

Phased pangenomes empower potato breeding

Lin Cheng



Supervisors: Prof. Sanwen Huang & Ludivine Lassois

Year: 2025

COMMUNAUTÉ FRANÇAISE DE BELGIQUE
UNIVERSITÉ DE LIÈGE – GEMBLoux AGRO-BIO TECH

Phased pangenomes empower potato breeding

Lin Cheng

Exploiter un pangenome phasé pour la conception d'haplotypes de pommes de terre
hybrides

Promoteurs: Prof. Sanwen Huang & Ludivine Lassois
Année civile: 2025

Copyright. Cette œuvre est sous licence Creative Commons. Vous êtes libre de reproduire, de modifier, de distribuer et de communiquer cette création au public selon les conditions suivantes:

- paternité (BY): vous devez citer le nom de l'auteur original de la manière indiquée par l'auteur de l'œuvre ou le titulaire des droits qui vous confère cette autorisation (mais pas d'une manière qui suggérerait qu'ils vous soutiennent ou approuvent votre utilisation de l'œuvre);

- pas d'utilisation commerciale (NC): vous n'avez pas le droit d'utiliser cette création à des fins commerciales;

- partage des conditions initiales à l'identique (SA): si vous modifiez, transformez ou adaptez cette création, vous n'avez le droit de distribuer la création qui en résulte que sous un contrat identique à celui-ci. À chaque réutilisation ou distribution de cette création, vous devez faire apparaître clairement au public les conditions contractuelles de sa mise à disposition. Chacune de ces conditions peut être levée si vous obtenez l'autorisation du titulaire des droits sur cette œuvre. Rien dans ce contrat ne diminue ou ne restreint le droit moral de l'auteur.

© Lin Cheng 30/06/2025

Abstract

Lin Cheng (2025). “Phased pangenomes empower potato breeding” (PhD Dissertation in English).

Gembloux, Belgium, Gembloux Agro-Bio Tech, University of Liege.

96 pages, 26 figures, 6 tables.

Summary:

Potato is the most widely consumed tuber crop globally, yet breeding progress is limited by its complex, highly heterozygous, and often tetraploid genome. Traditional clonal propagation increases costs and disease risks, while the lack of inbred lines makes hybrid breeding difficult. The Upotato Plan aims to transform potato into a seed-propagated diploid crop that can be improved through modern hybrid breeding strategies. However, hidden deleterious mutations in the heterozygous genome have posed a major obstacle.

To address this, we assembled a phased pangenome using 31 genetically diverse diploid potato accessions, including wild species and indigenous cultivars. This resource reveals extensive haplotype diversity and enables accurate detection of deleterious single nucleotide polymorphism (dSNPs) and deleterious structural variants (dSV). Notably, we discovered a “broken window” effect in the genome, where deleterious SNPs accumulate near harmful structural variants, compounding genetic burden. To reduce this load, we computationally designed “ideal haplotypes” with minimal harmful mutations, offering a powerful framework for precise parental selection and accelerated breeding.

Our work demonstrates that phased pangenomes are essential for understanding potato genomic diversity and for guiding next-generation breeding strategies that can overcome long-standing barriers in potato improvement.

Key words: Potato, genome, pangenome, haplotype, structural variations (SVs), deleterious mutations, deleterious SVs.

Résumé

Lin Cheng (2025). "Phased pangenomes empower potato breeding" (thèse de doctorat en anglais).

Gembloux, Belgique, Gembloux Agro-Bio Tech, Université de Liège.

96 pages, 26 figures, 6 tableaux.

Résumé:

La pomme de terre est le tubercule le plus consommé au monde. Pourtant, les progrès de la sélection sont limités par son génome complexe, hautement hétérozygote et souvent tétraploïde. La propagation clonale traditionnelle augmente les coûts et les risques de maladies, tandis que l'absence de lignées consanguines rend la sélection hybride difficile. Le Plan Upotato vise à transformer la pomme de terre en une culture diploïde propagée par semences, améliorable grâce à des stratégies modernes de sélection hybride. Cependant, des mutations délétères cachées dans le génome hétérozygote constituent un obstacle majeur.

Pour y remédier, nous avons assemblé un pangénome phasé à partir de 31 accessions diploïdes de pomme de terre génétiquement diverses, incluant des espèces sauvages et des cultivars indigènes. Cette ressource révèle une grande diversité d'haplotypes et permet une détection précise des polymorphismes mononucléotidiques délétères (dSNP) et des variants structuraux délétères (dSV). Nous avons notamment découvert un effet de « fenêtre brisée » dans le génome, où les SNP délétères s'accumulent à proximité des variants structuraux nocifs, aggravant ainsi la charge génétique. Pour réduire cette charge, nous avons conçu par ordinateur des « haplotypes idéaux » présentant un minimum de mutations néfastes, offrant ainsi un cadre performant pour une sélection parentale précise et une sélection accélérée.

Nos travaux démontrent que les pangénomes phasés sont essentiels à la compréhension de la diversité génomique de la pomme de terre et à l'élaboration de stratégies de sélection de nouvelle génération, capables de surmonter les obstacles persistants à l'amélioration de la pomme de terre.

Mots clés: Pomme de terre, génome, pangénome, haplotype, variations structurelles (SVs), mutations délétères, SV délétères.

Acknowledgments

First and foremost, I am deeply thankful to Prof. Sanwen Huang (黄三文) & Ludivine Lassois, who have jointly guided me in my Ph.D. studies 🌹. I would also like to express my sincere gratitude to all those who have supported me throughout the course of this research.

I would also like to express my heartfelt gratitude to my tutor in China, Prof Sanwen, who gave me the opportunity to participate in cutting-edge research and taught me how to become a scientist. He supported me throughout my entire Ph.D. journey. To be honest, without his expert guidance, I would not have achieved what I have today. His profound insights into science have truly left a lasting impression on me throughout my research. Prof. Sanwen is a great teacher, one who does not impose rigid rules on students. Instead, he encourages them to express their ideas and shows deep respect for their thoughts. He teaches students according to their individual characteristics. Additionally, he provided me with an excellent platform for studying and discussing scientific matters.

Many thanks to my Prof. in Belgium, Ludivine Lassois, who selected me into her Ph.D. program, giving me the opportunity to study abroad and explore life outside of China. I still deeply remember my first day in Belgium. When I arrived and couldn't find my office, she kindly came and guided me step by step—not only telling me how to adapt to life in a foreign country but also providing all the necessary information about our lab. She introduced me to everyone in the lab and invited me to team-building activities. All of these efforts helped me integrate into this academic family quickly and made me feel at home, even while being far away from my homeland.

Thanks to my thesis jury for their helpful suggestions to improve my thesis: Gengler Nicolas; Martine Schroyen; Massart Sébastien; MUHOVSKI Yordan; Adrien Blum and Qian Zhou (周倩).

I would like to express my sincere gratitude to Mingjun Zhang (张明军). Without his invaluable support during my enrollment process, I would not have been able to complete my Ph.D. studies.

Thanks to this excellent lab (Solab; Solanaceae Laboratory), I have had the privilege of meeting many mentors who have taught me a great deal. These include Thomas Städler at ETH Zurich, Yongfeng Zhou (周永锋) and Suhua Li (李素华) at CAAS, Nils Stein at IPK, Chunzhi Zhang (张春芝) at CAAS, Yaoyao Wu (吴瑶瑶) at Nanjing Agricultural University, and Erik Garrison at the University of Tennessee Health Science Center, and William J. Lucas at the University of California.

I am also grateful to my colleagues, who share the same enthusiasm for science and with whom I have had the pleasure of working. Together, we strive for scientific progress. Special thanks to Dié Tang (唐蝶), Zhiyang Zhang (张智洋), Nan Wang

(王楠), Zhigui Bao (鲍志贵), Zhaonian Dong (董兆年), Hongbo Li (李宏博), Pei Wang (王培), Yi Zheng (郑毅), Pingxian Zhang (张平贤), Lian Qun (练群), Yuting Yang (杨雨婷), Zhaoxu Ma (马召旭), and Yong Hu (胡勇).

Additionally, I would like to express my gratitude to the friends I met in Belgium, Maria Osipenko and Elise Lhoste. They are interesting girls, and we usually work and eat together. Blum Adrien, taught me many things, including culture and French. I still remember that my very first French word, "Bonjour," came from this amazing language teacher. BIN Jimmy and Quievreux Martin were always warm-hearted and ready to help with everything. HIM Sok Lay amazed me with her incredible cooking skills, and De Clerck Caroline, Delaplace Pierre, LAGERQUIST Elsa, Dhommée Roxane, Jean Paul, ROJAS VARGAS Esther, BULAMBO Pacifique, BERGENHUIZEN Lucas, Serkan, Cecere Andrea, Guidosse Quentin, BIVER Simon, and many others whose names I may not know, also contributed to making my time in Belgium truly unforgettable.

Thanks to my house owner, Lam, who is a migrant from Vietnam. She is very kind to me, and we worked together to paint her house. It was a really interesting experience. When she found out that my kitchen was not working, she bought another one for me. I was deeply touched by her kindness.

Thanks to my Chinese friends whom I met in Belgium, we have helped each other to make our lives more colorful. Special thanks to Zhen Wang (王真), Tao Sun (孙涛), Menglu Wang (王梦露), Xiaofang Ji (纪小芳), Zixuan Han (韩紫璇), Jin Xie (谢晋), Simin Fan (范思敏), Ling Li (李玲), Fangzhou Wang (王方舟), Xiaoxian Liu (刘笑先), Ning Ding (丁宁), Han Wang (王瀚), Zhiguo Guo (郭治国), Kai Wang (王凯), Yue Zhang (张越) and Hui Wu (吴辉).

Finally, I would like to extend my thanks to my parents and family members for their unwavering support and companionship throughout my journey. I wish I can make them proud in the near future. Additionally, it is a great honor to be a student at the Chinese Academy of Agricultural Sciences and the University of Liège.

Thank you all !!!

Lin Cheng
30/06/2025 in Gembloux, Belgium

Content

ABSTRACT	I
RÉSUMÉ	III
ACKNOWLEDGMENTS	V
LIST OF FIGURES.....	IX
LIST OF TABLES.....	XI
LIST OF ABBREVIATIONS	XIII
CHAPTER I GENERAL INTRODUCTION	1
1. POTATO - GENERAL INTRODUCTION	3
2. HISTORICAL GENETIC IMPROVEMENT OF POTATO	4
3. BEYOND CONVENTIONAL BREEDING: INNOVATIONS DRIVING THE FUTURE OF POTATO IMPROVEMENT	5
4. INTRODUCTION TO THE GENOME.....	6
5. INTRODUCTION TO THE REFERENCE GENOME	7
6. INTRODUCTION TO HAPLOTYPE-RESOLVED/PHASED GENOMES	8
7. DEVELOPMENT OF SEQUENCING TECHNOLOGY	9
8. INTRODUCTION TO PLANT GENOME ASSEMBLY	9
8.1 <i>Simple plant genome assembly</i>	10
8.2 <i>Highly heterozygous genome assembly</i>	10
8.3 <i>Highly repetitive genome assembly</i>	10
8.4 <i>High ploidy genome assembly</i>	11
9. POTATO GENOME ASSEMBLY	12
9.1 <i>Clonal propagation shape genome architectures</i>	12
9.2 <i>Progress in potato genome assembly</i>	12
9.3 <i>Difficulty in potato genome assembly</i>	12
10. INTRODUCTION TO PANGENOMES AND PANGENOME GRAPHS	13
10.1 <i>General definition of pangenomes</i>	13
10.2 <i>Progress in pangenome research</i>	13
10.3 <i>Methods for pangenomes construction</i>	14
10.4 <i>The application of pangenomes in crop research</i>	15
11. OUR CURRENT PROGRESS FOR POTATO BREEDING.....	17
12. HYBRID POTATO BREEDING REQUIRE A PHASED PANGENOME TO PURGE DELETERIOUS VARIATIONS.....	18
12.1 <i>Phased pangenomes for heterozygous potato</i>	18
12.2 <i>Identification of deleterious variations</i>	19
13. CONCLUSIONS.....	20
14. REFERENCES	21
CHAPTER II THESIS OBJECTIVES	25
1. <i>Objectives</i>	27

2. RESEARCH ROADMAP AND OUTLINE	27
2.1. Research roadmap	27
2.2. Outline	28
CHAPTER III LEVERAGING A PHASED PANGENOME FOR HAPLOTYPE DESIGN OF HYBRID POTATO	31
1. INTRODUCTION	36
2. RESULTS	37
2.1 The potato pangenome graph (PPG-v1.0)	37
2.2 TEs drive structural variation (SV) formation	48
2.3 Domestication boosts heterozygosity	51
2.4 Fate of deleterious SVs (dSVs) under breeding.....	57
2.5 The 'broken-window' effect of dSVs.....	64
2.6 Genome design of ideal haplotypes.....	66
3. DISCUSSION	70
4. MATERIALS AND METHODS	71
4.1 Sequencing Materials.....	71
4.2 De novo genome assembly of 60 haplotypes	71
4.3 Evaluation of the genome assemblies.....	72
4.4 Annotation of repetitive elements	72
4.5 Prediction of protein-coding genes	72
4.6 Construction of the potato pangenome	73
4.7 Pangenome size and growth.....	73
4.8 Phylogeny and synteny analysis	73
4.9 Identification of Single Nucleotide Polymorphisms (SNPs), InDels and SVs.....	74
4.10 Haplotype diversity and genome heterozygosity by sequence (GHSL) statistic ...	74
4.11 SV dynamics in potato domestication.....	74
4.12 Identification of TEMRs	75
4.13 Haplotype recombination in inbred lines	75
4.14 Identification of dSVs and dSNPs.....	75
4.15 Estimation of the 'broken window' effect.....	76
4.16 The definition and construction of ideal potato haplotypes (IPHs).	76
4.17 Design an ideal potato chromosome and aggregating 12 ideal chromosomes.	77
5. REFERENCES	79
CHAPTER IV GENERAL CONCLUSION AND PERSPECTIVES.....	85
1. GENERAL CONCLUSION	87
1.1 Phased pangenomes capture comprehensive genomic diversity.....	87
1.2 TEs are major drivers of genomic diversity	88
1.3 Deleterious variants and the limits of conventional breeding.....	89
1.4 Precision breeding in highly heterozygous crops.....	90
2. GENERAL FUTURE PERSPECTIVES	91
3. REFERENCES	93
APPENDIX-PUBLICATIONS	96

List of Figures

Figure 1: Architecture of potato plant.....	3
Figure 2: Example of genome sequence data.....	7
Figure 3: Example of reference genome and reads mapping to two different genome regions (colored).	7
Figure 4: Example for phased diploid assembly and chimeric assembly.	9
Figure 5: Strategies for pan-genome construction.	14
Figure 6: Deleterious mutations result in severe inbreeding depression in inbred lines.	17
Figure 7: Research roadmap for the thesis.	28
Figure 8: Phased potato pangenome reference (PPG-v1.0).	38
Figure 9: Phylogenetic reconstruction of diverse potato accessions.....	39
Figure 10: The location of sequenced potato accessions.	40
Figure 11: The (High-throughput Chromosome Conformation Capture) Hi-C map for final genome assembly.	45
Figure 12: Assembly strategy and validation of the potato pangenome.	46
Figure 13: Visualization of the potato chromosome 10 pangenome graph.	48
Figure 14: TE activity drives structural variants in potato haplotypes.	50
Figure 15: Estimation of heterozygosity by SNP-based, K-mer-based, and Genome heterozygosity by sequence length (GHSL)-based methods.....	52
Figure 16: Potato domestication enhanced haplotype divergence.	53
Figure 17: Different chromosomes are distinguished by alternating light gray and dark gray.....	55
Figure 18: An example of a heterozygous inversion in the PG6244 genome assembly.	56
Figure 19: Pairwise Jaccard similarity analysis of 60 potato haplotypes.....	57
Figure 20: The characteristics of structural variations (SVs), deleterious SNPs (dSNPs) and deleterious SVs (dSVs) in the potato genome.	58
Figure 21: Fate of deleterious SVs (dSVs) in domestication and inbred-line development.	60
Figure 22: A 51 kb deleterious deletion eliminated three genes in potato cultivar PG6090 genome.	62
Figure 23: The recombination map of two inbred lines.....	63
Figure 24: Statistical significance of deleterious SNP (dSNP) enrichment in the deleterious SV (dSV) coupling phase.	65
Figure 25: The schematic diagram of ideal potato haplotypes (IPHs) for inbred lines.	67
Figure 26: Ideal potato haplotypes (IPHs) designed from the phased potato pangenome graph (PPG).	69

