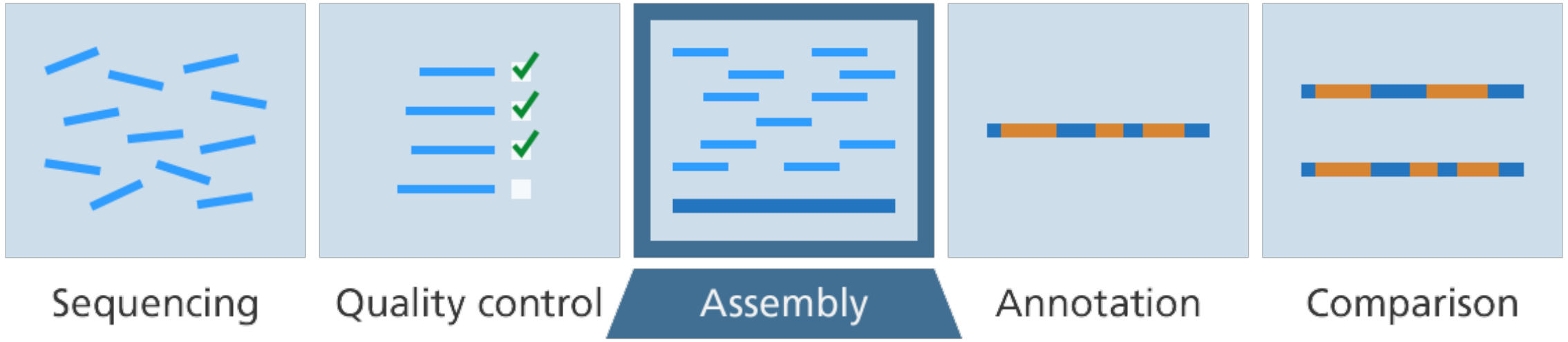


Genome assembly and annotation

Jun Huang; Tej Man Tamang; Bliss Betzen

Photo from https://www.yourgenome.org

1. Investigate the properties of the genome you study

- Genome size

❑ (the bigger the genome, the more data is needed: 60x for illumina)

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

Photo from BioNinja

2. Extract high quality DNA

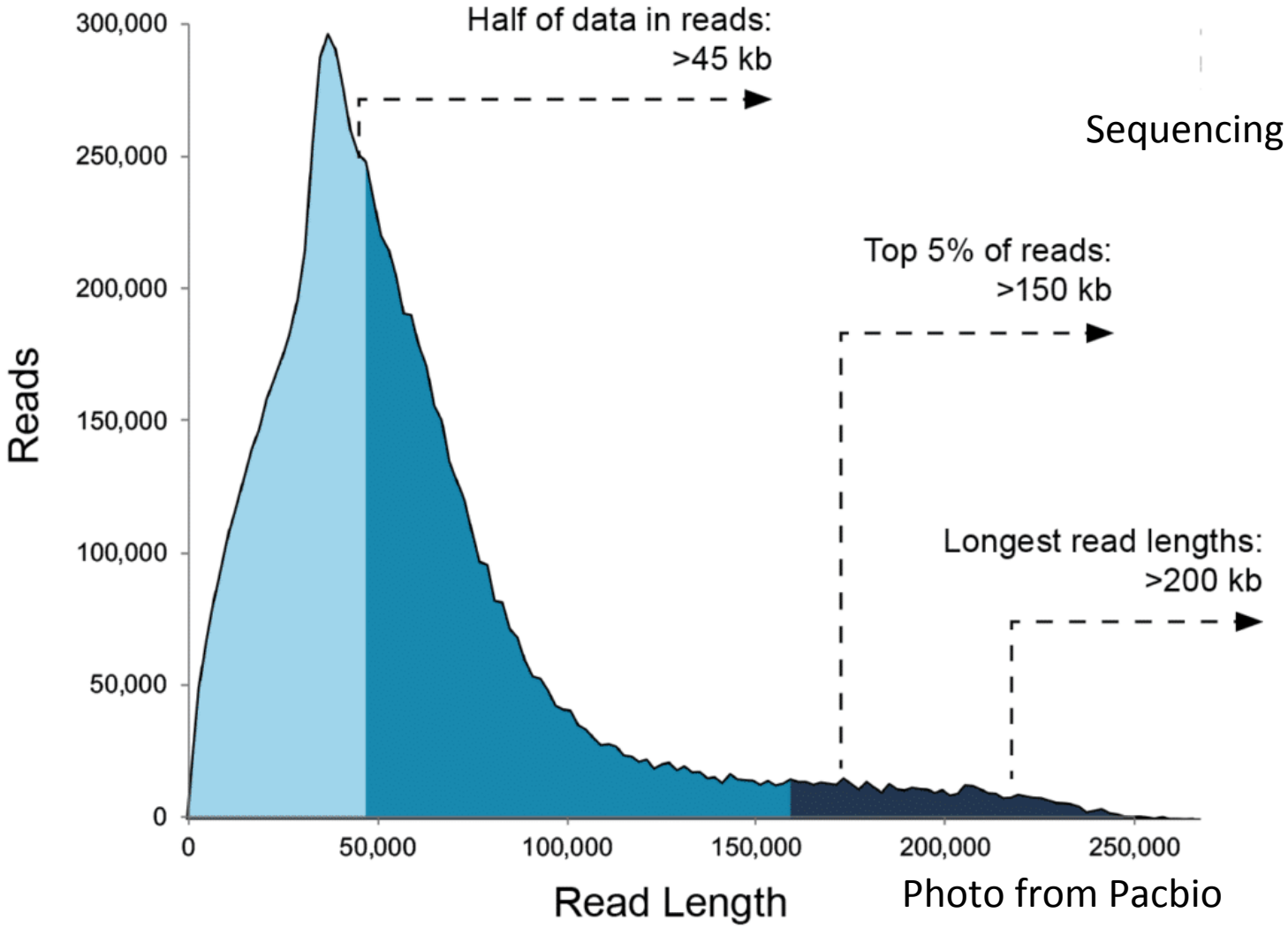
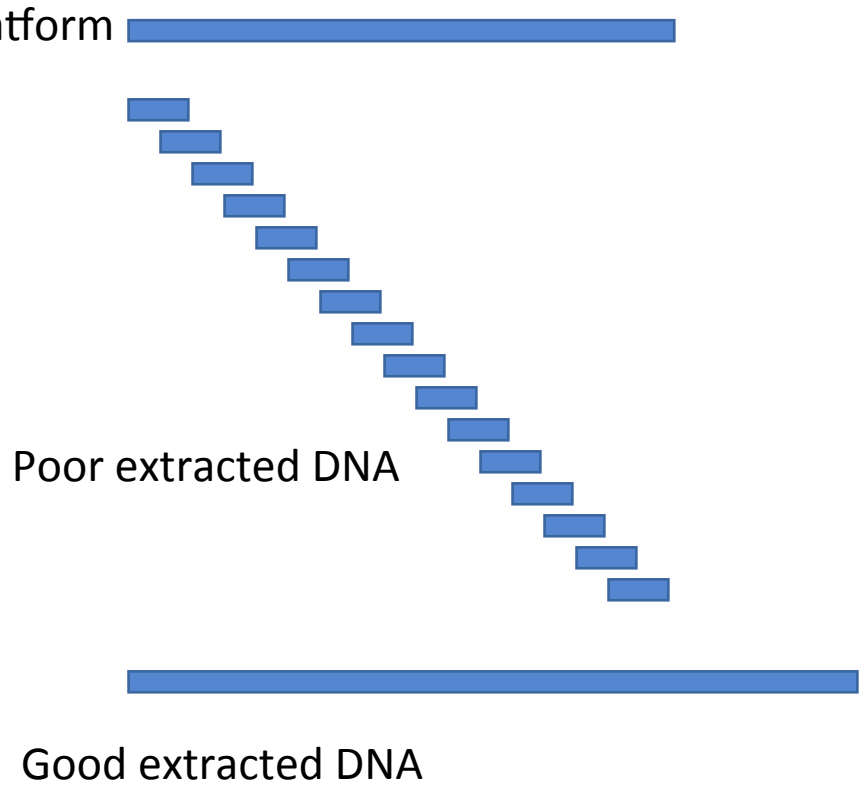


Photo from Pacbio



Read length data shown above from a 35 kb size-selected human library using the SMRTbell Express Template Prep Kit on a Sequel System

3. Choose an appropriate sequencing technology

Platform	Amplification Method	Sequencing Method	Detection Method	Average read length
Illumina (HiSeq, MiSeq etc)	bridge PCR	sequencing by synthesis	Light	100-200 bp
Life Tech Ion Torrent / Proton	emulsion PCR	Ion semiconductor sequencing	pH	200-400 bp
Roche 454	emulsion PCR	Pyrosequencing, cleavage of released pyrophosphate	light	700 bp
Life Tech SOLiD	emulsion PCR	sequencing by ligation of hybridizing labeled oligos	light	100 bp
Pacific Biosciences PacBio / Sequel System	No amplification, single-molecule sequencing	polymerase incorporating colored NTPs	light	10 kb, 50% >20 kb, 5% > 30 kb
Oxford Nanopore MinION	No amplification, single molecule nanopore sequencing	DNA molecule traverses pore	current	> 5.4 kb

Further reading, great lecture: Sequencing technology - Past, Present and Future, http://www.molgen.mpg.de/899148/OWS2013_NGS.pdf

A near finished genome assembly for *M. oryzae* was acquired by Nanopore sequencing

Table 2. Summary statistics for *M. oryzae* genomes assembled from nanopore reads

M. oryzae isolate	Host	CANU version[1]	# Contigs	Assemby length (bp)	N25 (bp)	N50 (bp)	N75 (bp)	Max length (bp)	Mean length (bp)	Min length (bp)	GenBank Accession[2]
BTJP4-1[3]	*Triticum aestivum*	1.7	59	44,506,712	6,840,169	4,344,896	3,373,527	7,174,201	754,351	13,054	GCA_900474225.2
BTMP13-1[3]	*Triticum aestivum*	1.6	16	43,978,087	7,837,192	6,037,509	4,385,994	10,783,101	2,748,630	7,390	GCA_900474375.2
BTGP1-b[4]	*Triticum aestivum*	1.7	74	44,406,102	3,690,742	2,814,025	1,269,883	6,505,875	600,082	5,533	GCA_900474635.2
BTGP6-f[4]	*Triticum aestivum*	1.7	57	44,234,333	5,243,043	3,705,381	2,027,069	6,048,575	776,041	8,312	GCA_900474435.2
BR32	*Triticum aestivum*	1.6	21	41,471,325	11,366,628	5,047,693	3,895,412	11,366,628	1,974,825	18,099	GCA_900474545.2
FR13	*Oryza sativa*	1.7	46	46,415,940	6,634,785	5,357,033	2,121,955	7,257,380	1,009,042	19,712	GCA_900474655.2
US71	*Setaria italica*	1.7	84	45,673,611	3,535,243	2,015,667	979,014	4,788,334	543,733	6,882	GCA_900474175.2
CD156	*Eleusine indica*	1.7	44	43,859,562	6,040,961	4,257,479	3,430,372	6,066,300	996,808	8,777	GCA_900474475.2

[1]Genome assembly was performed by Future Genomics Technologies using Canu assembly software
[2]Sequence assemblies were deposited at European Nucleotide Archive (ENA) with study accession PRJEB27137
[3]*M. oryzae* isolates collected from wheat during 2016 epidemic in Bangladesh
[4]*M. oryzae* isolates collected from wheat during 2017 epidemic in Bangladesh

Win et al.,2019

4. Estimate the necessary computational resources

SPAdes:
Assembly of small genome using short reads.
Smartdenovo:
De novo assembler for Pacbio and Nanopore data.
REPET:
Software suite dedicated to detect, classify and annotate repeats
Eugene:
Open gene finder for genome

Device specifications

Device name	DESKTOP-7KAJT8L
Processor	Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz 2.60 GHz
Installed RAM	16.0 GB (15.9 GB usable)
Device ID	A8521EEB-F12E-4520-8837-CB4C293FD721
Product ID	00325-95800-00000-AAOEM
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

Table 1. **Examples of time and computer resources used by software dedicated to assembly and annotation.** SPAdes is an assembler designed for the assembly of small genomes using short reads. Smartdenovo is a *de novo* assembler for PacBio and Oxford Nanopore (ONT) data. The REPET package is a software suite dedicated to detect, classify and annotate repeats. EuGene is an open integrative gene finder for eukaryotic and prokaryotic genomes. Processing time and RAM used will be affected by amount of input data, complexity of data, and genome size.

