Outils d'assemblage de génomes

Anaïs LOUIS

Encadrante : Sèverine Bérard

Tuteur pédagogique : Alban Mancheron

Master Sciences et Numériques pour la Santé
Spécialité Bioinformatique, Connaissances, Données

16 janvier 2020

Séquençage

- Séquençage -> avancées technologiques depuis le début des années 2000.
- Enjeu -> assemblage de génome.

Assembler = Reconstruire une séquence initiale à partir des *reads* obtenus par le séquençage de cette même séquence.

Assemblage de génome

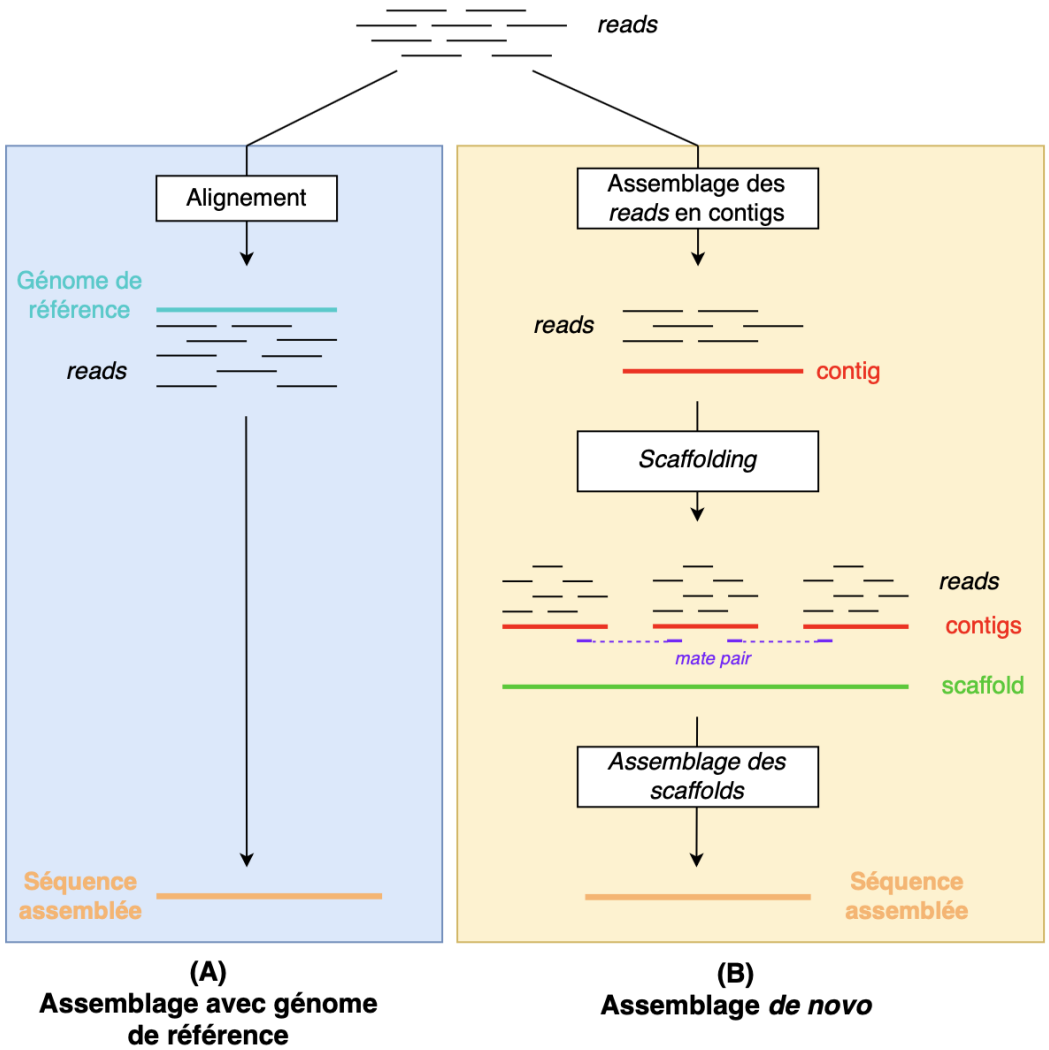


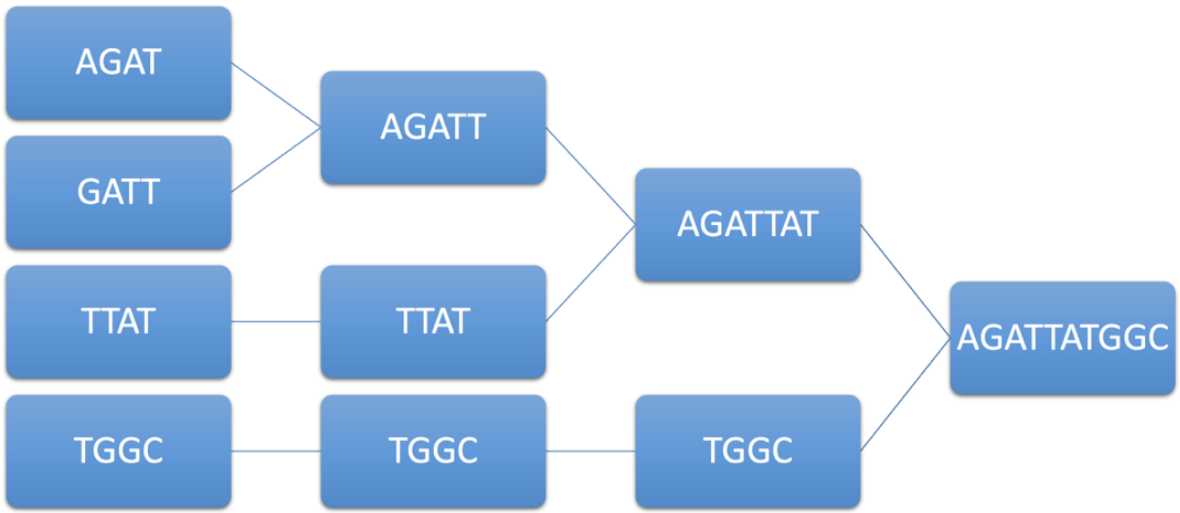
Figure – Les différents types d'assemblage.

Objectifs du stage

- Présentation succinte des stratégies d'assemblage
- Aperçu des outils d'assemblage de génome *de novo* existants
- Réaliser un guide d'aide à la décision dans le choix d'un outil d'assemblage de génome *de novo*

