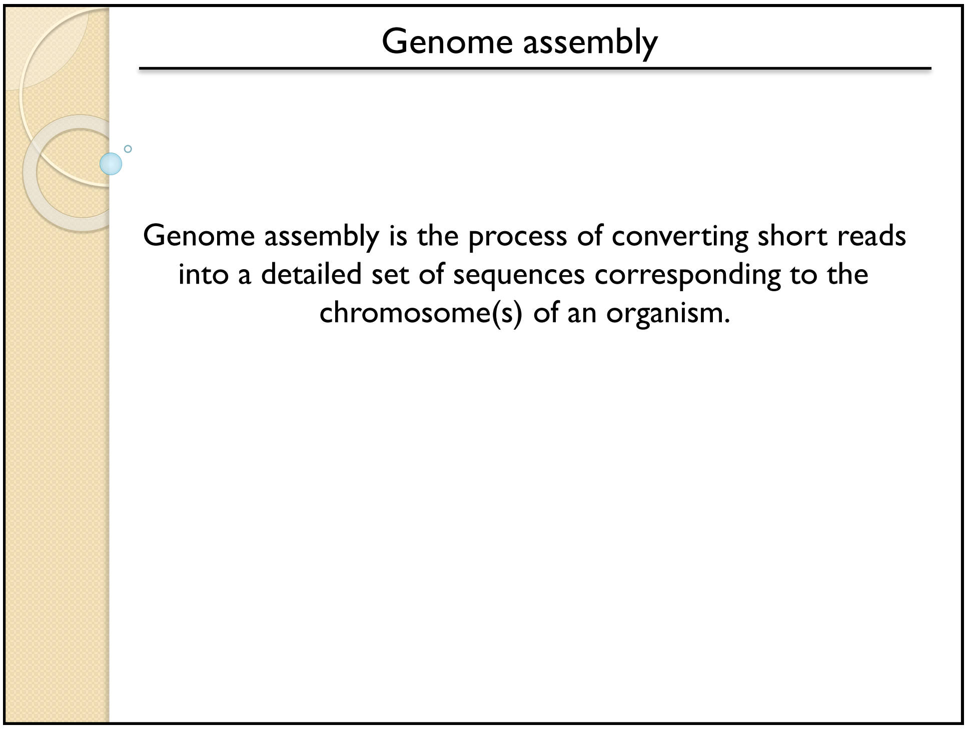
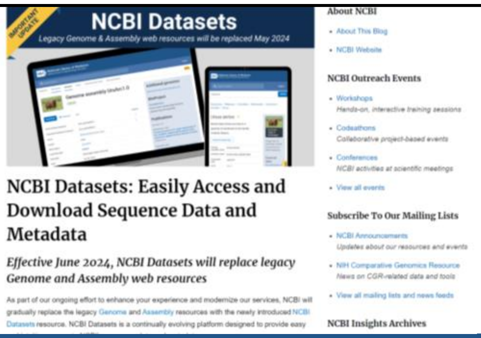
Genome assembly



Genome assembly

Genome assembly is the process of converting short reads into a detailed set of sequences corresponding to the chromosome(s) of an organism.



NCBI Datasets

Legacy Genome & Assembly web resources will be replaced May 2024

NCBI Datasets: Easily Access and Download Sequence Data and Metadata

Effective June 2024, NCBI Datasets will replace legacy Genome and Assembly web resources

As part of our ongoing effort to enhance your experience and modernize our services, NCBI will gradually replace the legacy Genome and Assembly resources with the newly introduced NCBI Datasets resource. NCBI Datasets is a continually evolving platform designed to provide easy

About NCBI

- About This Blog
- NCBI Website

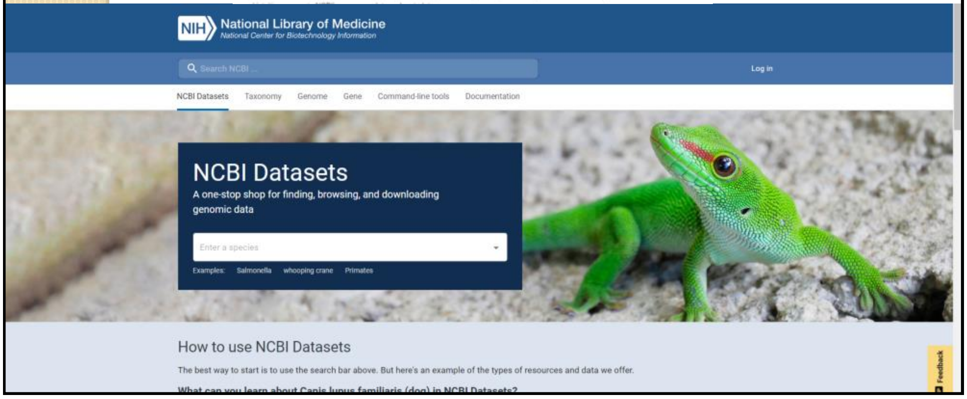
NCBI Outreach Events

- Workshops
  Hands-on, interactive training sessions
- Codeathons
  Collaborative project-based events
- Conferences
  NCBI activities at scientific meetings
- View all events

Subscribe To Our Mailing Lists

- NCBI Announcements
  Updates about our resources and events
- NIH Comparative Genomics Resource
  News on CGR-related data and tools
- View all mailing lists and news feeds

NCBI Insights Archives



NIH
National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
Log in
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation

NCBI Datasets

A one-stop shop for finding, browsing, and downloading genomic data

Enter a species
Examples: Salmonella whooping crane Primates
Feedback

How to use NCBI Datasets

The best way to start is to use the search bar above. But here's an example of the types of resources and data we offer.

What can you learn about Canis lupus familiaris (dog) in NCBI Datasets?

3

Genome assembly: relevance

- Genome assembly is needed when a genome is first sequenced. We can relate reads to chromosomes.
- For the human genome, the assembly is "frozen" as a snapshot every few years. The current assembly is GRCh38. (GRC refers to Genome Reference Consortium at http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/)
- For most human genome work we do not need to do "de novo" (from anew) assembly. Instead we map reads to a reference genome—one that is already assembled.
- Genome assembly is a crucial behind-the-scenes part of calling human genome (or other) variants.

