# **The TAIR12 genome assembly:**
# a new reference for *Arabidopsis thaliana*

Xiao Dong, Raúl Wijfjes, Korbinian Schneeberger*

*LMU Munich, Germany*

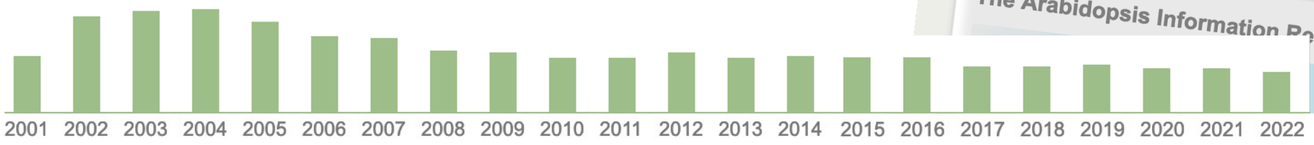
*MPI for Plant Breeding Research, Cologne, Germany*

*Community Consensus Arabidopsis thaliana*
*Reference Genome Assembly Consortium*

The Arabidopsis Genome Initiative, *Nature*, 2000

# The impact of the *A. thaliana* reference sequences

- Strengthened the model species: Col-0 acts as reference line for functional analysis
- Tool to access the genome: primer and marker design, physical map for mapping
- Acts as community-wide accepted standard for annotations and as knowledge basis (transferred to many other plant genomes)



Google scholar: >11,000 citations

The Arabidopsis Information Resource

Breaking News

**See you in San Diego at PAG31!**

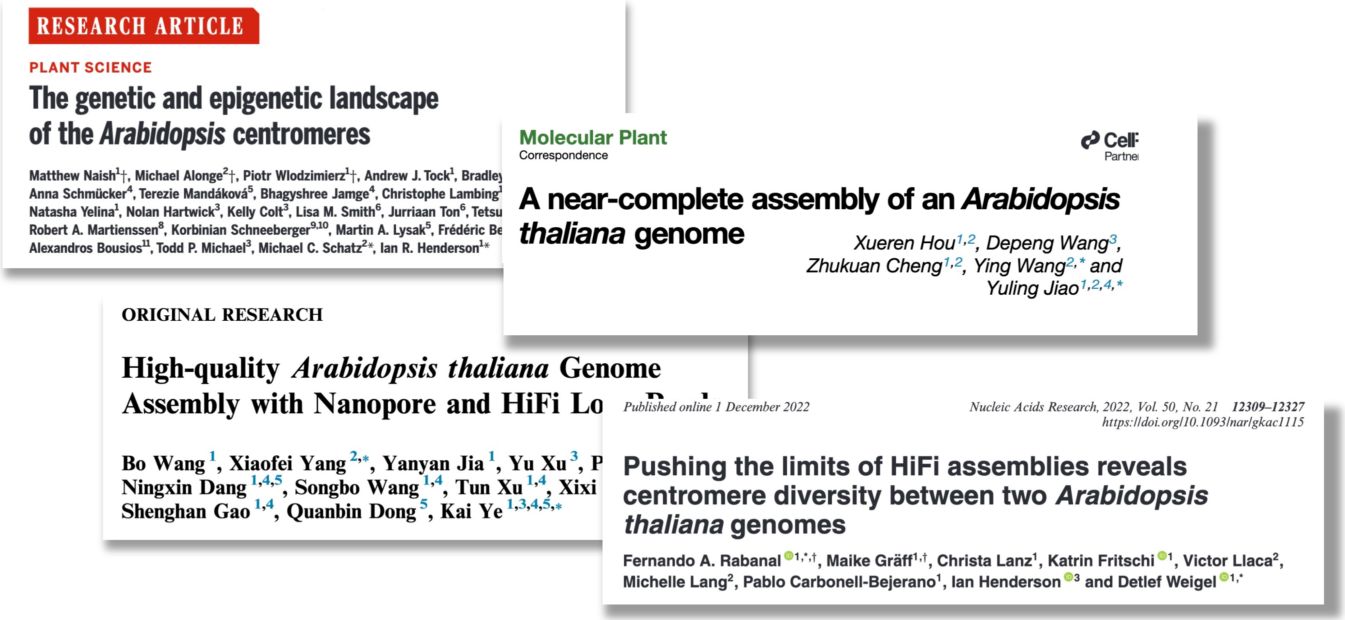
[Jan 4, 2024]

TAIR/Phoenix/AgBioData will be at PAG31 in San Diego. Come to our workshops on **Arabidopsis Informatics** and **AgBioData**. Also visit us at booth #422.

