

Correction de données de séquençage de troisième génération

Pierre MORISSE

Directeur de thèse : Thierry LECROQ

Co-encadrant : Arnaud LEFEBVRE

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

26 septembre 2019





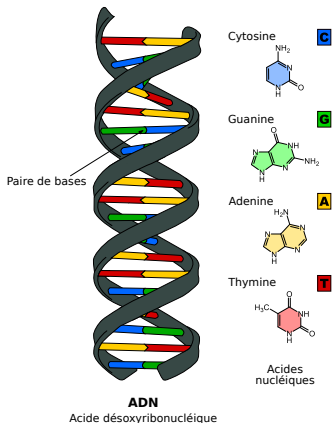
Plan

- 1 Contexte
- 2 Correction de *reads* longs
- 3 Évaluation des méthodes existantes : ELECTOR
- 4 Correction hybride : HG-CoLoR
- 5 Auto-correction : CONSENT
- 6 Conclusion



- 1 **Contexte**
- 2 Correction de *reads* longs
- 3 Évaluation des méthodes existantes : ELECTOR
- 4 Correction hybride : HG-CoLoR
- 5 Auto-correction : CONSENT
- 6 Conclusion

L'ADN



- Stocke l'information biologique
- Partie variable : composée de 4 acides nucléiques
- Certaines régions forment des gènes
- Organisé en plusieurs structures : chromosomes



L'ADN

- Ensemble du matériel génétique des chromosomes : génome
- Plusieurs types d'organismes : haploïdes et polyploïdes
- Contenu des génomes structuré : répétitions
- Génomes : tailles variables

Génome	Organisme	Taille
<i>Escherichia coli</i>	Bactérie	4,6 Mbp
<i>Saccharomyces cerevisiae</i>	Levure	13 Mbp
<i>Caenorhabditis elegans</i>	Nématode	100 Mbp
<i>Drosophila melanogaster</i>	Drosophile	180 Mbp
<i>Homo sapiens</i>	Humain	3 300 Mbp



Le séquençage de l'ADN

- Mis au point en 1977
- Retranscrire l'information contenue dans un échantillon d'ADN sous forme de chaînes de caractères sur l'alphabet ADN
 $\Sigma = \{A, C, G, T\}$
- Ces chaînes de caractères sont appelées *reads*
- Représentent des fragments de la séquence originale d'ADN
- Utilisés pour résoudre divers problèmes en biologie



Première génération

- 1977 : apparition des premières technologies de séquençage
- Petites quantités de *reads* de 500 à 1 000 bases
- Séquençage long (< 70 k bases / heure)
- Taux d'erreurs négligeables (0,001% à 0,1%)
- Adoptées par les biologistes pendant près de 30 ans



Deuxième génération

- 2005 : deuxième génération de technologies de séquençage
- Séquençage haut débit (NGS)
- Millions de *reads* courts (100 - 300 bases)
- Séquençage plus rapide (20 M bases / heure)
- Faibles taux d'erreurs ($\sim 1\%$)



Troisième génération

- 2010 : troisième génération de technologies de séquençage
- Deux principaux acteurs : Pacific Biosciences (PacBio) et Oxford Nanopore Technologies (ONT)
- *Reads* plus longs (milliers à centaines de milliers de bases)
- Taux d'erreurs très importants (de 10 à plus de 30%)
- Prometteurs pour l'assemblage de génomes longs et complexes



Alignement

- Recherche de similarités entre séquences (par exemple, entre les *reads* et un génome de référence)
- Mettre en évidence des variations, des mutations

A	T	T	G	-	A
A	T	-	C	C	A



Alignement multiple

- Peut être généralisé à un ensemble de séquences
- On parle alors d'alignement multiple
- Utilisé pour définir une séquence consensus à partir d'un ensemble de séquences

A	T	A	G	-	A
A	T	A	C	C	A
T	A	-	G	C	A
T	T	-	C	C	A
A	T	-	G	C	T



Alignement multiple

- Peut être généralisé à un ensemble de séquences
- On parle alors d'alignement multiple
- Utilisé pour définir une séquence consensus à partir d'un ensemble de séquences

A	T	A	G	-	A
A	T	A	C	C	A
T	A	-	G	C	A
T	T	-	C	C	A
A	T	-	G	C	T

A T - G C A



Alignement multiple

- Peut être généralisé à un ensemble de séquences
- On parle alors d'alignement multiple
- Utilisé pour définir une séquence consensus à partir d'un ensemble de séquences

A	T	A	G	-	A
A	T	A	C	C	A
T	A	-	G	C	A
T	T	-	C	C	A
A	T	-	G	C	T

A T - G C A



Alignement multiple

- Peut être généralisé à un ensemble de séquences
- On parle alors d'alignement multiple
- Utilisé pour définir une séquence consensus à partir d'un ensemble de séquences

A	T	A	G	-	A
A	T	A	C	C	A
T	A	-	G	C	A
T	T	-	C	C	A
A	T	-	G	C	T
A	T	-	G	C	A



Assemblage

- Génomes de référence pas disponibles pour toutes les espèces
- *Reads* plus courts que les chromosomes originaux
- Reconstruire les chromosomes du génome, en recherchant des chevauchements entre les *reads*



Assemblage

```

                A T A G C A
          C T G G A T   G C A C C T
T A T C T G                               C T C A T G

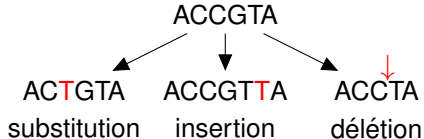
T A T C T G G A T A G C A C C T C A T G
  
```

- Séquence produite à partir de plusieurs *reads* : contig
- But de l'assemblage : produire un contig par chromosome



Correction

- *Reads* imparfaits : présence d'erreurs de séquençage





Correction

- Ces erreurs compliquent les analyses
- Nécessaire de les corriger
- Cœur des travaux de ma thèse



- 1 Contexte
- 2 Correction de *reads* longs**
- 3 Évaluation des méthodes existantes : ELECTOR
- 4 Correction hybride : HG-CoLoR
- 5 Auto-correction : CONSENT
- 6 Conclusion



Intérêt des *reads* longs

- Présence de répétitions dans les génomes
- *Reads* courts : plus courts que les répétitions
- *Reads* longs : couvrent les répétitions

Exemple





Intérêt des *reads* longs

- Présence de répétitions dans les génomes
- *Reads* courts : plus courts que les répétitions
- *Reads* longs : couvrent les répétitions

Exemple





Intérêt des *reads* longs

- Présence de répétitions dans les génomes
- *Reads* courts : plus courts que les répétitions
- *Reads* longs : couvrent les répétitions

Exemple





Intérêt des *reads* longs

- Présence de répétitions dans les génomes
- *Reads* courts : plus courts que les répétitions
- *Reads* longs : couvrent les répétitions

Exemple





Intérêt des *reads* longs

- Présence de répétitions dans les génomes
- *Reads* courts : plus courts que les répétitions
- *Reads* longs : couvrent les répétitions

Exemple





Problématique

- Nombreuses méthodes de correction de *reads* courts
- Pas applicables aux *reads* longs
- État de l'art, deux principales approches :
 - Correction hybride (*reads* longs + *reads* courts)
 - Auto-correction (*reads* longs uniquement)



Correction hybride

- *Reads* courts + *reads* longs séquencés pour le même individu
- Utilise l'information des *reads* courts pour corriger les *reads* longs
- Quatre approches :
 - 1 Alignement de *reads* courts
 - 2 Alignement de contigs
 - 3 Utilisation de graphes de de Bruijn
 - 4 Utilisation de modèles de Markov cachés



Alignement de *reads* courts

- Première approche de correction hybride
- Définition d'alignements multiples à partir des *reads* courts
- Utilisation des alignements multiples pour calculer une séquence consensus