List of Tables

Table 1: Information of accessions in this study.....	41
Table 2: Statistics of raw HiFi data of 31 potato accessions.....	42
Table 3: Statistics of raw HiC data of 31 potato accessions.	43
Table 4: The proportion of TE-mediated rearrangement	51
Table 5: The distribution of heterozygous SVs in the potato genome	59
Table 6: The distribution of dSVs in the potato genome.	61

List of Abbreviations

AGI	Arabidopsis Genome Initiative
BAC	Bacterial artificial chromosome
BP	Base paired
BUSCO	Benchmarking Universal Single-Copy Orthologs
CND	<i>Solanum candolleanum</i>
DELs	Deletions
DNA	Deoxyribonucleic Acid
dSNPs	Deleterious single nucleotide polymorphisms
dSVs	Deleterious structural variants
DUPs	Duplications
DM	Double monoploid
Dis-coupling	Distance to coupling phase
Dis-repulsion	Distance to repulsion phase
FAO	Food and Agriculture Organization of the United Nations
Hi-C	High-throughput chromosome conformation
HiFi	High-fidelity
HGP	Human Genome Project
H1	Haplotype 1
H2	Haplotype 2
InDels	Insertion and deletion
INSs	Insertions
INVs	Inversions
IPHs	Idea potato haplotypes
IPHs-A	Ideal potato haplotypes of inbred line A
IPHs-E	Ideal potato haplotypes of inbred line E
GB	Gibabase
GHSL	Genome heterozygosity by sequence length
GERP	Genomic evolutionary rate profiling
GO	Gene Ontology
GON	<i>Solanum tuberosum</i> Group Goniocalyx
GS	Genome selection
KB	Kilobase
LTRs	Long terminal repeats
MB	Megabase
NLR	Leucine-rich repeat
nSNPs	Non-synonymous SNPs
ONT	Oxford Nanopore Technologies
PCA	Principal component analysis
PHU	<i>Solanum tuberosum</i> Group Phureja
PGGB	Pangenome Graph Builder
PPG	Potato pangenome graph
SDs	Segmental duplications

SFS	Site frequency spectrum
SNP	Single nucleotide polymorphism
sSNPs	Synonymous SNPs
STN	<i>Solanum tuberosum</i> Group Stenotomum
SVs	Structural variations
TB	Terabase
TEs	Transposable elements
TEMRs	Transposable elements-mediated rearrangements
TIP	TE insertions polymorphisms
TRs	Tandem repeats
VCF	Variants calling format
MC	Minigraph-Cactus

Chapter I General Introduction

1. Potato - General introduction

The cultivated potato (*Solanum tuberosum* L.) is a tuber-based, clonally propagated crop with four sets of chromosomes (tetraploid) that has shaped civilizations and transformed global food systems (Stokstad et al., 2019 & Spooner et al., 2014). It was domesticated over 8,000 years ago in the Andean highlands of South America. Often referred to as a “hidden treasure beneath the soil,” the potato offers a potent combination of high caloric density, rich nutritional value, and adaptability, making it a vital component of global food security (**Figure 1**).

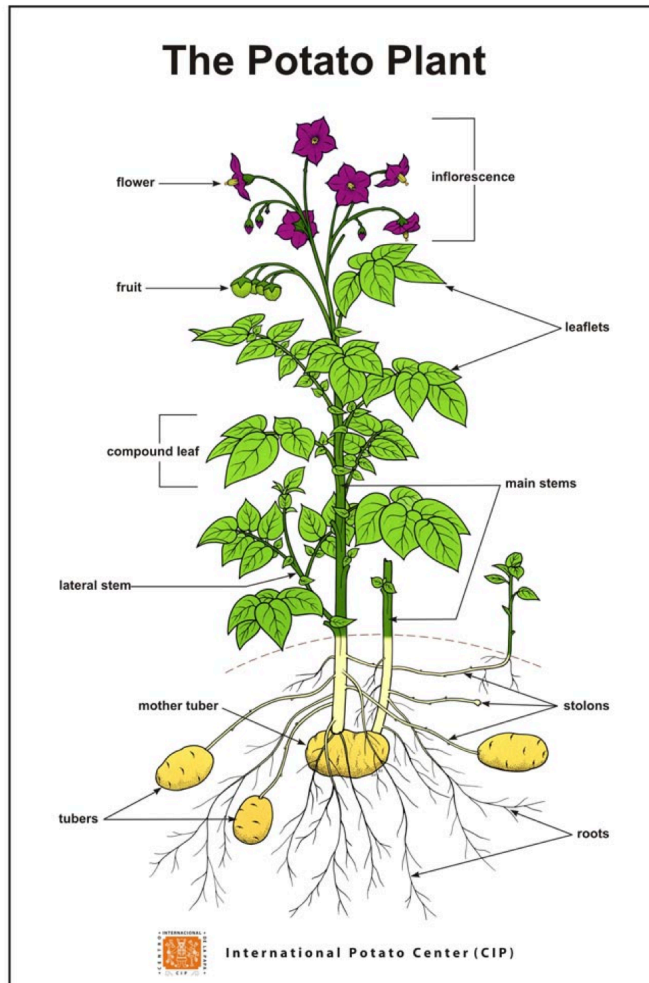


Figure 1: Architecture of potato plant.

This diagram illustrates the botanical structure of the potato (*Solanum tuberosum*), a member of the Solanaceae family (picture from <https://cipotato.org/>).

Potatoes have both above-ground and below-ground parts. The above-ground part consists of green stems and leaves that grow from the main plant. The leaves are broad and help the plant absorb sunlight for photosynthesis. The stems are strong and support the plant's structure.

Below the ground, the potato plant forms tubers, which are thick, fleshy storage organs. These tubers grow at the tips of specialized underground stems called stolons. The tubers are usually round or oval with a rough skin, which can be brown, red, or purple. Inside, the flesh is typically white, yellow, or purple, depending on the variety. The tubers store nutrients, primarily starch, which helps the plant survive through the winter and can be harvested for consumption.

According to the FAO (Food and Agriculture Organization of the United Nations; <https://www.fao.org>) annual global crop production data (2022), potatoes rank 6th with 369,753,096.64 tons, following sugarcane, maize (corn), rice, wheat, and oil palm fruit. Potatoes are the most widely grown non-cereal crop globally. However, prolonged domestication has significantly reduced the genetic diversity of cultivated varieties (Tang et al., 2022), limiting their adaptability to climate change—an issue projected to reduce potato yields by the end of the century (Raymundo et al., 2018). The extent of these impacts and their associated uncertainties will vary across regions. To address this challenge, exploring and utilizing genetic diversity is essential for the genetic improvement of potatoes.

2. Historical genetic improvement of potato

Tetraploid potatoes, which possess four sets of chromosomes, represent the predominant type cultivated worldwide. Breeding these potatoes presents unique challenges due to their genetic complexity. Traits in tetraploid populations often segregate over extended breeding periods, impeding the rapid fixation of desirable characteristics. For instance, skin color remains variable in tetraploid populations even after centuries of breeding (Van et al., 1994), thereby hindering progress in the selection of other important traits.

The formal history of potato breeding began in the early 19th century, with the first documented hybridization between cultivars reported by Knight in 1807 (Knight et al., 1807). In the following decades, farmers, hobbyists, and seedsmen contributed significantly to cultivar development. Notably, Chilean *tuberosum* landraces were introduced into breeding programs—such as ‘Daber’ in Germany (1830) (Plaisted et al., 1989) and ‘Rough Purple Chili’ in the USA (1851) (Goodrich et al., 1863)—and played a foundational role in shaping modern cultivars. The famous ‘Russet Burbank’ (Ortiz et al., 2001), North America’s most widely grown cultivar, was derived from ‘Rough Purple Chili’ through open pollination and selection, released in 1914.

As potato cultivation expanded globally, formal breeding initiatives were launched in countries like China and India in the 1930s, both of which are now among the world’s leading potato producers (Gaur et al., 2000 and Jin et al., 2004). Despite substantial progress—with over 4000 cultivars documented globally by 2005—the genetic base of cultivated *S. tuberosum* remains narrow, stemming from its historical derivation from a limited gene pool outside Latin America.

Conventional breeding focused primarily on phenotypic selection within *S. tuberosum*, often intercrossing cultivars with complementary traits. However, yield improvement through such intra-tuberosum crosses proved limited. The narrow genetic base and emerging challenges like disease mutations and climate stress highlighted the need for broader germplasm sources and more diverse breeding strategies.

Recent advancements have focused on decoding the tetraploid potato genome, aiming to address these challenges to enhance breeding efficiency and cultivar development (Sun et al., 2022 and Bao et al., 2022). However, the complex inheritance patterns associated with autotetraploidy—such as polysomic segregation and high heterozygosity—make genetic analysis and trait selection challenging. In addition, clonal propagation leads to slow breeding cycles, limited genetic gain, and increased vulnerability to pathogens. Despite these constraints, tetraploid breeding has successfully incorporated resistance traits from wild *Solanum* species and continues to serve as the foundation of global potato production. Nevertheless, to meet the growing demands for sustainable agriculture and climate resilience, innovative breeding approaches are needed to overcome the intrinsic limitations of the tetraploid system.

3. Beyond conventional breeding: innovations driving the future of potato improvement

Conventional potato breeding has predominantly relied on the tetraploid *Solanum tuberosum*, but the complex genetics of tetraploids—such as polysomic inheritance, linkage drag, and the masking of deleterious alleles—significantly limit the efficiency of genetic improvement. As a result, breeding strategies are increasingly shifting towards diploid potato systems, where undesirable alleles are more likely to be expressed in homozygous form and can be structurally purged from breeding populations (Jansky et al., 2016).

While the idea of hybrid potato breeding dates back over 50 years (Hawkes et al., 1956), progress was historically hindered by self-incompatibility and severe inbreeding depression in diploid germplasm (Ye et al., 2018). Selfing often results in low-yielding, poor-quality inbred lines due to the exposure of deleterious recessive mutations (Phumichai et al., 2005). A breakthrough came with the discovery of the *Sli* gene, originating from *Solanum chacoense*, which confers self-compatibility in diploid potatoes and enables the development of inbred lines (Hosaka et al., 1998).

Through successive selfing and selection, breeders have been able to eliminate large-effect deleterious mutations. Although many small- to moderate-effect deleterious alleles remain, their negative impacts are largely masked in F1 hybrids, which exhibit strong heterosis and vigorous growth (Zhang et al., 2019). This supports the foundation of a new F1 hybrid potato breeding system.

This emerging system has the potential to revolutionize potato breeding, effectively creating what can be considered a “new crop.” Desired traits can be rapidly enriched through backcrossing, preserving known agronomic or processing qualities. Furthermore, trait stacking—the combination of multiple desirable traits such as

disease resistances—becomes feasible within just a few years. Although commercial tetraploids currently outperform diploids in yield and adaptation, this is likely due to the much greater historical investment in tetraploid breeding. With focused breeding efforts and successful purging of harmful alleles, diploid potatoes may ultimately surpass their tetraploid counterparts (Zhang et al., 2021).

Notably, the world's three most important staple crops—rice, wheat, and maize—are all cultivated as diploids. As we continue to understand the genetic and physiological impacts of ploidy, diploid potato breeding offers a promising pathway for more efficient, targeted, and sustainable potato improvement.

4. Introduction to the genome

Since Gregor Mendel uncovered the basic laws of inheritance through his experiments with pea plants in the mid-19th century (Mendel et al., 1866), our understanding of genetics has steadily advanced. A major breakthrough came in 1953, when James Watson and Francis Crick discovered the double-helix structure of DNA, revealing the molecular basis of heredity. From that point onward, the once-enigmatic nature of genes began to be gradually unveiled. Over the decades, continued research in molecular biology, genomics, and biotechnology has transformed genetics from a theoretical concept into a powerful tool for understanding life and improving agriculture, medicine, and other areas of human advancement.

Genes are arranged along chromosomes, and the complete set of chromosomes makes up the genetic material that determines the traits of a living organism. This genetic material, represented by the sequence of A (adenine), T (thymine), C (cytosine), and G (guanine), is the fundamental resource for crop breeding (Kersey et al., 2019). Therefore, genome information provides powerful tools to improve the efficiency and precision of developing new plant or animal varieties with desirable traits, such as higher yield, disease resistance, drought tolerance, or specific physical and biochemical characteristics (Venter et al., 2001).

Genomes (**Figure 2-2**) are crucial for deciphering the evolutionary history of organisms, a lineage dating back nearly a billion years and encompassing almost half a million living species (Avise et al., 1997). This history, known as phylogeny, serves as both a record of past life and a powerful tool for advancing basic and applied plant sciences. Integrating phylogenetic and genomic studies can illuminate the mechanisms behind species emergence and extinction. Additionally, phylogenies can

direct efforts to enhance crop plants, identify novel medicines, and design effective conservation strategies.

```
>ADK8
GGGGAAGGGACTCAGGCTCAGTTCATCATGGAGAAATATGGTATTCGCAAATCTCCAC
TGGCGATATGCTGCGTGCTGCGGTCAAATCTGGCTCCGAGCTGGGTAAACAAGCAAAAGA
CATTATGGATGCTGGCAAACCTGGTCACCGACGAACTGGTGATCGCGCTGGTTAAAGAGCG
CATTGCTCAGGAAGACTGCCGTAATGGTTTCCTGTTGGACGGCTTCCCGCGTACCATTCC
GCAGGCAGACGCGATGAAAGAAGCGGGCATCAATGTTGATTACGTTCTGGAATTTGATGT
ACCGGACGAACTGATCGTTGACCGTATCGTGGGTGCGCGTGTTACGCGCCGCTGCGCCG
TGTTTATCACGTTAAATTCAATCCGCCGAAAGTAGAAGGTAAAGACGACGTTACCGGTGA
AGAGCTGACCACCCGTAAAGATGATCAGGAAGAAACCGTACGTAAACGTCTGGTTGAATA
```

Figure 2: Example of genome sequence data.

Genome sequence data consists of long strings of the nucleotide bases A (adenine), T (thymine), C (cytosine), and G (guanine), which together carry the genetic information of an organism. For example, a sequence like "ATCG" represents a small fragment of DNA. This data is essential for identifying genes, understanding genetic variation, and studying how traits are inherited. In crops like potato, genome sequencing helps breeders uncover important genes related to yield, quality, and disease resistance, supporting more targeted and efficient breeding strategies.

5. Introduction to the reference genome

A reference genome is a standard DNA sequence that represents the typical genetic makeup of a species. It serves as a baseline for comparing individuals, identifying genetic variations, and studying gene functions. Constructed from the DNA of one or a few individuals, a reference genome acts as a "map" of the species' genome, highlighting key genes, regulatory regions, and structural elements (**Figure 3**).



Figure 3: Example of reference genome and reads mapping to two different genome regions (colored).

The first human genome was released in 2001 by the Human Genome Project (HGP), which was launched in 1990 as an international research effort to sequence and map all the DNA in the human genome, consisting of approximately three billion base pairs (Venter et al., 2001). In 2000, a draft sequence was announced, marking a milestone in biological science. The project revealed that humans have about 20,000 to 25,000

genes, fewer than previously expected. The completed human genome sequence has provided invaluable insights into genetics, disease mechanisms, and human evolution, paving the way for advances in personalized medicine, genetic diagnosis, and biotechnology. The initial findings were published in landmark papers in 2001 in *Nature* and *Science*, highlighting the collaborative achievement that transformed biological research worldwide.

