FC bio-informatique Fev 2015

NGS : des reads aux SNPs

(Vincent Ranwez – Montpellier SupAgro)

NGS : des reads aux SNPs

- Introduction NGS
  - Exemples d'applications
  - Rapide comparatif des technologies
  - Les données reçues
  - Nettoyage des reads
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

Applications des NGS

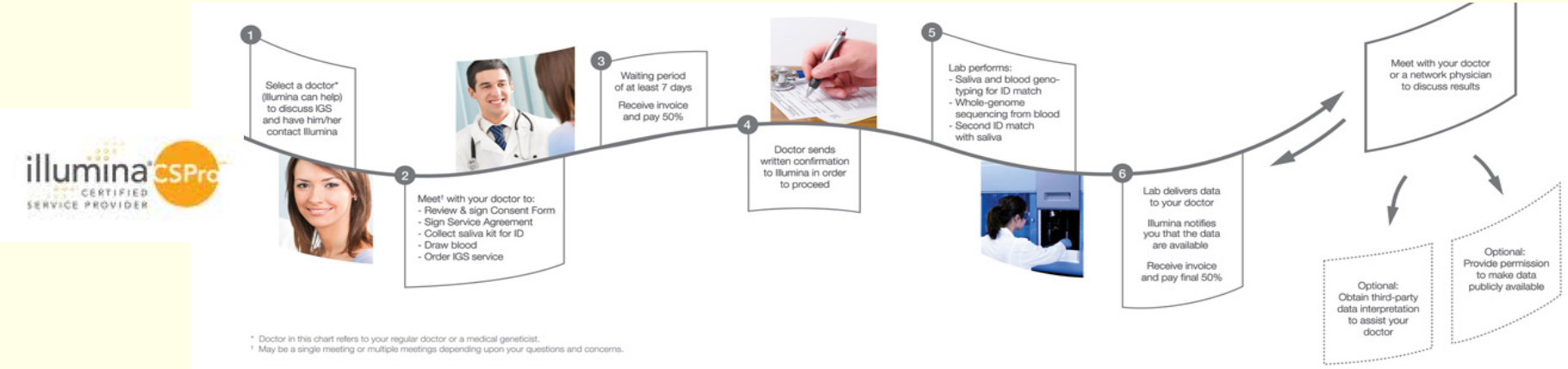
- Santé (humaine) : lier une maladie à une zone du génome
  - Reséquençage massif 1000 génomes, 2500 génomes etc.

http://www.1000genomes.org/

Applications des NGS

- Santé humaine :
  - médecine personnalisée ☺
  - Assurance personnalisée ☹

⇨ Nombreuses questions médicales, juridiques et éthiques

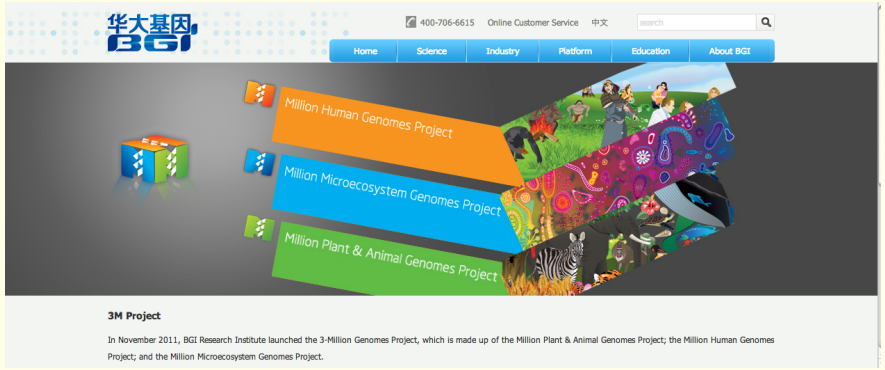


Applications des NGS

- Etude de la diversité inter et intra espèces
  - Phylogénie, flux de gènes, protection de la bio-diversité

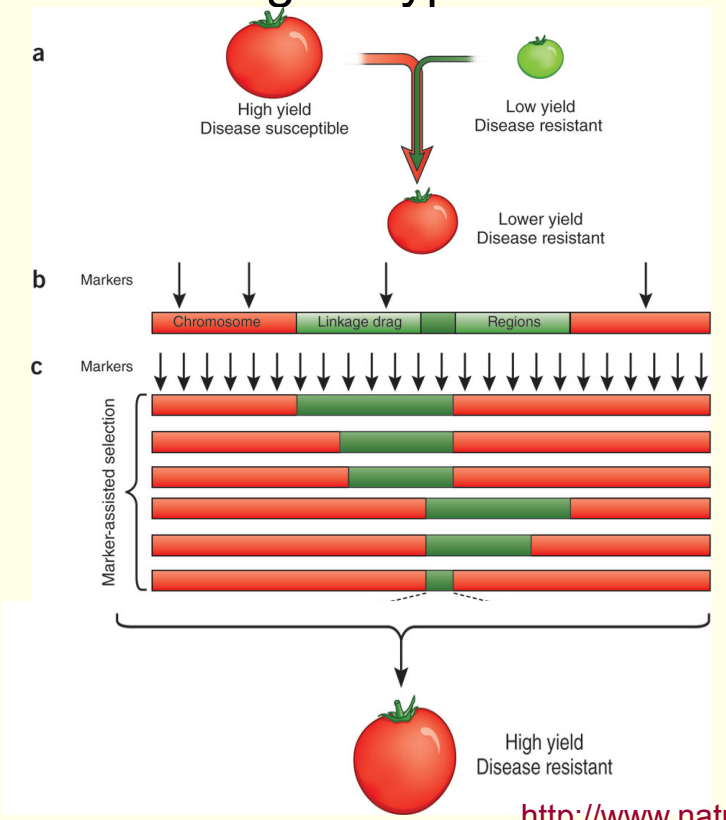
http://www.1001genomes.org/

http://genome10k.soe.ucsc.edu/



Applications des NGS

- Identification de marqueurs génomiques sur l'ensemble du génomes : SNP
  - Sélection assistée par marqueurs
  - lien génotype X phénotype
  - interactions génotype X environnement

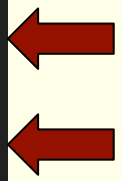


http://www.nature.com/nbt/journal/v30/n8/fig_tab/nbt.2319_F1.html

Rapide comparatif des technologies

NGS platforms overview

Technology	454			Solexa			SOLiD	
Platform:	GS 20	FLX	Ti	GA	GA II	1	2	
Reads:	500 k	500 k	1 M	28 M	80 M		40 M	115 M
Fragment								
Read length:	100	200	350	35	50	75	25	35
Run time:	6 hr	7 hr	9 hr	3 d	3 d	4 d	6 d	5 d
Yield:	50 Mb	100 Mb	400 Mb	1 Gb	4 Gb	6 Gb	1 Gb	4 Gb
Yield per hour:	8,3 Mb/hr	14,3 Mb/hr	44,4 Mb/hr	13,9 Mb/hr	55,6 Mb/hr	62,5 Mb/hr	6,9 Mb/hr	33,3 Mb/hr
Images:	11 GB	13 GB	27 GB	500 GB	1.1 TB	1.7 TB	1.8 TB	2.5 TB
PA Disk:	3 GB	3 GB	15 GB	175 GB	300 GB	350 GB	300 GB	750 GB
PA CPU:	10 hr	140 hr	220 hr	100 hr	70 hr	100 hr	NA	NA
SRA:	500 MB	1 GB	4 GB	30 GB	50 GB	75 GB	100 GB	140 GB



http://www.clcbio.com/index.php?id=1479

Rapide comparatif des technologies

www.biomedcentral.com – Table

www.biomedcentral.com/1471-2164/13/341/table/T1

Quail et al. BMC Genomics 2012

Cette page est en anglais Voulez-vous la traduire ? Non Traduire Ne jamais traduire les pages rédigées en anglais

Table 1

Technical specifications of Next Generation Sequencing platforms utilised in this study