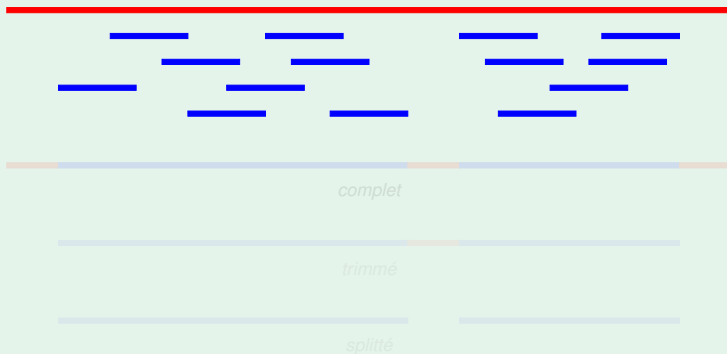
Alignement de *reads* courts

Exemple



Alignement de *reads* courts

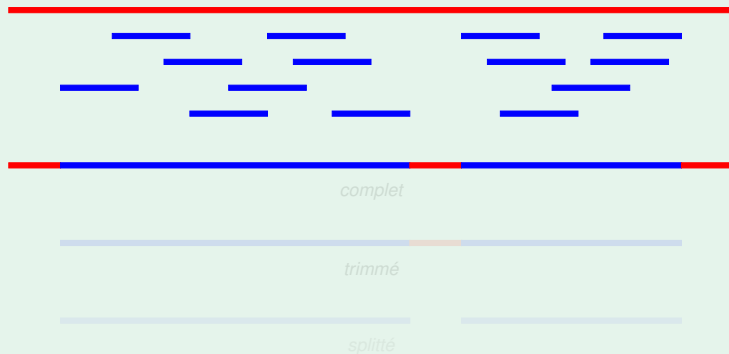
Exemple





Alignement de *reads* courts

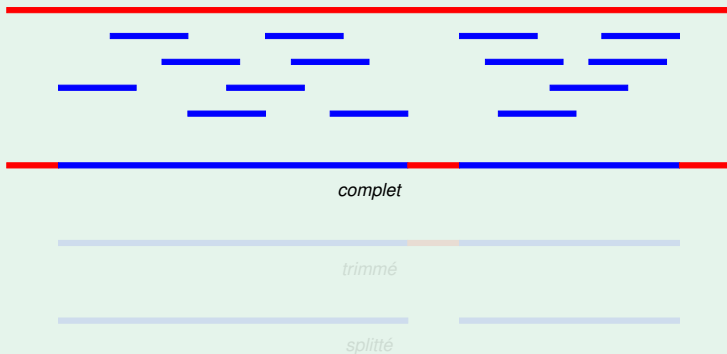
Exemple





Alignement de *reads* courts

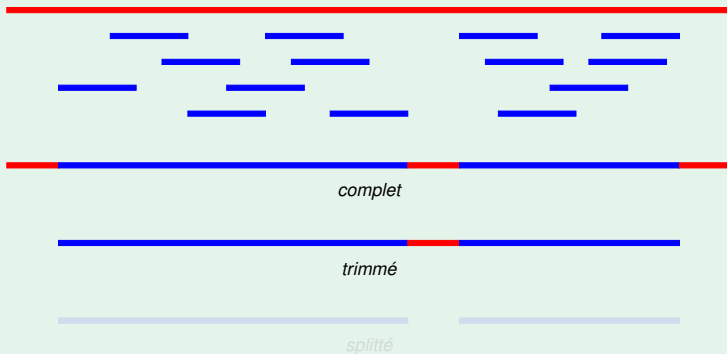
Exemple





Alignement de *reads* courts

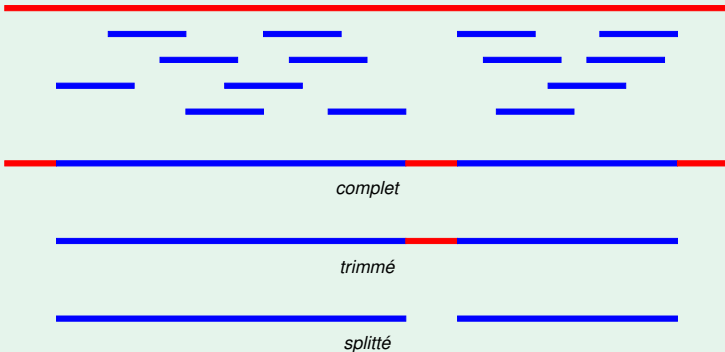
Exemple





Alignement de *reads* courts

Exemple





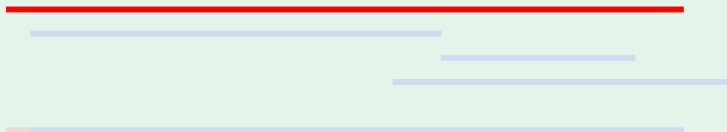
Alignement de contigs

- Contigs : plus longs que les *reads* courts
- Plus aisé de couvrir les régions très bruitées des *reads* longs
- Définition d'alignements multiples et calcul d'une séquence consensus



Alignement de contigs

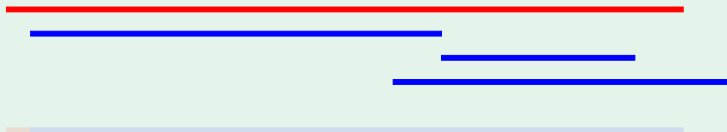
Exemple





Alignement de contigs

Exemple





Alignement de contigs

Exemple





Utilisation de graphes de de Bruijn

- Graphe de de Bruijn (DBG) : graphe orienté
- Sommets : k -mers (facteurs de longueur k) des *reads*
- Arcs : chevauchements préfixe / suffixe de longueur $k - 1$ entre les k -mers

Utilisation de graphes de de Bruijn

Exemple

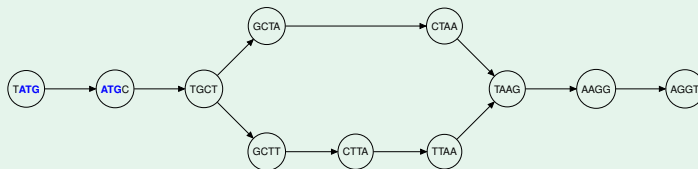


FIGURE – Graphe de de Bruijn d'ordre 4 pour l'ensemble de séquences $S = \{TATGCT, TGCTAA, GCTTAA, TAAGGT\}$.

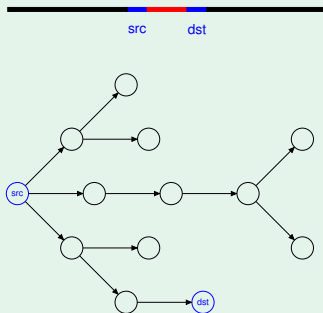


Utilisation de graphes de de Bruijn

- Construction du graphe à partir des k -mers solides des *reads* courts
- Ancrage des k -mers solides des *reads* longs sur le graphe
- Correction des régions composées de k -mers faibles en parcourant le graphe

Utilisation de graphes de de Bruijn

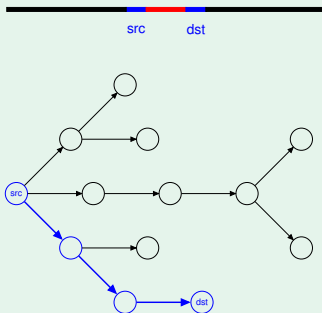
Exemple





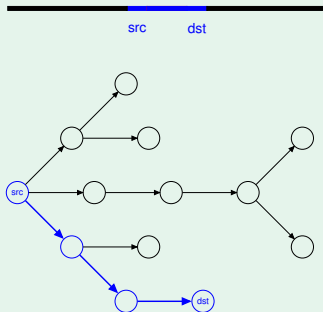
Utilisation de graphes de de Bruijn

Exemple



Utilisation de graphes de de Bruijn

Exemple





Auto-correction

- Utilise uniquement l'information des *reads* longs
- Importants développements récents
- Technologies de troisième génération : évolution rapide
 - Réduction des taux d'erreurs (10-12%)
 - Augmentation de la longueur (*ultra-long reads* ONT > 1 M bases)



Auto-correction

- État de l'art :
 - Calcul des chevauchements entre les *reads* longs
 - Calcul d'une séquence consensus à partir des chevauchements
- Deux approches :
 - 1 Pseudo alignement multiple
 - 2 Utilisation de graphes de de Bruijn



Pseudo alignement multiple

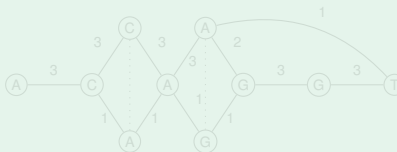
- Construction d'un graphe orienté acyclique (DAG) résumant les alignements
- DAG : représente un pseudo alignement multiple
- Recherche du chemin de poids maximum pour produire la séquence consensus



Pseudo alignement multiple

Exemple

ACCA A GGT	R ₁
ACA A GGGT	R ₂
ACCA AG GT	R ₁
ACCA A ..T	R ₃

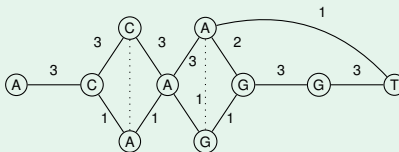




Pseudo alignement multiple

Exemple

ACCA A GGT	R ₁
ACA A GGGT	R ₂
ACCA AG GT	R ₁
ACCA A ..T	R ₃

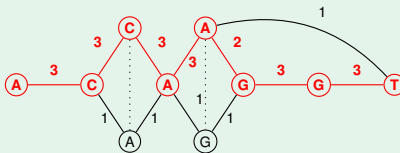




Pseudo alignement multiple

Exemple

ACCA**A**GGT R₁
 ACA**A**GGGT R₂
 ACCAA**GGT** R₁
 ACCAA...T R₃





Utilisation de graphes de de Bruijn

- Division des alignements en petites fenêtres
- Construction d'un graphe de de Bruijn pour chaque fenêtre
- Correction des fenêtres à l'aide des graphes de de Bruijn



Utilisation de graphes de de Bruijn

Exemple

.GATCGGG..TAT.TGCCCGTGTTTATGCGTGTG	R ₁
TGTTCAGGCAAATATG...GAAACAAGGCCTG..	R ₂
GAT...CGGGTATTGCCCGTGTTTATGCGTG..TG	R ₁
TATTTCTG..AT.GCGC.TGACTTTTCTTGGCAG	R ₃



Utilisation de graphes de de Bruijn

Exemple

.GATCGGG..TAT.TGCCCGTGTTTATGCGTGTG	R ₁
TGTTCAGGCAAATATG...GAAACAAGGCCTG..	R ₂
GAT..CGGGTATTGCCCGTGTTTATGCGTG..TG	R ₁
TATTTCTG..AT.GCGC.TGACTTTTCTTGGCAG	R ₃