The first plant genome, *Arabidopsis thaliana*, was released in 2000 by the Arabidopsis Genome Initiative (AGI) (The Arabidopsis Genome Initiative, 2000), a collaborative international effort to sequence and map the entire genome of this model plant species. The genome consists of approximately 125 million base pairs and about 25,000 genes. The completion of this genome marked the first time a plant genome was fully sequenced, providing a critical resource for plant biology and genetics. This milestone advanced our understanding of plant development, gene function, and evolution, and laid the foundation for agricultural and biotechnological research. The results were published in a landmark paper in *Nature* in 2000, highlighting the global collaboration that significantly impacted plant science worldwide.

The first potato reference genome was constructed by The Potato Genome Sequencing Consortium in 2011 (The Potato Genome Sequencing Consortium, 2011). This genome was sequenced and assembled using a homozygous doubled-monoploid potato clone, covering 86% of the 844-megabase genome. A total of 39,031 protein-coding genes were predicted, and evidence was presented showing that potentially deleterious mutations occur frequently and are likely contributors to inbreeding depression. The potato genome sequence provides a valuable foundation for the genetic improvement of this essential crop. Since then, the genetic analysis and research of potato has increased dramatically, particularly in areas such as molecular breeding, genome editing, and genomic selection. (Felcher et al., 2012; Das et al., 2019 and Zhang et al., 2021). While it remains a powerful tool for research in agriculture and evolutionary, it does not capture the full genetic diversity of a species.

6. Introduction to haplotype-resolved/phased genomes

Recent advances in sequencing technologies and computational methods have enabled genome assemblies to be generated at the haplotype level in both animals and plants. Haplotype-resolved, or allele-aware, assembly focuses on accurately separating allelic variations to reconstruct haplotype sequences (**Figure 4a**), addressing limitations of consensus genome assemblies. In plants, this approach has been particularly applied to complex genomes, such as autopolyploid or highly heterozygous diploid species. For instance, allele-defined genome sequences have uncovered chromosomal inversions in autohexaploid sugarcane (Healey et al., 2024). Additionally, precise characterization of allelic variation in the alfalfa genome has facilitated genome-editing applications, highlighting the potential of haplotype-resolved data for functional analysis and breeding research (Shen et al., 2020).

A haplotype-resolved genome is a comprehensive representation of two sets of genetic information, inherited from each parent of a diploid organism. Instead of blending the genetic data into a single consensus sequence (**Figure 4b**), this approach distinguishes and organizes the genome into two distinct parental haplotypes. (This

process will be discussed in more detail in Chapter 3, in the Methods section on phased potato genome assembly).

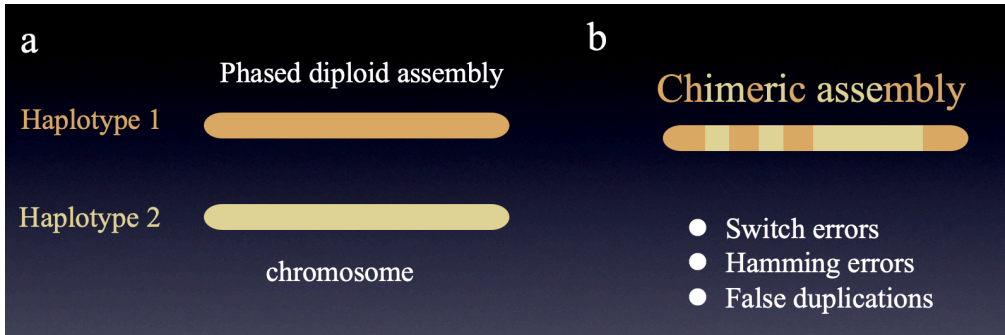


Figure 4: Example for phased diploid assembly and chimeric assembly.

(a) Haplotype-resolved (phased) diploid assembly and (b) chimeric assembly, which merges two haplotypes into a single consensus genome. Chimeric assemblies often introduce various types of errors, including switch errors, Hamming errors, and false duplications (Liao et al., 2023).

7. Development of sequencing technology

The development of sequencing technologies has transformed genomics research dramatically. Starting with Illumina sequencing, which provides short but highly accurate reads at an unprecedented throughput (Quail et al., 2008), researchers were able to decode genomes quickly and cost-effectively. However, its limitations in read length made it challenging to resolve complex regions and structural variations. To address this, long-read technologies such as PacBio (Rhoads et al., 2015) and Oxford Nanopore (Jain et al., 2016) emerged, enabling the sequencing of much longer fragments, though initially at the cost of lower accuracy. Recent innovations, such as PacBio High-fidelity sequencing (HiFi; Wenger et al., 2019 and Hu et al., 2021) and Duplex sequencing (Kennedy et al., 2014), combine the advantages of long reads with ultra-high accuracy by independently sequencing both strands of a DNA molecule and comparing them to eliminate errors. This evolution from short-read Illumina platforms to high-fidelity long-read Duplex technologies marks a significant leap forward in accuracy, resolution, and the ability to detect rare variants and complex genomic features.

8. Introduction to plant genome assembly

The purpose of genome assembly is to correctly assemble a large number of short/long DNA sequence fragments (reads) obtained by sequencing into a complete genome sequence so that we can fully understand the genetic information of an organism.

The key to the assembly process is to identify the overlapping relationship between adjacent sequenced fragments on the genome. In addition to sequencing technology,

the heterozygosity and repetitive sequences of the genome have the greatest impact on the assembly effect (Sun et al., 2022). Plant genomes often undergo local duplication, whole genome doubling, and repetitive sequence expansion, resulting in a large number of similar or homologous fragments in the genome, which generates redundant overlapping relationships during assembly and increases the difficulty of assembly. Since plant genomes are very diverse, assembling new species based on a few published species remains a significant challenge.

8.1 Simple plant genome assembly

Genomes with a size of no more than 1Gb, a homozygous or heterozygous degree of less than 0.5%, and a repetitive sequence content of less than 50% can be classified as simple genomes. *Arabidopsis thaliana* which was the first model species to complete their genome maps (The Arabidopsis Genome Initiative, 2000), are simple genomes. Simple genomes can be assembled using second-generation sequencing data, a mixture of second- and third-generation sequencing data, or pure third-generation sequencing data.

8.2 Highly heterozygous genome assembly

Self-incompatibility and asexual reproduction are common in plants in nature, resulting in heterozygous characteristics of the genome. The highly heterozygous genome has a heterozygosity above 1%, that is, the sequence difference of homologous fragments reaches 1%-2%, resulting in the inability to fully merge the reads of the homologous regions during assembly, resulting in a large number of branch structures, which seriously affects the continuity of the assembly and subsequent analysis.

The assembly of highly heterozygous genomes has always been a difficult problem in the field of genomic assembly, and there are still no relatively simple methods and tools. With the application of HiFi read data and the development of assembly software such as hifiasm (Cheng et al., 2021) that comprehensively utilizes the advantages of third-generation sequencing and Hi-C data, the rapid progress of heterozygous plant genome analysis will be promoted.

8.3 Highly repetitive genome assembly

Repetitive sequences play an indispensable role in species evolution and functional regulation and are an important part of the genome. The high sequence similarity, different lengths, and wide range of copy number variations of repetitive sequences have always been a difficult problem in assembly. Compared with the second-generation sequencing technology, the third-generation long-read sequencing can cross the repetitive sequence region, improve the discrimination of repetitive sequences, and significantly improve the continuity of assembly and the integrity and accuracy of repetitive sequence assembly. This advantage is fully reflected in the maize genome, in which 85% of the sequences are derived from transposon amplification. The maize B73 (Liu et al., 2020) genome assembled with PacBio data has an improvement in contig continuity compared to the previous version based on the second-generation assembly, and corrects the assembly errors in the centromere

region, greatly improving the annotation of gene functional regions and the evolutionary analysis of transposons.

8.4 High ploidy genome assembly

Since hybridization and genome doubling lead to the existence of polyploid plants, some important crops such as wheat, cotton, and potatoes are polyploid. The analysis of their genomes is an important factor affecting the progress of crop breeding. Polyploid species are divided into allopolyploids and autopolyploids according to their formation mechanism (Comai et al., 2005). In allopolyploids, chromosomes originate from different ancestors, and subgenomes can be distinguished within the genome, which has less interference with assembly; while in autopolyploids, multiple sets of chromosomes are highly similar, which is equivalent to a highly heterozygous genome, and assembly is extremely difficult.

- Allopolyploid genome assembly

Allopolyploids can usually be assembled as homozygous genomes, and the focus is on distinguishing subgenomes after assembly. When the International Wheat Sequencing Consortium analyzed the genome of hexaploid cultivated wheat (*Triticum aestivum*, AABBDD), flow cytometer separation technology was used to separate 21 chromosomes, and BAC libraries were constructed for sequencing and assembly (Brenchley et al., 2012). Compared with second-generation sequencing data, third-generation sequencing data can better distinguish similar sequences and assemble contigs with longer continuity, which can then be combined with whole-genome genetic maps or Hi-C maps to distinguish heterologous chromosomes. The tetraploid cotton TM-1 (*Gossypium hirsutum*, AADD) genome published in 2015 was assembled from 100,000 BAC clones and genetic maps (Li et al., 2015). The new versions of the TM-1 genome published in 2019 and 2020 were completed using PacBio data, Hi-C maps, and optical maps, which improved the quality of the reference genome and provided a more efficient and lower-cost polyploid assembly method.

- Autopolyploid genome assembly

Compared to allopolyploids produced by natural hybridization, autopolyploids are formed by chromosome doubling. Genetically, multiple sets of chromosomes can be associated, and the homologous regions in the sequence have a high degree of similarity, which interfere with each other during the assembly process. In the era dominated by second-generation sequencing data, in order to construct the reference genome of species, only haploid materials can be sequenced to reduce the difficulty of assembly or tolerate and merge heterozygous regions. The hexaploid sweet potato genome (*Ipomoea batatas*, B1B1B2B2B2B2; Yang et al., 2017) reported for the first time the haploid reference genome of an autopolyploid plant and the typing results of 30% of the genome region. Tetraploid alfalfa used the highly accurate PacBio HiFi data for polyploid assembly for the first time and obtained good contig continuity results. Although the assembly and typing of autopolyploids have been successful in many species, the typing software based on Hi-C still relies on the haploid reference genome, and the differentiation effect is not obvious when dealing with homologous chromosomes with small differences. The analysis of complex autopolyploid

genomes still requires the exploration of multiple types of data and technology integration.

9. Potato genome assembly

9.1 Clonal propagation shape genome architectures

Genome architecture is profoundly influenced by the mode of propagation, with sexual and asexual reproduction exerting distinct evolutionary pressures. Sexual reproduction, as seen in crops such as rice and maize, involves meiotic recombination and the fusion of gametes from two individuals, resulting in genetic reshuffling and increased diversity across generations. This process facilitates the elimination of deleterious mutations and promotes genome stability through natural selection. In contrast, asexual (clonal) reproduction, observed in species such as cassava and potato, bypasses recombination, transmitting identical genetic material from one generation to the next. While this ensures the preservation of advantageous genotypes, it can also lead to the accumulation of mutations, structural variations, and epigenetic changes over time. Consequently, the propagation strategy adopted by an organism plays a critical role in shaping its genome structure, variability, and long-term adaptability.

9.2 Progress in potato genome assembly

Although the first potato genome, derived from a double monoploid, was published in Nature in 2011 (The Potato Genome Sequencing Consortium, 2011), the first haplotype-resolved tetraploid genome was not assembled until 2022 through a collaborative effort involving multiple institutions and various advanced sequencing technologies, such as PacBio HiFi, ultra-long Oxford Nanopore reads, and Hi-C (Sun et al., 2022). The complexity of the tetraploid potato genome has posed major challenges to assembly and has hindered progress in genetic analysis. In contrast, over 70% of wild potato species are diploid, and they exhibit substantial genetic diversity, offering valuable resources for developing climate-resilient cultivars. Diploid genome assembly is significantly less complex than that of tetraploids, making it a more tractable model for genetic analysis. Therefore, simplifying the genomic landscape by focusing on diploid potato species can greatly facilitate genomic research and accelerate molecular breeding and crop improvement.

9.3 Difficulty in potato genome assembly

The assembly of the potato genome presents considerable challenges due to its complex genetic characteristics. Cultivated potato (*Solanum tuberosum*) is a highly heterozygous autotetraploid species ($2n = 4x = 48$), meaning it contains four copies of each chromosome with extensive allelic variation. This polyploidy complicates genome assembly because homologous sequences from different haplotypes are difficult to distinguish, often resulting in chimeric assemblies or collapsed regions. Furthermore, the genome is rich in repetitive elements, such as Transposable element (TEs), Tandem repeats (TR) and Segment duplications (SDs), those hinder accurate contig extension and scaffolding. High heterozygosity and repeat element also leads to difficulties in haplotype phasing and increases the likelihood of assembly errors, such as switch errors and misjoins. As a result, producing a high-quality, haplotype-

resolved assembly of the potato genome requires advanced sequencing technologies, such as long-read sequencing and HiFi reads, combined with sophisticated assembly algorithms capable of handling polyploidy and complex repeat structures.

10. Introduction to pangenomes and pangenome graphs

As a single genome or individual does not represent the entire population, broader sampling is necessary. The pangenome, first introduced by Tettelin et al., (2005) in a study of *Streptococcus agalactiae*, refers to the entire set of genes present within all individuals of a species. Compared to single reference genomes, pangenomes reduce reference bias and offer a more accurate and comprehensive representation of genomic diversity within a species. A pangenome graph, on the other hand, is a graphical data structure used to represent this genetic diversity in a more comprehensive way. Instead of a single linear reference genome, a pangenome graph encodes sequences as nodes and their connections as edges, allowing multiple alternative paths to represent different variants across individuals. This graph structure enables more accurate representation and analysis of complex genetic variation compared to a traditional linear genome (Eizenga et al., 2020).

10.1 General definition of pangenomes

The pangenome encompasses the representative set of DNA sequences within a population, offering a more accurate and diverse representation of global genomic variation while reducing reference bias. In the context of a pan-genome, genes are typically categorized into core and dispensable (also called variable or accessory) gene sets. Core genes are those that are present in all individuals of a species and are often essential for basic cellular functions, development, and viability. In contrast, dispensable genes are present in only a subset of individuals and may contribute to phenotypic diversity, environmental adaptation, or stress responses. The relative proportions of core and dispensable genes vary across species and are influenced by population diversity, genome size, and evolutionary history (Golicz et al., 2016). The identification and functional characterization of these gene sets provide critical insights into genome plasticity, adaptation, and the molecular basis of trait variation within a species. For instance, in polyploid pangenomes, 66% of genes were categorized as core in wheat and 42% in broomcorn millet (Jiao et al., 2024 and Walkowiak et al., 2020).

10.2 Progress in pangenome research

Since the publication of the first plant pan-genome in soybean (*Glycine max*) (Li et al., 2014), research on diploid plant species has expanded rapidly, encompassing key crops such as rice (*Oryza sativa*) (Zhao et al., 2018), and maize (*Zea mays*) (Hi et al., 2014). These studies have provided important insights into structural variation, gene presence/absence variation (PAV), and the distribution of core versus variable genes, which are often associated with traits like stress resistance, yield, and environmental adaptation.

Diploid pan-genomes are more tractable than those of polyploids due to the simpler genomic architecture and reduced haplotype complexity. Advanced approaches such

as graph-based genotyping have been widely used in diploid pan-genomes to discover structural variants, infer domestication events, and link genomic variation to phenotype. For instance, in broomcorn millet (*Panicum miliaceum*) (Chen et al., 2023), graph-based analyses have enabled the identification of variants related to domestication centers and key agronomic traits.