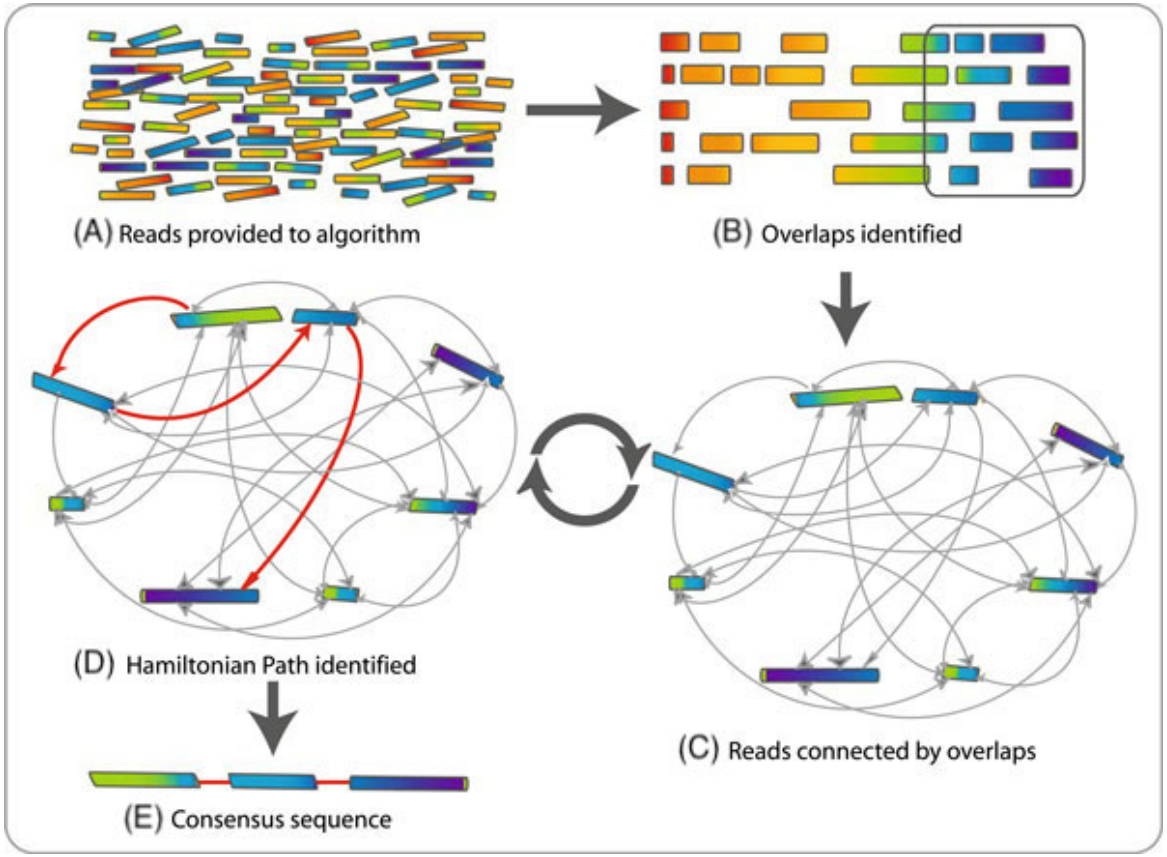
Algorithmes d'assemblage de génome

Algorithmes d'assemblage de génome - Méthode gloutonne



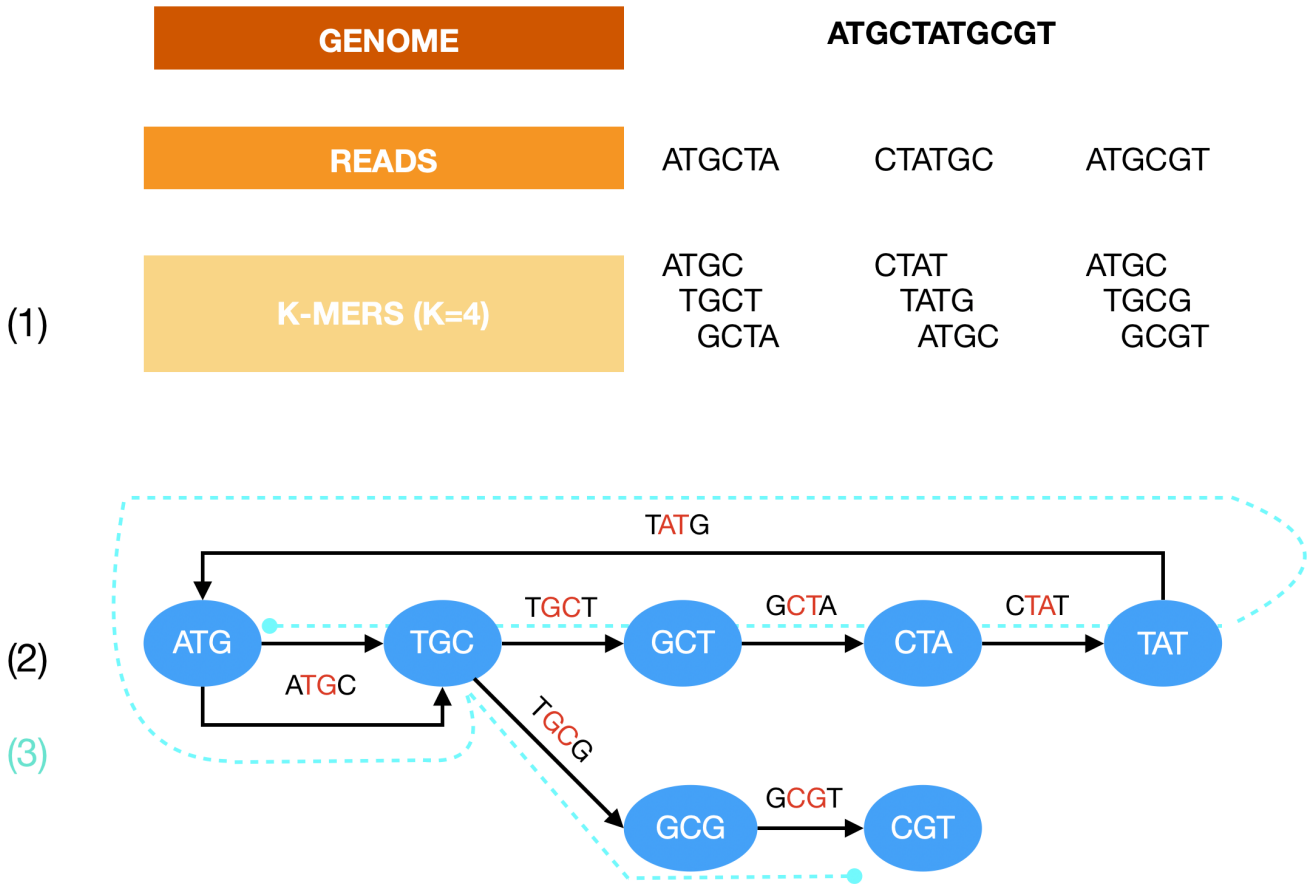
1 : Calcul de scores de chevauchement entre les *reads*
2 : Assemblage de la paire de *reads* avec le score le plus élevé
3 : Refaire les étapes 1 et 2 jusqu'à ce que tous les *reads* soient assemblés

