# Premiers résultats obtenus sur le PacBio RSII de la plateforme GET PLAGE

## *Focus sur l'assemblage 100% PacBio de génomes « pas simples »*

Jérôme Gouzy *et al.*

# Nos modèles: une bactérie, un champignon, une plante

- **Xanthomonas: bactérie pathogène de plantes; génome d'environ 5Mb; 1 chromosome avec ou sans micro plasmides; ~65 %GC; plusieurs opérons ribosomiques; plusieurs dizaines de copies de séquences d'insertions de plusieurs familles; des gènes très intéressants de plusieurs kb, composés de répétitions en tandem presque exactes.**

- **Microbotryum: champignon pathogène de plante; génome d'environ 30Mb; ~55 %GC; des chromosomes sexuels de plusieurs mégabases, très intéressants d'un point de vue évolutif … mais principalement composés de séquences répétées imbriquées.**

- **Tournesol: plante; génome d'environ 3.6Gb; génotype quasi homozygote; 6 ans de gros efforts du consortium pour obtenir un assemblage 454/illumina … largement perfectible ➔ le génome est largement composé d'éléments transposables très conservés et ça c'est un très gros problème !**

# Objectifs

1. **On veut obtenir 100 équivalent génome avec les dernières chimies PacBio:**
   **convaincus par les résultats de Sergey Koren qui depuis début 2014 « contigue » les bras chromosomiques sur Arabidopsis et la Drosophile https://sites.google.com/site/sergekoren/**
   - **1 SMRT cell pour Xanthomonas (quelques centaines d'euros)**
   - **~5 SMRT cells pour Microbotryum (quelques milliers d'€)**
   - **~500 SMRT cells pour le tournesol (plusieurs centaines de milliers d'€ à financer!)**

2. **Pour envisager de traiter les problèmes les plus complexes on veut construire une banque avec les inserts les plus longs possibles** ***(a modérer si l'on s'attend a de petits chromosomes/plasmides que l'on pourrait perdre au sizing)***

   **Cf présentation Gerrit sur les problématiques de qualité d'ADN et de maîtrise du protocole de construction des librairies**

3. **Maîtrise du protocole d'assemblage**
   - **Très facile pour Xanthomonas et Microbotryum: smartanalysis de PacBio (web/cli) ou PBcR(cli)**
   - **Beaucoup plus compliqué pour le Tournesol où il faut gérer les temps de calcul et les problèmes de stockage (ex: comparer 2a2 beaucoup de séquences répétées pose « quelques » soucis pratiques)**

# Assemblage de génomes 100% PacBio - Préambule

- **Les erreurs contenues dans les lectures PacBio:**
  - Présentent un taux variable selon la polymérase utilisée
  - Ont un taux d'erreur de l'ordre d'environ 16% pour le protocole P6/C4
  - Bien plus d'INDEL que de mismatch
  - Sont aléatoires (très bonne nouvelle!)

➔**Les lectures doivent d'abord être corrigées avant d'être assemblées**

- Avec des lectures courtes illumina
  - ➔pas si facile que cela de faire un mapping spécifique sur des données si bruitées surtout lorsque le génome présente des répétitions légèrement divergentes
- Avec des lectures longues PacBio
  - ➔Nécessité de développer des programmes de mapping pour aligner des séquences aussi distantes et avec un tel modèle d'erreur

➔**Le résultat de l'assemblage doit également être corrigé**

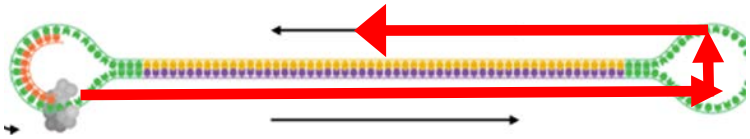
# Principe général de l'assemblage 100 % PacBio

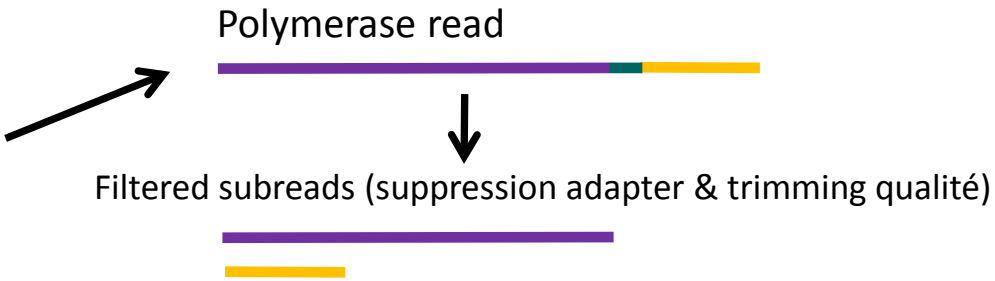
- 2 pipelines très proches conceptuellement et techniquement
  - Les lectures sont corrigées avant d'être assemblées par WGS(CABOG). Les consensus sont corrigés avec quiver de PacBio
  - Les différences se situent dans les paramétrages par défaut et les versions

