Formation
« Bioinformatique pour le traitement de données deséquençage (NGS) »

Annabelle Haudry

Equipe Le Cocon

LBBE

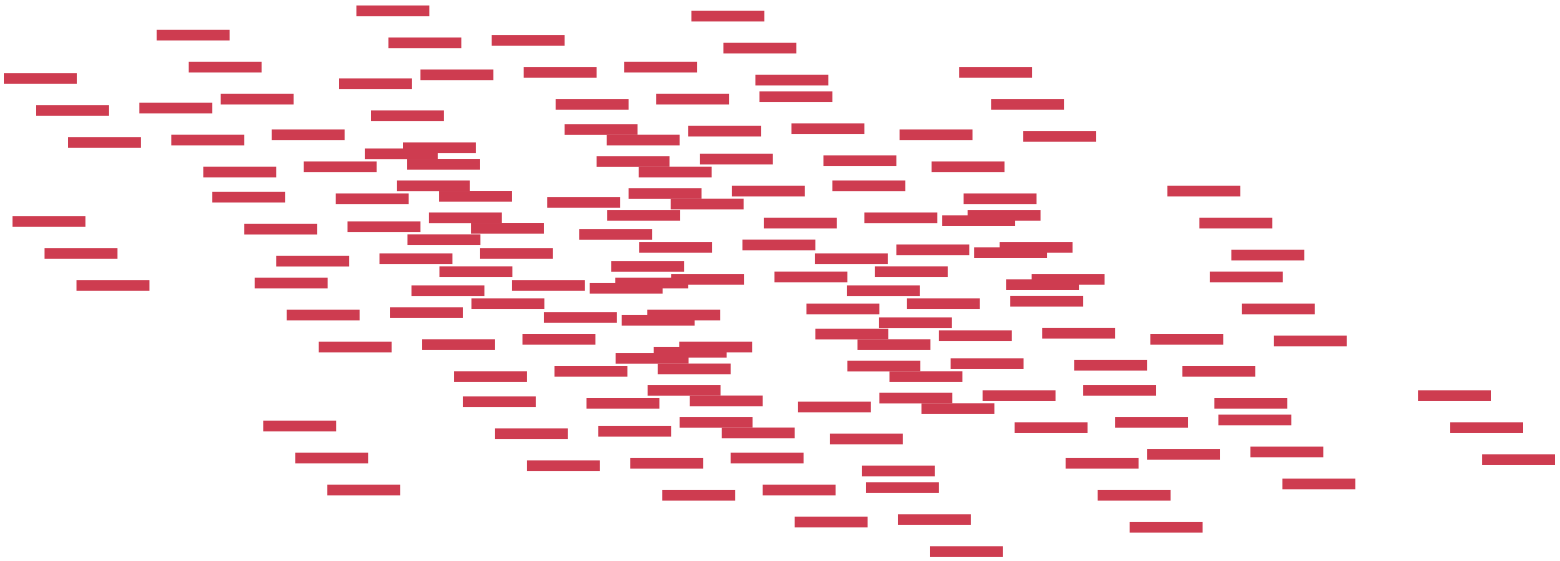
cnrs

Université Claude Bernard Lyon 1

Les techniques de séquençage NGS produisent des millions de courtes séquences (short reads).

Les assembler sans génome de référence représente un véritable challenge.

Plusieurs assembleurs *de novo* ont été développés pour s'atteler à la tâche, reposant sur plusieurs algorithmes.



Pourquoi réaliser un assemblage *de novo*?

- Absence d'un génome de référence
- Découverte de nouveaux gènes (polymorphisme/ divergence espèce proche)
- Polymorphisme d'inversions chromosomique
- En parallèle du mapping sur génome de référence ?

Plan

— Les principes des méthodes d'assemblage

— Principaux assembleurs

— Correction : besoin, principes, méthodes, outils

— Cas des reads courts (paired-ends et mate pairs)

— Méthode et pipeline pour reads longs et courts par graphe de chevauchements

Point de départ

Short reads:

Lectures au format fastq, séquences de 100-150pb pour HiSeq, 250pb pour MiSeq.

Les différents algorithmes

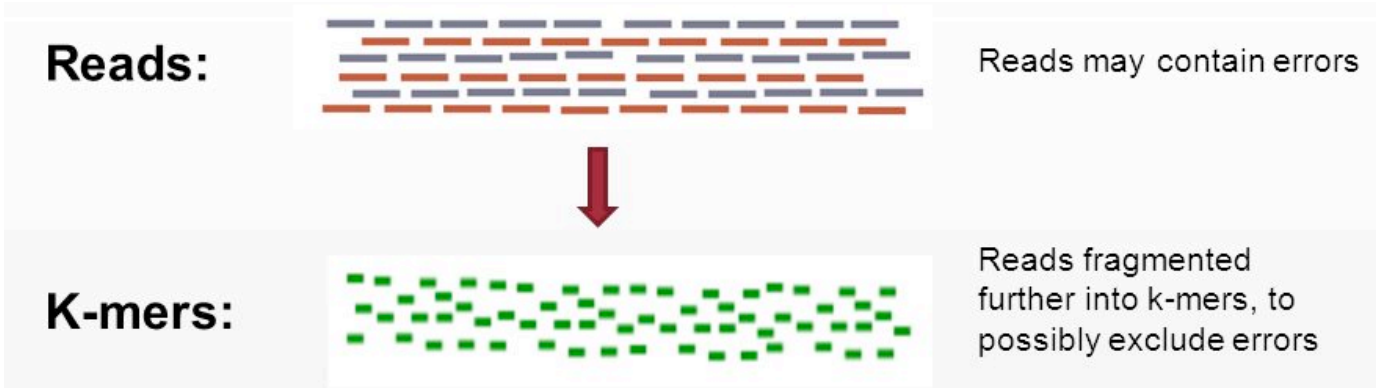
- De Bruijn graph

Basé sur une approche par *k-mers*

Les différents algorithmes

- De Bruijn graph

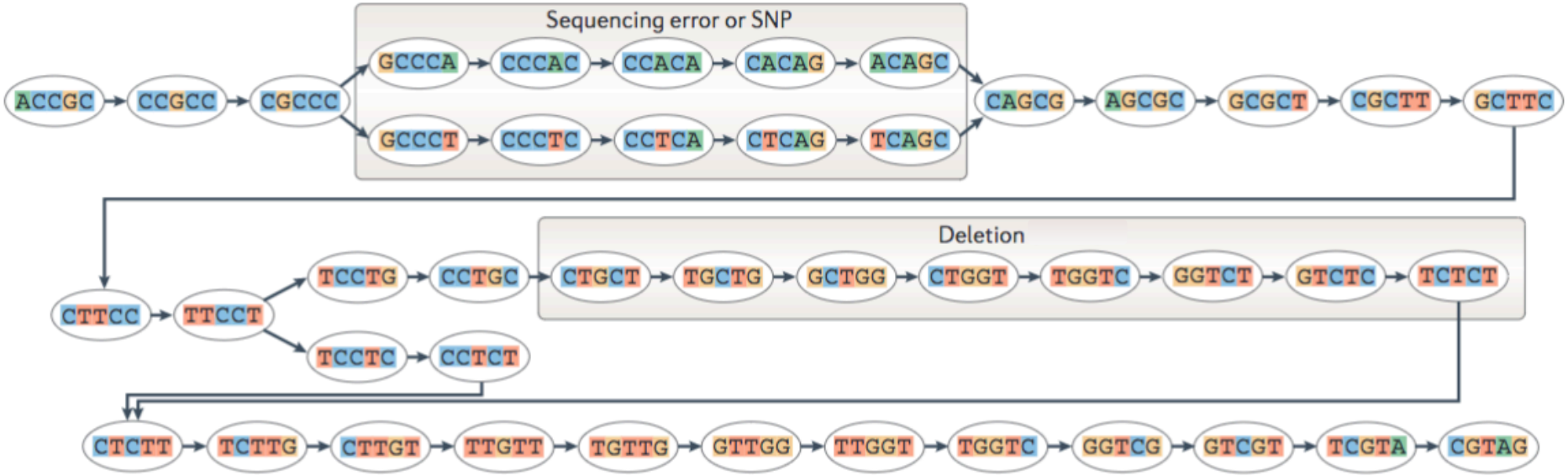
Basé sur une approche par *k-mers*



Les différents algorithmes

- De Bruijn graph

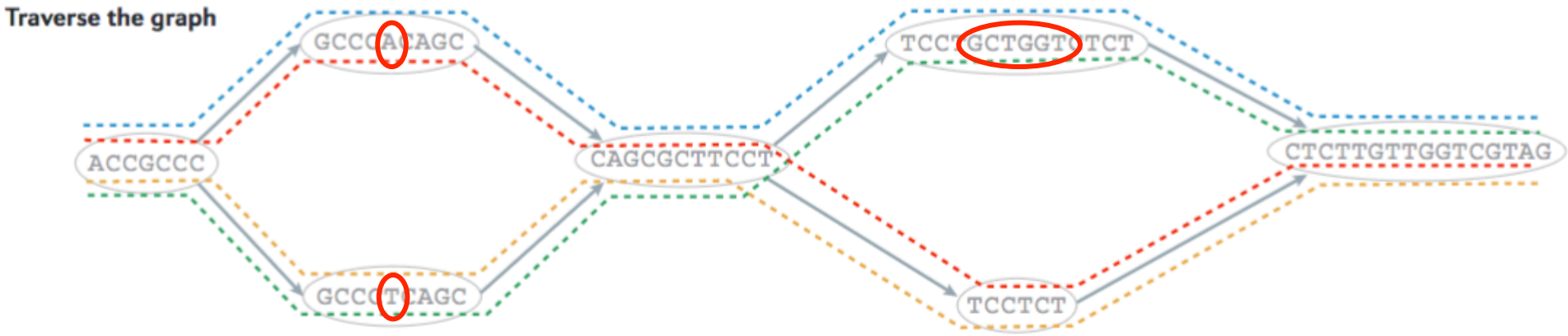
Génération du graphe par overlap des *k-mers* de *k -1*