Méthodes disponibles

Correction hybride

Méthode	Approche	Année
PBcR	Alignement <i>reads</i> courts	2012
LSC	Alignement <i>reads</i> courts	2012
ECTools	Alignement contigs	2014
LoRDEC	Graphes de de Bruijn	2014
Proovread	Alignement <i>reads</i> courts	2014
Nanocorr	Alignement <i>reads</i> courts	2015
NaS	Alignement <i>reads</i> courts	2015
CoLoRMap	Alignement <i>reads</i> courts	2016
Jabba	Graphes de de Bruijn	2016
LSCplus	Alignement <i>reads</i> courts	2016
HALC	Alignement contigs	2017
HECIL	Alignement <i>reads</i> courts	2017
Hercules	Modèles de Markov cachés	2017
FMLRC	Graphes de de Bruijn	2018
MiRCA	Alignement contigs	2018

15 méthodes

Auto-correction

Méthode	Approche	Année
PBcR-BLASR	Alignement multiple	2013
PBDAGCon	Alignement multiple	2013
Sprai	Alignement multiple	2014
PBcR-MHAP	Alignement multiple	2015
FalconSense	Alignement multiple	2016
Sparc	Alignement multiple	2016
Canu	Alignement multiple	2017
Daccord	Graphes de de Bruijn	2017
LoRMA	Graphes de de Bruijn	2017
MECAT	Alignement multiple	2017
FLAS	Alignement multiple	2018

11 méthodes



Contributions

- **ELECTOR** : un outil d'évaluation des méthodes de correction
- HG-CoLoR : un outil de correction hybride destiné aux *reads* longs extrêmement bruités ($> 30\%$ d'erreurs)
- CONSENT : un outil d'auto-correction destiné aux *ultra-long*s *reads* (> 50 k bases)



Contributions

- ELECTOR : un outil d'évaluation des méthodes de correction
- HG-CoLoR : un outil de correction hybride destiné aux *reads* longs extrêmement bruités ($> 30\%$ d'erreurs)
- CONSENT : un outil d'auto-correction destiné aux *ultra-long*s *reads* (> 50 k bases)



Contributions

- ELECTOR : un outil d'évaluation des méthodes de correction
- HG-CoLoR : un outil de correction hybride destiné aux *reads* longs extrêmement bruités ($> 30\%$ d'erreurs)
- CONSENT : un outil d'auto-correction destiné aux *ultra-long*s *reads* (> 50 k bases)



- 1 Contexte
- 2 Correction de *reads* longs
- 3 Évaluation des méthodes existantes : ELECTOR**
- 4 Correction hybride : HG-CoLoR
- 5 Auto-correction : CONSENT
- 6 Conclusion



Contexte

- Nombreuses méthodes
- Propriétés des jeux de données : impact sur la correction
- Qualité de la correction : impact important sur les analyses
- Quelle méthode utiliser pour un cas donné
- Besoin d'évaluer précisément la qualité des méthodes existantes



Objectifs

Les méthodes d'évaluation doivent :

- être rapides
- passer à l'échelle sur de larges génomes
- être faciles d'utilisation
- permettre de réaliser des benchmarks
- être informatives



Principe

- Comparaison des trois versions (triplet) d'un même *read*
 - 1 Version de référence (issue du génome de référence)
 - 2 Version originale (contenant des erreurs)
 - 3 Version corrigée (après application d'un outil de correction)
- Dédution de métriques décrivant la correction



État de l'art

- Un seul outil disponible : LRCstats
- Lent
- Nécessite de nombreux pré-traitements
- Calcule peu de métriques :
 - taux d'erreurs avant / après correction
 - comptes des différents types d'erreurs



LRCstats

référence R

ATTGCTGAT . . .

non corrigé O

ATGTCTT . . .

corrigé C

ATTGCAGAT . . .

ATTG-CTGAT . . .

AT-GTCT--T . . .

ATTGCTGAT . . .

ATTGCAGAT . . .

Aligner sur
read de
référence



ELECTOR

- Collaboration avec C. Marchet, A. Limasset, L. Lecompte et P. Peterlongo
- Facile d'utilisation
- Rapide
- Passage à l'échelle
- Calcul de métriques additionnelles



Métriques

	ELECTOR	LRCstats
Taux d'erreurs avant/après correction	✓	✓
Délétions	✓	✓
Insertions	✓	✓
Substitutions	✓	✓
Rappel	✓	✗
Précision	✓	✗
<i>Reads trimmés / splittés</i>	✓	✗
Taille manquante	✓	✗
<i>Reads étendus</i>	✓	✗
Taille d'extension	✓	✗
Ratio longueur homopolymers	✓	✗
Statistiques d'alignement	✓	✗
Statistiques d'assemblage	✓	✗

Principe

référence R

ATTGCTGAT . . .

non corrigé O

ATGTCTT . . .

corrigé C

ATTGCAGAT . . .

Alignement
multiple

ATTG-CTGAT . . .

AT-GTCT--T . . .

ATTG-CAGAT . . .



Alignement multiple

- Calcul avec POA (Partial Order Alignment)
- Avantages : plus précis, métriques additionnelles
- Inconvénient : coûteux (NP complet)
- Passage à l'échelle difficile
- Solution : stratégie de segmentation



Stratégie de segmentation

Calcul des k -mers communs entre les trois séquences





Stratégie de segmentation

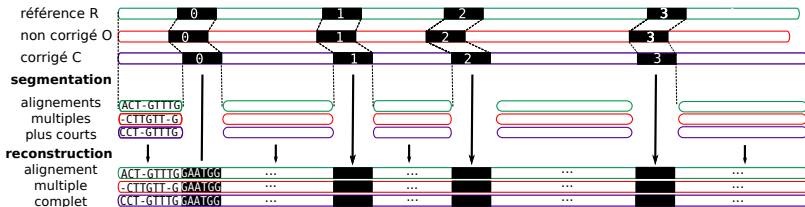
Recherche de la plus longue chaîne de k -mers communs





Alignement multiple segmenté

ensemble de graines communes et colinéaires : 0 1 2 3





Temps d'exécution

Outil	HALC	Canu
<i>E. coli</i>		
Correction	24min	12min
LRCstats	4h58	4h05
ELECTOR	28min	11min
<i>S. cerevisiae</i>		
Correction	1h19	31min
LRCstats	10h56	10h46
ELECTOR	1h55	32min
<i>C. elegans</i>		
Correction	5h59	4h33
LRCstats	83h29	85h08
ELECTOR	32h35	4h19



Métriques

Métriques	Reads non corrigés		Reads corrigés avec HALC	
	ELECTOR	LRCstats	ELECTOR	LRCstats
Nombre de bases	238 309 333	237 655 341	212 266 193	214 152 119
Taux d'erreurs (%)	14,0333	17,5129	0,4212	0,2261
Délétions	5 235 890	8 991 984	1 035 978	120 743
Insertions	28 772 841	32 589 970	100 874	215 507
Substitutions	4 058 953	16 331 23	198 853	221 646
Rappel (%)	-	-	99,9734	✗
Précision (%)	-	-	99,5872	✗
<i>Reads trimmés</i> ou <i>splittés</i>	-	-	12 043	✗
Longueur manquante moyenne (bases)	-	-	577,5	✗
<i>Reads étendus</i>	-	-	71	✗
Longueur moyenne d'extension (bases)	-	-	53,2	✗
<i>Reads</i> de mauvaise qualité	-	-	160	✗
<i>Reads</i> trop courts	-	-	3 436	✗
Ratio longueur homopolymers	-	-	1,0580	✗



Synthèse

- Évaluation rapide des résultats d'un outil de correction
- Nombreuses métriques (rappel, précision, ...)
- Stratégie innovante de segmentation d'alignement multiple
- Plus rapide et meilleur passage à l'échelle que LRCstats



- 1 Contexte
- 2 Correction de *reads* longs
- 3 Évaluation des méthodes existantes : ELECTOR
- 4 Correction hybride : HG-CoLoR**
- 5 Auto-correction : CONSENT
- 6 Conclusion



Contexte

- Premières expériences : *reads* très bruités ($> 30\%$ d'erreurs)
- Auto-correction inefficace sur de tels taux d'erreurs
- Méthodes hybrides existantes applicables, mais :
 - très coûteuses
 - mauvais résultats d'assemblage



HG-CoLoR

- Combine différentes approches de l'état de l'art
- Adopte une approche *seed and extend*
 - 1 Alignement des *reads* courts sur les *reads* longs : graines (BLASR)
 - 2 Extension des graines à l'aide d'un DBG d'ordre variable (PgSA)



Workflow

Première étape

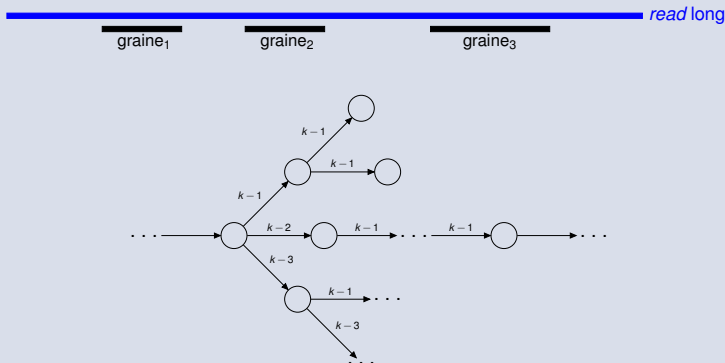
Alignement des *reads* courts sur les *reads* longs



