DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

TECHNOLOGIES DE SÉQUENÇAGE

ET STRATÉGIES D'ASSEMBLAGE *DE NOVO*

Sébastien FAYE | Genoscope/CNS

30 JUIN 2015

OUTLINE

- **Context**
- **Objectives**
- **K-mer Count modelling**
- **The k-mer spectrum**
- **Genome assembly methods**
  - Greedy approaches
  - Graph-based: Overlap-Layout-Consensus (OLC)
  - Graph-based: De Bruijn Graph (DBG)
- **Technologies for long reads sequencing**
  - What can we do with long reads ?
  - OAK Genome assembly

CONTEXT

Genome complexity	*DNA abundance and quality*	*Type of library*	*Sequencing technology*	*Assembly method*	*Assembly QC*
Ploïdy	Material unlimited	Overlapping reads	Sanger	*de bruijn* Graph (DGB)	Continuity
Zygoty	Very small organisms	Single reads	454	Overlap-Layout-Consensus (OLC)	Missamblies
Repeats	Monoclonal population	Short fragments	Illumina	Scaffolding	Physical map
Genome size	High molecular weight DNA	Long fragments	Ion Torrent	Gapclosing	Genetic map
	Nuclear DNA contamination	BAC	Moleculo		
			PacBio		
			Nanopore		

OBJECTIVES

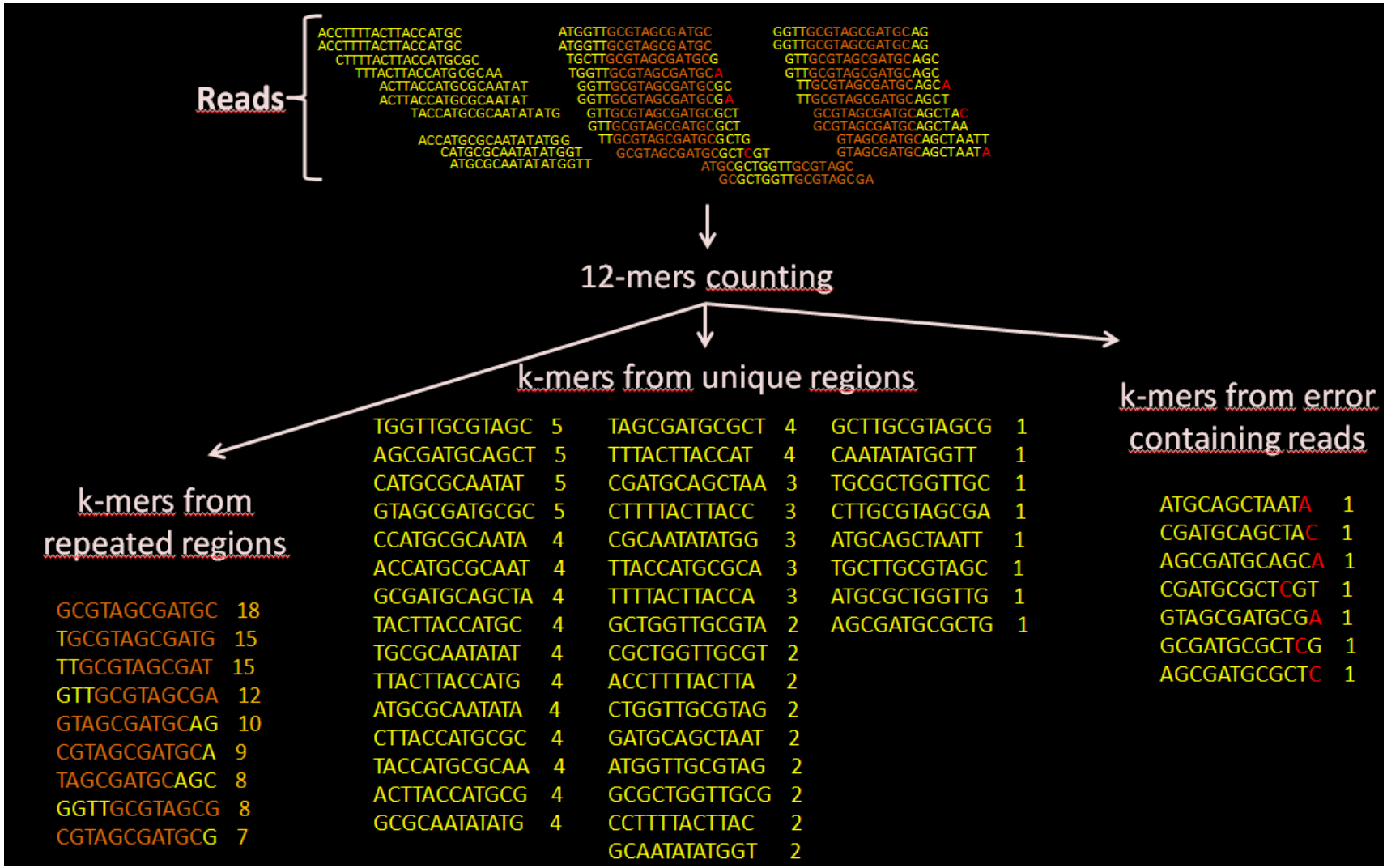
- More than 95% of the genes
- A N50 contig > the gene size
- A N50 scaffold > 1 Mb
- Less than 5% of undetermided bases

- Prior knowledges about the genome complexity
  - Size of the genome
  - Heterozygous part of the genome
  - Repeated part of the genome

