High-quality genome assembly of *Nicotiana tabacum*

using linked reads sequencing and contact map scaffolding

Mohamed Zouine

Toulouse INP

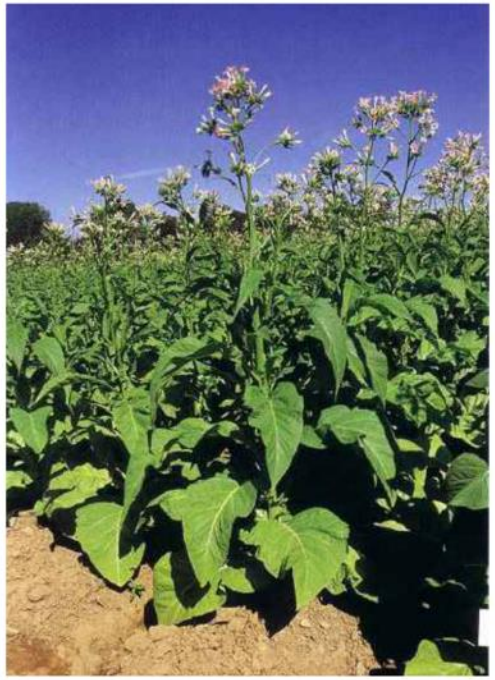
Background

- Tobacco (*Nicotiana tabacum*) is an important plant model system.
- Played a key role in the early development of molecular plant biology.
- The tobacco genome is large, allotetraploid, likely arising from hybridisation between diploid *N. sylvestris* and *N. tomentosiformis* ancestors.
- A genome assembly exists but still need improvement because of the high level of fragmentation due to the genome complexities.

N. tabacum genome

N. Tomentosiformis
2.7Gb
(2n = 24)

X



N. tabacum
Genome size **4.5Gb**
allotetraploid (2n = 4 × = 48)

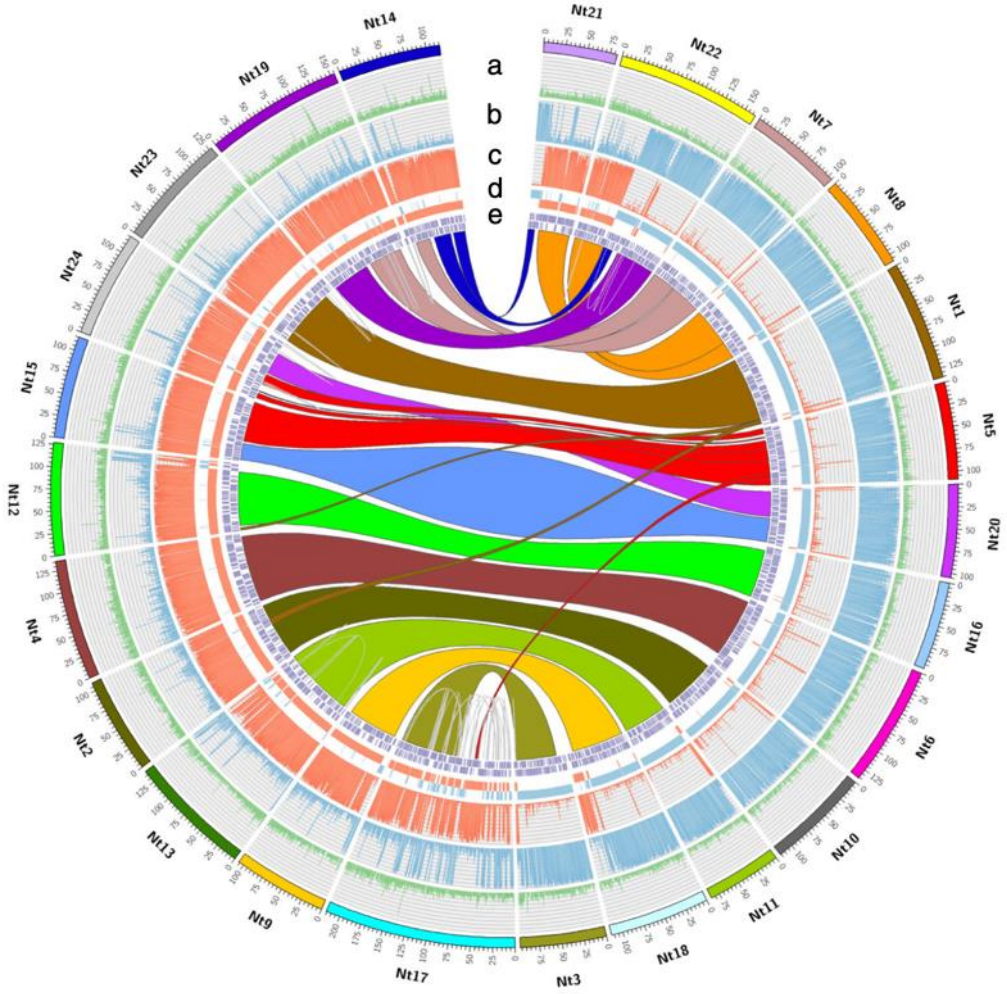
N. Sylvestris
2.6Gb
(2n = 24)

