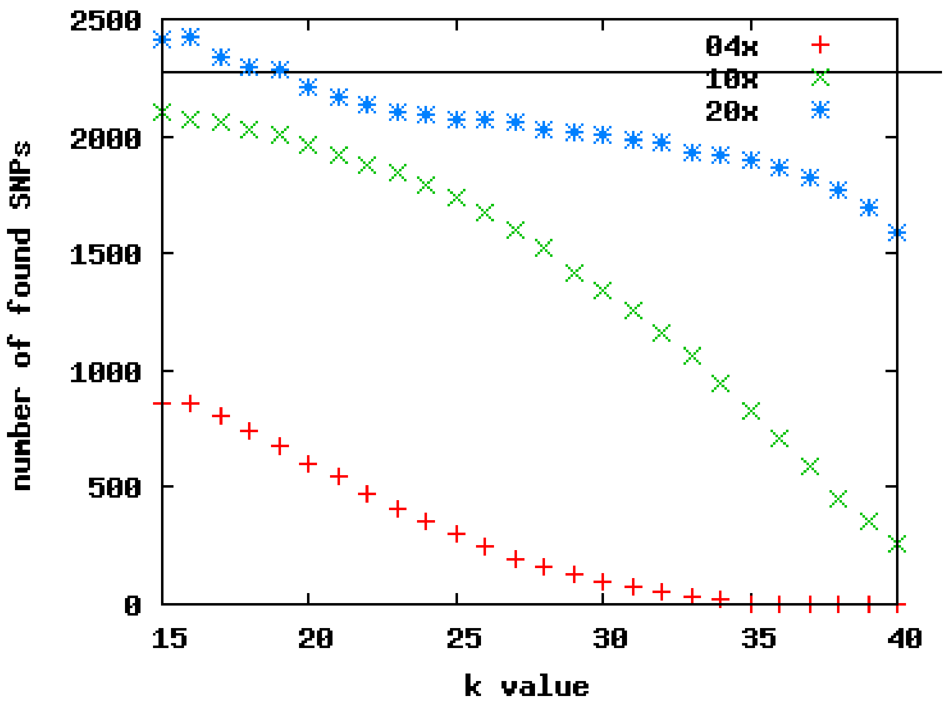
Quelques resultats

- Sur données réelles et simulées
- 2 paramètres k et δ
- Qualité des données
  - couverture
  - erreurs
  - répétitions

Données simulées

Apperçu global

MC58 sequence (22 Mb)

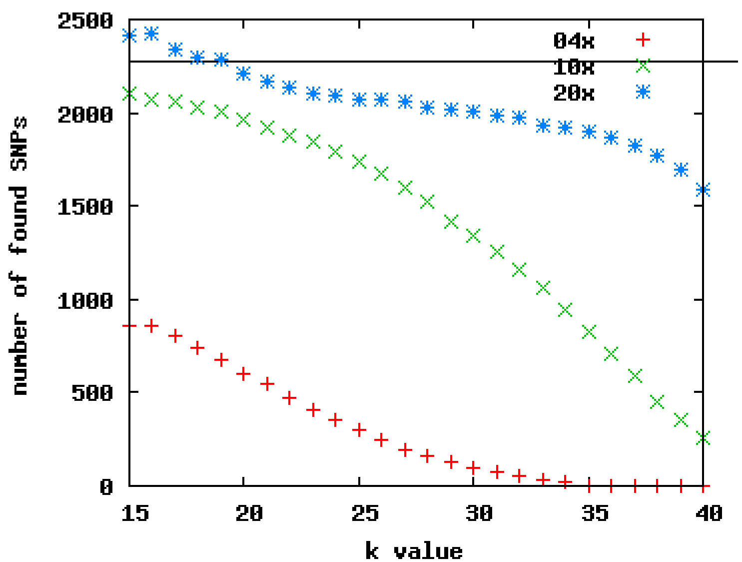


- avec *k* « assez gros (≥20) » : aucun faux positifs
- jeux de paramètres robuste (ici δ = 20)