Learn about the re-annotation of the Arabidopsis genome

# Long-read sequencing improved genome assembly

Reference sequence not complete (119 Mb vs. ~140 Mb est. genome size)



RESEARCH ARTICLE

PLANT SCIENCE

The genetic and epigenetic landscape of the *Arabidopsis* centromeres

Matthew Naish[1]†, Michael Alonge[2]†, Piotr Wlodzimierz[1]†, Andrew J. Tock[1], Bradley
Anna Schmücker[4], Terezie Mandáková[5], Bhagyshree Jamge[4], Christophe Lambing[1
Natasha Yelina[1], Nolan Hartwick[3], Kelly Colt[3], Lisa M. Smith[6], Jurriaan Ton[6], Tetsu
Robert A. Martienssen[8], Korbinian Schneeberger[9,10], Martin A. Lysak[5], Frédéric Be
Alexandros Bousios[11], Todd P. Michael[3], Michael C. Schatz[2]*, Ian R. Henderson[1]*

Molecular Plant
Correspondence

Cell Partne

**A near-complete assembly of an *Arabidopsis thaliana* genome**

*Xueren Hou[1,2], Depeng Wang[3], Zhukuan Cheng[1,2], Ying Wang[2,*] and Yuling Jiao[1,2,4,*]*

ORIGINAL RESEARCH

**High-quality *Arabidopsis thaliana* Genome Assembly with Nanopore and HiFi Lo**

**Bo Wang[1], Xiaofei Yang[2,*], Yanyan Jia[1], Yu Xu[3], P
Ningxin Dang[1,4,5], Songbo Wang[1,4], Tun Xu[1,4], Xixi
Shenghan Gao[1,4], Quanbin Dong[5], Kai Ye[1,3,4,5,*]**

*Published online 1 December 2022*

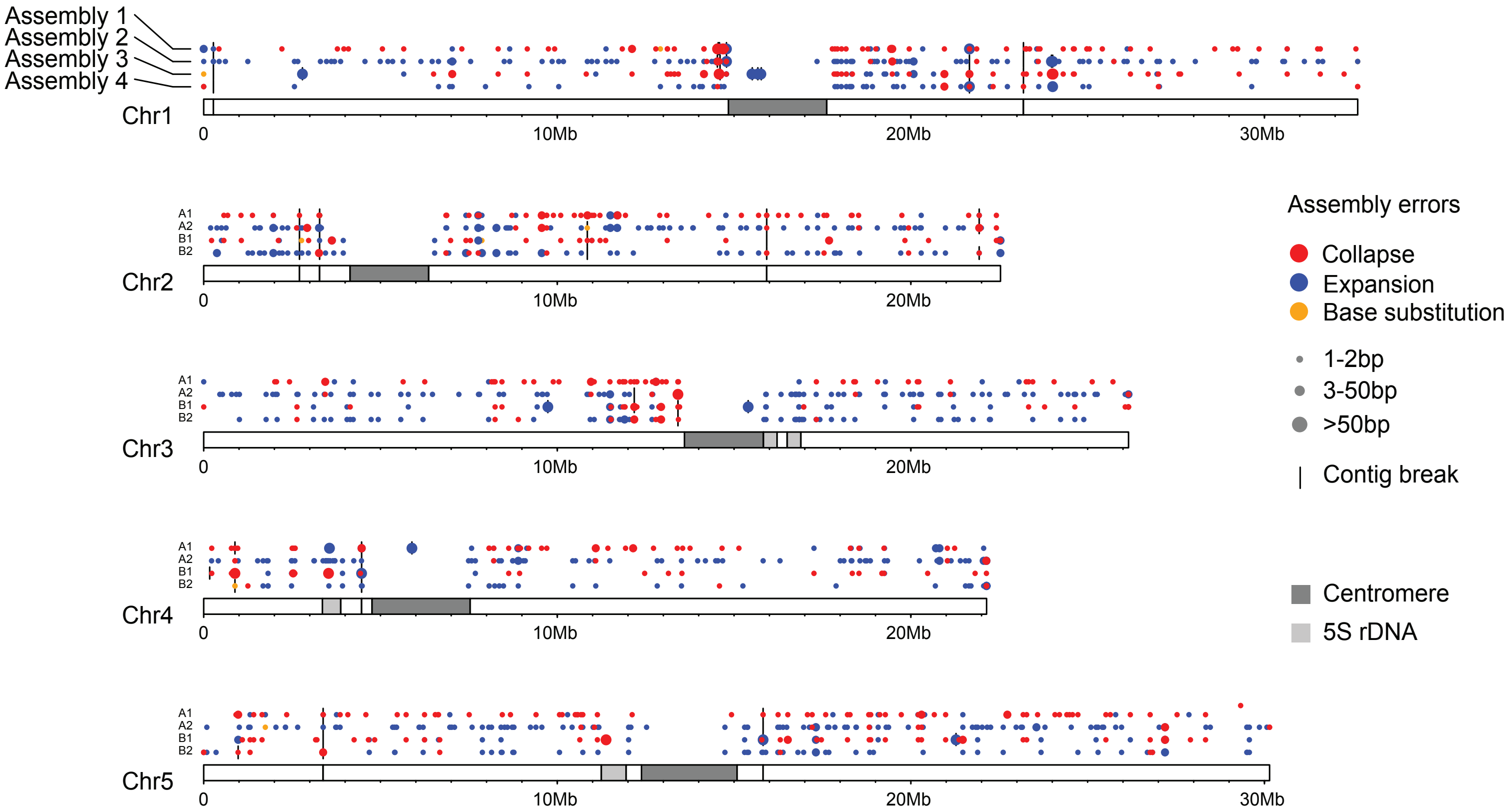
*Nucleic Acids Research, 2022, Vol. 50, No. 21* ***12309–12327***
*https://doi.org/10.1093/nar/gkac1115*