Reference Genome	Size	Software	Input (space used on disk)	CPU/RAM Available	Real time	Max RAM Used
Alivibrio wodanis	4.97Mb	SPAdes v3.10	200x Illumina reads (760 MB)	4 CPU/16GB RAM	2h17m3s	2,94GB
				12 CPU/256GB RAM	38m8s	9,37GB
Caenorhabditis elegans	100.2Mb	Smartdenovo	20x Pacbio P6C4 Corrected long reads (1,9 GB)	8 CPU/16GB RAM	24m47s	1,92GB
			80x Pacbio P6C4 Corrected long reads (7,6 GB)	8 CPU/16GB RAM	5h38m16s	7,29GB
		REPET v2.5	*C. Elegans* genome (100 MB) Repbase aa 20.05 (20 MB) Pfam 27 (GypsyDB) (1,2 GB) rRNA from eukaryota (2,6 MB)	8 CPU/16 GB RAM	1h53m11s + 19h9m40s	8,96GB
		Eugene v4.2a	*C. Elegans* genome (100 MB) Repbase aa 20.05 (20 MB) Proteins sequences (swissprot) (2,8 MB) ESTs sequences (29 MB)	8 CPU/32 GB RAM	5h2m30s	16,94GB
Arabidopsis thaliana	134.6Mb	Smartdenovo	20x Pacbio P5C3 corrected long reads (2,7 GB)	8 CPU/16GB RAM	1h16m20s	2,4GB
		REPET v2.5	*A. Thaliana* genome (130 MB) Repbase aa 20.05 (20 MB) Pfam 27 (GypsyDB) (1,2 GB) rRNA from eukaryota (2,6 MB)	8 CPU/16 GB RAM	5h6m23s + 33h10m34s	10,25GB
		Eugene v4.2a	*A. Thaliana* genome (130 MB) Repbase aa 20.05 (20 MB) Proteins sequences (swissprot) (9,2 MB) ESTs sequences (31 MB)	8 CPU/32 GB RAM	6h17m18s	17,25GB
Theobroma cacao	324.7Mb	Eugene v4.2a	*T. Cacao* genome (315 MB) Repbase aa 20.05 (20 MB) Proteins sequences (swissprot) (31 MB) ESTs sequences (103 MB)	8 CPU/188 GB RAM	41h27m13s	72,5GB

5. Assemble your genome

F1000Research 2018, 7(ELIXIR):148 Last updated: 18 FEB 2019

Victoria Dominguez Del Angel et al.,2019

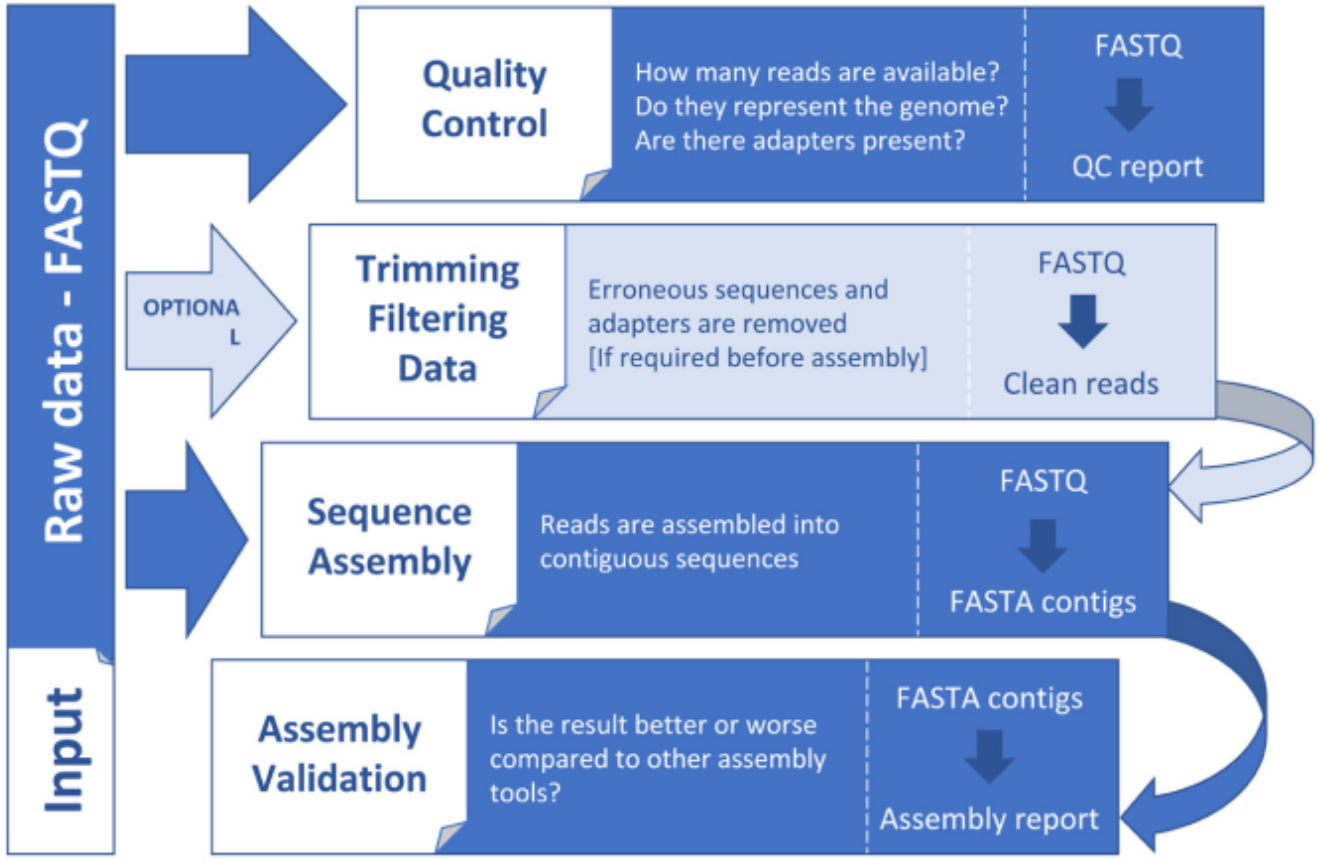
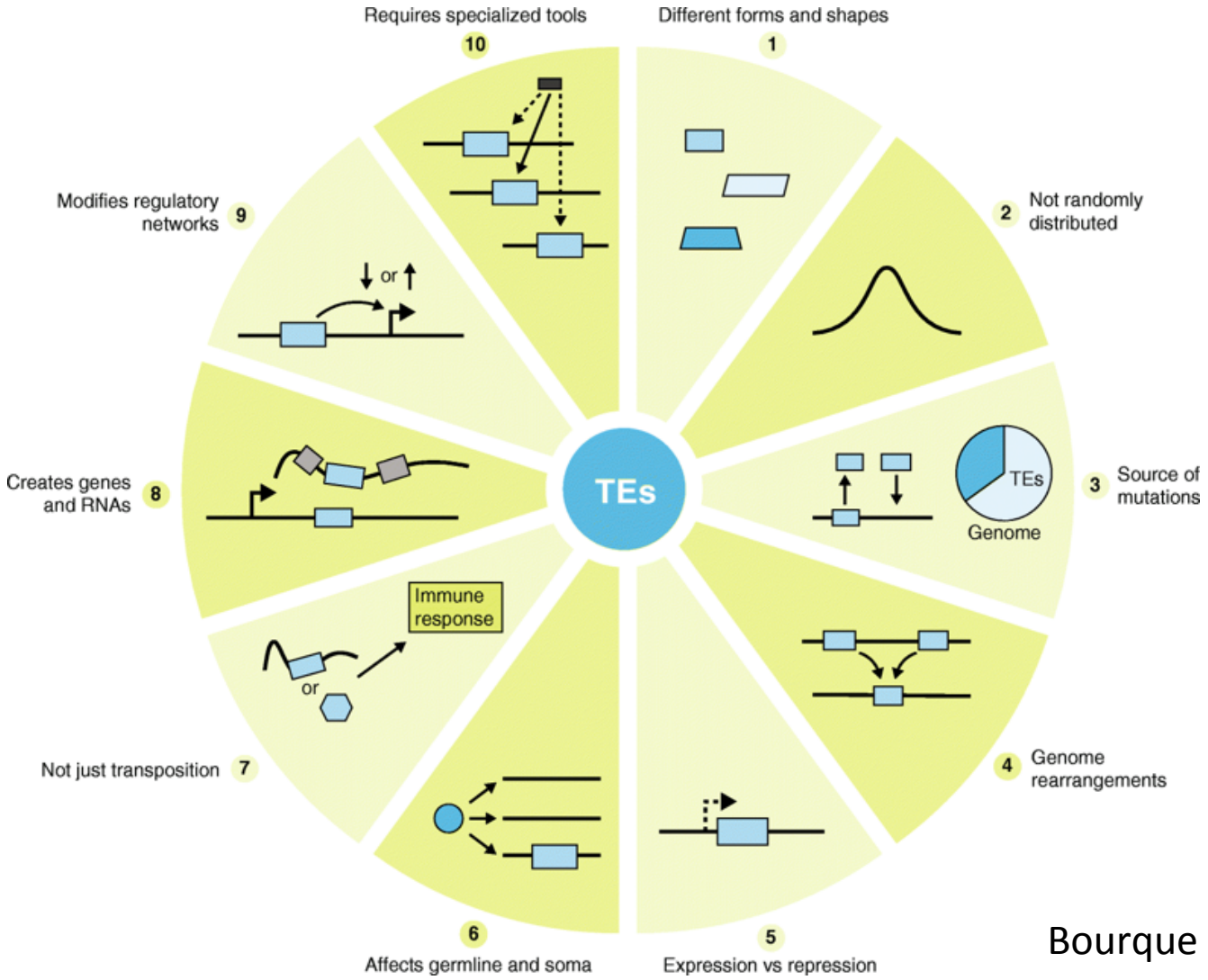


Figure 2. General steps in a genome assembly workflow. Input and output data are indicated for each step.

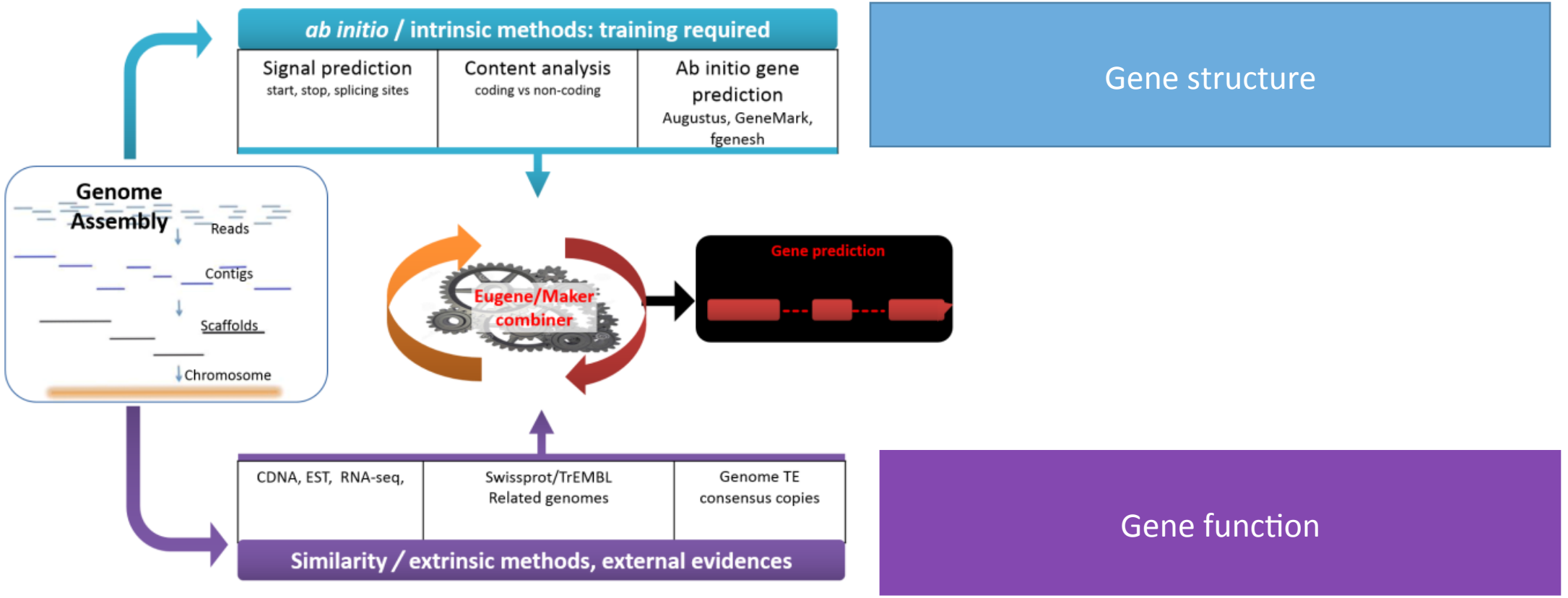
6. Do not neglect to annotate Transposable Elements

Two useful tools for TE annotation
1. RepeatMasker
2. REPET package



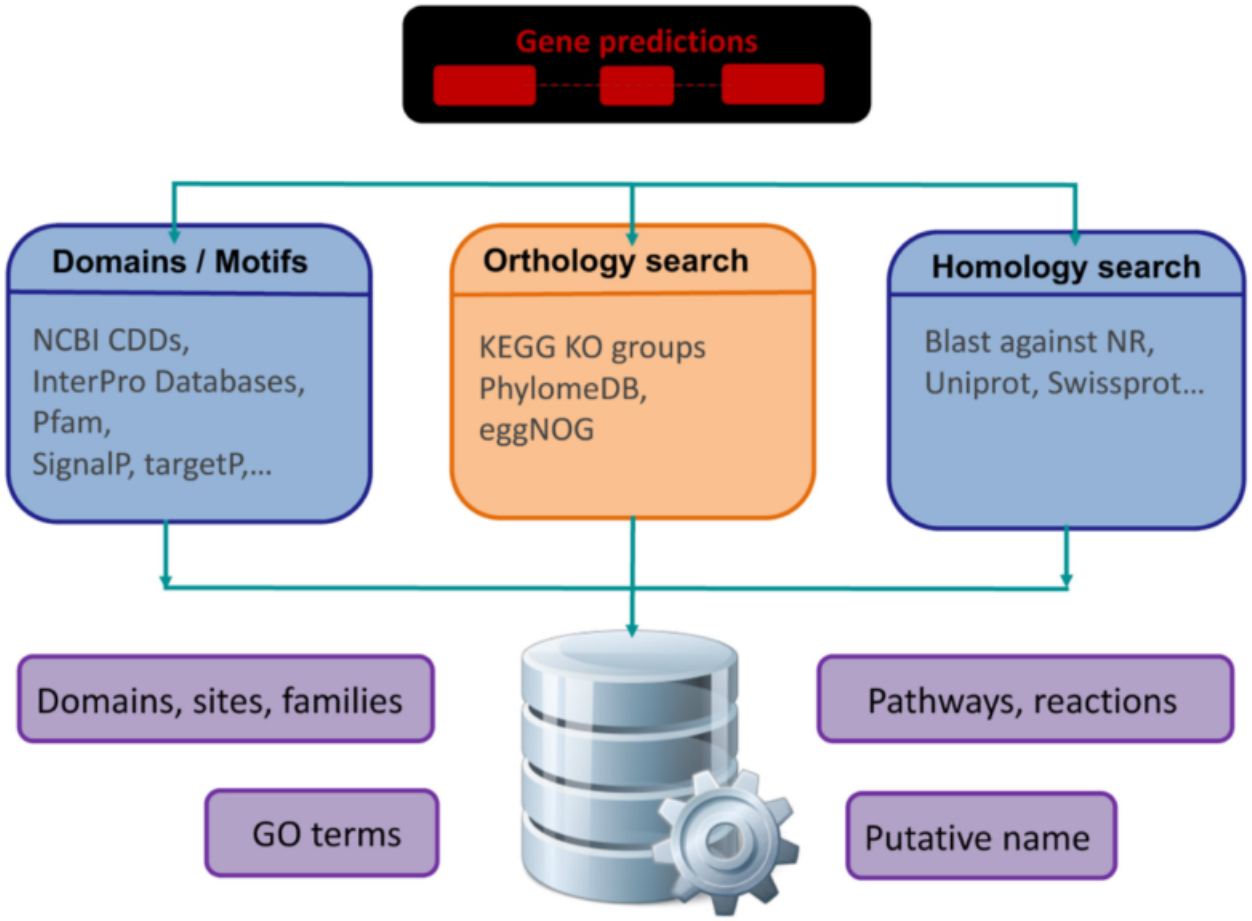
Bourque et al., 2018

7. Annotate genes with high quality experimental evidence



Victoria Dominguez Del Angel et al.,2019

Functional Annotation Pipelines.



