Seminário de Afinidades em Genômica e Bioinformática (SAGB)

30 de agosto de 2023

USP

cena ESALQ

Mais informações: https://bit.ly/3surnsP

Introdução ao seminário

https://labbces.netlify.app/project/sagb/

Agenda

Periodicidade e local: A cada dois meses, na última quarta-feira do mês às 10h. Um encontro no CENA/USP, o seguinte na ESALQ/USP, e repete.

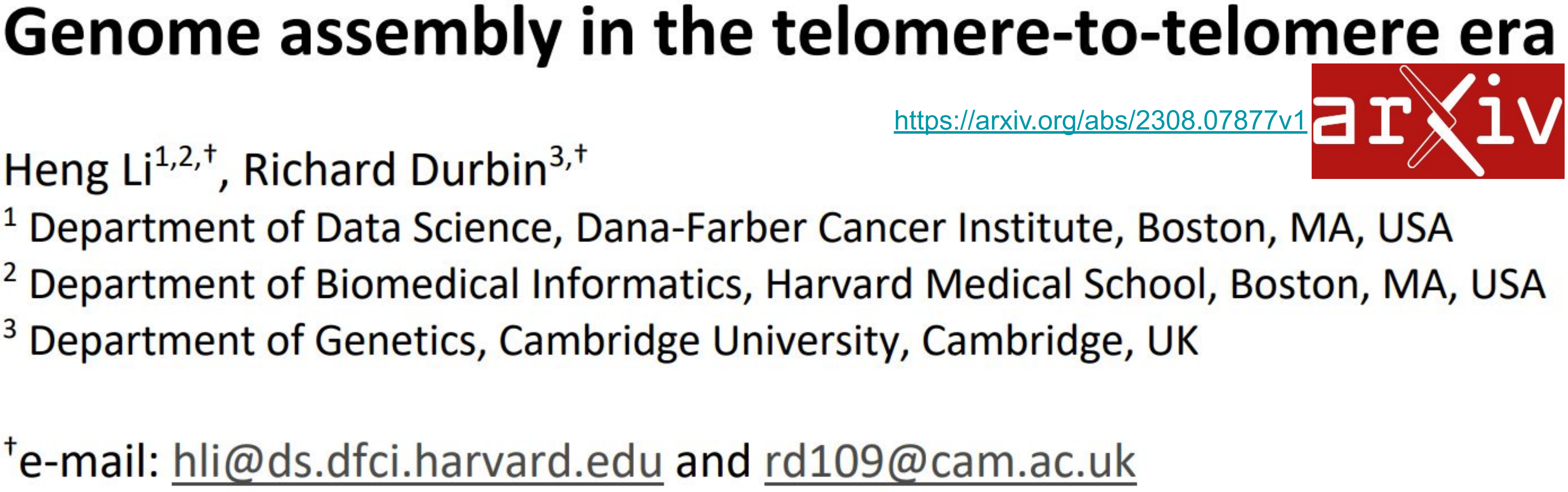
Metodologia: Em cada encontro duas pessoas liderarão a discussão do artigo selecionado, uma pessoa da ESALQ e outra do CENA. No final de cada encontro serão definidos os líderes do próximo encontro, tentando que sejam voluntários.

Objetivos

- Criar um espaço de conversa informal e aberto a todos os alunos (graduação e pós-graduação) e pesquisadores do campus LQ, em temas relacionados a Genômica e Bioinformática.
- Servir como um instrumento para difundir o interesse e conhecimento de Genômica e Bioinformática no campus LQ
- Fomentar as relações inter-pessoais entres pesquisadores e alunos das duas unidades do campus LQ: CENA e ESALQ

Organização

- Dr. Renato Augusto Corrêa dos Santos (CENA)
- Dr. Thais Dal'Sasso (ESALQ)
- Prof. Dr. Claudia Vitorello (ESALQ)
- Prof. Dr. Douglas Silva Domingues (ESALQ)
- Prof. Dr. Diego M. Riaño-Pachón (CENA)



Genome assembly in the telomere-to-telomere era

Heng Li[1,2,†], Richard Durbin[3,†]

[1] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA

[2] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[3] Department of Genetics, Cambridge University, Cambridge, UK

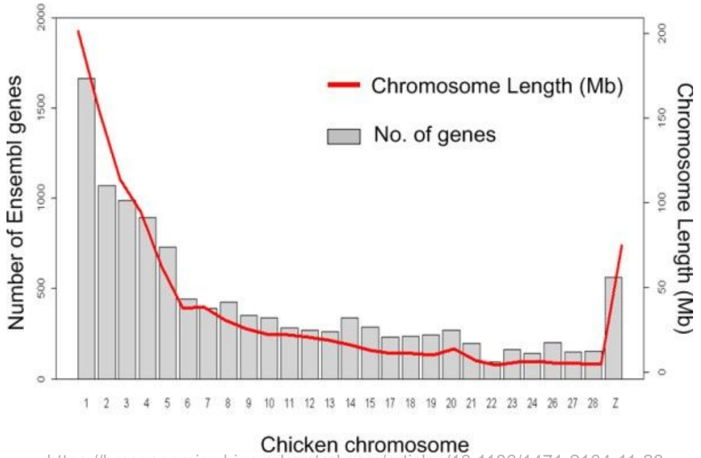
[†]e-mail: hli@ds.dfci.harvard.edu and rd109@cam.ac.uk

https://arxiv.org/abs/2308.07877v1

Algumas perguntas

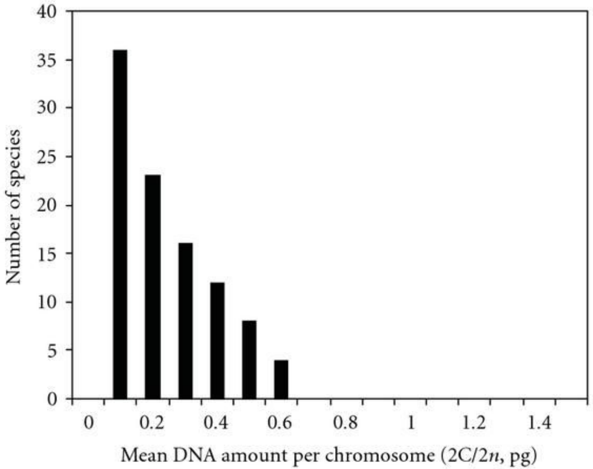
Porque precisamos da montagem de genomas?
Precisamos de montagens telômero a telômero?
É verdade que as tecnologias de sequenciamento de Terceira geração geram leituras de baixa qualidade?
Porque são importantes as sequências repetitivas na montagem de genomas, e como elas impactam as montagens?
O que é um montagem boa?