10.3 Methods for pangenomes construction

There are three primary strategies for constructing a pan-genome (Figure 5):

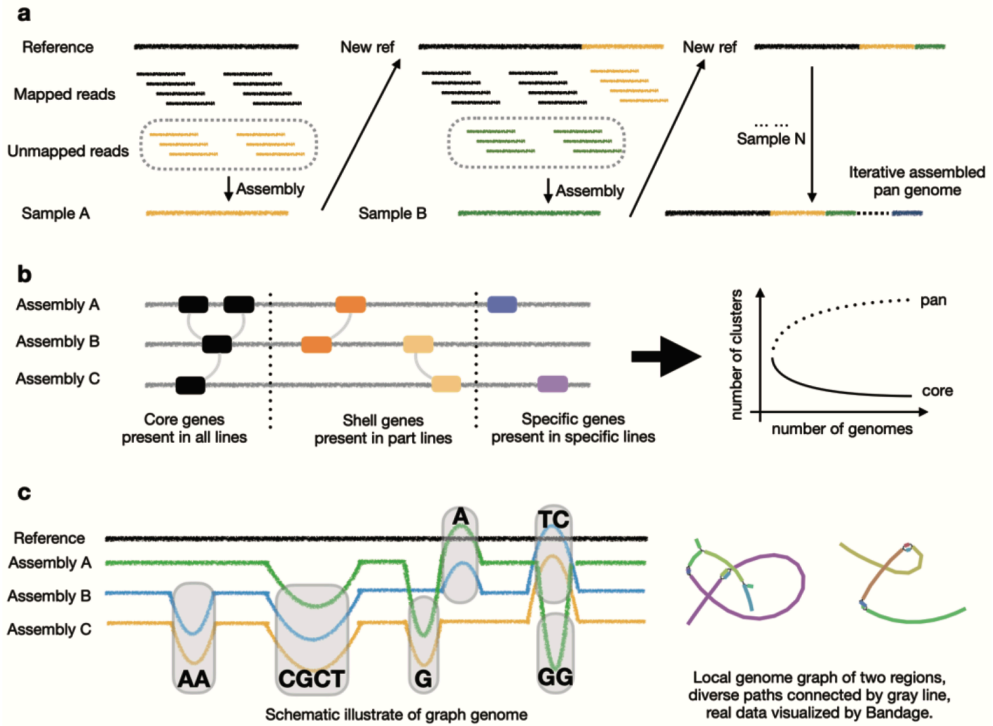


Figure 5: Strategies for pan-genome construction.

(a) Iterative assembly-based pan-genome. Sequencing reads from multiple individuals are aligned to a reference genome. Unmapped reads are assembled into novel contigs and iteratively incorporated into the reference genome, progressively expanding its sequence content. (b) De novo assembly-based pan-genome. High-quality genomes from multiple individuals are independently assembled and annotated. Genes are clustered into pan-gene families using sequence similarity, and classified into core and dispensable gene sets based on their frequency across all samples. Pan- and core-genome curves are modeled using non-linear regression. (c) Graph-based pan-genome. A genome graph is constructed by integrating genomic variants from multiple assemblies into a reference genome backbone. Alternative paths (gray boxes) represent sequences that differ from the reference. The right

panel illustrates the actual structures of two regions in the graph, revealing complex structural variation.

- Iterative Mapping and Assembly (Map-to-Pan Strategy; **Figure 5a**)

Early pan-genome studies, constrained by limited sequencing data, relied on mapping resequenced individuals to a single reference genome. Unmapped reads were extracted, assembled, and iteratively added to the reference to capture novel genomic content. This approach, known as iterative assembly, often results in fragmented pan-genomes and is less effective at detecting large structural variations. A representative example is the 2018 study of 3,010 rice (*Oryza sativa* L.) accessions, where researchers proposed the "map-to-pan" strategy and recovered 268 Mb of non-redundant novel sequences to supplement the Nipponbare reference genome (Wang et al., 2018). However, this method can introduce redundancy and lacks assembly contiguity.

- De Novo Assembly of Individual Genomes (**Figure 5b**)

A more accurate but resource-intensive method involves de novo assembly of high-quality genomes from multiple individuals, followed by pan-genome construction. This approach enables chromosome-level comparisons and precise detection of presence/absence variations (PAVs) and structural rearrangements. Examples include pan-genomes of cultivated and wild rice (Zhao et al., 2018), barley (*Hordeum vulgare* L.) (Jayakodi et al., 2022), and wheat (*Triticum aestivum* L.) (Walkowiak et al., 2020), which used second-generation sequencing data, with Hi-C scaffolding for chromosome-level resolution in some cases. The barley pan-genome, for instance, revealed two frequent chromosome translocations associated with domestication and breeding, demonstrating how high-quality assemblies enhance the resolution and applicability of pan-genome analyses in crop improvement.

- Graph-based Pan-genomes (**Figure 5c**)

Graph-based genome representations have rapidly emerged as a powerful strategy (Li et al., 2020). In this model, genomic sequences from different individuals are represented as paths through a genome graph, capturing both shared and divergent sequences. Graph pan-genomes are typically constructed by aligning multiple de novo assembled genomes to a linear reference, identifying variants, removing redundancy, and integrating all sequences into a unified graph. This approach preserves individual-specific sequences while maintaining coordinate consistency across genomes, improving the representation of structural variation and facilitating complex population-level analyses. A landmark example is the 2020 soybean (*Glycine max*) pan-genome, which constructed the first high-quality graph reference genome in crops (Liu et al., 2020). Using third-generation sequencing, researchers assembled 26 representative accessions with a contig N50 of 22.6 Mb, achieving unprecedented assembly continuity and setting a new standard for future pan-genomic studies

10.4 The application of pangenomes in crop research

The application of pangenomes in crop research is revolutionizing the identification and utilization of genetic variation for breeding, such as grape and rice (Schreiber et al., 2024, Liu et al., 2024 and Guo et al., 2025). Unlike reference genomes that

represent a single genotype, pangenomes integrate core genes, dispensable and private genes. This comprehensive view enables the discovery of novel variants, particularly structural variants (SVs) such as gene deletions, duplications, inversions, and translocations, which are often missed in single-reference analyses but can have profound effects on important agronomic traits.

One of the most significant uses of crop pangenomes is in genetic mapping and selection of trait-associated variants. Traditional mapping approaches focused largely on SNPs and small indels, but pangenomic analyses have revealed that SVs, such as gene deletions or copy number variants, are also tightly linked to key traits like flowering time and disease resistance (Schiessl et al., 2019). Furthermore, large inversions, as seen in maize (Fang et al., 2012), may underlie local adaptation but are difficult to detect or interpret without access to multiple high-quality genome assemblies.

Pangenomes also support genomic selection (GS) (Bayer et al., 2021), a technique that predicts phenotypes using genome-wide marker data. Including SVs derived from pangenomes can increase the accuracy of GS models, especially for complex quantitative traits controlled by many small-effect loci. Pangenome-informed models can better impute sparse genotyping data and improve phenotype prediction beyond what SNP-based models allow.

In the context of disease resistance, pangenomes have enabled the construction of "resistance gene atlases" by cataloging variation in gene families such as nucleotide-binding leucine-rich repeat (NLR) genes (Barragan et al., 2021). These genes frequently occur in highly variable clusters, and their arrangement and copy number can directly influence resistance efficacy. Targeted sequencing approaches and, increasingly, full-genome assemblies are used to survey these regions, and efforts are underway to define the full "NLRome" of major crops for durable resistance breeding.

Additionally, pangenomics facilitates the utilization of crop wild relatives through the construction of "super-pangenomes" — pangenomic frameworks extended across species or genera. These efforts uncover wild alleles for traits like stress tolerance, self-compatibility, and tuber development, which can be introgressed into elite lines. For example, pangenomic analysis of *Solanum* species has identified key resistance and yield-enhancing genes in tomato and potato, despite challenges from clonal propagation and polyploidy (Tang et al., 2022).

As sequencing technologies become more affordable and analytical tools more powerful, crop pangenomes are evolving into haplotype catalogues — comprehensive genomic resources that can guide the selection of parental lines, improve marker-assisted selection, and accelerate precision breeding strategies. In this light, pangenomics is not just a tool for understanding diversity but a cornerstone of next-generation crop improvement.

The development of phased pangenomes, supported by high-fidelity long-read sequencing and advanced genome assembly algorithms, has fundamentally reshaped the possibilities for potato breeding. These tools enable precise identification of structural and deleterious variants, deeper insights into heterozygous genomes, and more efficient trait selection. In species like potato—historically constrained by

polyploidy, clonal propagation, and limited genetic gain—such innovations offer a transformative path forward.

11. Our current progress for potato breeding

Building on these recent genomic breakthroughs, our team initiated a diploid potato breeding program aimed at developing inbred lines and true F1 hybrids. In 2021, we successfully cultivated the first seed-propagated potato hybrid (Zhang et al., 2021). However, due to long-term asexual propagation and the accumulation of deleterious mutations, our diploid inbred lines showed poor growth and harbored numerous harmful alleles, as confirmed by selfing and crossing experiments (Zhang et al., 2019).

These deleterious mutations contribute significantly to inbreeding depression, posing a major challenge to the development of vigorous inbred lines and the advancement of hybrid potato breeding. While many large-effect deleterious mutations have been removed during inbred line development, minor and moderate-effect mutations still substantially reduce plant fitness (**Figure 6**). In the field, inbred lines often grow poorly due to homozygous expression of these mutations, whereas hybrids grow vigorously as the harmful alleles are masked in the heterozygous state.

Eliminating remaining deleterious variants will be essential for the continued progress of potato breeding.

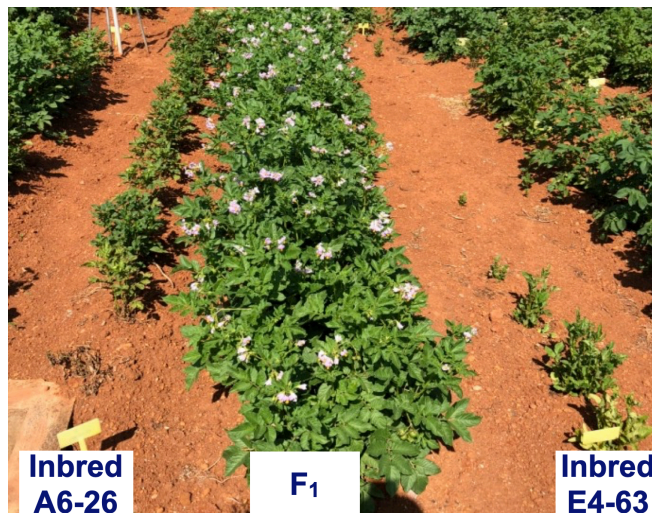


Figure 6: Deleterious mutations result in severe inbreeding depression in inbred lines.

From left to right: inbred line E4-64, F1 hybrid (inbred line E4-63 crossed with A6-26), and inbred line A6-26. The other lines are cultivated tetraploid potatoes.

12. Hybrid potato breeding require a phased pangenome to purge deleterious variations

12.1 Phased pangenomes for heterozygous potato

Potato is a species characterized by significant genetic diversity and high levels of heterozygosity (The Potato Genome Sequencing Consortium, 2011). To accurately represent this diversity and facilitate precision breeding, the use of phased (haplotype-resolved) pangenomes is crucial (Zhou et al., 2020). The following outlines the key reasons why phased pangenomes are necessary for highly heterozygous potato genomes:

- High level of Heterozygosity

The potato genome exhibits high levels of heterozygosity, meaning there is considerable genetic variation within individual plants. This high degree of heterozygosity complicates genome assembly, particularly in regions where distinguishing between parental alleles is challenging. Phased genomes help researchers separate these alleles, thereby more accurately capturing genetic variation (Sun et al., 2022 and Bao et al., 2022).

- Genetic Diversity in Breeding

As a predominantly outcrossing species, potatoes undergo frequent genetic exchange between individuals, resulting in a broad spectrum of genetic diversity (Bao et al., 2022). This diversity is essential for breeding programs aimed at developing varieties with desirable traits such as disease resistance, drought tolerance, or increased yield. By incorporating genome-wide data from multiple potato varieties, phased pangenomes allow for a more comprehensive understanding of the genetic diversity (Li et al., 2023) present across different potato cultivars.

- Identification of Structural Variants (SVs)

Heterozygous regions in the potato genome often harbor structural variants, such as insertions (INs), deletions (DELs), inversions (INVs) or duplications (DUPS). Phased pangenomes are essential for accurately mapping these structural variants (Li et al., 2023, Sun et al., 2022 and Bao et al., 2022), as they maintain the integrity of haplotypes (sets of alleles inherited together), thereby providing a clearer view of the genome's structure and facilitating a better understanding of its functional elements.

- Resolving Assembly Ambiguities

Without phasing, distinguishing between alleles derived from different parental genomes in heterozygous regions becomes difficult, which may lead to errors or incomplete genome assemblies (Sun et al., 2020 and Li et al., 2023). Phased genome assemblies ensure that both parental genomes are accurately represented, addressing ambiguities and enhancing the quality and completeness of the genome sequence.

- Applications in Precision Breeding

Phased pangenomes are instrumental for marker-assisted breeding, such as introgression segment from wild species (Sun et al., 2025). A comprehensive and accurate understanding of the genetic diversity within the potato genome, such as European potato, enables breeders to identify specific loci associated with important

agronomic traits and diversity, accelerating the development of new, improved potato varieties for agriculture.

12.2 Identification of deleterious variations

Phased pangenomes also contribute to the identification of deleterious variations (mutations that negatively affect an organism's fitness), which impeded potato breeding (Zhang et al., 2019, Zhou et al., 2020 and Wu et al., 2023). Here's a detailed explanation of how phased pangenomes influence the study of deleterious variations:

- Accurate Identification of Deleterious Variants

A phased pangenome represents both copies of an individual's genome (haplotypes), allowing for the distinction between alleles inherited from each parent. This enables more precise identification of deleterious variants (Zhou et al., 2020). When these variants are located in heterozygous regions, it is crucial to determine whether the mutation is present in one or both alleles. A phased genome can clarify whether a deleterious variant is carried in a single copy (heterozygous) or both copies (homozygous), which is important for understanding the potential impact on phenotype and fitness.

- Distinguishing Between Harmful and Neutral Variants

Phased pangenomes allow researchers to differentiate between variants that are potentially deleterious and those that are benign. When both copies of a gene carry a deleterious mutation, the organism may express a harmful phenotype, particularly in recessive conditions (Zhang et al., 2019). Conversely, a heterozygous state might not result in a phenotype but could still carry a deleterious variant that may contribute to reduced fitness under certain environmental conditions. Phasing helps resolve these subtleties by clearly associating each allele with its effect.

- Haplotypic Context of Deleterious Variants

Many deleterious variants exist in specific haplotypes or genomic regions that might be inherited together due to linkage disequilibrium (the non-random association of alleles at different loci, Zhou et al., 2020). Phased pangenomes allow for the resolution of these linked regions, providing insights into how specific deleterious mutations are inherited in combination with other genetic variations. This information is crucial for understanding genetic interactions, potential epistatic effects, and the risk of expressing deleterious traits.

- Impact on Genetic Counseling and Breeding Programs

In breeding programs, understanding the inheritance of deleterious mutations is essential for avoiding the introduction or propagation of harmful variants (Wu et al., 2023). Phased pangenomes enable breeders to track deleterious alleles across different cultivars and select against them more effectively. In genetic counseling, knowing the phase of a variant allows for more accurate predictions of whether an individual will express a harmful phenotype, particularly when both parents carry a heterozygous deleterious allele.

- Tracking Deleterious Variants in Diverse Populations

In highly heterozygous populations, such as those in outcrossing species like cassava (Esuma et al., 2017), deleterious mutations might be common and present in different individuals. A phased pangenome, which includes multiple genomes from diverse individuals, provides a broader view of the distribution of deleterious variants across populations. This is essential for understanding how these variants might affect the population as a whole and for evaluating the potential risks of genetic erosion or inbreeding depression.