Les différents algorithmes

- De Bruijn graph

Représentation des séquences possibles sur le graphe

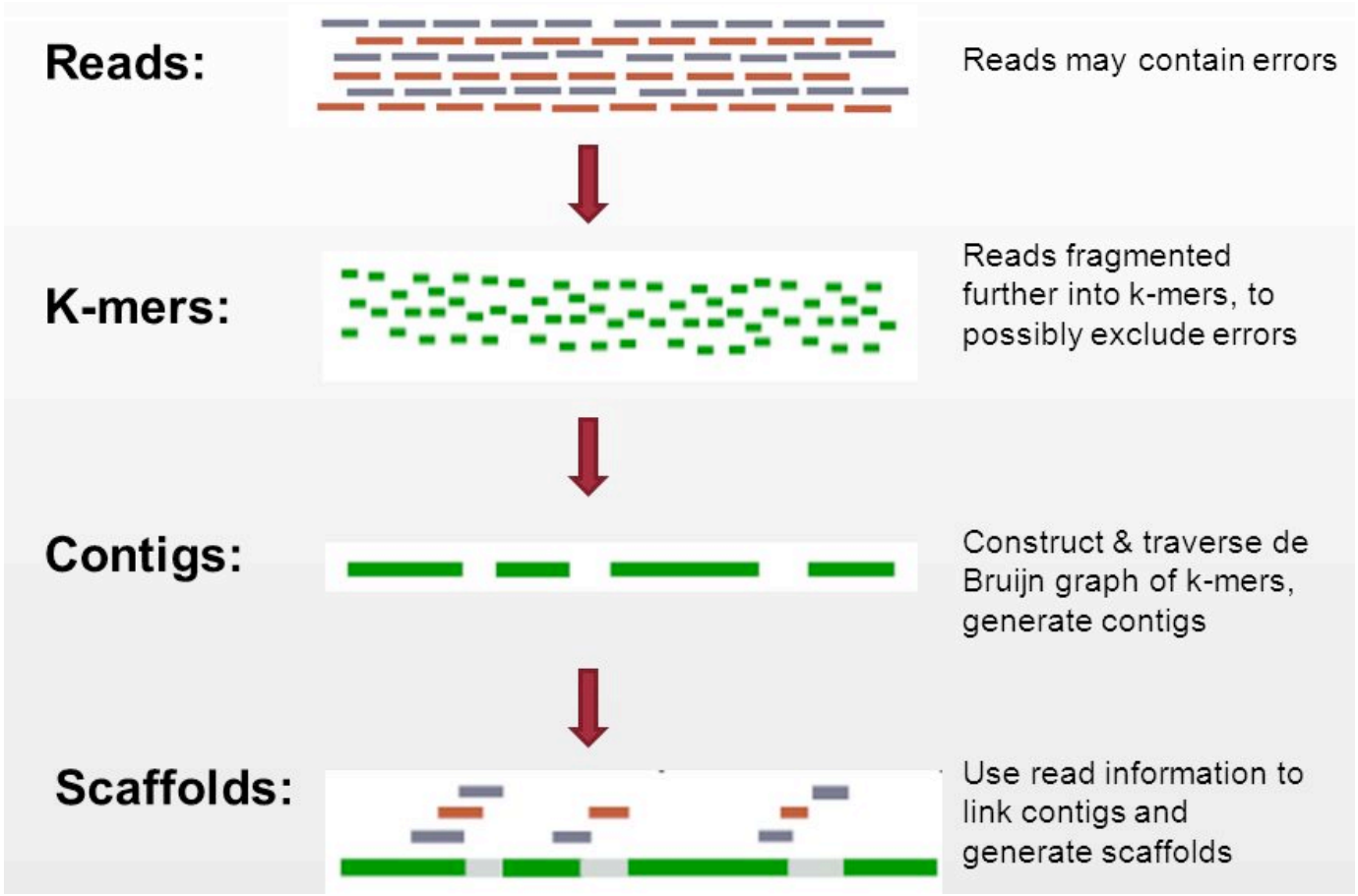


Génération des séquences « *contigs* »

Les différents algorithmes

- De Bruijn graph

Basé sur une approche par *k-mers*



Les différents algorithmes

- De Bruijn graph
- Overlap layout consensus

Algorithme également basé sur des graphes de superposition des séquences similaires (complètes). Approche plus ancienne, utilisée sur des jeux de données réduits.

DBG: plus rapide d'exécution

OLC: meilleure exécution pour les longues séquences

Les différents algorithmes

- De Bruijn graph
- Overlap layout consensus
- String graph: une variante d'OLC qui élimine les séquences inutiles.

Les différents algorithmes

- De Bruijn graph
- Overlap layout consensus
- String graph
- Greedy algorithm

Commence par joindre les reads qui se superposent le mieux pour produire des contigs

Les différents algorithmes

- De Bruijn graph
- Overlap layout consensus
- String graph
- Greedy algorithm
- Hybrid algorithm

Approches diverses qui mixent les précédents algorithmes afin de réduire le erreurs et le nombre de contigs

Les différents algorithmes

- De Bruijn graph: ABySS, Velvet, SOAPdenovo
- Overlap layout consensus: Edena
- String graph: SGA
- Greedy algorithm: SSAKE, Perga
- Hybrid algorithm: Ray

Quel assembleur choisir ?

Une vraie question,

Un vrai casse-tête !

A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective

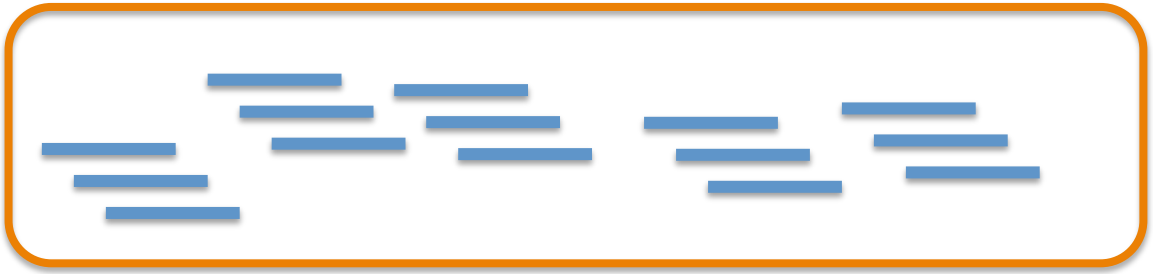
Khan et al. 2018 Evolutionary Bioinformatics

Quel assembleur choisir ?

Les critères à prendre en compte:

- Qualité de l'assemblage obtenu

1 chromosome (2 bras chromosomiques) assemblés en n scaffolds



Quel assembleur choisir ?

Les critères à prendre en compte:

- Qualité de l'assemblage obtenu
- Temps et puissance de calcul : différences non négligeables !!