Victoria Dominguez Del Angel et al.,2019

8. Use well-established output formats and submit your data to suitable repositories

9. Ensure your methods are computationally repeatable and reproducible

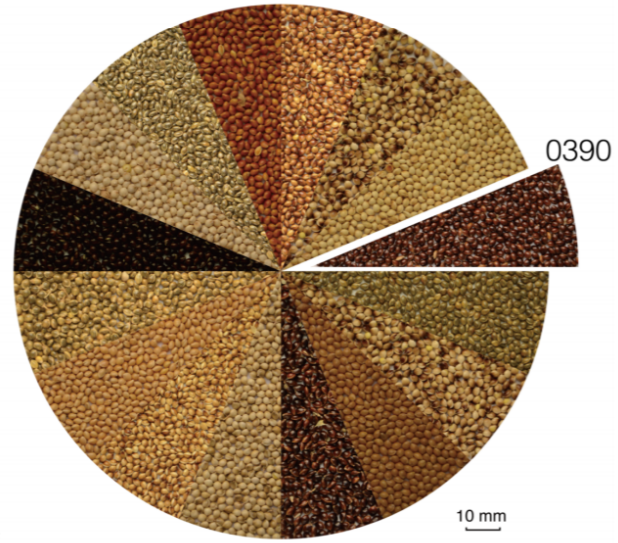
10. Investigate, re-analyse, re-annotate

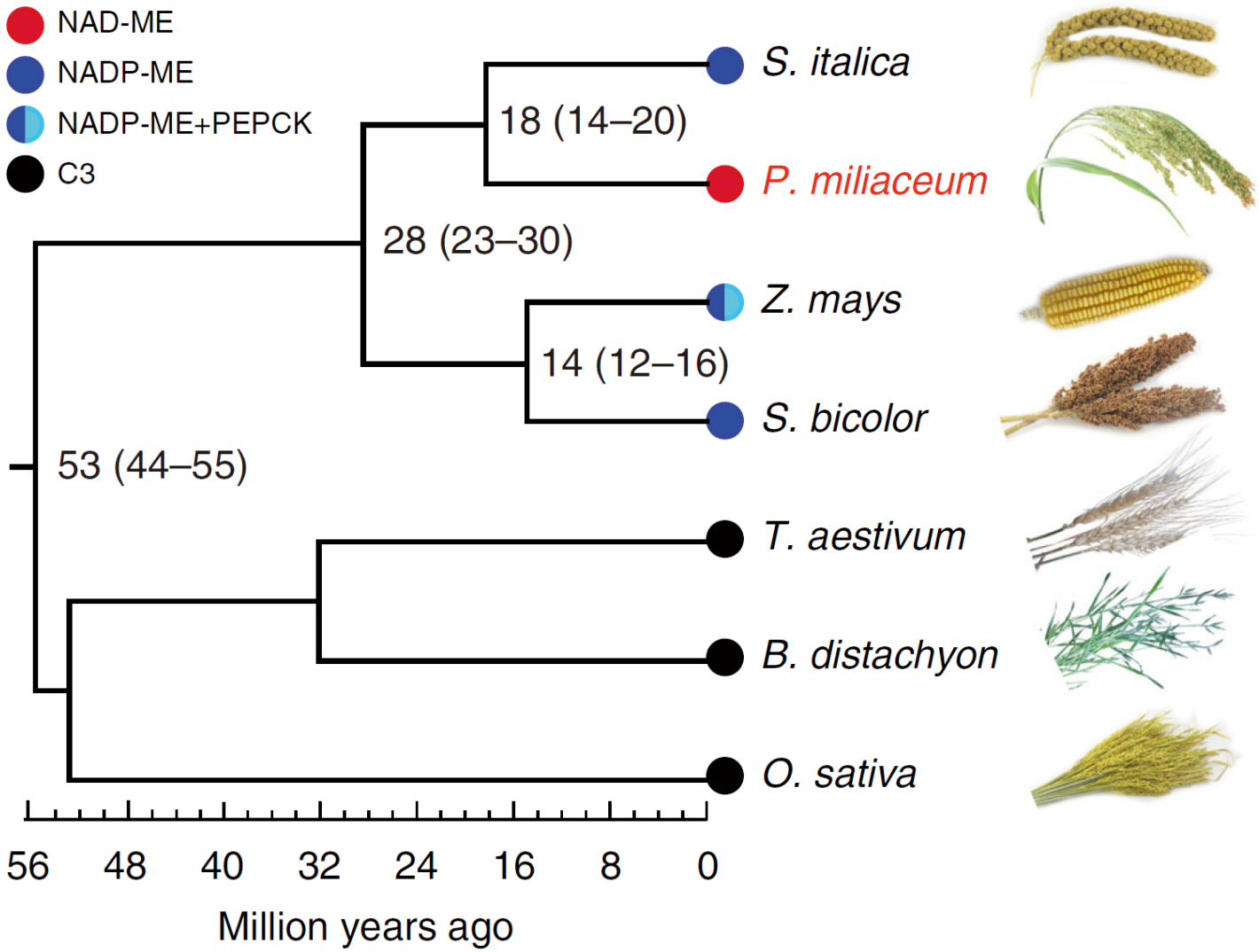
Case Study I: The genome of broomcorn millet

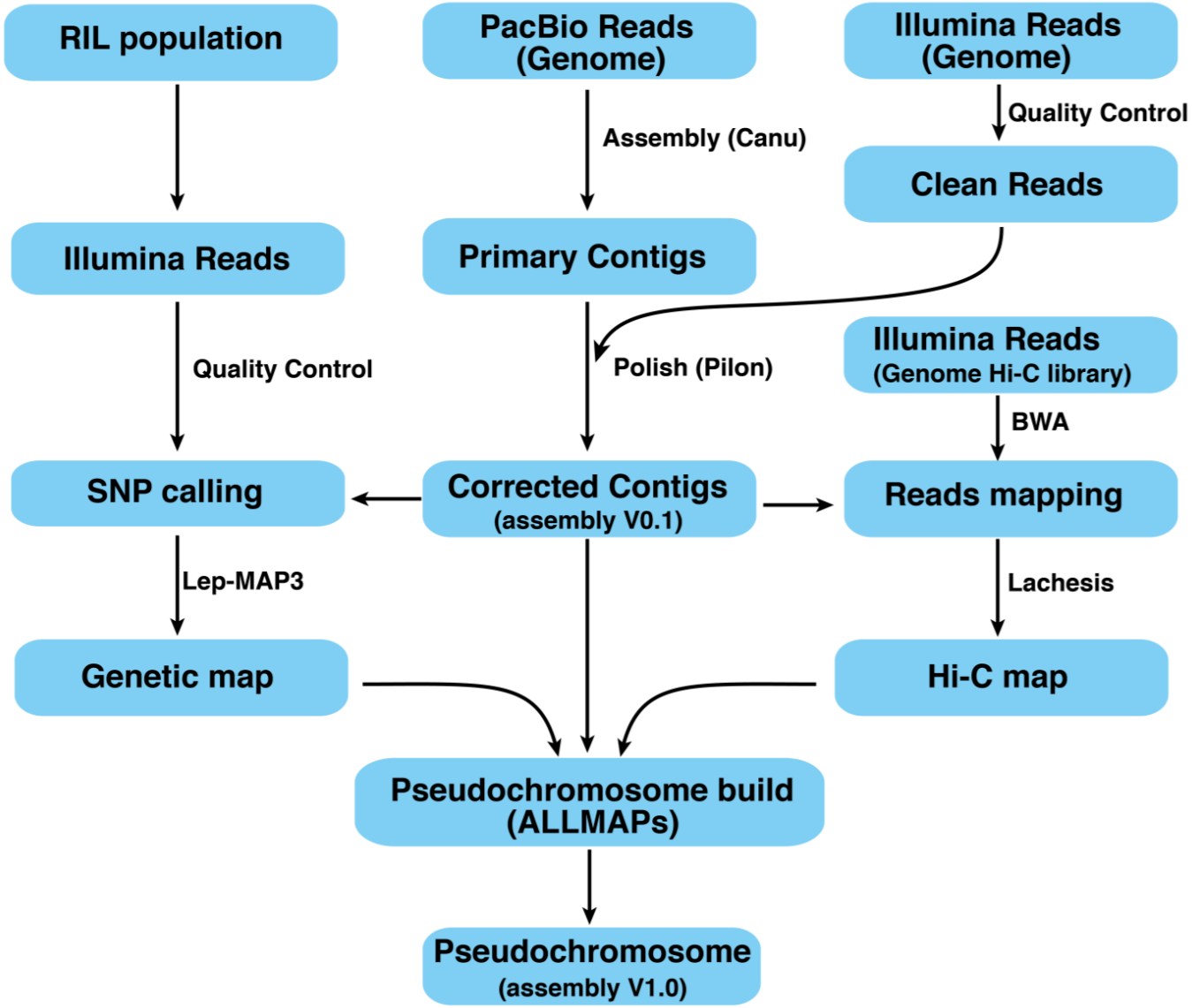
Zhou et al. 2019, Nature communications

Broomcorn millet *(Panicum Miliaceum)*

- Also known as common millet, proso millet and hog millet
- Drought-tolerant cereal
- Gluten-free, nutritious, higher protein content, minerals and antioxidants
- Origin: Northern China
- C_4 photosynthesis
- Allotetraploid with 36 chromosomes




NAD-ME
NADP-ME
NADP-ME+PEPCK
C3
S. italica
P. miliaceum
Z. mays
S. bicolor
T. aestivum
B. distachyon
O. sativa
18 (14–20)
28 (23–30)
14 (12–16)
53 (44–55)
56
48
40
32
24
16
8
0
Million years ago


RIL population
PacBio Reads (Genome)
Illumina Reads (Genome)
Quality Control
Assembly (Canu)
Clean Reads
Illumina Reads
Primary Contigs
Quality Control
Polish (Pilon)
Illumina Reads (Genome Hi-C library)
BWA
SNP calling
Corrected Contigs (assembly V0.1)
Reads mapping
Lep-MAP3
Lachesis
Genetic map
Hi-C map
Pseudochromosome build (ALLMAPs)
Pseudochromosome (assembly V1.0)

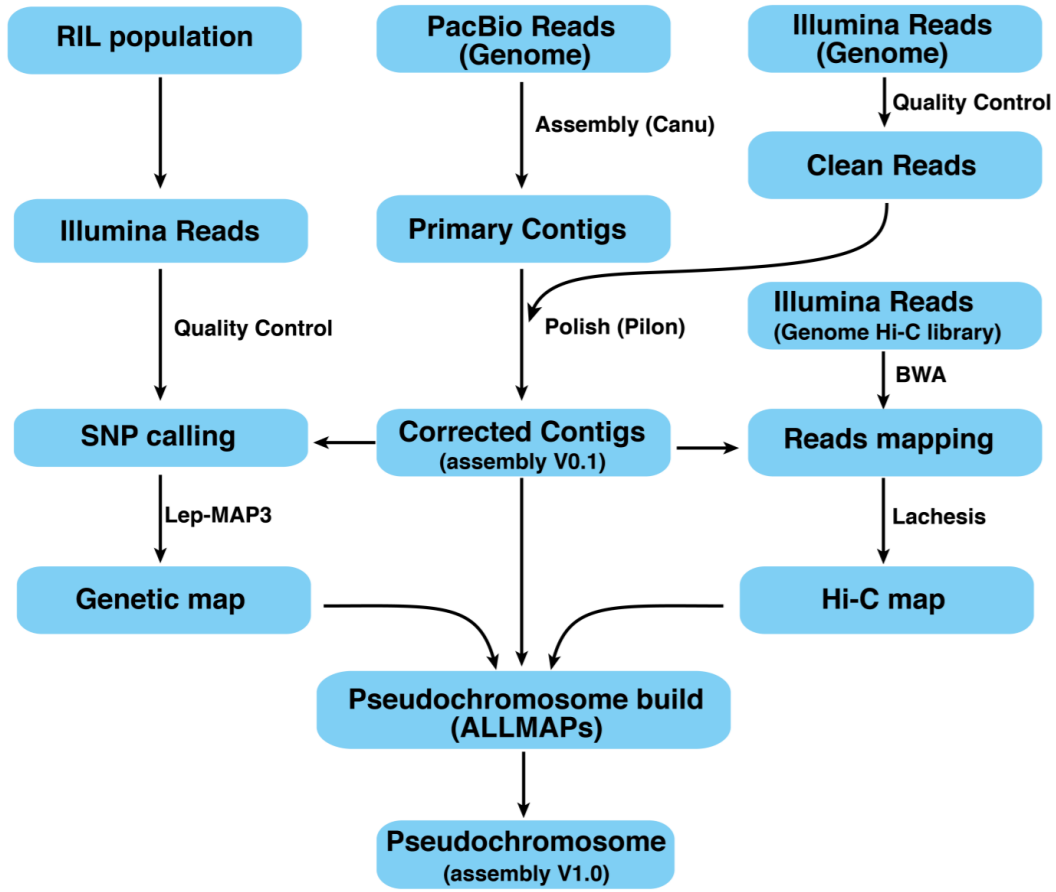
- 2 g of 10-day-old broomcorn millet seedlings were fixed in 1% formaldehyde solution.
- The nuclei/chromatin was extracted from the fixed tissue and digested with HindIII
- The overhangs filled in by biotin-14-dCTP and the Klenow enzyme
- Genomic DNA was extracted and sheared to a size of 300−500 bp
- The biotin-labeled DNA fragments were enriched using streptavidin beads and subject to library preparation.