| | HGAP 3 (PacBio=PB) | PBcR (Koren *et al.*) |
|---|---|---|
| **Correction des lectures** | | |
| Alignement | PB/BLASR | MHAP (Berlin et al.) ou PB/BLASR |
| Correction | PB/dagcon | PBcR (PB/falconcns\|PB/dagcon) |
| **Contigage** | | |
| Overlap | CA/overlap | CA/overlap |
| Layout | CA/unitigger | CA/unitigger (bogart) |
| Consensus | CA/utgcns | CA/utgcns (pbutgcns) |
| **Correction de l'assemblage** | | |
| Polishing | PB/Quiver | PB/Quiver |

# Xanthomonas (5Mb) 1/2

- **Echantillon:** Laurent Noel (LIPM) & Baptiste Mayjonade (LIPM, PIA SUNRISE)
- **Données**



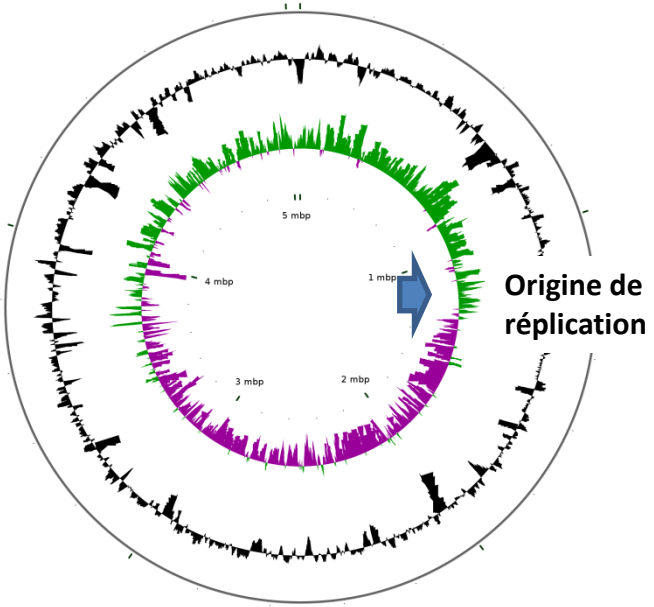



| « filtered_subreads » P6/C4 | # | MIN | MAX | N50 BP | N50 NUM | MEAN | BP |
|---|---|---|---|---|---|---|---|
| Xantho 0.1 nM | 81 112 | 50 | 42 598 | 9 599 | 24 690 | 7 886 | 639Mb |
| Xantho 0.2 nM | 113 545 | 50 | 47 821 | 9 662 | 33 755 | 7 821 | 888Mb |

- **Assemblage (Xantho 0.1nM):** Gérald Salin (GET PLAGE), **smrtanalysis HGAP3 (2h30)**
- **2 contigs tels qu'attendus: le chromosome de 5Mb et un plasmide d'environ 80kb**
  - **Avec le protocole P6/C4, la longueur des lectures permet de traverser sans problème les répétitions de plusieurs kb que l'on trouve chez les bactéries (ex: opérons ribosomiques)**

# Xanthomonas (5Mb) 2/2

- Le « GC skew » du chromosome étant également tel qu'attendu




- L'assemblage est parfait, nous cherchons maintenant à améliorer l'étape de correctior du consensus sur « les gènes très intéressant de plusieurs kb, composés de répétitions en tandem presque exactes. »

Sébastien Carrere & Laurent Noel (LIPM)

Avec le protocole standard de correction du consensus on ne corrige pas toujours très bien les strechs d'homopolymers ➔ frameshifts dans les gènes « très intéressants »

# Microbotryum (30Mb)

- **Echantillon: Stephanie Le-prieur, Hélène Badouin, Tatiana Giraud (ESE U-PSUD/CNRS)**
- **Actuellement seules 3 SMRT cells ont été produites pour 675Mo (22x)**
  - **Faible rendement dû à un souci de robotique lors du run**
  - **+ Les statistiques concernant la taille des lectures sont très correctes lorsqu'on les compare à un (très bon) fournisseur de données PacBio**

| « filtered_subreads » P6/C4 | # | MIN | MAX | N50 BP | N50 NUM | MEAN | BP |
|---|---|---|---|---|---|---|---|
| IGM | 101553 | 50 | 44500 | 14626 | 27313 | 10329 | 1Gb |
| GET PLAGE | 89420 | 50 | 41634 | 10870 | 23804 | 7550 | 675Mb |