4

Whereas early genome assembly projects were often aided by clone maps or other mapping data, many current assembly projects forego these scaffolding data and only assemble genomes into smaller segments. Recently, new technologies have been invented that allow chromosome-scale assembly at a lower cost and faster speed than traditional methods.

Many new technologies can now be used to create chromosome-scale assemblies without costly and time-consuming methods such as BAC-end sequencing and physical mapping.

Rice and Green, 2019. New Approaches for Genome Assembly and Scaffolding

5

Consequently, the contiguity of new genome assemblies decreased as high-throughput sequencing was widely adopted (Figure 1b,c)

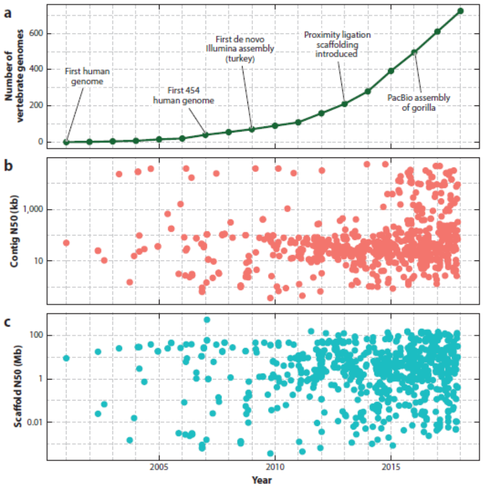


Figure 1

Timeline and statistics of vertebrate genome assemblies deposited in the National Center for Biotechnology Information's Genbank. Although second-generation sequencing has allowed more genomes to be published each year by making sequencing faster and cheaper, it has not increased the contiguity of published genomes. (*a*) Number of vertebrate genome assemblies available on Genbank at the end of each year, showing accelerating growth over the past decade. (*b*) Contig and (*c*) scaffold N50s of all vertebrate genome assemblies deposited in Genbank per year.

6

Genome Contig Assembly

No technology currently exists that can read DNA from one end to the other of even moderately sized chromosomes, which are typically tens or hundreds of millions of base pairs long. All current approaches for genome assembly read many segments that are considerably shorter than chromosomes.

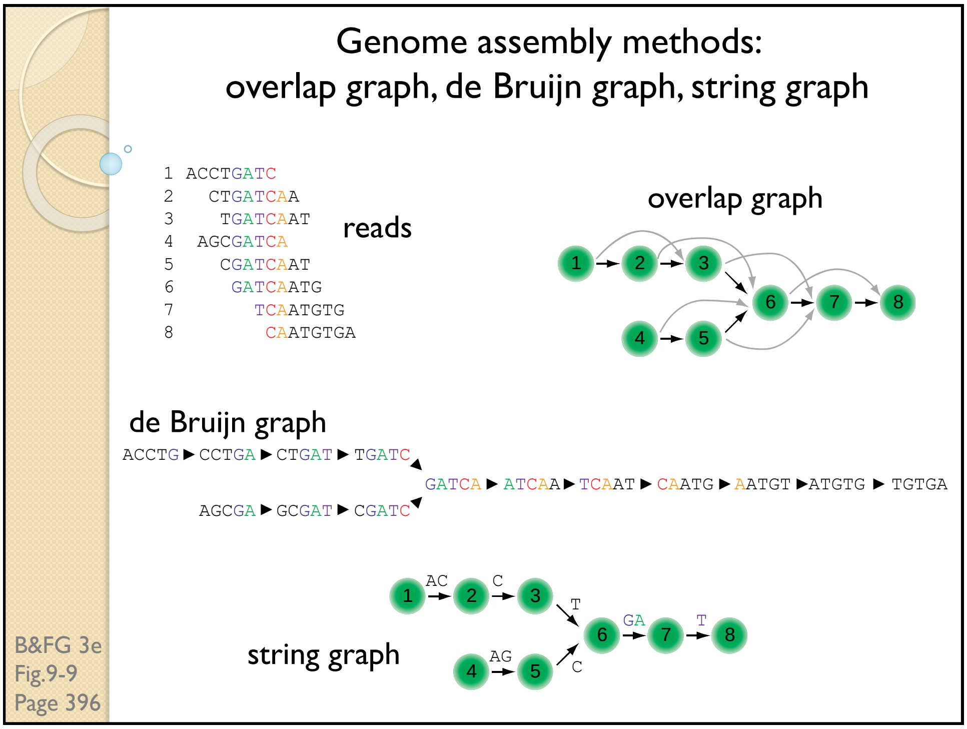
Both long-read sequencing technologies implement single-molecule sequencing methods and generate reads with a distribution of lengths that, for assembly purposes, target a range of tens to hundreds of kilobases (kb)- typically 10–25 kb for PacBio HiFi reads (also circular consensus sequencing, CCS), 10–40 kb for PacBio continuous long reads (CLR) and 10 kb-2 Megabases (Mb) for ONT, where the upper limit is constrained principally by properties of the input material (Payne et al., 2018).

7

Assembly algorithms

- overlap-layout-consensus: input DNA sequence reads are compared, all versus all, in the overlap step. The overlap-layout-consensus algorithm is based on identifying overlapping regions between reads and using these overlaps to construct longer contiguous sequences (contigs).
- de Bruijn graph: short words (k-mers) that are observed in the reads are the nodes of the graph, and edges are added when these k-mers are adjacent in sequence reads. In this process, each read is used to populate the graph but not compared directly to all the other reads.
- hybrid assembly