Algorithmes d'assemblage de génome - Méthode OLC



(A) *Reads* donnés en entrée à l'algorithme
(B) Identification des chevauchements entre chaque *reads*
(C) Construction d'un graphe de chevauchement
(D) Recherche d'un chemin Hamiltonien
(E) Construction d'une séquence consensus

Algorithmes d'assemblage de génome - Méthode GDB



(1) Décomposition des *reads* en k-mers
(2) Construction du graphe de De Bruijn
(3) Recherche du chemin eulérien

Recensement des outils

Recensement des outils

Outil	Sortie	Type *reads*	Algorithme	Outil	Sortie	Type *reads*	Algorithme
Flye	2019	L	Repeat graph	Mapsembler	2012	S	OLC
Peregrine	2019	L	SHIMMER	Metavelvet	2012	S	GDB
Ra	2019	L	OLC	Minia	2012	S	GDB
Shasta	2019	L	?	Newbler	2012	Pyro	OLC
Wtdbg2	2019	L	FBG	IMR/DENOM	2011	S	?
Canu	2017	L	OLC	Locas	2011	S	OLC
Smartdenovo	2017	L	OLC	Soapdenovo	2009	S	GDB
Unicycler	2017	S,L,H	GDB/OLC	Abyss	2009	S	GDB
Miniasm	2016	L	OLC	Allpaths	2008	S	GDB
NOVOPlast	2016	S	Extension par graines	Euler-SR	2008	S	GDB
IVA	2015	S	OLC	Velvet	2008	S	GDB
Megahit	2015	S	GDB	Edena	2008	S	OLC
Falcon	2013	L	HGAP	SSAKE	2007	S	Glouton
Masurca	2013	S,H	OLC/GDB	VCAKE	2007	S	Glouton
SGA	2012	S	String graph	Euler	2001	Sanger	GDB
Spades	2012	S	GDB	CELERA	2000	Sanger	OLC
IDBA	2012	S	GDB				

Table – Caractéristiques des outils d'assemblage de génome

Sigles : S : *short read*, L : *long read*, H : hybride, GDB : graphe de De Bruijn, OLC : *Overlap Layout Consensus*, SHIMMER : Sparse HIereachical MimiMizER, FBG : Fuzzy Bruijn Graph.