- **Avec seulement 22x, on est loin des (très) bons résultats obtenus par ailleurs (180x, chimie P5) mais on a déjà des métriques correctes …. pour 3h de bioinformatique (… vs plusieurs mois !)**

| « filtered_subreads » P6/C4 | # | MIN | MAX | N50 BP | N50 NUM | MEAN | BP |
|---|---|---|---|---|---|---|---|
| 20x 454 Broad (A1) | 550 | 1000 | 867 920 | 194 032 | 41 | 46651 | 25,6Mb 3,7%N |
| Illumina PE, MP3,8, »20 » Fontanillas *et al.* 2015 (A1+A2) | 401 | 1001 | 1,358Mb | 313 520 | 25 | 68 987 | 27,6Mb 5%N |
| GET PLAGE 22x (PBcR paramètres « low cov ») | 401 | 6673 | 607 534 | 135 514 | 68 | 75 399 | 30Mb 0%N |

# Tournesol (3,6 Gb) 1/2

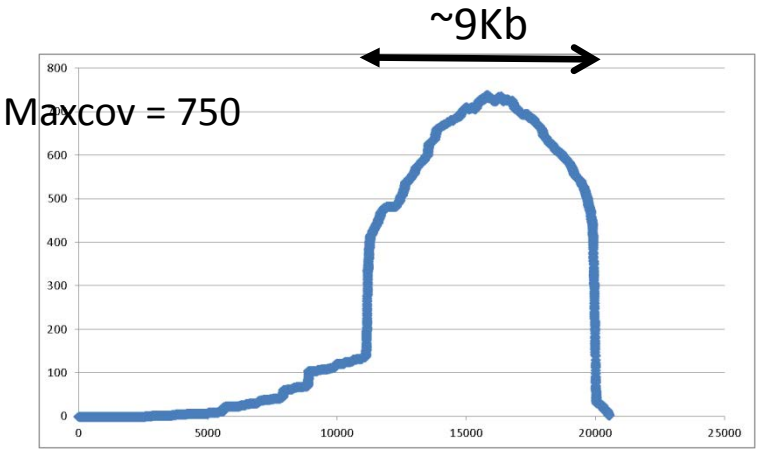
- **Financement PIA SUNRISE:**
  - **Coordinateur:** Nicolas Langlade (LIPM)
  - **WP Génomique:** Stéphane Muños & Baptiste Mayjonade (LIPM); Hélène Bergès *et al.* (CNRGV); **WP Bioinformatique:** Jérôme Gouzy *et al.* (LIPM)

- **Echantillon et optimisation de la construction des banques et du chargement:** Baptiste Mayjonade **(SUNRISE/LIPM, détaché GET Plage pour 6 mois) conseillé par** Gerrit Kuhn **(Pacbio)**

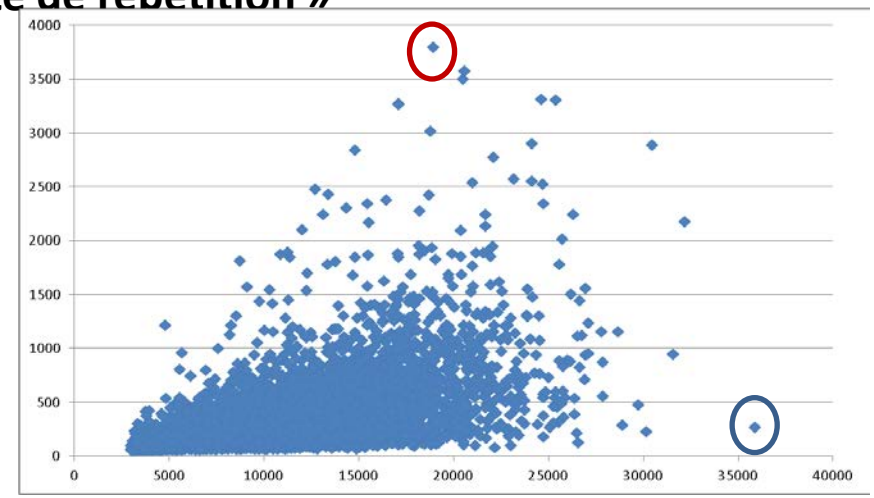
| | # | MIN | MAX | N50 BP | N50 NUM | MEAN | BP |
|---|---|---|---|---|---|---|---|
| Moyenne sur 39 smrt cells IGM | 111698 | 50 | 46606 | 12321 | 32 211 | 9 263 | 1Gb |
| GET Plage: Training | 65287 | 50 | 51820 | 12567 | 18 062 | 9 407 | 614Mb |
| GET Plage: Quelques jours d'optimisation plus tard | 89345 | 50 | 51311 | 15185 | 23 175 | 10 415 | 930Mb |