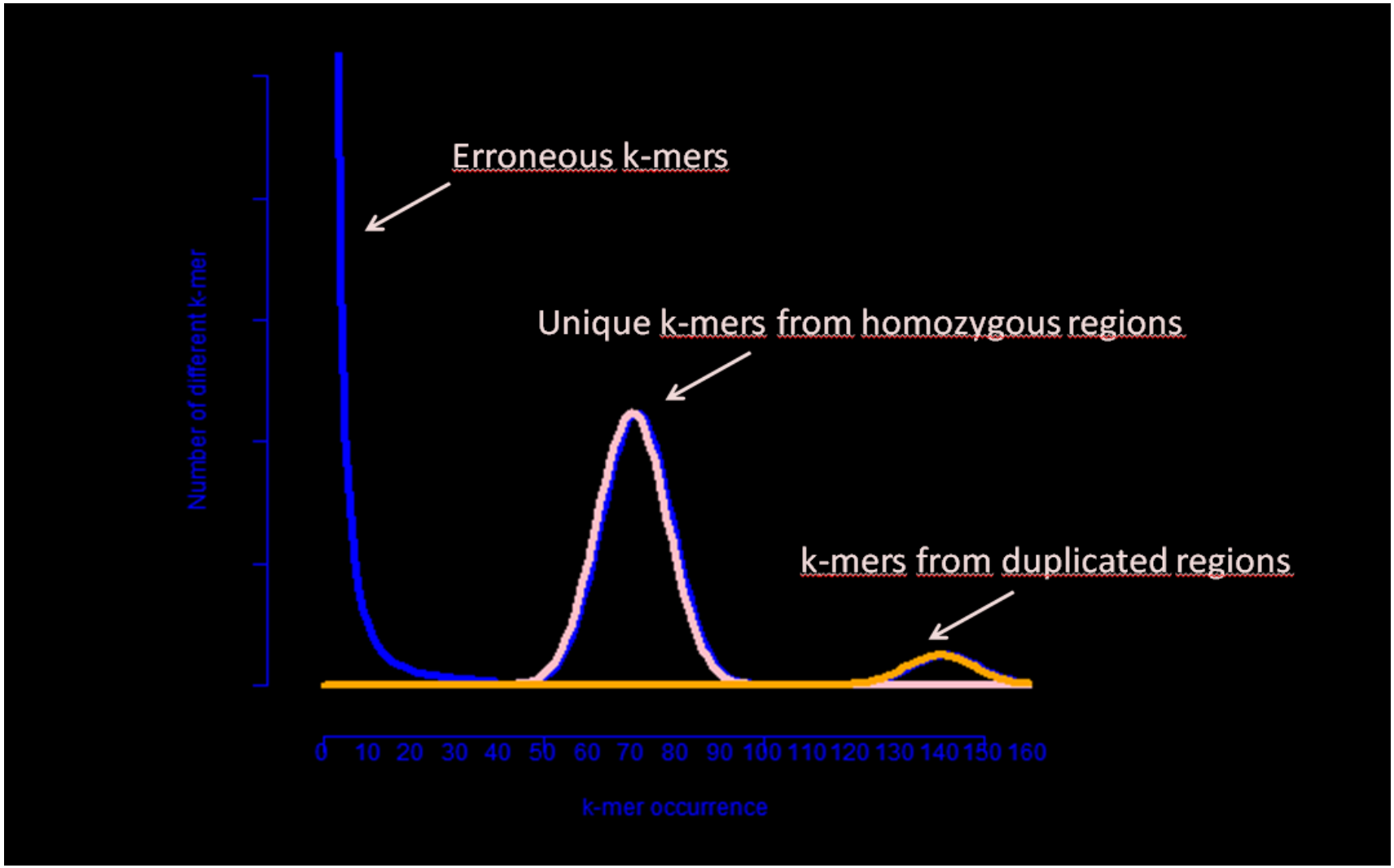
  ➔ **Use K-mers count modelling**

- Need to decide the sequencing and assembly strategy
  - Do I have HMW DNA? => long reads, long fragments, optical maps
  - « easy » genome : Illumina only
  - « hard » genome : hybrid strategy
  - *Sequencing technology access ? Cost?*

K-MERS COUNTING



THE K-MER SPECTRUM



THE K-MER SPECTRUM

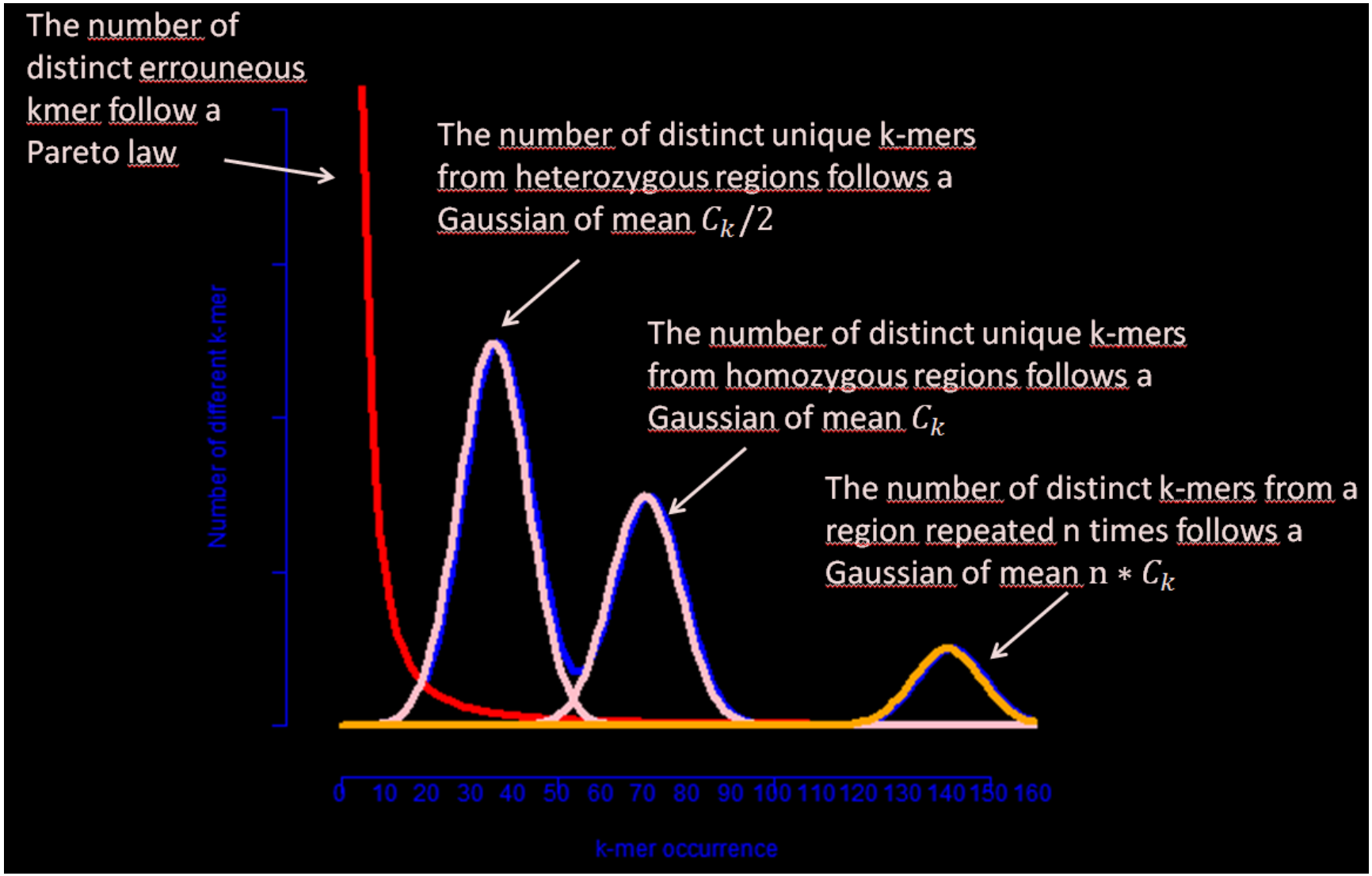
What is the expected k-mer coverage?