8



Genome assembly methods: overlap graph, de Bruijn graph, string graph

```
1 ACCTGATC
2   CTGATCAA
3    TGATCAAT
4  AGCGATCA
5    CGATCAAT
6     GATCAATG
7       TCAATGTG
8        CAATGTGA
```

reads

overlap graph

de Bruijn graph

ACCTG ▶ CCTGA ▶ CTGAT ▶ TGATC

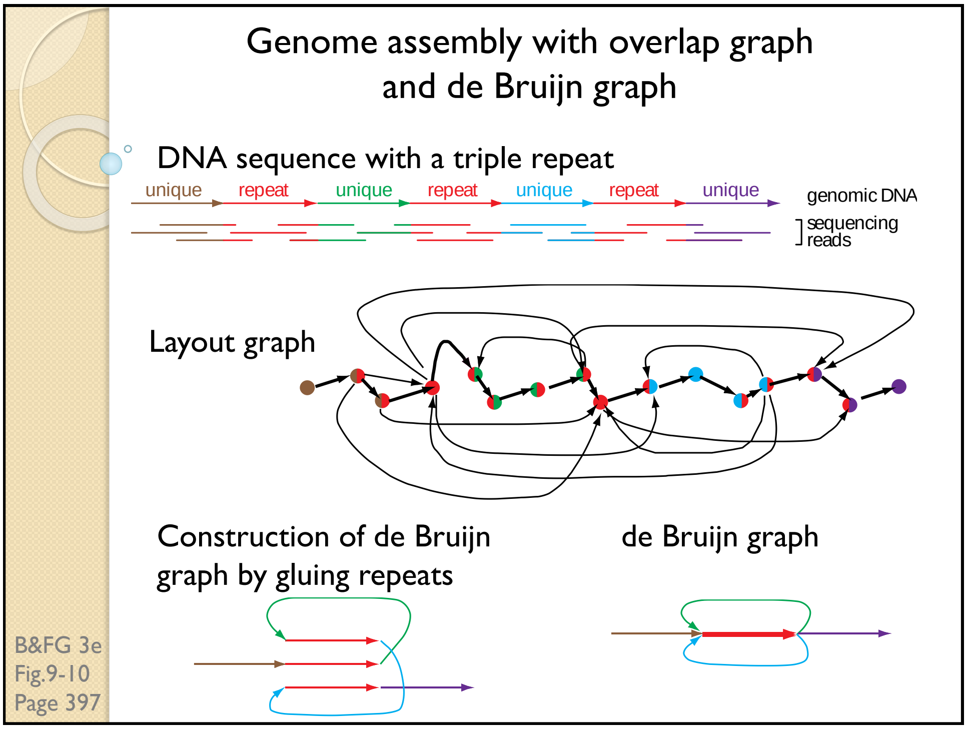
GATCA ▶ ATCAA ▶ TCAAT ▶ CAATG ▶ AATGT ▶ ATGTG ▶ TGTGA

AGCGA ▶ GCGAT ▶ CGATC

string graph

B&FG 3e
Fig.9-9
Page 396

9



Genome assembly with overlap graph and de Bruijn graph

DNA sequence with a triple repeat

Layout graph

Construction of de Bruijn graph by gluing repeats

de Bruijn graph

B&FG 3e
Fig.9-10
Page 397

10

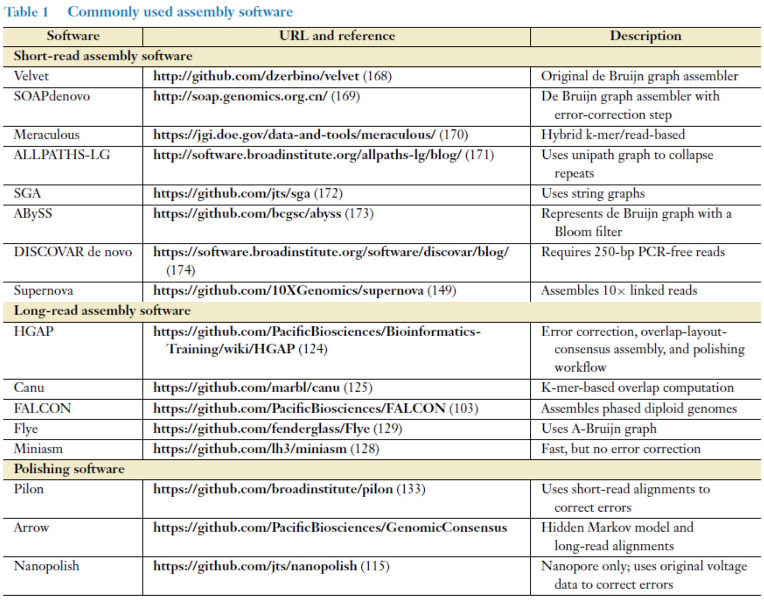


Table 1 Commonly used assembly software

Software	URL and reference	Description
Short-read assembly software		
Velvet	**http://github.com/dzerbino/velvet** (168)	Original de Bruijn graph assembler
SOAPdenovo	**http://soap.genomics.org.cn/** (169)	De Bruijn graph assembler with error-correction step
Meraculous	**https://jgi.doe.gov/data-and-tools/meraculous/** (170)	Hybrid k-mer/read-based
ALLPATHS-LG	**http://software.broadinstitute.org/allpaths-lg/blog/** (171)	Uses unipath graph to collapse repeats
SGA	**https://github.com/jts/sga** (172)	Uses string graphs
ABySS	**https://github.com/bcgsc/abyss** (173)	Represents de Bruijn graph with a Bloom filter
DISCOVAR de novo	**https://software.broadinstitute.org/software/discovar/blog/** (174)	Requires 250-bp PCR-free reads
Supernova	**https://github.com/10XGenomics/supernova** (149)	Assembles 10× linked reads
Long-read assembly software		
HGAP	**https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP** (124)	Error correction, overlap-layout-consensus assembly, and polishing workflow
Canu	**https://github.com/marbl/canu** (125)	K-mer-based overlap computation
FALCON	**https://github.com/PacificBiosciences/FALCON** (103)	Assembles phased diploid genomes
Flye	**https://github.com/fenderglass/Flye** (129)	Uses A-Bruijn graph
Miniasm	**https://github.com/lh3/miniasm** (128)	Fast, but no error correction
Polishing software		
Pilon	**https://github.com/broadinstitute/pilon** (133)	Uses short-read alignments to correct errors
Arrow	**https://github.com/PacificBiosciences/GenomicConsensus**	Hidden Markov model and long-read alignments
Nanopolish	**https://github.com/jts/nanopolish** (115)	Nanopore only; uses original voltage data to correct errors

Spades??