Dr. Julio Thesis

Current tobacco reference genome

Edwards et al. BMC Genomics (2017) 18:448
DOI 10.1186/s12864-017-3791-6

Sequencing technology: short reads
Hybrid assembly: optical map
Genome Size: 4 Gb
N50: 2.17 Mb
Anchored genome: 64 %



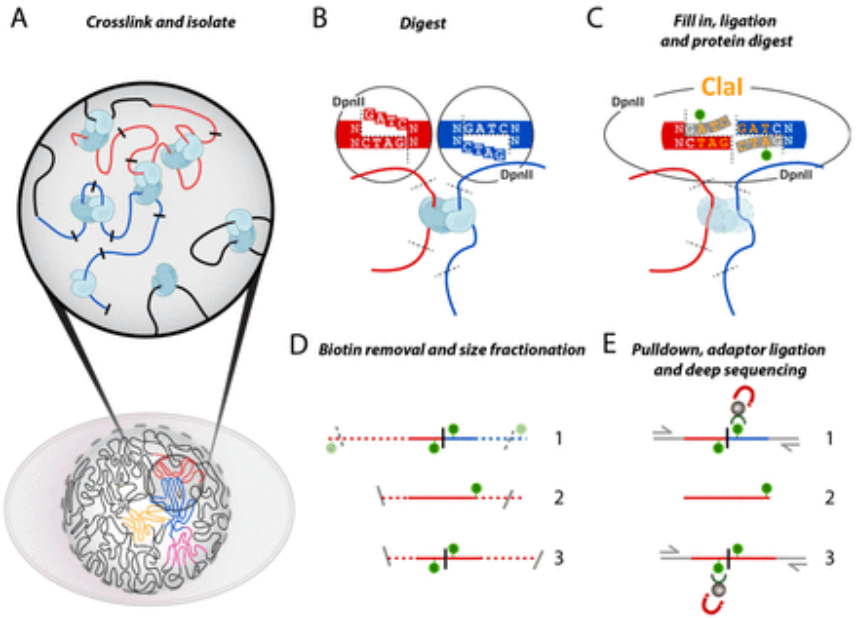
Toward the improvement of the actual genomic resource

