# Towards ultra-accurate and complete chromosome-scale Solanaceae reference genomes

Michael Alonge[1], Melanie Kirsche[1], Matthias Benoit[2], Srividya Ramakrishnan[1], Jia He[2], Joyce Van Eck[3,4], Zachary B. Lippman[2,5], Michael C. Schatz[1,2,6]

1. Department of Computer Science, Johns Hopkins University, Baltimore, MD, 2. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 3. Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, 4. Boyce Thompson Institute, Ithaca, NY, 5. Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 6. Department of Biology, Johns Hopkins University, Baltimore, MD

## Modern Plant Genome Assembly

Despite recent advances in long-read sequencing technologies, plant genome assembly remains challenging. In particular, modern assemblies often fail to accurately reconstruct genomic repeats that are pervasive in plant genomes. Large initiatives such as the Vertebrate Genomes Project and the Human Telomere-To-Telomere (T2T) consortium have established best practices and industry standards for assembling animal genomes, but it remains unclear how these approaches translate to plant genomes[1,2]. Here, we use PacBio Circular Consensus accurate long-reads (CCS), Oxford Nanopore ultra-long reads (ONT), and chromatin conformation capture (Hi-C) to establish platinum-level genome assemblies for tomato and groundcherry. We share initial results, new scaffolding methods, and current best practices and recommendations.

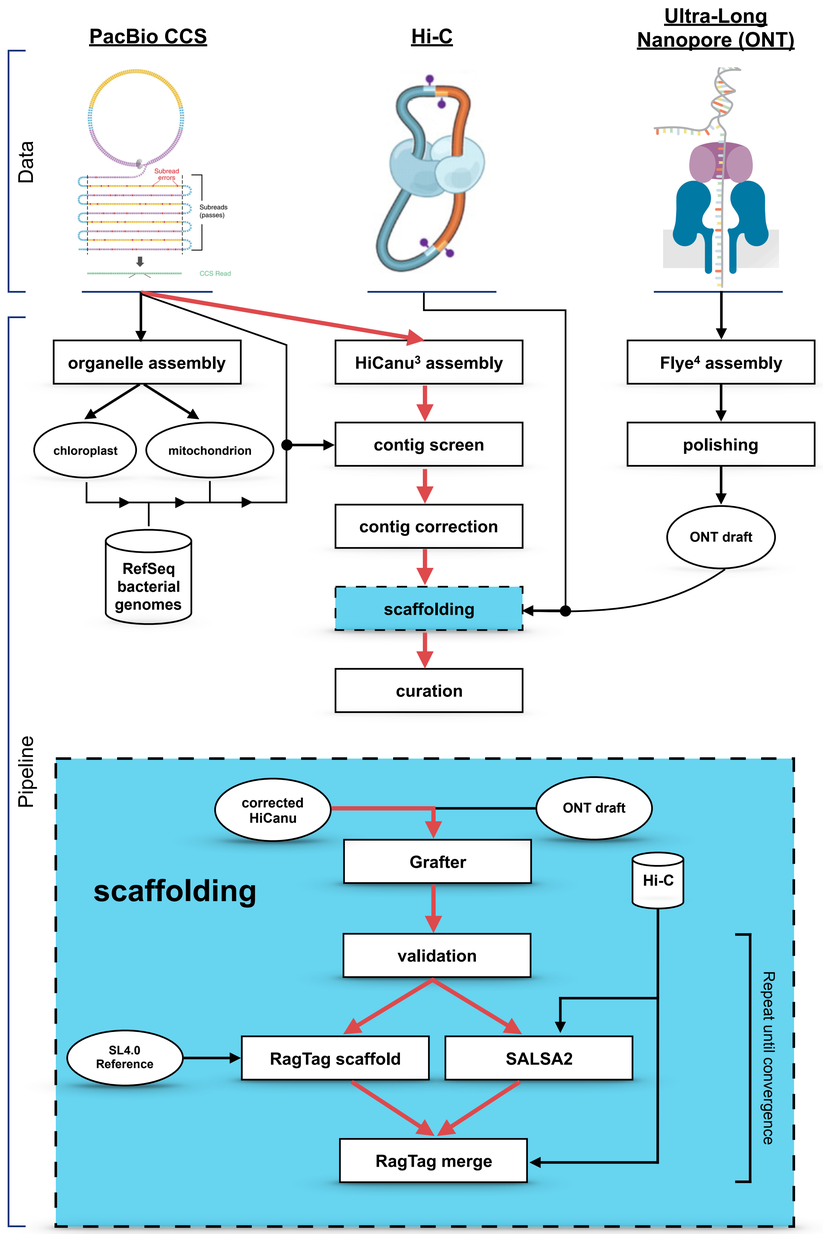


**Figure 1:** A workflow diagram depicting the assembly process. The top "data" panel shows the major data types used in the pipeline[5,6,7]. The "pipeline" panel first shows the general assembly pipeline (top) while the blue portion (bottom) depicts the scaffolding step in more detail. Red arrows indicate the path of the primary, and ultimately final assembly.