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	$128 K	$80 K**	$695 K	$256 K	$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	$502	$1000 (318 chip)	$2000	$148	$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 µg	50-1000 ng	50-1000 ng

http://www.biomedcentral.com/content/pdf/1471-2164-13-341.pdf

Rapide comparatif des technologies

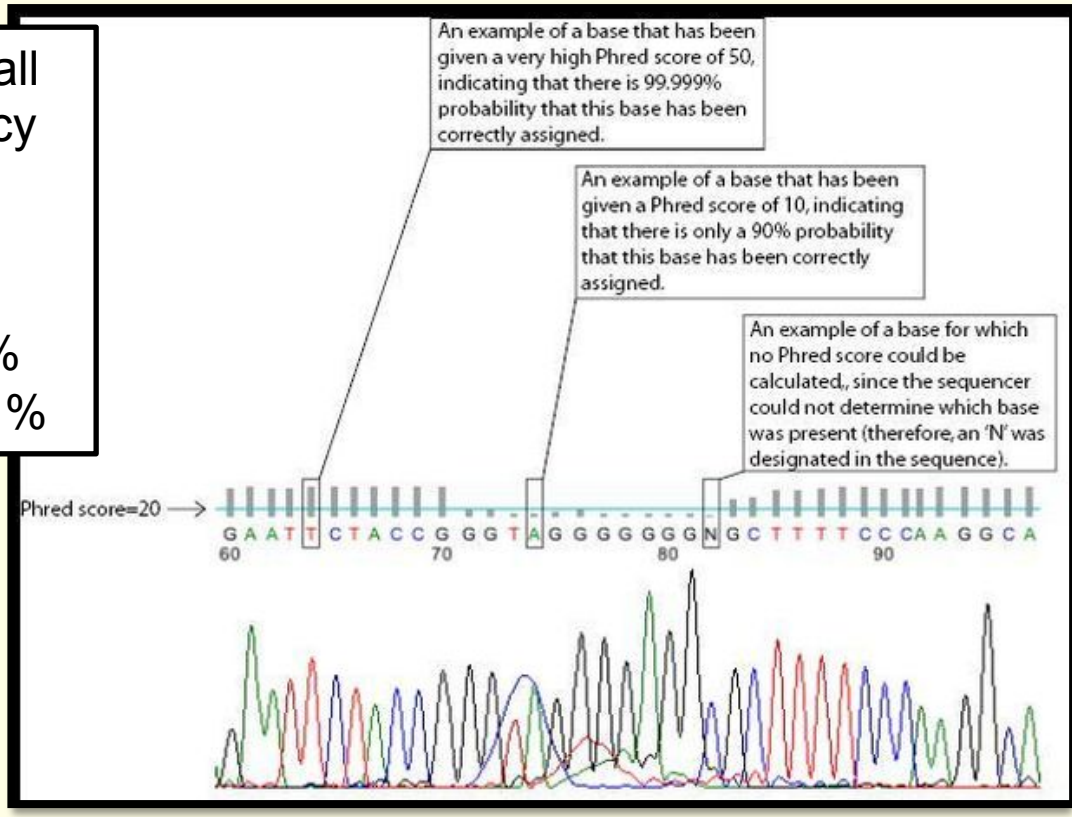
- Les comparatifs sont souvent obsolètes
- Chaque technologie a ses atouts
  - HiSeq, forte couverture reads fiables mais…. reads très courts
  - PacBio, reads très longs mais … fort taux d'erreurs
  - ⇨ Rien n'oblige à n'utiliser qu'une technologie !
- Comment choisir ?
  - En fonction de la question que vous souhaitez traiter !
  - Le prix par base ne fait pas tout
  - Mélange de science et de pression commerciale
  - Tous les prestataires ne sont pas égaux...

Discuter avec des utilisateurs/clients avant de faire un choix

Ce que l'on reçoit

- Plein de séquences avec pour chaque nucléotide de chaque séquence un indice de qualité (Q)
- Q est obtenu à partir de la probabilité P que la base soit erronée : **Q = -10 log(P)**

Q	P	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %



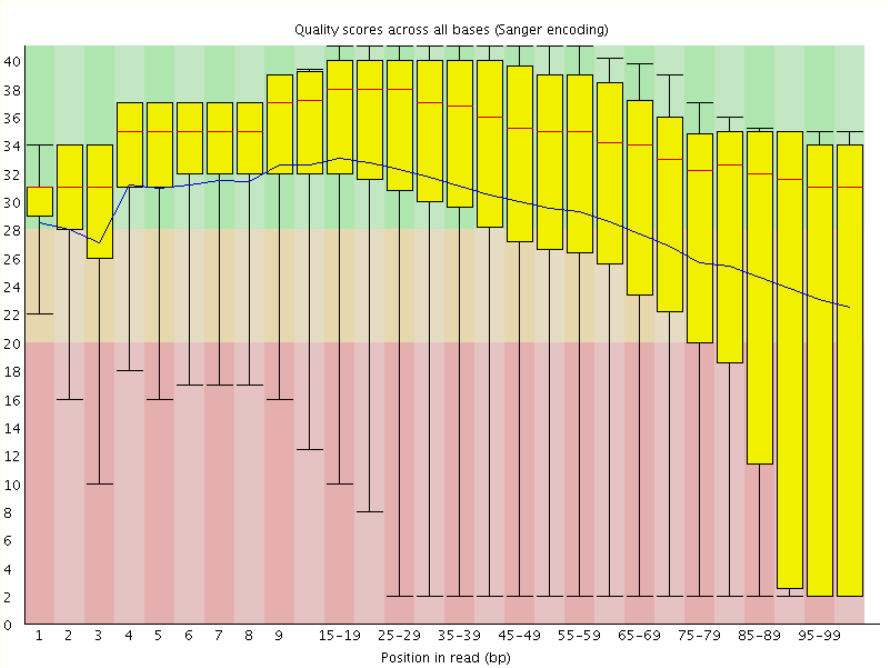
http://en.wikipedia.org/wiki/Phred_quality_score

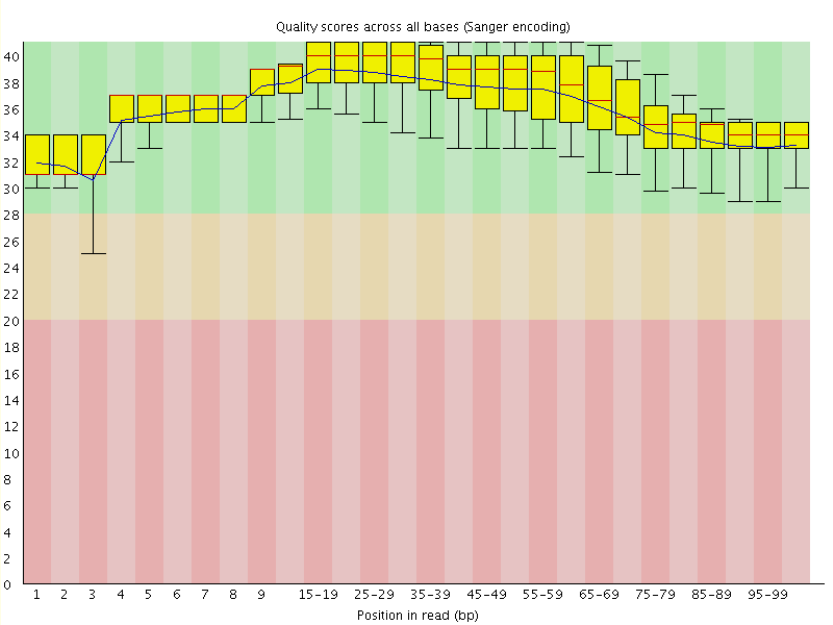
Nettoyage des reads

- Adaptateurs
  - Normalement enlevés avant livraison
- Les extrémités des reads sont souvent de moins bonne qualité
  - Elimination des bases extrêmes de faible qualité
  - Masquages des bases internes de faible qualité
- `ACCTTCTTTTCCACGTCTT  (seuil 5)`
  `23878788828889884 33`
- `  CTTCTTTNCCACGT`

Nettoyage des reads

- S'assurer de la qualité avant toute analyse
  - ⇨ FastQC
- Exemple : Riz avant et après nettoyage