13. Conclusions

As a major tuber crop, potato plays a vital role in global food security. However, most cultivated varieties are tetraploid and clonally propagated, resulting in genetic complexity that hampers breeding progress. Polyploidy complicates allele tracking and trait inheritance, while clonal propagation limits recombination, slows genetic gain, and increases susceptibility to biotic and abiotic stresses—contrasting with seed-propagated crops like rice, maize, and tomato (Zhang et al., 2021).

In recent years, diploid breeding has emerged as a promising alternative. Yet, the development of vigorous inbred lines remains challenging due to numerous deleterious alleles masked in heterozygous form (Zhou et al., 2020). Accurate mapping and characterization of these variants are essential for achieving successful diploid hybrid breeding. In this context, **phased pangenomes** offer a transformative framework. By resolving haplotypes inherited from each parent, they enable precise interpretation of genetic diversity, structural variation, and inheritance patterns—thereby informing rational breeding decisions.

To address these challenges, the diploid potato breeding initiative (Jansky et al., 2016; Zhang et al., 2019; Bethke et al., 2022) seeks to transition from clonally propagated tetraploids to true seed-propagated diploid hybrids. This shift promises faster breeding cycles, greater genetic control, and enhanced selection for key traits such as yield, disease resistance, and climate resilience.

Although the current monoploid reference genome (~800 Mb, 12 chromosomes; The Potato Genome Sequencing Consortium, 2011) has supported key discoveries, it fails to capture the full spectrum of genetic variation in cultivated and wild potato populations. Furthermore, the complex architecture of the potato genome continues to complicate gene discovery and trait mapping. Recent advances in diploid genomics, particularly those leveraging pangenomic frameworks (Tang et al., 2022), provide the foundation for overcoming these limitations.

Decoding the phased potato pangenome is thus essential—not only to accelerate breeding efficiency and reduce production costs, but also to align with the principles of sustainable agriculture. A deeper understanding of genomic diversity and the distribution of deleterious mutations will be key to guiding the next generation of breeding strategies and securing the future of potato cultivation under global change.

14. References

- The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
- Avise J C, Wollenberg K. Phylogenetics and the origin of species. *Proceedings of the National Academy of Sciences*, **94**(15): 7748-7755 (1997).
- Bao Z, Li C, Li G, et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular plant*, **15**(7): 1211-1226 (2022).
- Barragan A C, Weigel D. Plant NLR diversity: the known unknowns of pan-NLRomes. *The Plant Cell*, **33**(4): 814-831 (2021).
- Bayer P E, Petereit J, Danilevicz M F, et al. The application of pangenomics and machine learning in genomic selection in plants. *The Plant Genome*, **2021**, 14(3): e20112.
- Bethke, P.C., Halterman, D.A., Francis, D.M. et al. Diploid Potatoes as a Catalyst for Change in the Potato Industry. *Am. J. Potato Res.* **99**, 337–357 (2022).
- Brenchley R, Spannagl M, Pfeifer M, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**(7426): 705-710 (2012).
- Chen, J. et al. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nat. Genet.* **55**, 2243–2254 (2023).
- Cheng H, Concepcion G T, Feng X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, **18**(2): 170-175 (2021).
- Comai L. The advantages and disadvantages of being polyploid. *Nature reviews genetics*, **6**(11): 836-846 (2005).
- Das Dangol S, Barakate A, Stephens J, et al. Genome editing of potato using *CRISPR* technologies: current development and future prospective. *Plant Cell, Tissue and Organ Culture (PCTOC)*, **139**(2): 403-416 (2019).
- Eizenga, J. M. et al. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
- Fang Z, Pyhäjärvi T, Weber A L, et al. Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, **191**(3): 883-894 (2012).
- Felcher K J, Coombs J J, Massa A N, et al. Integration of two diploid potato linkage maps with the potato genome sequence. *PloS one*, **7**(4): e36347 (2012).
- Golicz A A, Batley J, Edwards D. Towards plant pangenomics. *Plant biotechnology journal*, **14**(4): 1099-1105 (2016).
- Goodrich CE et al. The origination and test culture of seedling potatoes. *Trans New York State Agric Soc* **23**:89–134 (1863)
- Ortiz R et al. The state of the use of potato genetic diversity. In: Cooper HD, Spillane C, Hodgkin T (eds) Broadening the genetic base of crop production. CABI Publishing, Wallingford, pp 181–200 (2001)
- Gaur PC, Pandey SK et al. Potato improvement in sub-tropics. In: Paul Khurana SM, Shekhawat GS, Singh BP, Pandey SK (eds) *Potato, global research and development*, vol 1, pp 52–63. Indian Potato Association, Shimla, India (2000)

- Guo, D., Li, Y., Lu, H. et al. A pangenome reference of wild and cultivated rice. *Nature* (2025). <https://doi.org/10.1038/s41586-025-08883-6>.
- Hawkes JG et al. Taxonomic studies on the tuber-bearing solanums. 1. *Solanum tuberosum* and the tetraploid species complex. *Proc Linn Soc London* **166**:97–144 (1956).
- Healey A L, Garsmeur O, Lovell J T, et al. The complex polyploid genome architecture of sugarcane. *Nature*, **628**(8009): 804-810 (2024).
- Hirsch, C. N. et al. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**, 121–135 (2014).
- Hosaka K, Hannernan RE Jr et al. Genetics of self-compatibility in a self-incompatible wild diploid potato species *Solanum chacoense*. I. Detection of an S locus inhibitor (Sli) gene. *Euphytica* **99**:191–197 (1998)
- Jain M, Olsen H E, Paten B, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, **17**: 1-11(2016).
- Jansky, S. H. et al. Reinventing potato as a diploid inbred line–based crop. *Crop Sci.* **56**, 1412–1422 (2016).
- Jayakodi M, Padmarasu S, Haberer G, et al. The barley Pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**(7837):284-289 (2020).
- Jiao, C. et al. Pan-genome bridges wheat structural variations with habitat and breeding. *Nature* **637**, 384–393 (2024).
- Jin LP, Qu DY, Xie KY, Bian CS, Duan SG et al. Potato germplasm, breeding studies in China. In: *Proceedings of the Fifth world potato congress*, Kunming, China, 26–31st March 2004, **pp** 175–177 (2004)
- Knight TA. On raising of new and early varieties of the potato (*Solanum tuberosum*). *Trans Hort Soc London* **1**:57–59 (1807)
- Li, Y. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
- Liao W W, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*, **617**(7960): 312-324 (2023).
- Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol*, **21**(1): 265 (2020).
- Li F, Fan G, Lu C, et al. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature biotechnology*, **33**(5): 524-530 (2015).
- Liu Y, Du H, Li P, et al. Pan-genome of wild and cultivated soybeans. *Cell*, **182**(1):162-176. e13 (2020).
- Liu, Z., Wang, N., Su, Y. et al. Grapevine pangenome facilitates trait genetics and genomic breeding. *Nat Genet* **56**, 2804–2814 (2024).
- Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).

- Wang W, Mauleon R, Hu Z, et al. Genomic variation in 3, 010 diverse accessions of Asian cultivated rice. *Nature*, **557**(7703):43-49 (2018).
- Wenger A M, Peluso P, Rowell W J, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, **37**(10): 1155-1162 (2019).
- Hu T, Chitnis N, Monos D, et al. Next-generation sequencing technologies: An overview. *Human immunology*, **82**(11): 801-811 (2021).
- Kennedy S R, Schmitt M W, Fox E J, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature protocols*, **9**(11): 2586-2606 (2014).
- Kersey P J. Plant genome sequences: past, present, future. *Current opinion in plant biology*, **48**: 1-8 (2019).
- Liu J, Seetharam A S, Chougule K, et al. Gapless assembly of maize chromosomes using long-read technologies. *Genome Biology*, **21**: 1-17 (2020).
- Mendel, G. Versuche über pflanzenhybriden. *Verh. Naturforsch. Vereines Brünn* **4**, 3–47 (1866).
- Quail M A, Kozarewa I, Smith F, et al. A large genome center's improvements to the Illumina sequencing system. *Nature methods*, **5**(12): 1005-1010(2008).
- Phumichai C, Mori M, Kobayashi A, Kamijima O, Hosaka K et al. Toward the development of highly homozygous diploid potato lines using the self-compatibility controlling Sb gene. *Genome* **48**:977–984(2005).
- Plaisted RL, Hoopes RW et al. The past record and future prospects for the use of exotic potato germplasm. *Am Potato J* **66**:603–627 (1989).
- Rhoads A, Au K F. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, **13**(5): 278-289 (2015).
- Ramu P, Esuma W, Kawuki R, et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature genetics*, 2017, **49**(6): 959-963.
- Schiessl S V, Katche E, Ihien E, et al. The role of genomic structural variation in the genetic improvement of polyploid crops. *The Crop Journal*, **7**(2): 127-140 (2019).
- Schreiber, M., Jayakodi, M., Stein, N. & Mascher, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* **25**, 563–577 (2024).
- Shen C, Du H, Chen Z, et al. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Molecular Plant*, **13**(9): 1250-1261 (2020).
- Stokstad, E. The new potato. *Science* **363**, 574–577 (2019).
- Spooner, D. M., Ghislain, M., Simon, R., Jansky, S. H. & Gavrilenko, T. Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. *Bot. Rev.* **80**, 283–383 (2014).
- Sun, X., Jiao, C., Schwaninger, H. et al. Phased diploid genome assemblies and pangenomes provide insights into the genetic history of apple domestication. *Nat Genet* **52**, 1423–1432 (2020).

- Sun H, Jiao W B, Krause K, et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature genetics*, **54**(3): 342-348 (2022).
- Sun Y, Shang L, Zhu Q H, et al. Twenty years of plant genome sequencing: achievements and challenges. *Trends in Plant Science*, **27**(4): 391-401 (2022).
- Sun H, Tusso S, Dent C I, et al. The phased pan-genome of tetraploid European potato. *Nature*, **642**, 389–397 (2025).
- Tang D, Jia Y, Zhang J, et al. Genome evolution and diversity of wild and cultivated potatoes. *Nature*, **606**(7914): 535-541 (2022).
- Tettelin H, Masignani V, Cieslewicz M J, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences*, **102**(39): 13950-13955 (2005).
- The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000)
- The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
- Van Eck H J, Jacobs J M E, Van Den Berg P M M M, et al. The inheritance of anthocyanin pigmentation in potato (*Solanum tuberosum* L.) and mapping of tuber skin colour loci using RFLPs. *Heredity*, **1994**, 73(4): 410-421.
- Venter J C, Adams M D, Myers E W, et al. The sequence of the human genome. *science*, **291**(5507): 1304-1351(2001).
- Wu, Y. *et al.* Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328.e15 (2023).
- Yang J, Moeinzadeh M H, Kuhl H, et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature plants*, **3**(9): 696-703 (2017).
- Ye, M., Peng, Z., Tang, D. et al. Generation of self-compatible diploid potato by knockout of S-RNase. *Nature Plants* **4**, 651–654 (2018).
- Zhang, C. *et al.* The genetic basis of inbreeding depression in potato. *Nat. Genet.* **51**, 374–378 (2019).
- Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
- Zhang, C. *et al.* Genome design of hybrid potato. *Cell* **184**, 3873–3883 (2021).

2

Chapter II Thesis Objectives

1. Objectives

To accelerate the genetic improvement of potato, the “**Upotato Plan**” was launched by **Solab (Solanaceae Laboratory) at the Agricultural Genomics Institute at Shenzhen (AGIS)**, in collaboration with other major research and development partners involved in potato breeding in China. This genome-enabled breeding initiative aims to transform the crop from a **clonally propagated tetraploid** into a **seed-propagated diploid hybrid**.

Similar to hybrid maize, the first critical step in developing hybrid potato varieties is the creation of inbred lines. However, this process is particularly challenging in potato due to the long-term accumulation of deleterious mutations masked in the heterozygous state through clonal propagation.

The **global objective of this thesis** is to contribute to the improvement of diploid potato inbred lines by leveraging high-resolution genomic tools.

Overcoming the challenges associated with inbred line development requires a deep understanding of both genome-wide diversity and the distribution of hidden deleterious mutations in potato genomes. Accordingly, this research focuses on the following specific five key objectives:

1. Fully decipher the diploid potato pangenome across diverse populations.
2. Construct a comprehensive map of genomic variation, including SNPs and structural variants (SVs).
3. Analyze heterozygosity patterns and their contribution to potato domestication.
4. Identify deleterious variants, including deleterious SNPs (dSNPs) and deleterious SVs (dSVs).
5. Design an optimal diploid potato genotype with a minimal load of deleterious variants for use in breeding programs.

These objectives serve as the foundation of the research framework, detailed in the following roadmap and chapter outline.

2. Research roadmap and outline

2.1. Research roadmap

The overall structure of this thesis project is illustrated in **Figure 7**.

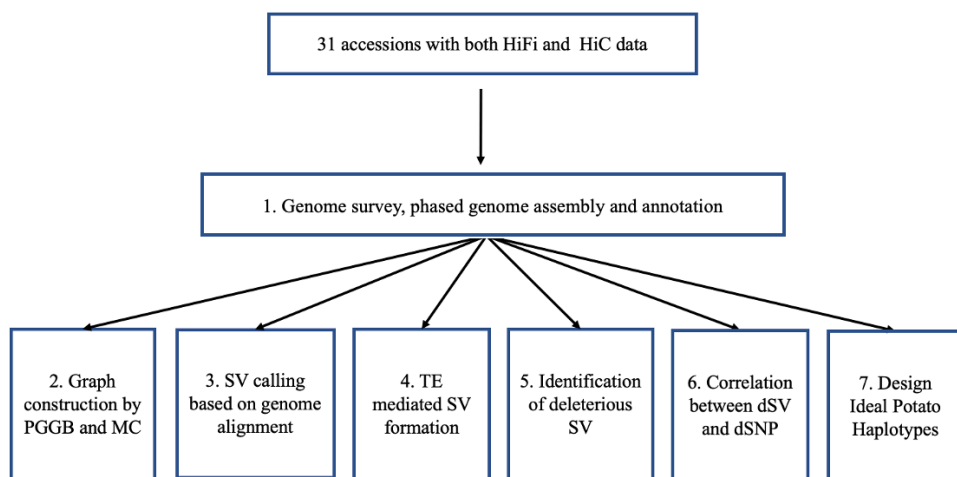


Figure 7: Research roadmap for the thesis.

HiFi: high-fidelity reads. HiC or Hi-C: High-throughput Chromosome Conformation

Capture. PGGB: Pangenome Graph Builder. MC: Minigraph-Cactus. SV: Structural

variations. TE: Transposable element. dSV: deleterious structural variation. dSNPs:

deleterious single nucleotide polymorphism

2.2. Outline

Chapter I provides a literature review on recent advances in potato genomics, including haplotype-resolved genomes, pangenomes, and graph-based pangenomes used for variation detection and statistical analysis. This chapter also highlights how pangenomic approaches contribute to our understanding of crop diversity and genetic improvement, with a specific focus on the highly heterozygous nature of the potato genome.

Chapter II outlines the general objectives and structure of the thesis, setting the stage for the main research questions and methodological framework.

Chapter III presents the core of this thesis, which consists of a comprehensive genomic analysis aimed at elucidating the patterns and effects of deleterious variation in diploid potato genomes:

- We sequenced 31 diploid potato accessions and assembled a total of 60 haplotype-resolved genomes.
- Using these assemblies, we identified extensive structural variants (SVs).
- We investigated transposable elements (TEs) and found that most SVs are associated with TEs, suggesting that TE activity is a major driver of genomic diversity in potato.

- Through phased pangenomes, we explored the patterns of heterozygosity in wild and cultivated potato genomes and revealed the evolutionary forces contributing to increased heterozygosity in domesticated varieties.
- We demonstrated that highly heterozygous genomes can mask deleterious variants, with cultivated potatoes showing more heterozygous deleterious SVs (dSVs) but fewer homozygous dSVs than wild relatives.
- We observed a biased accumulation of deleterious SNPs (dSNPs) near dSVs, a pattern we describe as the “broken windows” effect.
- Finally, we used this information to design **ideal potato haplotypes (IPHs)** with minimal deleterious variation, offering a genome-informed approach to guide breeding efforts.