Type	Library	Platform	Mean Fragment size (bp)	Read length (bp)	Raw data (Gp)	Raw coverage (x)	Effective data (Gp)	Effective coverage (x)
Genome	PCR-free	Illumina	420	250-250	80.27	87.00	79.69	86.36
	20-kb single molecule	PacBio	-	6,540*	81.03	87.79	81.03	87.79
	Hi-C	Illumina	-	125-125	64.98	70.40	20.88	22.62
RNAseq	1-week seedlings	Illumina	300	125-125	25.70		24.50	
	3-week shoot	Illumina	300	125-125	20.10		18.60	
	8-week leaf blade	Illumina	300	125-125	38.60		36.50	
	8-week leaf sheath	Illumina	300	125-125	46.30		43.90	
	8-week inforescence	Illumina	300	125-125	16.90		25.40	
	8-week stem	Illumina	300	125-125	22.10		20.60	
	8-week root	Illumina	300	125-125	18.20		16.80	
	Mature seeds	Illumina	300	125-125	24.20		22.80	

* Mean subread length

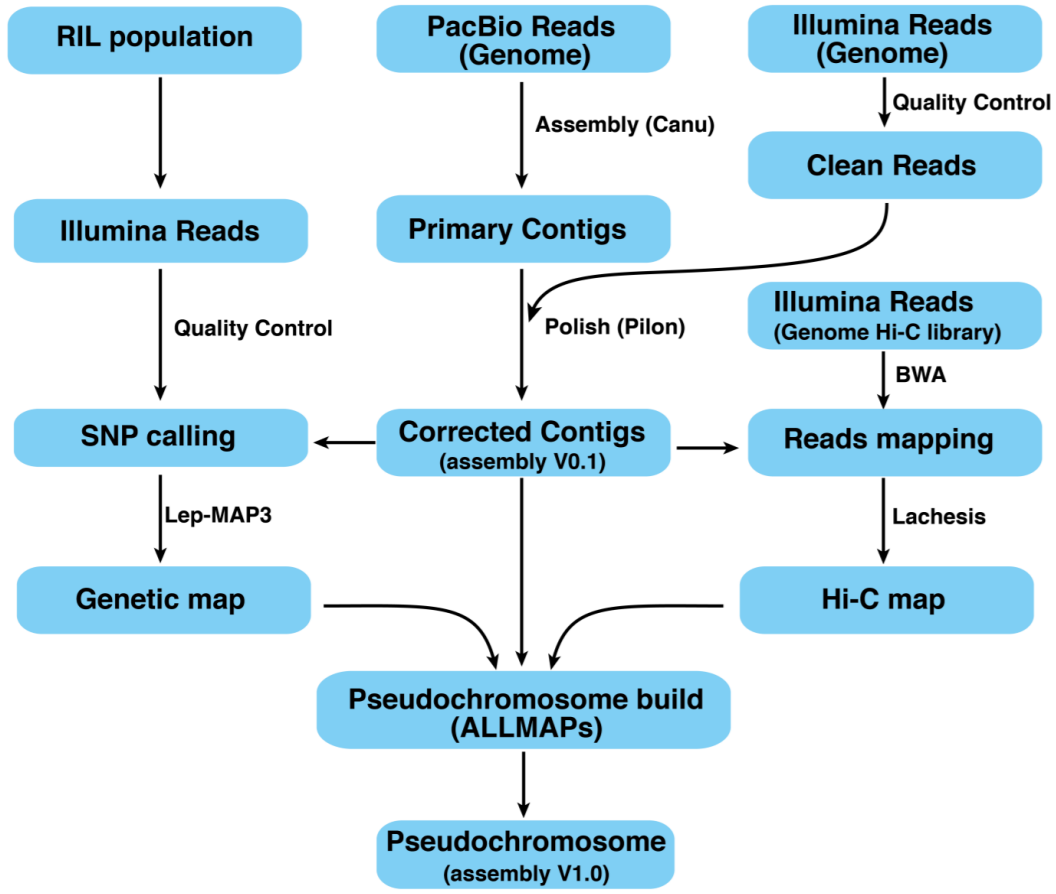
Genetic map construction

- 132 RIL (F6) obtained from biparental cross along with the two parents were genotyped by WGS
- 222,081 biallelic SNP were called using bcftools(v1.7)
- Lep-MAP3(v0.2) used for genetic mp construction.
- A LOD score set to 13 and recombination fraction 0.03
- 18 Linkage Group
- Final genetic map: 221,787 SNP marker
- Genetic length: Maternal parent: 2811 cm
  Paternal parent: 3092 cm

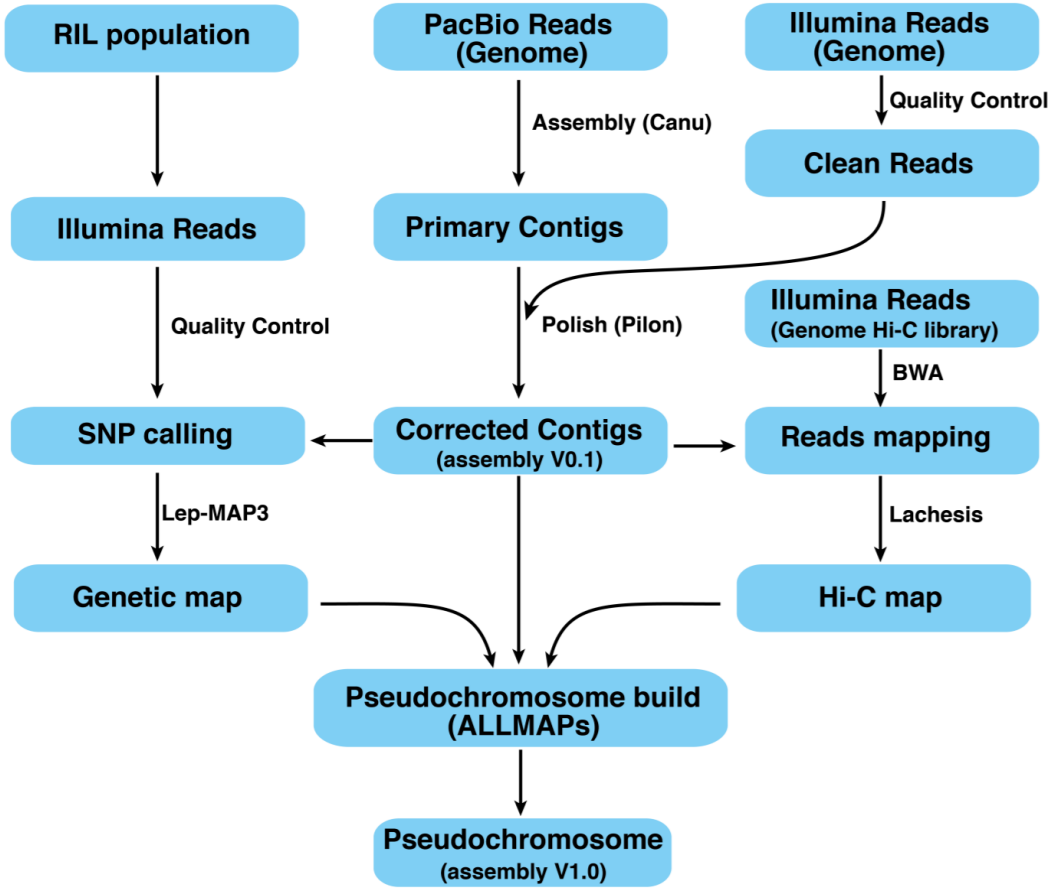


Assembly

- Filtered subreads (81.03 Gb) used for Assembly using canu
    - Genome Size Parameter: 900M
    - Error rate: 0.013
- Primary contigs polished using Pilon (v1.22)



- HiC library preprocessed and aligned to Pm_0390_v0.1 assembly using aln and sampe commands from bwa (v0.7.17)
- Resulting bam files and contigs from Pm_0390_v0.1 used in LACHESIS with cluster set to 18
- HiC map converted to 100- cm pseudomap
- HiC map, genetic map from two parents and Pm_0390_v0.1 contigs were used in ALLMAPS (v0.8.4) to generate 18 pseudomaps



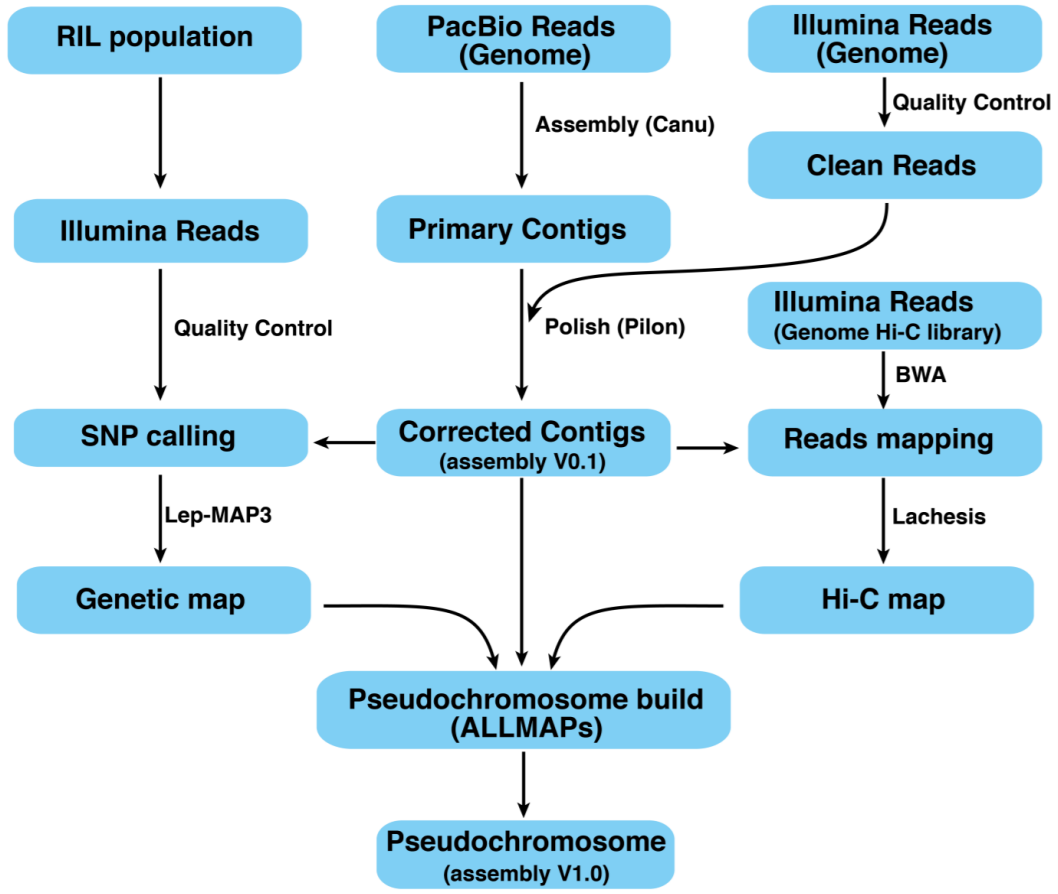
Global statistic of *P. miliacem* genome assembly and annotation

Version	v0.1	v1.0
Data source	PacBio + Illumina	v0.1 + Genetic map + Hi-C
Total assembly size (bp)	839,022,999	854,674,422
Number of scaffolds (≥ 1000 bp)		1,309
Longest scaffold (bp)		66,884,923
Scaffold N50 (bp)		46,661,915
Scaffold L50		8
Scaffold N90 (bp)		32,167,407
Scaffold L90		17
Number of contigs	5,541	5,541
Longest contig (bp)	5,222,262	5,222,262
Contig N50 (bp)	368,640	368,640
Contig L50	423	423
Missing bases	0	16,924,001 (1.98%)
Single-base error rate	0.004%	0.004%

	Number	Size
Assembly feature		
Estimated genome size		923 Mb
Total scaffolds (≥1000 bp)	1,309	855 Mb
Undetermined bases	1.98%	16.9 Mb
Scaffold N50	8	46,662 kb
Longest scaffold		66,885 kb
Pseudochromosomes	18	822 Mb
Anchored contigs	4,146	805 Mb
Anchored and oriented contigs	3,242	722 Mb
Total contigs	5,541	838 Mb
Contig N50	423	369 kb
Longest contig		5222 kb
GC content	46.8%	
Genome annotation		
Repetitive sequences	58.2%	495 Mb
Protein-coding genes	55,930	181 Mb
Genes in pseudochromosomes	55,527 (99.3%)	
Noncoding RNAs	9643	1.5 Mb

Assessment of genome assembly

- PE250 reads was preprocessed and aligned to Pm_0390_v1 using bwa mem.
- Samtools and GATK used for SNP calling and summarization
- 20 kb PacBio library created from a 40-kb fosmid library
- Falcon v0.3.0 with default parameters was used for the de novo assembly of fosmid sequences.
- the contigs were then aligned to Pm_0390_v1
- Assembly completeness assessed using both transcriptome data and BUSCO