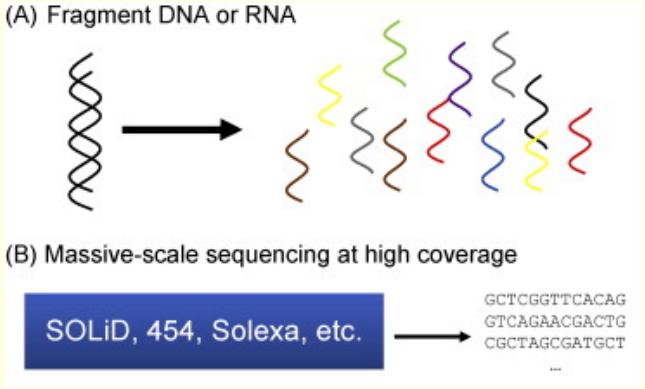


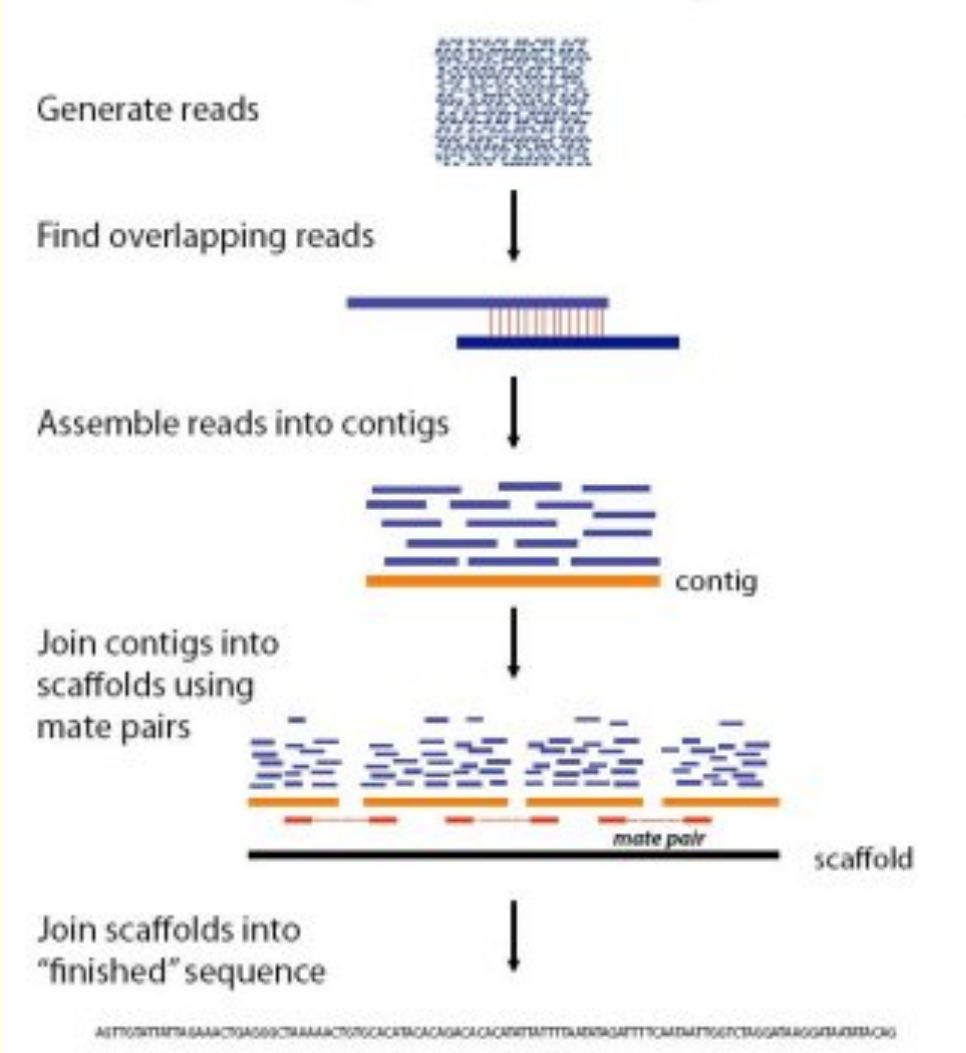
NGS : des reads aux SNPs

- Introduction NGS
- Assemblage de-novo
  - Intuition, problématique
  - Premières approches « gloutonnes »
  - Approches utilisant des Graphes
  - Apports des paired-end
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

Intuition, problématique

- Vision globale

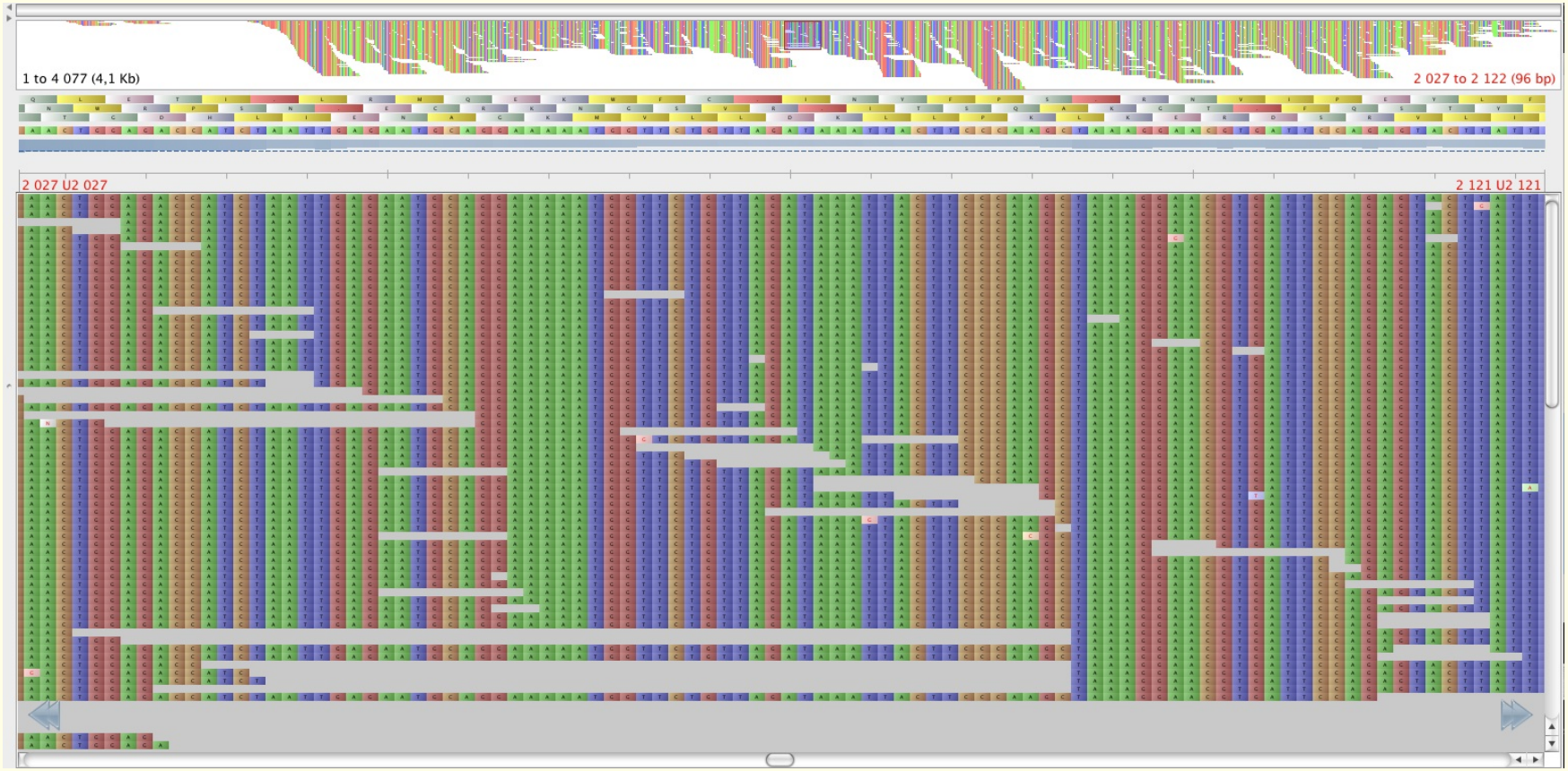




http://fhqdddddd.blog.163.com/blog/static/186991542009101904922370/

Intuition, problématique

- Vision globale



Intuition, problématique

- Supposons que l'on séquence un tout petit (fragment de) génome

  avec des reads de 10pbs on peut avoir :

  CTCCCTGTCA
  GTCATCTGTC
  ACCCTCCCTG

- Comment retrouver le génome de départ ?
- CTCCCTGTCA

  ACCCTCCCTG

  GTCATCTGTC

  CTCCCTGTCACCCTCCCTGTCATCTGTC (génome 1)

- ACCCTCCCTGTCATCTGTC (génome 2)

Premières approches gloutonnes
(e.g. SSAKE)

- Idées de base : construire le génome petit à petit en ajoutant un read après l'autre :

  `CTCCCTGTCA`

  `GTCATCTGTC`

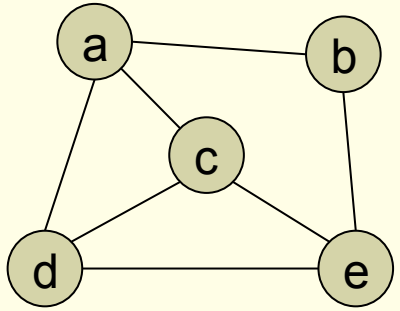
  `ACCCTCCCTG`

- Supposons que l'on parte de : `CTCCCTGTCA` (ajout?)