This chapter emphasizes the power of pangenomics and variation mapping for understanding deleterious variation and provides a blueprint for precision breeding in potato.

Chapter IV presents perspectives and general conclusions, summarizing the key findings and future directions of the thesis.

3

Chapter III Leveraging a phased pangenome for haplotype design of hybrid potato

Article

Leveraging a phased pangenome for haplotype design of hybrid potato

Authors

Lin Cheng^{1,9,14}, Nan Wang^{1,2,14}, Zhigui Bao^{1,3,14}, Qian Zhou⁴, Andrea Guarracino⁵, Yuting Yang¹, Pei Wang¹, Zhiyang Zhang¹, Dié Tang^{1,6}, Pingxian Zhang¹, Yaoyao Wu^{1,7}, Yao Zhou^{1,8}, Yi Zheng¹, Yong Hu¹, Qun Lian¹, Zhaoxu Ma¹, Ludivine Lassois⁹, Chunzhi Zhang¹, William J. Lucas¹⁰, Erik Garrison⁵, Nils Stein^{11,12}, Thomas Städler¹³, Yongfeng Zhou^{1,2} and Sanwen Huang^{1,2,*}

Affiliations

¹National Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China

²National Key Laboratory of Tropical Crop Breeding, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan 571101, China

³Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

⁴School of Agriculture and Biotechnology, Sun Yat-Sen University, Shenzhen 518107, China

⁵Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, 71 S Manassas St, Memphis, TN 38163, USA

⁶Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA

⁷College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

⁸Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences; University of Chinese Academy of Sciences, Beijing 100093, China.

⁹Plant Genetics and Rhizosphere Processes Laboratory, TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech, Gembloux, Belgium

¹⁰Department of Plant Biology, College of Biological Sciences, University of California, Davis, CA 95616, USA

¹¹Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany

¹²Crop Plant Genetics, Institute of Agricultural and Nutritional Sciences, Martin-Luther-University of Halle-Wittenberg, Halle (Saale), Germany

¹³Institute of Integrative Biology and Zurich-Basel Plant Science Center, ETH Zurich, 8092 Zurich, Switzerland

¹⁴Those authors contribute equally

*Corresponding author

Reference: Lin Cheng *et al.* Leveraging a phased pangenome for haplotype design of hybrid potato. **640**, 408–417 *Nature* (2025).

Abstract

The tetraploid genome and clonal propagation of the cultivated potato (*Solanum tuberosum* L.) dictate a slow, non-accumulative breeding mode of the most important tuber crop. Transitioning potato breeding to a seed-propagated hybrid system based on diploid inbred lines has the potential to greatly accelerate its improvement (Zhang et al., 2021). Crucially, the development of inbred lines is impeded by manifold deleterious variants; elucidating their nature and ways to eliminate them is the current focus of hybrid potato research. However, most published diploid potato genomes are unphased, concealing crucial information on haplotype diversity and heterozygosity. Here, we develop a phased potato pangenome graph of 60 haplotypes from cultivated diploids and the ancestral wild species, finding evidence for the prevalence of transposable elements in generating structural variants (SVs). Compared to the linear reference, the graph pangenome represents a 3.1-fold expansion (3,076 Mb versus 742 Mb). Notably, we observe enhanced heterozygosity in cultivated diploids compared to wild ones, indicating extensive hybridization during potato domestication (14.0% versus 9.5%). Using conservative criteria, we identify 19,625 putatively deleterious SVs (dSVs) and reveal a biased accumulation of deleterious SNPs (dSNPs) around dSVs in coupling phase ('broken-window effect'). Based on the graph pangenome, we computationally design ideal potato haplotypes (IPHs) with minimal dSNPs and dSVs. These advances provide critical insights into the genomic basis of clonal propagation and will guide breeders to develop a suite of promising inbred lines.

Keywords: potato, genome, pangenome, TEs, deleterious variations, SV.

1. Introduction

Potato (*Solanum tuberosum* L.) is the most important tuber crop, annually feeding ~1.3 billion people in over 120 countries (Stokstad et al., 2019; Spooner et al., 2014 and Zhang et al., 2021). Recent studies have succeeded in decoding the genomes of monoploid, diploid, and tetraploid potatoes, to break self-incompatibility, and to dissect the genetic basis of inbreeding depression (The Potato Genome Sequencing Consortium, 2011; Zhou et al., 2020; Lian et al., 2019; Bao et al., 2022; Jansky et al., 2016; Zhang et al., 2019; Wu et al., 2023; Van et al., 2020; Freire et al., 2021; Tang et al., 2022; Sun et al., 2022 and Ye et al., 2018). These efforts led to the development of the first generation of highly homozygous inbred lines and subsequently, the first uniform hybrid as a proof-of-concept (Zhang et al., 2021), demonstrating the potential of genome design to revolutionize potato breeding towards a fast, iterative mode.

A central focus of hybrid potato breeding programs is a better understanding of deleterious variants (The Potato Genome Sequencing Consortium, 2011; Zhou et al., 2020; Lian et al., 2019; Bao et al., 2022; Jansky et al., 2016; Zhang et al., 2019 and Wu et al., 2023), which compromise growth and overall fitness (Zhang et al., 2019). Deleterious SNPs have been widely studied in many species such as humans, dogs, and rice (Chun et al., 2009; Marsden et al., 2016; Liu et al., 2017 and Zhang et al., 2021). Structural variants (SVs), however, are expected to affect more genomic regions and are more likely to have stronger impacts on fitness than SNPs; nonetheless, deleterious SVs have rarely been studied. For deleterious variants closely linked in repulsion phase in diploid potatoes (Zhou et al., 2020), phenotype-based selection is neither efficient nor accurate in purging them. Haplotype-based selection based on in-depth understanding of deleterious variants and knowledge-guided recombination strategies is essential for integrating superior genomic fragments from different donors, paving the way for the eventual development of ideal potato haplotypes.

Thus far, our knowledge on haplotype diversity in potatoes is limited. Only a few genomes have been published and most diploid potatoes are assembled into a single chimeric haploid genome, with only a few examples of haplotype-resolved assemblies (Van et al., 2020; Freire et al., 2021 and Tang et al., 2022). Collapsed assemblies of two haplotypes from diploid genomes lose valuable information on haplotype diversity and the phase state, introducing multiple types of errors, including switch errors, false duplications, and nucleotide consensus errors (Zhang et al., 2021 and Jarvis et al., 2022). Haplotype-resolved assemblies provide a more accurate representation of genetic diversity and promise to become essential for future potato breeding (Zhou et al., 2020; Bao et al., 2022 and Sun et al., 2022). Concurrently, there is a need to shift from single reference genomes to phased pangenome references that better capture diversity across multiple populations (Schreiber et al., 2024).

In this study, we de novo assembled 60 haplotypes from 31 diploid potatoes, including 20 haplotypes from 10 wild *S. candolleianum* accessions, 38 haplotypes from 19 domesticated diploid accessions, as well as two haplotypes from two inbred lines. From these 60 potato haplotypes, we constructed a potato pangenome graph, allowing us to uncover extensive non-reference sequences and to document multiallelic SVs. We further reveal the origin and fate of SVs in evolution and

domestication and design ideal potato haplotypes (IPHs) for future breeding. Our study provides in-depth understanding of the haplotype diversity of a clonal crop, establishing the theoretical foundation for the reinvention of potato.

2.Results

2.1 *The potato pangenome graph (PPG-v1.0)*

To capture and characterize haplotype diversity in potato populations, 31 accessions were selected based on previous studies of genetic diversity and principal component analysis (PCA) of 193 accessions (Zhang et al., 2021; Zhou et al., 2020; Zhang et al., 2019 and Tang et al., 2022) (**Figure 8a**; **Figure 9** and **Figure 10**). These comprise 19 landraces covering three indigenous diploid cultivated groups (*S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), two inbred lines (A6-26 and E4-63, the 5th-generation selfed offspring of diploid PG6359 and E86-69, respectively) (Zhang et al., 2021), and 10 wild accessions from *S. candolleianum* (CND), considered to be the progenitor of domesticated potatoes (Hardigan et al., 2024) (**Table 1**). We generated an average of 28.15 Gb (~38-fold) PacBio high-fidelity (HiFi) reads and 110.43 Gb (~149-fold) high-throughput chromosome conformation capture technique (Hi-C) reads (**Table 2** and **Table 3**). Genome surveys using the HiFi reads revealed an average of ~1.55% genomic heterozygosity across cultivated diploid potato genomes (excluding inbred lines; **Table 4**). Our initial de novo assembled haplotypes have an average genome size of 811 Mb, with contig N50 of 12.25 Mb (**Figure 8b**). Paired haplotypes of each heterozygous diploid were further scaffolded based on Hi-C contact maps (**Figure 11**), resulting in pseudo-chromosome assemblies with an average anchor rate of 95.17%. Assessment using Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao et al., 2015) demonstrates a high level of completeness with an average of 99.2%.

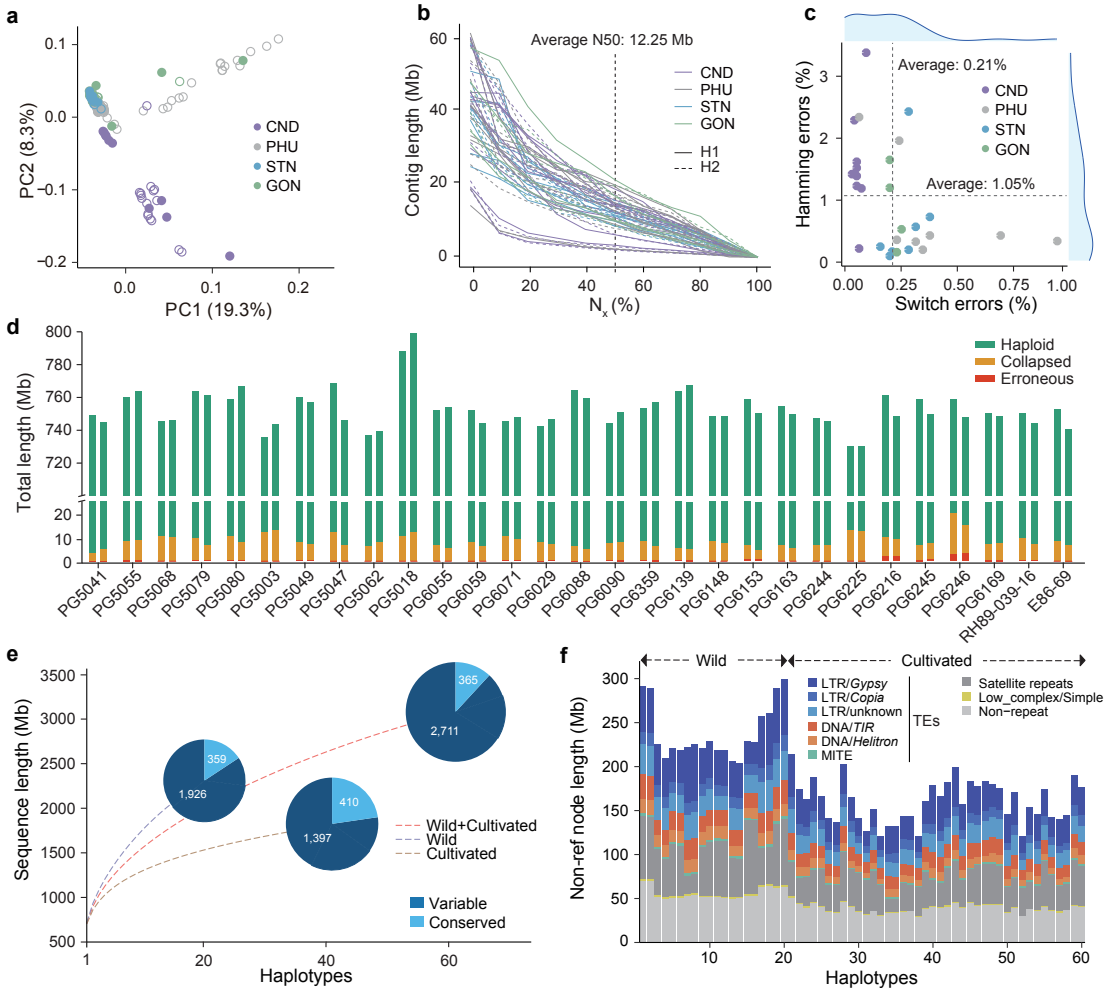


Figure 8: Phased potato pangenome reference (PPG-v1.0).

a, Principal component analysis from 193 potato accessions; the 31 selected accessions are highlighted with solid points. *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND). b, Assessment of genome contiguity by contig N50 for haplotype 1 (H1, solid lines) and haplotype 2 (H2, dashed lines). c, Assessment of correctness by switch errors and Hamming errors for each diploid assembly. d, Reliability of the 29 diploid assemblies (excluding 2 inbred lines). For each haplotype-resolved assembly, the left bar represents H1 and the right bar represents H2. Regions identified as haploid are considered trustworthy (green). The y-axis is broken to illustrate both the prominence of the reliable haploid component and the

stratification of the unreliable segments. e, Pangenome growth curves for wild, cultivated, and wild+cultivated groups. Pie plots represent conserved sequence (present in at least 90% of the haplotypes) and variable sequence in the Pangenome graph builder (PGGB) pangenome graph. Numbers within the pie plots represent sequence length (Mb). f, Non-reference nodes (>50 bp) in the potato PGGB pangenome were annotated as four primary types, including Transposable elements (TEs), satellite repeats, low-complex (simple repeats) and non-repetitive sequences. Six different TE families were identified. om left to right: inbred line E4-64, F1 hybrid (inbred line E4-63 crossed with A6-26), and inbred line A6-26. The other lines are cultivated tetraploid potatoes.

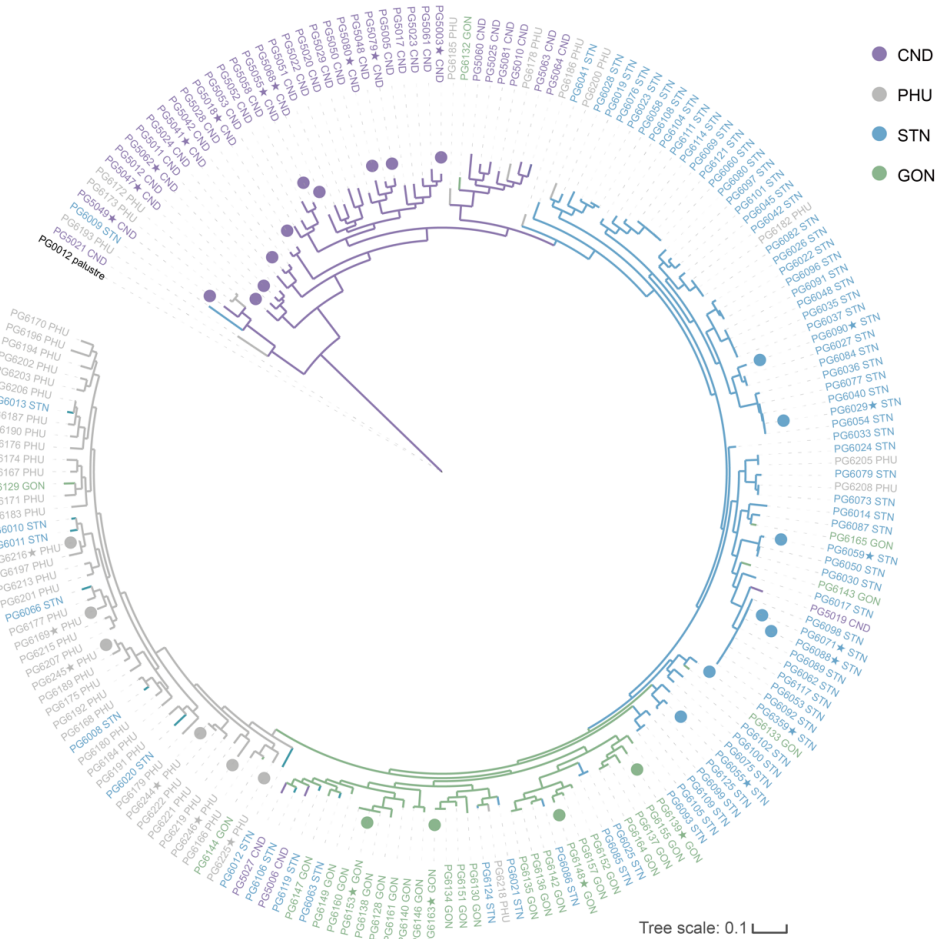


Figure 9: Phylogenetic reconstruction of diverse potato accessions.