Chromosomes are huge



https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-28

Monocots

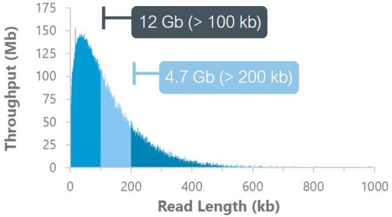


https://www.hindawi.com/journals/jb/2010/862516/

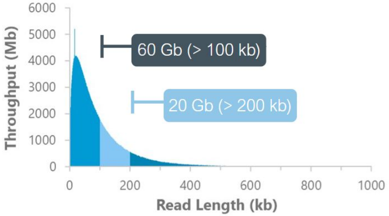
Sequencing technologies DO NOT read full chromosomes

Ultra-long reads ONT

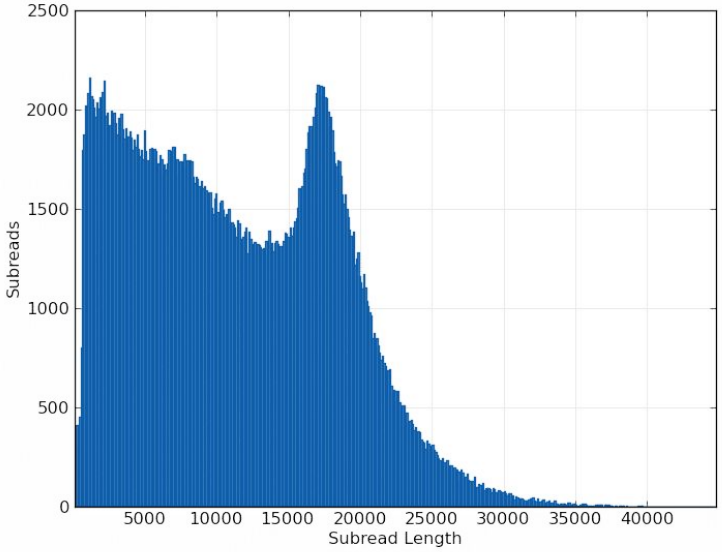
HG02723 on MinION
24 Gb, 97 kb N50, 1.2 Mb max



HG02723 on PromethION
130 Gb, 93 kb N50, 1.0 Mb max



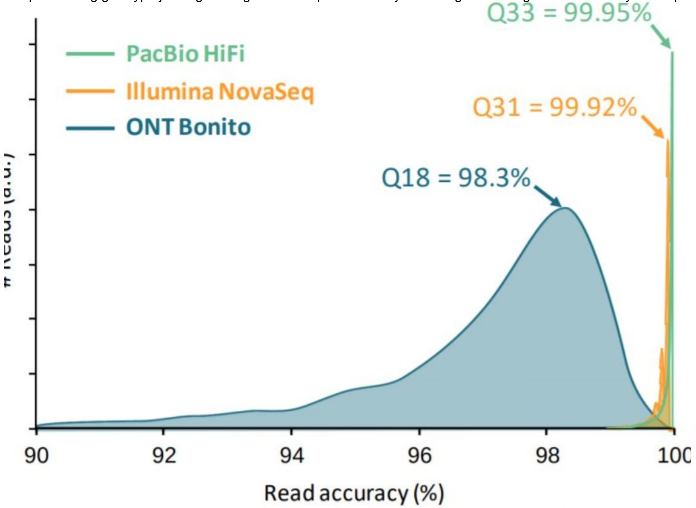
PacBio reads



Modern sequencing technologies are very accurate

2020

https://training.galaxyproject.org/training-material/topics/assembly/tutorials/get-started-genome-assembly/slides-plain.htr



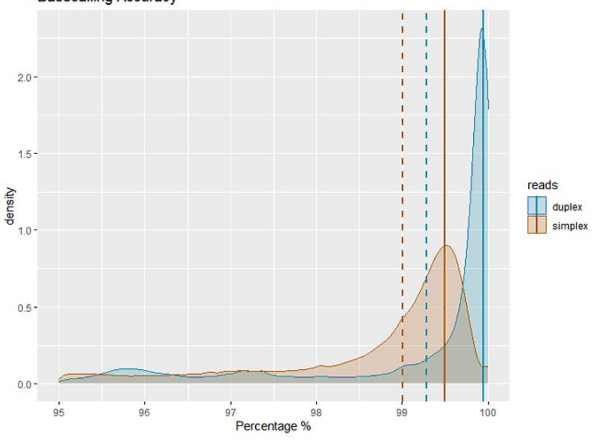
PacBio HiFi: HG003 18 kb library, Sequel II System Chemistry 2.0, precisionFDA Truth Challenge V2

Illumina: HG002 2×150 bp NovaSeq library, precisionFDA Truth Challenge V2

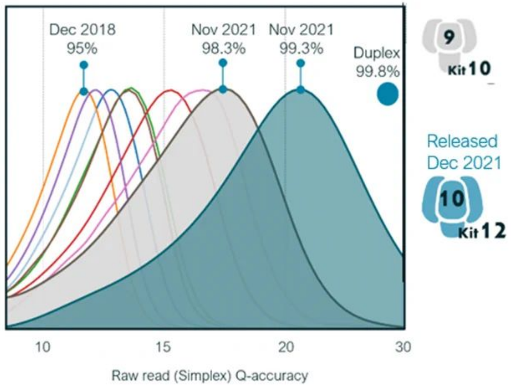
ONT: Bonito NCM Nanopore Tech Update Dec. 2020 and Bonito Basecalling with R9.4.1

2022

https://nanoporetech.com/about-us/news/22-highlights-remember-2022

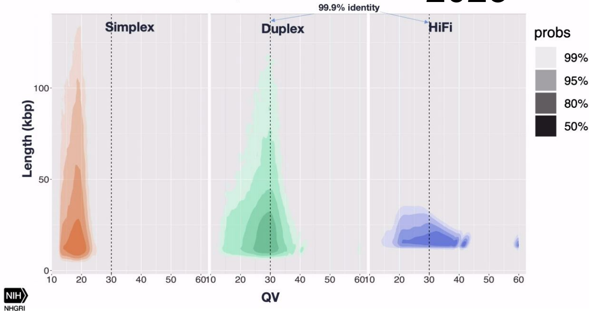


https://twitter.com/nanopore/status/1480996225029652483

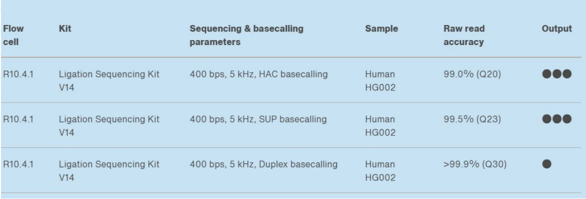


HiFi vs ONT reads (tomato)