11

CREATING MORE CONTIGUOUS ASSEMBLIES WITH LONG READS

- Pacific Biosciences (SMRT, 2009)
  - The incorporation of fluorescently labeled nucleotides is detected and reveals the sequence of the analysed DNA strand.
  - PacBio offers Continuous Long Reads (CLR) and Circular Consensus Sequencing (CCS) reads also called High-Fidelity (HiFi).
- Oxford Nanopore Sequencing (2005, 1 channel flow cell, etc.)
  - It works by monitoring changes to an electrical current as nucleic acids are passed through a protein nanopore. The resulting signal is decoded to provide the specific DNA or RNA sequence.

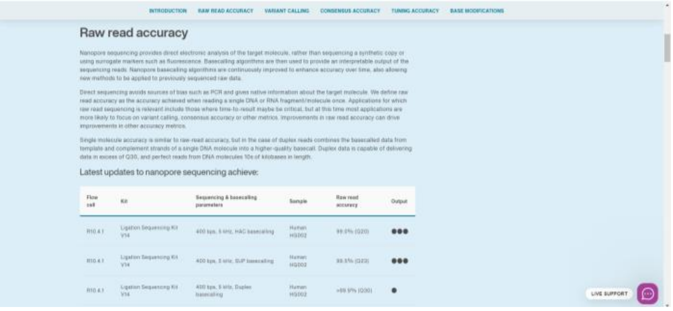
12

Hybrid assembly

- The accuracy of the short reads is used to decrease the error rate of the long reads from up to 20% to as low as 0.1%. Then, the corrected long reads are assembled using an algorithm such as overlap-layout-consensus.

Is it still necessary with new chemistry used by ONT and PacBio?

13



Raw read accuracy

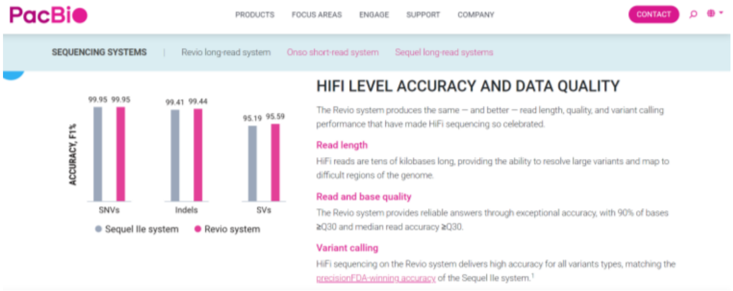
Nanopore sequencing provides direct electronic analysis of the target molecule, rather than sequencing a synthetic copy or using surrogate markers such as fluorescence. Basecalling algorithms are then used to provide an interpretable output of the sequencing reads. Nanopore basecalling algorithms are continuously improved to enhance accuracy over time, also allowing new methods to be applied to previously sequenced raw data.

Direct sequencing avoids sources of bias such as PCR and gives native information about the target molecule. We define raw read accuracy as the accuracy achieved when reading a single DNA or RNA fragment/molecule once. Applications for which raw read sequencing is relevant include those where time-to-result maybe be critical, but at this time most applications are more likely to focus on variant calling, consensus accuracy or other metrics. Improvements in raw read accuracy can drive improvements in other accuracy metrics.

Single molecule accuracy is similar to raw-read accuracy, but in the case of duplex reads combines the basecalled data from template and complement strands of a single DNA molecule into a higher-quality basecall. Duplex data is capable of delivering data in excess of Q30, and perfect reads from DNA molecules 10s of kilobases in length.

Latest updates to nanopore sequencing achieve:

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)	●



HIFI LEVEL ACCURACY AND DATA QUALITY

The Revio system produces the same — and better — read length, quality, and variant calling performance that have made HiFi sequencing so celebrated.

Read length

HiFi reads are tens of kilobases long, providing the ability to resolve large variants and map to difficult regions of the genome.

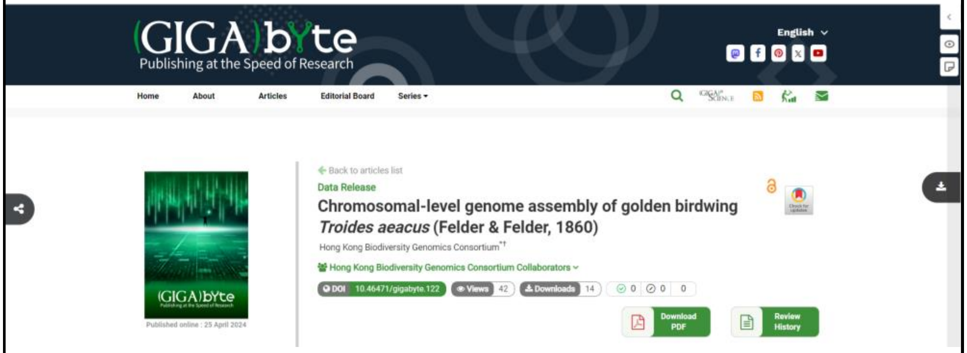
Read and base quality

The Revio system provides reliable answers through exceptional accuracy, with 90% of bases ≥Q30 and median read accuracy ≥Q30.

Variant calling

HiFi sequencing on the Revio system delivers high accuracy for all variants types, matching the precisionFDA-winning accuracy of the Sequel IIe system.[1]

14



Published online : 25 April 2024

Data Release

Chromosomal-level genome assembly of golden birdwing *Troides aeacus* (Felder & Felder, 1860)

Hong Kong Biodiversity Genomics Consortium*†

Hong Kong Biodiversity Genomics Consortium Collaborators

DOI 10.46471/gigabyte.122 | Views 42 | Downloads 14 | 0 | 0 | 0

https://gigabytejournal.com/articles/122

15

NEW APPROACHES FOR LONG-RANGE GENOME SCAFFOLDING

- method called Hi-C, Omni-C (Hi-C is a chromosome conformation capture (3C)-based technology to detect pair-wise chromatin interactions genome-wide)
- Linked-Read Sequencing (single-tube long fragment reads (stLFR) and haplo-tagging (Meier et al., 2020; Wang et al., 2019)
- Optical maps
- Synteny-Based Methods