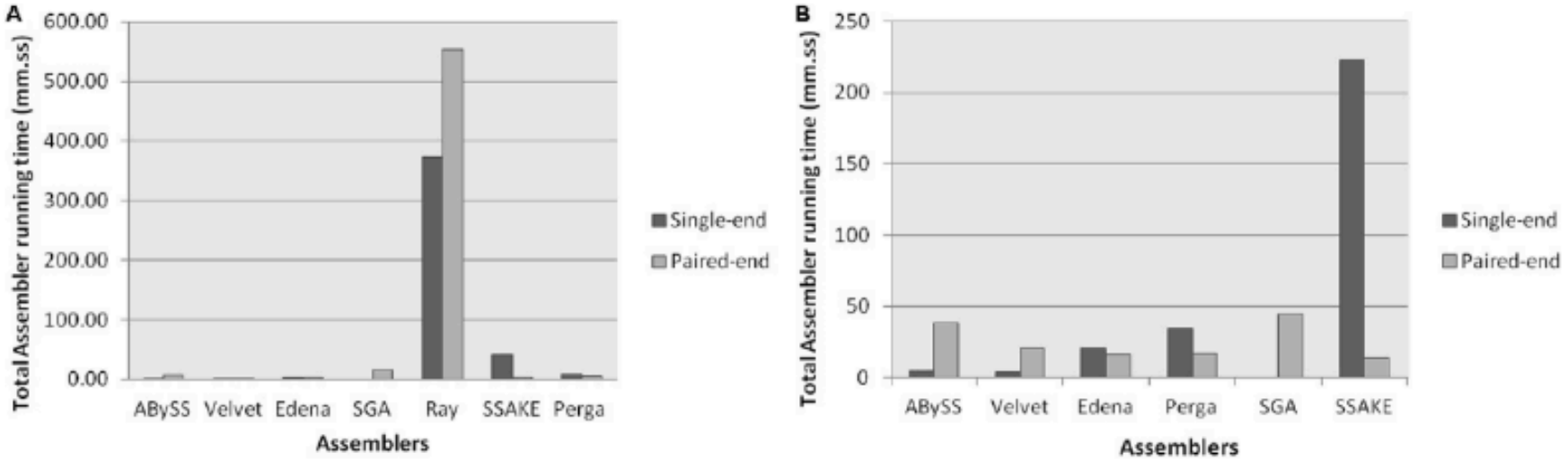
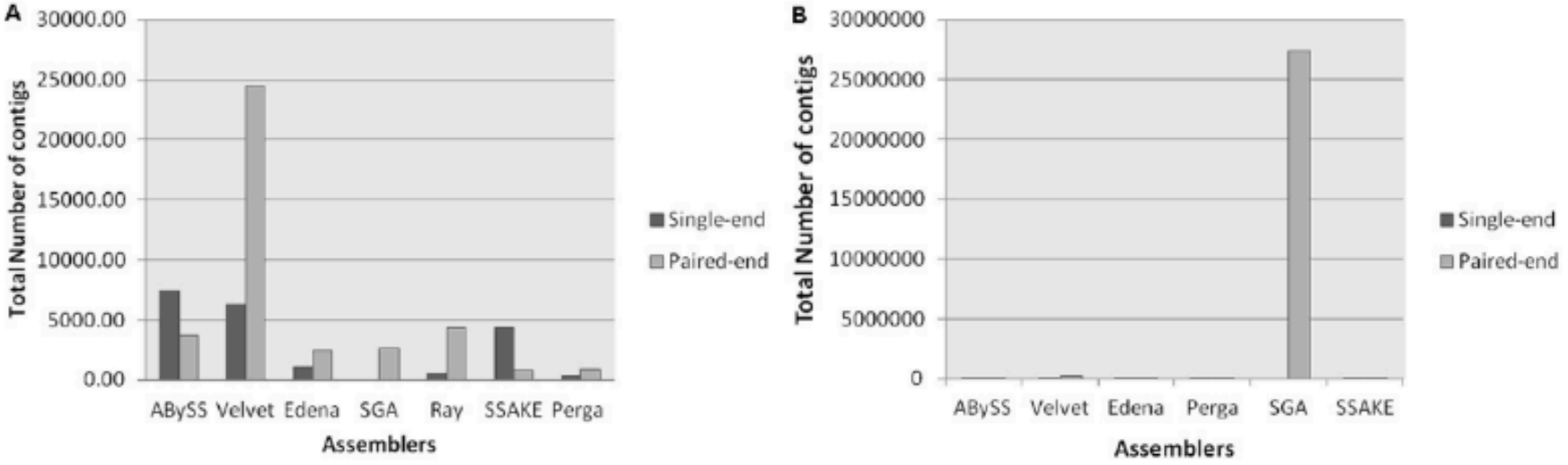


Figure 1. The comparison of total median assembling time of each assembler for (A) paired-end and single-end prokaryotic data sets and (B) paired-end and single-end eukaryotic data sets.

Quel assembleur choisir ?

Les critères à prendre en compte:

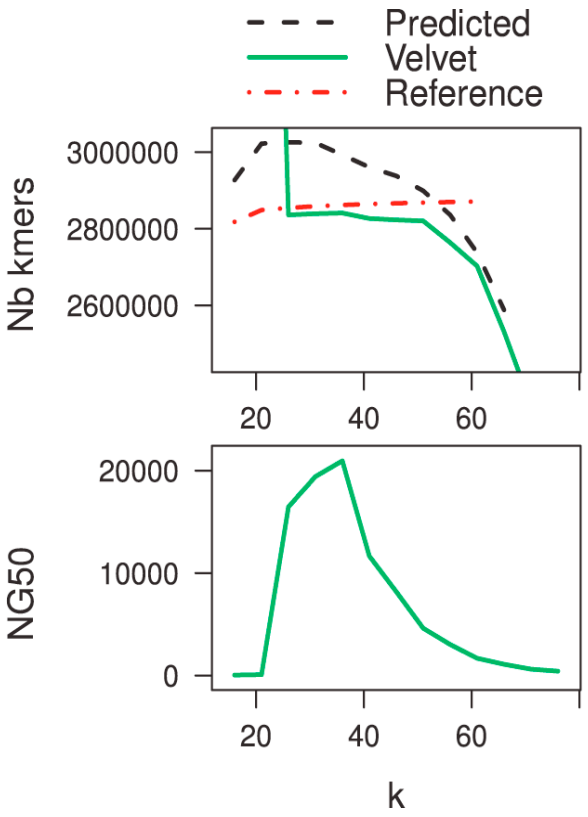
- Qualité de l'assemblage obtenu
- Temps et puissance de calcul : différences non négligeables Justifiées??