Recensement des outils

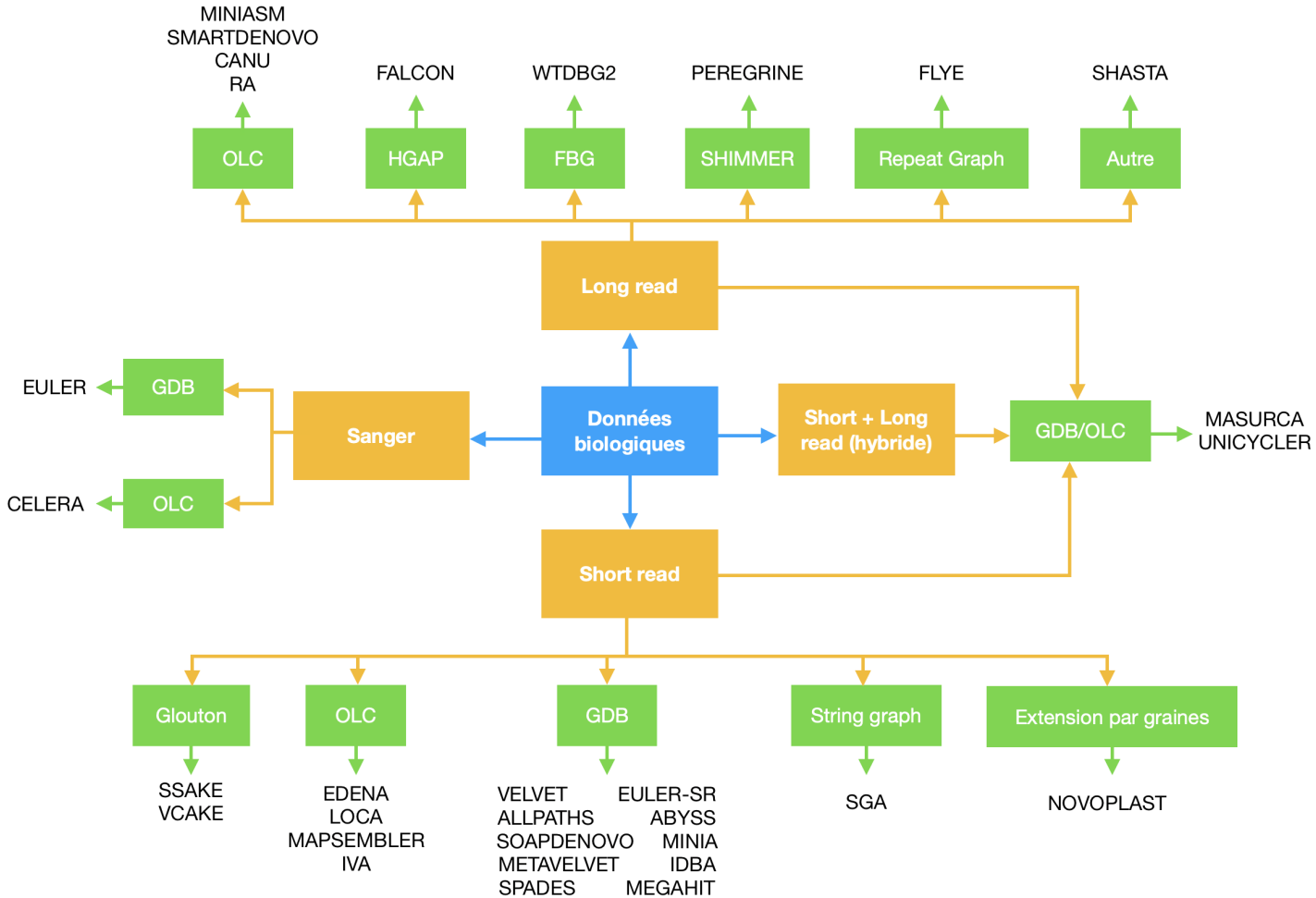