  `CTCCCTGTCA`

- Choisir le read (et la position) qui minimise le risque
  - ⇨ Plus grand chevauchement
  - ⇨ Construisez un exemple ou cela ne marche pas
- Choisir le read qui minimise le risque (alternatives)
  - ⇨ Plus petit ajout, plus grand rapport (chevauchement/ajout)

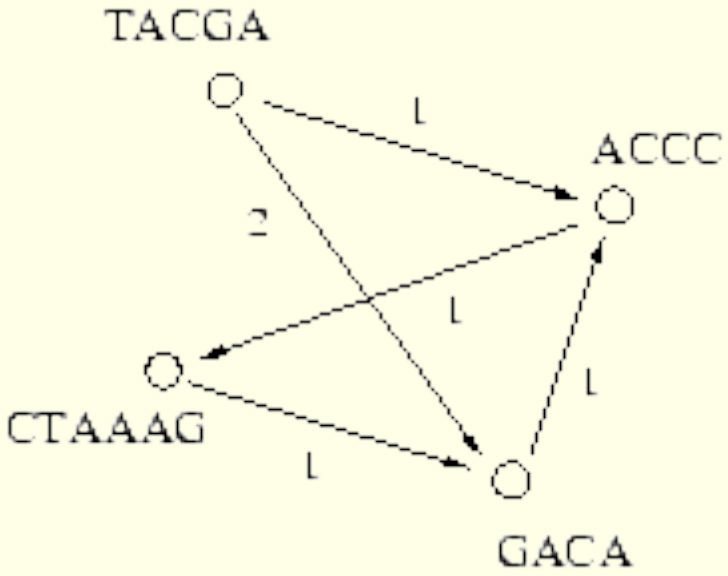
Approche par graphe de chevauchement
(e.g. CAP & Celera Assembler)



- Un graphe G= (V,E) est composé de :
  - V : un ensemble de sommets/nœuds (Vertices)
  - E : un ensemble darcs/arêtes (Edges)
- Un arc e = (u,v) est une paire de sommets
  - On peut associer des valeurs aux nœuds et aux arêtes
  - Les arêtes peuvent être ou non orientées
- Dans un graphe de chevauchement
  - Chaque séquence à assembler est un nœud
  - On met une arête entre deux nœuds si leurs séquences se chevauchent
  - Valeur d'une arête = taille du chevauchement
- Construisez le graphe de chevauchement pour E = {TACGA,ACCC,CTAAAG,GACA }

Approche par graphe de chevauchement
(e.g. CAP & Celera Assembler)

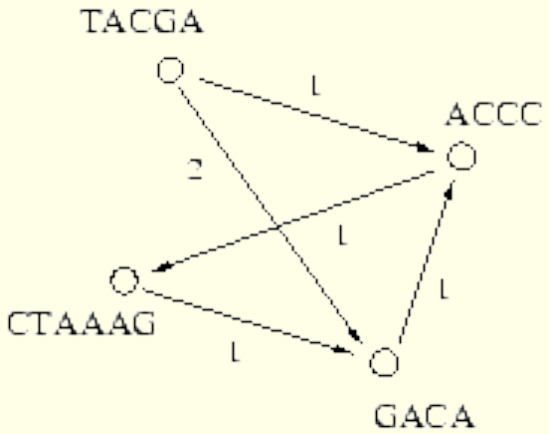
- Graphe pour E = {TACGA,ACCC,CTAAAG,GACA }
- Comment se traduit le problème d'assemblage, que faut-il chercher dans le graphe de chevauchement ?



On cherche un chemin qui passe une et une seule fois par chaque sommet et qui soit de poids maximal

Approche par graphe de chevauchement
(e.g. CAP & Celera Assembler)

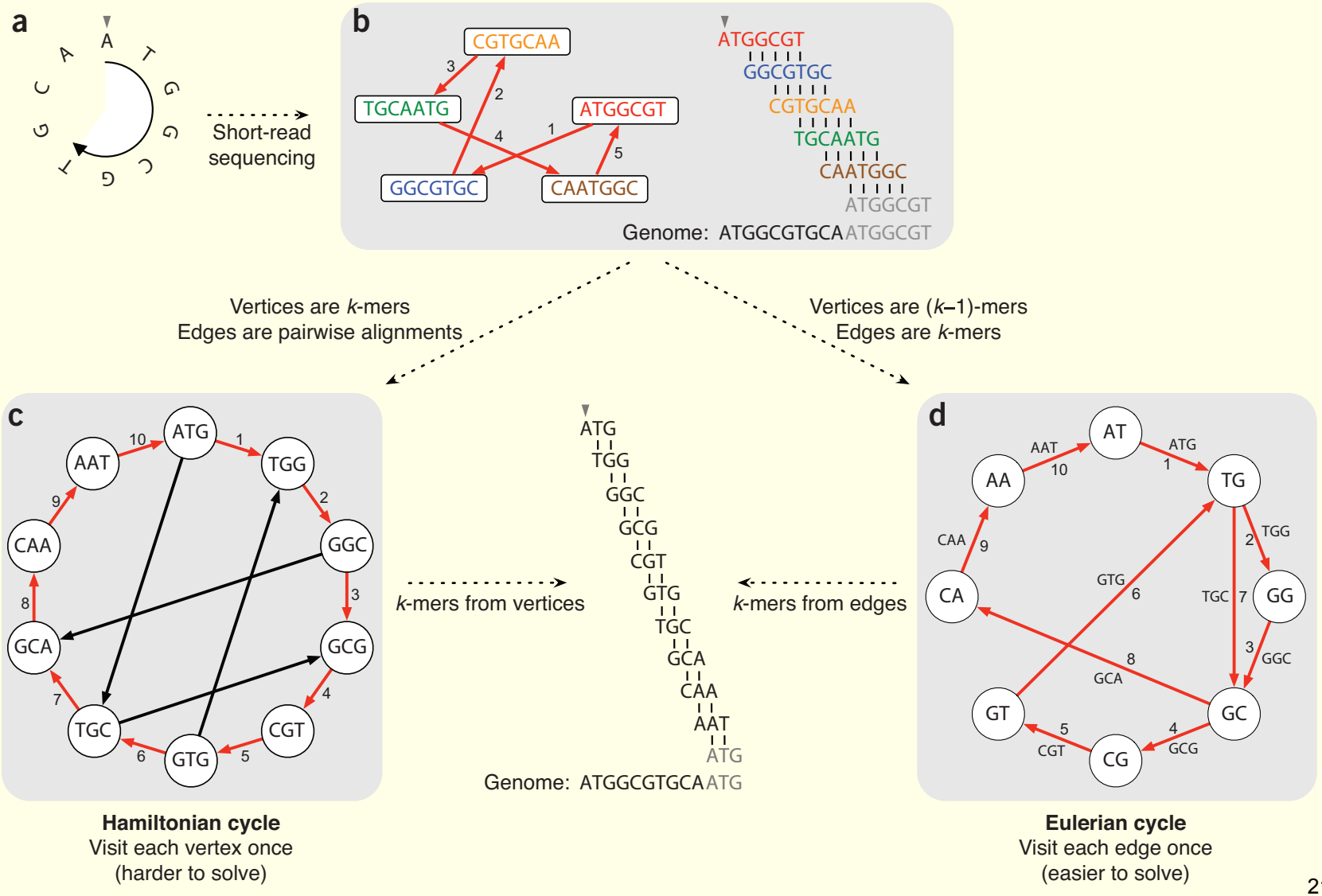
- Limites de l'approche
  - La construction du graphe est coûteuse en temps de calculs combien de comparaisons pour 10 million de reads ?
  - La recherche d'un chemin passant une et une seule fois par chaque nœud est un problème difficile



- Idée : décomposer les reads en fragments de taille k fixée à l'avance : k-mer

Approche par graphe de « De Bruijn »

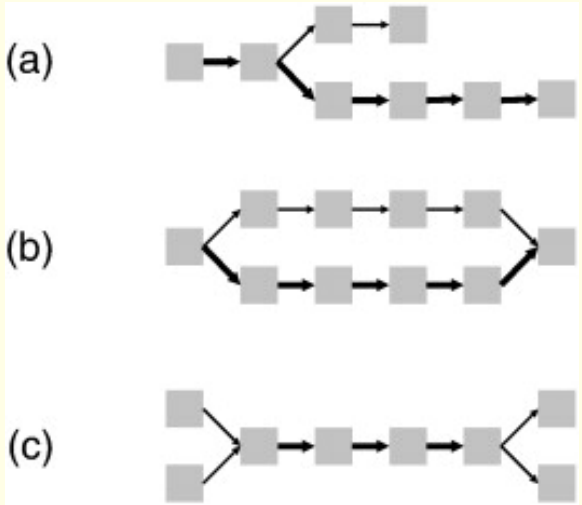
(e.g. ABySS, trinity, velvet …)

