The 1001G+ project: A curated collection of *Arabidopsis thaliana* long-read genome assemblies to advance plant research

The 1001 Genomes Plus Consortium*†

*For correspondence: Magnus Nordborg (magnus.nordborg@gmi.oeaw.ac.at) and Detlef Weigel (weigel@weigelworld.org).

†Alphabetical list of authors**:** Carlos C. Alonso-Blanco[1], Haim Ashkenazy[2], Pierre Baduel[3], Zhigui Bao[2], Claude Becker[4], Erwann Caillieux[3], Vincent Colot[3], Duncan Crosbie[4], Louna De Oliveira[3], Joffrey Fitz[2], Katrin Fritschi[2], Elizaveta Grigoreva[5], Yalong Guo[6], Anette Habring[2], Ian Henderson[7], Xing-Hui Hou[6], Yiheng Hu[8], Anna Igolkina[5], Minghui Kang[9], Eric Kemen[8], Paul J. Kersey[10], Aleksandra Kornienko[5], Qichao Lian[11], Haijun Liu[5], Jianquan Liu[9], Miriam Lucke[2], Baptiste Mayjonade[12], Raphaël Mercier[11], Almudena Mollá Morales[5], Andrea Movilli[2], Kevin D. Murray[2], Matthew Naish[7], Magnus Nordborg[5], Fernando A. Rabanal[2], Fabrice Roux[12], Niklas Schandry[4], Korbinian Schneeberger[4,11], Rebecca Schwab[2], Gautam Shirsekar[2], Svitlana Sushko[2], Yueqi Tao[2], Luisa Teasdale[2], Sebastian Vorbrugg[2], Detlef Weigel[2], Wenfei Xian[2]

[1]Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid-28049, Spain
[2]Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[3]Institut de Biologie de l'École Normale Supérieure (IBENS), ENS, 75005, Paris, France
[4]Ludwig Maximilian University München, 82152 Planegg-Martinsried, Germany
[5]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna Biocenter (VBC), 1030 Vienna, Austria
[6]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[7]University of Cambridge, Cambridge, CB2 3EA, United Kingdom
[8]IMIT/ZMBP, University of Tübingen, 72076 Tübingen, Germany
[9]Lanzhou University, State Key Laboratory of Herbage Improvement and Grassland Agro-ecosystems, Lanzhou 730000, China
[10]Royal Botanic Gardens, Kew, Richmond, London TW9 3AE, United Kingdom
[11]Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany
[12]LIPME, INRAE, CNRS, Université de Toulouse, 3!26 Castanet-Tolosan, France

Abstract

Arabidopsis thaliana was the first plant for which a high-quality genome sequence became available. The publication of the first reference genome sequence almost 25 years ago was already accompanied by genome-wide data on sequence polymorphisms in another accession, or naturally occurring strain. Since then, inventories of genome-wide diversity have been generated at increasingly precise levels. High-density genotype data for *A. thaliana*, including those from the 1001 Genomes Project, were key to demonstrating the enormous power of GWAS in inbred populations of wild plants, and the comparison of intraspecific polymorphism with interspecific divergence has illuminated many aspects of plant genome evolution. Over the past decade, an increasing number of nearly complete genome sequences have been published for many more accessions. Here, we highlight the diversity of a curated collection of previously published and so far unpublished genome sequences assembled using different types of long reads, including PacBio Continuous Long Reads (CLR), PacBio High Fidelity (HiFi) reads, and Oxford Nanopore Technologies (ONT) reads. This 1001 Genomes Plus (1001G+) resource is being made available at http://1001genomes.org. We invite colleagues with yet unpublished genome assemblies from *A. thaliana* accessions to contribute to this effort.

Introduction

The human HapMap and 1000 Genomes Projects paved the way for the species-wide description of variation across the genome (The International Hapmap Consortium 2003; Birney and Soranzo 2015), but plants have not been far behind, with *Arabidopsis thaliana* being the second species for which genome-wide SNP data became available in 2007 (Kim et al. 2007). The same year saw the initiation of the 1001 Genomes Project for *A. thaliana*, an international, informal collaborative effort that demonstrated the practicality and power of resequencing large, publicly available collections of inbred lines that are perfect for genome-wide association studies (GWAS) (Atwell et al. 2010; 1001 Genomes Consortium 2016).