Workflow

Deuxième étape

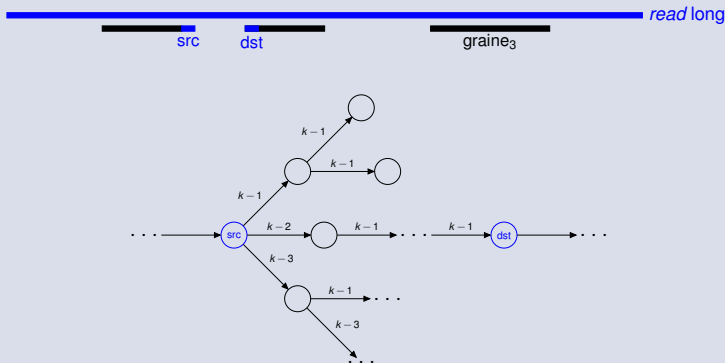
Construction du graphe à partir des k -mers des *reads* courts



Workflow

Deuxième étape

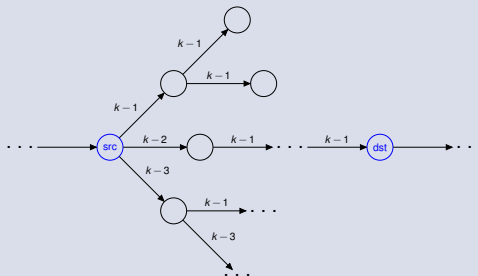
Utilisation des graines comme points d'ancrage sur le graphe



Workflow

Troisième étape

Parcours du graphe afin de relier les graines

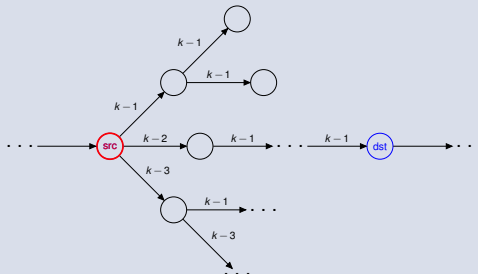




Workflow

Troisième étape

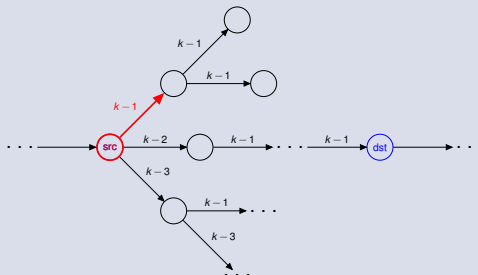
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

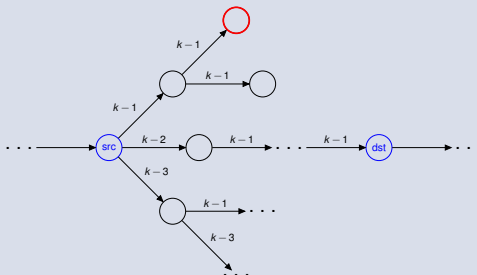
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

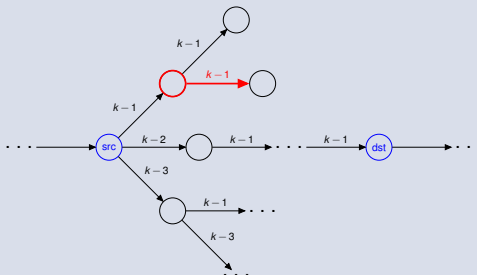
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

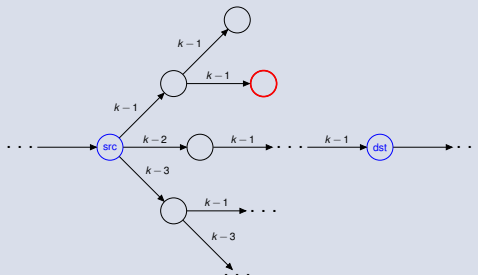
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

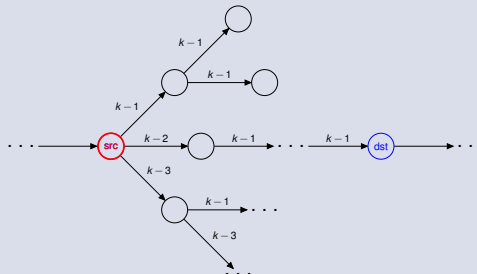
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

Parcours du graphe afin de relier les graines

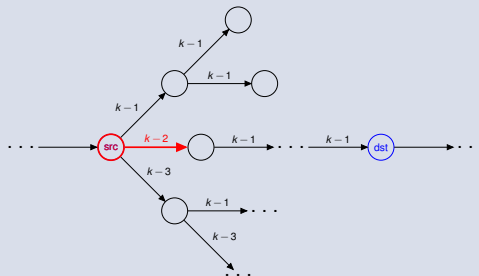




Workflow

Troisième étape

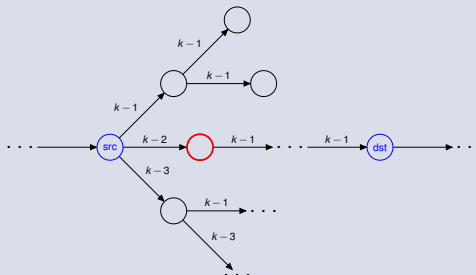
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

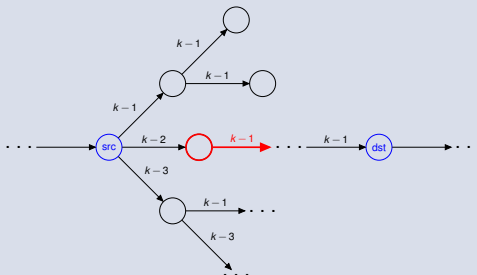
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

Parcours du graphe afin de relier les graines

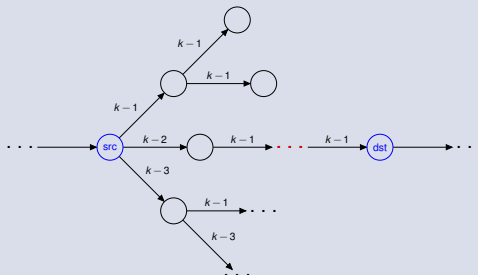




Workflow

Troisième étape

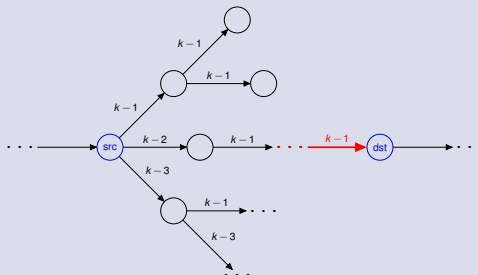
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

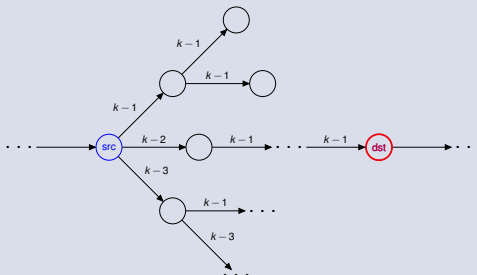
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

Parcours du graphe afin de relier les graines

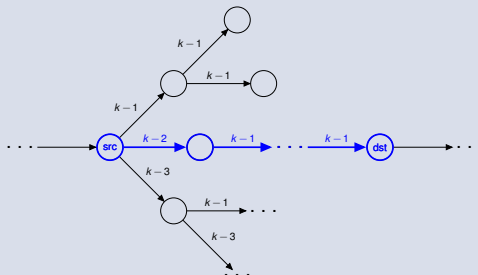




Workflow

Troisième étape

Parcours du graphe afin de relier les graines

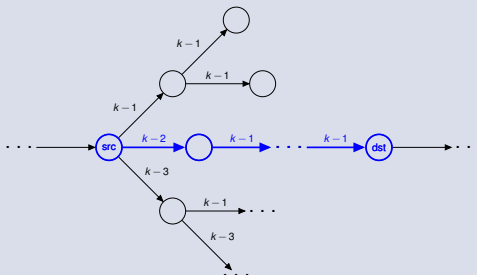




Workflow

Troisième étape

Parcours du graphe afin de relier les graines



Workflow

Troisième étape

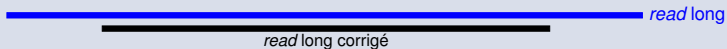
Parcours du graphe afin de relier les graines



Workflow

Troisième étape

Parcours du graphe afin de relier les graines

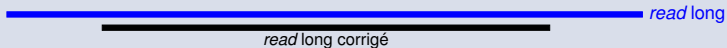




Workflow

Quatrième étape

Extension des extrémités





Comparaison à l'état de l'art

- Outils comparés :
 - CoLoRMap
 - LoRDEC
 - NaS
 - MECAT
- Évaluation avec le second module d'ELECTOR



Jeu de données

<i>S. cerevisiae</i>	
Organisme de référence	
Souche	W303
Séquence de référence	scf7180000000{084-13}
Taille du génome (Mbp)	12,2
Reads réels ONT	
Accession	Genoscope
Nombre de reads	205 923
Longueur moyenne	5 698
Nombre de bases (M)	1 173
Couverture	95x
Taux d'erreurs (%)	44,51
Reads Illumina	
Accession	Genoscope ⁶
Nombre de reads	2 500 000
Longueur	250
Nombre de bases (M)	625
Couverture	50x