G = Genome size
C = Genome coverage
n = Number of reads
l = Reads size
C_k = Expected k-mer coverage

if $G \gg k$ then $G - k \simeq G$ and

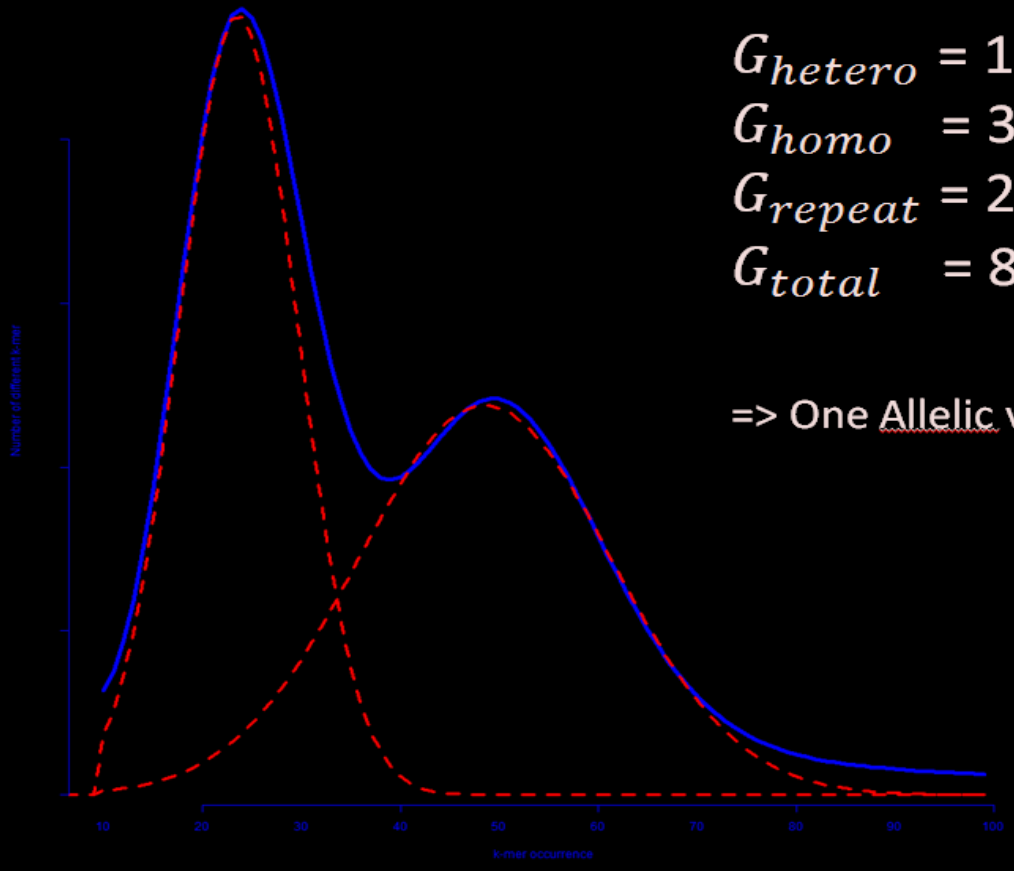
$$C_k = \frac{C\,(l-k+1)}{l} = \frac{n\,(l-k+1)}{G}$$

THE K-MER SPECTRUM



EXAMPLE K-MERS COUNT MODELLING

Step 1: Illumina sequencing of 100 bp reads of *Quercus robur*
Step 2: 31-mers counting with Jellyfish (Marçais *et al.*, Bioinformatics (2011)
Step 3 : 31-mers spectrum modelling



G_{hetero} = 177 Mb (20%)
G_{homo} = 380 Mb (46%)
G_{repeat} = 262 Mb (34%)
G_{total} = 819 Mb