16

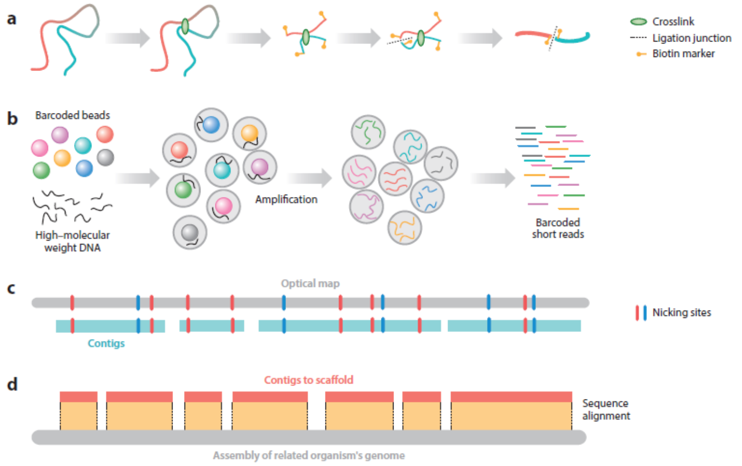


Figure 4

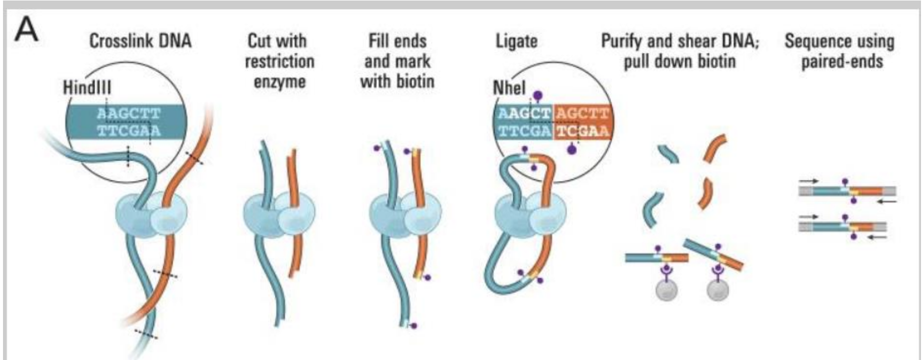
Overview of methods for long-range scaffolding. (*a*) In proximity ligation, chromatin is crosslinked and then restriction digested, ligated, and fragmented to create reads containing sequence from two different parts of the same chromosome. (*b*) In 10× linked-read sequencing, high–molecular weight DNA is combined with barcoded beads in oil droplets and then undergoes barcoding and amplification inside the droplets, resulting in reads with the same barcode that came from the same initial fragment of DNA. (*c*) BioNano optical maps are created by nicking high–molecular weight DNA with multiple nicking enzymes and attaching fluorescent markers at the nick sites. Contigs can then be aligned to the optical map by lining up nicking sequences in the contigs with the locations of fluorescent markers in the map. (*d*) In synteny-based approaches, contigs are mapped to the assembled genomes of one or more related species. These alignments imply the order and orientation of the aligned contigs.

17

As of April 2021, four biochemical companies (Arima Genomics, Dovetail Genomics, Phase Genomics, and Qiagen) manufacture Hi-C kits, which are formulated with different components and protocols. In general, conventional Hi-C kits employ a restriction enzyme or a cocktail of multiple restriction enzymes, whereas Omni-C employs a sequence-independent endonuclease (Table 1). In Omni-C, to capture more proximal contacts, disuccinimidyl glutarate (DSG) and formaldehyde are used for sample fixation (Nowak et al., 2005), which is now provided as a kit by Dovetail Genomics.

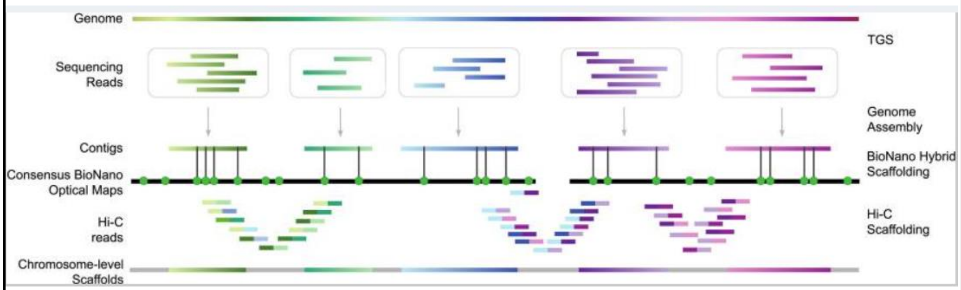
https://onlinelibrary.wiley.com/doi/full/10.1111/mec.16146

18


A
Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
HindIII
AAGCTT
TTCGAA
NheI
AAGCT AGCTT
TTCGA TCGAA

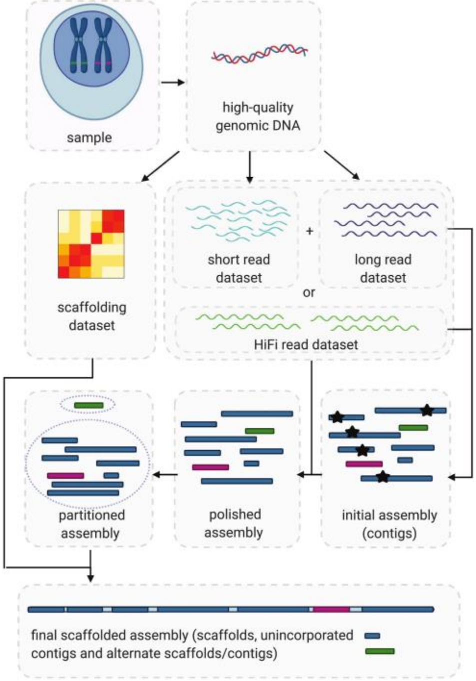
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2858594/

19


Genome
Sequencing Reads
Contigs
Consensus BioNano Optical Maps
Hi-C reads
Chromosome-level Scaffolds
TGS
Genome Assembly
BioNano Hybrid Scaffolding
Hi-C Scaffolding

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6926122/

20



https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13312

21

Key approaches for genome assembly that are generally recommended in all species include the following:

(a) Genome assemblies should include long-read sequencing except in rare cases where it is effectively impossible to acquire adequately preserved samples needed for HMW DNA standards.