10x GENOMICS®

Long read sequencing
Chromium 10x

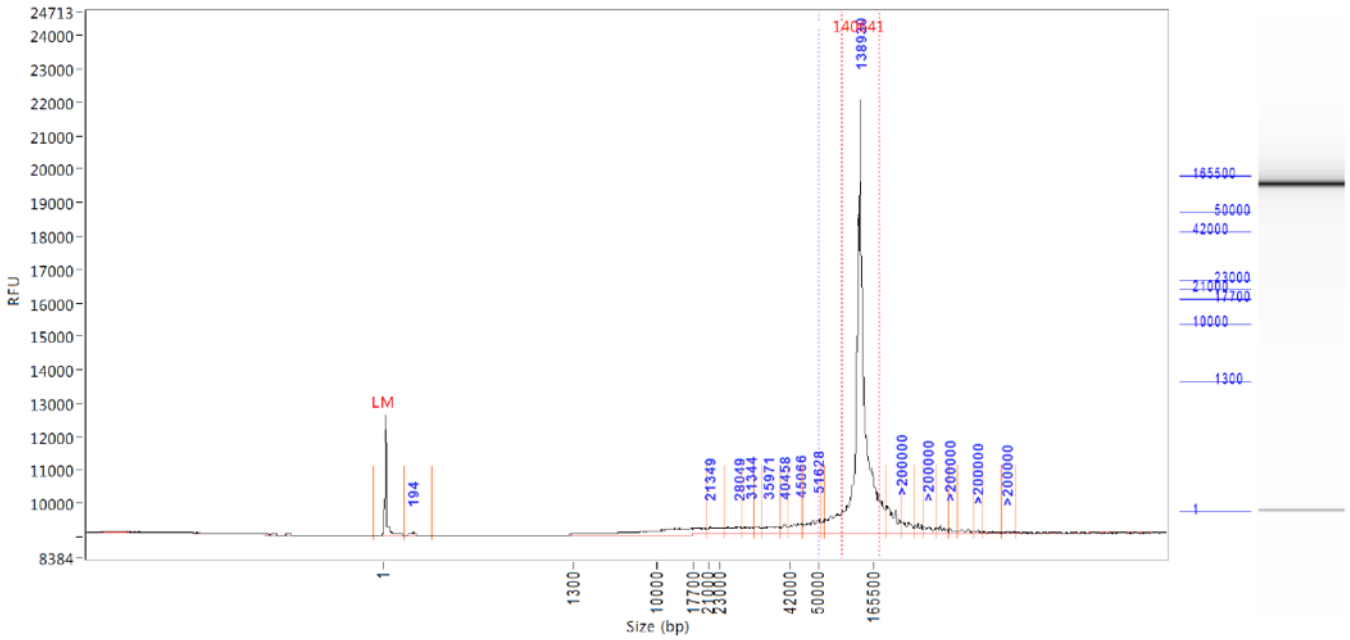
+

Hi-C

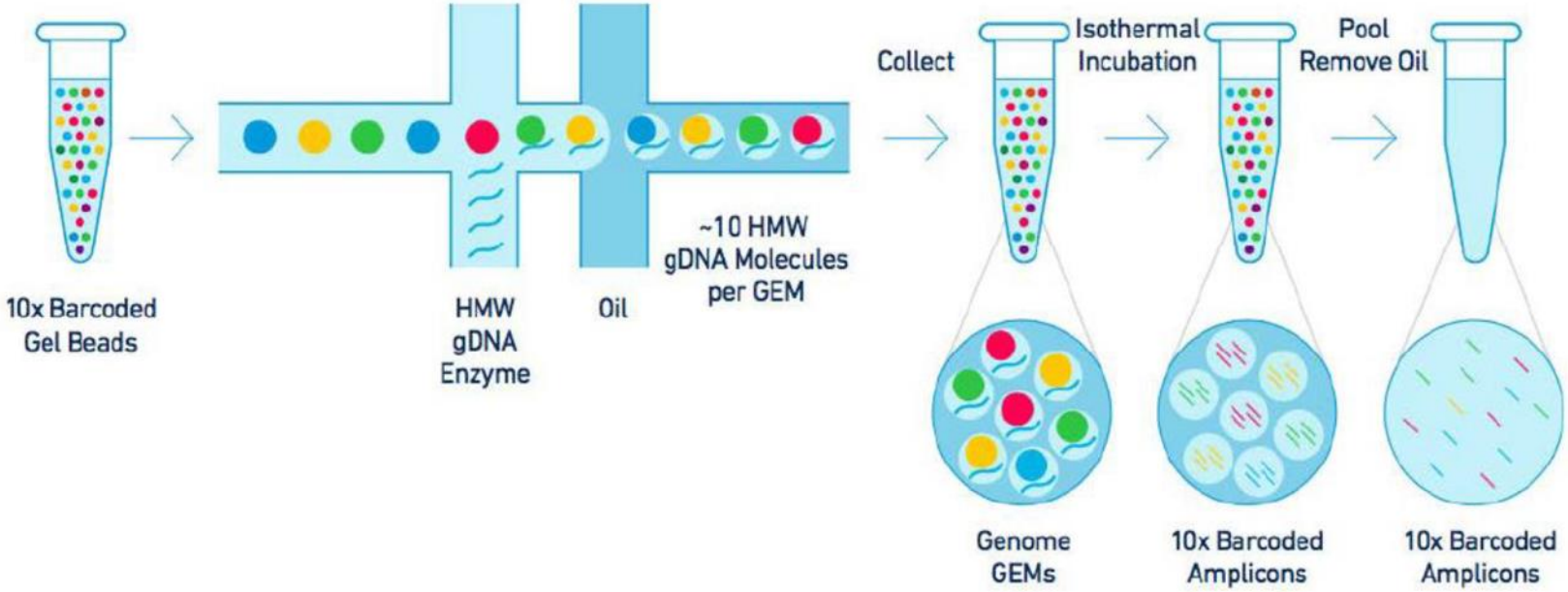


doi: https://doi.org/10.1101/090001

Long molecule DNA extraction



Linked reads library sequencing



40X coverage assembled by supernova software (32 cores, 600 Gb mem, 3 weeks)