=> One Allelic variant every 155 bp in average

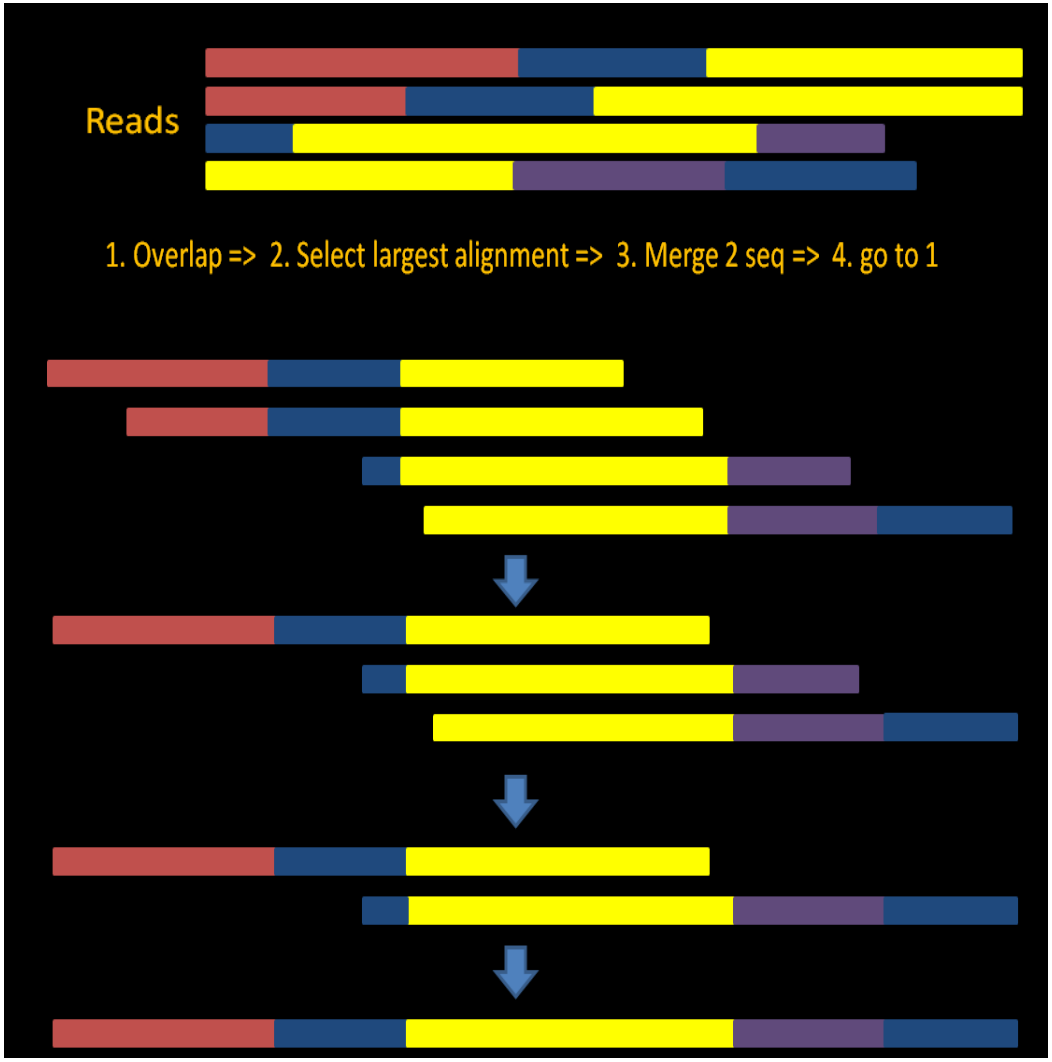
GENOME ASSEMBLY METHODS

- **Greedy approaches**
- **Graph-based: Overlap-Layout-Consensus (OLC)**
- **Graph-based: De Bruijn Graph (DBG)**

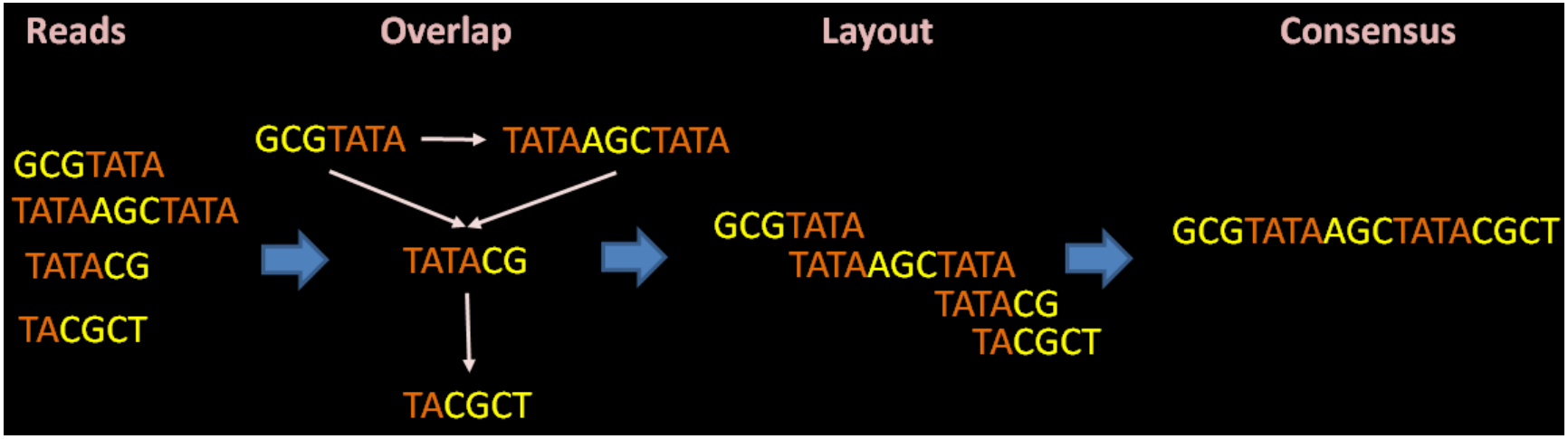
GREEDY APPROACHES

- Greedy assembler for Sanger data
  - PHRAP (1994)
  - The TIGR assembler (1995)
  - CAP3 (1999)

- Greedy assembler for short reads
  - SSAKE
  - SHARCGS
  - VCAKE



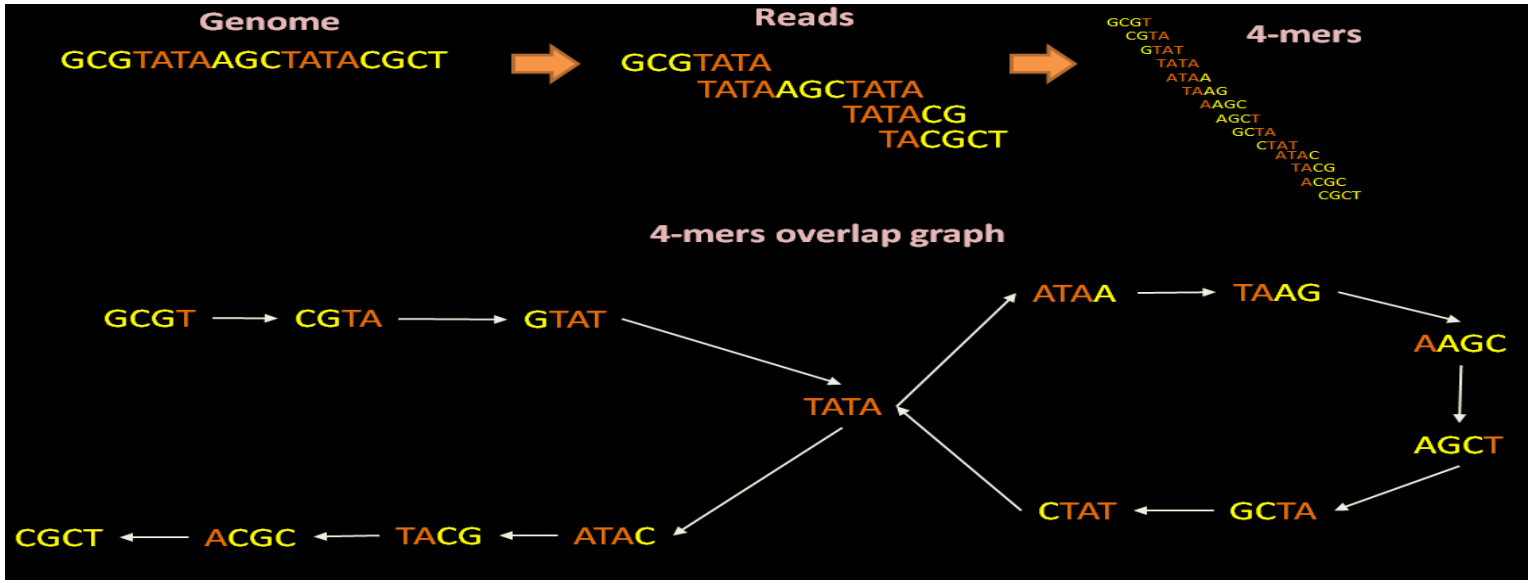
OVERLAP-LAYOUT-CONSENSUS (OLC)