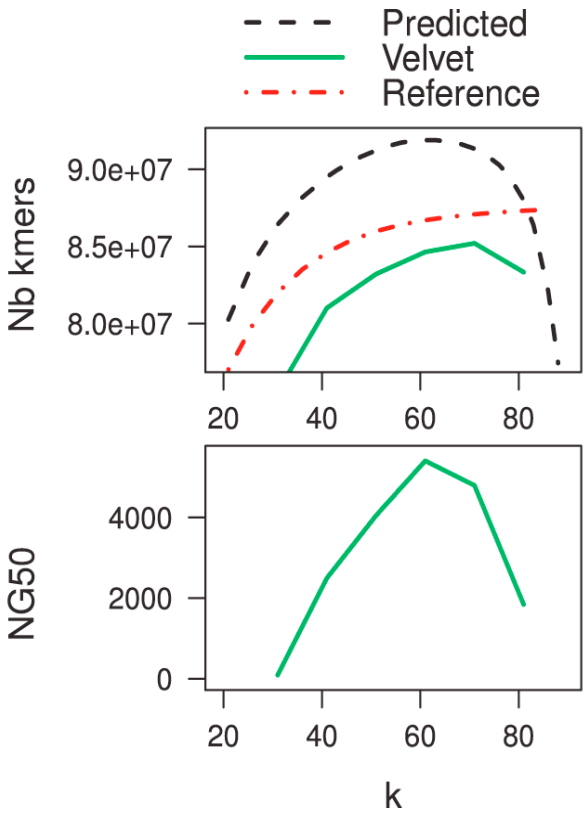
Si DBG approach

Il faut choisir une taille ou tester différentes tailles de *k-mers.*

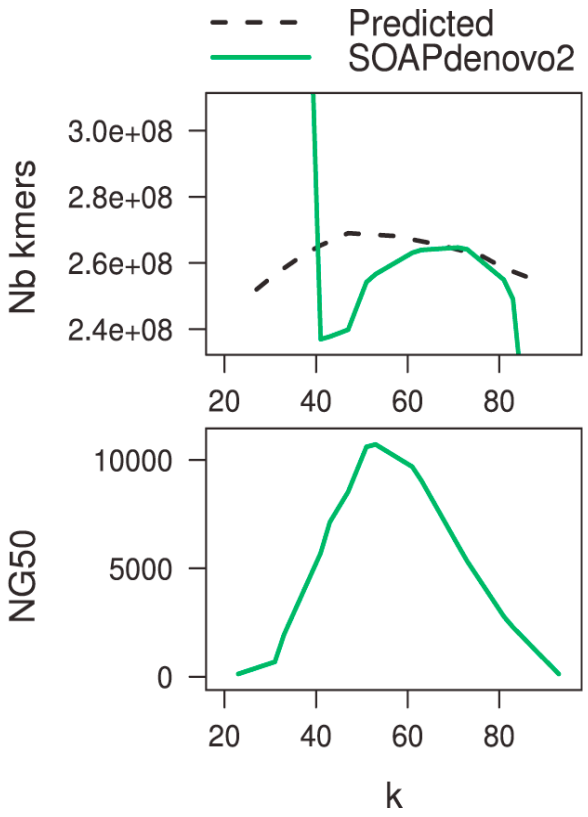
Effet de la taille de k-mer size



S. aureus



chr14



B. impatiens

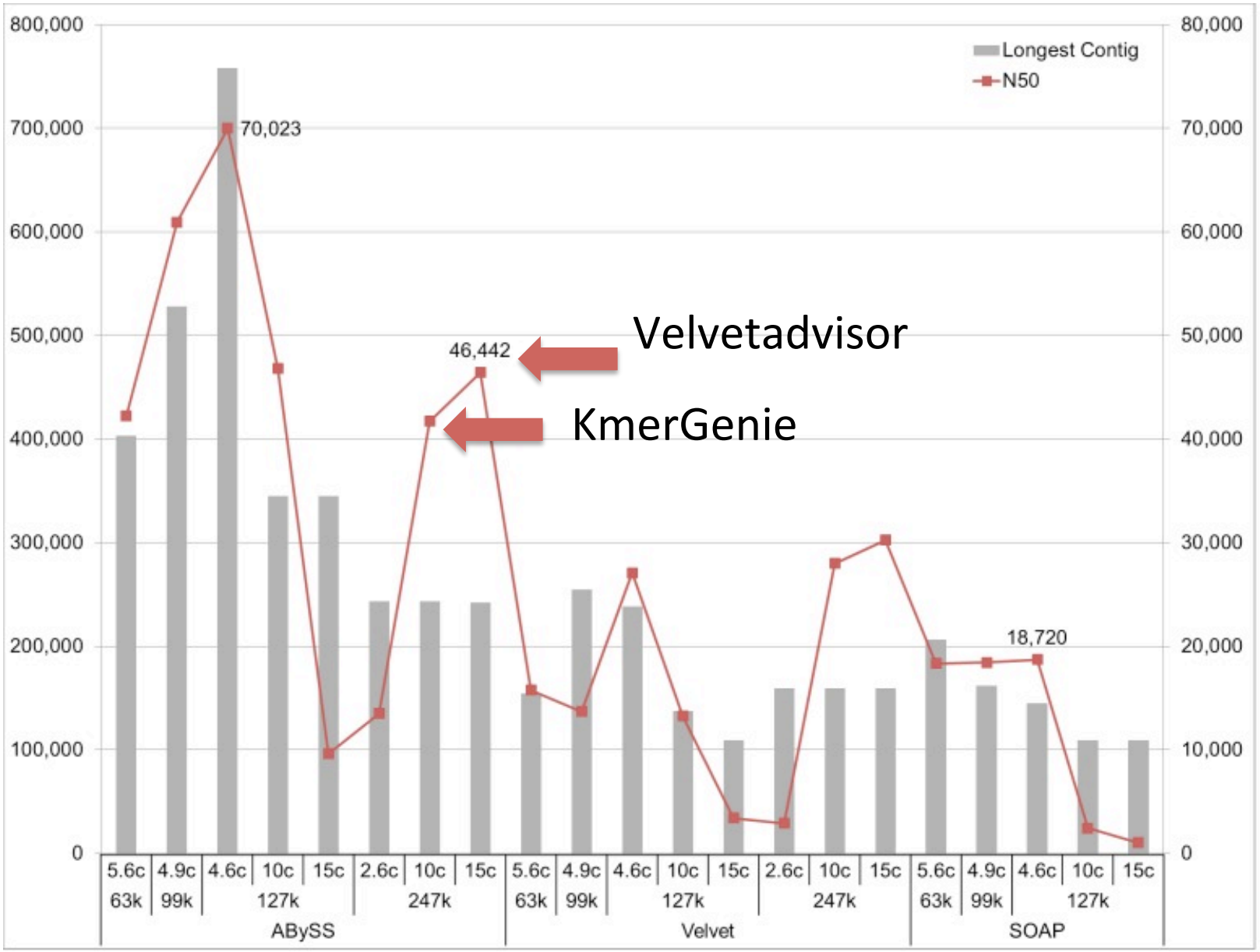
Si DBG approach

Il faut choisir une taille ou tester différentes tailles de *k-mers.*

KmerGenie, Velvetadvisor but see

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5237644/

Effet de la taille de k-mer size



La profondeur de séquençage est souvent hétérogène entre régions d'un même génome, ou si plusieurs génomes sont présents (méta-génomique).

La profondeur de séquençage est souvent hétérogène entre régions d'un même génome, ou si plusieurs génomes sont présents (méta-génomique).

La violation de l'hypothèse de profondeur de séquençage homogène est mal gérée par la plupart des assembleurs => problème pour reconstruire de longs contigs.

Un petit k entraine un excès de branches dans le DBG.

Un grand k entraine un excès de gaps.

Alternative intéressante

IDBA

Approche basée sur DBG, qui teste itérativement plusieurs tailles de *k-mers,* adapte le *k-mer* selon la profondeur de séquençage localement et propose un assemblage optimisé en contigs.

Article: Peng et al. 2012.

Alternative intéressante

IDBA

Approche basée sur DBG, qui teste itérativement plusieurs tailles de *k-mers* et propose un assemblage optimisé en contigs.

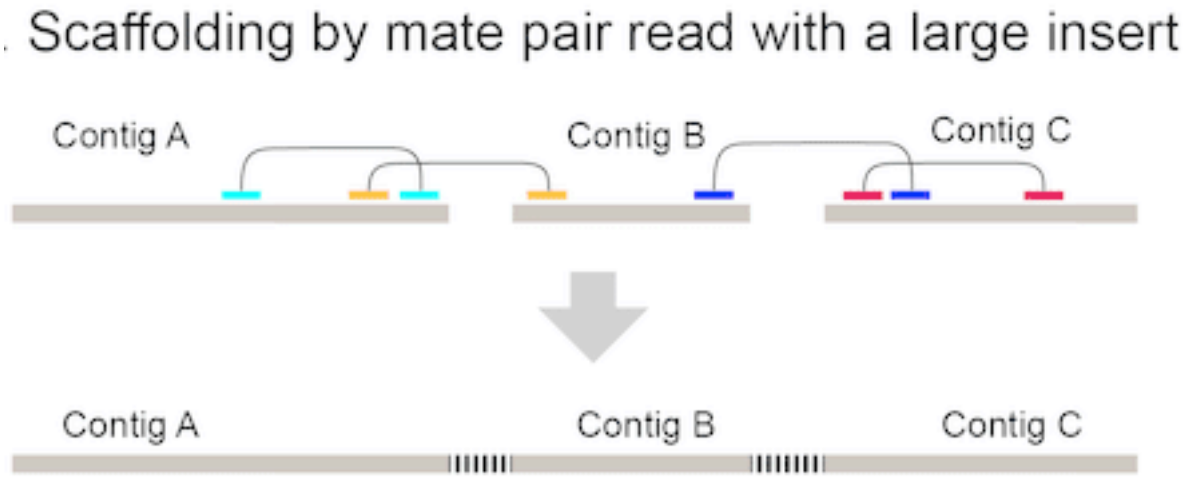
Donne de très bons résultats

Rapide

Mon conseil: comparer l'assemblage IDBA avec celui d'une ou deux autres méthodes

Possibilité d'utiliser plusieurs librairies pour assembler les contigs en scaffolds

1. Utilisation de librairies mate pair avec des « grandes » tailles d'insert



ALLPATH-LG, SSPACE

Possibilité d'utiliser plusieurs librairies pour assembler les contigs en scaffolds

2. Utilisation de librairies « longs reads »

Plusieurs outils développés récemment:

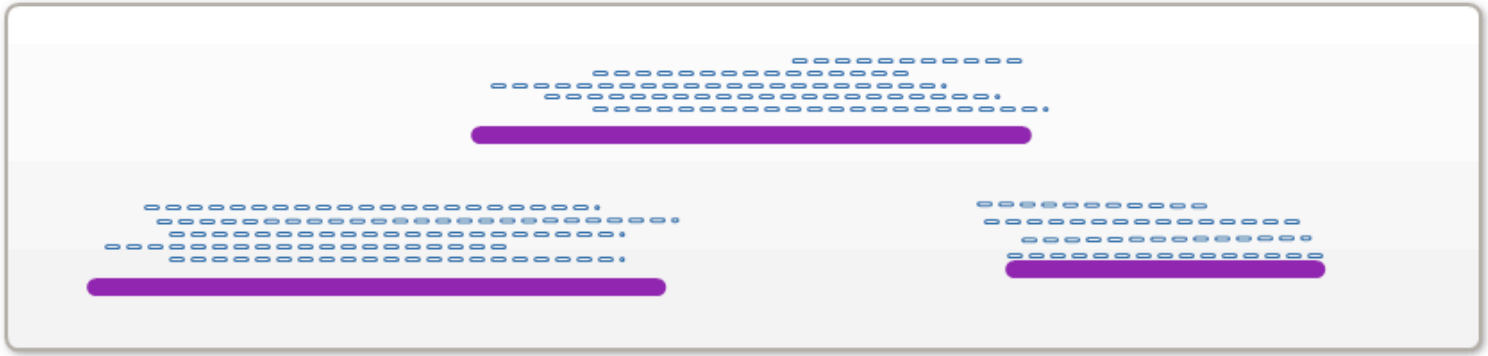
ALLPATHS pipeline maps uncorrected long reads to the graph of an assembly [26]

SSPACE-LongRead maps long reads to contigs assembled from short reads [27]

Alpaca (ALLPATHS and Celera Assembler)

Cas des long reads

- **Long reads-only *de novo* assembly**. Using **just** long reads from a long insert library, the reads are often preprocessed before being assembled using an Overlap-Layout-Consensus algorithm. The best known implementation of this is **HGAP.**



Cas des long reads

- **Long reads-only *de novo* assembly**.

- **Hybrid *de novo* assembly**. Using a **combination** of long and short read data, the reads are used together during assembly to generate a hybrid assembly.

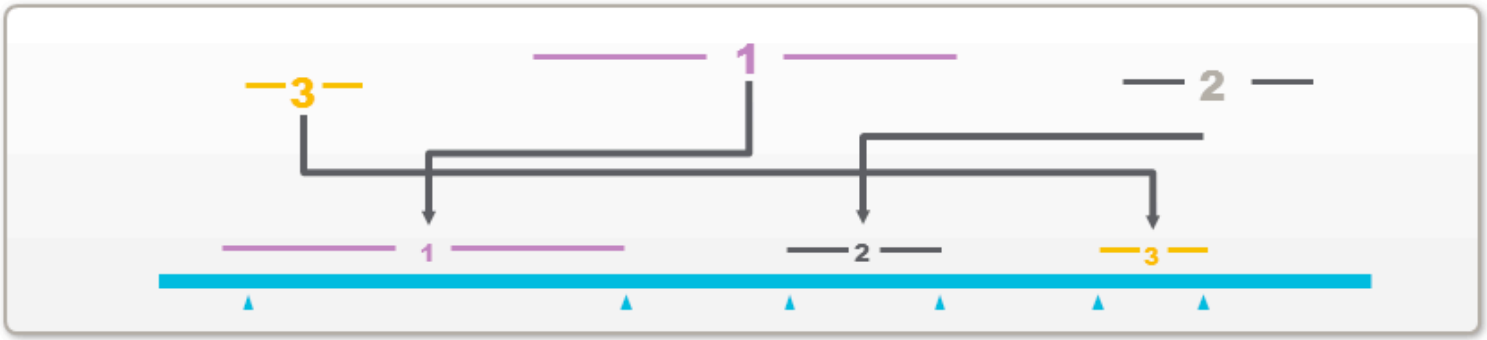
Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm
https://genome.cshlp.org/content/early/2017/01/27/gr.213405.116.abstract

Cas des long reads

- **Long reads-only *de novo* assembly**.
- **Hybrid *de novo* assembly**.
- **Gap filling**. Starting with an existing mate-pair based assembly, the internal gaps (consisting of Ns) inside the scaffolds are filled using long sequences.