**Pushing the limits of HiFi assemblies reveals centromere diversity between two *Arabidopsis thaliana* genomes**

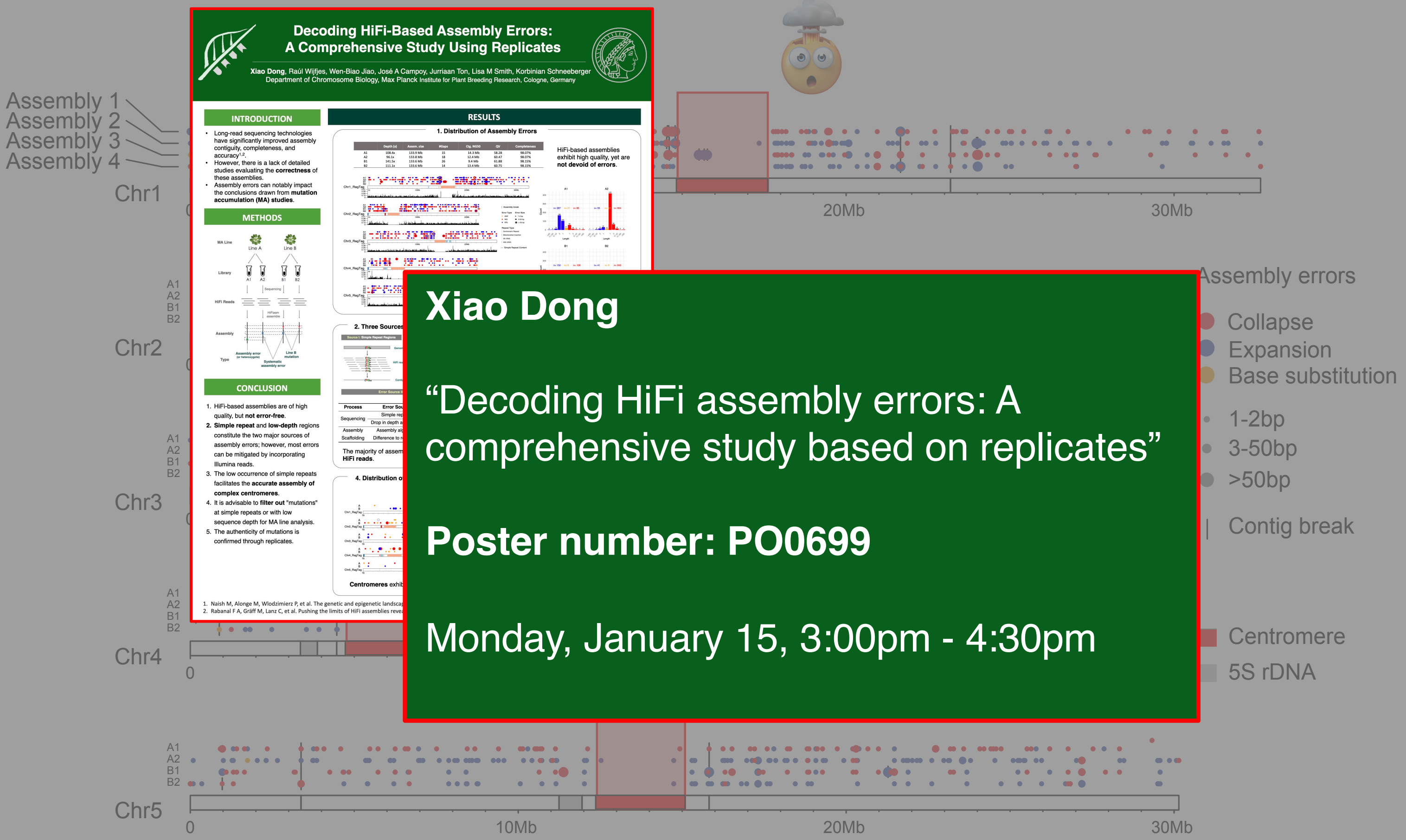
**Fernando A. Rabanal[1,*,†], Maike Gräff[1,†], Christa Lanz[1], Katrin Fritschi[1], Victor Llaca[2], Michelle Lang[2], Pablo Carbonell-Bejerano[1], Ian Henderson[3] and Detlef Weigel[1,*]**