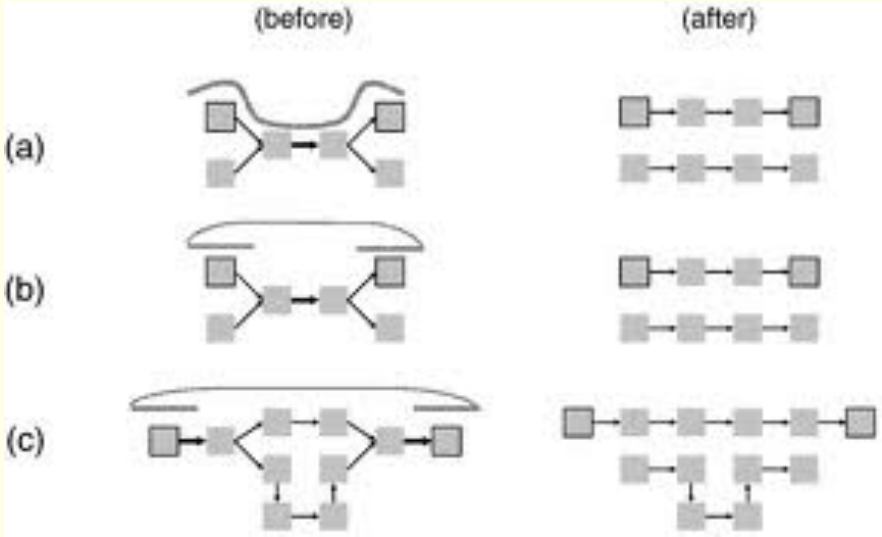


Compeau et al. Nature Biotechnologie 2011

Approche par graphe de « De Bruijn »
(e.g. ABySS, trinity, velvet …)

- Résoudre les ambiguïtés
  - Profondeur, nombre de reads
  - Paired-end
  - Matted pairs



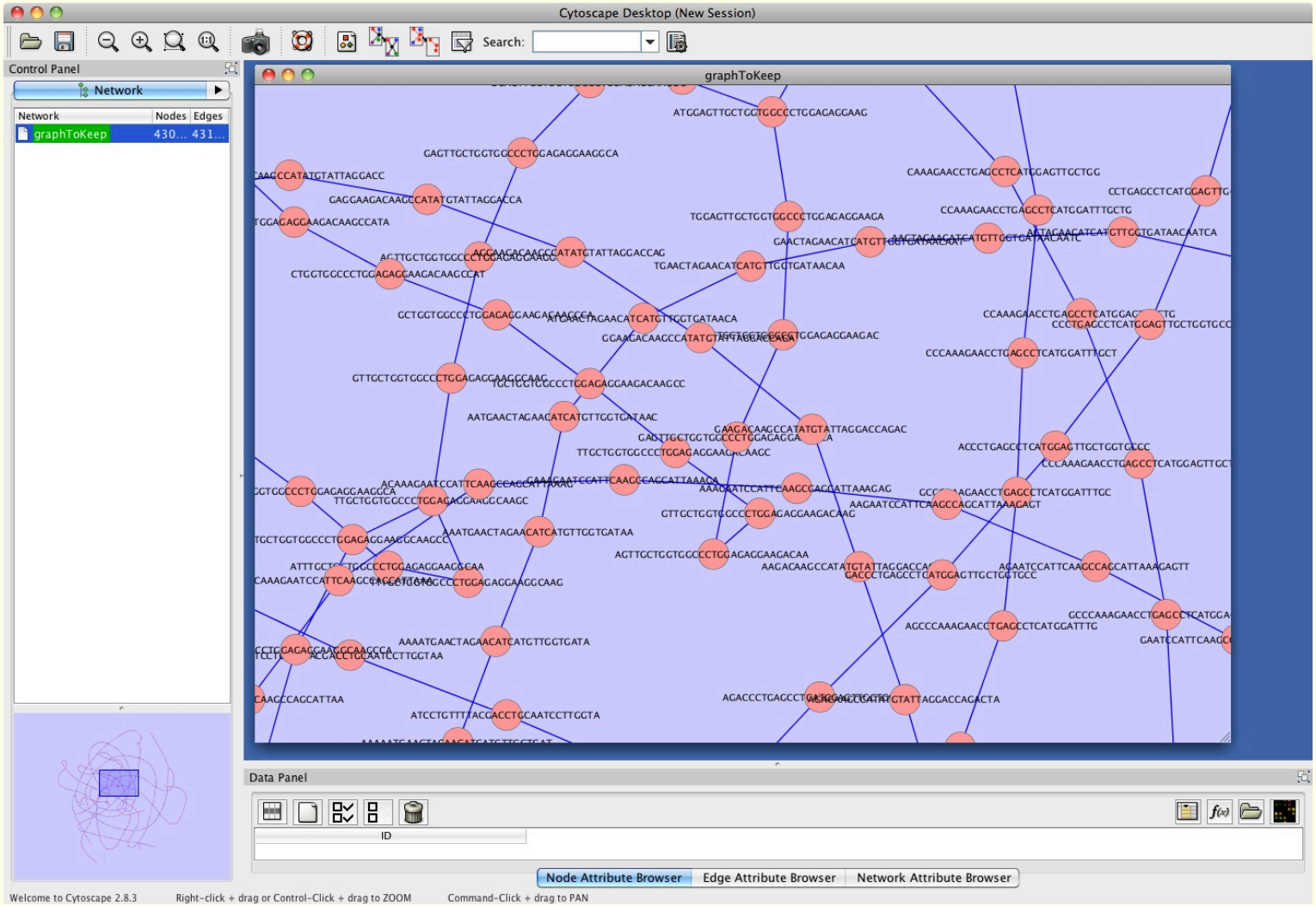


- Les erreurs de séquençage augmentent la taille du graphe
  - Essayer de corriger avant d'assembler (e.g. ABySS)

Miller et al. Genomics 2010
http://www.sciencedirect.com/science/article/pii/S0888754310000492

Approche par graphe de « De Bruijn »
(e.g. ABySS, trinity, velvet …)

- Tout petit exemple un peu plus réaliste

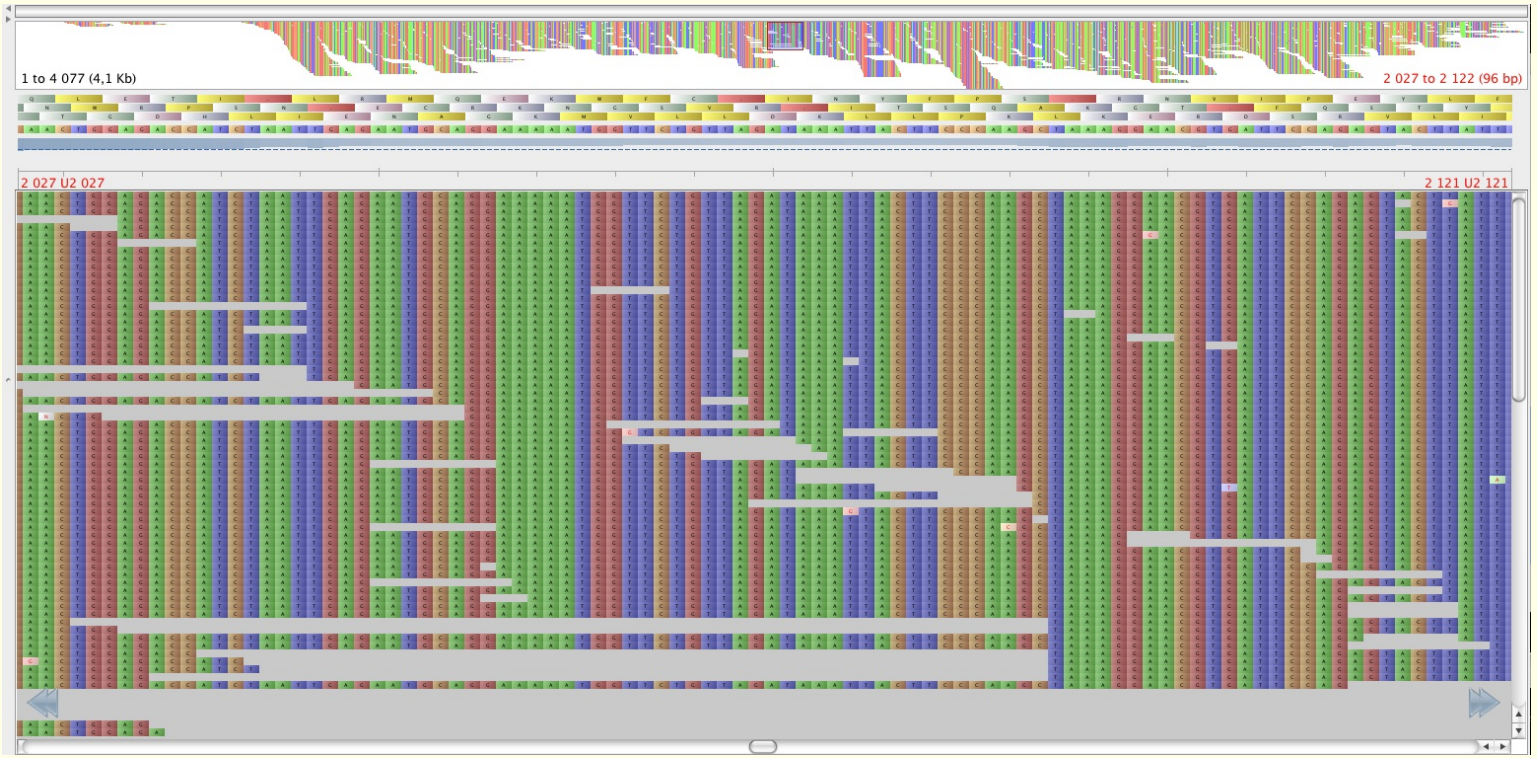


NGS : des reads aux SNPs

- Introduction NGS
- Assemblage de-novo
- Mapping sur un génome connu
  - Approches par graines (et arbre des suffixes)
  - Approches par table de suffixe
- Détection de SNP
- Conclusions et discussions