2023



https://nanoporetech.com/accuracy



Flow cell	Kit	Sequencing & basecalling parameters	Sample	Raw read accuracy	Output
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, HAC basecalling	Human HG002	99.0% (Q20)	●●●
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, SUP basecalling	Human HG002	99.5% (Q23)	●●●
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, Duplex basecalling	Human HG002	>99.9% (Q30)	●

Which datasets are used in genome assembly?

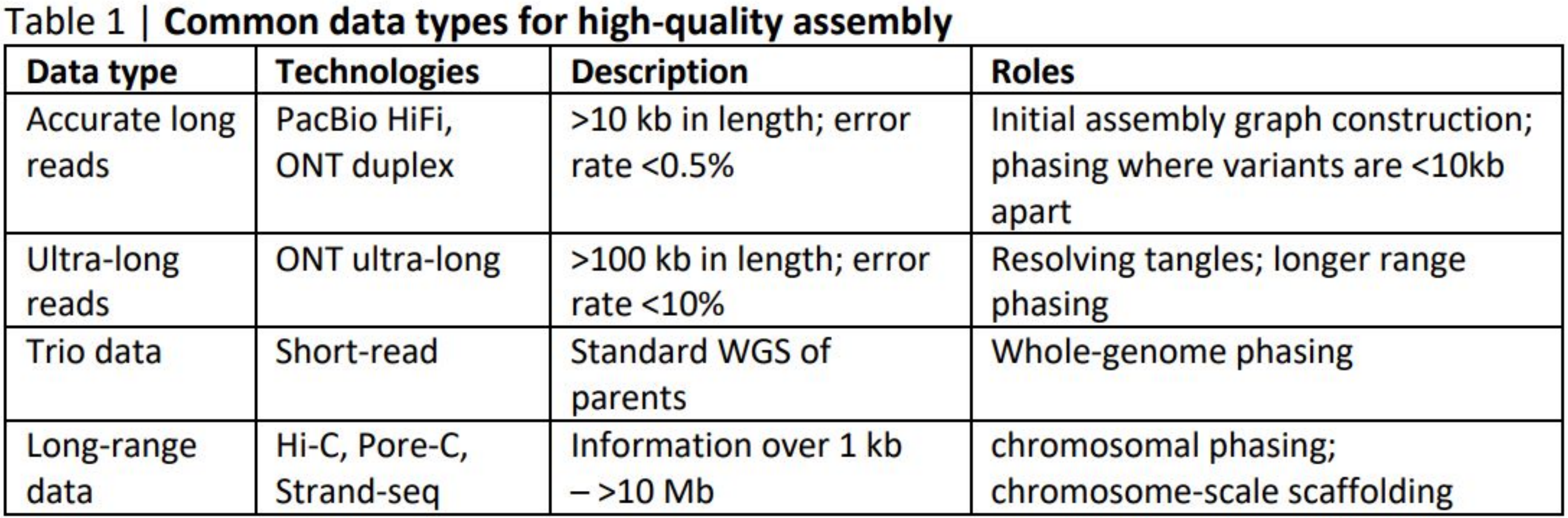
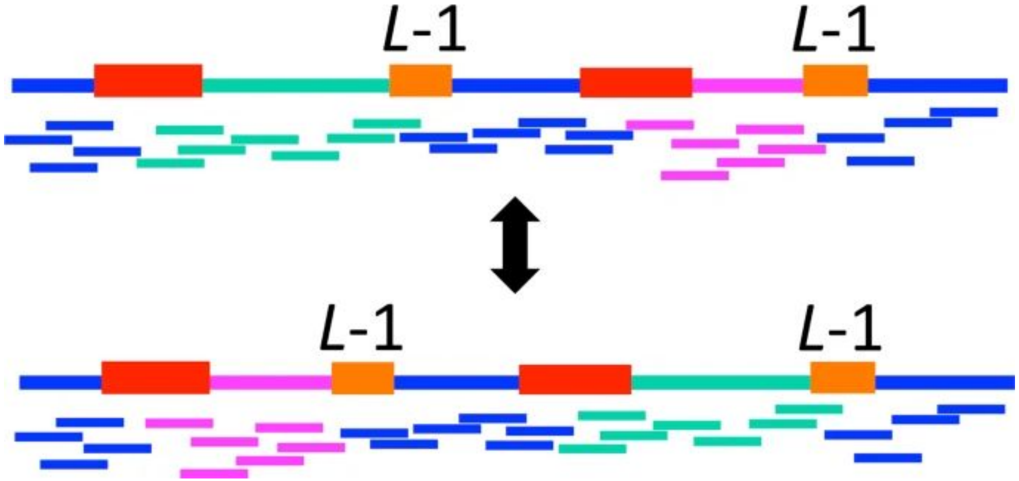


Table 1 | **Common data types for high-quality assembly**

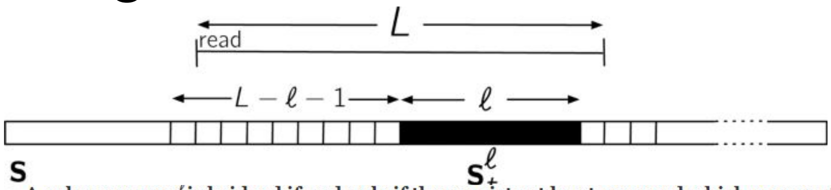
Data type	Technologies	Description	Roles
Accurate long reads	PacBio HiFi, ONT duplex	>10 kb in length; error rate <0.5%	Initial assembly graph construction; phasing where variants are <10kb apart
Ultra-long reads	ONT ultra-long	>100 kb in length; error rate <10%	Resolving tangles; longer range phasing
Trio data	Short-read	Standard WGS of parents	Whole-genome phasing
Long-range data	Hi-C, Pore-C, Strand-seq	Information over 1 kb – >10 Mb	chromosomal phasing; chromosome-scale scaffolding

But, keep an eye on ONT Duplex and putty sequencing

Why repetitive sequences make assembly difficult?
How difficult depends on the read length

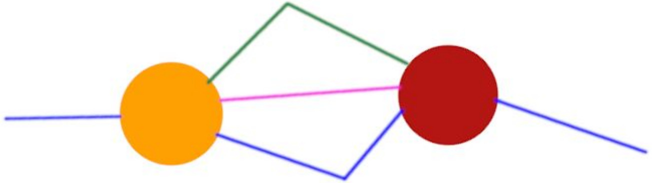


The likelihood of observing the reads under two possible sequences (the green and magenta segments swapped) is the same. Here, the two red subsequences form a repeat and the two orange subsequences form another repeat.



