

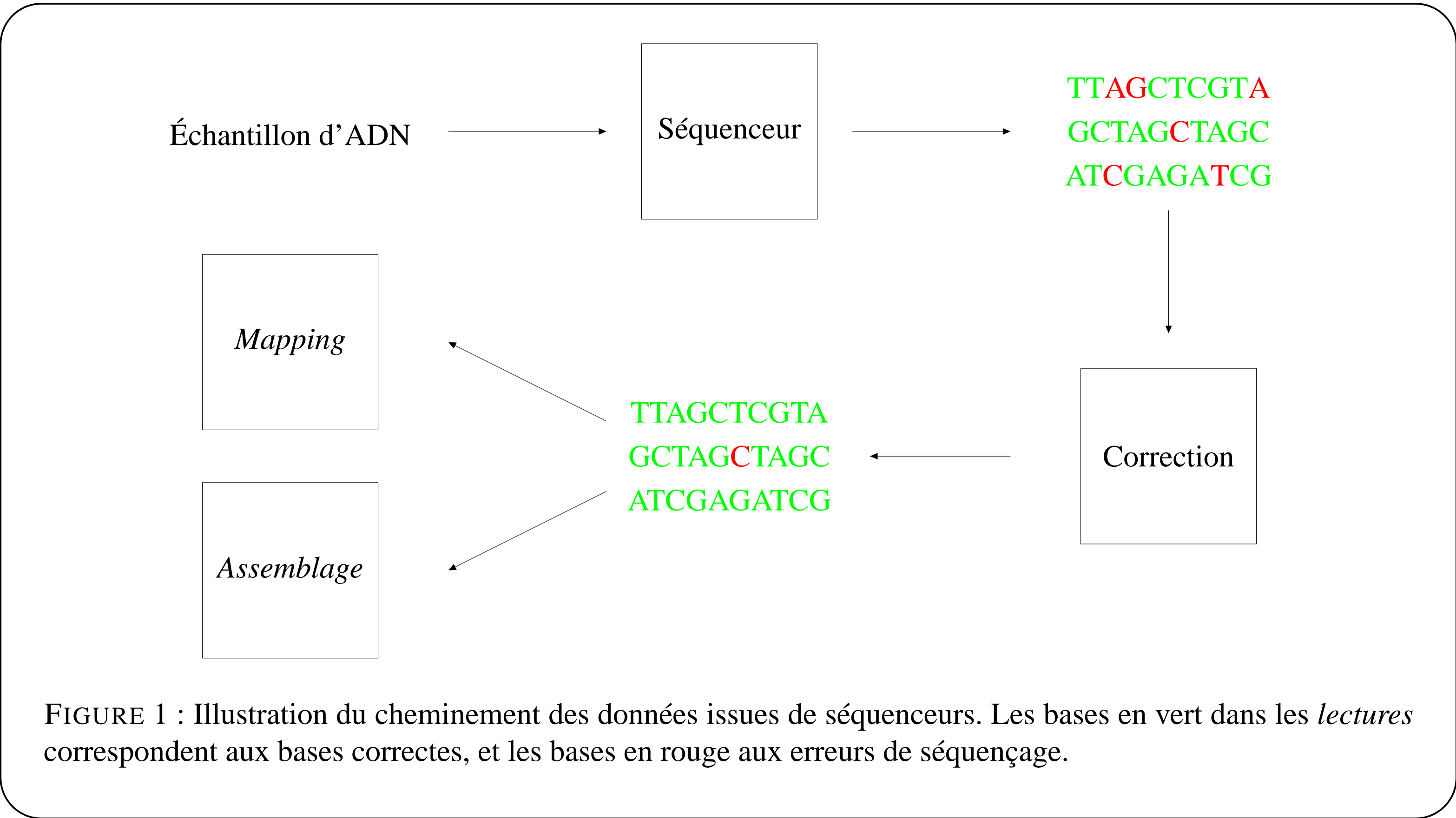
Pierre Morisse
Doctorat en Informatique 1^o année
Spécialité Bioinformatique

Introduction

Depuis le milieu des années 2000, les séquenceurs à très haut débit se développent et permettent de séquencer l'ADN d'un individu sous forme de courtes séquences appelées *lectures*, utilisées pour résoudre divers problèmes de génomique, notamment de *mapping* et d'*assemblage*.



Lors de du séquençage de l'ADN, les séquenceurs peuvent introduire des erreurs dans les lectures produites. Nous appelons ces erreurs des erreurs de séquençage, et il est donc souvent nécessaire de faire subir aux *lectures* une procédure de correction avant de les utiliser, afin d'améliorer leur précision.