There was, however, a “dirty” secret underlying these all efforts, namely that the majority of genetic variation was largely ignored (Igolkina et al. 2024). While single nucleotide polymorphisms (SNPs) or small insertions and deletions (indels) are, with some caveats, accessible to short-read sequencing, the field has been mostly blind to major structural variants (SVs), such as large deletions, inversions, duplications, or polymorphic insertions of transposable elements (TEs). The importance of such major sequence differences has been documented in numerous functional studies (Weigel and Nordborg 2015; Alonge et al. 2020; Simon et al. 2022; Zhou et al. 2022; Jayakodi et al. 2024).

In response to the short-comings of short-read resequencing, several long-read technologies have been developed. In the last few years, dramatic advances in accuracy and cost of long-read sequencing have been made, allowing for population-scale analyses (De Coster et al. 2021). In *A. thaliana*, initial reports focused on individual genome sequences from non-reference strains, followed by larger and larger collections of long-read genome assemblies (Zapata et al. 2016;

Jiao and Schneeberger 2020; Kang et al. 2023; Wlodzimierz et al. 2023b; Igolkina et al. 2024; Lian et al. 2024).

We describe ongoing efforts to generate a curated collection of *A. thaliana* genome sequences assembled from long reads, including both previously published and so far unpublished assemblies. We refer to this curated collection as the 1001 Genomes Plus (1001G+) resource. We are in the process of making these assemblies available at http://1001genomes.org, and we invite colleagues with collections of yet unpublished assemblies to contribute to this effort.

Results

We collected published *A. thaliana* chromosome-level assemblies that were primarily generated with either long reads from Pacific Biosciences (PacBio) platforms (Continuous Long Reads [CLR], Consensus Circular Sequencing [CCS]/High Fidelity [HiFi] reads) or the Oxford Nanopore Technology (ONT) platform (Berlin et al. 2015; Chin et al. 2016; Zapata et al. 2016; Michael et al. 2018; Goel et al. 2019; Pucker et al. 2019; Jiao and Schneeberger 2020; Barragan et al. 2021; Naish et al. 2021; Wang et al. 2021, 2023; Hou et al. 2022; Rabanal et al. 2022; Wibowo et al. 2022; Christenhusz et al. 2023; Jaegle et al. 2023; Kang et al. 2023; Wlodzimierz et al. 2023b; Igolkina et al. 2024; Jiang et al. 2024; Kileeg et al. 2024; Lian et al. 2024; Tao et al. 2024; Teasdale et al. 2024). We also reached out to colleagues who we knew were still in the process of generating long-read assemblies. An overview of the number of genome assemblies from the different sources is provided in **Table 1**, and their diversity is indicated in **Figure 1**. In general, PacBio CLR-based assemblies reach excellent accuracy in the chromosome arms but are lacking most of the centromeres. PacBio HiFi-based assemblies have excellent accuracy and often span entire centromeres and 5S ribosomal RNA gene (rDNA) clusters within single contigs (Rabanal et al. 2022). ONT based assemblies have very good accuracy and include centromeric sequences, but these tend to be fragmented and do not span over chromosome arms. Finally, 45S rDNA arrays are very difficult to reconstruct completely; their assembly requires dedicated efforts that combine a variety of approaches (Fultz et al. 2023). The canonical chromosome number in *A. thaliana* is five, and many assemblies include telomeres for the eight chromosome arms that have non-repetitive sequences at their ends. The ends of two chromosome arms, 2p and 4p, consist of very large 45S rDNA arrays; telomeres can only be manually anchored to these two chromosome arms (Tao et al. 2024).

Table 1. Overview of *A. thaliana* genome assemblies in the 1001G+ collection.