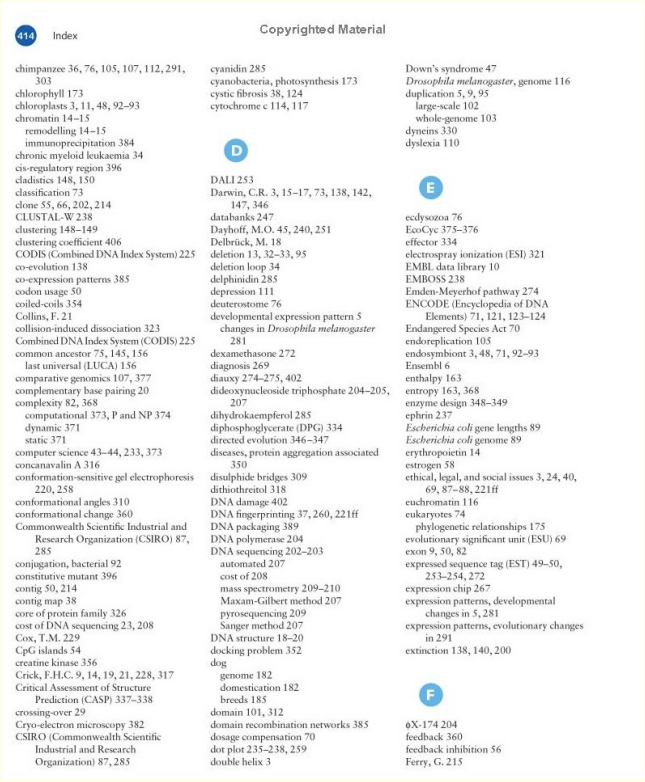
Problématique

- Problématique
  - On dispose d'un génome/transcriptome de référence
  - On veut positionner les reads dessus



Approches par graines et arbre de suffixe
(e.g. SSAHA, SOAP, BLAT)

- Pour chaque read
  - Parcourir tout le génome et tester l'égalité read/portion génome
  - ⇨ Trop looooooong
- Chercher dans un livre de 500 pages où le mot «séquence » apparaît ☹
  - ⇨ Avec un index ? ☺
- Pour chaque k-mer
  - Stocker ses position dans le génome
- Reads de 60 pb
  - Prendre des K-mer de 60 ?
  - ⇨ Lourd à stocker
  - Prendre des K-mer de 20 ?
  - ⇨ Tolérance aux erreurs
  - ⇨ Mieux, mais reste lourd à stocker

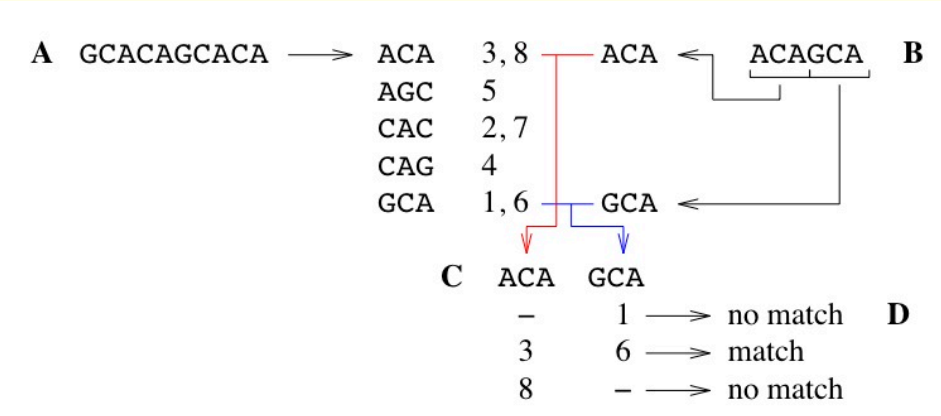


414 Index

Copyrighted Material

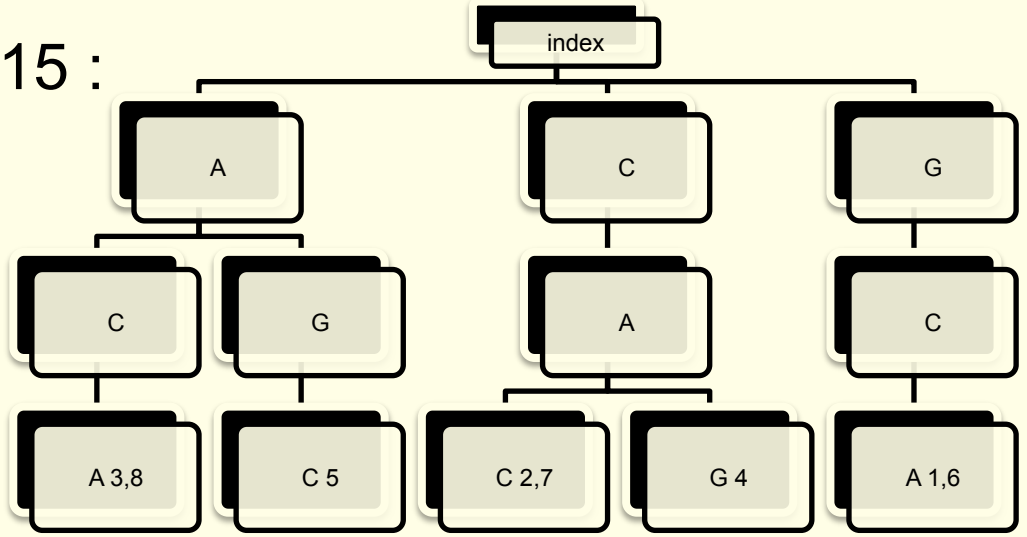
Approches par graines et arbre de suffixe
(e.g. SSAHA, SOAP, BLAT)

- Exemple
  - Génome GCACAGCACA
  - Indexation des 3-mers



- Recherche du read :
  - Décomposition en 2 3-mers
  - Recherche dans l'index de ces 3-mers
  - Identification des zones avec les deux k-mers

- Compression :
  12 lettres au lieu de 15 :
  ⇨ 20%de gain



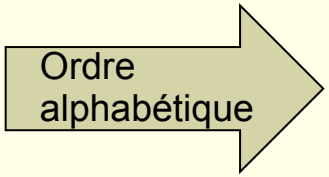
Approches par table de suffixe

- Pour chaque read
  - Parcourir tout le génome et tester l'égalité read/portion génome
  - ⇨ Trop looooooong
- Chercher dans un livre de 500 pages où le mot «séquence » apparaît ☹
- Chercher la définition de « séquence » dans un dictionnaire ☺
- Pour cela on va construire une table des suffixes
  - Lister tous les suffixes du génome
  - Trier les suffixes par ordre alphabétique
  - Ex table des suffixes de **`CATTATTAGGA $`**

Approches par table de suffixe

- Table des suffixes du génome CATTATTAGGA ?

1	CATTATTAGGA
2	ATTATTAGGA
3	TTATTAGGA
4	TATTAGGA
5	ATTAGGA
6	TTAGGA
7	TAGGA
8	AGGA
9	GGA
10	GA
11	A



1	11	A
2	8	AGGA
3	6	ATTAGGA
4	2	ATTATTAGGA
5	1	CATTATTAGGA
6	10	GA
7	9	GGA
8	7	TAGGA
9	4	TATTAGGA
10	6	TTAGGA
11	3	TTATTAGGA

- Trouver les positions des occurrences de TTA, de GGT ?
- En quoi le tri accélère t-il la recherche ?
  - ⇨ Est-ce un gain important ?
  - ⇨ Comment stocker la table en économisant l'espace ?