# Tournesol (3,6 Gb) 2/2

- **Pourquoi il très important d'essayer d'obtenir des lectures les plus longues possibles ?**
  - **Mapping de 1x de données sur 2x de lectures longues (>= 20Kb)**
  - **Analyse de la couverture des lectures « longues » en ne considérant que les hits de plus de 3kb**
  - **Recherche d'un pattern de type « unité de répétition »**




Exemple de pattern recherché




X=longueur de « l'unité de répétition » (max 36kb)
Y=maxcov de chaque unité (max=3750)

- **Avec seulement 18x de données et 2jours de calcul (PBcR 8.3rc1) sur la PF Bioinformatique Genotoul on a déjà des métriques équivalentes à nos assemblages illumina/454: #contigs: 87346 (3kb->318Kb); N50: 19,9Kb; #BP: 1,5Gb**

# Conclusions du moment

- **Sous la condition d'avoir des données de bonne qualité et en quantité suffisante (100x), on va revenir à l'ère des vrais génomes de référence**
  - **L'outil smrtanalysis est très facile d'utilisation et permet d'assembler en routine des petits génomes.**
  - **L'assemblage de génomes nécessitant moins d'une dizaine de SMRT cells ne pose aucun problème bioinformatique**
- **L'assemblage d'un très gros génome très répété comme le tournesol**
  - **Pose quelques problèmes de stockage et d'adaptation ou de paramétrage des pipelines existants mais cela devrait être réglé d'ici l'obtention des 100x (= d'ici cet été)**
  - **La longueur des lectures obtenues sur la PF GET Plage nous laisse très optimiste sur notre capacité à assembler le génome**
- **Une fois optimisé et validé sur le tournesol, Baptiste Mayjonade publiera son protocole d'extraction d'ADN pour la construction de banques long inserts.**

# Perspectives (en génomique)

- Une fois les protocoles bioinformatiques pour tous les types de génomes homozygotes en place, il faudra valider ou mettre au point l'assemblage de génomes hétérozygotes ➔ évaluer PB/Falcon
- Evaluer les outils proposés par Gene Meyers
- Exploiter les caractéristiques de la technologie PacBio pour générer en même temps un génome et un épigénome

# Biblio

- HGAP
  - Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013 Jun;10(6):563-9. Chin CS[1], Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. (surtout les pages 21-37 de la suppl note 1 pour les méthodes DAGcon & quiver)
- PBcR:
  - Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology 14:R101 (2013). Sergey Koren, Gregory P Harhay, Timothy P. Smith, James L. Bono, Dayna M Harhay, D Scott McVey, Diana Radune, Nicholas H. Begman, Adam M. Phillippy (Attention: données PacBio chimie C2)
  - https://sites.google.com/site/sergekoren/
- MHAP
  - Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James Drake, Jane M Landolin, Adam M Phillippy
  - http://biorxiv.org/content/early/2014/08/14/008003
    - résultats chimie P5/C3 sur la drosophile, arabidopsis, l'humain
    - Comparaison avec Moleculo/illumina sur la capacité a retrouver les repeats de la droso
- Pour les problématiques d'assemblage, attention a bien prendre en compte le cocktail enzyme/chimie lorsque l'on interprète : P4-C2 (ok pour les bactéries sans trop de repeats), P5-C3 (ok pour les génomes <=120Mb), P6-C4 (cela devient jouable pour les très gros génomes)