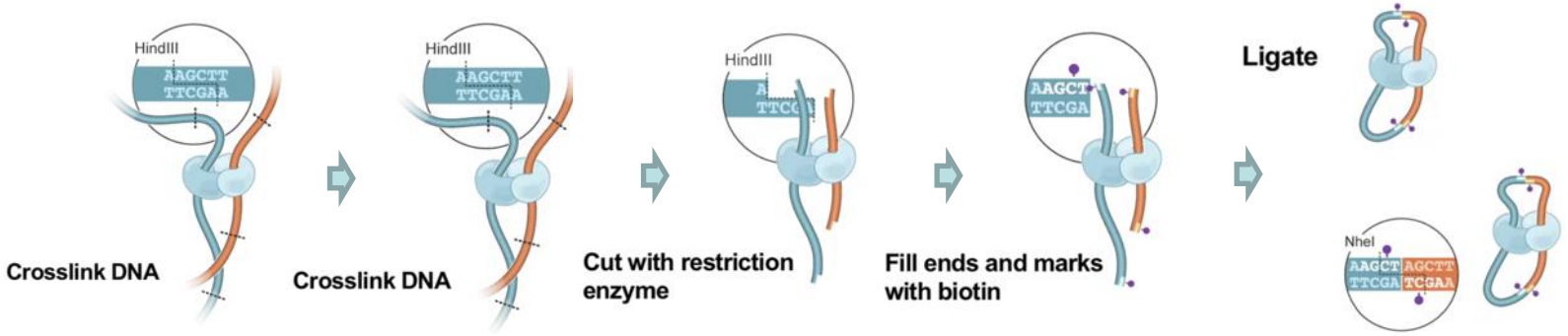
Scaffolds: 105 132

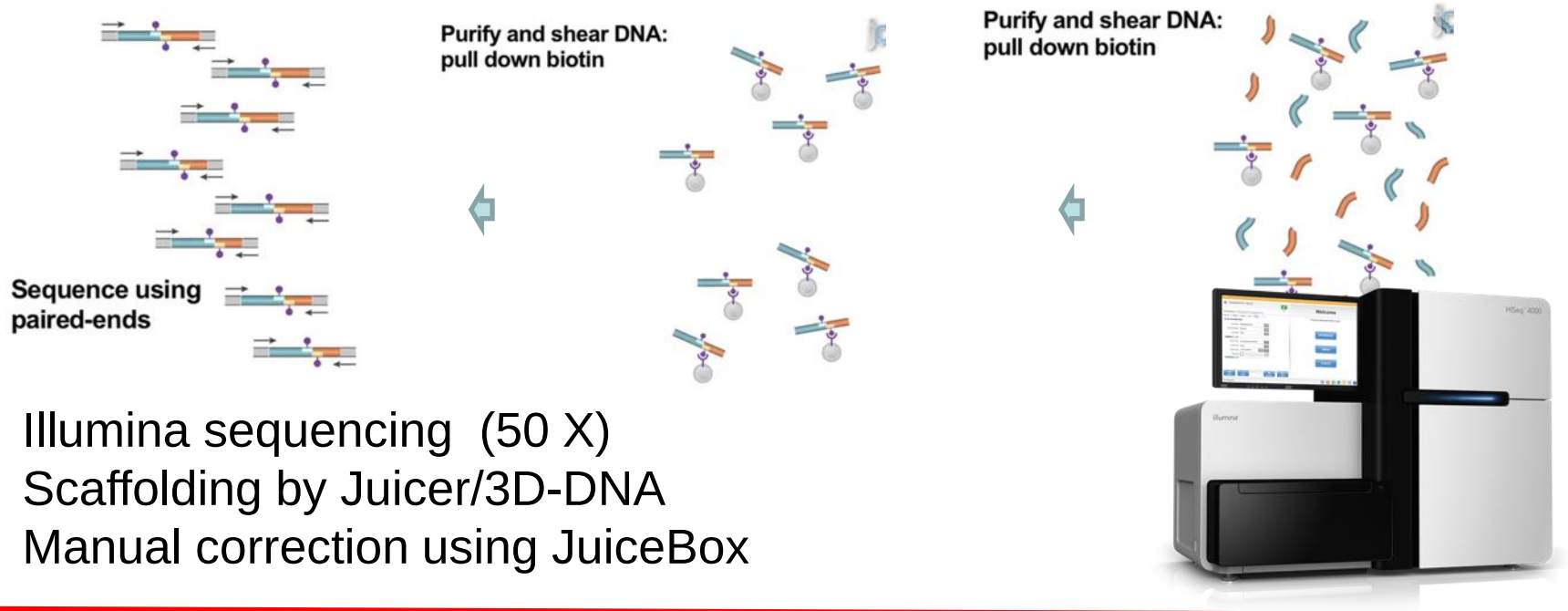
Genome in assembly: 4.2 Gb

N50: 2.1 Gb

L50: 569

Chromosome level scaffolding using Hi-C

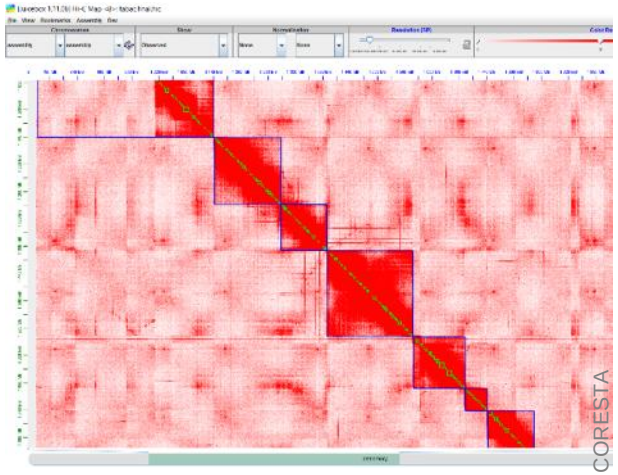




Illumina sequencing (50 X)
Scaffolding by Juicer/3D-DNA