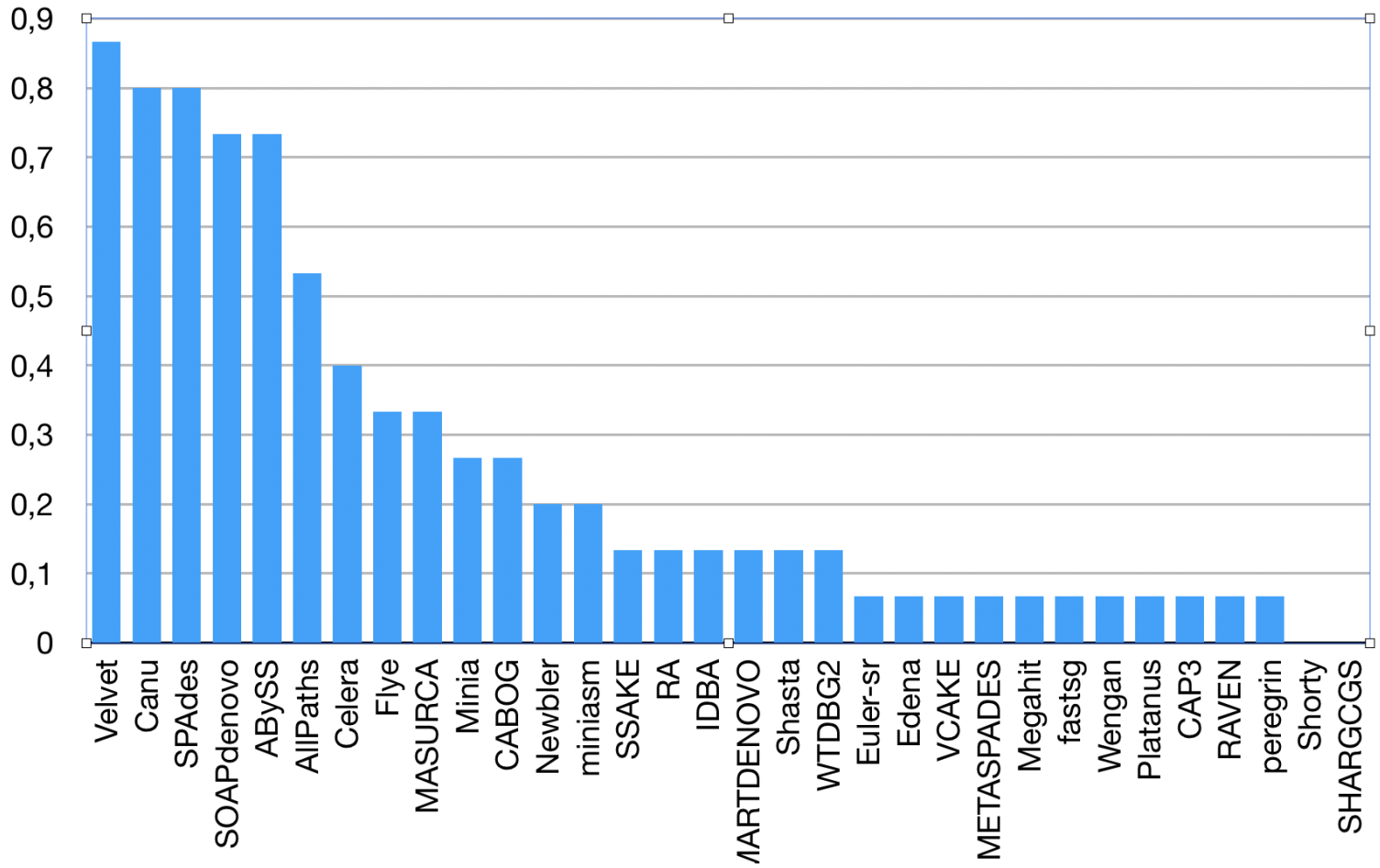