# Intuitions sur les algorithmes

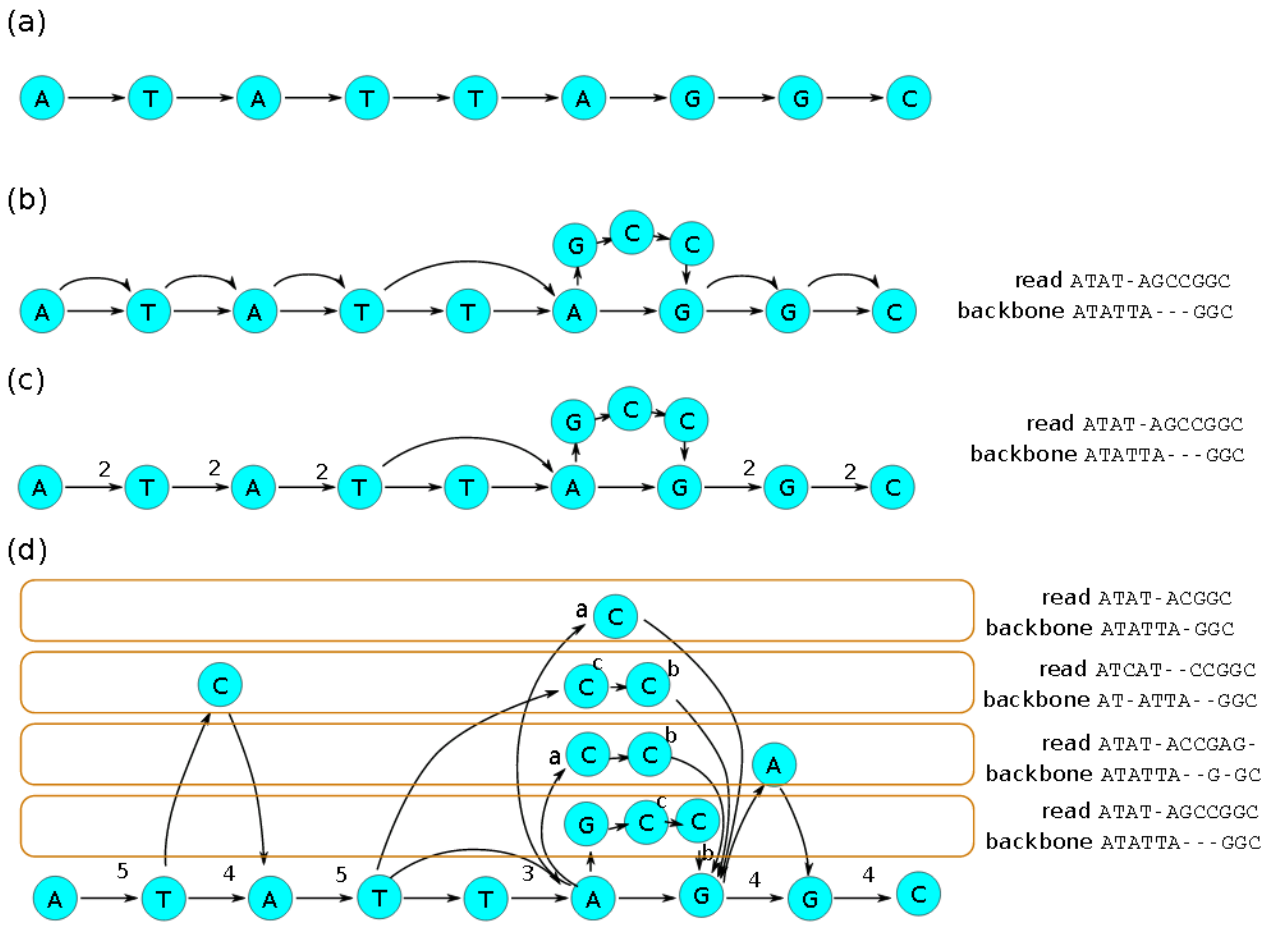
- MHAP: alignement 2a2 de séquences bruitées
- DAGcon: correction des lectures
- Quiver: édition du consensus

# Biblio

- **HGAP**
  - **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013 Jun;10(6):563-9. Chin CS[1], Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. (surtout les pages 21-37 de la suppl note 1 pour les méthodes DAGcon & quiver)**
- **PBcR:**
  - **Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology 14:R101 (2013). Sergey Koren, Gregory P Harhay, Timothy P. Smith, James L. Bono, Dayna M Harhay, D Scott McVey, Diana Radune, Nicholas H. Begman, Adam M. Phillippy (Attention: données PacBio chimie C2)**
  - **https://sites.google.com/site/sergekoren/**
- **MHAP**
  - **Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James Drake, Jane M Landolin, Adam M Phillippy**
  - **http://biorxiv.org/content/early/2014/08/14/008003**
    - **résultats chimie P5/C3 sur la drosophile, arabidopsis, l'humain**
    - **Comparaison avec Moleculo/illumina sur la capacité a retrouver les repeats de la droso**
- **Pour les problématiques d'assemblage, attention a bien prendre en compte le cocktail enzyme/chimie lorsque l'on interprète : P4-C2 (ok pour les bactéries sans trop de repeats), P5-C3 (ok pour les génomes <=120Mb), P6-C4 (cela devient jouable pour les très gros génomes)**