Naish *et al*, Science, 2021
Wang *et al*, GPB, 2022
Hou *et al*, Mol Plant, 2022
Rabanal *et al*, NAR, 2022

# PacBio HiFi assemblies still include (a few) errors



# PacBio HiFi assemblies still include (a few) errors



## Decoding HiFi-Based Assembly Errors: A Comprehensive Study Using Replicates

**Xiao Dong**, Raúl Wijfjes, Wen-Biao Jiao, José A Campoy, Jurriaan Ton, Lisa M Smith, Korbinian Schneeberger
Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany

### INTRODUCTION

- Long-read sequencing technologies have significantly improved assembly contiguity, completeness, and accuracy[1,2].
- However, there is a lack of detailed studies evaluating the **correctness** of these assemblies.
- Assembly errors can notably impact the conclusions drawn from **mutation accumulation (MA) studies**.

### METHODS

MA Line: Line A, Line B
Library: A1, A2, B1, B2
Sequencing
HiFi Reads
HiFiasm assemble
Assembly
Type: Assembly error (or heterozygote); Systematic assembly error; Line B mutation

### CONCLUSION

1. HiFi-based assemblies are of high quality, but **not error-free**.
2. **Simple repeat** and **low-depth** regions constitute the two major sources of assembly errors; however, most errors can be mitigated by incorporating Illumina reads.
3. The low occurrence of simple repeats facilitates the **accurate assembly of complex centromeres**.
4. It is advisable to **filter out** "mutations" at simple repeats or with low sequence depth for MA line analysis.
5. The authenticity of mutations is confirmed through replicates.

### RESULTS

#### 1. Distribution of Assembly Errors

| | Depth (x) | Assem. size | #Gaps | Ctg. NG50 | QV | Completeness |
|---|---|---|---|---|---|---|
| A1 | 108.4x | 133.9 Mb | 15 | 14.3 Mb | 58.28 | 98.07% |
| A2 | 96.1x | 133.8 Mb | 18 | 12.4 Mb | 60.47 | 98.07% |
| B1 | 141.5x | 133.6 Mb | 26 | 9.4 Mb | 61.88 | 98.15% |
| B2 | 111.1x | 133.6 Mb | 14 | 13.4 Mb | 60.75 | 98.15% |

HiFi-based assemblies exhibit high quality, yet are **not devoid of errors**.

#### 2. Three Sources

Source I: Simple Repeat Regions

| Process | Error Sou |
|---|---|
| Sequencing | Simple rep<br>Drop in depth a |
| Assembly | Assembly alg |
| Scaffolding | Difference to r |

The majority of assem **HiFi reads**.

#### 4. Distribution o

**Centromeres** exhib

1. Naish M, Alonge M, Wlodzimierz P, et al. The genetic and epigenetic landscap
2. Rabanal F A, Gräff M, Lanz C, et al. Pushing the limits of HiFi assemblies revea

## Xiao Dong

“Decoding HiFi assembly errors: A comprehensive study based on replicates”

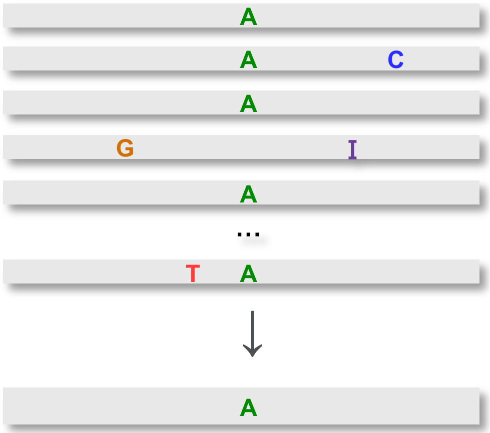
**Poster number: PO0699**

Monday, January 15, 3:00pm - 4:30pm

# A community consensus assembly strategy

## Consensus of 13 independent Col-0 assemblies

- (6 different stocks from 5 different labs)

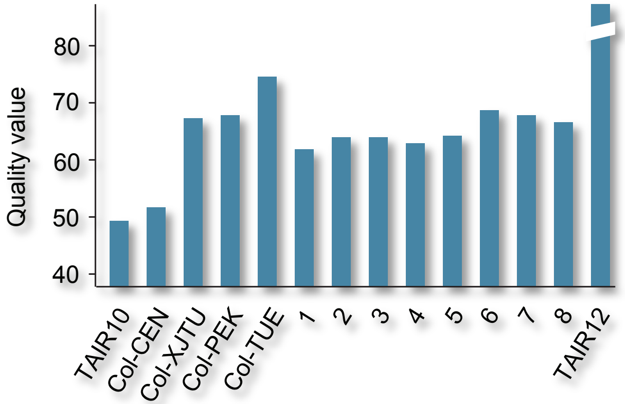


Col-0 community-consensus assembly (**Col-CC**)