- The OLC problem: the time
  - Overlap step: $O(N^2)$ computation time (N = number of sequenced bases)
  - Use K-mers to **Seed & extend** overlaps.

- Major OLC programs
  - Celera Assembler
  - Newbler (optimised for 454, no more supported and limited read size to 2kb)

DE BRUIJN GRAPH (DBG)



- Sequencing errors: method for reads corrections
- Cannot span repeats with length < K
- Future development: long reads integration (PacBio and Nanopore):
  reads alignment to the graph => haplotype phasing, repeat resolution
- Major DBG programs
  Velvet
  Soapdenovo2 (Luo et al 2012): very popular
  Abyss, SGA (Simpson 2009, Simpson and Durbin 2011): low memory
  Allpath-LG (Gnerre et al., 2010): good continuity with special recipe
  Minia (Chikhi and Rizk 2012): very low memory
  Spades (Bankevich et al., 2012): high continuity

TECHNOLOGIES FOR LONG READS SEQUENCING

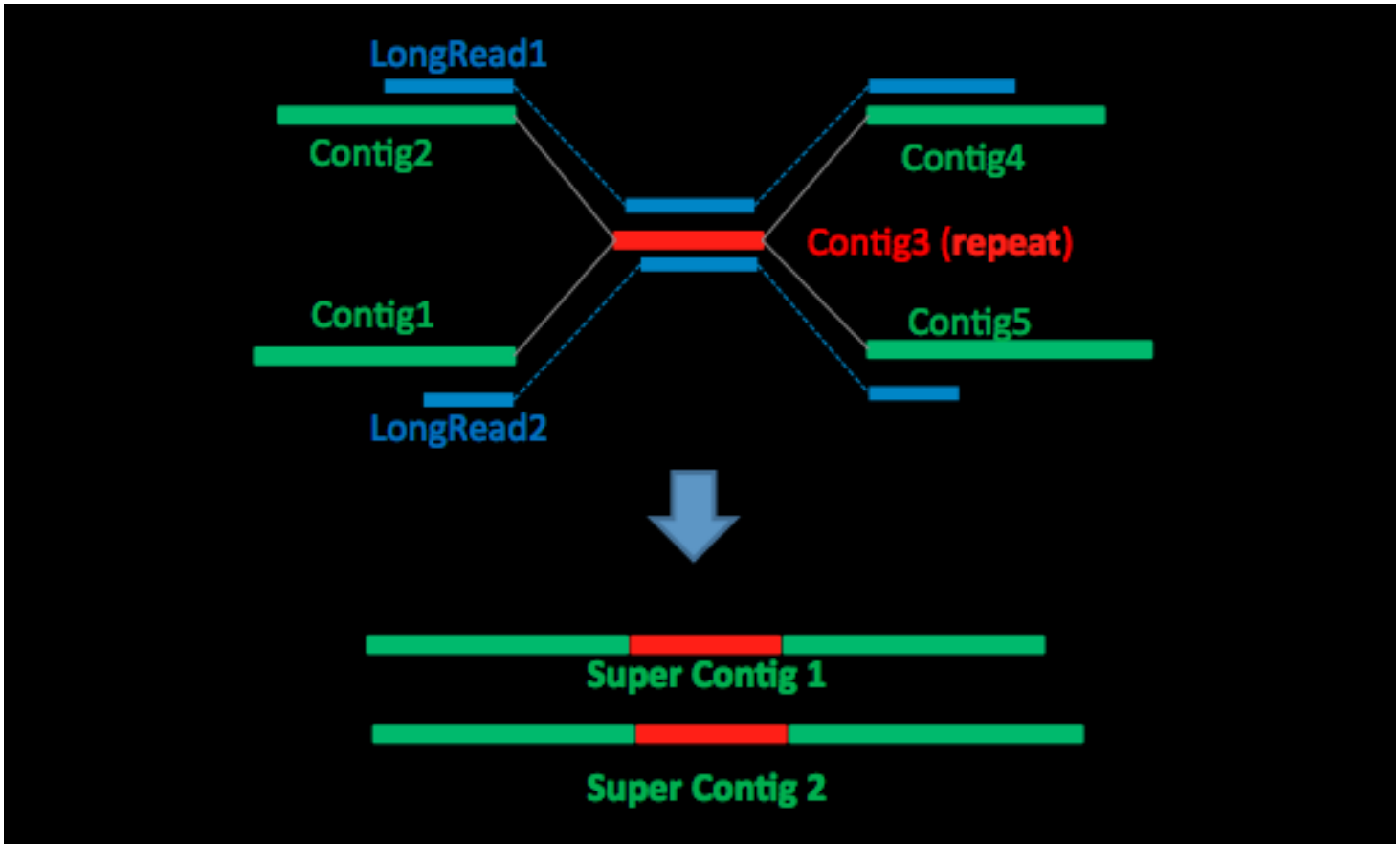
- PacBio reads from Pacific Bioscience RS II
- Moleculo reads from Illumina Truseq Long synthetic reads
- MinION reads from Oxford Nanopore Technology