Alignement

Méthode	Originaux	HG-CoLoR	CoLoRMap	LoRDEC	NaS	MECAT
Nombre de <i>reads</i>	205 923	76 193	72 017	196 091	71 793	14 740
Nombre de bases (M)	1 173	512	165	221	426	84
Longueur moyenne (bases)	5 698	6 725	2 294	1 125	5 938	5 668
<i>Reads</i> alignés (%)	22,7313	99,7126	99,7403	98,5094	99,8231	99,4573
Identité moyenne (%)	55,4937	99,7176	99,6958	98,8168	99,8985	80,0763
Couverture du génome (%)	71,1573	99,5341	99,1528	98,8934	98,6795	92,6533
Temps	N/A	8 h 51 min	10 h 44min	1 h 09 min	> 16 jours	14 min
Mémoire (Mo)	N/A	11 575	18 241	797	-	7 374



Assemblage

Méthode	Originaux	HG-CoLoR	CoLoRMap	LoRDEC	NaS	MECAT
Nombre de contigs	143	53	71	66	159	102
Contigs alignés (%)	99,3007	100	95,7746	100	100	100
NGA50 (bases)	3 355	470 355	204 052	361 726	85 205	3 163
NGA75 (bases)	3 355	212 761	88 468	178 604	43 313	3 163
Couverture du génome (%)	22,8369	95,3867	87,1140	90,0706	89,7248	15,1658



Synthèse

- Correction hybride, pour les *reads* très fortement bruités
- Combinaison de stratégies de l'état de l'art
 - ❶ Alignement de *reads* courts
 - ❷ Utilisation d'un graphe de de Bruijn d'ordre variable
- Plus efficace que les méthodes de l'état de l'art
- Bon compromis entre qualité et consommation de ressources



- 1 Contexte
- 2 Correction de *reads* longs
- 3 Évaluation des méthodes existantes : ELECTOR
- 4 Correction hybride : HG-CoLoR
- 5 Auto-correction : CONSENT**
- 6 Conclusion



Contexte

- Évolution des technologies de troisième génération :
 - réduction des taux d'erreurs (10-12%)
 - augmentation de la longueur (*ultra-long reads* ONT > 1 M bases)
- La correction est toujours nécessaire
- L'auto-correction est une alternative efficace



Contexte

- État de l'art : calcul des chevauchements, puis
 - Pseudo alignement multiple
 - Utilisation de graphes de de Bruijn
- Correction d'*ultra-long reads* : challenge
- Les méthodes de l'état de l'art ne passent pas à l'échelle



CONSENT

- Collaboration avec C. Marchet et A. Limasset
- Combine les deux approches précédentes (MSA + DBG)
- Calcule de *vrais* alignements multiples
- Généralise la stratégie de segmentation d'ELECTOR
- Passe à l'échelle sur les *ultra-long reads*
- Permet de corriger des assemblages



Pré-traitement

Calcul des chevauchements

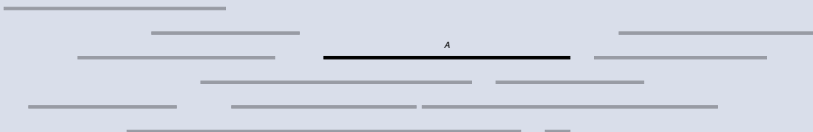
- Avec Minimap2





Première étape : calcul des piles d'alignement

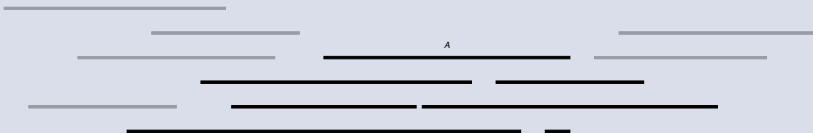
Sélection d'un *read* à corriger





Première étape : calcul des piles d'alignement

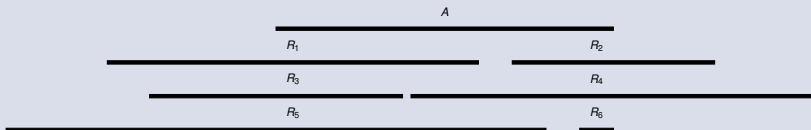
Récupération des *reads* chevauchants





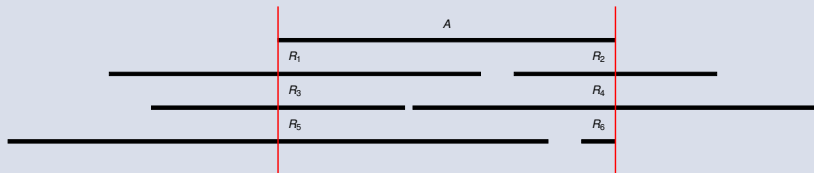
Première étape : calcul des piles d'alignement

Obtention de la pile



Première étape : calcul des piles d'alignement

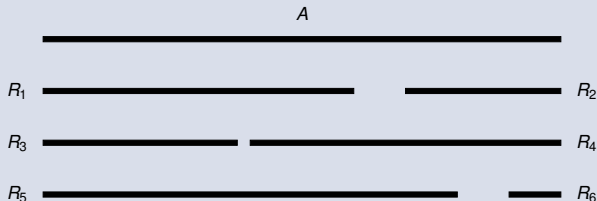
Trimmer la pile





Première étape : calcul des piles d'alignement

Trimmer la pile





Deuxième étape : division de la pile en fenêtres

On considère uniquement les fenêtres :

- De longueur ℓ fixée
- Supportées par au moins c reads

Exemple

Sur la pile précédente, avec $c = 4$:





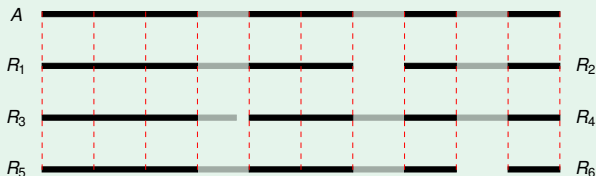
Deuxième étape : division de la pile en fenêtres

On considère uniquement les fenêtres :

- De longueur ℓ fixée
- Supportées par au moins c reads

Exemple

Sur la pile précédente, avec $c = 4$:





Troisième étape : consensus d'une fenêtre

Calcul du consensus

- Alignement multiple des séquences
- Contrairement aux autres méthodes, calcul d'un *vrai* alignement multiple
- $\Rightarrow$ POA
- Intérêt : construction directe du graphe représentant l'alignement multiple



Troisième étape : consensus d'une fenêtre

Stratégie de segmentation

- En pratique, fenêtres de quelques centaines de bases
- POA est coûteux, même sur de telles fenêtres
- Généralisation de la stratégie de segmentation d'ELECTOR
- Calcul d'alignements multiples pour de plus petites séquences



Troisième étape : consensus d'une fenêtre

Stratégie de segmentation

1. Calcul des k -mers communs entre les séquences des fenêtres





Troisième étape : consensus d'une fenêtre

Stratégie de segmentation

2. Recherche de la plus longue chaîne de k -mers :

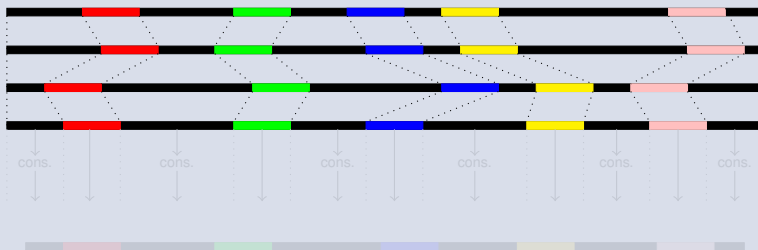




Troisième étape : consensus d'une fenêtre

Stratégie de segmentation

3. Calcul des alignements multiples et des consensus locaux

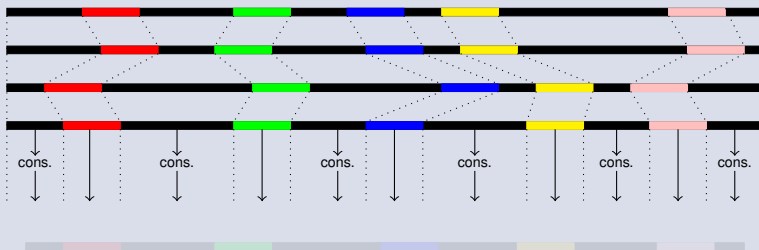




Troisième étape : consensus d'une fenêtre

Stratégie de segmentation

3. Calcul des alignements multiples et des consensus locaux

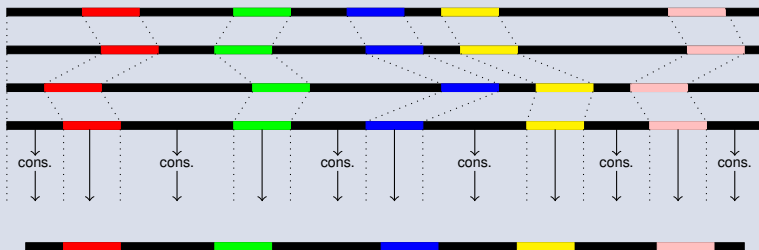




Troisième étape : consensus d'une fenêtre

Stratégie de segmentation

3. Calcul des alignements multiples et des consensus locaux





Quatrième étape : raffinement du consensus

Approche

- Consensus : *k*-mers solides en majuscules, *k*-mers faibles en minuscules

GATCGGGTcatTGCCCGTGTTTATGCGTgtg

- Construction d'un DBG avec les *k*-mers des séquences de la fenêtre
- Correction des régions en minuscules : similaire à HG-CoLoR