Figure – Guide d'aide à la décision dans le choix d'un assembleur en fonction des données à disposition et du type d'algorithme implémenté

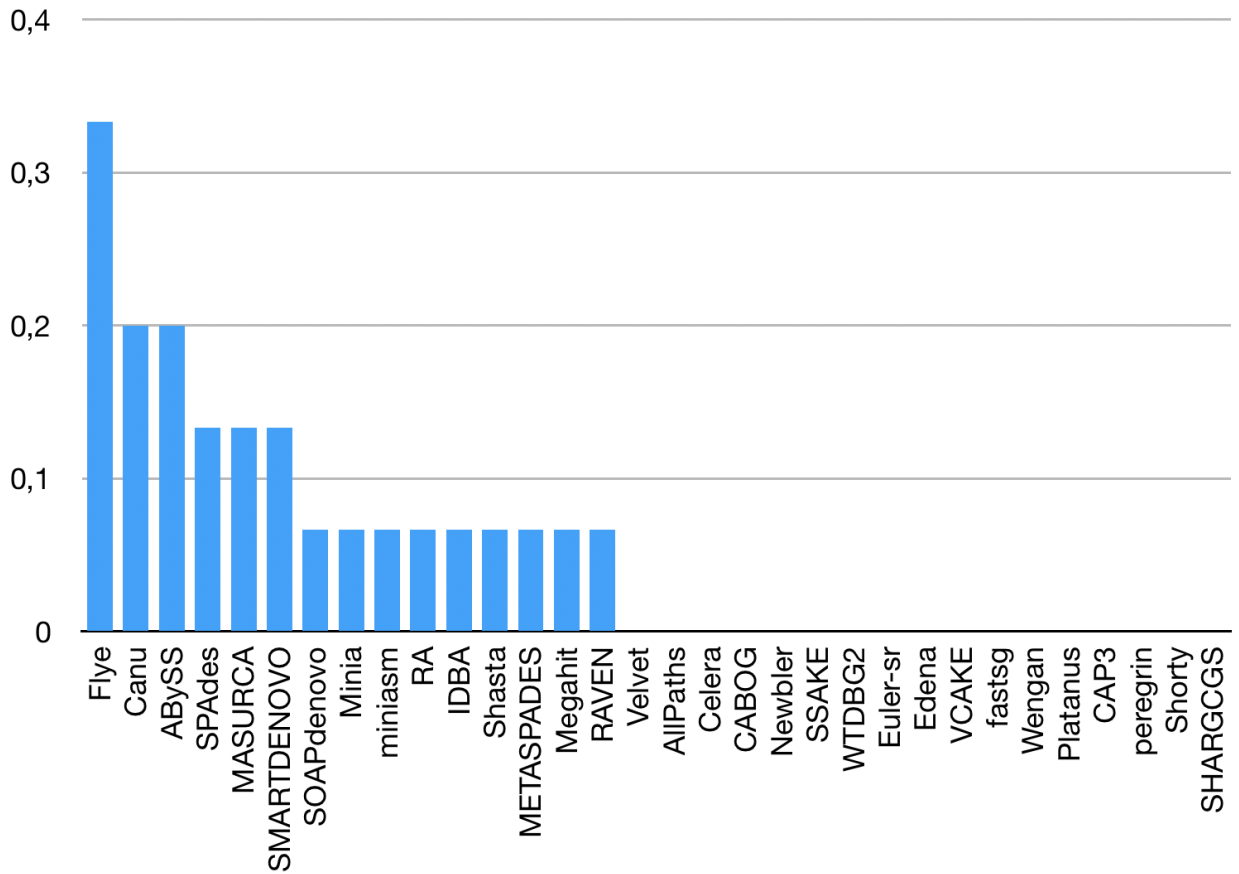
Sondage sur l'utilisation des outils d'assemblage de génome