A subsequence s_t^ℓ is bridged if and only if there exists at least one read which covers at least one base on both sides of the subsequence, i.e. the read arrives in the preceding length L - ℓ -1 interval.

Oftentimes repeats are collapsed, and the assembly is fragmented



Luckily . . .

Given long error-free reads, we can distinguish different repeat copies and successfully assemble them. Reads are never all entirely error-free, but when the read error rate is low enough and sequencing errors are sufficiently independent, we can correct most errors and achieve high-quality assembly (Li & Durbin, 2023).

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S5-S18

Recipes

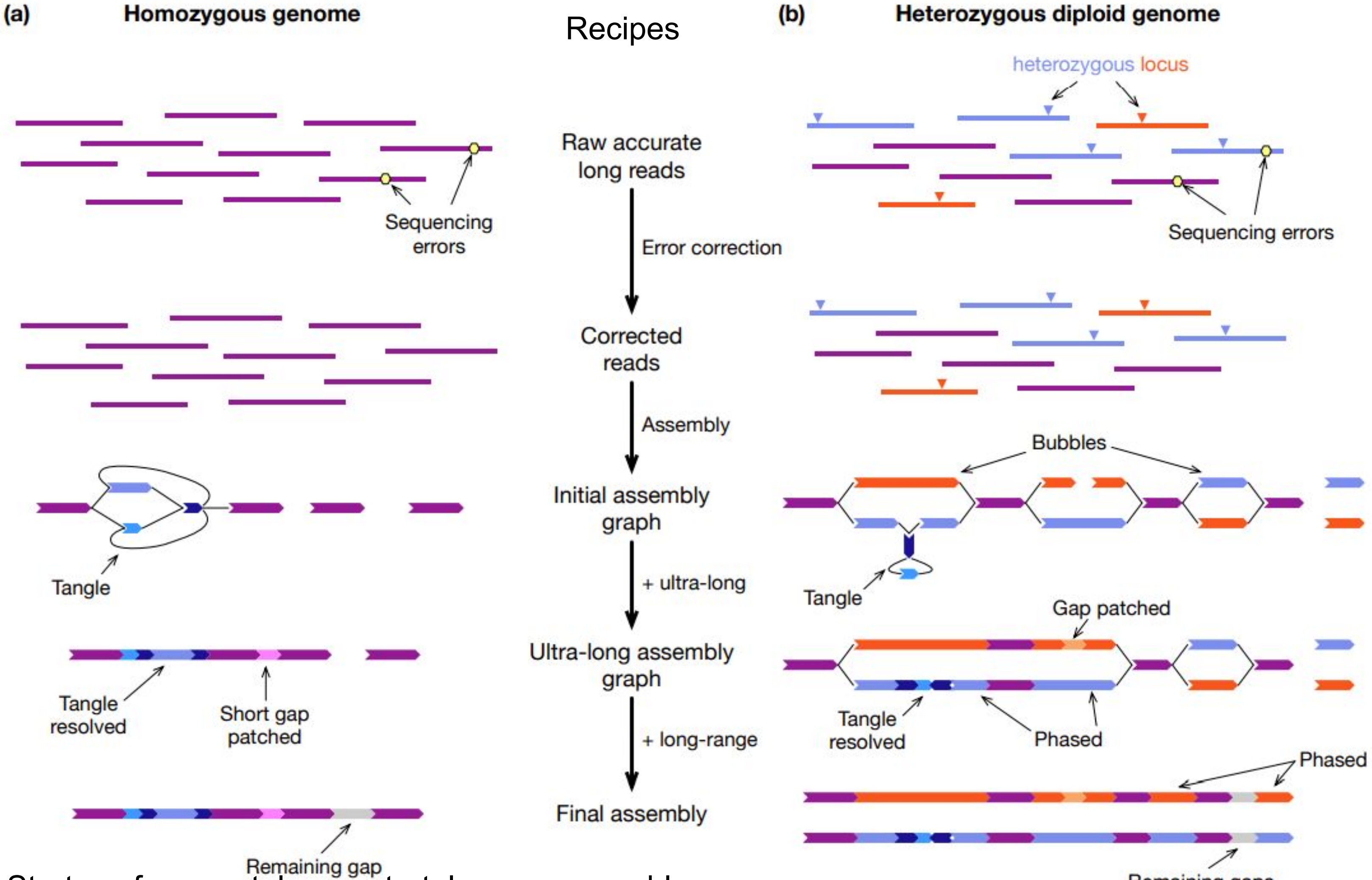


Figure 1.Strategy for near telomere-to-telomere assembly

Chromosome level phasing required more than long-reads. Long-range data is very important!

