# A Strategy for Balanced Haplotype-Resolved De Novo Assembly of the Autotetraploid Genome of Medicago sativa

Adela Poublan-Couzardot, Bernadette Julier, Marie Pégard, Simon De Givry, Christine Gaspin, Fabrice Legeai, Frédéric Choulet, Christophe Klopp

▶ **To cite this version:**

Adela Poublan-Couzardot, Bernadette Julier, Marie Pégard, Simon De Givry, Christine Gaspin, et al.. A Strategy for Balanced Haplotype-Resolved De Novo Assembly of the Autotetraploid Genome of Medicago sativa. JOBIM 2025 (Journées ouvertes de Bio-Informatique 2025), Jul 2025, Bordeaux, France. ⟨hal-05208435⟩

**HAL Id: hal-05208435**

**https://hal.inrae.fr/hal-05208435v1**

Submitted on 13 Aug 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

# A Strategy for Balanced Haplotype-Resolved De Novo Assembly of the Autotetraploid Genome of Medicago sativa


**Adela Poublan-Couzardot[1]*, Bernadette Julier[2], Marie Pegard[2], Simon de-Givry[3], Christine Gaspin[1], Fabrice Legeai[45], Frederic Choulet[6], Christophe Klopp[1]**

[1] INRAE MIAT, Genotoul-Bioinfo, Auzeville-Tolosane, France
[2] INRAE URP3F, Lusignan, France
[3] INRAE SaAB, MIAT, Auzeville-Tolosane, France
[4] INRAE BIPAA, IGEPP, Le Rheu, France
[5] INRIA, IRISA, GenOuest Core Facility, Rennes, France
[6] INRAE GDEC, Université Clermont Auvergne, France

*Corresponding: adela.poublan-couzardot@inrae.fr


## Background

- *Medicago sativa* (alfa-alfa, luzerne) is an **autotetraploid** forage crop with a large, **repeat-rich** genome.
  - Reference genome[1] : 4x8 CHRS; haplotype ~ 851 Mbp; N50 contig length = 459 kbp; L50 contig count = 1870
  - Haplotype-resolved assembly is difficult due to **high similarity between chromosomes**.
  - Assembly strategies usually adapted for diploids
- Hifiasm + Hi-C – enables de novo phased assembly also for polyploids

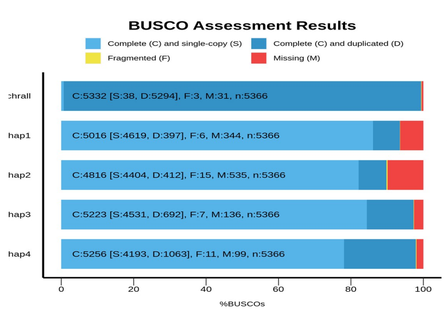
## Material & Methods

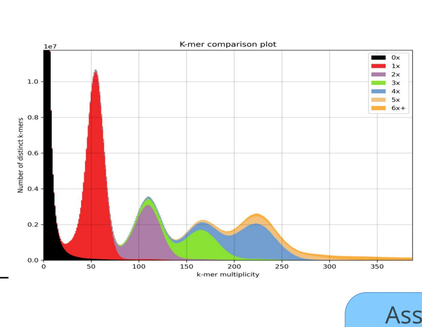
- Data: ~105× HiFi (PacBio) reads + 1 G Hi-C (OmniC) read pairs
- → Genomescope2[2]: ~3.2 Gbp, 86% repeats, 1.72% heterozygosity)
- **Contig assembly:** Hifiasm[3] v0.24.0 (with Hi-C integration, --*nhap* 4)
- l k-mer profile for completeness
- **Post-processing**: set of proteins aligned to contigs with minimap2[5] → contigs reassignment with Toulbar2[6] → Toulbar2 optimization[7] (modification of constraints)
- **Scaffolding**: Hi-C aligned with Juicer[8] (total Hi-C to all haplotypes) → kept within-haplotype aligned Hi-C → scaffolding with 3D-DNA[8] (only 1st round ; MAPQ=0) → manual curation in Juicebox[9]
- heck with DGenies[10] and against the reference genome[1]