A total of 31 accessions were used to construct the potato pangenome graph (PPG-v1.0), including 27 sequenced diploid potato accessions from the international potato center (CIP) accessions are highlighted in the figure, two diploid cultivars (RH89-039-16 and E86-69), and two inbred lines (A6-26 and E4-63). *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND)

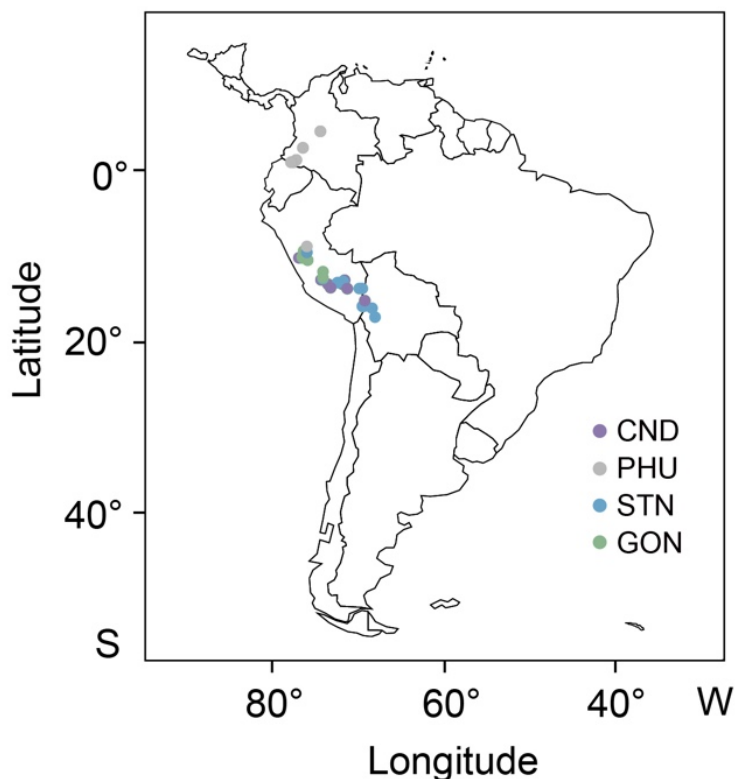


Figure 10: The location of sequenced potato accessions.

Geographic collection locations of 27 sequenced potatoes from the international potato center (CIP) accessions (<https://npgsweb.ars-grin.gov/gringlobal/search>). The base map was generated using the function mapBubbles() in the R package rworldmap. *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND).

Table 1: Information of accessions in this study.

Note,* represent the inbred lines. *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND).

ID	Accession	CIP Number	Type	Group	Data type
1	PG5041	PI 442697	Wild	CND	HiFi / Hi-C
2	PG5055	PI 545972	Wild	CND	HiFi / Hi-C
3	PG5068	PI 458379	Wild	CND	HiFi / Hi-C
4	PG5079	PI 310945	Wild	CND	HiFi / Hi-C
5	PG5080	PI 310946	Wild	CND	HiFi / Hi-C
6	PG5003	PI 458403	Wild	CND	HiFi / Hi-C
7	PG5049	PI 458380	Wild	CND	HiFi / Hi-C
8	PG5047	PI 210040	Wild	CND	HiFi / Hi-C
9	PG5062	PI 210055	Wild	CND	HiFi / Hi-C
10	PG5018	PI 365317	Wild	CND	HiFi / Hi-C
11	PG6055	CIP 700235	Cultivated	STN	HiFi / Hi-C
12	PG6059	CIP 700407	Cultivated	STN	HiFi / Hi-C
13	PG6071	CIP 702353	Cultivated	STN	HiFi / Hi-C
14	PG6029	CIP 703473	Cultivated	STN	HiFi / Hi-C
15	PG6088	CIP 703801	Cultivated	STN	HiFi / Hi-C
16	PG6090	CIP 703808	Cultivated	STN	HiFi / Hi-C
17	PG6359	CIP 705468	Cultivated	STN	HiFi / Hi-C
18	PG6139	CIP 702467	Cultivated	GON	HiFi / Hi-C
19	PG6148	CIP 703825	Cultivated	GON	HiFi / Hi-C
20	PG6153	CIP 704254	Cultivated	GON	HiFi / Hi-C
21	PG6163	CIP 705934	Cultivated	GON	HiFi / Hi-C
22	PG6244	CIP 703509	Cultivated	PHU	HiFi / Hi-C
23	PG6225	CIP 704120	Cultivated	PHU	HiFi / Hi-C
24	PG6216	CIP 704228	Cultivated	PHU	HiFi / Hi-C
25	PG6245	PI 225669	Cultivated	PHU	HiFi / Hi-C
26	PG6246	PI 283119	Cultivated	PHU	HiFi / Hi-C
27	PG6169	PI 225677	Cultivated	PHU	HiFi / Hi-C
28	RH89-039-16	-	Cultivated	PHU	HiFi / Hi-C
29	E86-69	-	Cultivated	PHU	HiFi / Hi-C

30	A6-26*	-	Inbred line	STN	HiFi / Hi-C
31	E4-63*	-	Inbred line	PHU	HiFi / Hi-C
32	DM1-3 516 R44 (DM v6.1)	-	Double haploid	PHU	Reference genome
33	-	-	Outgroup	-	-
34	-	-	Outgroup	-	-
35	-	-	Outgroup	-	-

Table 2: Statistics of raw HiFi data of 31 potato accessions.

Note,* represent the inbred lines.

ID	Accession	Read base (Gb)	Coverage	Read count
1	PG5041	31.46	42	2,034,252
2	PG5055	34.70	46	2,204,007
3	PG5068	19.71	26	1,003,663
4	PG5079	31.79	42	2,018,283
5	PG5080	35.32	47	2,261,561
6	PG5003	17.28	23	836,495
7	PG5049	32.07	43	1,938,057
8	PG5047	27.72	37	1,647,893
9	PG5062	32.94	44	1,818,355
10	PG5018	22.62	30	1,088,261
11	PG6055	25.70	34	1,306,318
12	PG6059	28.10	37	2,250,313
13	PG6071	30.15	40	1,815,881
14	PG6029	25.61	34	2,101,695
15	PG6088	28.89	39	1,662,954
16	PG6090	29.87	40	1,815,963
17	PG6359	29.14	39	2,193,724
18	PG6139	32.24	43	1,804,820
19	PG6148	24.51	33	1,670,567
20	PG6153	34.14	46	2,018,875
21	PG6163	21.20	28	1,176,656

22	PG6244	22.69	30	1,621,291
23	PG6225	15.88	21	1,232,314
24	PG6216	38.13	51	2,163,198
25	PG6245	32.09	43	2,109,160
26	PG6246	30.21	40	2,068,191
27	PG6169	29.30	39	1,575,644
28	RH89-039-16	30.92	41	2,322,217
29	E86-69	24.85	33	1,940,371
30	A6-26*	25.31	34	2,176,997
31	E4-63*	28.24	38	1,816,048

Table 3: Statistics of raw HiC data of 31 potato accessions.

Note,* represent the inbred lines.

ID	Accession	Read base (Gb)	Coverage	Read count
1	PG5041	110.06	148	733,748,716
2	PG5055	98.99	133	659,912,574
3	PG5068	132.54	179	884,773,600
4	PG5079	96.19	130	641,252,884
5	PG5080	115.74	156	771,628,998
6	PG5003	137.14	185	914,242,058
7	PG5049	99.42	134	662,785,354
8	PG5047	99.33	134	662,204,634
9	PG5062	116.54	157	776,908,368
10	PG5018	95.78	129	638,561,370
11	PG6055	100.52	135	670,143,078
12	PG6059	103.94	140	692,918,190
13	PG6071	101.88	137	679,212,294
14	PG6029	102.82	139	685,453,022
15	PG6088	97.76	132	651,766,516
16	PG6090	100.29	135	668,627,112
17	PG6359	146.40	197	977,261,468
18	PG6139	111.43	150	742,839,158

19	PG6148	100.58	136	670,538,062
20	PG6153	104.56	141	697,085,858
21	PG6163	102.75	138	685,014,636
22	PG6244	119.80	161	798,670,034
23	PG6225	99.24	134	661,623,468
24	PG6216	115.67	156	771,108,180
25	PG6245	99.42	134	662,832,290
26	PG6246	101.12	136	674,117,990
27	PG6169	118.34	159	788,939,706
28	RH89-039-16	136.92	185	912,817,118
29	E86-69	130.11	175	867,369,420
30	A6-26*	93.64	126	624,297,652
31	E4-63*	134.42	181	896,141,254

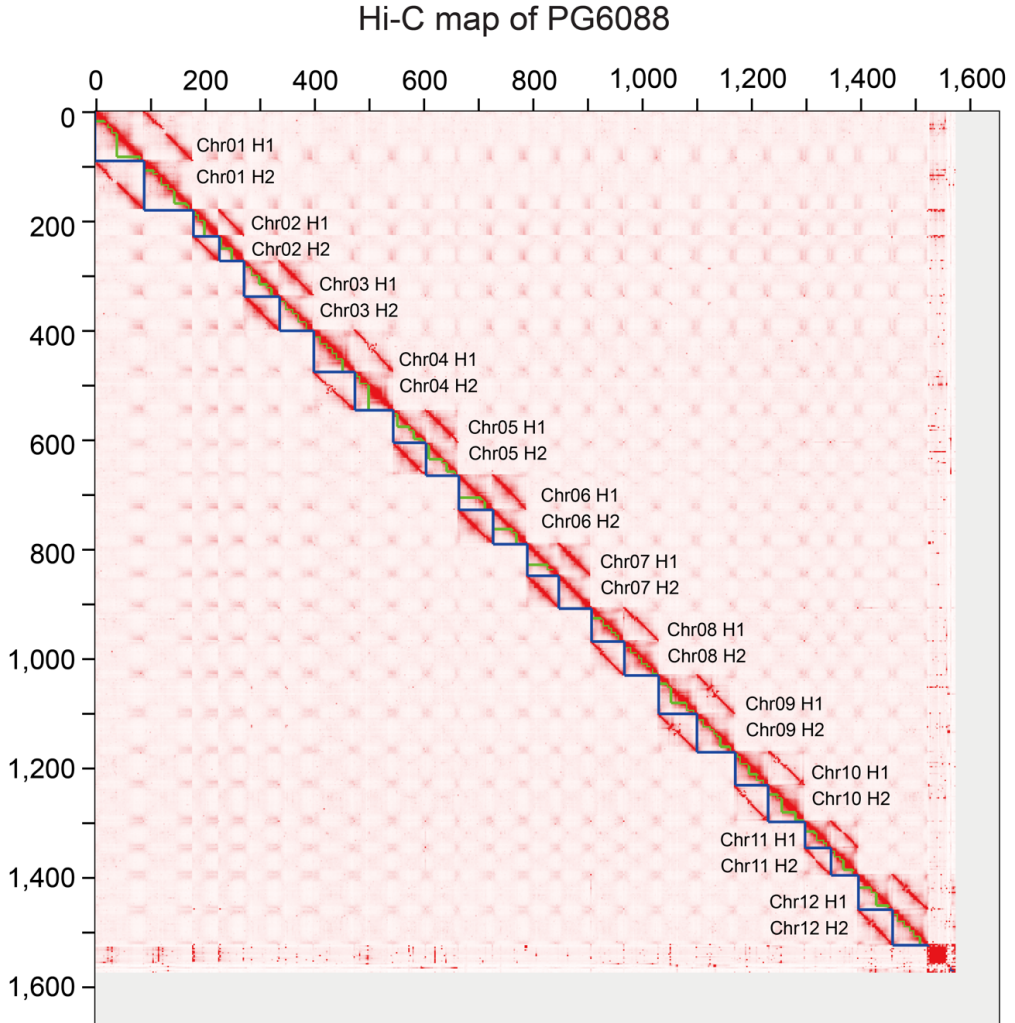


Figure 11: The (High-throughput Chromosome Conformation Capture) Hi-C map for final genome assembly.

The Hi-C map of diploid potato PG6088. The x- and y-axis represent physical distance in genome. Two haplotypes of each chromosome are clustered together. Blue color indicates chromosome boundary. Green color indicates contig boundary. Hi-C maps of all 29 heterozygous diploid potato assemblies have been upload to Figshare (<https://figshare.com/articles/dataset/IPHs/25846003>).

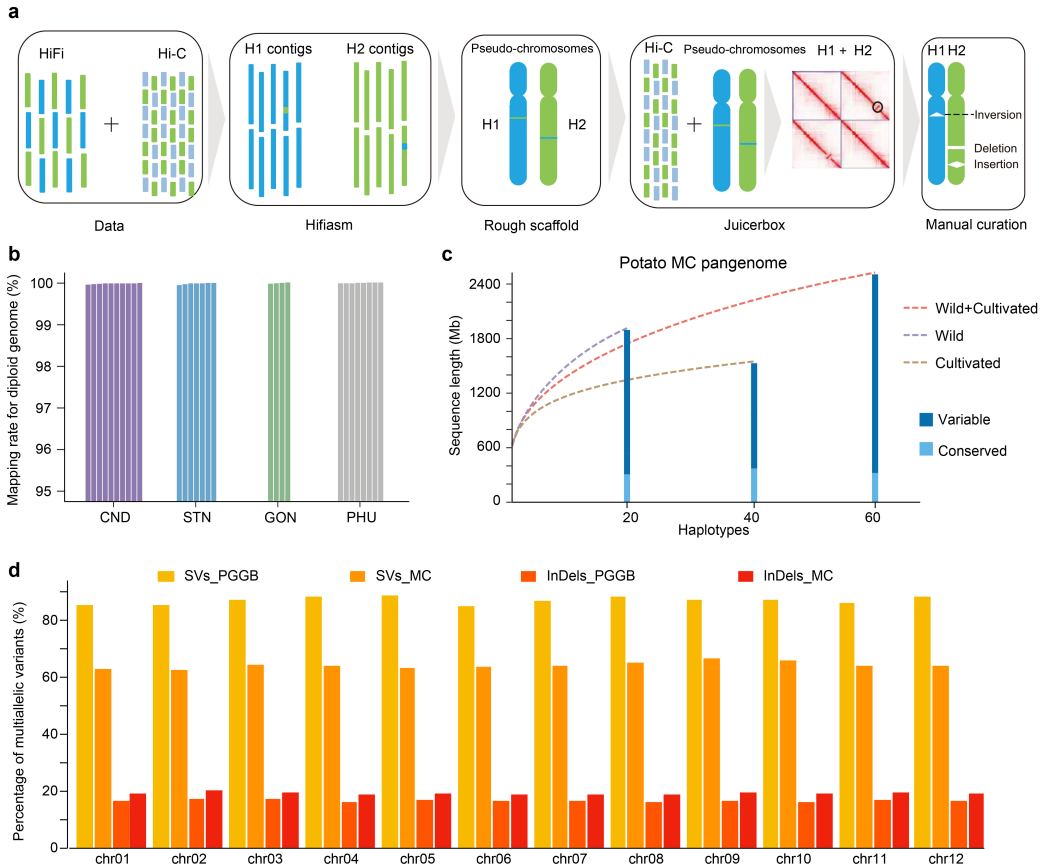


Figure 12: Assembly strategy and validation of the potato pangenome.

a, Schematic diagram of the assembly process of the haplotype-resolved potato genome. b, Assessment of genome completeness by the read-mapping rate. c, Pangenome growth curves for 61 haplotypes (plus DMv6.1 reference) based on the MC graph, including conserved (present in at least 90% haplotypes) and variable sequence (<90% of haplotypes). d, Percentage of multiallelic structural variations (SVs) and InDels for the pangenome graph builder (PGGB) and minigraph cactus (MC) potato pangenomes. *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND).