## Pros / cons of a consensus appraoch

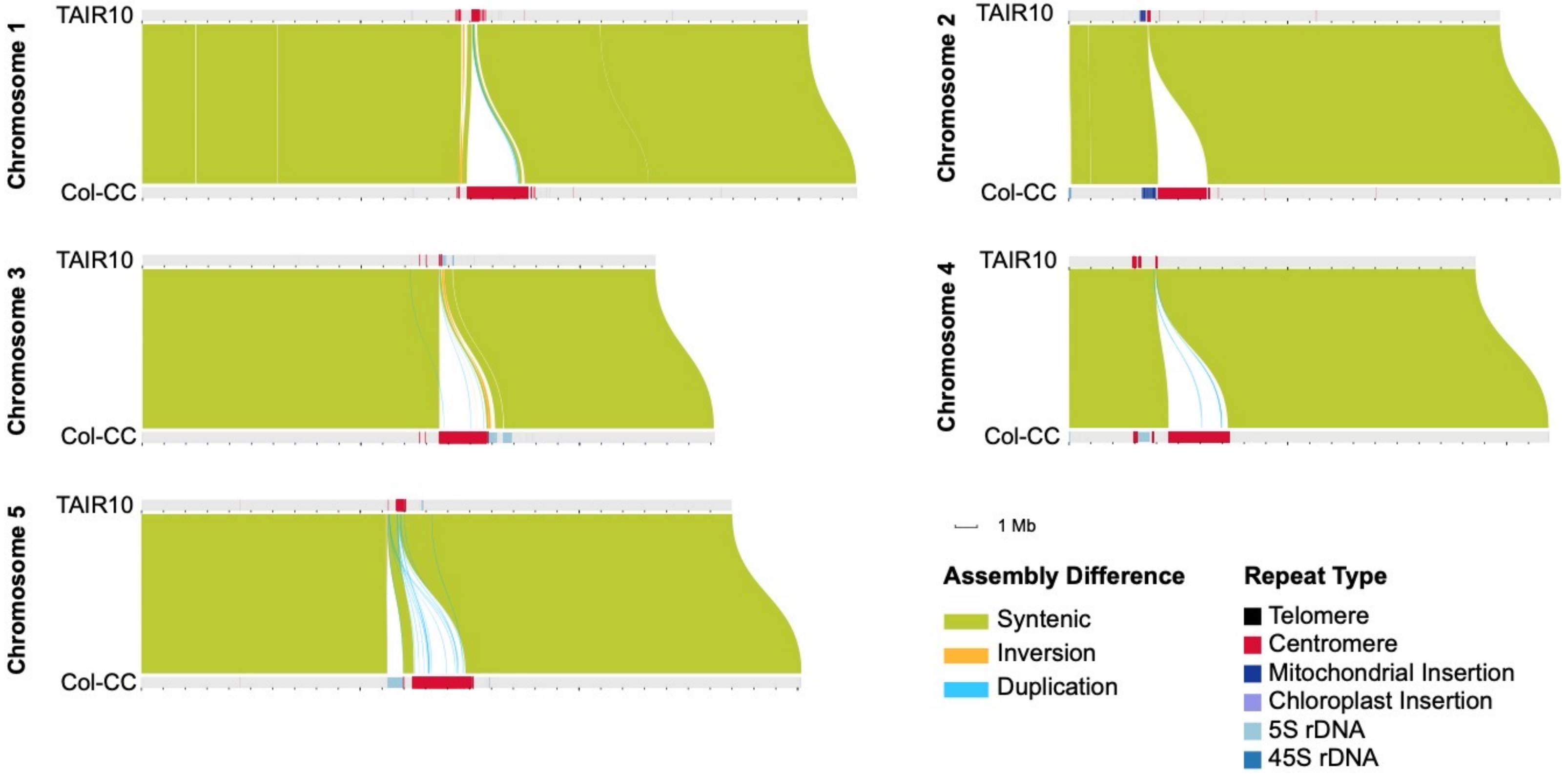
❖ No corresponding individual

❖ Reconstruction of the ancestral Col-0 genome

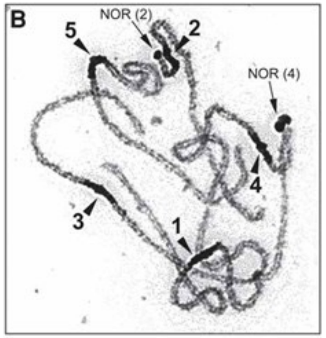


Naish et al, 2021; Wang et al, 2022; Hou et al, 2022; Rabanal et al, 2022

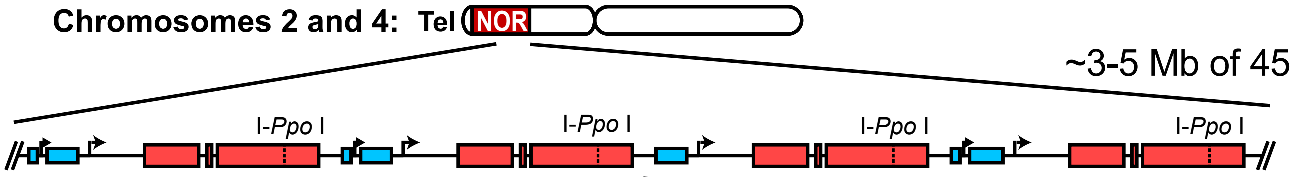
# Col-CC.v1: 133 Mb of exceptional quality



# The assembly of the nucleolus organizer regions (NORs)

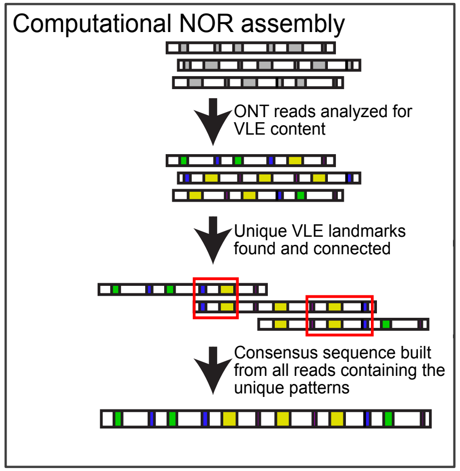


Lysak *et al*, *Methods in Mol Biol*, 2006



Chromosomes 2 and 4:

~3-5 Mb of 45S rDNA repeats


Computational NOR assembly
ONT reads analyzed for VLE content
Unique VLE landmarks found and connected
Consensus sequence built from all reads containing the unique patterns