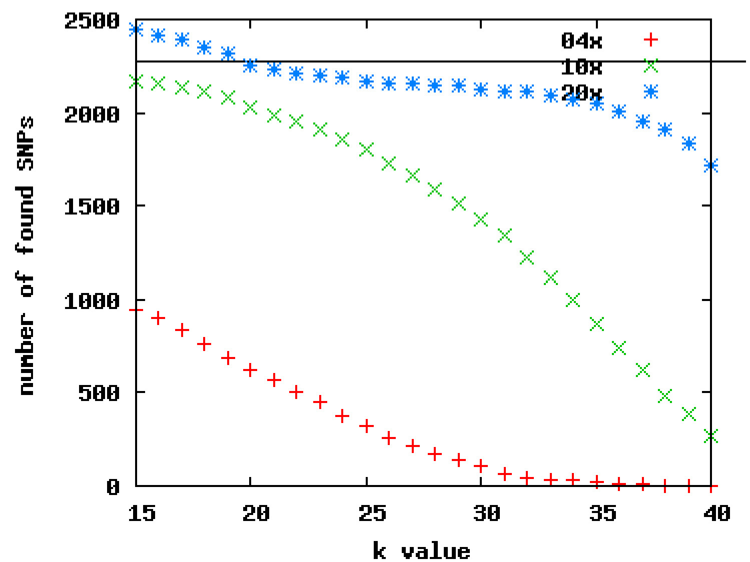
Et les répétitions ?

Données simulées

MC58 sequence



MC58 sequence shuffle

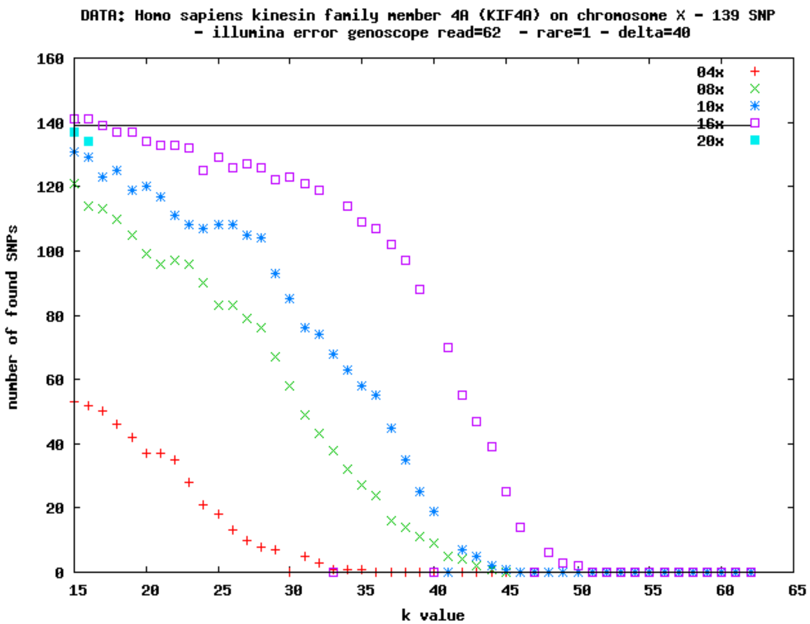


- Peu sensible aux répétitions

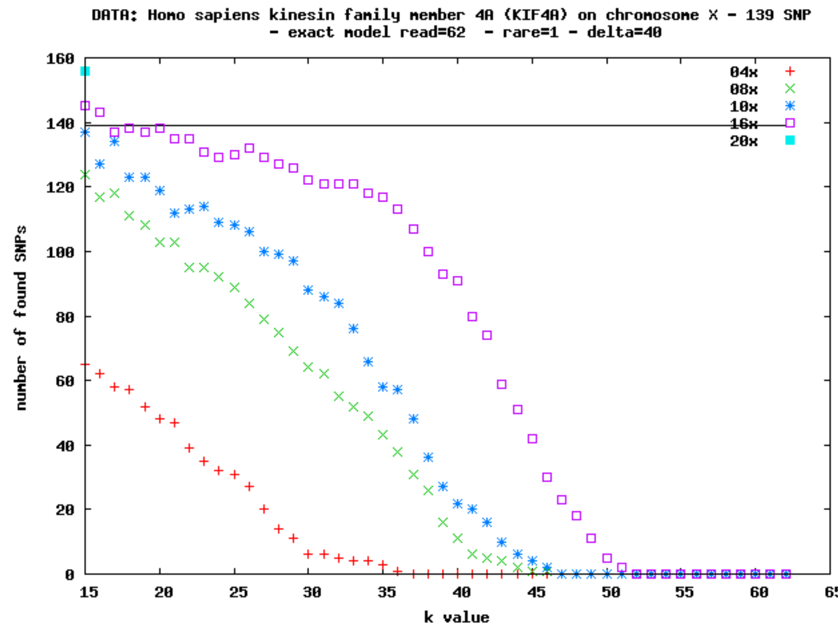
Données simulées

Et les erreurs de sequençage ?