Comparaison à l'état de l'art

- Outils comparés :
 - Canu
 - Daccord
 - FLAS
 - MECAT
- Évaluation avec le second module d'ELECTOR



Jeux de données

	<i>D. melanogaster</i>	<i>H. sapiens</i> , chr1 ¹
Organisme de référence		
Souche	BDGP Release 6	GRCh38
Séquence de référence	ISO1 MT/dm6	NC_000001.11
Taille du génome (Mbp)	144	249
Reads réels ONT		
Accession	SRX3676783	PRJEB23027
Nombre de <i>reads</i>	1 327 569	1 075 867
Longueur moyenne (bases)	6 828	6 744
Nombre de bases (M)	9 064	7 256
Couverture	63x	29x
Taux d'erreurs (%)	14,55	17,60

¹ contient des *ultra-long reads*

Alignement

Jeu de données	Outil	Nombre de reads	Nombre de bases (M)	N50 (bases)	Identité moyenne (%)	Couverture du génome (%)	Temps	Mémoire (Mo)
<i>D. melanogaster</i>	Original	1 327 569	9 064	11 853	85,43	98,47	N/A	N/A
	Canu	829 965	6 993	12 694	95,20	97,89	14 h 04 min	10 295
	Daccord	-	-	-	-	-	-	-
	FLAS	855 275	7 866	11 742	94,99	98,09	10 h 18 min	18 820
	MECAT	849 704	7 288	11 676	96,52	97,34	1 h 54 min	13 443
	CONSENT	1 065 621	8 178	12 297	96,72	98,20	38 h	51 361
<i>H. sapiens</i>	Original	1 075 867	7 256	10 568	82,40	92,46	N/A	N/A
	Canu ¹	-	-	-	-	-	-	-
	Daccord ¹	-	-	-	-	-	-	-
	FLAS ¹	670 708	5 695	10 198	91,00	92,37	4 h 57 min	14 957
	MECAT ¹	667 532	5 479	10 343	91,69	91,44	1 h 53 min	11 075
	CONSENT	869 462	6 349	10 839	93,00	92,40	8 h 30 min	45 869

¹ ultra-long reads filtrés



Assemblage

Jeu de données	Outil	Nombre de contigs	Contigs alignés (%)	NGA50 (bases)	NGA75 (bases)	Couverture du genome (%)
<i>D. melanogaster</i>	Original	423	96,4500	864 011	159 590	83,1900
	Canu	276	97,4638	4 517 869	1 293 578	92,1034
	Daccord	-	-	-	-	-
	FLAS	265	99,6226	4 220 587	830 179	92,1105
	MECAT	370	100,0000	853 544	278 555	89,5839
	CONSENT	321	98,7539	8 077 282	2 009 827	92,5688
<i>H. sapiens</i>	Original	201	93,5300	1 025 355	247 806	77,5700
	Canu ¹	-	-	-	-	-
	Daccord ¹	-	-	-	-	-
	FLAS ¹	259	99,6139	1 622 624	346 914	88,4068
	MECAT ¹	249	99,5984	1 783 022	462 658	88,7002
	CONSENT	248	98,7903	3 760 455	1 368 818	88,9587

¹ ultra-long reads filtrés



Synthèse

- Auto-correction, pour les *ultra-long reads*
- Combinaison des deux approches de l'état de l'art : MSA + DBG
- Stratégie de segmentation pour accélérer le calcul
- Comparable à l'état de l'art
- Seule méthode à passer à l'échelle sur les *ultra-long reads*
- Permet également la correction d'assemblages



- 1 Contexte
- 2 Correction de *reads* longs
- 3 Évaluation des méthodes existantes : ELECTOR
- 4 Correction hybride : HG-CoLoR
- 5 Auto-correction : CONSENT
- 6 Conclusion**



Conclusion

- Étude des problématiques liées à la correction de *reads* longs
- Développement d'outils variés
 - Évaluation de la correction : ELECTOR
 - Correction hybride : HG-CoLoR
 - Auto-correction : CONSENT
- Pour répondre à des problèmes rencontrés par les outils de l'état de l'art



Perspectives

ELECTOR

- Parallélisation du calcul des métriques
- Ajout de compléments au module d'assemblage



Perspectives

HG-CoLoR

- Remplacement de BLASR
- Remplacement de PgSA



Perspectives

CONSENT

- Optimisation des paramètres
- Filtrage des chevauchements



Perspectives

Générales

- Étude de la correction de jeux de données polyploïdes
- Étude de la correction de données ARN
- Étude de l'impact de l'ordre correction / assemblage



7 Suppléments



Publications et communications

Revue internationale avec comité de lecture

- Pierre Morisse, Thierry Lecroq, Arnaud Lefebvre. Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph. *Bioinformatics*, Volume 34, Issue 24, 15 December 2018, Pages 4213-4222.
- Camille Marchet, Pierre Morisse, Lolita Lecompte, Arnaud Lefebvre, Thierry Lecroq, Pierre Peterlongo, Antoine Limasset. ELECTOR : Evaluator for long reads correction methods. *NAR Genomics and Bioinformatics*, under revision.



Publications et communications

Conférences internationales avec comité de lecture

- Pierre Morisse, Camille Marchet, Antoine Limasset, Thierry Lecroq, Arnaud Lefebvre. CONSENT : Scalable self-correction of long reads with multiple sequence alignment. *RECOMB-SEQ 2019*, Washington, USA, mai 2019.



Publications et communications

Conférences nationales avec comité de lecture

- Pierre Morisse, Camille Marchet, Antoine Limasset, Thierry Lecroq, Arnaud Lefebvre. CONSENT : Scalable self-correction of long reads with multiple sequence alignment. *Actes des 20^{èmes} Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)*, pages 54-61, Nantes, France, juillet 2019.
- Pierre Morisse, Thierry Lecroq and Arnaud Lefebvre. HG-CoLoR : Hybrid-Graph for the error Correction of Long Reads. *Actes des 18^{èmes} Journées Ouvertes en Biologie, Informatique et Mathématique (JOBIM)*, pages 67-74, Lille, France, juillet 2017.



Publications et communications

Workshops internationaux

- Pierre Morisse, Camille Marchet, Antoine Limasset, Thierry Lecroq, Arnaud Lefebvre. CONSENT : Scalable self-correction of long reads with multiple sequence alignment. *Data Structures in Bioinformatics (DSB) 2019*, Dortmund, Allemagne, février 2019.
- Pierre Morisse, Thierry Lecroq and Arnaud Lefebvre. Hybrid correction of long reads using a variable-order de Bruijn graph. *Data Structures in Bioinformatics (DSB) 2018*, Helsinki, Finlande, mai 2018.
- Pierre Morisse, Thierry Lecroq and Arnaud Lefebvre. Enhanced de Bruijn Graphs. *Mathematical foundations in Bioinformatics (MatBio)*, Londres, Royaume-Uni, septembre 2017.



Publications et communications

Workshops nationaux

- Pierre Morisse, Antoine Limasset, Camille Marchet, Arnaud Lefebvre, Pierre Peterlongo, Thierry Lecroq. LoRSCo : Long Reads Self-Correction. *SeqBio 2018*, Rouen, France, novembre 2018.
- Camille Marchet, Pierre Morisse, Lolita Lecompte, Antoine Limasset, Arnaud Lefebvre, Thierry Lecroq, Pierre Peterlongo. ELECTOR : EvaLuator of Error Correction Tools for lOnG Reads. *SeqBio 2018*, Rouen, France, novembre 2018.
- Pierre Morisse, Thierry Lecroq and Arnaud Lefebvre. HG-CoLoR : enHanced de Bruijn Graph for the error Correction of Long Reads. *SeqBio 2017*, Lille, France, novembre 2017.

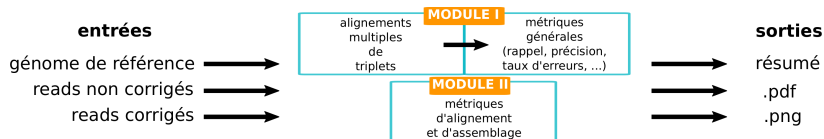


Publications et communications

Communications sous forme de posters

- Camille Marchet, Pierre Morisse, Lolita Lecompte, Antoine Limasset, Arnaud Lefebvre, Pierre Peterlongo, Thierry Lecroq.
ELECTOR : EvaLUation of Error Correction Tools for lOnG Reads.
Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM) 2018, Marseille, France, juillet 2018.