Transformée de Burrow Wheeler
(e.g. bwa, bowtie)

- Reprend l'idée de la table de suffixe mais en beaucoup plus compact

Format de sortie d'un mapping

- Les fichiers « SAM » (Sequence Alignment/Map)
  - Fichier au format texte très lourd
  - Id du read, qualité du mapping, position du mapping, description « CIGAR » du mapping etc.
- Format « CIGAR »
  - Description des évènements associés à un mapping
  - M (match or mutation) I (insertion) D (Délétion)

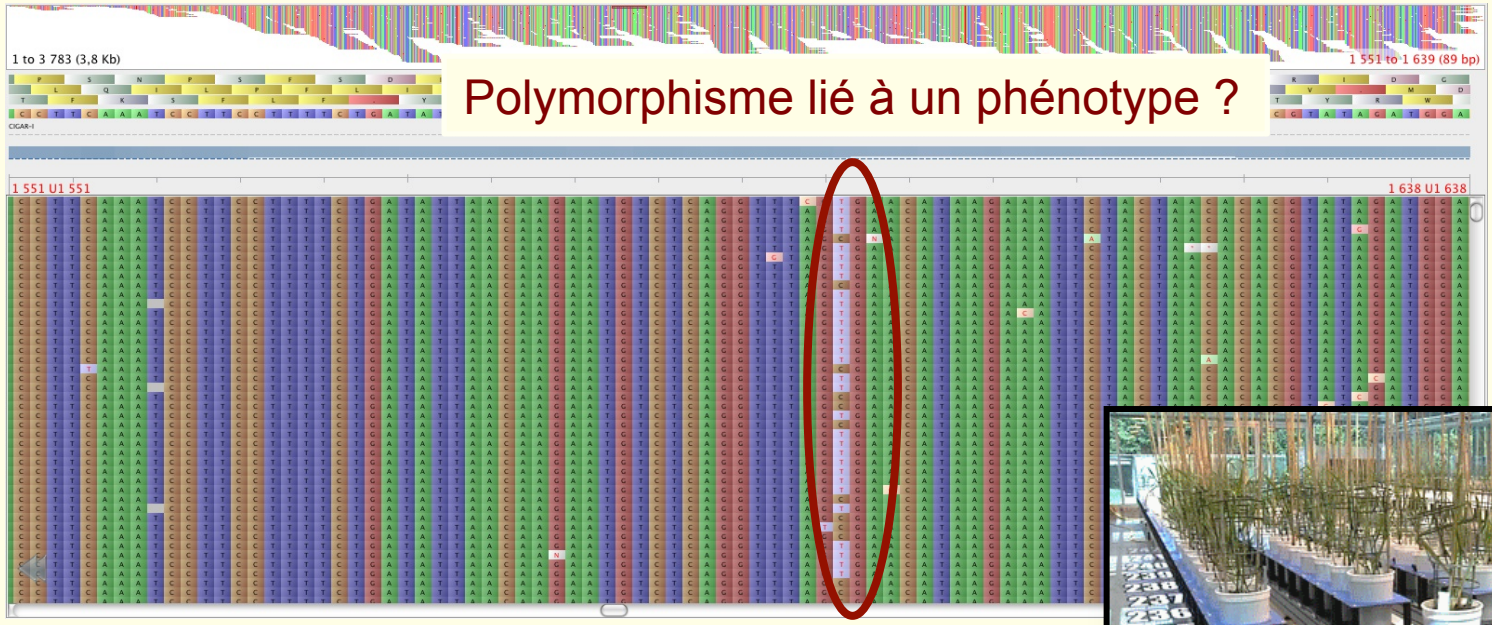
```
ref …CTTCATTACAG-TCTTTCG…
read     ATC-CAGCT

CIGAR : 3M1D3M1I1M (on décrit ce qui se passe dans le read)
```

- Les fichiers « BAM » versions compressées des SAM
  - Plus rapide et moins lourd que les fichiers SAM
  - Fichier binaire non lisible par un humain

- Introduction NGS
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
  - Approche par consensus
  - Approche probabiliste générale
  - Réglages et vérifications
- Conclusions et discussions

Problématique

- Problématique
  - On dispose d'un génome/transcriptome de référence
  - On à positionner les reads dessus
  - On veut identifier les variations/polymorphismes



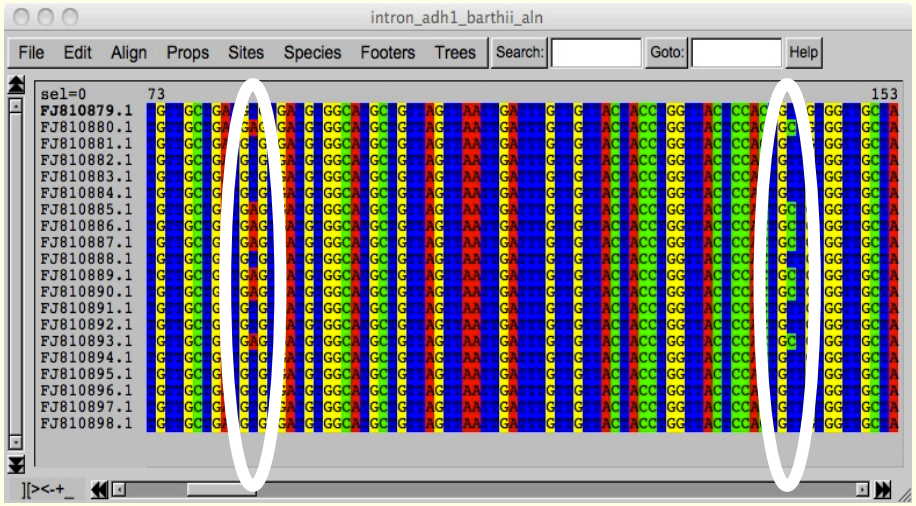
Approche par consensus
(e.g. avec dnaSP)

- Construire une séquence par individu
  - Assemblage par individu (pb de couverture ?)
  - Faire un consensus par individu après les avoir les mappés sur
    - Un génome de référence (l'idéal si possible)
    - Un assemblage fait en combinant les reads de tous les individus
  - Regarder pour chaque site si les individus diffèrent

Approche par consensus

(e.g. avec dnaSP)

- Limite de l'approche ?




```
   Indiv 1
TACAGTACGATC…
TACAGT
   AGTACGATC
   ATCAC
  CAGCACG
 ACATCA
```

```
   Indiv 2
TACAGTACGATC…
 ACAGT
   AGTACGATC
   AGTACG
   ATCA
TACAGCACG
 ACATCA
    TCAC
```

```
A   02050400100
C   00300030001
G   00003002000
T   10002000010

Indiv 1 : TACAGTACGATC…
Indiv 2 : TACAGTACGATC…

A   03060600100
C   00300040001
G   00004003000
T   10003000010
```

Approche probablisite
(e.g. GATK, samtools, reads2SNP)

- $G_{a,b}$: Génotype, ex $G_{C,C}$: homozygote C|C
- $D_{i,j}$: Données observées, fréquences des nucléotides pour l'individu i au site j
  - ⇨ $D_{i,j} = \{f_A, f_C, f_G, f_T\}$

$$P\left(G_{ab}\middle|D_{i,j}\right) = \frac{p\left(G_{ab}\right)p\left(D_{i,j}\middle|G_{ab}\right)}{p\left(D_{i,j}\right)}$$

$$P\left(D_{i,j}\right) = \sum_a \sum_b p\left(G_{ab}\right)p\left(D_{i,j}\middle|G_{ab}\right)$$ a,b = A,C,G ou T

$$P\left(G_{ab}\middle|D_{i,j}\right) = \frac{p\left(G_{ab}\right)p\left(D_{i,j}\middle|G_{ab}\right)}{\sum_a \sum_b p\left(G_{ab}\right)p\left(D_{i,j}\middle|G_{ab}\right)}$$

Approche probablisite
(e.g. GATK, samtools, reads2SNP)

Paramètres de la vraisemblance

- *Hardy- Weinberg*

$$p\left(G_{ab}\right) = 2 p_a p_b \text{ si a} \neq b$$
$$p\left(G_{ab}\right) = p_a^2 \text{ si } a = b$$

p_a, p_b : fréquences au site *j (tous individus confondus)*

$$p\left(D_{i,j} \middle| G_{ab}\right) = \begin{cases} P_{Multinomial}\left\{1-3\varepsilon, \varepsilon, \varepsilon, \varepsilon\right\} \text{ si } a = b = \text{A} \\ P_{Multinomial}\left\{\frac{1}{2}-\varepsilon, \frac{1}{2}-\varepsilon, \varepsilon, \varepsilon\right\} \text{ si } a = \text{A et } b = \text{C} \\ \text{etc...} \end{cases}$$

Approche probablisite

(e.g. GATK, samtools, reads2SNP)

- $G_{a,b}$: Génotype, ex $G_{C,C}$: homozygote C|C
- $D_{i,j}$: Données observées, fréquences des nucléotides pour l'individu i au site j
  - ⇨ $D_{i,j} = \{f_A, f_C, f_G, f_T\}$
- *Probabilité postérieure et « SNP calling »*