Cas des long reads

- **Long reads-only *de novo* assembly**.
- **Hybrid *de novo* assembly**.
- **Gap filling**.
- **Scaffolding**. long reads are used to join contigs of an existing assembly (based on short read data)



Today, assembly a genome using long reads from Oxford Nanopore Technologies is really interesting in particular to solve repeats and structural variants in prokaryotic as well as in eukaryotic genomes. Assemblies are increasing contiguity and accuracy.

The daily increase of data sequences obtained and the fact that more and more tools are being released or updated every week, many species are having their genomes assembled and that's is great ...

"*But which assembly tool could give the best results for your favorite organism?*"

CulebrONT can help you! CulebrONT is an open-source, scalable, modular and traceable Snakemake pipeline, able to launch multiple assembly tools in parallel, giving you the possibility of circularise, polish, and correct assemblies, checking quality. CulebrONT can help to choose the best assembly between all possibilities.

About CulebrONT

- From assembly to correction

https://culebront-pipeline.readthedocs.io/en/latest/ABOUT/

Attention !! les long reads (PacBio ou MinIon) ont des taux d'erreur élevés (6-7%).

$\Rightarrow$ Correction des lectures nécessaires :

1. hybrid strategy, using two sets of reads:
the reference read set, whose error rate is assumed to be small (Illumina)
the long read set, which is then corrected using the reference set.
LoRDEC : http://www.atgc-montpellier.fr/lordec/

2. High coverage

3. 2D reads MinIon

Comparer des assemblages

Encore un casse-tête

Assemblathon2 (2013)

- 3 espèces (1 oiseau, 1 poisson, 1 serpent)
- 43 assemblages
- 21 équipes participantes

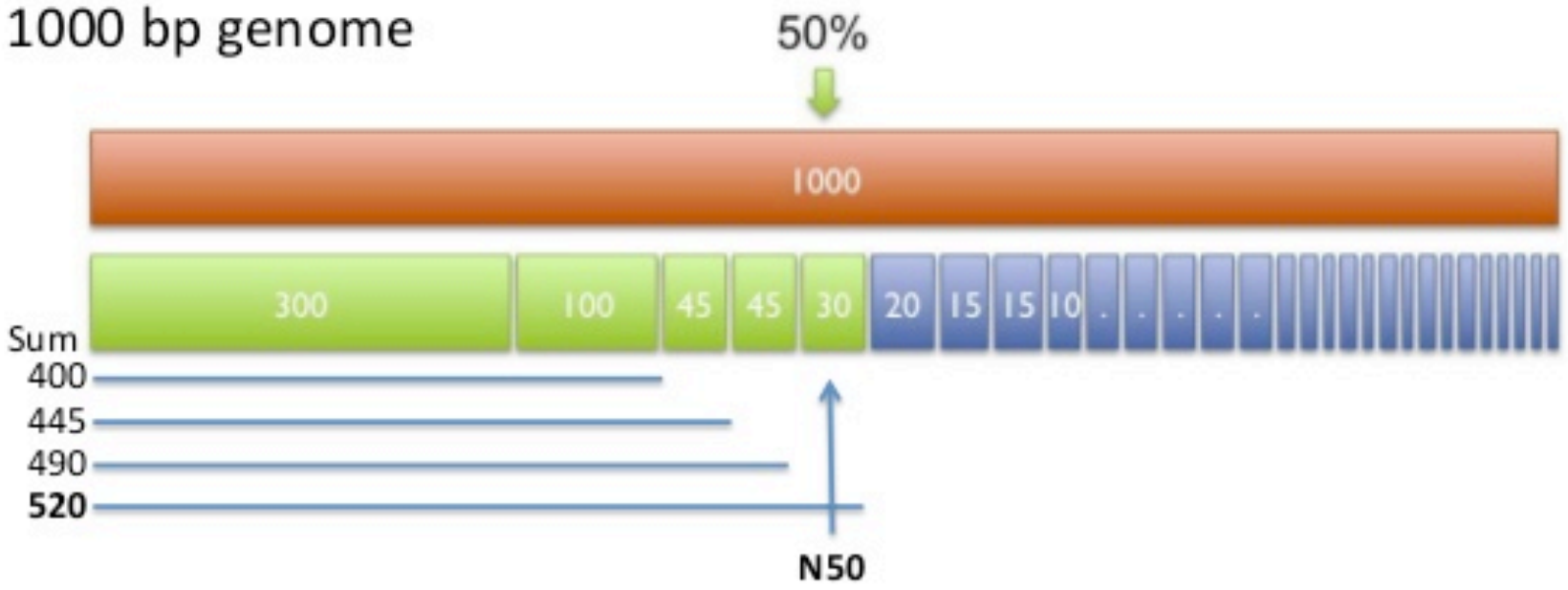
Etablissement de plusieurs métriques (en plus du N50):

NG50

OrthoDB genes found (Busco)

N50

50% of the genome is in contigs as large as the N50 value



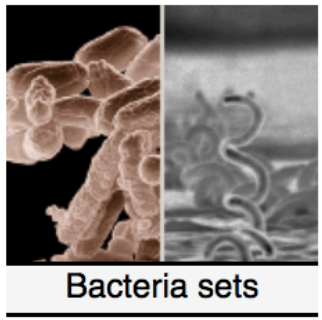
Courtesy of Michael Schatz, CSHL

Busco / Augustus

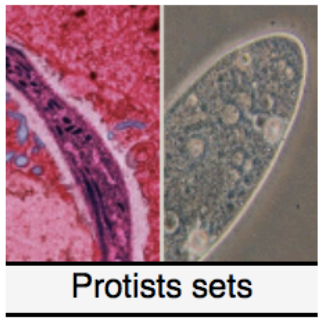
Principe:

- *de novo* gene annotation (augustus)
- Recherche de similarité de séquence avec des « core » gènes (hmmer)

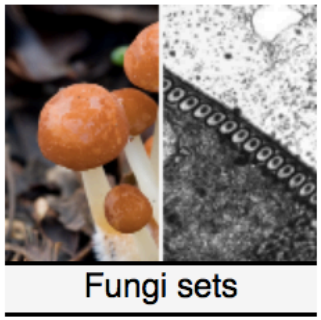
=> orthoDB datasets available on busco website