Les séquenceurs produisant des millions de *lectures*, il est nécessaire de développer des outils informatiques adaptés au traitement de telles quantités de données, afin de permettre la résolution des différents problèmes.



Évolution des séquenceurs

Les séquenceurs se développent et évoluent très rapidement. Ils deviennent moins imposants, moins coûteux, et visent ainsi à être plus accessibles au grand public. Ils produisent également des *lectures* de plus en plus longues, très utiles pour résoudre des problèmes complexes, bien que peu précises.

2007 - Illumina HiSeq	2015 - Oxford Nanopore MinION
	
1 milliard de lectures	Environ 70 000 lectures
Lectures courtes (100 - 150)	Lectures longues (> 10 000)
< 1% d'erreurs	30% d'erreurs
2 - 11 jours	48 heures
470 000 \$	1 000 \$

Principaux objectifs

Côté informatique :

- Développer des structures de données permettant de stocker et de traiter les grandes quantités de données formées par les *lectures*
- Développer des outils permettant aux biologistes de manipuler facilement les *lectures*

Côté biologie :

- Analyser les *lectures* afin détecter des mutations dans l'ADN d'un individu, et ainsi d'éventuelles pathologies
- Générer de nouveaux génomes de référence à partir des *lectures*, afin de les utiliser par la suite dans d'autres problèmes

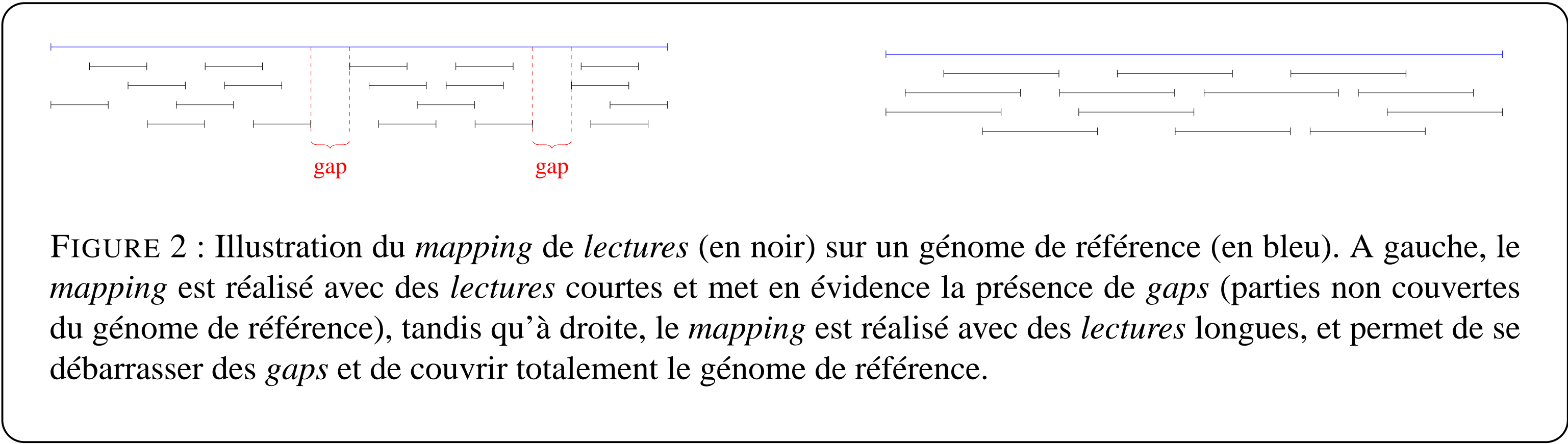
3 principaux problèmes

1. Correction :

- Présence d'erreurs de séquençage dans les *lectures*, très nombreuses dans les *lectures* longues
- Nécessité de réduire le taux d'erreur afin d'améliorer la précision des *lectures*, et de faciliter leur utilisation
- Différentes approches (Comparaison des *lectures* entre elles, analyse des *facteurs* des lectures, etc)