# PB/DAGcon (1/2)
## Chin *et al.* Nat. Meth. 2013, suppl note 1

- **DAGCon: A Directed Acyclic Graph Based Consensus Algorithm**
- **Basé sur une séquence de référence**
- **Merging rules: on merge les noeuds qui ont le même label et qui ont les mêmes arêtes entrantes et sortantes**



# PB/DAGcon (2/2)
## Chin *et al.* Nat. Meth. 2013, suppl note 1

**Creation du Consensus**

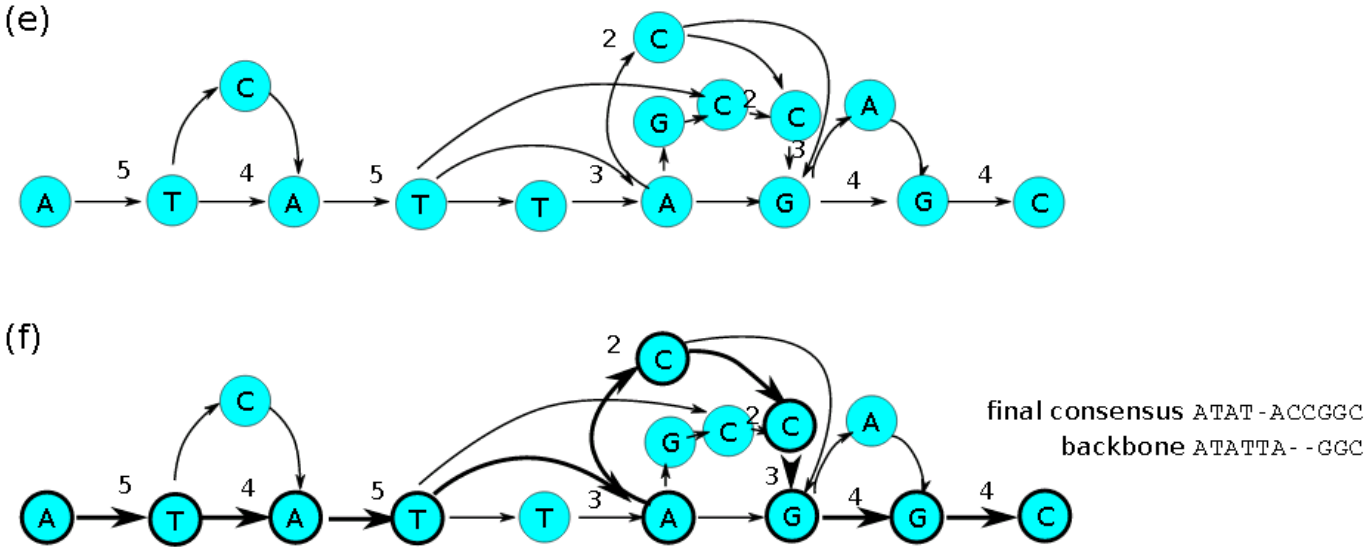
- Un score a chaque noeud
- Chemin qui maximise la somme des scores

Score: dépend des poids sur les arêtes sortantes

Score > 0 si une arete représente plus de la moitié de la couverture locale (poids max autour du noeud)

Score < 0 dans le cas contraire

Consensus = chemin de score maximum (programmation dynamique et backtracking)