Techn.	Institution	Publ.	Unpubl.	References for published assemblies
Legacy	Community	1		Lamesch et al. 2012
PacBio CLR	Max Planck Institute (MPI) for Plant Breeding Research	9		Zapata et al. 2016; Goel et al. 2019; Jiao and Schneeberger 2020
	Bielefeld Univ.	1		Pucker et al. 2019
	MPI for Biology Tübingen	2	5	Barragan et al. 2021; Wibowo et al. 2022
	Gregor Mendel Institute of Molecular Plant Biology (GMI)	6	135	Jaegle et al. 2023
	University of Chinese Academy of Sciences	2		Wang et al. 2023; Jiang et al. 2024
	GMI/MPI for Biology Tübingen	27		Igolkina et al. 2024
	Institute of Botany, Chinese Academy of Sciences		2	
	Univ.of Tübingen		16	
PacBio HiFi	Community		2	https://phoenixbioinformatics.atlassian.net/wiki/spaces/COM/pages/42215873/2024-01-15+PAG31+Summary
	Darwin Tree of Life	1		Christenhusz et al. 2023
	MPI for Biology Tübingen	72	87	Rabanal et al. 2022; Wlodzimierz et al. 2023; Tao et al. 2024; Teasdale et al. 2024
	Lanzhou Univ.	32		Kang et al. 2023
	MPI for Plant Breeding Research	48		Lian et al. 2024
ONT	Univ.of Cambridge	1	1	Naish et al. 2021
	Xi’an Jiaotong Univ.	1		Wang et al. 2021
	Univ. of Chinese Academy of Sciences	1		Hou et al. 2022
	MPI for Plant Breeding Research	24		Lian et al. 2024
	Univ. of Toronto	8		Kileeg et al. 2024
	Institut de Biologie de l'Ecole Normale Superieure (IBENS)		89	
	Ludwig Maximilian Univ. Munich		26	
Total		**238**	**361**	

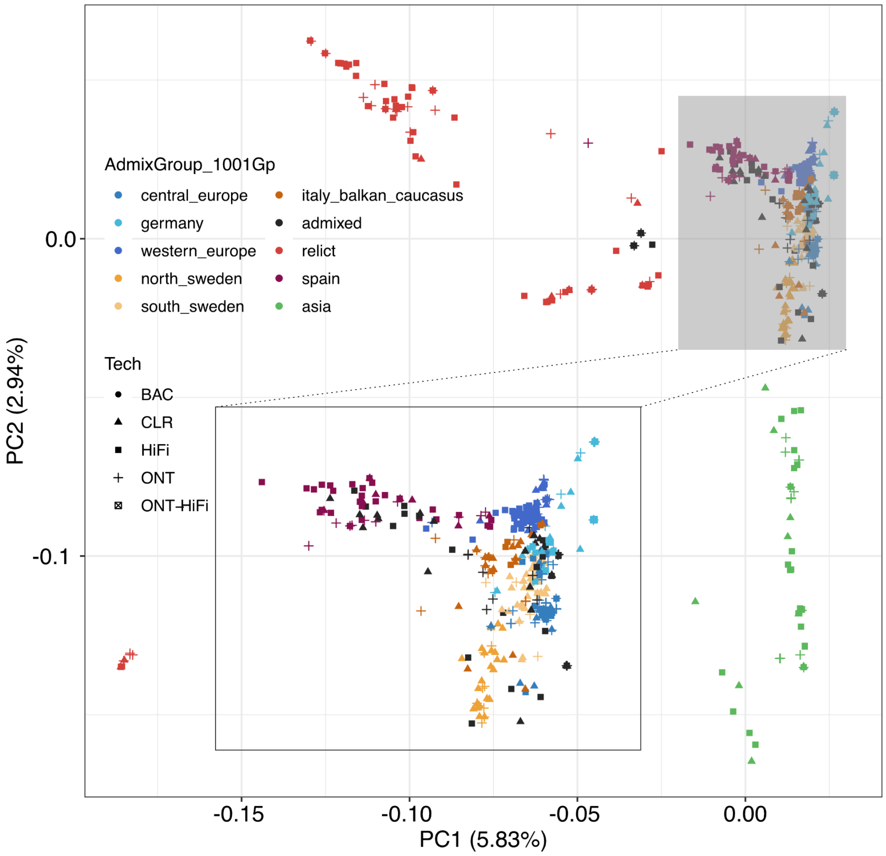

Figure 1. Diversity of *A. thaliana* assemblies. A Principal Component (PC) Analysis was performed based on bialleic SNPs from whole genome alignments of 581 assemblies to Col-CC (GCA_028009825.2, TAIR12). Colors indicate previously defined admixture groups (1001 Genomes Consortium 2016), shapes sequencing platforms. BAC, TAIR10 reference genome; ONT-HiFi, hybrid assembly from multiple Col-0 accessions. A few diverse assemblies, which are still being curated, are not yet included.