Bacteria sets


Eukaryota sets


Protists sets

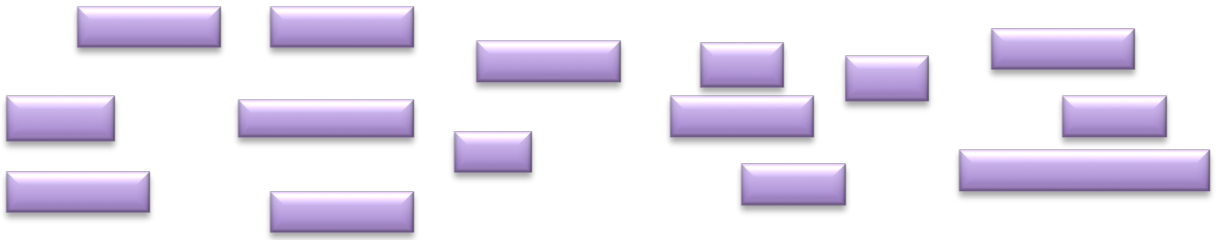

Metazoa sets


Fungi sets


Plants set

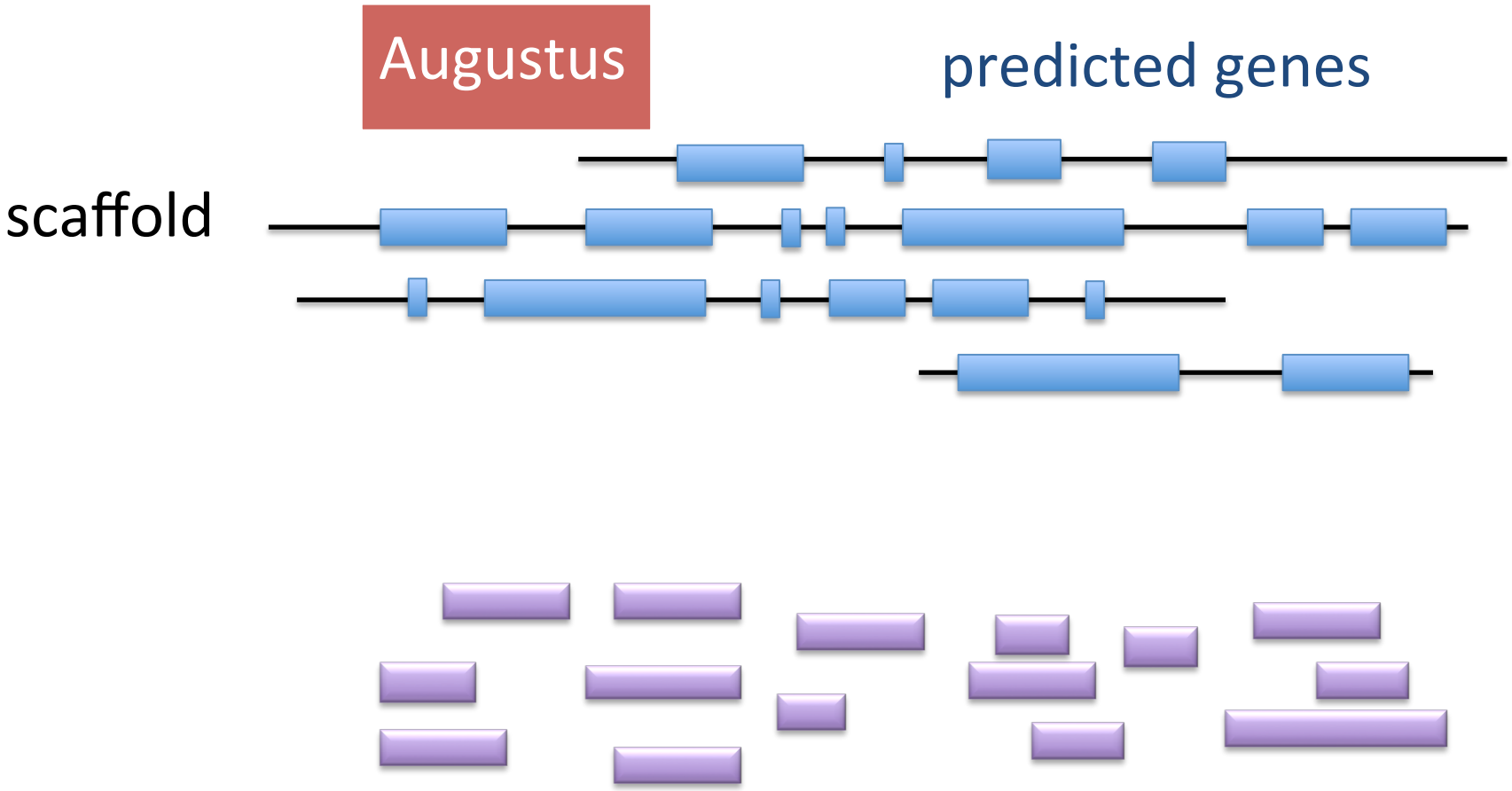
BUSCO

scaffolds

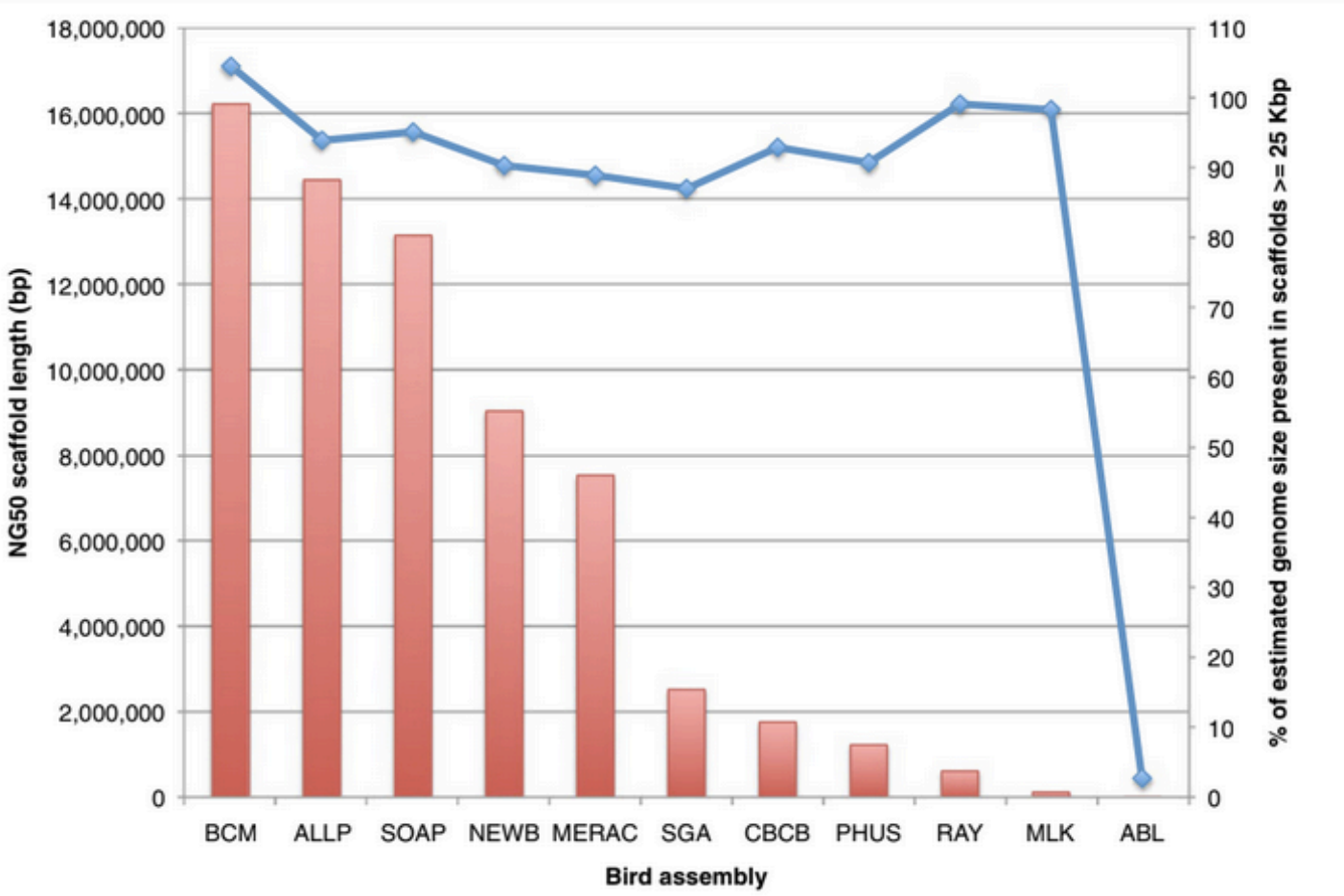


set of « core » genes (orthoDB)