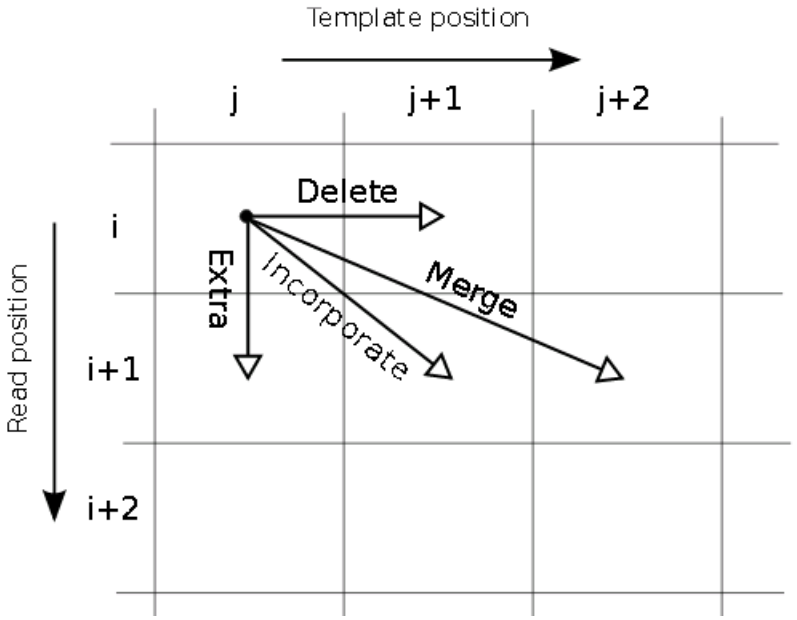


# Quiver (1/2)
## Chin *et al.* Nat. Meth. 2013, suppl note 1

- Given a vector of reads R from a single (unknown) template T, Quiver uses a greedy algorithm to maximize the likelihood Pr(R | T) for the unknown T. We develop a likelihood function Pr(R | T) which encodes the sequencing error model and is specific to a particular sequencing chemistry and enzyme. The parameters within the model are derived using a training step that learns an error model from SMRT sequencing data on a known template.
- For a long reference, we process the consensus with tiling windows across the reference to limit the amount of memory used.

# Quiver (2/2)
## Chin *et al.* Nat. Meth. 2013, suppl note 1

- On identifie les reads correspondants
  a une fenêtre, on crée un consensus, on l'édite d'un nucléotide tant que la vraissemblance augmente
- Pour calculer la vraissemblance on
  **aligne** les lectures contre le template
- Le modèle d'erreur intègre
  - la qualité des bases
  - les "pulse metrics"
    - InsertionQV
    - SubstitutionQV
    - DeletionQV
    - MergeQV
- « Merge » améliore la correction des homopolymers



# smrtanalysis/current/analysis/lib/python2.7/GenomicConsensus/quiver/resources/2014-09/GenomicConsensus/QuiverParameters.ini

```
[P5-C3.AllQVsMergingByChannelModel]
Match             =   0.184656435394
Mismatch          =   -0.380508126527
MismatchS         =   -0.0519773778309
Branch            =   -0.0178687456208
BranchS           =   -0.0865415022309
DeletionN         =   -0.928673177809
DeletionWithTag   =   -0.255381037375
DeletionWithTagS  =   0.0173271990056
Nce               =    0.303359662376
NceS              =   -0.0980869366241
Merge_A           =   -0.0402618414395
Merge_C           =    0.445432915183
Merge_G           =    0.256569746054
Merge_T           =    0.353800996389
MergeS_A          =   -0.118145654186
MergeS_C          =   -0.0471922787923
MergeS_G          =   -0.032653869882
MergeS_T          =   -0.0596571606945
```

```
[P6-C4.AllQVsMergingByChannelModel]
Match                = 0.262756
Mismatch             = -1.71623
MismatchS            = -0.00961684
Branch               = -0.400811
BranchS              = -0.0577744
DeletionN            = -1.39515
DeletionWithTag      = -0.232547
DeletionWithTagS     = -0.0235445
Nce                  = -0.237657
NceS                 = -0.0459215
Merge_A              = -1.13237
Merge_C              = 1.08894
Merge_G              = 0.570111
Merge_T              = -0.570049
MergeS_A             = -4.03641e-05
MergeS_C             = -0.107432
MergeS_G             = -0.0801512
MergeS_T             = -0.058112
```

# MHAP (1/3) Berlin *et al.*

- **Problématique: on veut identifier des overlaps entre deux séquences très bruitées**
  - **BlasR: Chaisson and Tesler BMC Bioinformatics 2012**
    - **On cherche une régions qui partage un certain nombre d'« ancres » (ex: 10) sans erreur (ex: 15mers)**
    - **On aligne proprement <u>les meilleurs candidats</u> (! Pb pour les génomes très répétés)**