(b) At least one scaffolding approach should be included with genome assembly such as Hi-C mapping or optical mapping (linked-read data is also appropriate but may not be available for most future projects).

(c) Short-read data should be included for genome polishing, error correction, k-mer analyses, and estimating the percent of reads that map back to assembly.

https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13312

22

Validation of genome assembly

- BUSCO (Benchmarking Universal SingleCopy Orthologs) OrthoDB (BUSCO uses set of genes which are present in 90 % of species in one copy only)
- QUAST

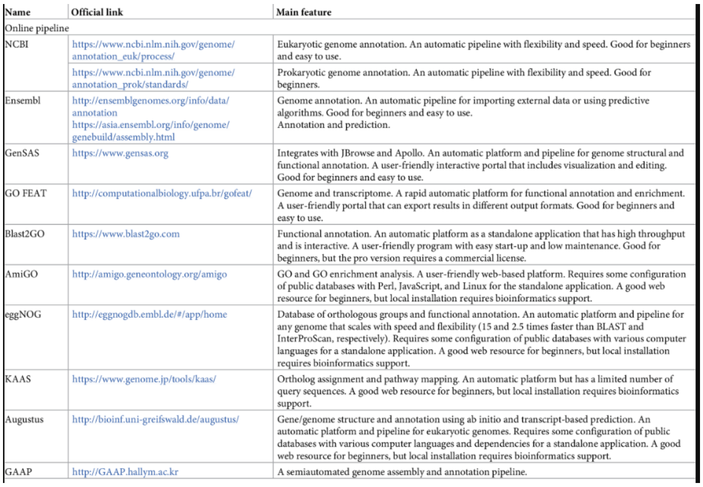
23

Genome annotation

- Genome annotation is the process of identifying and labeling functional elements within the genome, such as genes, regulatory regions, and repetitive elements.

24

Table 3. Commonly used genome annotation tools and programs.

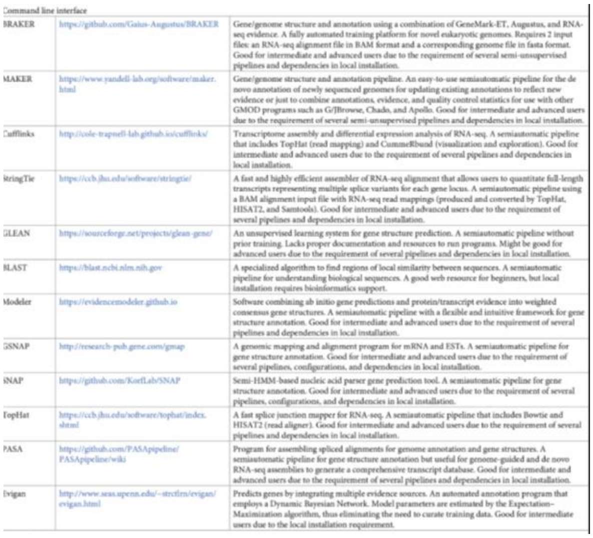


Name	Official link	Main feature
Online pipeline		
NCBI	https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/	Eukaryotic genome annotation. An automatic pipeline with flexibility and speed. Good for beginners and easy to use.
	https://www.ncbi.nlm.nih.gov/genome/annotation_prok/standards/	Prokaryotic genome annotation. An automatic pipeline with flexibility and speed. Good for beginners.
Ensembl	http://ensemblgenomes.org/info/data/annotation https://asia.ensembl.org/info/genome/genebuild/assembly.html	Genome annotation. An automatic pipeline for importing external data or using predictive algorithms. Good for beginners and easy to use. Annotation and prediction.
GenSAS	https://www.gensas.org	Integrates with JBrowse and Apollo. An automatic platform and pipeline for genome structural and functional annotation. A user-friendly interactive portal that includes visualization and editing. Good for beginners and easy to use.
GO FEAT	http://computationalbiology.ufpa.br/gofeat/	Genome and transcriptome. A rapid automatic platform for functional annotation and enrichment. A user-friendly portal that can export results in different output formats. Good for beginners and easy to use.
Blast2GO	https://www.blast2go.com	Functional annotation. An automatic platform as a standalone application that has high throughput and is interactive. A user-friendly program with easy start-up and low maintenance. Good for beginners, but the pro version requires a commercial license.
AmiGO	http://amigo.geneontology.org/amigo	GO and GO enrichment analysis. A user-friendly web-based platform. Requires some configuration of public databases with Perl, JavaScript, and Linux for the standalone application. A good web resource for beginners, but local installation requires bioinformatics support.
eggNOG	http://eggnogdb.embl.de/#/app/home	Database of orthologous groups and functional annotation. An automatic platform and pipeline for any genome that scales with speed and flexibility (15 and 2.5 times faster than BLAST and InterProScan, respectively). Requires some configuration of public databases with various computer languages for a standalone application. A good web resource for beginners, but local installation requires bioinformatics support.
KAAS	https://www.genome.jp/tools/kaas/	Ortholog assignment and pathway mapping. An automatic platform but has a limited number of query sequences. A good web resource for beginners, but local installation requires bioinformatics support.
Augustus	http://bioinf.uni-greifswald.de/augustus/	Gene/genome structure and annotation using ab initio and transcript-based prediction. An automatic platform and pipeline for eukaryotic genomes. Requires some configuration of public databases with various computer languages and dependencies for a standalone application. A good web resource for beginners, but local installation requires bioinformatics support.
GAAP	http://GAAP.hallym.ac.kr	A semiautomated genome assembly and annotation pipeline.

Jung H, Ventura T, Chung JS, Kim WJ, Nam BH, et al. (2020) Twelve quick steps for genome assembly and annotation in the classroom. PLOS Computational Biology 16(11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008325

25

Table 3. Commonly used genome annotation tools and programs.