## Results

### 1. Contig assembly

| Exp = 750 Mbp | Length (Mbp) |
|---|---|
| HAP1 | 784 |
| HAP2 | 733 |
| HAP3 | 866 |
| HAP4 | 899 |



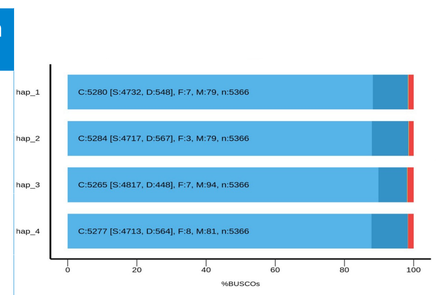



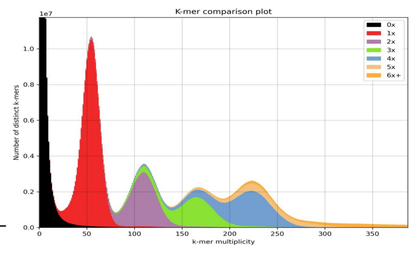

- Hifiasm produced :
  - ✓ Complete assembly – 4 haplotypes with no missing k-mers
  - ✗ Unbalanced haplotypes (varying haplotype lengths and BUSCO scores)

Assumption : Haplotypes in autopolyploids do not vary significantly between each other.

### plotype balancing post-processing

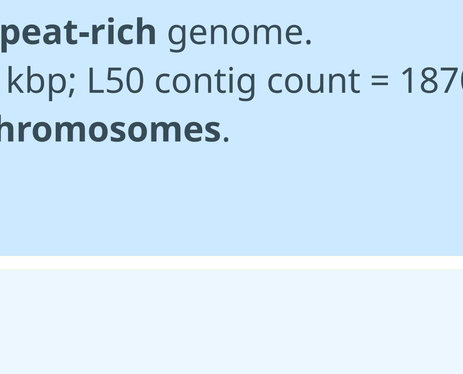
| Exp = 750 Mbp | Length (Mbp) |
|---|---|
| HAP1 | 798 |
| HAP2 | 785 |
| HAP3 | 787 |
| HAP4 | 800 |



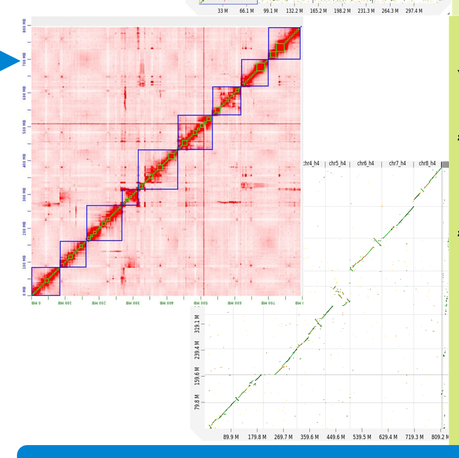

- Reassignment of contigs to haplotypes based on shared protein content :
  - ✓ More even haplotype lengths, better BUSCO balance
  - ✓ Some contigs 'lost', but no effect on the k-mer profile

### 2. Manual curation



- ✗ Chromosome arms not so clear (structure often broken), CHR4 almost completely missing in HAP4
- ✗ Dotplot with many small and big inversions

=> Extensive manual curation needed



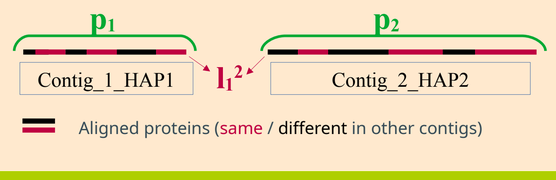
- ✓ All chromosomes present
- ✗ Chromosome arms even less visible, large missarangements
- ✗ Many large invertions, but less small ones