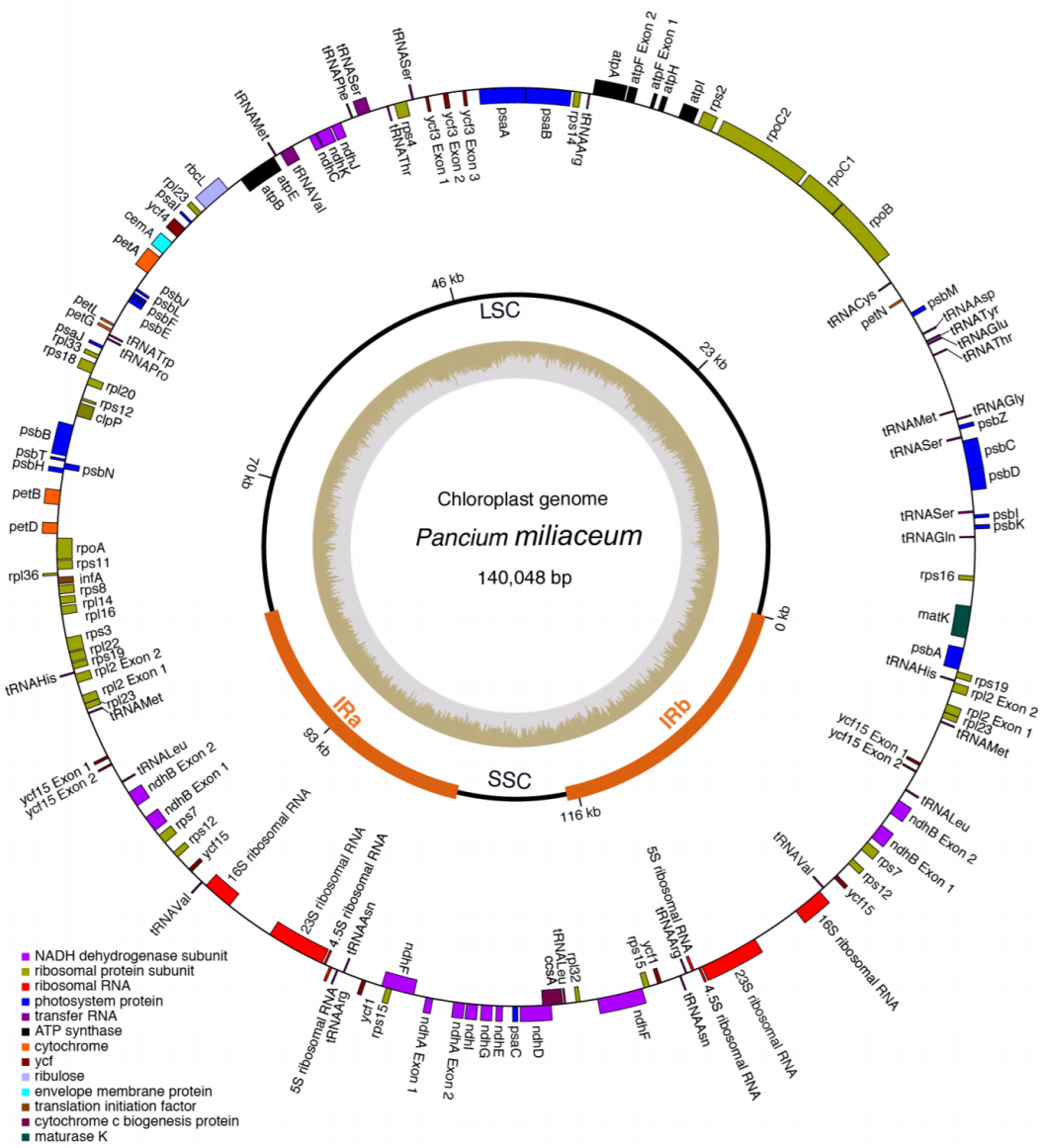


Assessment of genome assembly

- Alignment of the fosmid sequences to Pm_0390_v1 revealed no structural errors and high sequence identity rates (99.53–100%)
- A total of 305,520 transcripts were de novo assembled from 241 Gb of mRNA-seq data.
- More than 98% of the transcript sequences could be mapped to Pm_0390_v1
- In addition, 1411 (98%) of the 1440 plant single-copy orthologs from BUSCO v223 were identified in the broomcorn millet genome

Gene Annotation

- Chloroplast genome annotate using DOGMA and CpGA-VAS.
- Output from two softwares integrated and longer ORF were retained
- Predicted start-stop codon and exon-intron boundaries manually examined and curated
- GenomeVx used for generating chloroplast genome



Repeat Annotation

- 3 complimentary softwares were used
    - LTR_Finder (v1.06)
    - Piler (v1.0)
    - RepeatModeler (v4.0.6)
- De novo repeat library was used for homology search of repeats using Repeatmasker (v1.0.10)

Results

- 92.1% consists of transposable elements
- most TE sequences are retrotransposons
- 112,158 SSR with a mean occurrence frequency of 22.5 per Mb
- Most SSRs were composed of di- and tri-nucleotide motifs with an average length of ~22 bp

	Number	Size
Assembly feature		
Estimated genome size		923 Mb
Total scaffolds (≥1000 bp)	1,309	855 Mb
Undetermined bases	1.98%	16.9 Mb
Scaffold N50	8	46,662
Longest scaffold		66,885
Pseudochromosomes	18	822 Mb
Anchored contigs	4,146	805 Mb
Anchored and oriented contigs	3,242	722 Mb
Total contigs	5,541	838 Mb
Contig N50	423	369 kb
Longest contig		5222 kb
GC content	46.8%	
Genome annotation		
Repetitive sequences	58.2%	495 Mb
Protein-coding genes	55,930	181 Mb
Genes in pseudochromosomes	55,527 (99.3%)	
Noncoding RNAs	9643	1.5 Mb

Gene Prediction in repeat-masked genome

- Three approaches:
  - *Ab initio gene predition:* AUGUSTUS (v2.5.5), Genescan (v1.0), SNAP (version 2006-07-28), GlimmerHMM (v3.0.3), and Fgenesh
  - *Homology-based gene prediction:*
    - aligning the protein sequences of six grass species and *Arabidopsis thaliana* to Pm_0390_v1 using TBLASTN
    - Gene models were generated using GeneWise (v2.4.1)
  - *Transcriptome-assisted gene prediction:*
    - TopHat (v2.1.1) was used to map filtered mRNA-seq reads to Pm_0390_v1
    - Cufflinks (v2.2.1) was then used to assemble the alignments into transcripts
- GLEAN (v1.0.1) to generate the final gene set.

Method	Gene number	Average Length (bp)				Number of Exons per Gene	Source	Version
		Gene	CDS	Exon	Intron			
Ab initio								
Augustus	69,693	2,238	818	245	605	3.3		
Genescan	83,305	4,571	721	221	1,696	3.3		
GlimmerHMM	229,575	2,273	518	199	1,093	2.6		
SNAP	108,528	4,078	637	204	1,617	3.1		
Fgenesh	67,227	2,848	1,113	230	452	4.8		
Homology								
B. distachyon	79,246	2,587	1,131	314	561	3.6	Phytozome 12	314_v3.0
O. sativa	94,948	2,012	1,021	346	508	3	Phytozome 12	323_v7.0
S. italic	83,430	2,629	1,090	301	587	3.6	Phytozome 12	312_v2
S. bicolor	89,416	2,314	1,150	338	484	3.4	Phytozome 12	454_v3.0.1
A. thaliana	81,912	1,466	750	276	418	2.7	Phytozome 12	TAIR10
T. aestivum	61,160	2,812	1,251	329	557	3.8	Phytozome 12	296_v2.2
Z. mays	77,753	2,407	1,093	327	562	3.3	Phytozome 12	284_AGPv3
mRNA-seq	30,214	3,771	1,466	216	398	6.8		
GLEAN	**55,930**	**3,260**	**1,172**	**248**	**461**	**4.7**		

Type	Copy Number	Average Length (bp)	Total Length (bp)	Proportion of genome (%)
miRNA	339	141.5	47,984	0.01
tRNA	1,420	75.1	106,645	0.01
rRNA				
18S	161	1469.8	236,642	0.03
28S	531	142.4	75,597	0.01
5.8S	124	157.3	19,504	<0.01
5S	824	103.9	85,616	0.01
snRNA				
CD-box	2,050	105.2	215,756	0.02
HACA-box	89	129.3	11,512	<0.01
splicing	163	150.7	24,557	<0.01
Total	**5,701**		**823,813**	**0.09**

Functional annotation of gene models

- Predicted proteins searched against six protein/function database
- Interpro, Go, KEGG, KOG, Swiss-Prot, TrEMBL
- Results from six databases searches are concatenated

	Database	Number	Percent (%)
Total Genes		55,930	-
Annotation	InterPro	36,513	65.3
	GO	46,973	83.9
	KEGG	28,158	50.3
	KOG	53,474	95.6
	Swissprot	36,737	65.7
	TrEMBL	53,097	94.9
Total Annotated		54,003	96.6
Total Unannotated		1,927	3.4

Case Study II: Wheat genome annotation and assembly

"Shifting the limits in wheat research and breeding using a fully annotated reference genome," IWGSC 2018, Science

Chinese Spring Wheat

- Hexaploid
- Estimated genome size: ~15.76 Gb
- 21 chromosomes
- Three genomes
  - A: *Triticum Urartu*
  - B: *Aegilops speltoides*
  - D: *Aegilops tauschii*

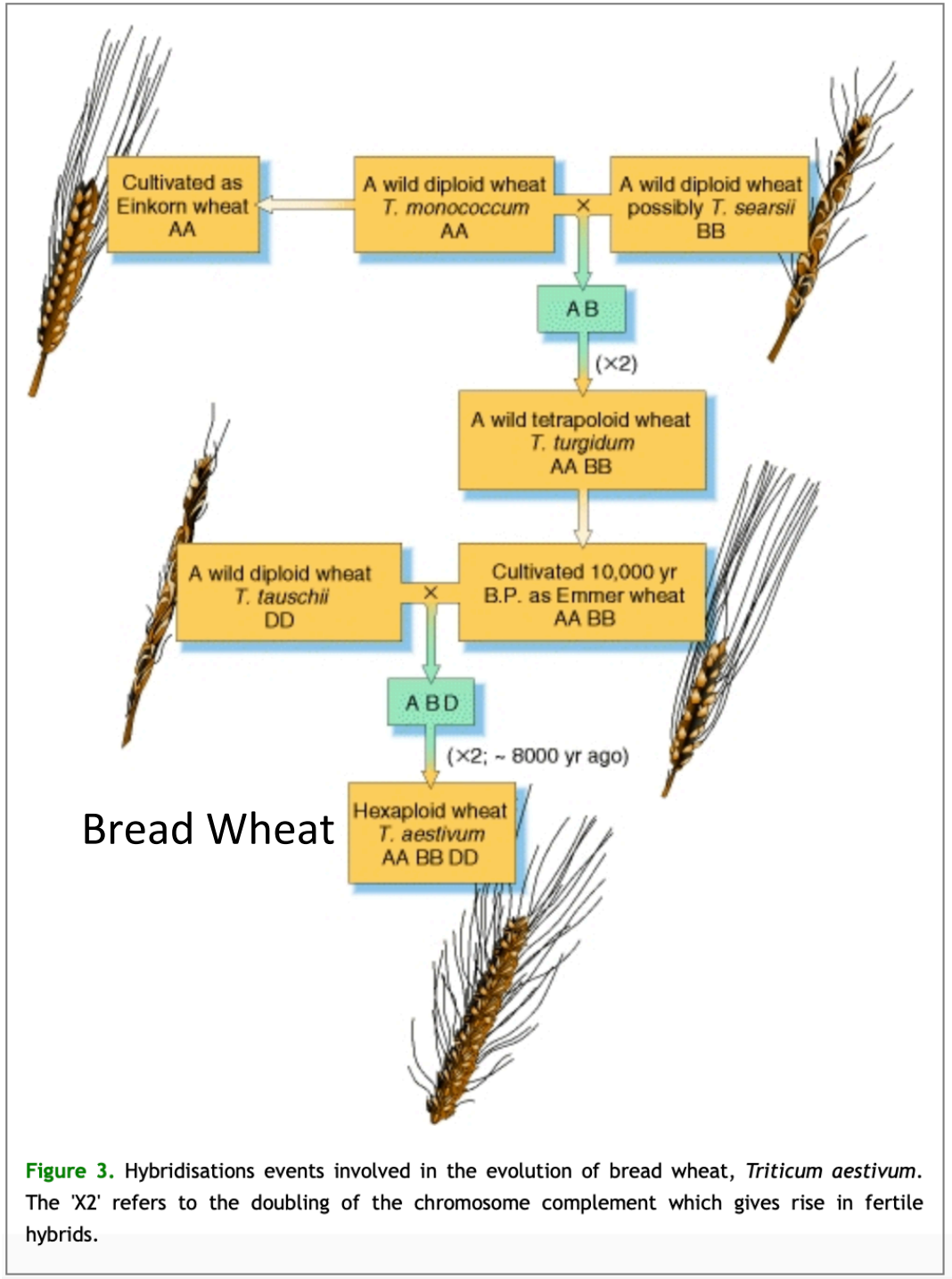


Figure 3. Hybridisations events involved in the evolution of bread wheat, *Triticum aestivum*. The 'X2' refers to the doubling of the chromosome complement which gives rise in fertile hybrids.