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

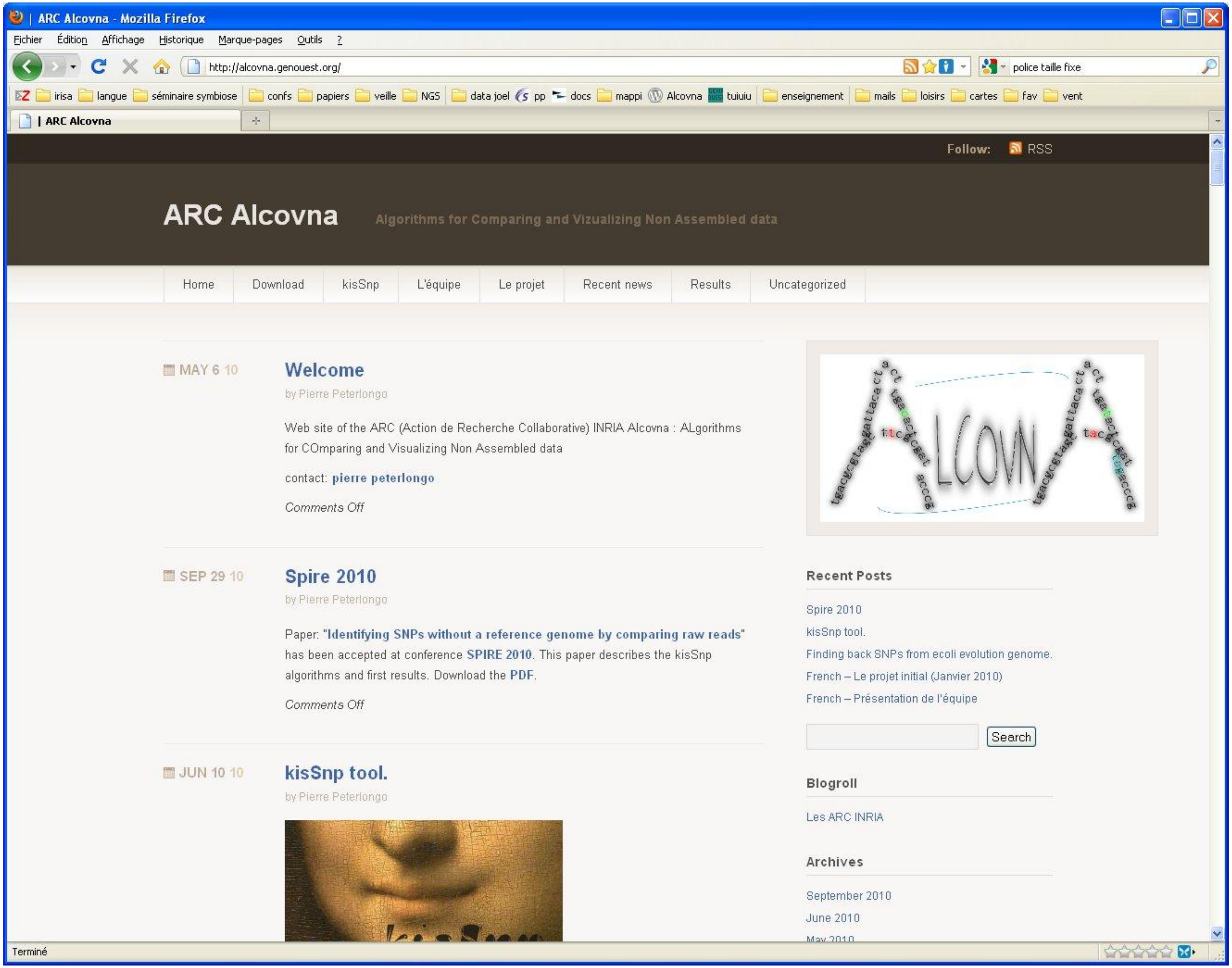
Bons points:

- 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/



Groupe NGS - IRISA

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

Experiment data

SNP simulation

ACGTCGTAGTGCTAGTCGTACGT → ACGTCGTAGTGGTAGTCGTACGT

Sequencing simulation with MetaSim

ACGTCG TCGTAG TAGTGC
AGTCGT
GTACGT AGTGCT GCTAGT

ACGTCG TCGTAG GTAGTG
AGTGGT TAGTGG
GTACGT AGTCGT GGTAGT

Cutting reads into overlapping k-mers

ACG CGT TCG TAG CTA
GTA GTG
AGT TGC GCT

ACG CGT TCG TAG GTA
GTA GTG
AGT TGG GGT

ALCOVNA

SNP number and sequence associated

GTGCTAG GTGGTAG

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
$\sim$12	$<$1	$>$20	$\sim$14
Easy	Easy	Easy	Easy

Single-molecule DNA sequencing technologies for future genomics research Gupta, TB 26(11), 2008