Sondage - Connaissance des outils



32 outils : 16 proposés dans la question + 16 donnés en réponse "autre"

Sondage – Outils les plus utilisés



15 outils les plus utilisés. Pourquoi une telle répartition ?

-> ***"Tout dépend de l'objectif"***

Conclusion et Perspective(s)

Conclusion

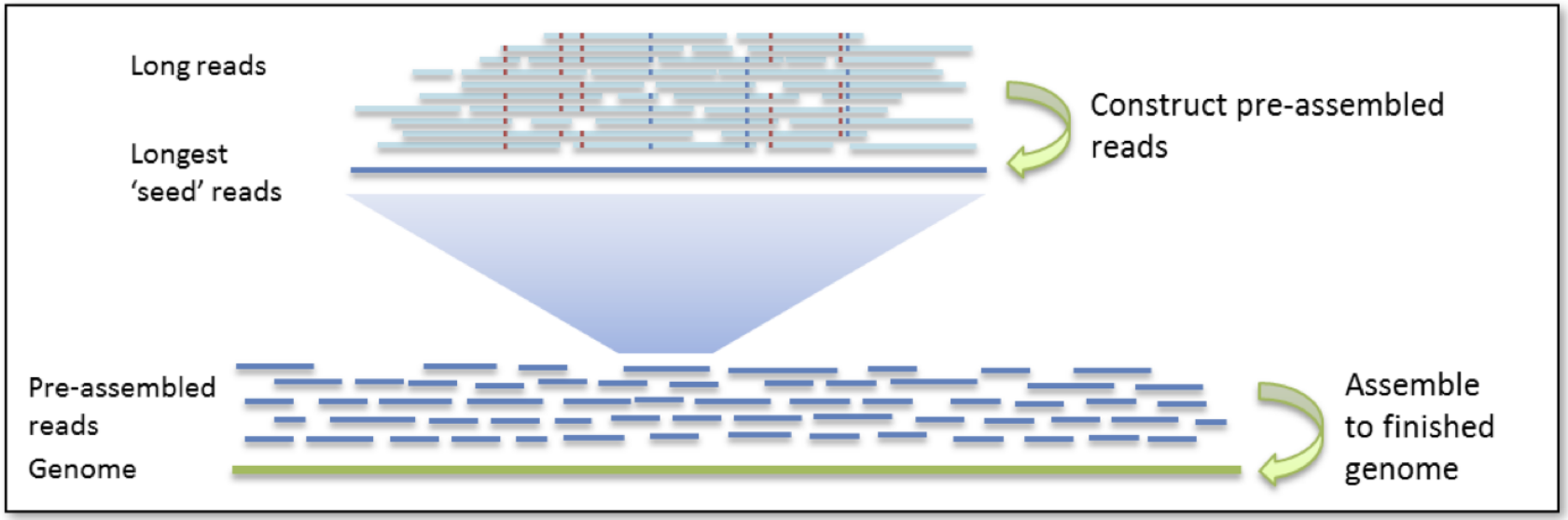
- De nombreux outils disponibles
- Évolution des outils en même temps que les améliorations technologiques
- Choix d'un outil en fonction des données à disposition

Perspective(s)