http://www.cerealsdb.uk.net/cerealgenomics/WheatBP/Documents/DOC_Evolution.php

Assembly

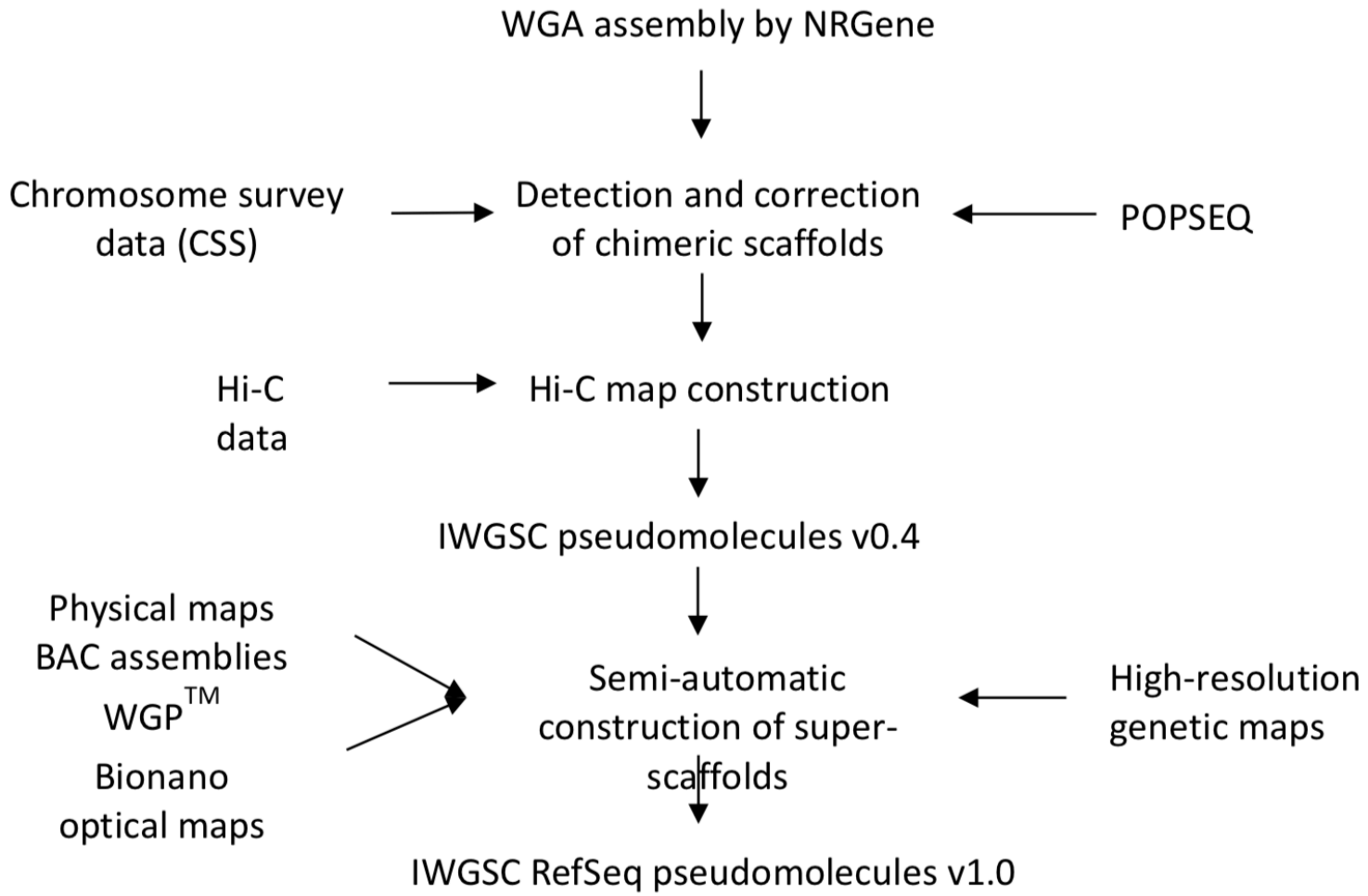


Fig. S1

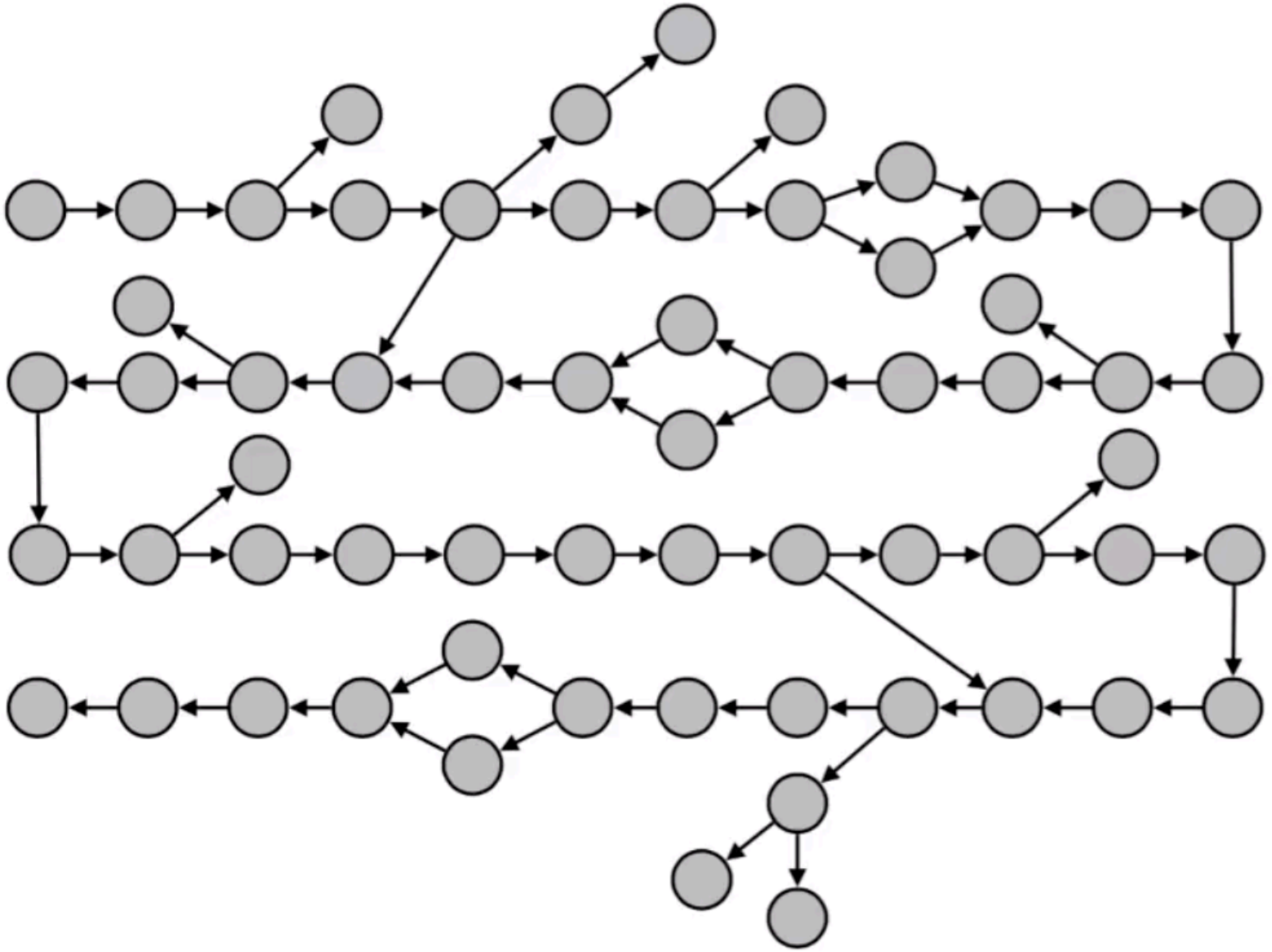
Data integration pipeline for the assembly of IWGSC RefSeq v1.0. The whole genome assembly (WGA) used DNA from the cultivar CS.

Sequencing Technologies and Assembly

- Short reads: Illumina
  - 450 bp- 10 kb range

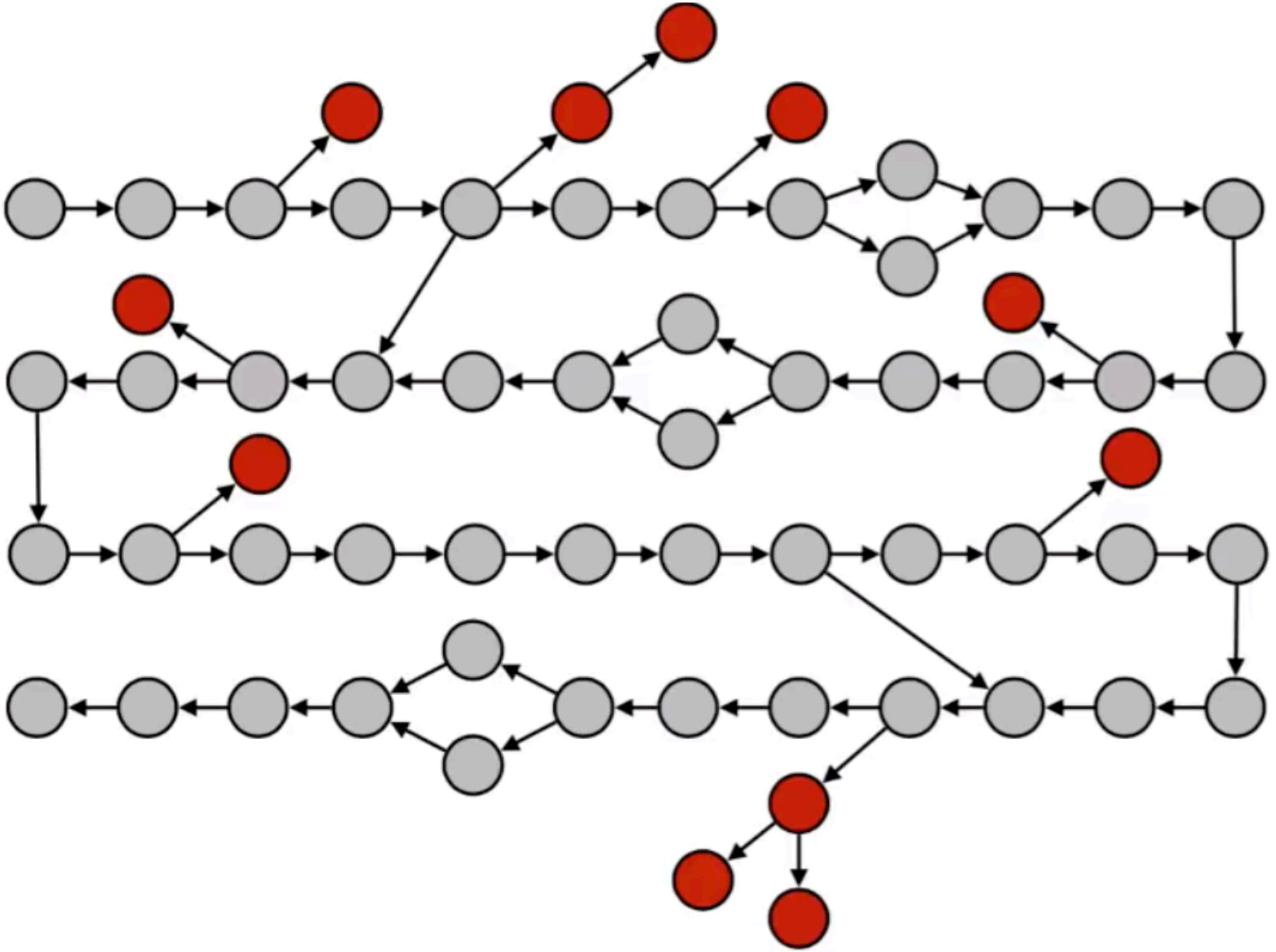
- De novo Whole-Genome Assembly (WGA)
  - **DenovoMAGIC2TM (NRGene, Nes Ziona, Israel)**
  - Genetic data
  - Physical data
  - Sequence data

DenovoMAGIC2TM- Example of De Bruijn Graph



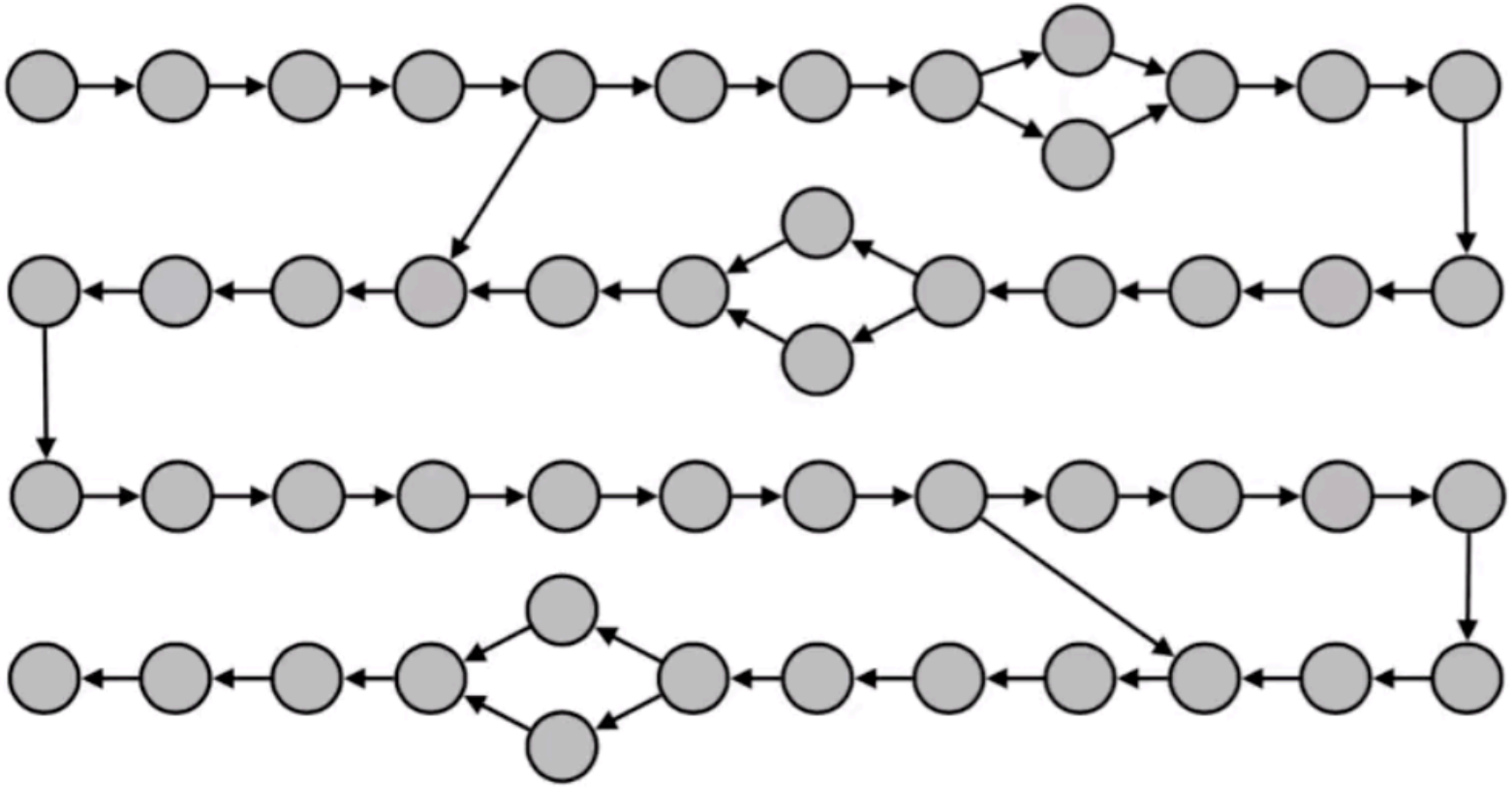
https://www.youtube.com/watch?v=5wvGapmA5zM