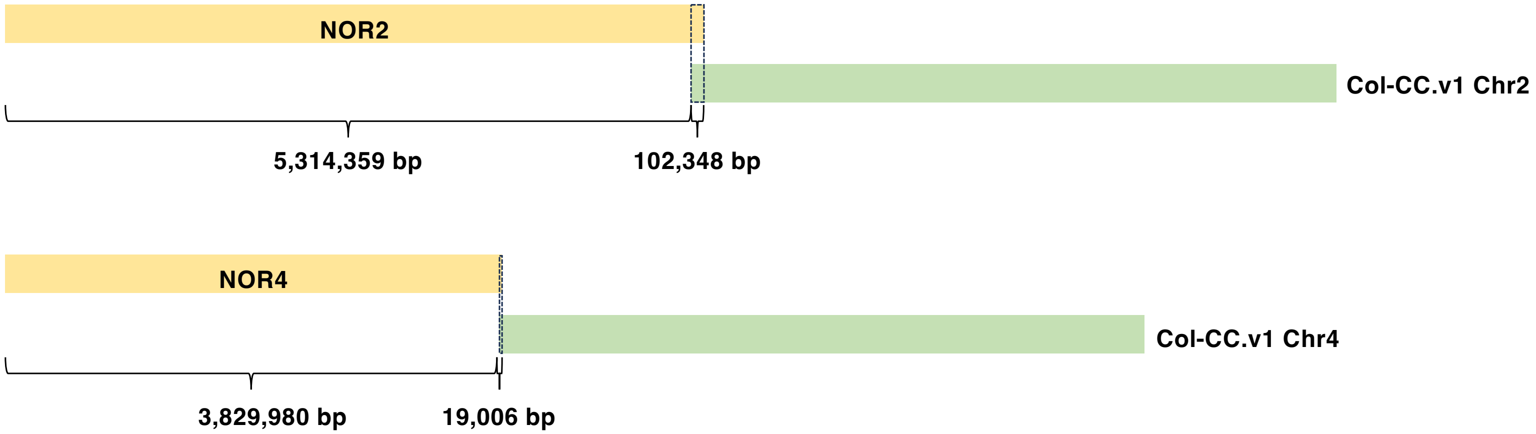
Fultz et al, *Science Advances*, 2023

Assembly of both NOR sequences

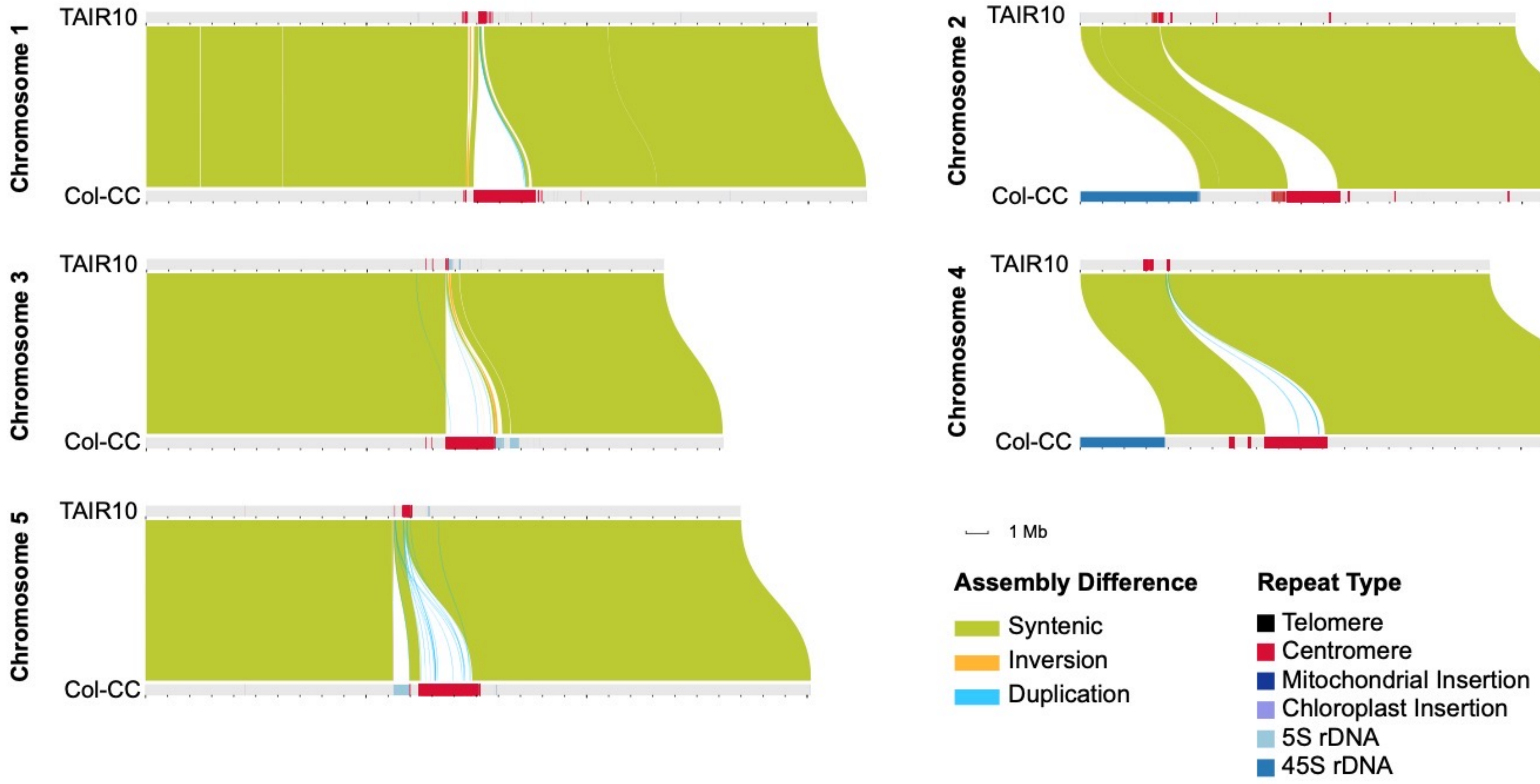
- from telomeres to euchromatin

# Integration of NORs sequences

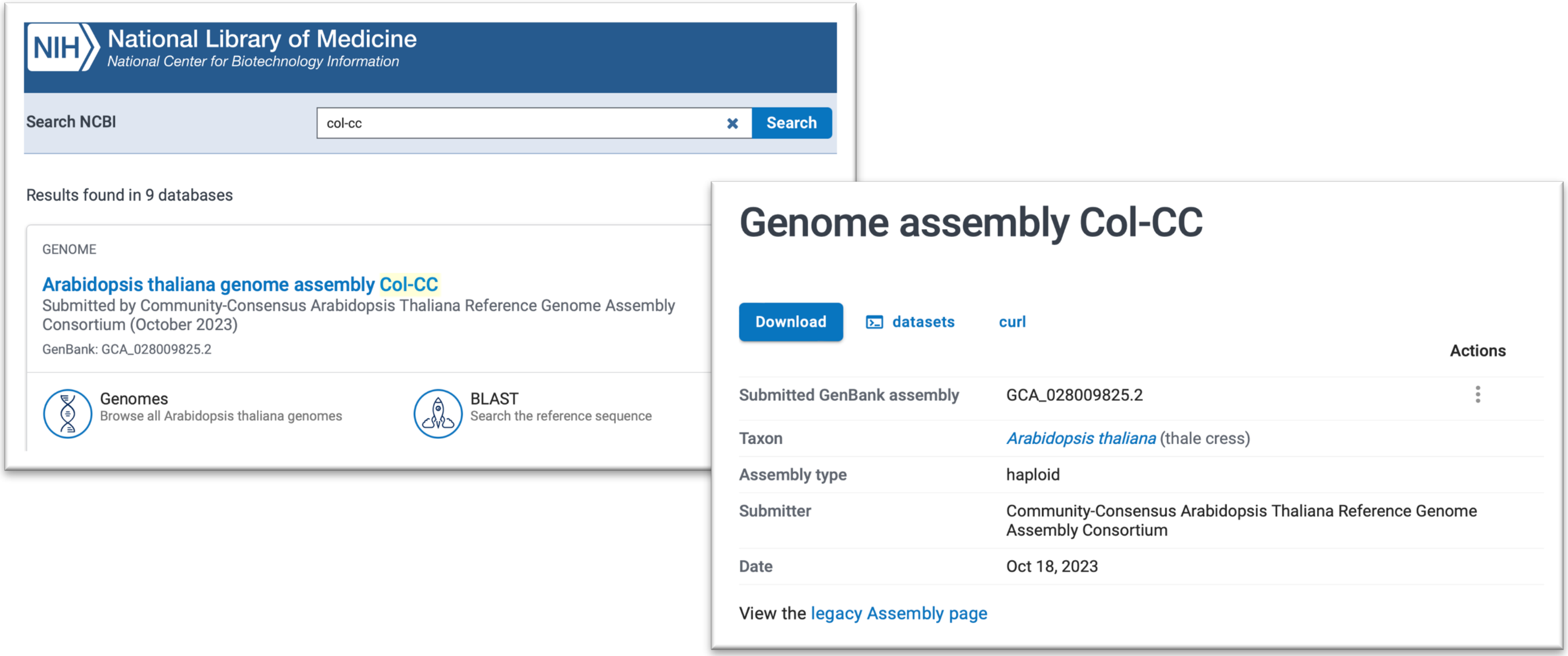
- NOR2 (5.42 Mb) and NOR4 (3.85 Mb) assemblies overlap with Col-CC
- Extension of Col-CC.v1 at the start of Chr2 and Chr4



# Col-CC.v2: first real gap-less T2T assembly of Arabidopsis



# Col-CC.v2 (assembly of TAIR12): publicly available at NCBI



NIH National Library of Medicine