- Approfondir l'étude pour améliorer le guide d'aide à la décision

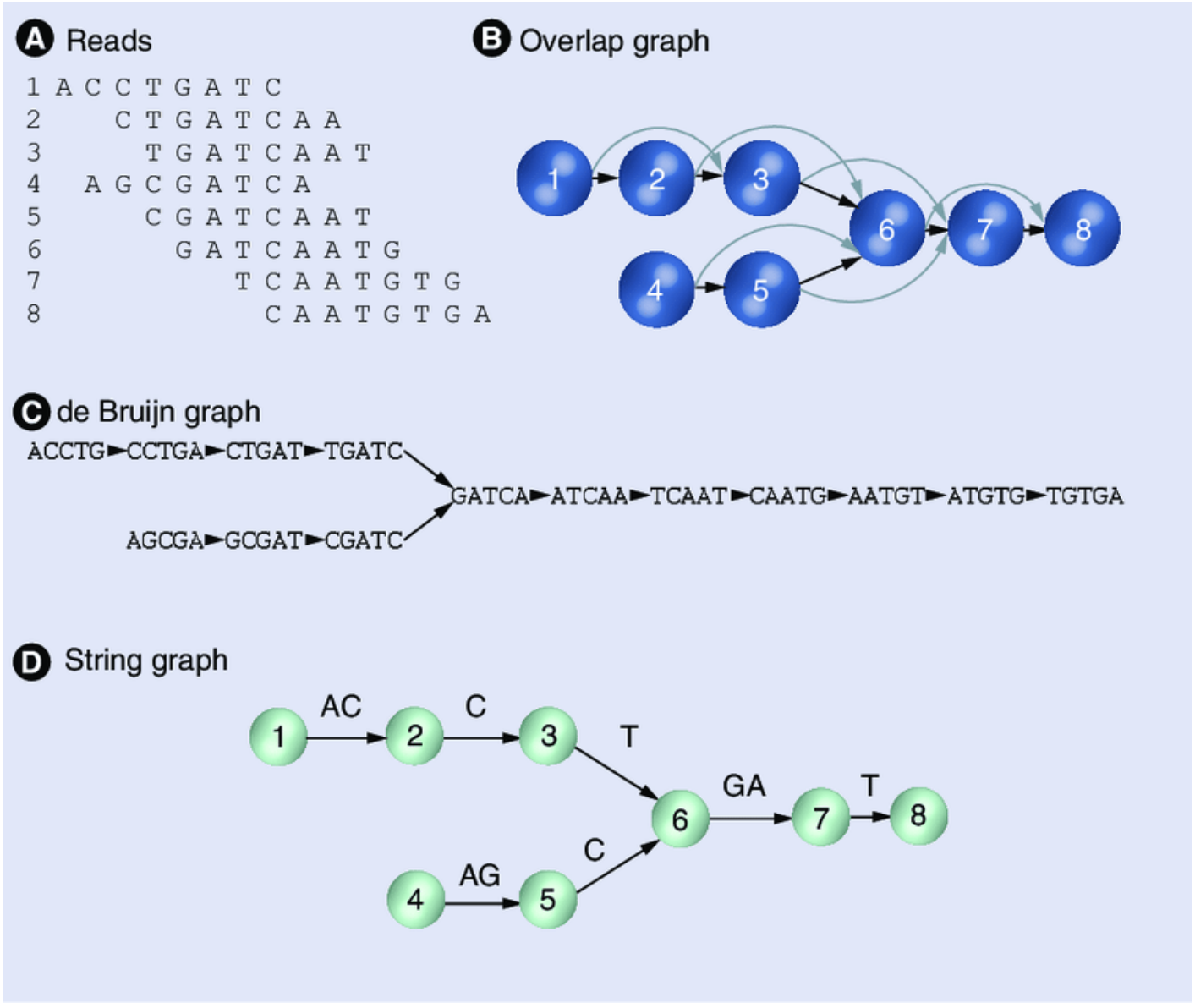
Merci de votre attention

Méthode HGAP



https ://pb-falcon.readthedocs.io/en/latest/about.html

Méthode *String graph*



Sondage

1. Quel(s) outil(s) d'assemblage connaissez-vous ? (Si vous en connaissez d'autres qui ne sont pas dans la liste ci-dessous, les ajouter dans la case "autre").
Propositions : Canu, SPAdes, Minia, SOAPdenovo, AllPaths, ABySS, Velvet, Euler-sr, Shorty, Edena, CABOG, Celera, Newbler, VCAKE, SHARGCGS, SSAKE, Autre(s).

2. Parmi ces outils, lequel utilisez-vous le plus régulièrement ?

3. Pourquoi utilisez-vous cet outil plutôt qu'un autre ?

4. Sur quel(s) type(s) de données et sur quel(s) génome(s) l'utilisez-vous ?

5. Quels sont les points forts de cet outil ?

6. Quels sont les points faibles de cet outil ?

7. Quels paramètres utilisez-vous ? (Paramètres par défaut ou paramètres plus spécifiques).

8. Si vous avez des commentaires ou références à ajouter, veuillez les écrire ci-dessous. Si vous souhaitez être informé des résultats de cette enquête, vous pouvez laisser en plus votre adresse mail.