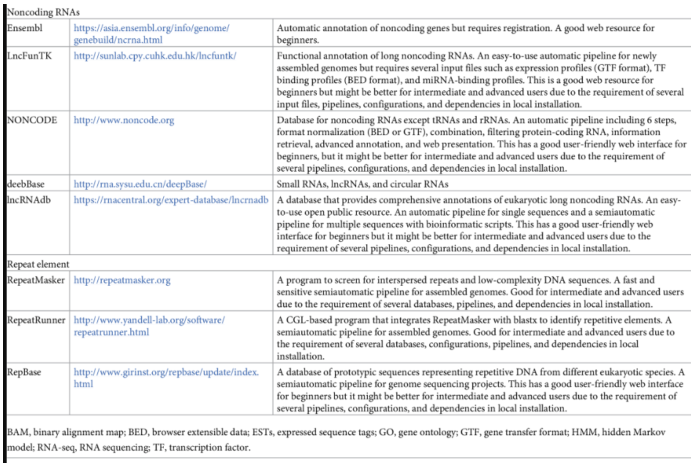


Command line interface		
BRAKER	https://github.com/Gaius-Augustus/BRAKER	Gene/genome structure and annotation using a combination of GeneMark-ET, Augustus, and RNA-seq evidence. A fully automated training platform for novel eukaryotic genomes. Requires 2 input files: an RNA-seq alignment file in BAM format and a corresponding genome file in fasta format. Good for intermediate and advanced users due to the requirement of several semi-unsupervised pipelines and dependencies in local installation.
MAKER	https://www.yandell-lab.org/software/maker.html	Gene/genome structure and annotation pipeline. An easy-to-use semiautomatic pipeline for the de novo annotation of newly sequenced genomes for updating existing annotations to reflect new evidence or just to combine annotations, evidence, and quality control statistics for use with other GMOD programs such as G/JBrowse, Chado, and Apollo. Good for intermediate and advanced users due to the requirement of several semi-unsupervised pipelines and dependencies in local installation.
Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/	Transcriptome assembly and differential expression analysis of RNA-seq. A semiautomatic pipeline that includes TopHat (read mapping) and CummeRbund (visualization and exploration). Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
StringTie	https://ccb.jhu.edu/software/stringtie/	A fast and highly efficient assembler of RNA-seq alignment that allows users to quantitate full-length transcripts representing multiple splice variants for each gene locus. A semiautomatic pipeline using a BAM alignment input file with RNA-seq read mappings (produced and converted by TopHat, HISAT2, and Samtools). Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
GLEAN	https://sourceforge.net/projects/glean-gene/	An unsupervised learning system for gene structure prediction. A semiautomatic pipeline without prior training. Lacks proper documentation and resources to run programs. Might be good for advanced users due to the requirement of several pipelines and dependencies in local installation.
BLAST	https://blast.ncbi.nlm.nih.gov	A specialized algorithm to find regions of local similarity between sequences. A semiautomatic pipeline for understanding biological sequences. A good web resource for beginners, but local installation requires bioinformatics support.
Modeler	https://evidencemodeler.github.io	Software combining ab initio gene predictions and protein/transcript evidence into weighted consensus gene structures. A semiautomatic pipeline with a flexible and intuitive framework for gene structure annotation. Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
GSNAP	http://research-pub.gene.com/gmap	A genomic mapping and alignment program for mRNA and ESTs. A semiautomatic pipeline for gene structure annotation. Good for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
SNAP	https://github.com/KorfLab/SNAP	Semi-HMM-based nucleic acid parser gene prediction tool. A semiautomatic pipeline for gene structure annotation. Good for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
TopHat	https://ccb.jhu.edu/software/tophat/index.shtml	A fast splice junction mapper for RNA-seq. A semiautomatic pipeline that includes Bowtie and HISAT2 (read aligner). Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
PASA	https://github.com/PASApipeline/PASApipeline/wiki	Program for assembling spliced alignments for genome annotation and gene structures. A semiautomatic pipeline for gene structure annotation but useful for genome-guided and de novo RNA-seq assemblies to generate a comprehensive transcript database. Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
Evigan	http://www.seas.upenn.edu/~strctlrn/evigan/evigan.html	Predicts genes by integrating multiple evidence sources. An automated annotation program that employs a Dynamic Bayesian Network. Model parameters are estimated by the Expectation–Maximization algorithm, thus eliminating the need to curate training data. Good for intermediate users due to the local installation requirement.

Jung H, Ventura T, Chung JS, Kim WJ, Nam BH, et al. (2020) Twelve quick steps for genome assembly and annotation in the classroom. PLOS Computational Biology 16(11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008325

26

Table 3. Commonly used genome annotation tools and programs.



Noncoding RNAs		
Ensembl	https://asia.ensembl.org/info/genome/genebuild/ncrna.html	Automatic annotation of noncoding genes but requires registration. A good web resource for beginners.
LncFunTK	http://sunlab.cpy.cuhk.edu.hk/lncfuntk/	Functional annotation of long noncoding RNAs. An easy-to-use automatic pipeline for newly assembled genomes but requires several input files such as expression profiles (GTF format), TF binding profiles (BED format), and miRNA-binding profiles. This is a good web resource for beginners but might be better for intermediate and advanced users due to the requirement of several input files, pipelines, configurations, and dependencies in local installation.
NONCODE	http://www.noncode.org	Database for noncoding RNAs except tRNAs and rRNAs. An automatic pipeline including 6 steps, format normalization (BED or GTF), combination, filtering protein-coding RNA, information retrieval, advanced annotation, and web presentation. This has a good user-friendly web interface for beginners, but it might be better for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
deebBase	http://rna.sysu.edu.cn/deepBase/	Small RNAs, lncRNAs, and circular RNAs
lncRNAdb	https://rnacentral.org/expert-database/lncrnadb	A database that provides comprehensive annotations of eukaryotic long noncoding RNAs. An easy-to-use open public resource. An automatic pipeline for single sequences and a semiautomatic pipeline for multiple sequences with bioinformatic scripts. This has a good user-friendly web interface for beginners but it might be better for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
Repeat element		
RepeatMasker	http://repeatmasker.org	A program to screen for interspersed repeats and low-complexity DNA sequences. A fast and sensitive semiautomatic pipeline for assembled genomes. Good for intermediate and advanced users due to the requirement of several databases, pipelines, and dependencies in local installation.
RepeatRunner	http://www.yandell-lab.org/software/repeatrunner.html	A CGL-based program that integrates RepeatMasker with blastx to identify repetitive elements. A semiautomatic pipeline for assembled genomes. Good for intermediate and advanced users due to the requirement of several databases, configurations, pipelines, and dependencies in local installation.
RepBase	http://www.girinst.org/repbase/update/index.html	A database of prototypic sequences representing repetitive DNA from different eukaryotic species. A semiautomatic pipeline for genome sequencing projects. This has a good user-friendly web interface for beginners but it might be better for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.