DenovoMAGIC2TM- Quality Control



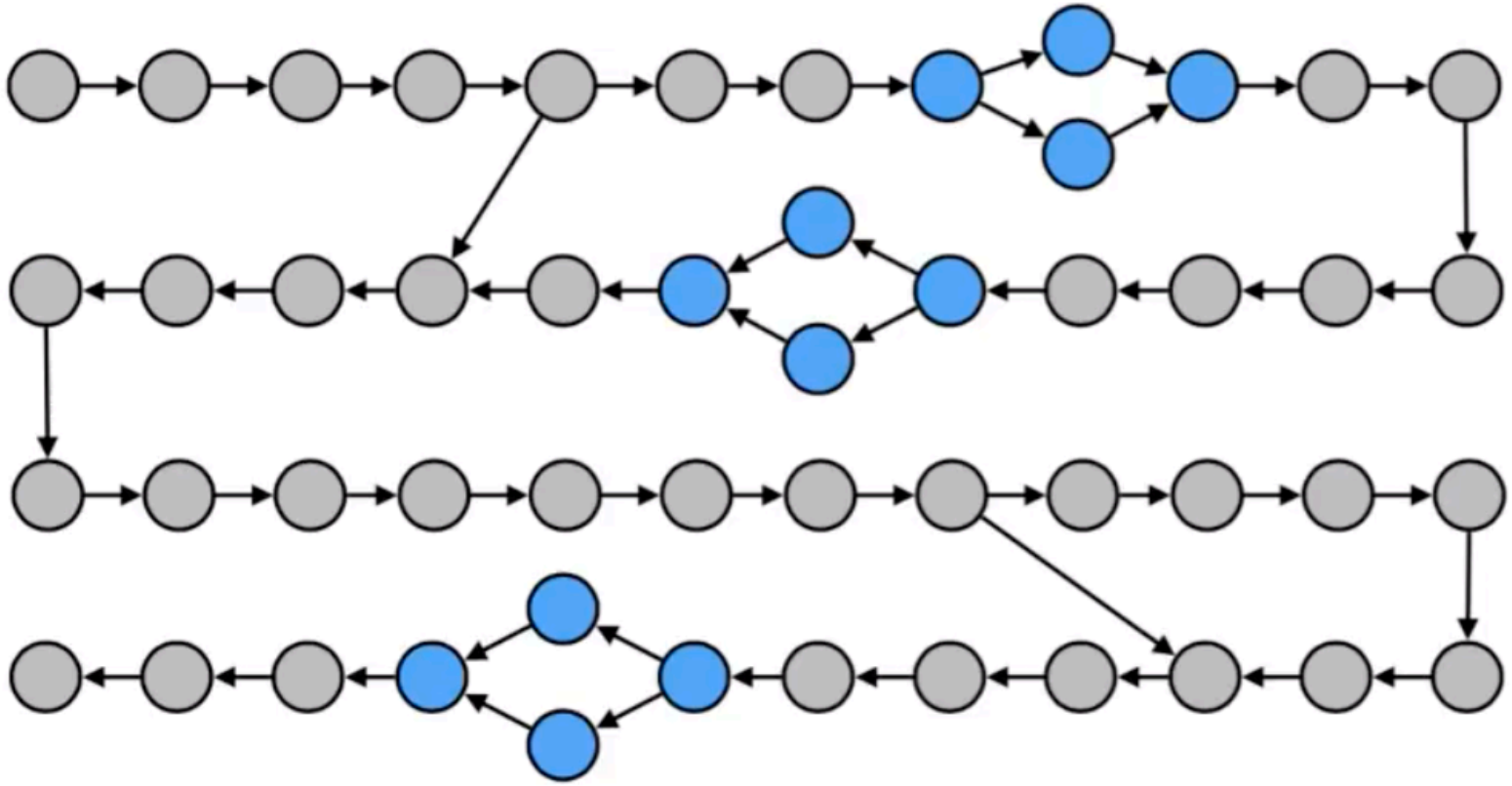
https://www.youtube.com/watch?v=5wvGapmA5zM

DenovoMAGIC2TM- Quality Control



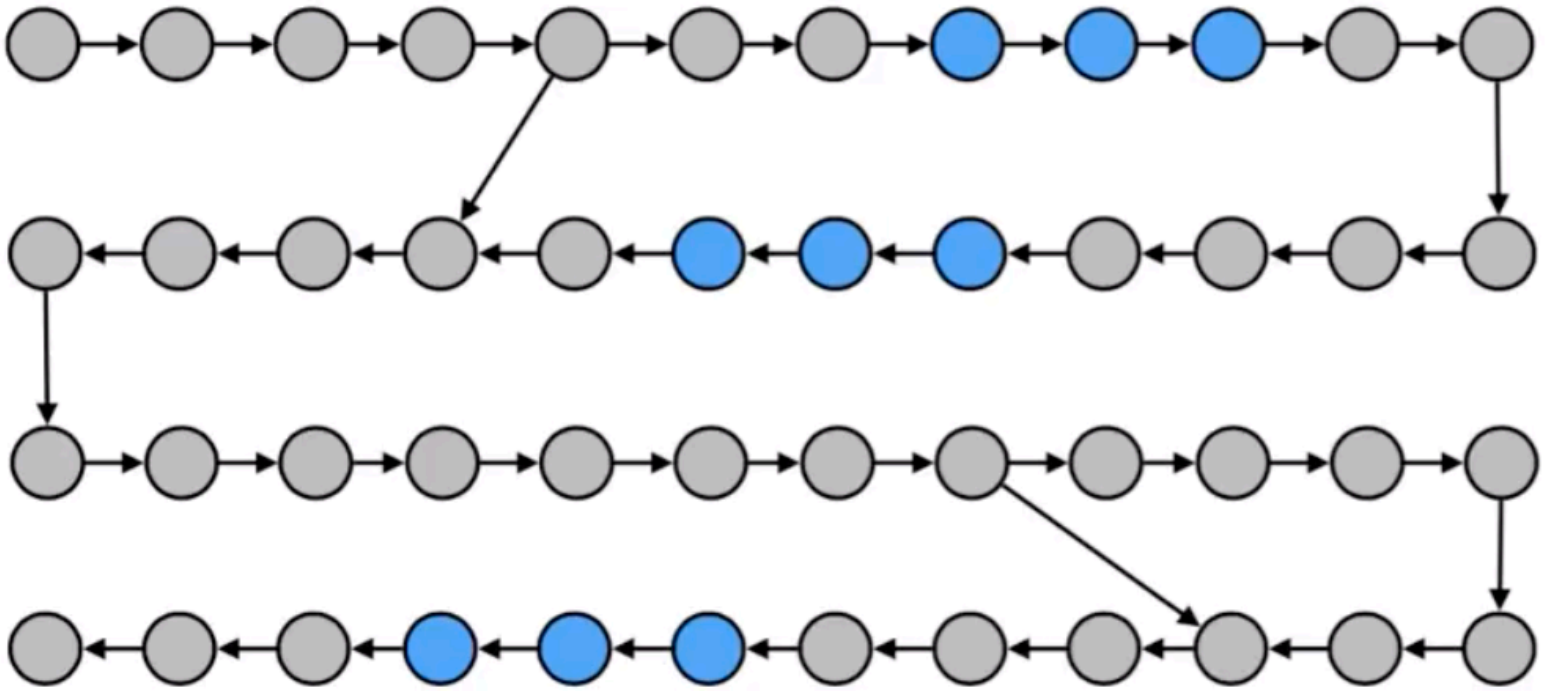
https://www.youtube.com/watch?v=5wvGapmA5zM

DenovoMAGIC2TM- Quality Control



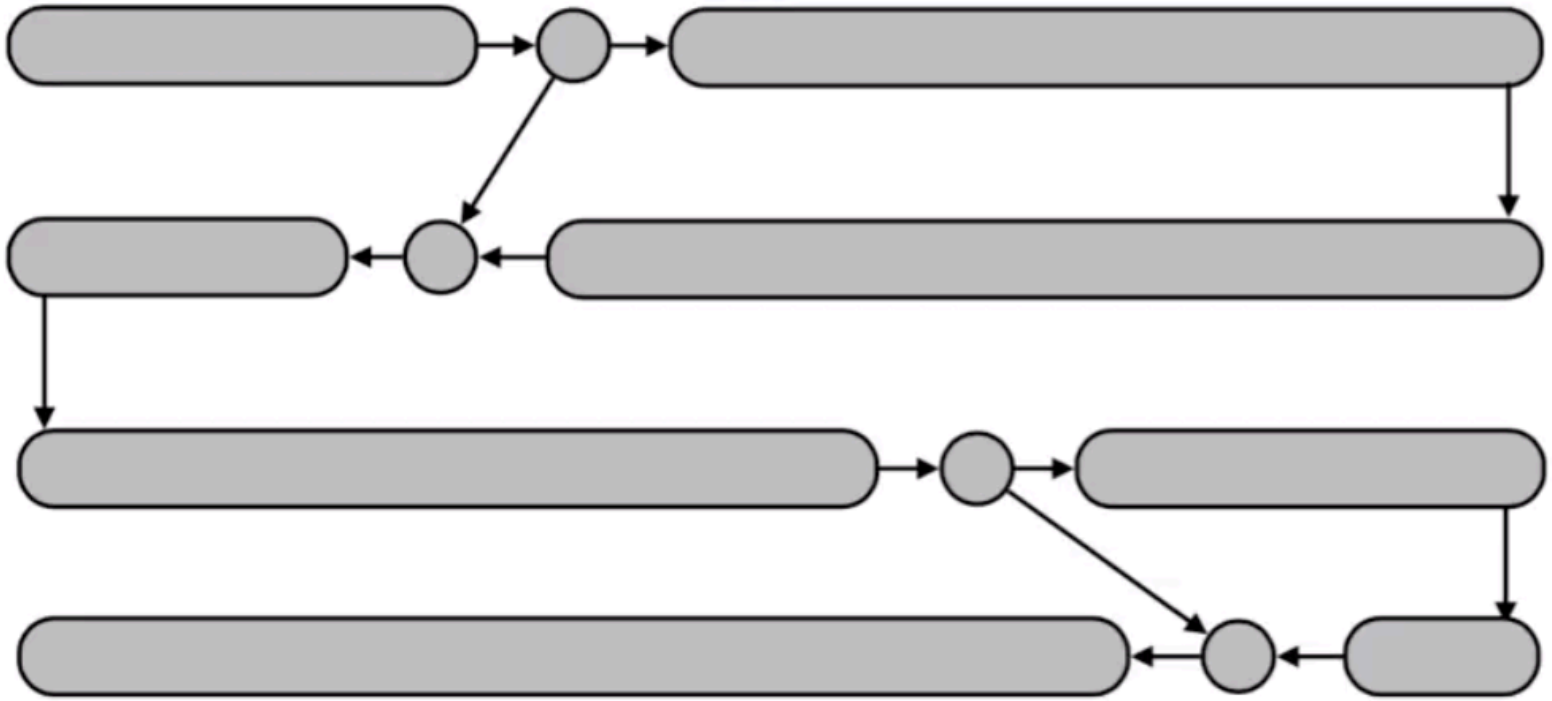
https://www.youtube.com/watch?v=5wvGapmA5zM

DenovoMAGIC2TM- Quality Control



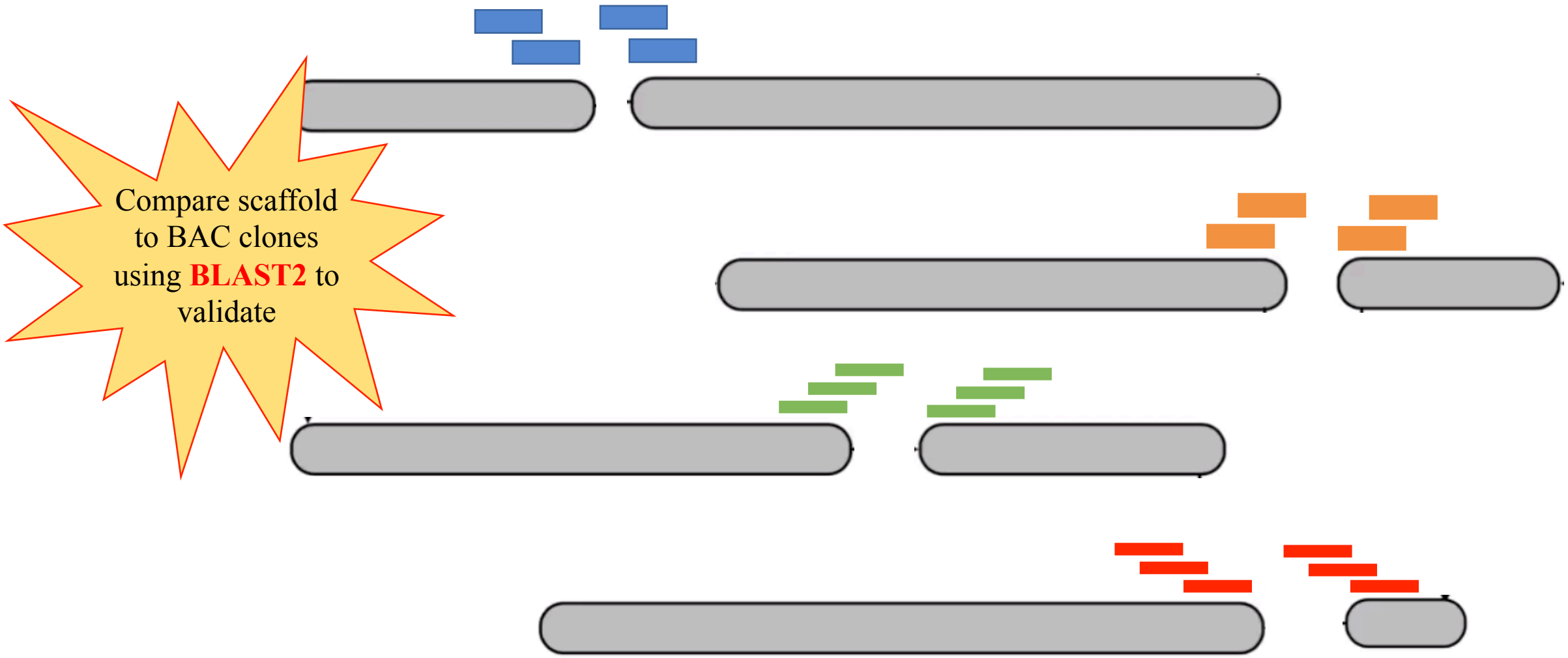
https://www.youtube.com/watch?v=5wvGapmA5zM