- **How can you integrate long reads?**

  1) Directly in the assembly process with a OLC method

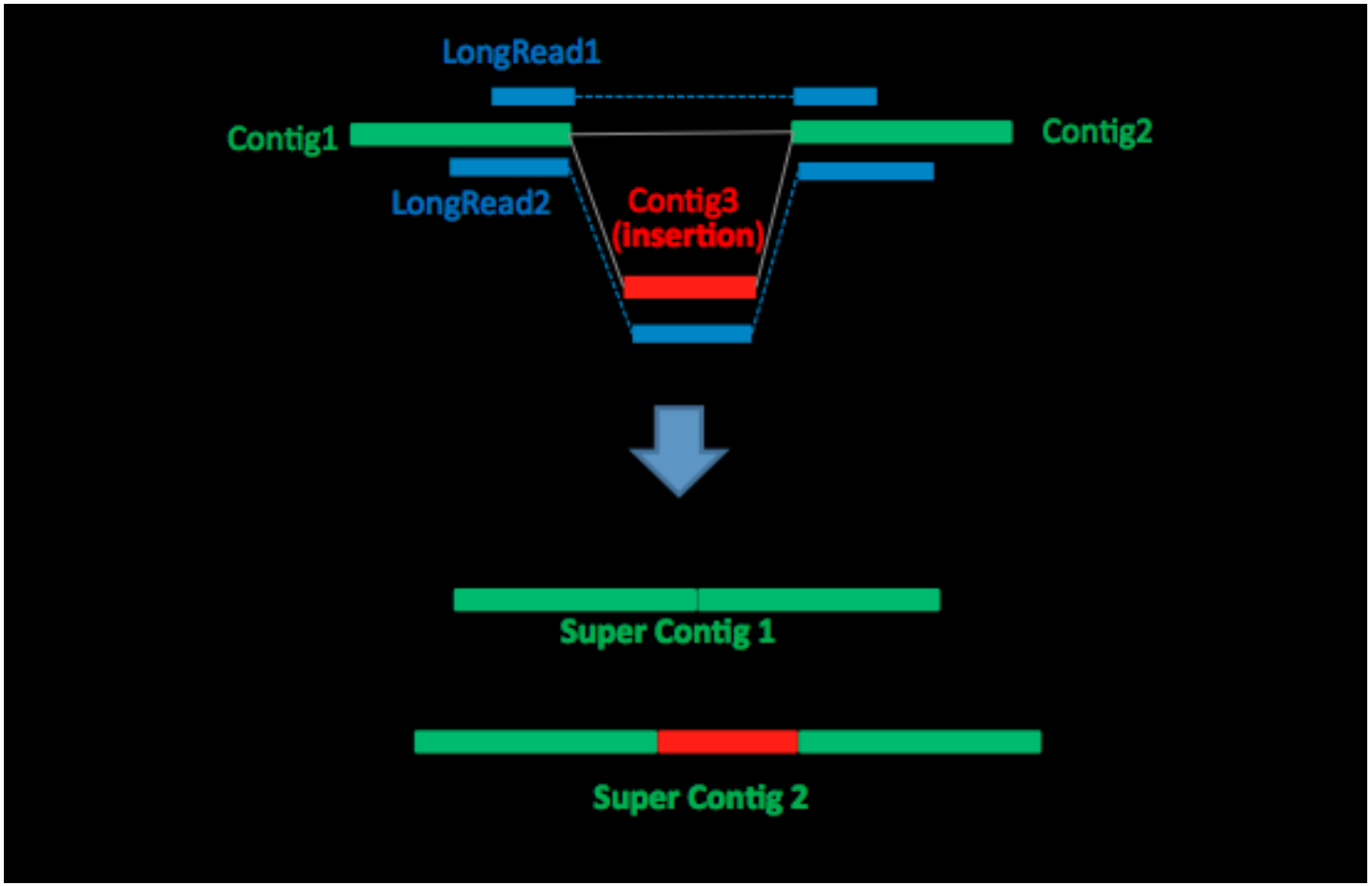
  2) After assembly by aligning the long reads to the contigs

WHAT CAN WE DO WITH LONG READS ?