*National Center for Biotechnology Information*

Search NCBI: col-cc ✖ Search

Results found in 9 databases

GENOME

**Arabidopsis thaliana genome assembly Col-CC**
Submitted by Community-Consensus Arabidopsis Thaliana Reference Genome Assembly Consortium (October 2023)
GenBank: GCA_028009825.2

Genomes
Browse all Arabidopsis thaliana genomes

BLAST
Search the reference sequence

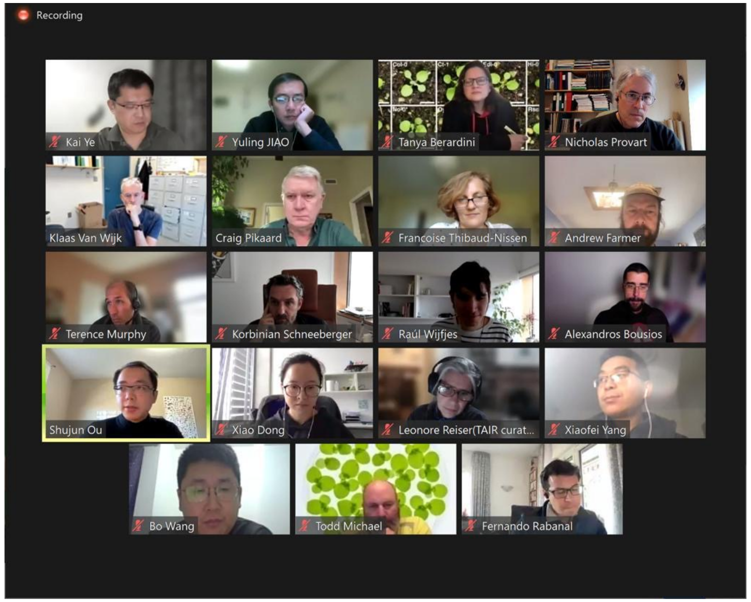
## Genome assembly Col-CC

Download | datasets | curl

| | | Actions |
|---|---|---|
| Submitted GenBank assembly | GCA_028009825.2 | ⋮ |
| Taxon | *Arabidopsis thaliana* (thale cress) | |
| Assembly type | haploid | |
| Submitter | Community-Consensus Arabidopsis Thaliana Reference Genome Assembly Consortium | |
| Date | Oct 18, 2023 | |

View the legacy Assembly page

# Acknowledgement



**Xiao Dong**, Raúl Wijfjes

Henderson, Jiao, Pikaard, Weigel, Ye labs

NCBI Terence Murphy

tair **Tanya Berardini**, Leonore Reiser

Annotation groups incl. >70 researchers world-wide contributing to the annotation

DFG Deutsche Forschungsgemeinschaft