Vue d'ensemble





Comptes et taux d'erreurs

référence	ACT-GTTTGA-CTTTG-CTGAT
non corrigé	-CTTATT-GAACT-GT-T--T
corrigé	ACTT-TT-GAACTTTGTCAGAT

■ délétion dans le *read* non corrigé

■ insertion dans le *read* non corrigé

■ substitution dans le *read* non corrigé

■ délétion dans le *read* corrigé

■ insertion dans le *read* corrigé

■ substitution dans le *read* corrigé



Rappel, précision

référence R

ACT-GTTTGA-CTTTG-CTGAT

non corrigé O

GCCTGT-TGGACT--GTCAG-T

corrigé C

ACTTGTTTGAATTTGTCAGAT

■ positions à corriger ($nt(R) \neq nt(O)$)

■ ■ positions traitées ($nt(R) \neq nt(O)$ OR $nt(C) \neq nt(R)$)

■ positions traitées telles que $nt(C) = nt(R)$

■ positions traitées telles que $nt(C) \neq nt(R)$

$$\text{Rappel} = \frac{\text{■}}{\text{■}}$$

$$\text{Précision} = \frac{\text{■}}{\text{■} + \text{■}}$$



Validation de la stratégie de segmentation

- Jeux de données simulées de la bactérie *E. coli*
- Reads de longueurs 1 000, 10 000, 100 000
- Taux d'erreurs de 10%
- Profondeur de 100x

Expérience	Rappel (%)	Précision (%)	Bases correctes (%)	Temps
"1k" sans segmentation	99,712	98,996	98,980	2 h 05 min
"1k" avec segmentation	99,769	98,992	98,979	28 min
"10k" sans segmentation	99,921	99,781	99,794	20 h 50 min
"10k" avec segmentation	99,921	99,795	99,793	29 min
"100k" sans segmentation	99,913	99,925	99,956	8 jours 18 h 38 min
"100k" avec segmentation	99,924	99,903	99,902	1 h 11 min



Comparaison à LRCstats

Jeu de données	<i>E. coli</i>	<i>S. cerevisiae</i>	<i>C. elegans</i>
Génome de référence			
Souche	K-12 substr. MG1655	W303	Bristol N2
Séquence de référence	NC_000913	scf718000000{084-13}	GCA_000002985.3
Taille du génome (Mbp)	4,6	12,2	100
Reads simulés PacBio			
Nombre de <i>reads</i>	11 306	30 132	244 277
Longueur moyenne (bases)	8 226	8 204	8 204
Nombre de bases (Mbp)	93	247	2 004
Profondeur	20x	20x	20x
Taux d'erreurs (%)	18,6	18,6	18,6
Reads Illumina			
Accession	Genoscope	Genoscope	Simulés avec ART
Nombre de <i>reads</i>	775 500	2 500 000	20 057 100
Longueur des <i>reads</i> (bases)	300	250	250
Nombre de bases (Mbp)	232	625	5 000
Profondeur	50x	50x	50x



Passage à l'échelle

- Jeux de données simulées de la bactérie *E. coli*
- Couverture de 100x
- Taux d'erreurs de 10%
- *Reads* de longueur 1 000, 10 000, 100 000, 1 000 000

Outil	Longueur des <i>reads</i> (bases)	Mémoire (Mo)	Temps	Temps CPU
LRCstats	1 000	1 803	42 min	8 h 18 min
ELECTOR	1 000	1 030	12 min	28 min
LRCstats	10 000	13 484	4 h 51 min	2 jours 22 h 38 min
ELECTOR	10 000	3 091	13 min	29 min
ELECTOR	100 000	12 231	28 min	1 h 11 min
ELECTOR	1 000 000	24 881	2 h 44 min	11 h 05 min

Graphe de de Bruijn

- Graphes de de Bruijn largement utilisés
- Rencontrent des difficultés :
 - grandes valeurs de k : arcs manquants dans le graphe
 - petites valeurs de k : nombreux embranchements dans le graphe

Graphe de de Bruijn d'ordre variable

- Représente tous les DBG jusqu'à un ordre maximal K
- Problème : représentations uniquement théoriques, ou implémentations non satisfaisantes
- Solution : développement d'une nouvelle représentation, reposant sur une structure d'index (PgSA)
 - Stockage des K -mers dans l'index
 - Interrogation de l'index pour retrouver les arcs du graphe



Représentation

Exemple

- $S = \{AGCTTACA, CTTACGTA\}$
- DBG d'ordre maximum $K = 6$: indexation des 6-mers des *reads*

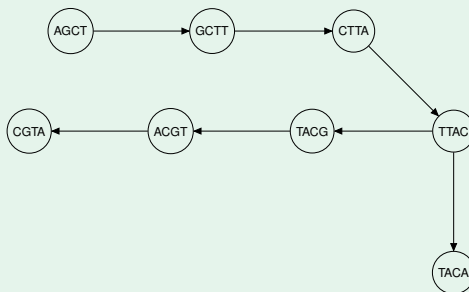
Ensemble des 6-mers

1 : AGCTTA
2 : CTTACA
3 : CTTACG
4 : GCTTAC
5 : TACGTA
6 : TTACGT

Représentation

Exemple

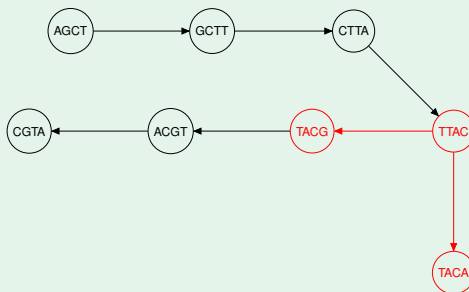
DBG d'ordre 4 de l'ensemble de *reads* précédent



Représentation

Exemple

DBG d'ordre 4 de l'ensemble de *reads* précédent



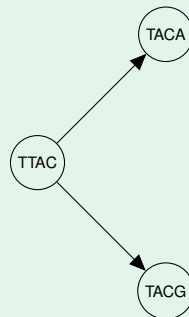
Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA



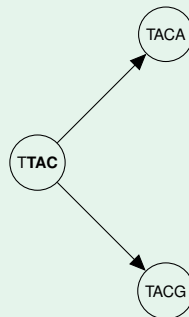
Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

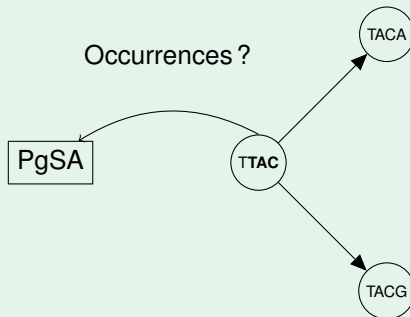


Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT



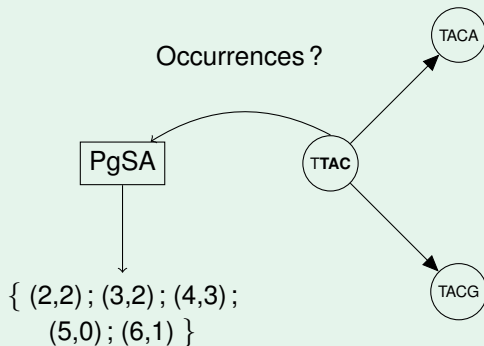
Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

Occurrences ?



Représentation

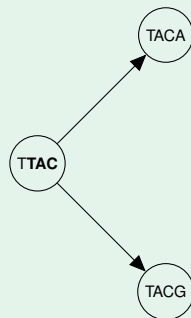
Exemple

Ensemble des 6-mers

1 : AGCTTA
2 : CTT**TACA**
3 : CTTACG
4 : GCTTAC
5 : TACGTA
6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

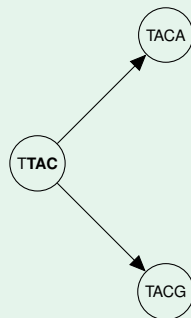
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTT**TAC**A
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

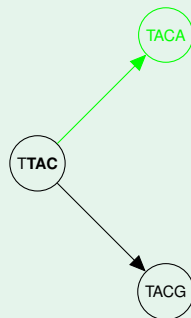
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTT**TAC**A
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

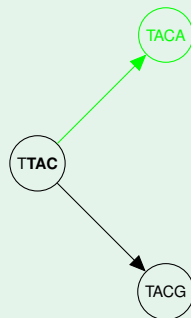
Exemple

Ensemble des 6-mers

1 : AGCTTA
 2 : CTTACA
 3 : CTT**TAC**G
 4 : GCTTAC
 5 : TACGTA
 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

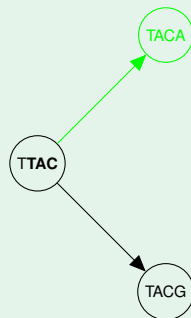
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTT**TACG**
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

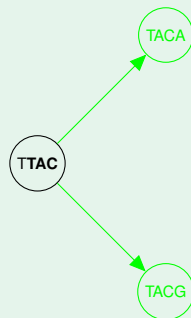
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTT**TACG**
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

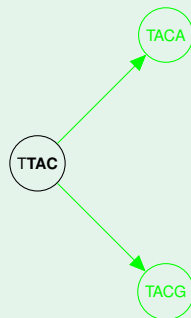
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCT**TAC**
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

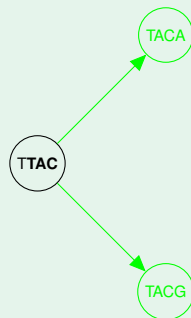
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCT**TAC**
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

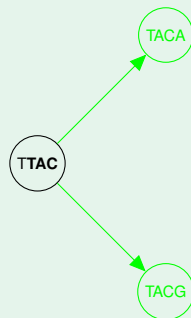
Exemple

Ensemble des 6-mers

1 : AGCTTA
 2 : CTTACA
 3 : CTTACG
 4 : GCTTAC
 5 : **TAC**GTA
 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

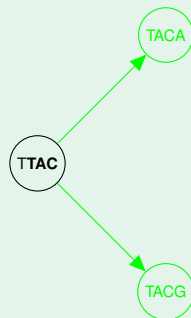
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : **TAC**GT**A**
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

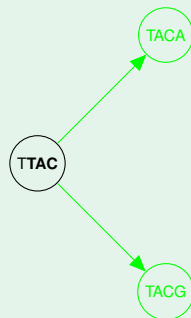
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : **TAC**TA
- 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