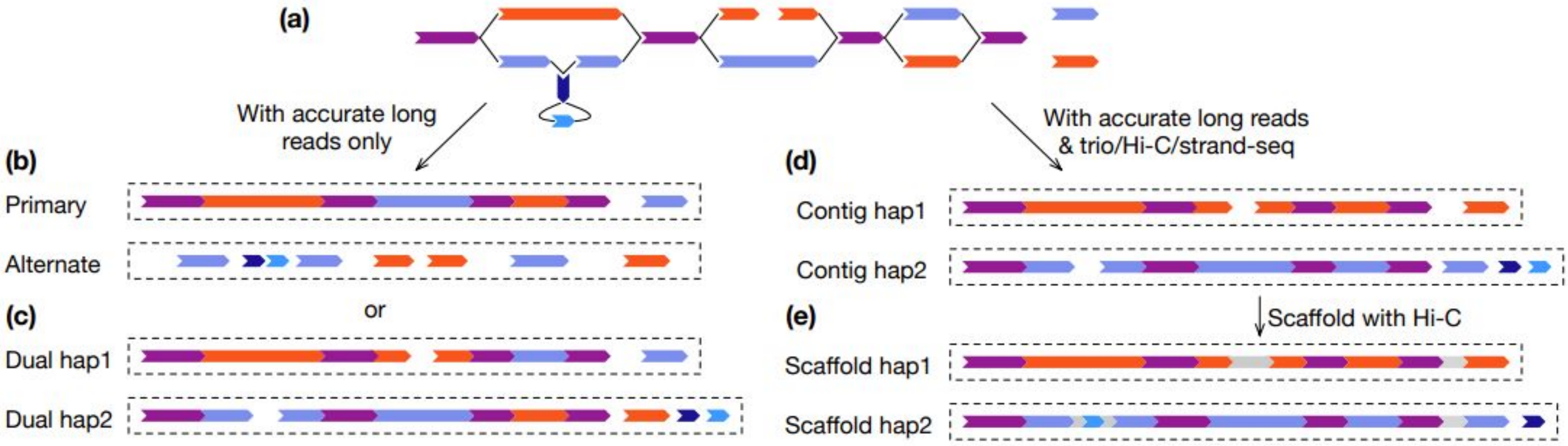
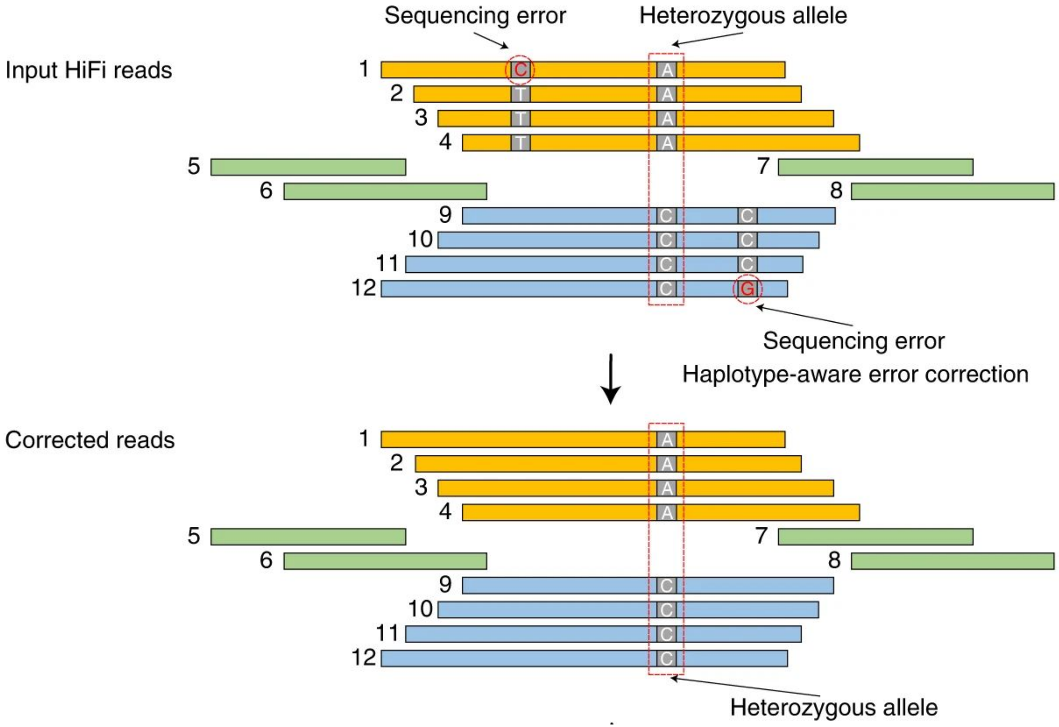
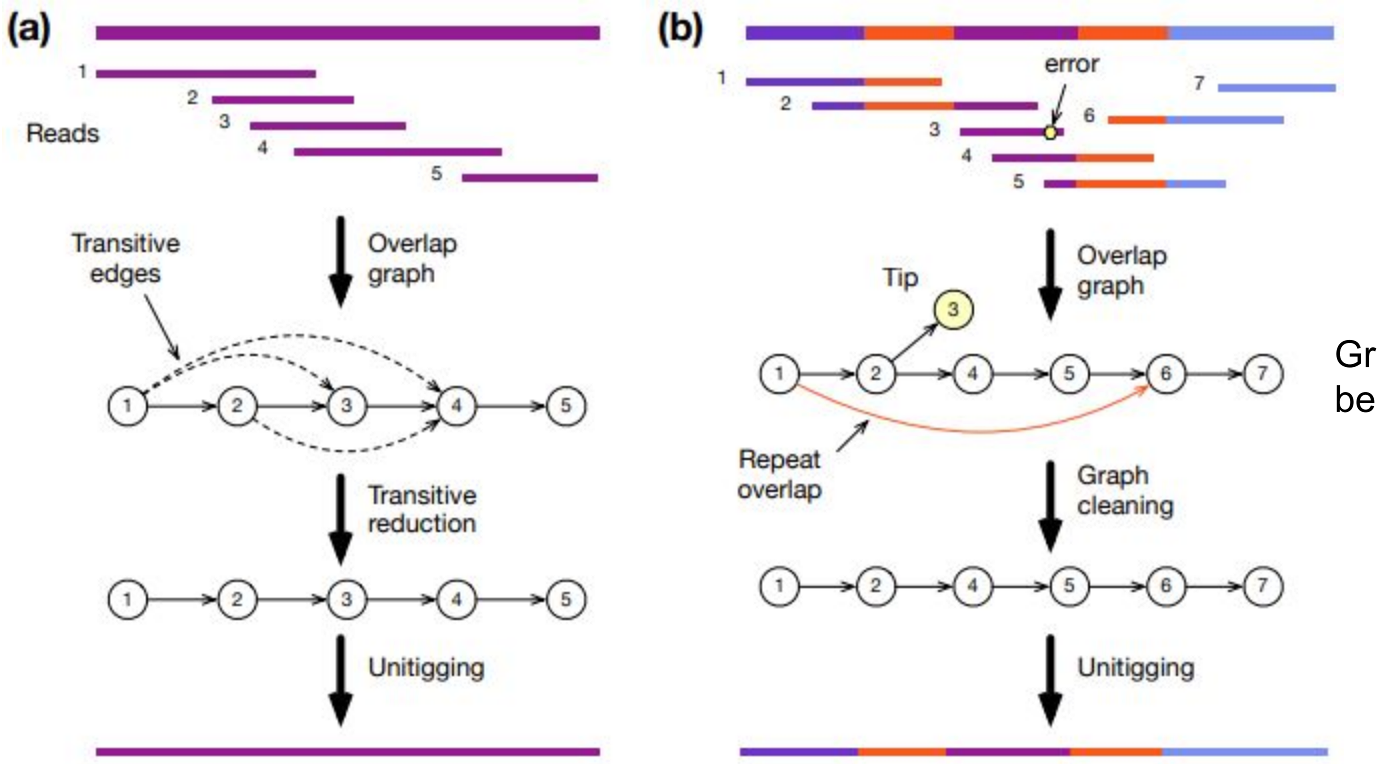