## ONT Scaffolding with Grafter

| Sample | Sum (Gbp) | n | N50 (Mbp) | L50 | QV | % Complete |
|---|---|---|---|---|---|---|
| **M82** | 0.822 | 1227 | 10.63 | 23 | 45.1815 | 98.577 |
| ***P. grisea*** | 1.363 | 861 | 18.99 | 23 | 52.2503 | 95.509 |

**Table 1.** Assembly stats for screened HiCanu contigs[8].

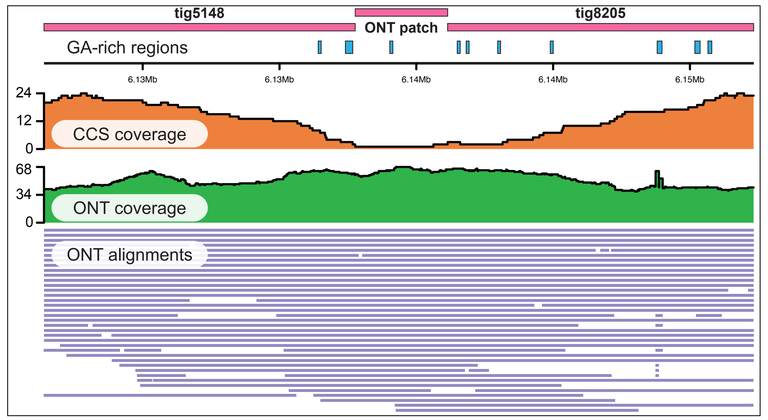


**Figure 2.** A genome browser view of a scaffolding patch made by Grafter[9]. A CCS coverage dropout initially caused a contig break that was subsequently patched and validated by ONT reads.

## Hybrid Scaffolding with RagTag

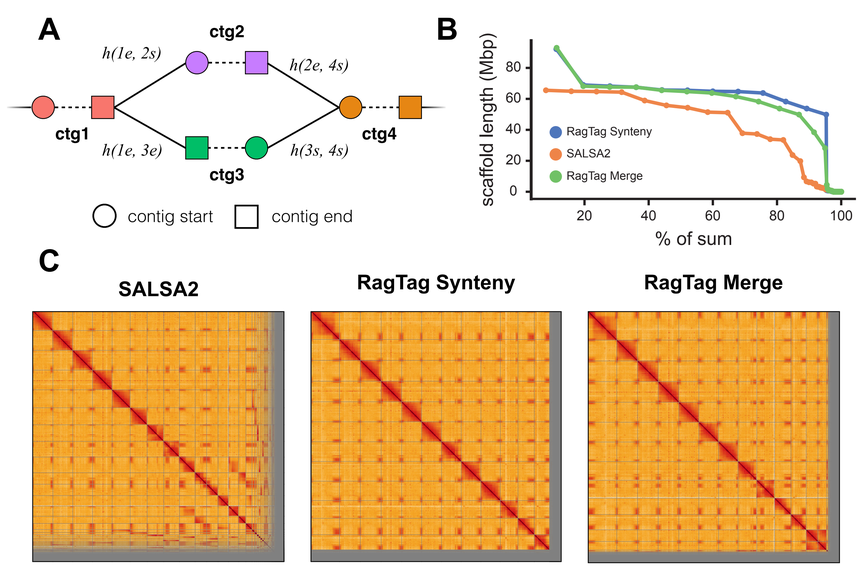


**Figure 3:** M82 HiCanu contigs were scaffolded using two different techniques: Hi-C-based *de novo* scaffolding (SALSA2) and synteny scaffolding (RagTag)[10,11]. A third set of scaffolds was derived by merging these assemblies with RagTag "merge". **(A)** The core data structure of RagTag "merge" represents scaffolding joins as edges in a graph. Each contig is represented by two nodes (start and end) which share an implicit edge (dotted lines). This example depicts a bubble in the graph caused by a scaffolding ambiguity. Edges connecting contigs are weighted by a Hi-C scoring function $h$. **(B)** An Nchart depicting the contiguity of these three assemblies. **(C)** Hi-C heat maps for the three assemblies. RagTag "merge" reaches near-chromosome-scale while leveraging Hi-C to resolve scaffolding ambiguities and avoid reference bias.