DenovoMAGIC2TM- Assembly contigs



https://www.youtube.com/watch?v=5wvGapmA5zM

DenovoMAGIC2TM- Scaffold assembly into WGA



https://www.youtube.com/watch?v=5wvGapmA5zM

DenovoMAGIC2TM

1. Read pre-processing and error correction
   - PCR with Illumina adaptors, Nextera linkers, and paired-end reads
   - From the paired-end libraries overlapping sequences were merged to create stitched reads
2. *De novo* assembly of contigs: graph of contigs
   - Able to identify repeats and non-repetitive sequences of contigs using k-mers
3. *De novo* assembly of scaffolds:
   - Map read ends to contigs- This creates overlapping read ends
   - Searching the graph for a unique path of contigs connecting pairs of reads mapping to two different non-repetitive contigs
   - Order scaffolds and link using the libraries- Accounting for gaps between contigs

Organelle Genome Assembly

- Remapped 150 million reads to the scaffold to create chloroplast and mitochondrial genome assemblies
- Assembly of organelle genome using **NUCLEAR** and **VISION**
- Annotated by aligning sequences to GenBank using **BLAST**

- Analyzed NUTPs and NUMTs using **NUCLEAR**
  - Segments of nuclear DNA deriving from plastid or mitochondrial DNA

Physical Mapping

- Hi-C
  - 3D chromosome conformation capture sequence data
- BAC library fingerprints
  - Using either High Information Content Fingerprinting (HICF) with **SNaPShot** technology or Whole Genome Profiling
- Bionano optical maps
  - Mapping chromosomes
- Radiation Hybrid (RH) Maps
  - Break chromosome via radiation
  - Clone fragments (ID SNPs)
  - Constructed using **Carthagene**

Genetic Mapping

- Genotype-by-sequencing (GBS)
  - Mapped two populations: double haploids and Recombinant Inbred Lines (RILs)
- Contigs of the chromosome shotgun sequence (CSS)
  - Mapped to WGA with **BWA mem**
  - SNPs called with **SAMtools**
- **POPSEQ**
  - Set chromosome assignments in genetic positions (cM coordinates)

Assembly

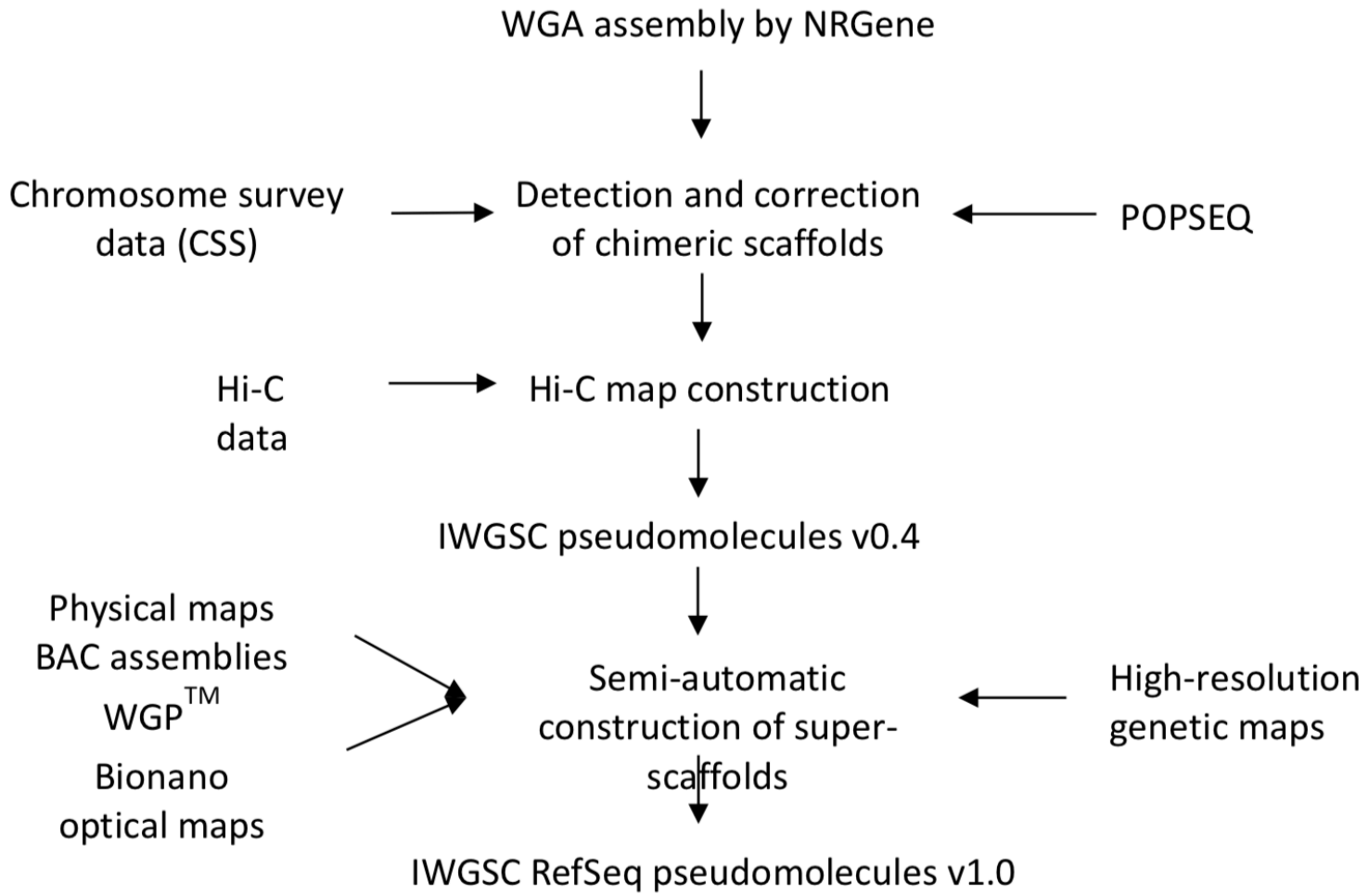


Fig. S1

Data integration pipeline for the assembly of IWGSC RefSeq v1.0. The whole genome assembly (WGA) used DNA from the cultivar CS.

Annotation

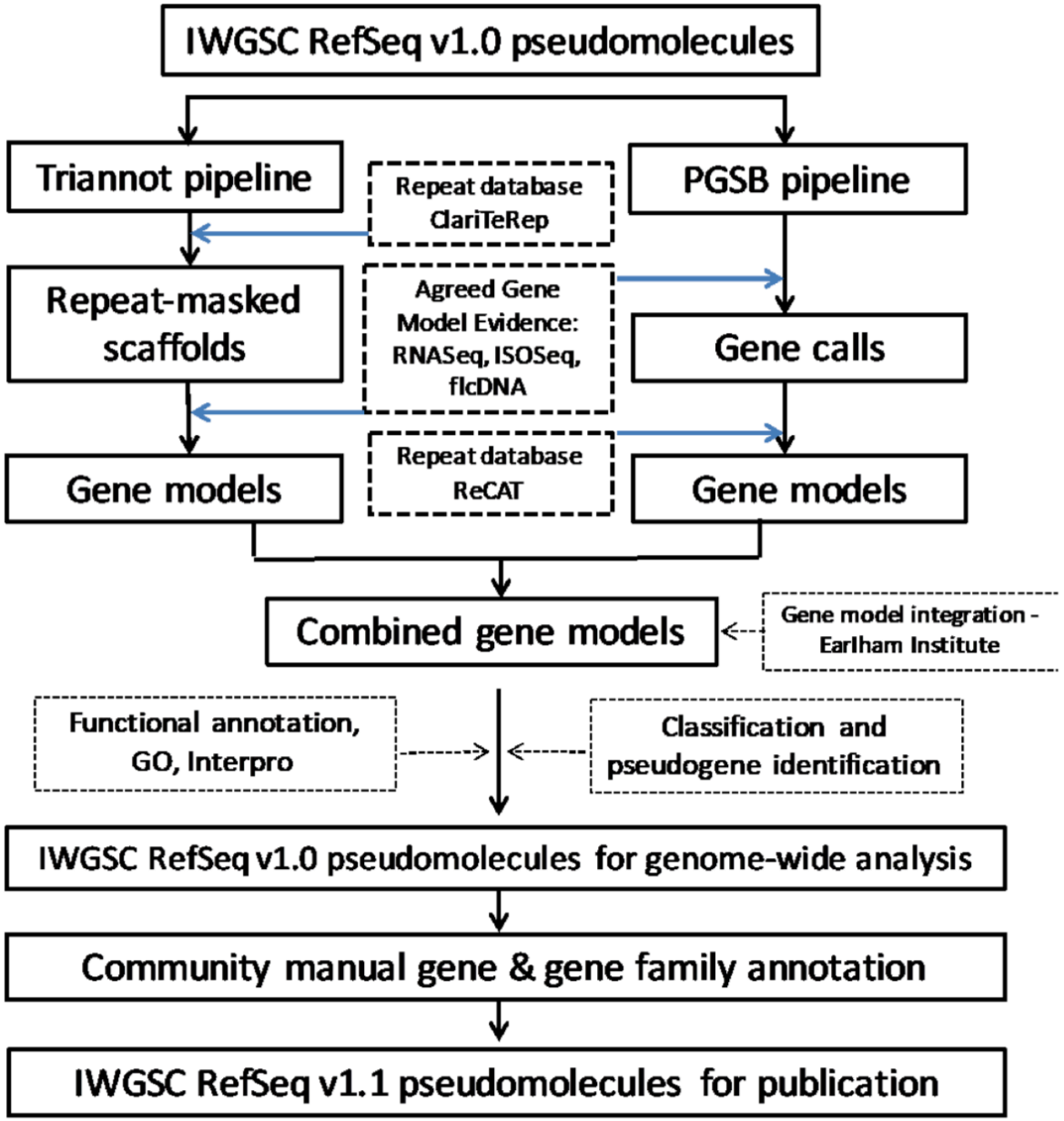


Fig. S8

IWGSC RefSeq v1.0 Genome annotation pipeline.

Annotation

- ChIP-seq data identified centromeric regions
- Reads mapped to CS pseudomolecules in SAM format using **BWA mem**
  - BAM format using **SAMtools**
  - Sorted with **Novosort**
  - Count reads using **BEDTools**
- **CLARITE** to model transposable elements
- Automated gene calling with two pipelines: **TriAnnot** and **PGSB**
- Annotation validated by comparing gene models with PacBio RNA-seq

Annotation Cont.

- miRNAs annotated by homology with known plant miRNAs using **BLAST**
- tRNAs annotated with **tRNAscan-SE**
- Functional annotation
  - Assign physiological function of genes using **AHRD**
  - Using other databases, they also used orthology-based functional annotation of gene by comparing gene models

Annotation

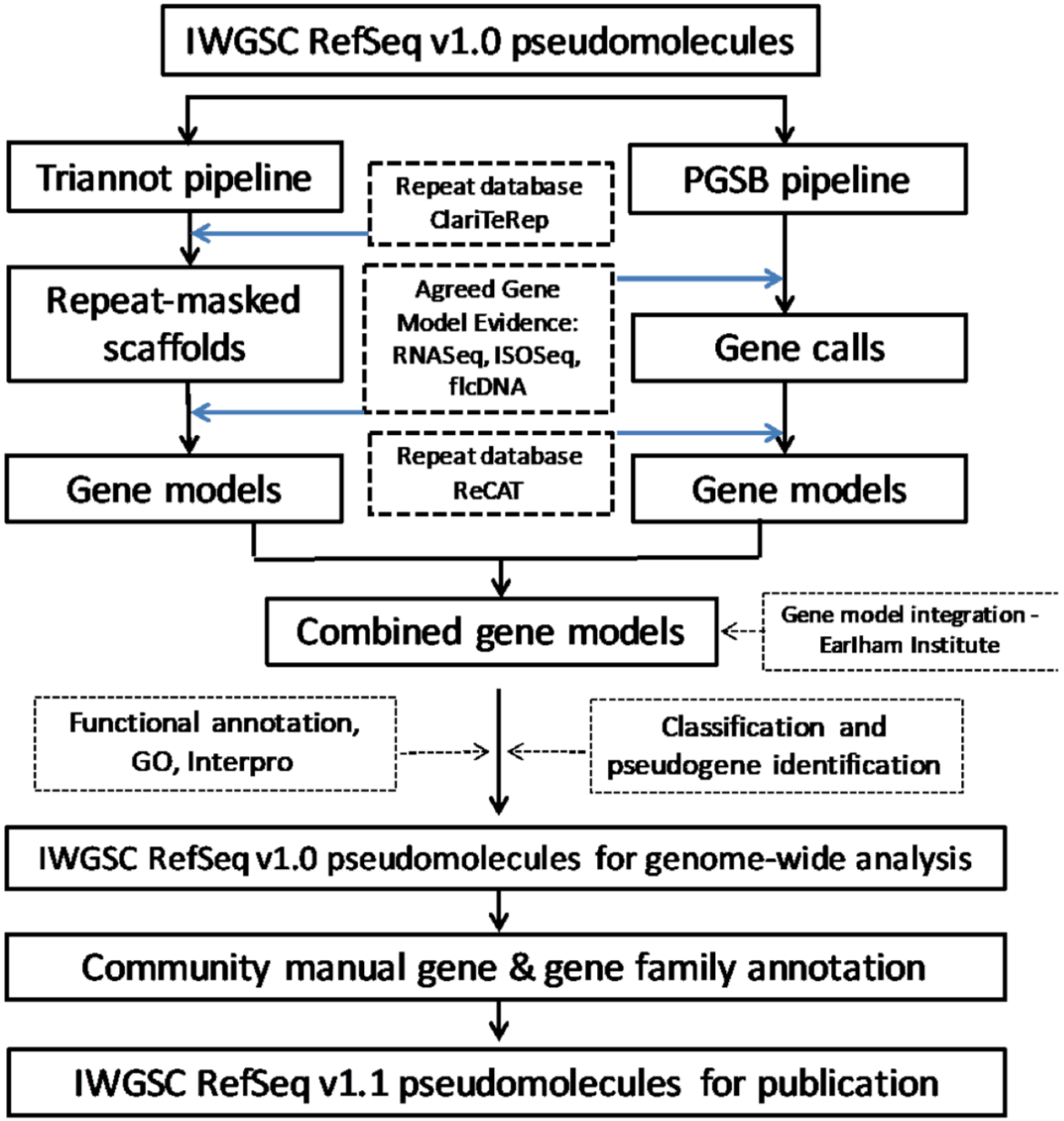


Fig. S8

IWGSC RefSeq v1.0 Genome annotation pipeline.


Thank You