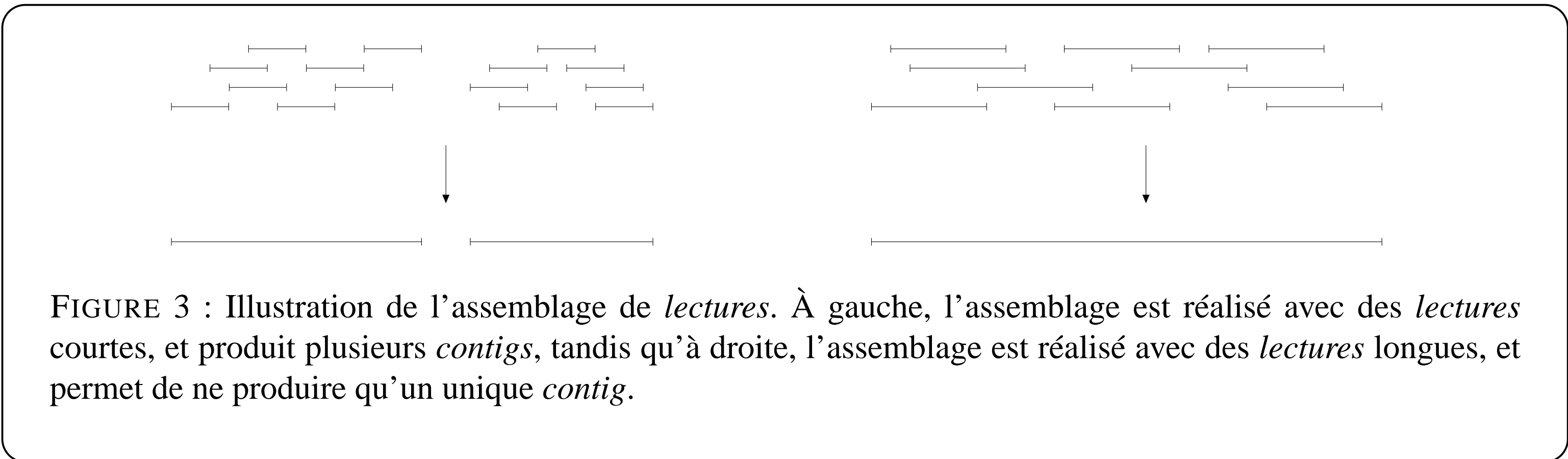
2. Mapping :

- Aligner les *lectures* séquencées sur un génome de référence
- Comparer l'ADN d'un individu à l'ADN du génome de référence
- Détecter des mutations et d'éventuelles pathologies



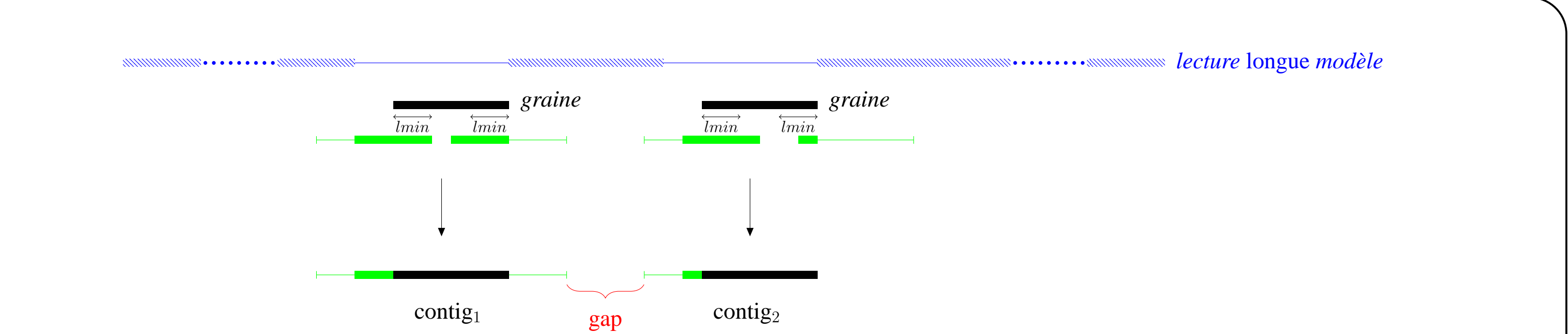
3. Assemblage :

- Aligner les *lectures* entre elles afin de trouver des chevauchements
- Assembler les *lectures* se chevauchant afin de créer des *contigs*
- Reconstruire ainsi le génome dont les *lectures* sont originellement issues



Mon travail actuel

Mon travail actuel porte sur le développement d'une méthode de correction de *lectures* longues. Plus précisément, de production de *lectures* longues dites *synthétiques*, car obtenues à partir d'un assemblage de *lectures* courtes. Pour cela, des *lectures* courtes sont alignées sur une *lecture* longue servant de *modèle*. Les *lectures* courtes totalement alignées servent alors de *graines*, et sont étendues à l'aide de *lectures* courtes partiellement alignées.



Une fois ces *contigs* obtenus, le problème restant est alors le remplissage des *gaps*. Il existe pour cela différentes méthodes, notamment l'alignement de toutes les *lectures* courtes entre elles afin de les assembler et ainsi produire de nouveaux *contigs*. De telles méthodes se montrent cependant très coûteuses en terme de temps, et au vu du grand nombre de *lectures* longues à corriger, nous cherchons actuellement une méthode permettant de combler ces *gaps* plus rapidement.