The assemblies are being made available in the Data Center of the 1001 Genomes project (https://1001genomes.org). None of the contigs in the original data sets are being reassembled, but we are performing quality checks for the scaffolding. Most of the available assemblies have been previously scaffolded based on reference genome information (Alonge et al. 2022), a process prone to errors. These errors often manifest as misplaced contigs, particularly those rich in repetitive sequences such as rDNAs and centromere satellite repeats, or organellar DNA contigs mistaken as true nuclear insertions (Rabanal et al. 2022). Another common error in publically available assemblies is the incorrect orientation of contigs in structurally variable regions. This is especially notable in the short arm of chromosome 4, where the reference accession Col-0 has a 1.7 Mb inversion spanning 1.17 Mb leading to the formation of a heterochromatic knob (Fransz et al. 2000), which is rare in the global *A. thaliana* population (Fransz et al. 2016). We therefore are visually inspecting all assemblies and using population information on structural variants to change contig joins that appear to be likely scaffolding errors due to reference bias. We are also identifying assemblies that are not from the accessions they are supposed to be based on previous short-read genotyping data, or that appear to have come from substantially identical material, sometimes with different accession identifiers. Finally, where original long reads are available, we are identifying regions of residual heterozygosity or potentially collapsed sequences in the assemblies. While the haploid assemblies are correct in the sense that the polymorphisms are present in the *A. thaliana* population, the heterozygous regions are reduced to a single haplophase that does not correspond to any haplotype that exists in the real world.

The total number of assemblies we have collected so far is 596 from 463 accessions. Minimizing also the number of accessions that are substantially related throughout their entire genomes (such as natural or lab-reared mutation accumulation lines (Exposito-Alonso et al. 2018; Monroe et al. 2022)), distinguished by fewer true SNPs than sequencing and assembly errors, the tally of unique assemblies currently stands at 438.

Outlook

We are planning to release the complete set of curated assemblies in the second quarter of 2025. Analyses that the Weigel and Nordborg labs are currently conducting with these assemblies include the following:

- Annotation of nuclear sequences using liftoff (Shumate and Salzberg 2021), Helixer (Stiehler et al. 2021), and a custom pipeline that makes use of Iso-seq full-length cDNA data from 16 accessions (Teasdale et al. 2024).
- Assembly and annotation of plastid genomes (Xian et al. 2024).
- Annotation of telomere repeats (Tao et al. 2024).
- Annotation of ribosomal RNA genes and centromere satellite repeats (Wlodzimierz et al. 2023a).
- Annotation of TEs (Ou et al. 2019; Igolkina et al. 2024; Sierra and Durbin 2024).
- Analysis of synteny blocks (Wang et al. 2024).
- Annotation of short tandem repeats (Readman et al. 2021).
- Development of a JBrowse 2-based genome browser that allows for choosing any of the genomes as base for comparison with all other genomes (Diesh et al. 2023).

We encourage colleagues who are generating additional genome assemblies of *A. thaliana* accessions or who are conducting either related or complementary analyses with available assemblies to join the 1001 Genomes Plus Project by contacting its coordinators, Magnus Nordborg (magnus.nordborg@gmi.oeaw.ac.at) and Detlef Weigel (weigel@weigelworld.org). Similarly to the original 1001 Genomes Project, which was enormously successful despite never having received coordinated funding, we hope that the 1001 Genomes Plus Project can showcase the power of a federated, community-driven bottom-up approach to the generation of powerful resources for evolutionary genomics.