Attendu (stat, risque d'erreurs)

$$P\left(G_{ab} \middle| D_{i,j}\right) = \frac{p\left(G_{ab}\right) p\left(D_{i,j} \middle| G_{ab}\right)}{\sum_a \sum_b p\left(G_{ab}\right) p\left(D_{i,j} \middle| G_{ab}\right)}$$

Attendu (génet des pops)

Réglages et vérifications
(e.g. GATK, samtools, reads2SNP)

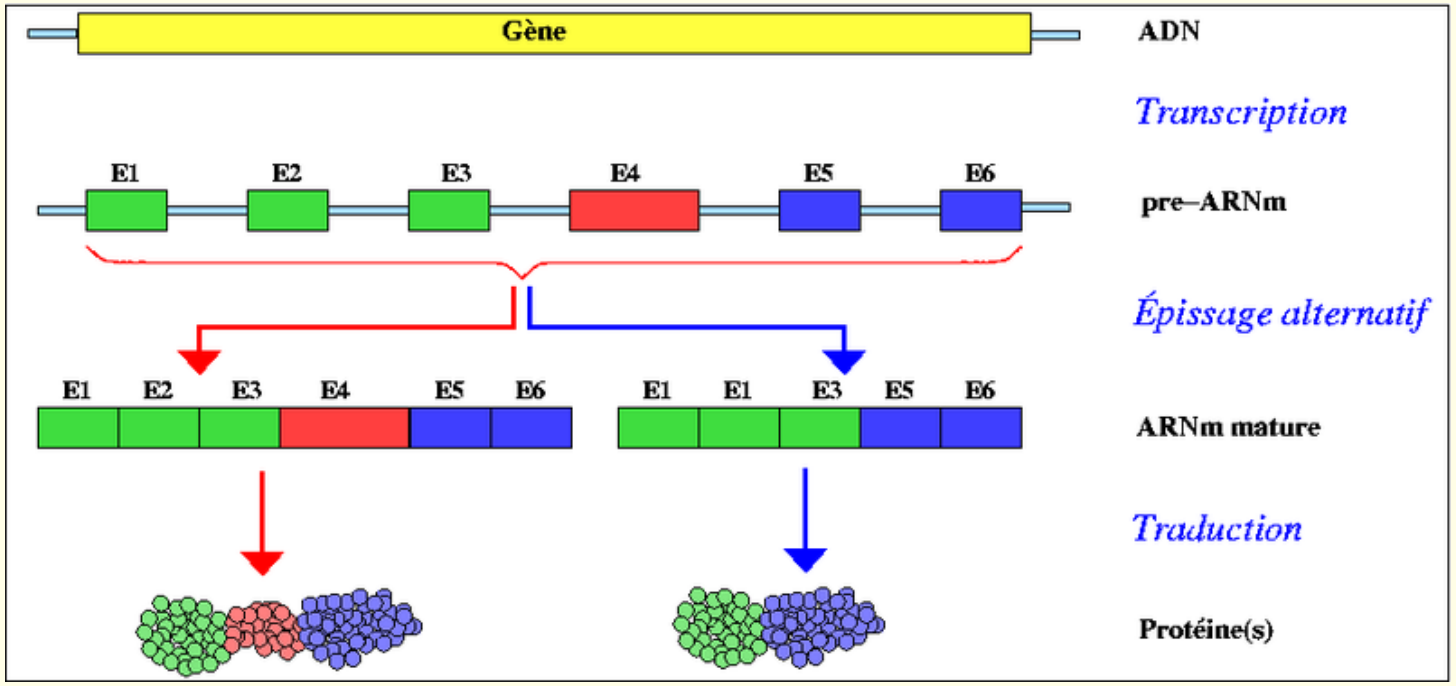
- Les fichiers vcf (variant call format)
  - Qualité du SNP, vraisemblance de chaque génotype etc.
- Fichiers bcf (binary call format)
  - Versions binaires compressées
- Qualité individuelle
  - Garder les SNP ayant une qualité > seuil (e.g. vcftools)
  - Garder les SNP isolés (3 SNP sur 10 sites consécutifs…???)
  - Etc.
- Qualité globale (comparaison aux attendus)
  - Ratio de SNP synonyme/non-synonyme (dN/dS << 1)
  - Ratio de SNP transition/transversion ~2
  - Etc.
- Calibration : utilisation d'un jeux de SNP connus (GATK)

NGS : des reads aux SNPs

- Introduction NGS
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

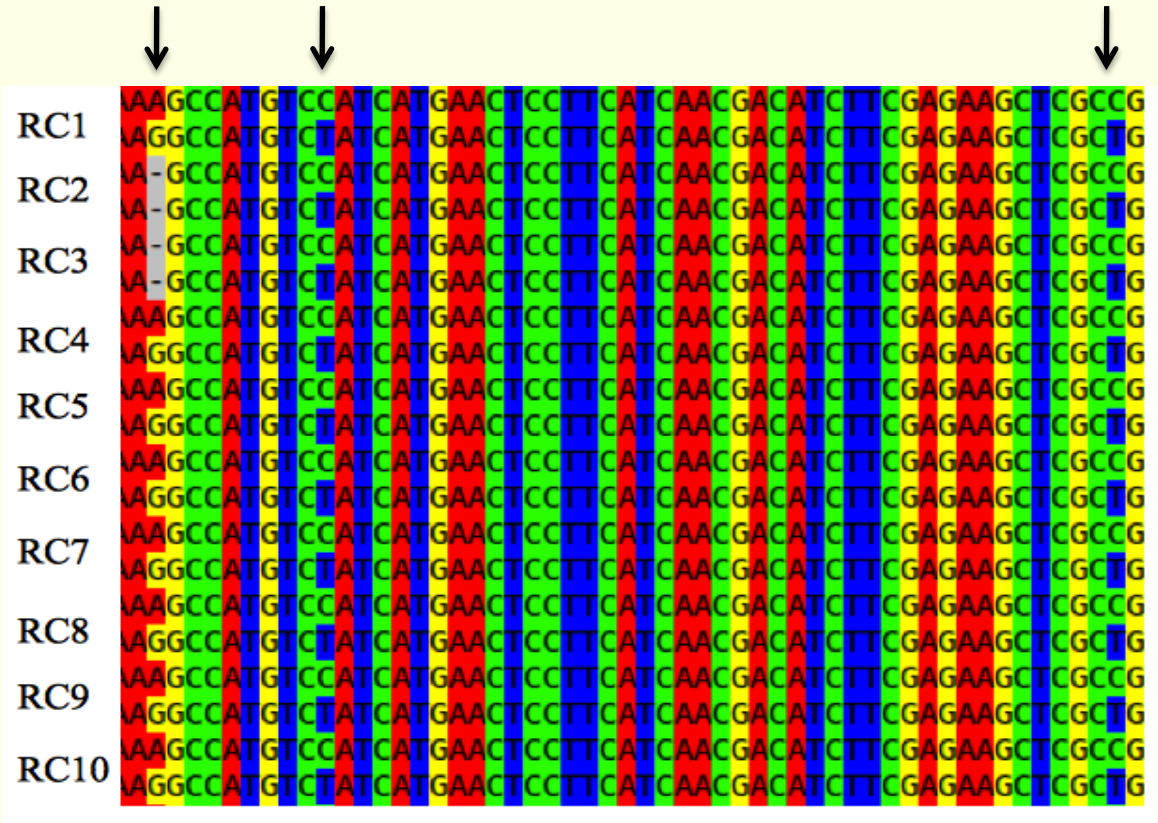
Quelques problèmes …

- Génomes de plantes
  - Beaucoup d'éléments répétés
  - ⇨ Complique l'assemblage et le mapping
- Transcriptomes plus simples ?



http://fr.wikipedia.org/wiki/%C3%89pissage

Quelques problèmes …



[0/8/0/40]
[0/7/0/27]
[0/5/0/30]
[0/10/0/35]
[0/2/0/10]
…

Exemple: Riz africain cultivé : très fortement autogame
A quoi cela peut-il être dû ?

Avant un projet NGS ?

- Séquencer pour quoi faire ?
  - Clarifier la question à laquelle on veut répondre
- Séquencer qui ?
  - Choix des individus, du type de tissu, de la date de prélèvement
  - Garder les individus vivants ou au moins de l'ADN (re-séquençage)
- Séquencer quoi ?
  - Génome ? Transcriptome ? Les deux ?
- Séquencer combien ?
  - Nombre d'individus => puissance statistique
  - Nombre de reads par individu => fiabilité, chevauchement
  - Choix de la technologie ...
- Evaluer les ressources nécessaires
  - CPU, stockage et … temps humain (CDD, partenaires, sous-traiter)

projet NGS ?

Ne pas avoir peur de
demander conseil,
demander de l'aide…