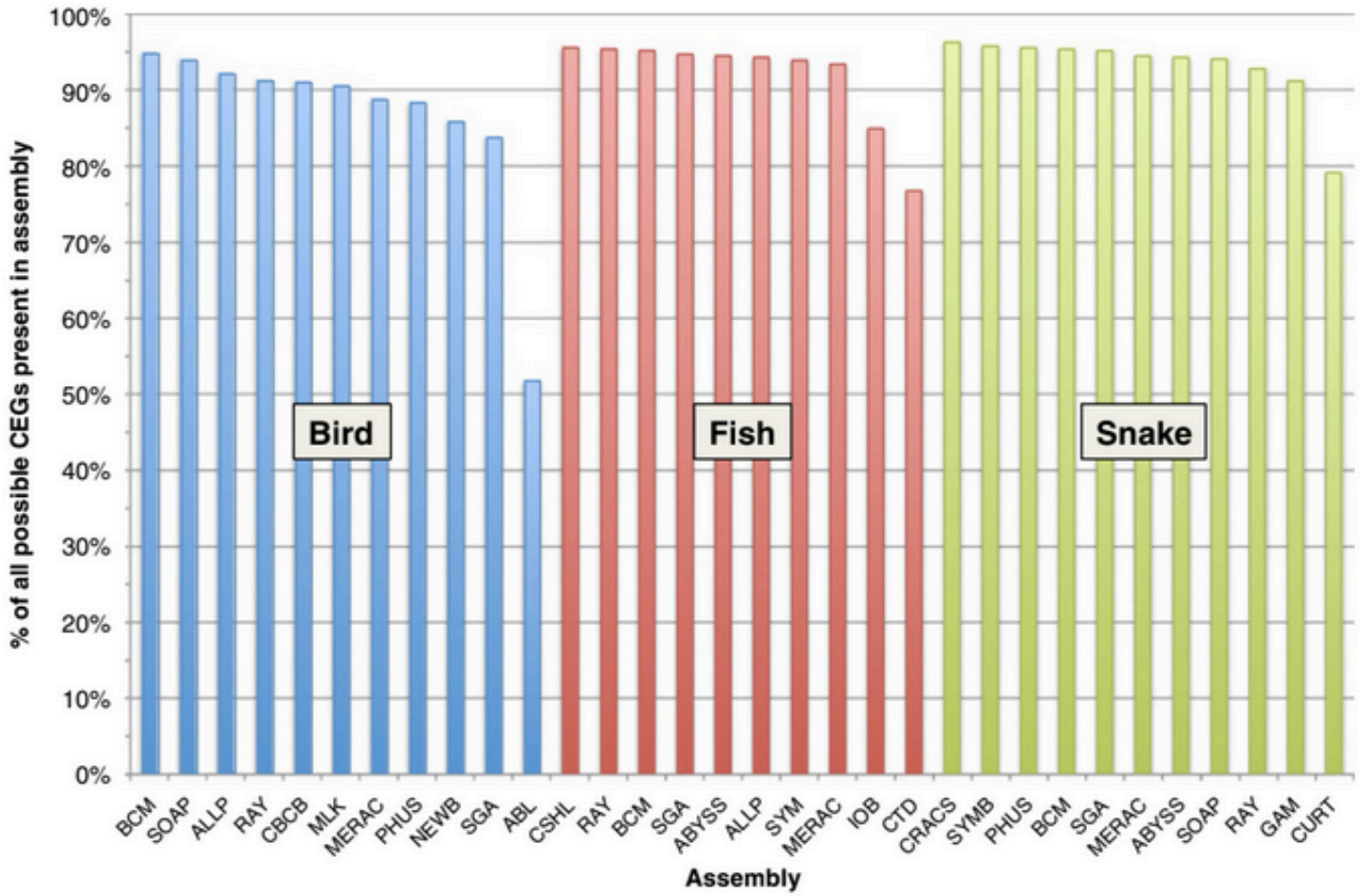
BUSCO

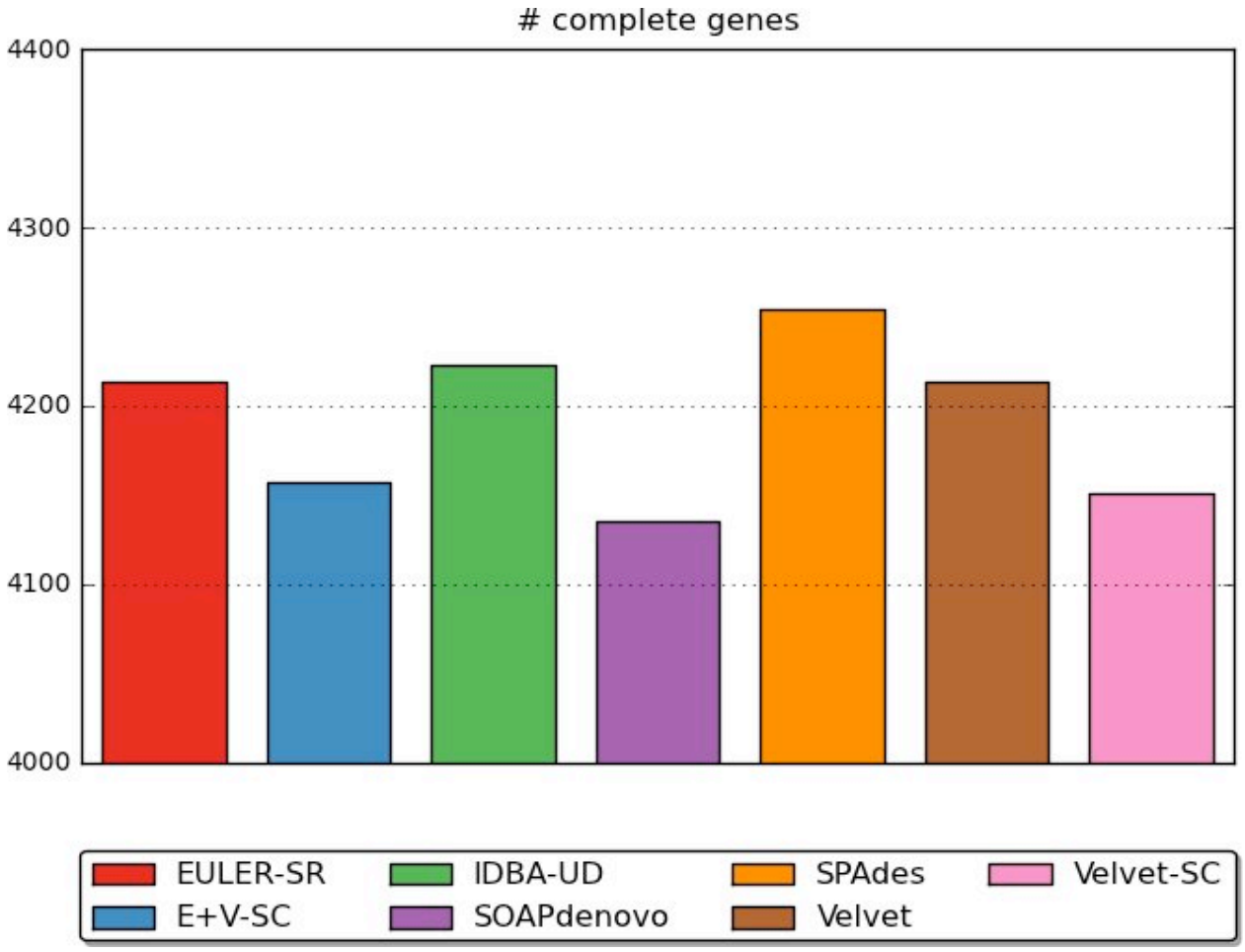

Augustus
predicted genes
scaffold
set of « core » genes (orthoDB)

Assemblathon2 (2013)

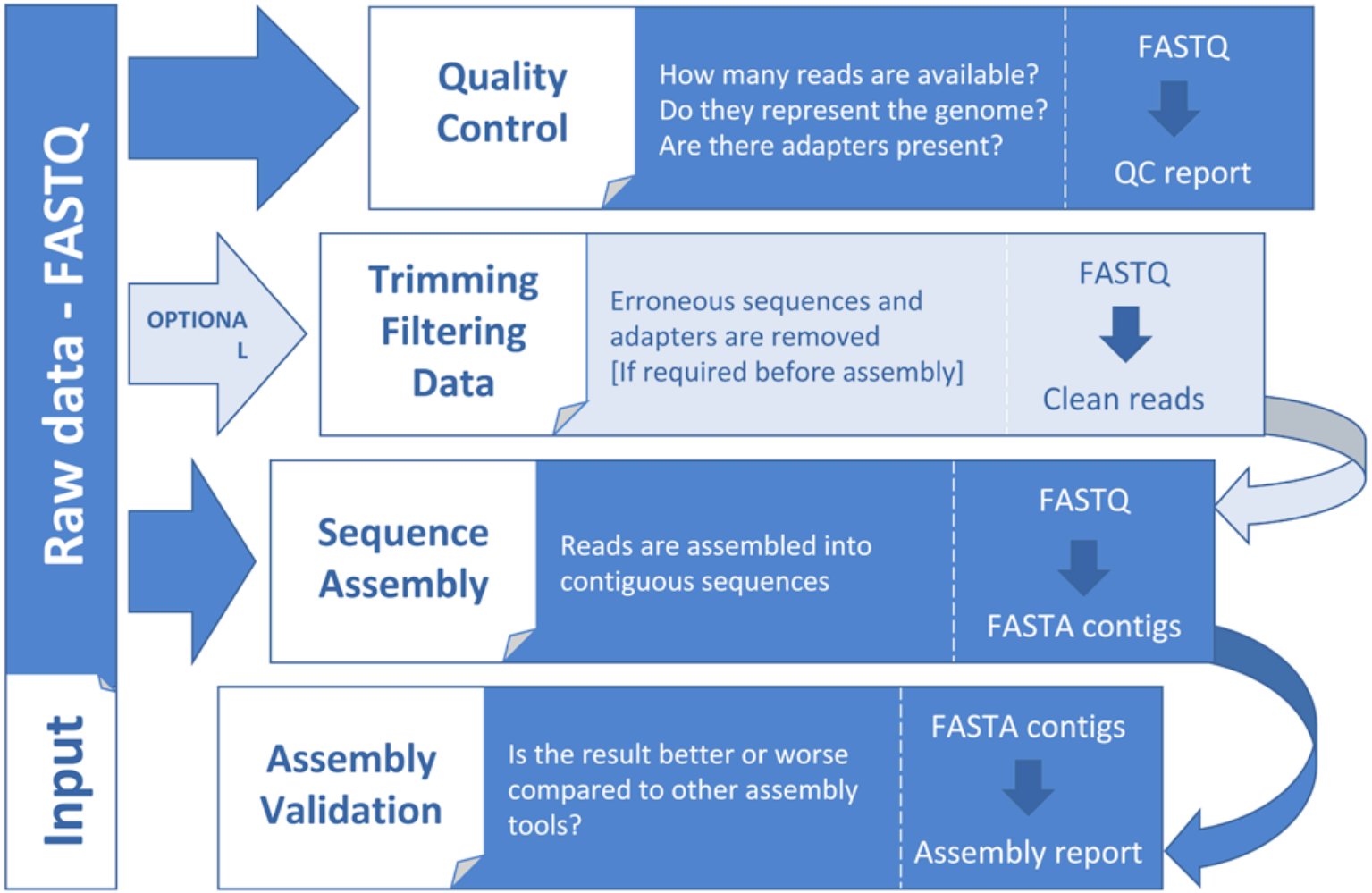


Assemblathon2 (2013)




complete genes
4400
4300
4200
4100
4000
EULER-SR
E+V-SC
IDBA-UD
SOAPdenovo
SPAdes
Velvet
Velvet-SC