Methods

Variant calling

We aligned 581 *A.thaliana* and two *A. lyrata* genome assemblies (Kolesnikova et al. 2023) to the Col-CC reference genome (GCA_028009825.2, TAIR12) using wfmash (v0.13) with parameters "-s 10k -p 90 -n 1". Variants were called across multiple samples with minipileup (https://github.com/lh3/minipileup, v1.0) using "-s0 -a0 -q0 -l 20000 -vcC -f Col-CC.fa". Variants in centromeric and rDNA regions were excluded using bcftools (Danecek et al. 2021). Genomes were also aligned to TAIR10, and genotype IDs were verified by scores >0.98 with SNPmatch (Pisupati et al. 2017) and the 1001G callset as the database.

Population genetics analysis

We created two biallelic SNP subsets: one including *A. lyrata* as an outgroup for a neighbor-joining (NJ) tree, and the other for kinship and principal component analysis (PCA). For the NJ tree, we included sites where both *A. lyrata* accessions shared the same allele and filtered for a minor allele frequency (MAF) >0.01 and a missing rate <20%. Distances were calculated with VCF2Dis (https://github.com/BGI-shenzhen/VCF2Dis, v1.50), and the tree was constructed using fneighbor in PHYLIP (Felsenstein 1989). For PCA and kinship analyses, we excluded *A. lyrata* and applied the same filtering thresholds. PCA was conducted using GCTA (Yang et al. 2011) (v1.94), retaining the first 20 principal components, and related individuals were identified with KING (Manichaikul et al. 2010) (–related). Admixture groups were assigned based on clustering in the neighbor-joining tree, with missing labels resolved through majority voting among labeled samples from SNPmatch-based IDs within each clade.

Acknowledgments

We thank all members of the global *A. thaliana* community who have directly or indirectly contributed to the collection of germplasm and the generation of genome sequences. Supported by grants PID2022-136893NB-I00 from the MCIN/AEI/10.13039/501100011033 and FEDER (EU) (Carlos Alonso-Blanco), TRR356 (DFG 491090170) (Niklas Schandry), ANR STEVE (ANR-21-CE45-0018) (Vincent Colot), the Austrian Academy of Sciences (Magnus Nordborg), the Novozymes Prize of the Novo Nordisk Foundation and the Max Planck Society (Detlef Weigel). The initial phase of the 1001 Genomes Plus Project in the Kersey, Nordborg and Weigel labs was supported by DFG, FWF and BBSRC through ERA-CAPS grant 1001G+.

Competing Interests

D.W. holds equity in Computomics, which advises plant breeders. D.W. also consults for KWS SE, a globally active plant breeder and seed producer. J.F. is an employee of Tropic TI, Lda. All other authors declare no competing interests.

References

1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016:166(2):481–491.

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, and Soyk S. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022:23(1):258.

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. 2020:182(1):145–161.e23.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010:465(7298):627–631.

Barragan AC, Collenberg M, Wang J, Lee RRQ, Cher WY, Rabanal FA, Ashkenazy H, Weigel D, and Chae E. A Truncated Singleton NLR Causes Hybrid Necrosis in Arabidopsis thaliana. Mol Biol Evol. 2021:38(2):557–574.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, and Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015:33(6):623–630.

Birney E and Soranzo N. Human genomics: The end of the start for population sequencing. Nature. 2015:526(7571):52–53.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016:13(12):1050–1054.

Christenhusz MJM, Twyford AD, Hudson A, Royal Botanic Gardens Kew Genome Acquisition Lab, Royal Botanic Garden Edinburgh Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, and Darwin Tree of Life Consortium. The genome sequence of thale cress, Arabidopsis thaliana (Heynh., 1842). Wellcome Open Res. 2023:8:40.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021:10(2). https://doi.org/10.1093/gigascience/giab008

De Coster W, Weissensteiner MH, and Sedlazeck FJ. Towards population-scale long-read sequencing. Nat Rev Genet. 2021:22(9):572–587.

Diesh C, Stevens GJ, Xie P, De Jesus Martinez T, Hershberg EA, Leung A, Guo E, Dider S, Zhang J, Bridge C, et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biol. 2023:24(1):74.

Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, Brachi B, Hagmann J, Grimm DG, Chen J, et al. The rate and potential relevance of new mutations in a colonizing plant lineage. PLoS Genet. 2018:14(2):e1007155.

Felsenstein J. PHYLIP - Phylogeny inference package - v3.2. https://doi.org/10.1111/j.1096-0031.1989.tb00562.x

Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, Zabel P, Bisseling T, and Jones GH. Integrated cytogenetic map of chromosome arm 4S of A. thaliana: structural organization of heterochromatic knob and centromere region. Cell. 2000:100(3):367–376.

Fransz P, Linc G, Lee C-R, Aflitos SA, Lasky JR, Toomajian C, Ali H, Peters J, van Dam P, Ji X, et al.

Molecular, genetic and evolutionary analysis of a paracentric inversion in Arabidopsis thaliana. Plant J. 2016:88(2):159–178.

Fultz D, McKinlay A, Enganti R, and Pikaard CS. Sequence and epigenetic landscapes of active and silent nucleolus organizer regions in Arabidopsis. Sci Adv. 2023:9(44):eadj4509.

Goel M, Sun H, Jiao W-B, and Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019:20(1):277.

Hou X, Wang D, Cheng Z, Wang Y, and Jiao Y. A near-complete assembly of an Arabidopsis thaliana genome. Mol Plant. 2022:15(8):1247–1250.

Igolkina A, Vorbrugg S, Rabanal F, Liu H-J, Ashkenazy H, Kornienko A, Fitz J, Collenberg M, Kubica C, Morales AM, et al. Towards an unbiased characterization of genetic polymorphism. bioRxiv. 2024:2024.05.30.596703. https://doi.org/10.1101/2024.05.30.596703

Jaegle B, Pisupati R, Soto-Jiménez LM, Burns R, Rabanal FA, and Nordborg M. Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity. Genome Biol. 2023:24(1):44.

Jayakodi M, Lu Q, Pidon H, Rabanus-Wallace MT, Bayer M, Lux T, Guo Y, Jaegle B, Badea A, Bekele W, et al. Structural variation in the pangenome of wild and domesticated barley. Nature. 2024. https://doi.org/10.1038/s41586-024-08187-1

Jiang J, Xu Y-C, Zhang Z-Q, Chen J-F, Niu X-M, Hou X-H, Li X-T, Wang L, Zhang YE, Ge S, et al. Forces driving transposable element load variation during Arabidopsis range expansion. Plant Cell. 2024:36(4):840–862.

Jiao W-B and Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 2020:11(1):1–10.

Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, Liu W, Chen C, Song Y, Tan L, et al. The pan-genome and local adaptation of Arabidopsis thaliana. Nat Commun. 2023:14(1):6259.

Kileeg Z, Wang P, and Mott GA. Chromosome-Scale Assembly and Annotation of Eight Arabidopsis thaliana Ecotypes. Genome Biol Evol. 2024:16(8):evae169.

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, and Nordborg M. Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat Genet. 2007:39(9):1151–1155.

Kolesnikova UK, Scott AD, Van de Velde JD, Burns R, Tikhomirov NP, Pfordt U, Clarke AC, Yant L, Seregin AP, Vekemans X, et al. Transition to self-compatibility associated with dominant S-allele in a diploid Siberian progenitor of allotetraploid Arabidopsis kamchatica revealed by Arabidopsis lyrata genomes. Mol Biol Evol. 2023:40(7):msad122.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012:40(Database issue):D1202–10.

Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger K, and Mercier R. A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. Nat Genet. 2024:56(5):982–991.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, and Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010:26(22):2867–2873.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, and Ecker JR. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun. 2018:9(1):541.

Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, Klein M, Hildebrandt J, Neumann M, Kliebenstein D, et al. Mutation bias reflects natural selection in Arabidopsis thaliana.

Nature. 2022:602:101–105.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, et al. The genetic and epigenetic landscape of the Arabidopsis centromeres. Science. 2021:374(6569):eabi7489.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019:20(1):275.