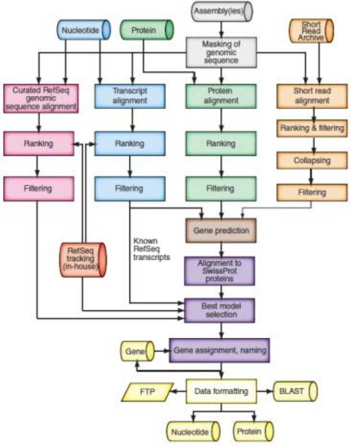
BAM, binary alignment map; BED, browser extensible data; ESTs, expressed sequence tags; GO, gene ontology; GTF, gene transfer format; HMM, hidden Markov model; RNA-seq, RNA sequencing; TF, transcription factor.

Jung H, Ventura T, Chung JS, Kim WJ, Nam BH, et al. (2020) Twelve quick steps for genome assembly and annotation in the classroom. PLOS Computational Biology 16(11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008325

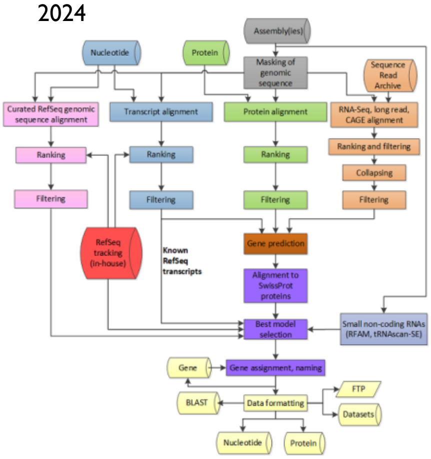
27

NCBI eukaryotic annotation pipeline

2015



2024



Conceptually similar to Ensembl pipeline

B&FG 3e
Fig. 15.14
Page 740

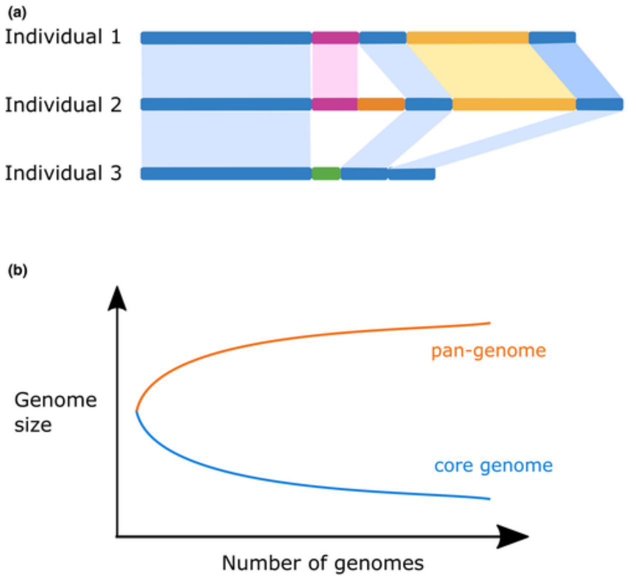
28

Biological challenges

- Repetitive regions (expanded gene families, complex repeats, highly repetitive regions such as centromeres and telomeres, and sex chromosomes, or at least portions of them.)
- Ploidy
- Pan and core genomes (The pan-genome represents all sequences among all of the DNA sequences that occur in a species whereas the core-genome is the DNA that is shared among all sequenced individuals.)
- *For example, the comparison of eight chromosome-level assemblies of Arabidopsis thaliana accessions revealed a core-genome, shared by all accessions, of ~105 Mb and ~24,000 genes, whereas the pan-genome was ~135 Mb in length and included ~30,000 genes (Jiao & Schneeberger, 2020), highlighting the vast amount of sequence data, including genes, that are missed by a single reference genome assembly.*

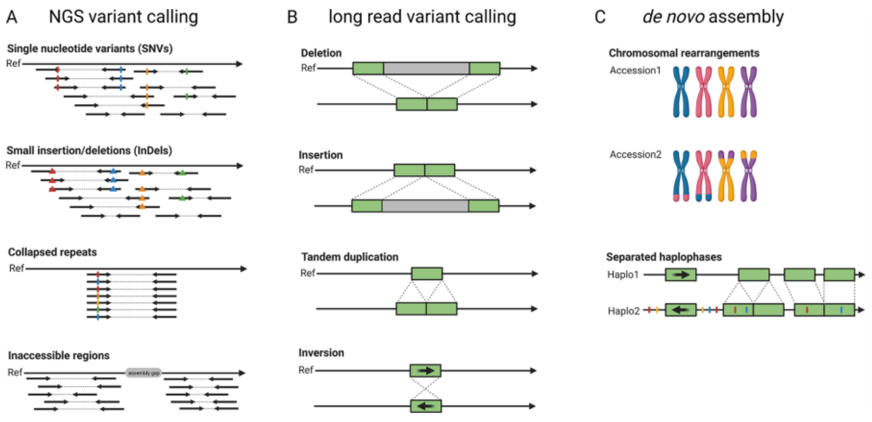
https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13312

29



https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13312

30



https://www.cambridge.org/core/journals/quantitative-plant-biology/article/plant-genome-sequence-assembly-in-the-era-of-long-reads-progress-challenges-and-future-directions/3932B956CC468D07D3E14525FEAFEC3E

31

Why Is Chromosome-Scale Assembly Important?

- Cis-regulatory elements and the complexity of regulatory architecture
- Recombination
- Genetic association studies
- Chromosome evolution

32