Manual correction using JuiceBox

Combining Long read sequencing technologies

	Linked reads	+ HiC
Scaffolds	105 132	101438
Genome Size	4.2 Gb	4.2 Gb
Biggest Scaffold	12 Mb	190 Mb
N50 scaffold	2.1 Mb	61 Mb
L50 scaffold	569	19



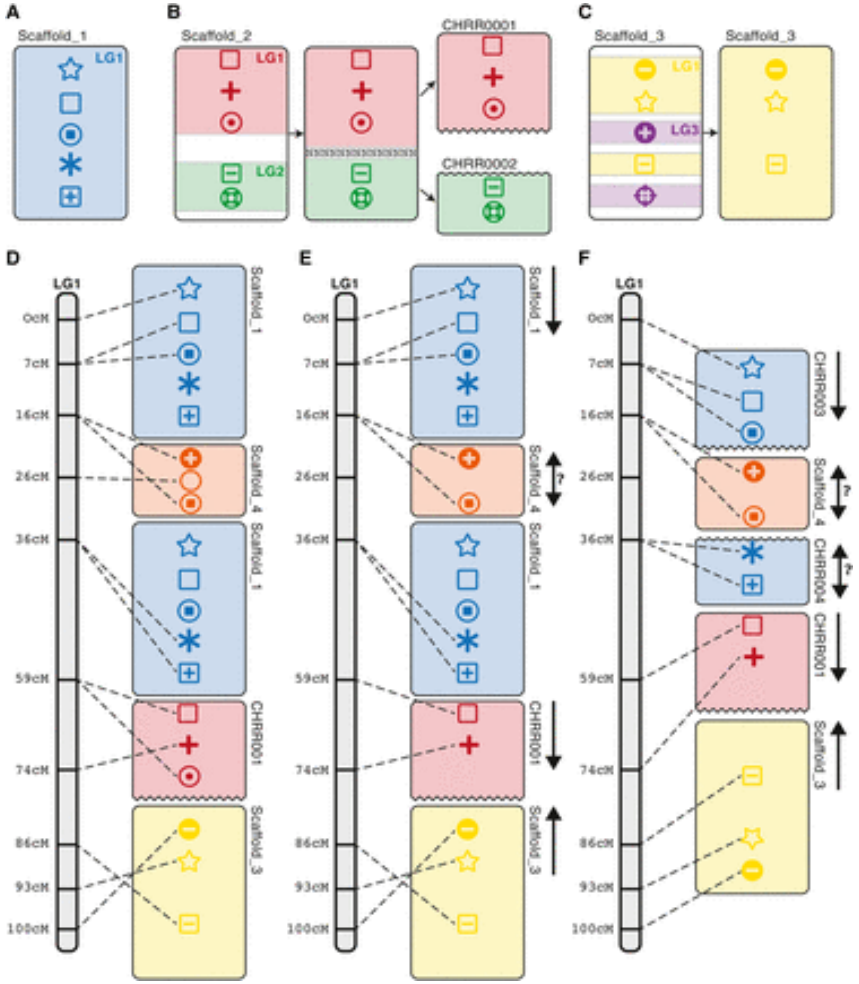
Pseudomolecules building using the genetic map