Figure 2. Types of phased assembly of diploid samples

As good as they are, modern TGS reads still need error correction

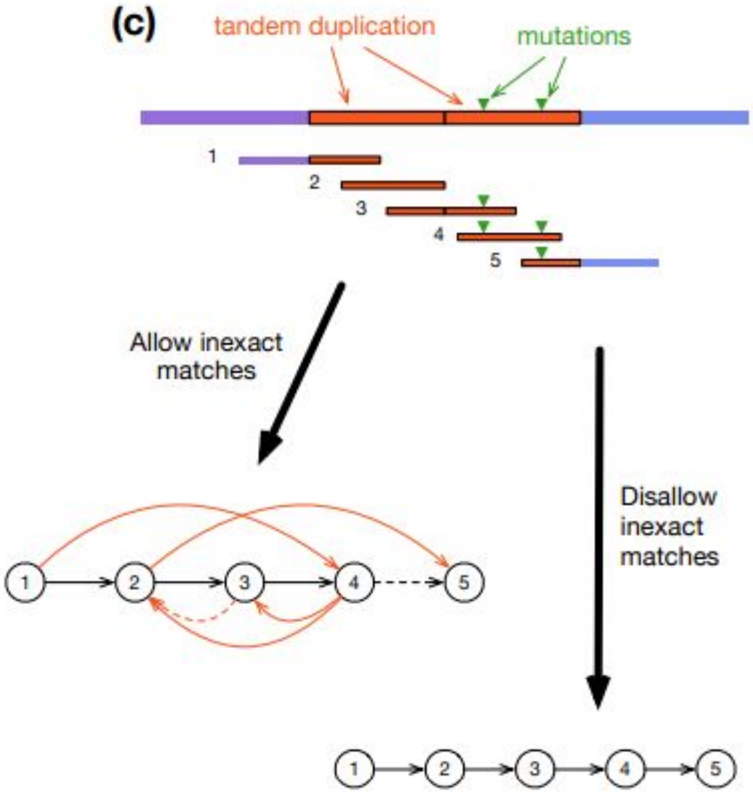


https://www.nature.com/articles/s41592-020-01056-5



Graphs need to be simplified

Figure 3. Assembly with overlap graphs



High quality reads (near perfect), allow to look “just” for perfect overlaps, which greatly simplest the overlap graph, and permits identifying repeat copies and

Figure 3. Assembly with overlap graphs

perfect overlaps also allow phasing.

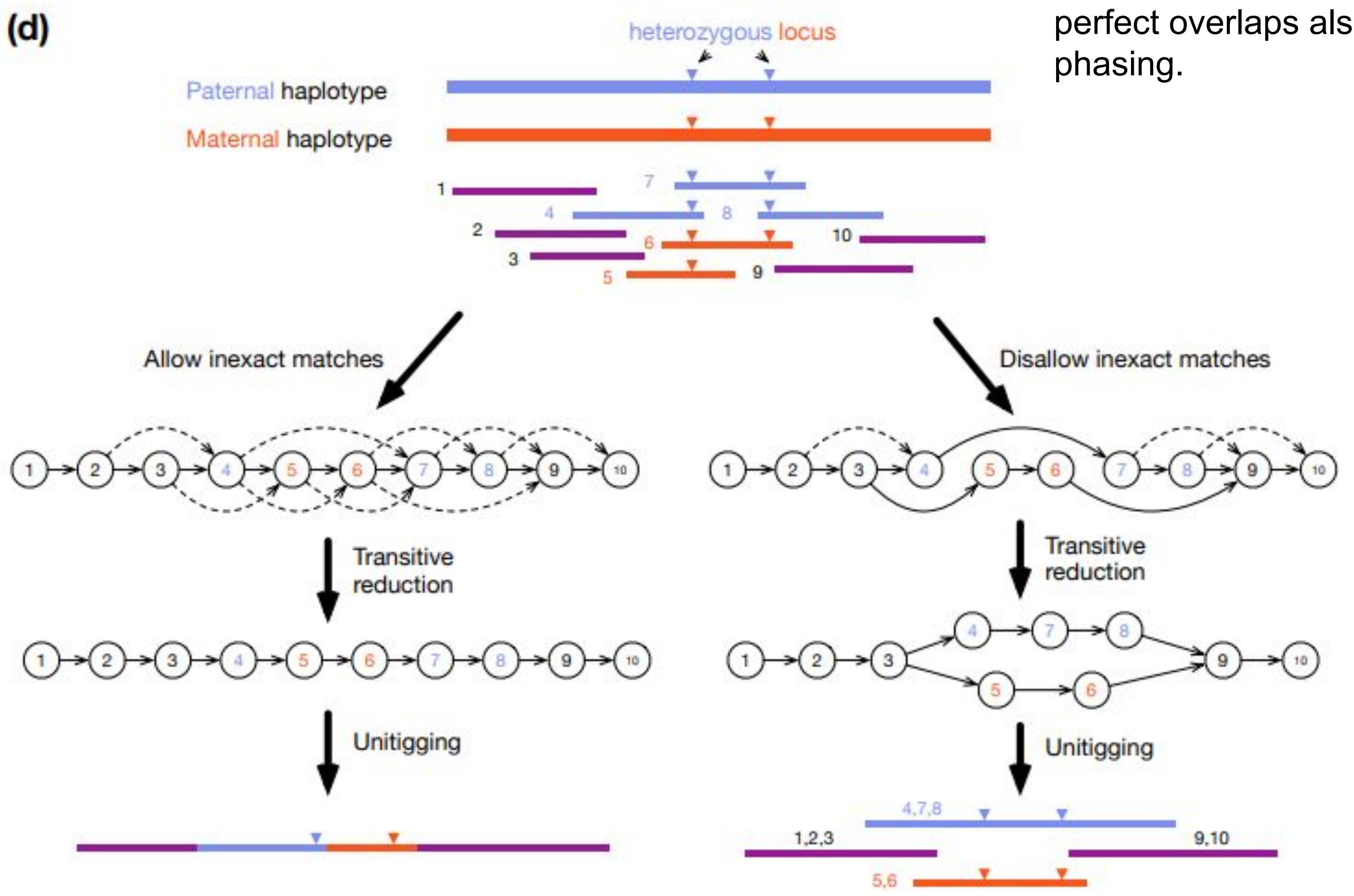
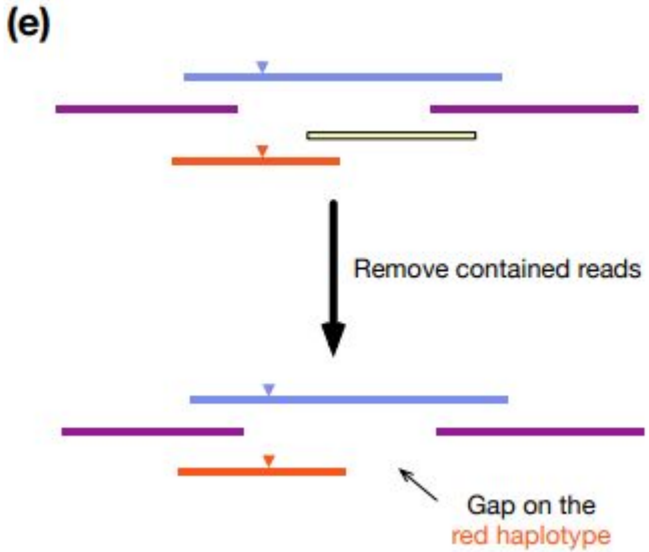
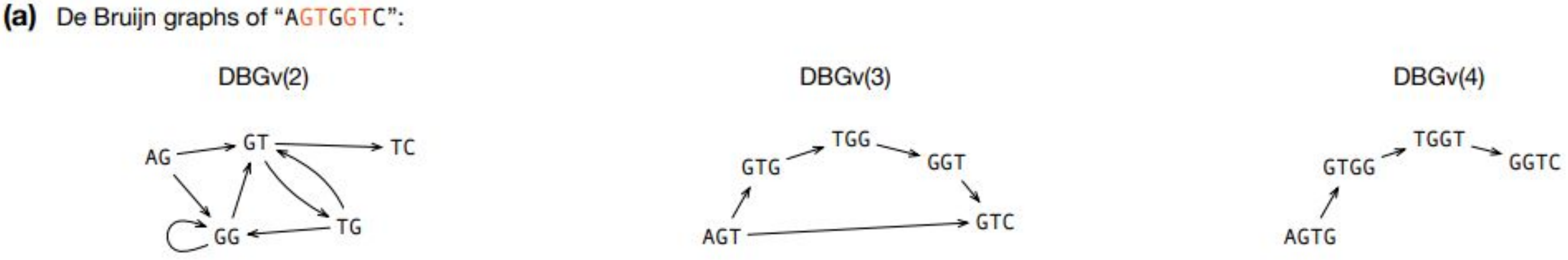