General workflow pour assemblage



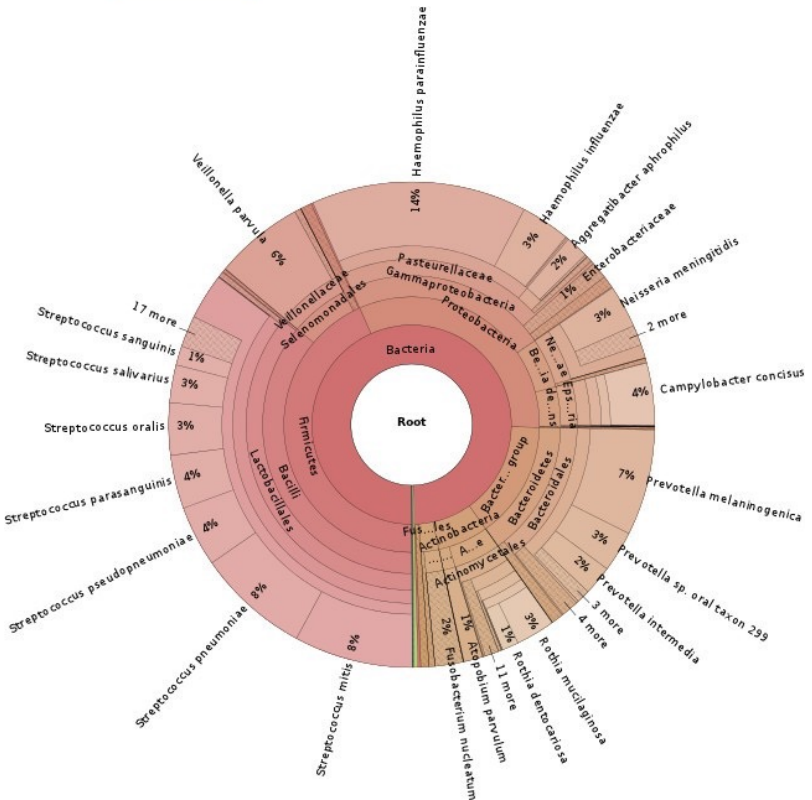
Quelques outils

Method | Open Access | Published: 03 March 2014

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood & Steven L Salzberg

Genome Biology **15**, Article number: R46 (2014) | Cite this article



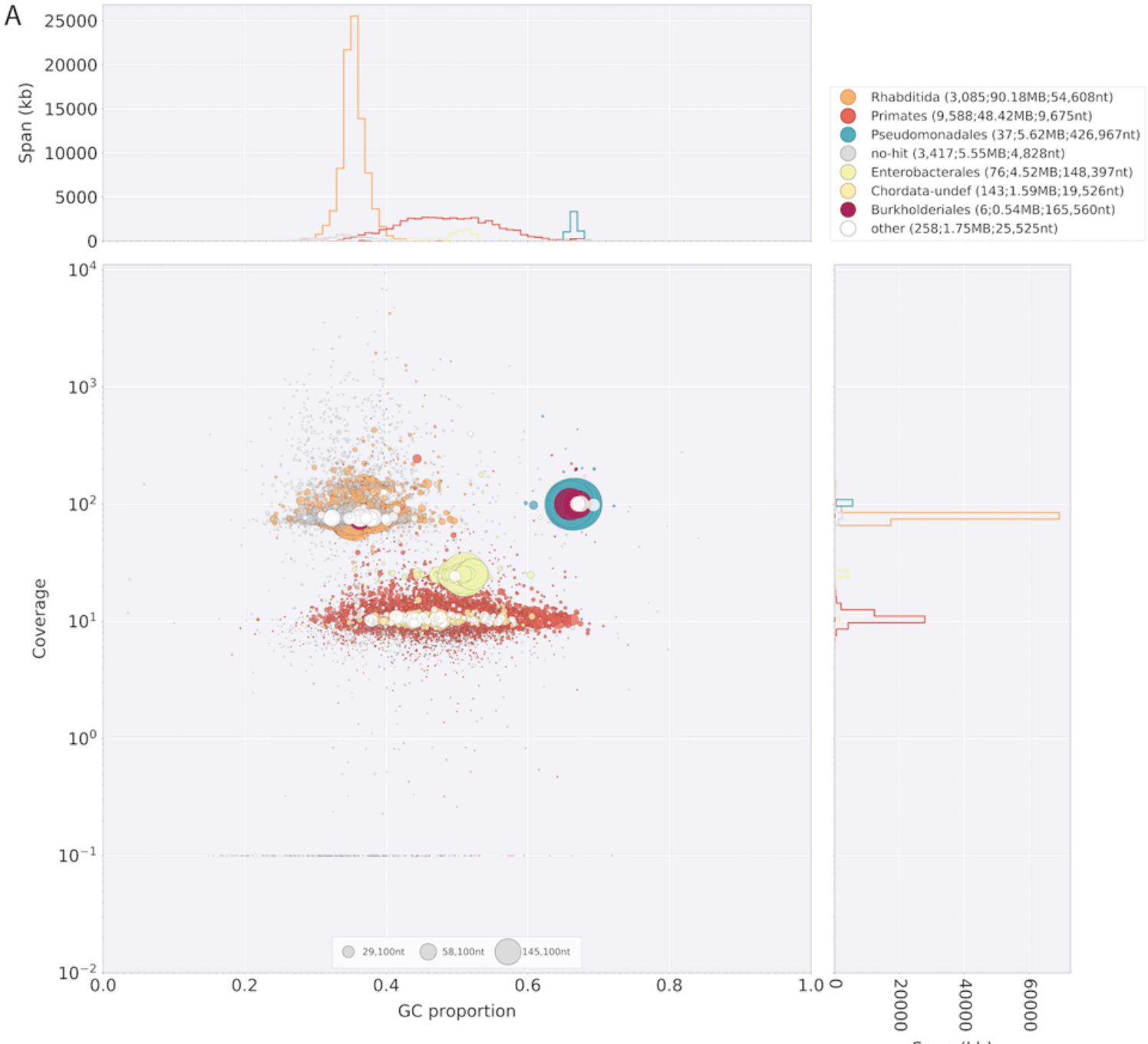
Taxonomic distribution of saliva microbiome reads classified by Kraken

Blobtools

Bowtie2

+

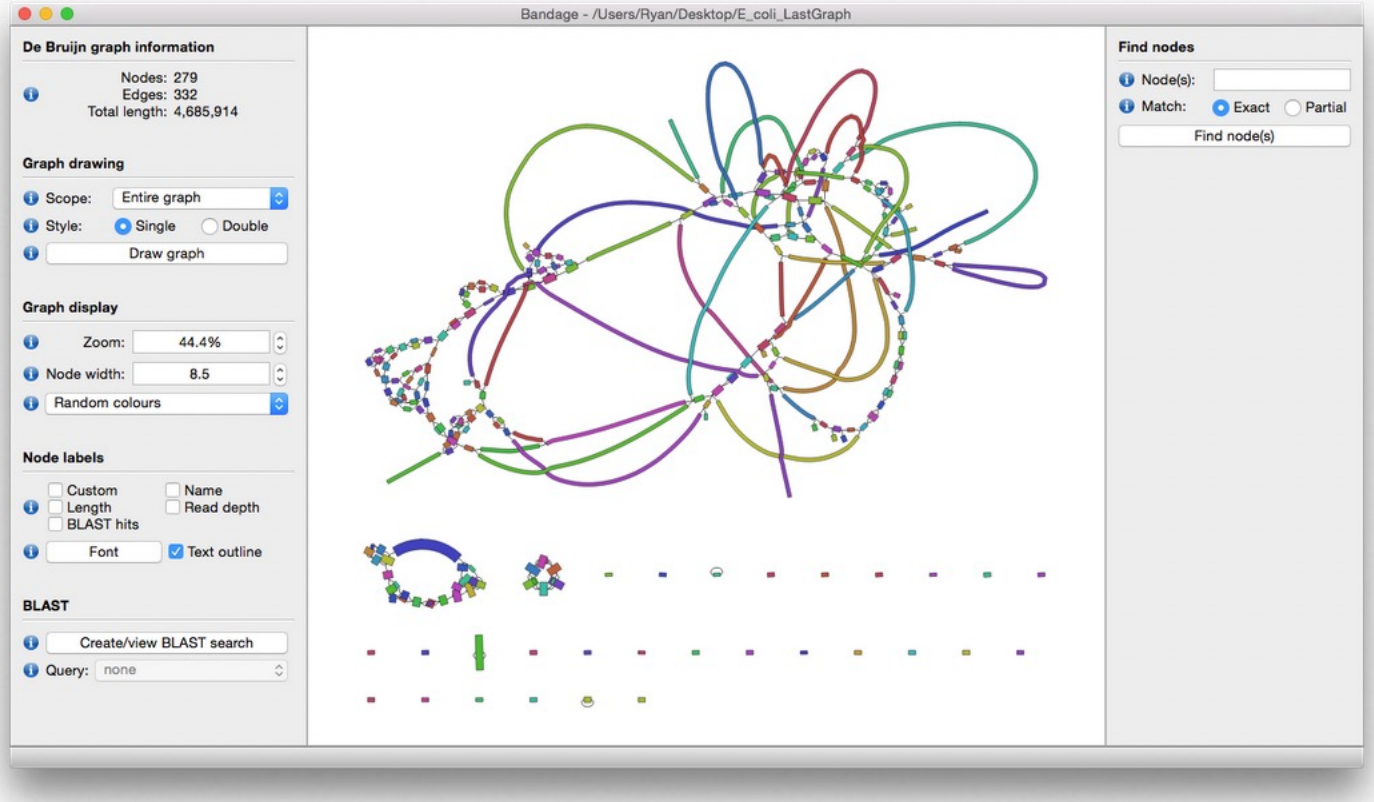
ncbi_blast



Bandage

a Bioinformatics Application for Navigating *De novo* Assembly Graphs Easily

Bandage is a program for visualising *de novo* assembly graphs. By displaying connections which are not present in the contigs file, Bandage opens up new possibilities for analysing *de novo* assemblies.



https://rrwick.github.io/Bandage/