Pisupati R, Reichardt I, Seren Ü, Korte P, Nizhynska V, Kerdaffrec E, Uzunova K, Rabanal FA, Filiault DL, and Nordborg M. Verification of Arabidopsis stock collections using SNPmatch, a tool for genotyping high-plexed samples. Sci Data. 2017:4(1):170184.

Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, and Weisshaar B. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLoS One. 2019:14(5):e0216233.

Rabanal FA, Gräff M, Lanz C, Fritschi K, Llaca V, Lang M, Carbonell-Bejerano P, Henderson I, and Weigel D. Pushing the limits of HiFi assemblies reveals centromere diversity between two Arabidopsis thaliana genomes. Nucleic Acids Res. 2022. https://doi.org/10.1093/nar/gkac1115

Readman C, Indhu-Shree R-B, Jan M F, and Inanc B. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol. 2021:22(1):224.

Shumate A and Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021:37(12):1639–1643.

Sierra P and Durbin R. Identification of transposable element families from pangenome polymorphisms. Mob DNA. 2024:15(1):13.

Simon M, Durand S, Ricou A, Vrielynck N, Mayjonade B, Gouzy J, Boyer R, Roux F, Camilleri C, and Budar F. APOK3, a pollen killer antidote in Arabidopsis thaliana. Genetics. 2022. https://doi.org/10.1093/genetics/iyac089

Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, and Denton AK. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics. 2021:36(22-23):5291–5298.

Tao Y, Xian W, Bao Z, Rabanal FA, Movilli A, Lanz C, Shirsekar G, and Weigel D. Atlas of telomeric repeat diversity in Arabidopsis thaliana. Genome Biology. 2024:25(1):1–18.

Teasdale LC, Murray KD, Collenberg M, Contreras-Garrido A, Schlegel T, van Ess L, Jüttner J, Lanz C, Deusch O, Fitz J, et al. Pangenomic context reveals the extent of intraspecific plant NLR evolution. bioRxiv. 2024. https://doi.org/10.1101/2024.09.02.610789

The International Hapmap Consortium. The International HapMap Project. Nature. 2003:426(6968):789–796.

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, et al. High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads. Genomics Proteomics Bioinformatics. 2021. https://doi.org/10.1016/j.gpb.2021.08.003

Wang W, Qin L, Zhang W, Tang L, Zhang C, Dong X, Miao P, Shen M, Du H, Cheng H, et al. WeiTsing, a pericycle-expressed ion channel, safeguards the stele to confer clubroot resistance. Cell. 2023:186(12):2656–2671.e18.

Wang Y, Tang H, Wang X, Sun Y, Joseph PV, and Paterson AH. Detection of colinear blocks and synteny and evolutionary analyses based on utilization of MCScanX. Nat Protoc. 2024:19(7):2206–2229.

Weigel D and Nordborg M. Population genomics for understanding adaptation in wild plant species. Annu

Rev Genet. 2015:49:315–338.

Wibowo AT, Antunez-Sanchez J, Dawson A, Price J, Meehan C, Wrightsman T, Collenberg M, Bezrukov I, Becker C, Benhamed M, et al. Predictable and stable epimutations induced during clonal plant propagation with embryonic transcription factor. PLoS Genet. 2022:18(11):e1010479.

Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E, Scott A, Mandáková T, Gorringe N, Tock AJ, Holland D, et al. Cycles of satellite and transposon evolution in Arabidopsis centromeres. Nature. 2023. https://doi.org/10.1038/s41586-023-06062-z

Xian W, Bezrukov I, Bao Z, Vorbrugg S, Gautam A, and Weigel D. TIPPo: A user-friendly tool for DE Novo assembly of organellar genomes with HiFi data. bioRxiv. 2024:2024.01.29.577798. https://doi.org/10.1101/2024.01.29.577798

Yang J, Lee SH, Goddard ME, and Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011:88(1):76–82.

Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, Patel V, Velikkakam James G, Koornneef M, Ossowski S, et al. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci U S A. 2016:113(28):E4052–60.

Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K, et al. Graph pangenome captures missing heritability and empowers tomato breeding. Nature. 2022:606(7914):527–534.