Figure 3. Assembly with overlap graphs



It is common to remove contained reads (yellow), i.e., a read contained in another one. However this could lead to assembly gaps, particularly when phasing. It is one of the main problems for overlap/string graphs assemblers.

Figure 3. Assembly with overlap graphs

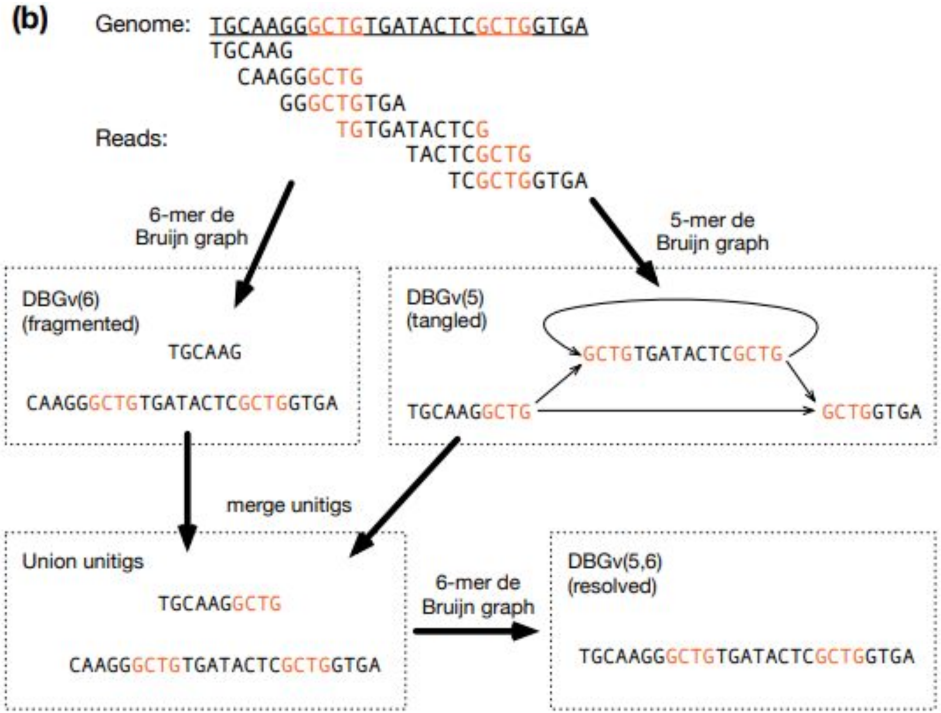


- The value of *k* is usually much smaller than the read length, which can lead to loss some information
- Smaller values of *k*, lead to more ambiguities, as shown in the figure. But much large values of *k* could lead to contig breakpoints in low coverage regions.

Figure 3. Assembly with de Bruijn graphs

There is no single best k for all situations, modern assemblers can use a mixture of k values for different regions of the genome, based on read coverage.