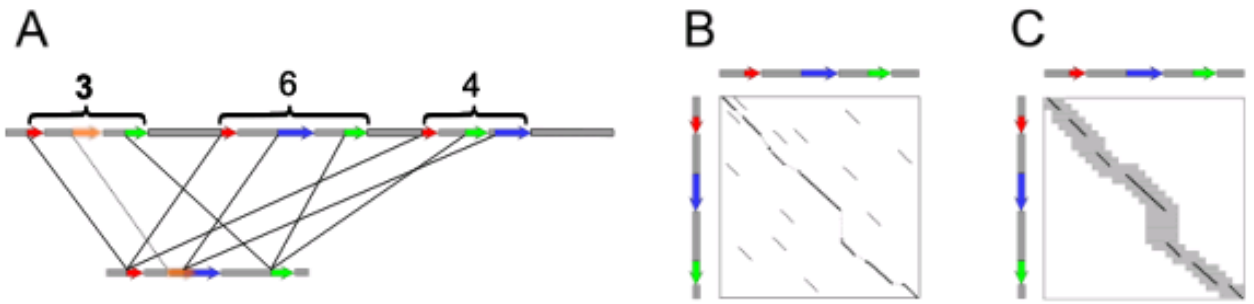
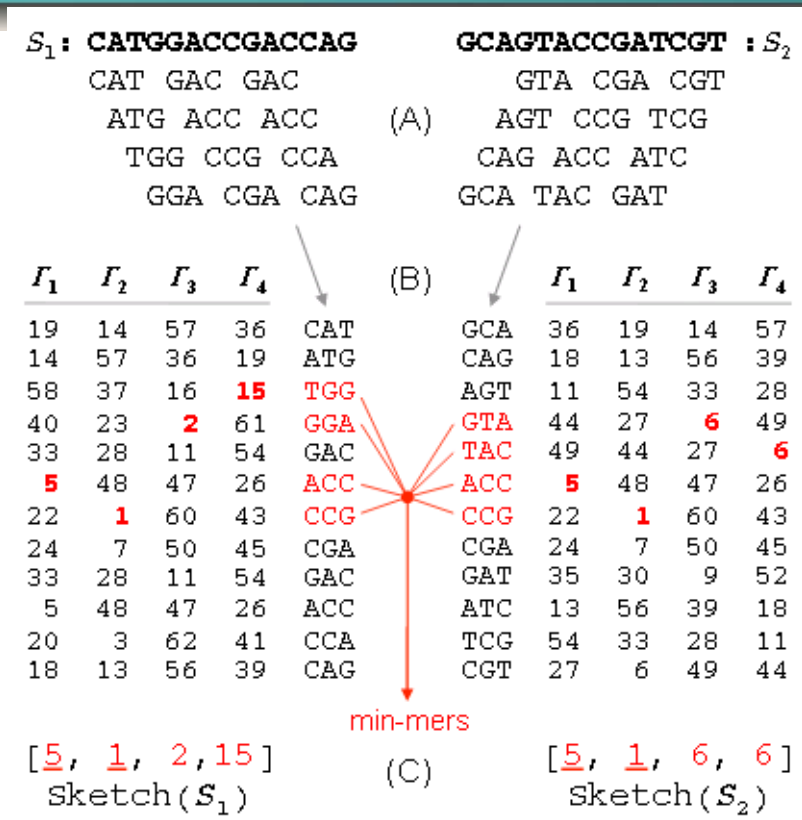



**Figure 9 Overview of the BLASR method.** (**A**) Candidate intervals are found by mapping short, exact matches as shown by colored arrows. Either a suffix array or BWT-FM index of the genome are used to find the exact matches. Intervals are defined over clusters of matches and are ranked; intervals with score 3, 6, and 4 are shown. (**B**) Matches scoring above a threshold are aligned using sparse dynamic programming on shorter exact matches. (**C**) Alignments that have a high-scoring sparse-dynamic programming score are realigned by dynamic programming over a subset of cells defined using the sparse dynamic programming alignment as a guide.

# MHAP (2/3) Berlin *et al.*

- MHAP beaucoup plus rapide que BlasR (600x plus rapide sur des données de Droso)
- Basé sur MinHash:
  - Utilisé pour comparer des pages web, des images, etc.
  - Réduit un texte a un petit ensemble de fingerprints appelé sketch (croquis)
- Sktechs de 1000 16-mers peuvent être utilisés pour détecter un chevauchement de 5Kb dans des lectures de 10Kb avec un taux d'erreur de 30%
- Le taux d'erreur de l'alignement et a peu prés additif du taux d'erreur des 2 lectures ➔ c'est plus facile de mapper sur une référence propre que de calculer un chevauchement entre deux lectures bruitées

# MHAP (3/3) Berlin *et al.*




Pour chaque séquence

-on fait la liste des kmers

-on encode en entier chaque kmer selon plusieurs de hash (ici 4)

-pour chaque fonction de hash on ne garde que la valeur min

-on calcule l'indice de similarité Jaccard entre les deux vecteurs de valeurs conservées (card(intersection)/card(union))

- on filtre

- localisation des min-mers partagés sur les séquences originales

- calcul de l'offset de l'overlap sur S1 et S2 *(and the median difference in their positions is computed to determine the overlap offset (0) for S1and S2)*