SNP markers (Imperial tobacco)

SSR markers (from bindler *et al.,* 2011) DOI: 10.1007/s00122-011-1578-8

Chromonomer

(Catchen et al.,2020)

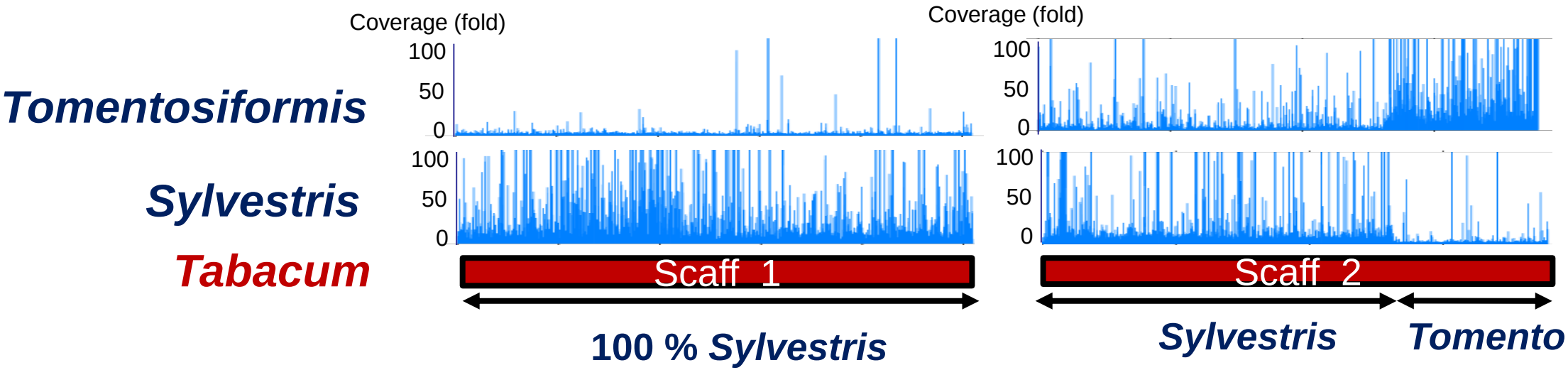


Combining Long read sequencing technologies

	Linked reads	+ Hi-C	+ Genetic Map
Scaffolds	105 132	101 438	24 + CH00
Genome Size	4.2 Gb	4.2 Gb	3,04 Gb (68 % Genome) Gaps (N) 8 %
Biggest Scaffold	12 Mb	190 Mb	200 Mb
N50 scaffold	2.1 Mb	61 Mb	126 Mb
L50 scaffold	569	19	12

Ancestors Origin of each *N. tabacum* chromosome

illumina short reads mapping on Nt chromosomes



Conclusion

- We generate an improved assembly for future research in tobacco plant model using linked reads sequencing, contact map scaffolding and genetic maps.

- Mapping genomic sequences of the its two ancestors allowed to gain insight into the origin of each *N. tabacum* chromosome.

- Genome annotation is ongoing.

Acknowledgments

GAYSSANT Harold
DARNIGE Eden
MAZA Elie
FRASSE Pierre
DJARI Anis
ZOUINE Mohamed

JULIO Emilie
COTUCHEAU Julien
DORLHAC DE BORNE François