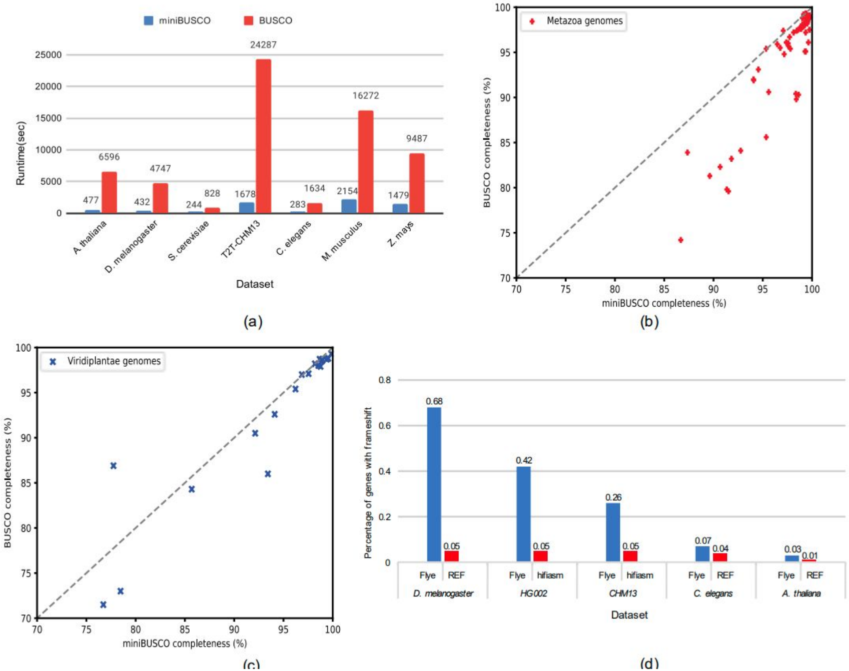
DBG assemblers were very common for short-read technologies. Now with near perfect long-reads they are coming back.



Evaluating sequence assemblies: Gene content

Looking for sets of conserved genes, the more you find the more complete assembly you have

- BUSCO
- Compleasm
- asmgene

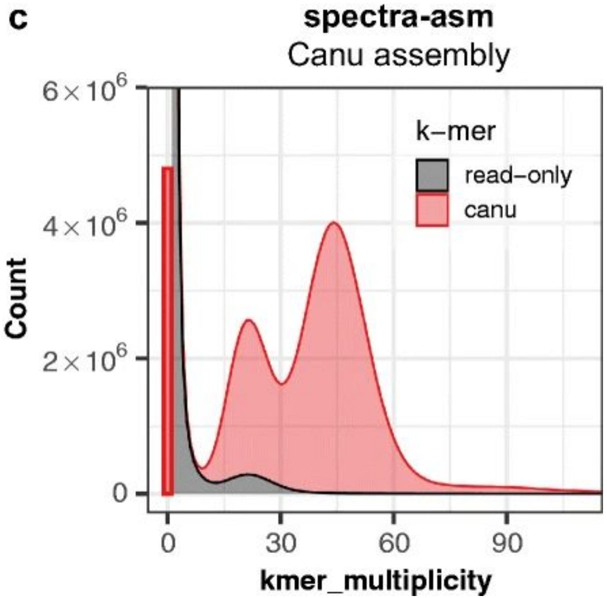


https://www.biorxiv.org/content/10.1101/2023.06.03.543588v1

Evaluating sequence assemblies: K-mers

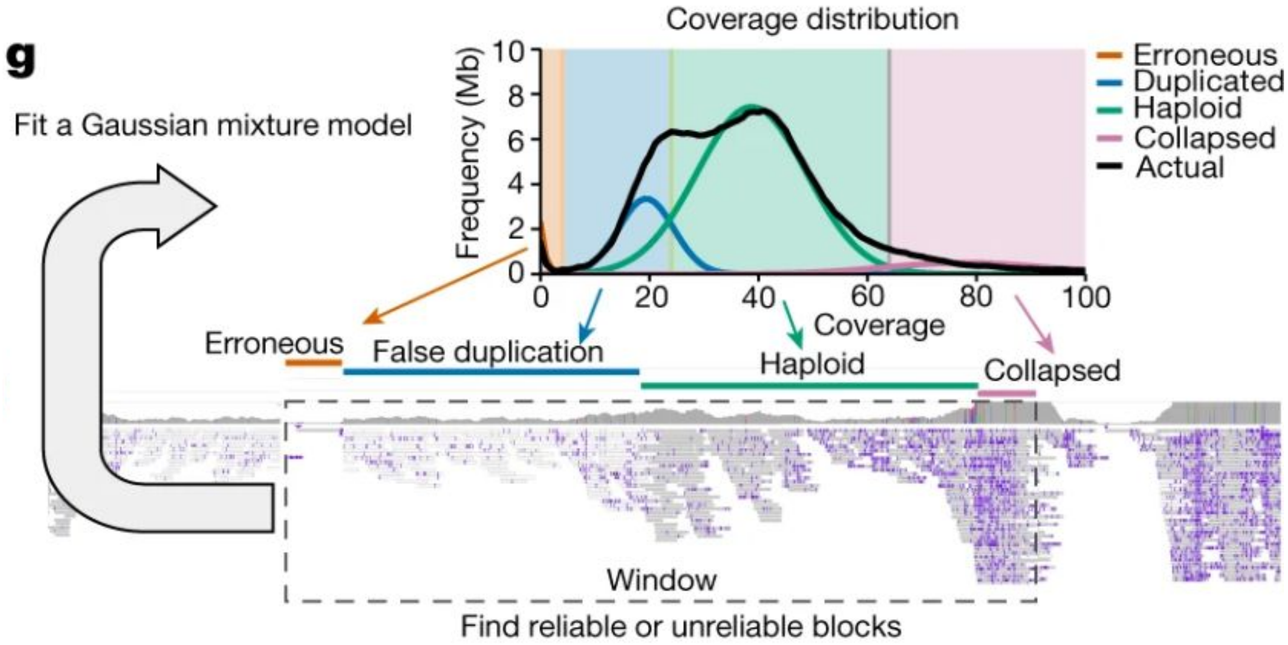
The catalog of kmers and their frequency should be similar between the assembly and the reads, deviation of this could suggest problems to be looked at:

- A kmer frequent in reads but absent in assembly suggest a part of the genome is missing in the assembly
- On the other hand, kmer more frequent in assembly than in reads, suggest a false supplication in assembly
- The phasing accuracy could be measured when trio data is available.



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02134-9

Evaluating sequence assemblies: Alignment



https://www.nature.com/articles/s41586-023-05896-x

Seminário de Afinidades em Genômica e Bioinformática (SAGB)

USP

cena ESALQ