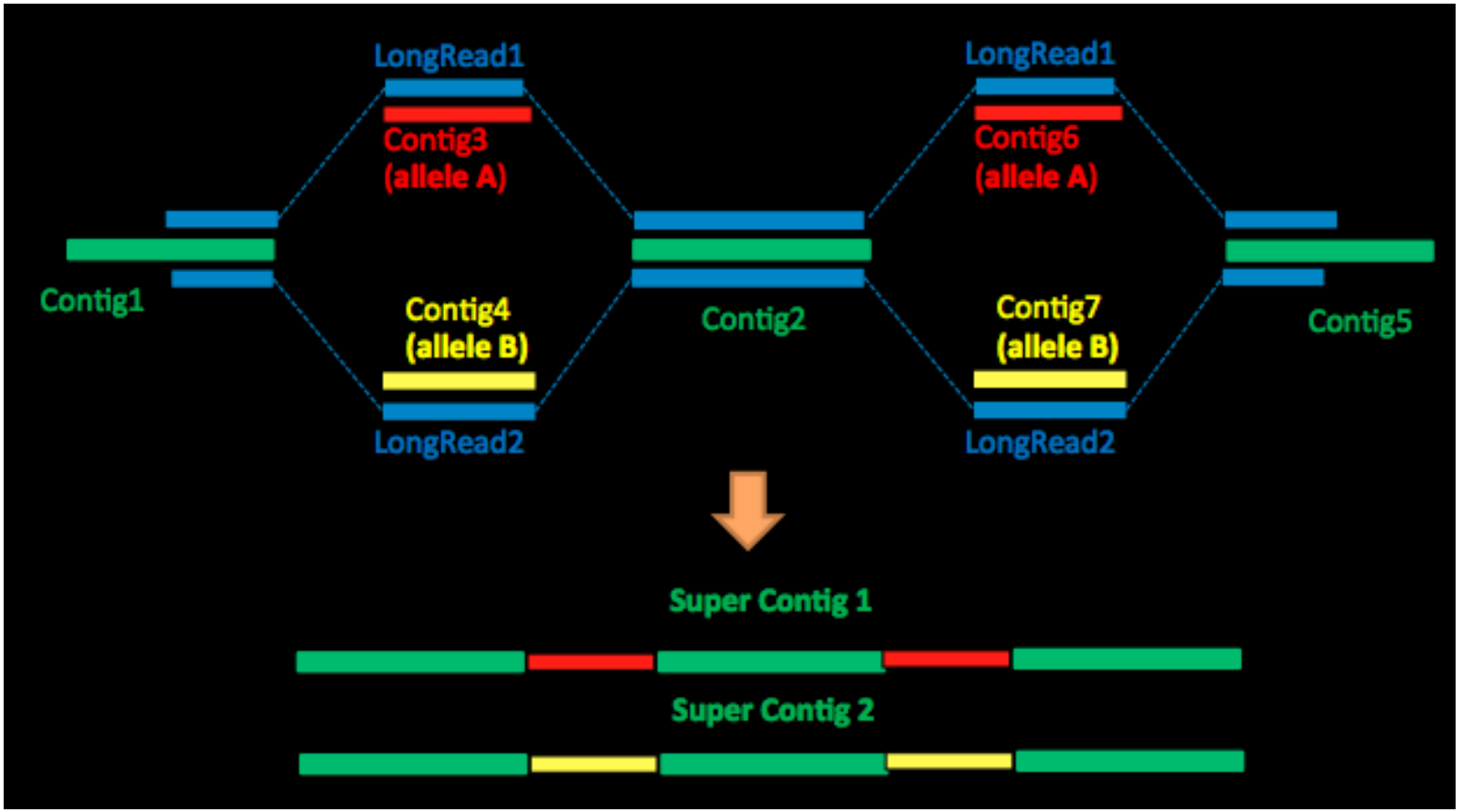
- **Solving repeated regions**

WHAT CAN WE DO WITH LONG READS ?



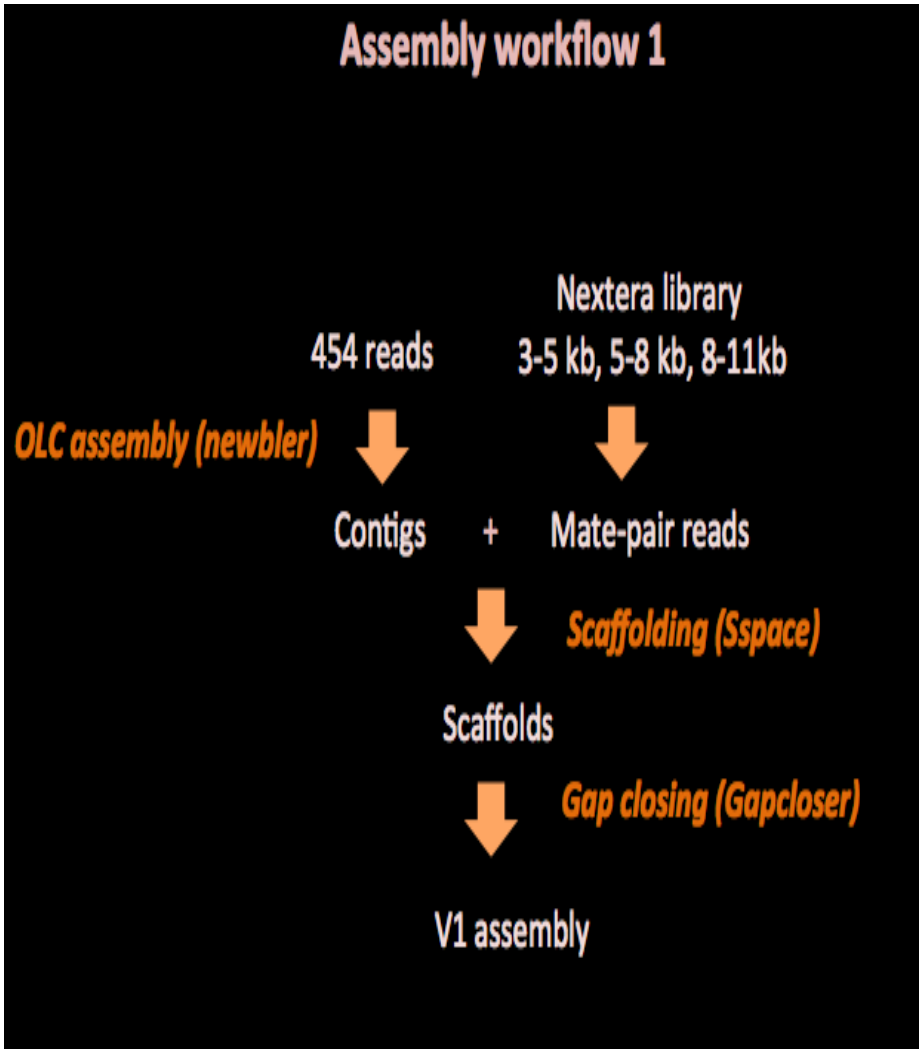
- **Solving of large insertions/deletions containing regions**