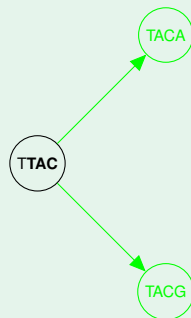
Exemple

Ensemble des 6-mers

1 : AGCTTA
 2 : CTTACA
 3 : CTTACG
 4 : GCTTAC
 5 : TACGTA
 6 : TTACGT

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

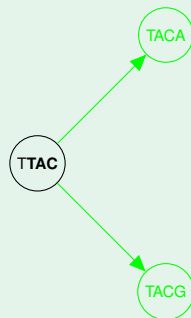
Exemple

Ensemble des 6-mers

1 : AGCTTA
 2 : CTTACA
 3 : CTTACG
 4 : GCTTAC
 5 : TACGTA
 6 : TTAC**GT**

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



Représentation

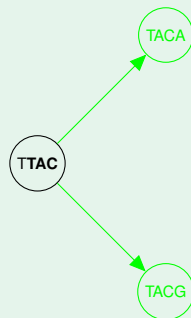
Exemple

Ensemble des 6-mers

1 : AGCTTA
 2 : CTTACA
 3 : CTTACG
 4 : GCTTAC
 5 : TACGTA
 6 : TTAC**GT**

PgSA

$\{ (2,2); (3,2); (4,3); (5,0); (6,1) \}$



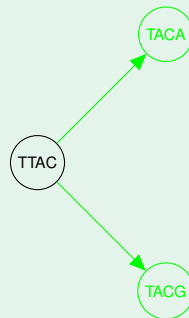
Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA



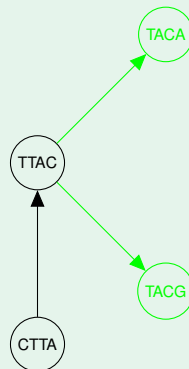
Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA



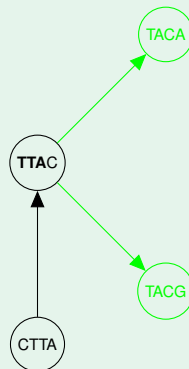
Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

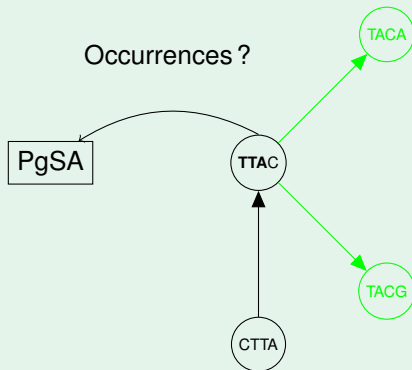


Représentation

Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT



Représentation

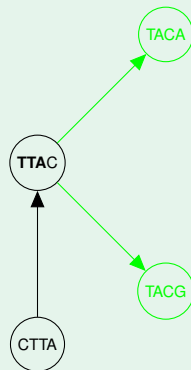
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

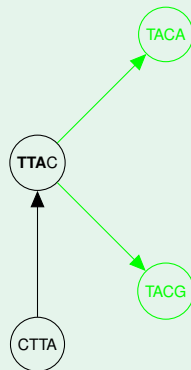
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

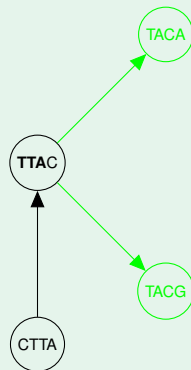
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

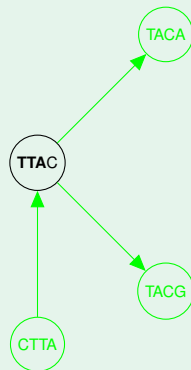
Exemple

Ensemble des 6-mers

- 1 : AG**C**TTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

{ (1,3); (2,1); (3,1);
(4,2); (6,0) }



Représentation

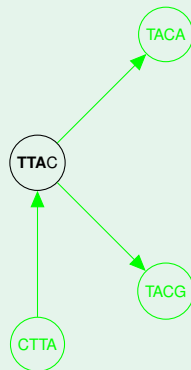
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

{ (1,3); (2,1); (3,1);
(4,2); (6,0) }



Représentation

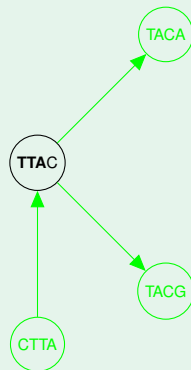
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : **CTTACA**
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

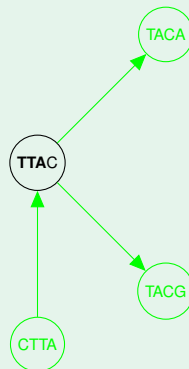
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : **CTTACA**
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

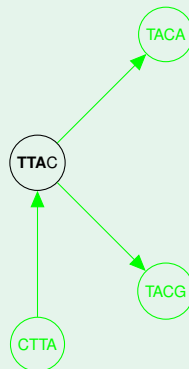
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

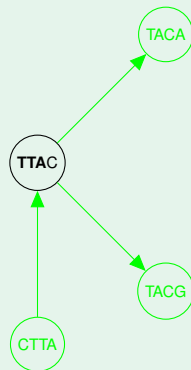
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : **CTT**ACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

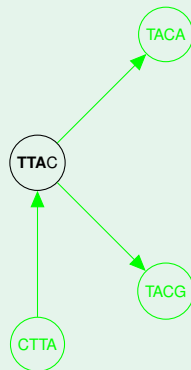
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : **CTTACG**
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

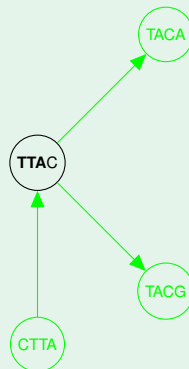
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

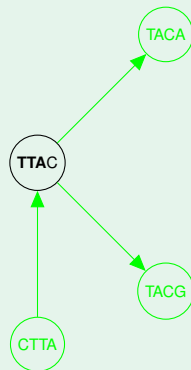
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : G**CTTAC**
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

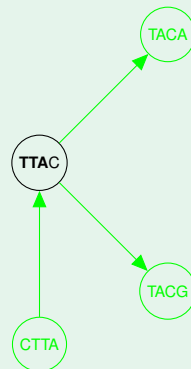
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : G**CTTAC**
- 5 : TACGTA
- 6 : TTACGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

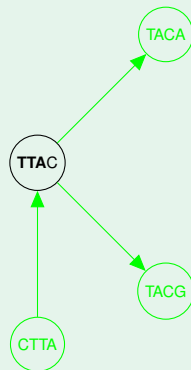
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : **TTACGT**

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Représentation

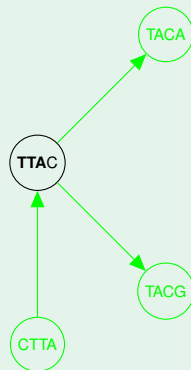
Exemple

Ensemble des 6-mers

- 1 : AGCTTA
- 2 : CTTACA
- 3 : CTTACG
- 4 : GCTTAC
- 5 : TACGTA
- 6 : **TTA**CGT

PgSA

$\{ (1,3); (2,1); (3,1); (4,2); (6,0) \}$



Workflow

Première étape

Alignement des *reads* courts sur les *reads* longs



Workflow

Deuxième étape

Combinaison des *reads* chevauchants



Cinquième étape : ancrage du consensus

Par alignement

- Alignement local, autour des positions de la fenêtre



- Répétition du processus avec les autres fenêtres



Validation de la stratégie de segmentation

- Jeu de données simulées de la bactérie *E. coli*
- Avec SimLoRD, 50x de couverture, 12% d'erreurs
- Évaluation avec le premier module d'ELECTOR



Validation de la stratégie de segmentation

Métrique	Sans segmentation	Avec segmentation
Nombre de bases	214 836 958	215 724 911
Taux d'erreurs	0,3394	0,1834
Délétions	592 740	302 677
Insertions	105 575	57 231
Substitutions	31 078	9 302
Rappel	99,9711	99,9816
Précision	99,6659	99,8196
<i>Reads trimmés ou splittés</i>	26 411	23 246
Longueur manquante moyenne	91,0	74,8
<i>Reads étendus</i>	1	0
Longueur moyenne d'extension	33,0	0,0
<i>Reads de mauvaise qualité</i>	0	0
<i>Reads trop courts</i>	0	0
Ratio longueur homopolymers	0,9936	1,0000
Temps	5 h 31 min	7 min
Mémoire (Mo)	750	675



Correction d'assemblages

- Permet de corriger des assemblages de *reads* bruts
- Immédiat : calcul des chevauchements entre contigs et *reads*
- Reste du processus : identique
- Premier outil d'auto-correction à proposer cette fonctionnalité

Résultats

- Jeux de données simulées de *C. elegans*
- Avec SimLoRD, 60x de couverture, 12% d'erreurs
- Comparaison de CONSENT à RACON
- Évaluation des assemblages avec QUAST-LG



Résultats

Métrique	Outil		
	Original	RACON	CONSENT
Nombre de contigs	47	47	47
Nombre de contigs alignés	47	47	47
NGA50 (bp)	5 495 235	5 073 456	5 019 450
NGA75 (bp)	2 656 350	2 349 027	2 286 190
Couverture du génome (%)	99,71	99,72	99,61
Erreurs / 100 kbp	10 611	819	330
Temps (secondes CPU)	N/A	136 325	30 907
Mémoire (Mo)	N/A	14 288	85 486