## M82 Introgressions Contain Large Inversions

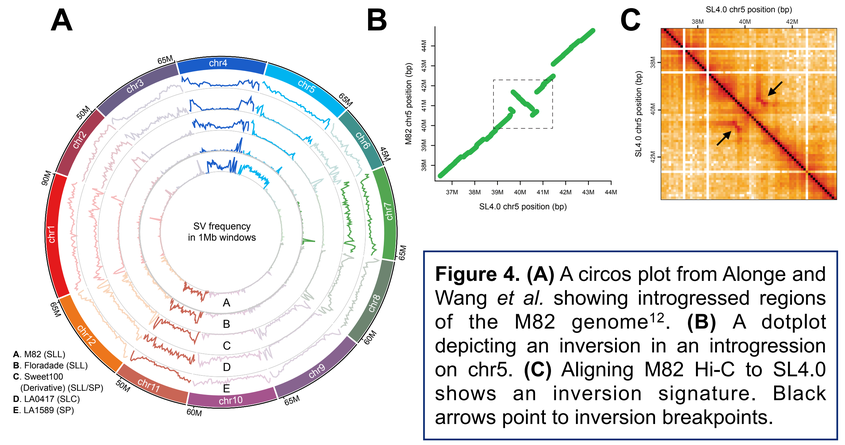


**Figure 4. (A)** A circos plot from Alonge and Wang *et al.* showing introgressed regions of the M82 genome[12]. **(B)** A dotplot depicting an inversion in an introgression on chr5. **(C)** Aligning M82 Hi-C to SL4.0 shows an inversion signature. Black arrows point to inversion breakpoints.

## A New Reference for *Physalis grisea*

We are using these assembly techniques to establish the first reference genome for *Physalis grisea* (groundcherry), a wild nightshade orphan crop. This work highlights the potential to rapidly establish high-quality reference assemblies for wild plant species.

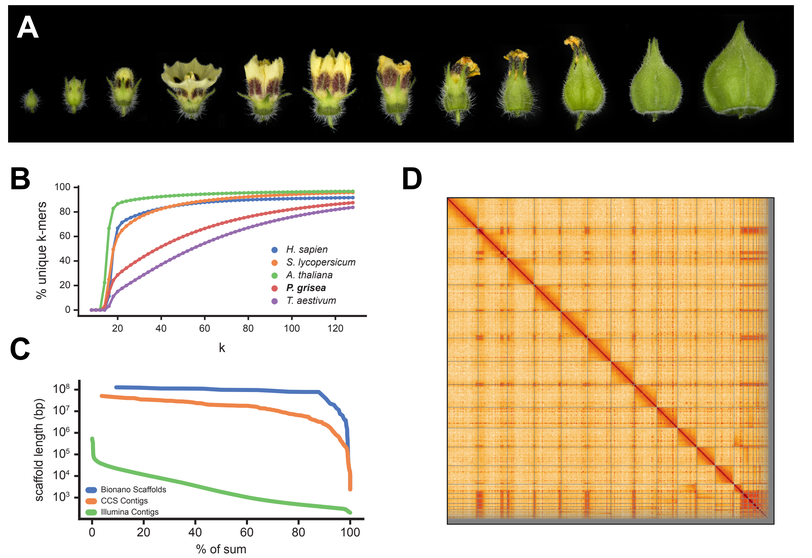


**Figure 5. (A)** Progressive stages of floral organ development for the G20 *P. grisea* accession. **(B)** The proportion of unique *k*-mers as a function of *k* from reference genomes representing multiple species. **(C)** An Nchart of multiple *P. grisea* assemblies[13]. **(D)** A Hi-C heat map of the CCS/Bionano scaffolds.

## References and Acknowledgements

**1.** Rhie, Arang, et al. "Towards complete and error-free genome assemblies of all vertebrate species." *bioRxiv* (2020). **2.** https://sites.google.com/ucsc.edu/t2tworkinggroup. **3.** Nurk, Sergey, et al. "HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads." *BioRxiv* (2020). **4.** Kolmogorov, Mikhail, et al. "Assembly of long, error-prone reads using repeat graphs." *Nature biotechnology* 37.5 (2019): 540-546. **5.** Wenger, Aaron M., et al. "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome." *Nature biotechnology* 37.10 (2019): 1155-1162. **6.** Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *science* 326.5950 (2009): 289-293. **7.** https://nanoporetech.com/sites/default/files/s3/literature/product-brochure.pdf. **8.** Rhie, Arang, et al. "Merqury: reference-free quality and phasing assessment for genome assemblies." *BioRxiv* (2020). **9.** https://github.com/mkirsche/Grafter. **10.** Ghurye, Jay, et al. "Integrating Hi-C links with assembly graphs for chromosome-scale assembly." *PLoS computational biology* 15.8 (2019): e1007273. **11.** https://github.com/malonge/RagTag. **12.** Alonge, Michael, et al. "Major impacts of widespread structural variation on gene expression and crop improvement in tomato." *Cell* 182.1 (2020): 145-161. **13.** Lemmon, Zachary H., et al. "Rapid improvement of domestication traits in an orphan crop by genome editing." *Nature plants* 4.10 (2018): 766-770. We thank Sergey Aganezov for helpful discussions. **We thank the NSF for funding this research.**