=> Manual curation still needed

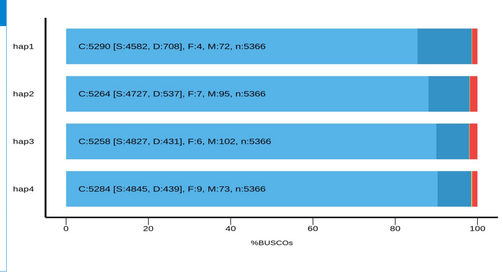
### Toulbar2 optimization

- **total cost = sum($I^2$) + sum(p)**

1. If two contigs are in the same haplotype => the cost is $I^2$ with I (number of shared proteins[2])
2. If a contig is moved from its haplotype of origin => the cost is p (number of proteins aligned in the contigs)



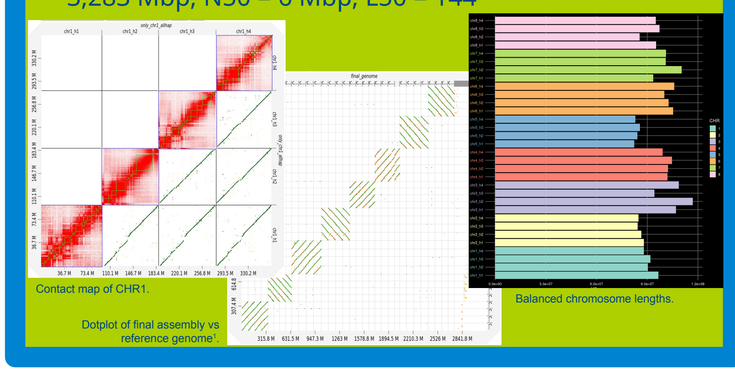

| Exp = 750 Mbp | Length (Mbp) |
|---|---|
| HAP1 | 770 |
| HAP2 | 780 |
| HAP3 | 760 |
| HAP4 | 766 |




- ➢ Toulbar2 is used to find the lowest cost possible.
- ➢ After modification of the two constraints :
- ➢ Toulbar2 reassigned between
- ➢ 21 and 43% contigs/per haplotype
- ➢ 8 and 17% haplotype length in bp
- ✓ Improved contig grouping
- ✓ Balanced haplotype lengths
- ✓ Retained all original contigs

### Final assembly

- ✓ Clear chromosome structure in all haplotypes
- ✓ Minimum inversions/misassemblies
- ✓ Improved collinearity with reference genome
- ➔ ~ 3,283 Mbp; N50 = 6 Mbp; L50 = 144




GITHUB REPO

PREPRINT[7]

## Conclusions

- ➢ Hifiasm + Hi-C produced unbalanced haplotype lengths and BUSCO scores, requiring extensive manual curation after scaffolding.
- ➢ Toulbar2 with added constraints enabled refined scaffolding requiring minimal curation.
- ➢ We obtained balanced haplotype lengths and collinearity despite the large, repetitive autotetraploid genome.

## References


1. Chen H, Zeng Y, Yang Y et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun. 2020;11(2494).
2. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications. 2020;11(1).
3. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods. 2021;18(2).
4. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: Assessing Genomic Data Quality and Beyond. Current Protocols. 2021;1(12).
5. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(3094–3100).
6. Ouali A, Allouche D, de Givry S, Loudni S, Lebbah Y, Loukil L, et al. Variable neighborhood search for graphical model energy minimization. Artificial Intelligence. 2020;278.
7. Klopp C, Durante V, Schiex T, de Givry S. Improving hifiasm haplotypes for autopolyploid genome assemblies using constraint programming. BioRxiv. 2025;04.01.646355.
8. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, Pham M, St Hilaire BG, Yao W, Stamenova E. The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Biorxiv. 2018;254797.
9. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;35(92–95).
10. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6(e4958).