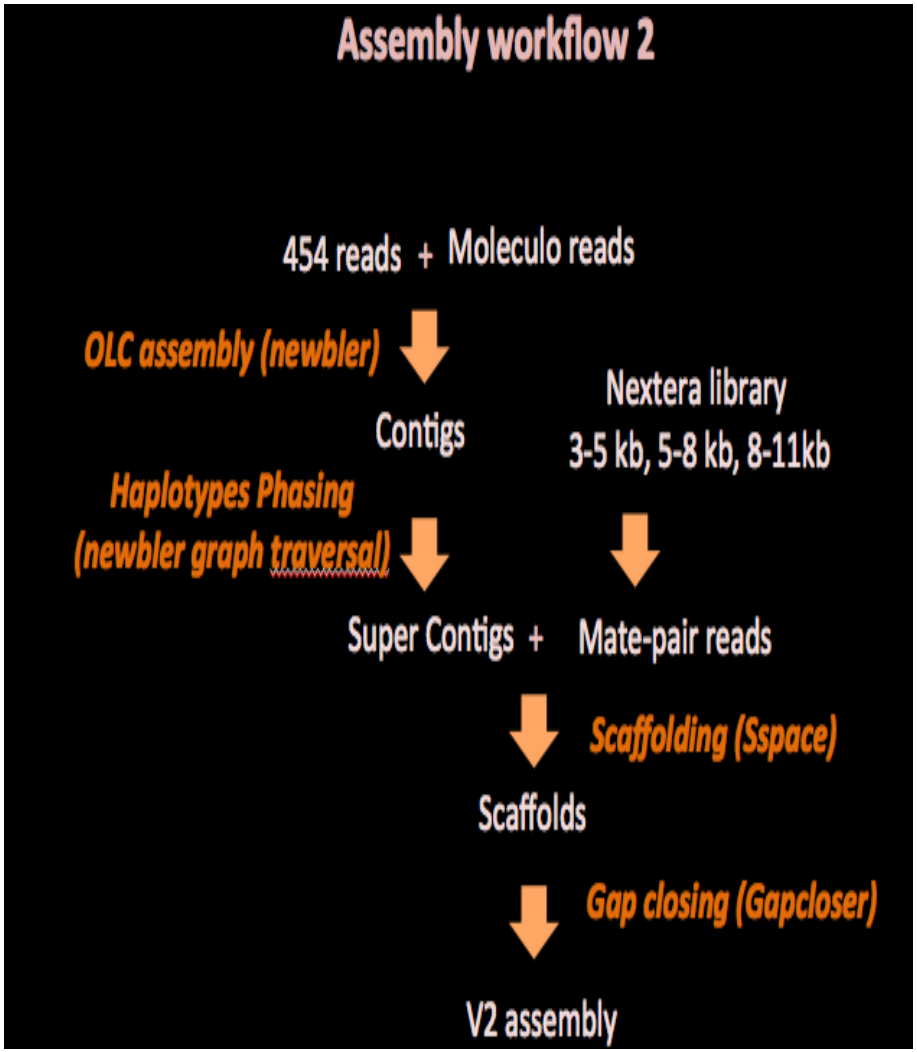
WHAT CAN WE DO WITH LONG READS ?



- **Haplotypes phasing**

OAK GENOME ASSEMBLY





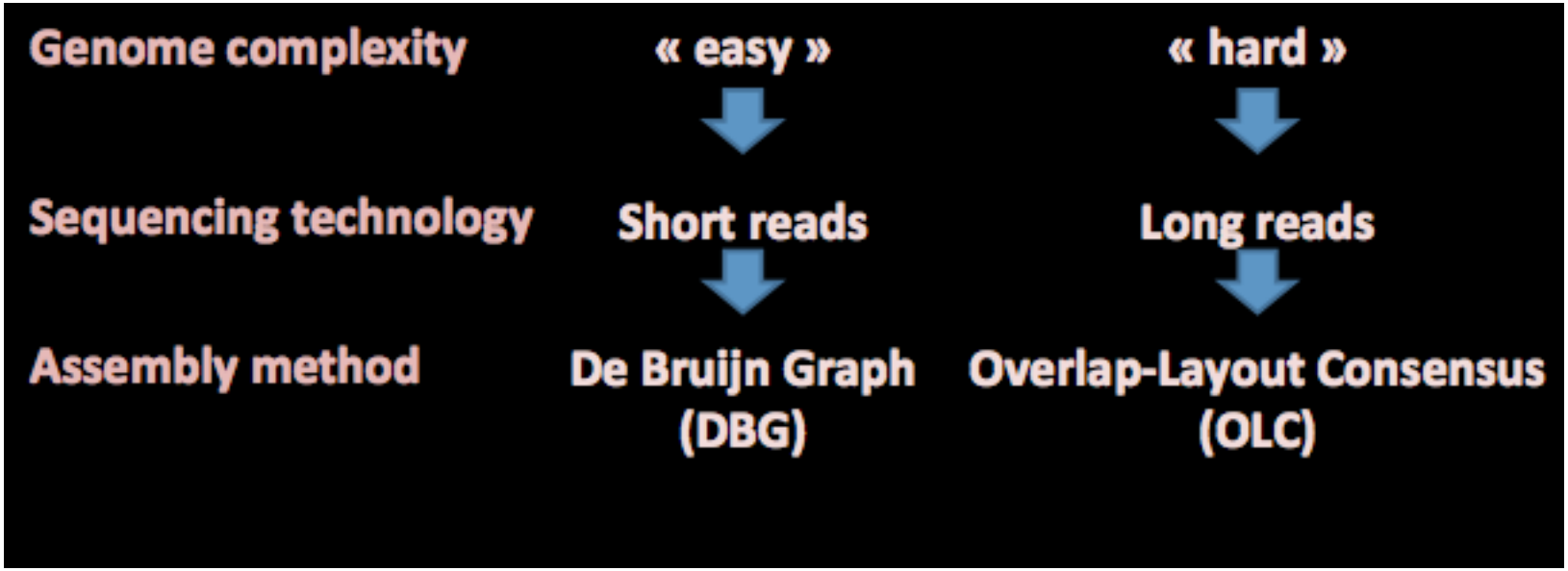
OAK GENOME ASSEMBLY

Assembly	Assemblies metrics V1 (454 + Nextera)	V2 (454 + Moleculo + Nextera)
N Contigs	99 761	53 551
N50 Contigs (Kb)	29.27	65.44
L50 Contigs	11 278	6 291
N90 Contigs (Kb)	5.82	15.37
L90 Contigs	45 896	22 470
Contigs Cumul Size (Gb)	1.19	1.39
N Scaffolds	17 910	9 025
N50 Scaffolds (Kb)	256.64	821.28
L50	1 468	538
N90 Scaffolds (Kb)	35.06	194.34
L90	6 626	1 893
% N	11.56%	4.63%
Scaffolds Cumul Size (Gb)	1.35	1.45

- **Improvment of the scaffolds continuity in V2**

CONCLUSION

- **What are the assembly strategies and how to decide ?**



R&D Bioinformatique et séquençage
Jean-Marc Aury
Sébastien Faye
Caroline Belser
Carole Dossat
Arnaud Couloux
Amin Madoui
Stefan Engelen

Laboratoire de Séquençage
Patrick Wincker
Arnaud Lemainque
Karine Labadie
Adriana Alberti

Laboratoire de Finition
Valérie Barbe
Sophie Mangenot

Commissariat à l'énergie atomique et aux énergies alternatives
Centre de Saclay | 91191 Gif-sur-Yvette Cedex
T. +33 (0)1 XX XX XX XX | F. +33 (0)1 XX XX XX XX

Direction
Département
Service

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019