Three additional independent analyses confirm the quality of our assemblies: (1) the low switch error rates (0.21% on average) and Hamming errors (1.05% on average) indicate the high correctness of our assemblies (**Figure 8c**); (2) the high read mapping rate (mean 99.99%) and consistent k-mer spectrum confirm the completeness of the

assemblies; and (3) to further estimate the reliability of our assemblies, we also used Flagger (Liao et al., 2023) to estimate the potentially erroneous regions. This confirmed that 98.9% of our assemblies are not collapsed or duplicated (**Figure 8d**). We annotated an average of 449.0 Mb (59.69%) as being composed of transposable elements (TEs) and predict 36,421–40,781 gene structure models for each haplotype.

To better represent non-reference sequences in potato genomes, we built two pangenome graphs for 60 haplotype assemblies plus the reference DMv6.1 (Pham et al., 2020) using the PanGenome Graph Builder (PGGB) (Garrison et al., 2024) and Minigraph-Cactus (MC) (Hickey et al., 2024), respectively. We found that the MC pipeline removed highly diverged and large inverted regions, resulting in the discarding of an average of ~14.5% of sequence per chromosome. Therefore, we utilize PGGB to build a potato pangenome graph, which reduces reference bias (Garrison et al., 2023) and is relatively more sensitive for variation calling than methods based on mapping to a linear reference genome. This unbiased potato pangenome graph (PPG-v1.0) comprises 248.64 million nodes and 345.61 million edges with total sequence length of 3,076 Mb. Wild potato accessions contribute a larger portion to the pangenome graph (2,286 Mb) compared to cultivated potato assemblies (1,807 Mb), indicating a higher pangenome graph growth rate in wild potatoes. We determined that PPG-v1.0 consists of 365 Mb conserved sequences (present in at least 90% of the haplotypes) and 2,711 Mb variable sequences (**Figure 8e**). Using the DMv6.1 reference as a baseline, PPG-v1.0 also reveals prevalent rearrangements (**Figure 13**). Furthermore, we annotated non-reference nodes in each haplotype path: 52.6% of them overlap with TEs, 22.7% with satellite repeats, and 23.8% are classified as non-repetitive sequences (**Figure 8f**). Notably, 13.3% of repetitive sequences exhibit characteristics of both transposons and satellites, potentially originating from atypical centromere or subtelomere repeats (Gong et al., 2012) in the potato genome.

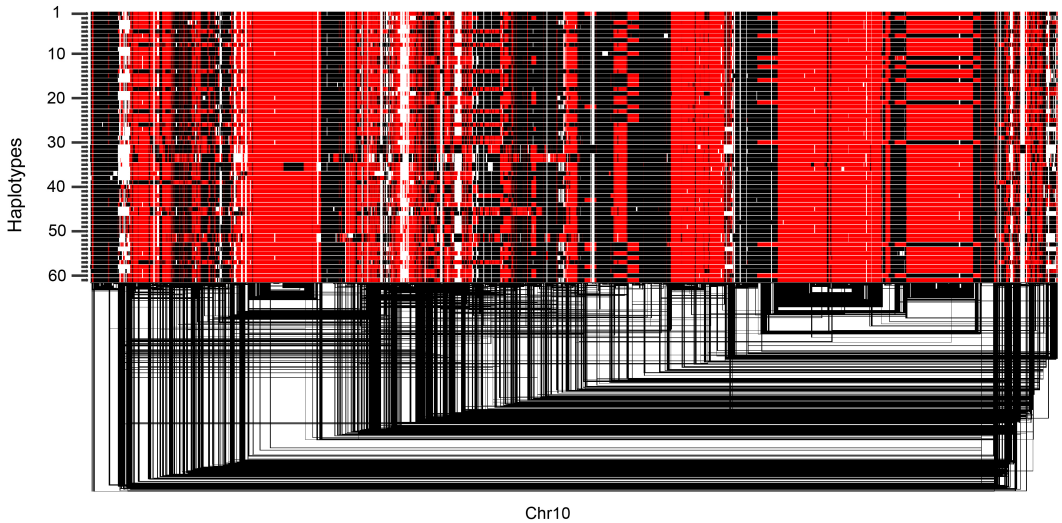


Figure 13: Visualization of the potato chromosome 10 pangenome graph.

The visualization utilizes a 1D representation of the pangenome graph, with the DMv6.1 reference genome serving as the baseline for orientation. Graph nodes are arranged from left to right corresponding to their positions within the reference genome. The black lines under the paths are the links, which represent the graph topology. Black and red bars indicate regions oriented in the forward and reverse direction relative to the pangenome sequence (the complete sequence formed by concatenating all graph nodes). Access to visualizations of the 1D pangenome graph builder (PGGB) pangenome graphs for all chromosomes can be found in Figshare (<https://figshare.com/articles/dataset/IPHS/25846003>).

Finally, we deconstructed genome variation from the graph structure for small variants (SNPs and InDels < 50 bp) and structural variants (SVs ≥ 50 bp) based on DMv6.1 coordinates. In total, the constructed potato haplotype variation map includes 46,012,502 SNPs, 3,587,850 InDels and 133,264 SVs. Interestingly, 16.9% of InDels and 87.0% of SVs are multiallelic, with three or more distinct alleles. The graph generated by the MC pipeline shows a similar trend (**Figure 12c,d**). Collectively, the PPG-v1.0 generated from phased diploid assemblies provides a fundamental resource for in-depth exploration of haplotype diversity and its consequences for potato breeding.

2.2 TEs drive structural variation (SV) formation

To reveal the mechanistic origins and dynamics of SVs in potato genomes, we analyzed the pangenome SV sequences and their residual flanking sequences (± 100 bp), which we annotated as one of three types: TEs, tandem repeats (TRs) and segmental duplications (SDs). We found that TEs account for 90.6% of repetitive

elements associated with SVs (**Figure 14a** and **Table 4**). Additionally, 28.2% of SVs are covered by a single TE sequence, suggesting that these features likely reflect TE insertion events (Bozan et al., 2023 and Dominguez et al., 2020). Recent reports suggest that TEs could serve as substrates for ectopic DNA repair leading to SV formation (Wicker et al., 2005 and Balachandran et al., 2022). Therefore, we examined SV breakpoints and found that 33.8% could be classified as being compatible with ectopic recombination events from TE-mediated rearrangement (TEMR), characterized by both breakpoints (100 bp) with the same TE class. This observed TEMR rate is significantly higher than the expected rate of 7.7% ($P < 2.2 \times 10^{-16}$, two-tailed Student's t-test), supporting a significant contribution of homologous TEs in mediating SV formation. TEMR events were further classified based on TE families, revealing that LTR/Gypsy-mediated TEMR (42.4%) is seven times more prevalent than that involving LTR/Copia (7.0%) (**Figure 14b**). Moreover, SVs mediated by LTR/Gypsy TEMR are significantly longer than those mediated by LTR/Copia TEMR (average of 7,005 bp versus 1,899 bp; $P < 2.2 \times 10^{-16}$, two-sided Student's t-test), and exhibit more recent and frequent activity (**Figure 14c**).

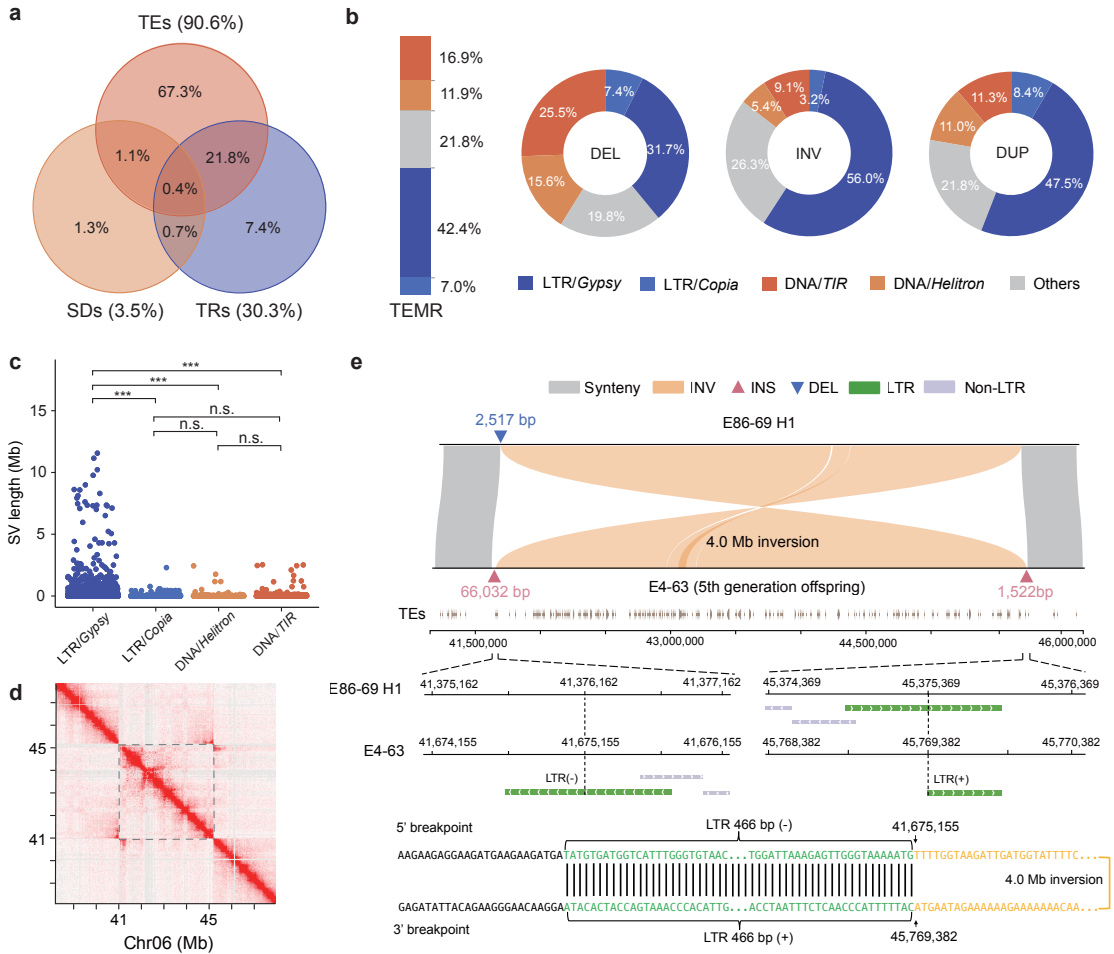


Figure 14: TE activity drives structural variants in potato haplotypes.

a, Proportion diagram of structural variation formation driven by different repetitive sequences, including transposable elements (TEs), tandem repeats (TRs), and segmental duplications (SDs). b, Left: Percentage of TE-mediated rearrangement (TEMR) through ectopic recombination by different families. Right: Percentage of different TE families in TEMR among the three SV types. DEL: deletion; INV: inversion; DUP: duplication. Inserted sequences (INS) can be formed by TE movement and do not fall under our definition of TEMR. c, Comparison of SV length of TEMR for different TE families. A two-sided Student's t-test was used to calculate the p-values. N.s., not significant; *** indicates $P < 2.2 \times 10^{-16}$, $n = 2,768$ (LTR/Gypsy), 361 (LTR/Copia), 596 (DNA/Helitron) and 1,044 (DNA/TIR). d, Validation for the 4.0-Mb de novo inversion using Hi-C data from potato

E86-69; the haplotype of inbred line E4-63 was used as the reference genome. e, Sequence analysis in the de novo inversion. Top: Synteny plot of the de novo inversion. Middle: Distribution of TEs in the de novo inversion. Bottom: The 466-bp homologous sequence stretch of the LTR mediating the de novo inversion.

Table 4: The proportion of TE-mediated rearrangement

Type	Expected rate under random shuffling (%)	Observed (%)
LTR/ <i>Copia</i>	0.03	2.5
LTR/ <i>Gypsy</i>	5.65	13.7
LTR others	1.69	7.2
<i>Helitron</i>	0.15	4.2
CACTA	0.00	1.0
hAT	0.02	1.3
Mutator	0.09	2.9
PIF	0.00	0.0
Tc1	0.01	0.9

To investigate the potential recent TE activity in potato genomes, we also compared two homozygous inbred lines with their heterozygous founders. Interestingly, we identified a ~4.0 Mb (chr06: 41.7-45.7 Mb) LTR-mediated de novo paracentric inversion (**Figure 14d,e**) in the 5th-generation inbred line E4-63, but not in the two founder haplotypes of the parental E86-69. Therefore, it is advisable to approach this de novo inversion with caution when implementing recurrent selection in hybrid potato breeding (Fang et al., 2012 and Berdan et al., 2021).

2.3 Domestication boosts heterozygosity

Previous studies suggested elevated heterozygosity in cultivated potatoes compared to wild ones (Jarvis et al., 2017), consistent with heterozygosity evaluated in our study based on k-mers statistics using HiFi reads. To better describe heterozygosity and the loss of collinearity between the two haplotypes within the same diploid individual, we introduce the concept of Genome Heterozygosity by Sequence Length (GHSL). This approach complements existing k-mer-based methods by quantifying the total sequence length of non-redundant variants (SNPs, InDels, and SVs) between two haplotypes in phased diploid assemblies (**Figure 15**). We found GHSL to be approximately 93.8 Mb in our phased assemblies, constituting 12.5% of the average haplotype length (**Figure 16a**). This estimate is similar to the levels estimated by read mapping in clonally propagated grapevine (Zhou et al., 2019). Moreover, our analysis

shows that inversions account for a large proportion of haplotype-specific SVs (59.4% on average).

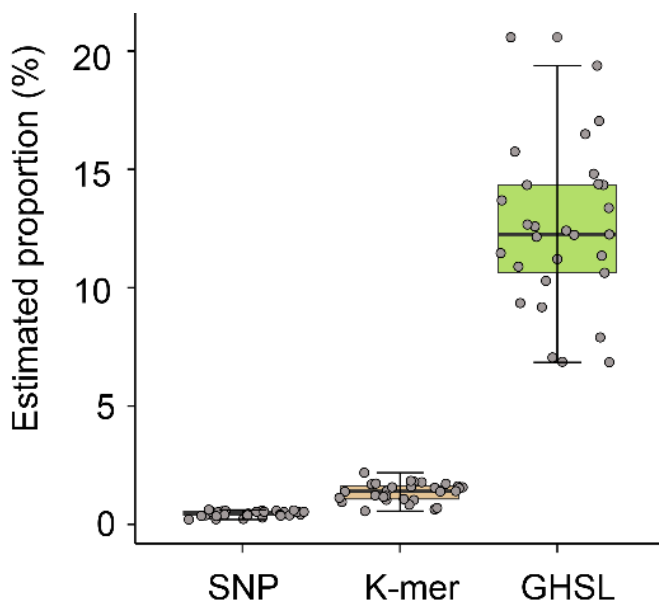


Figure 15: Estimation of heterozygosity by SNP-based, K-mer-based, and Genome heterozygosity by sequence length (GHSL)-based methods.

Three independent methods were employed to estimate genome heterozygosity (n=29 accessions). The SNP-based method calculated heterozygosity using observed heterozygous SNPs. The K-mer-based method considered k-mer frequency and distribution. GenomeScope (v2.0) was used to infer genome heterozygosity proportion. Finally, Genome Heterozygosity by Sequence Length (GHSL) was determined by summing non-redundant heterozygous variants. The boxes represent 75% and 25% quartiles, the central line indicates the median and the whiskers extend to 1.5 times the interquartile range

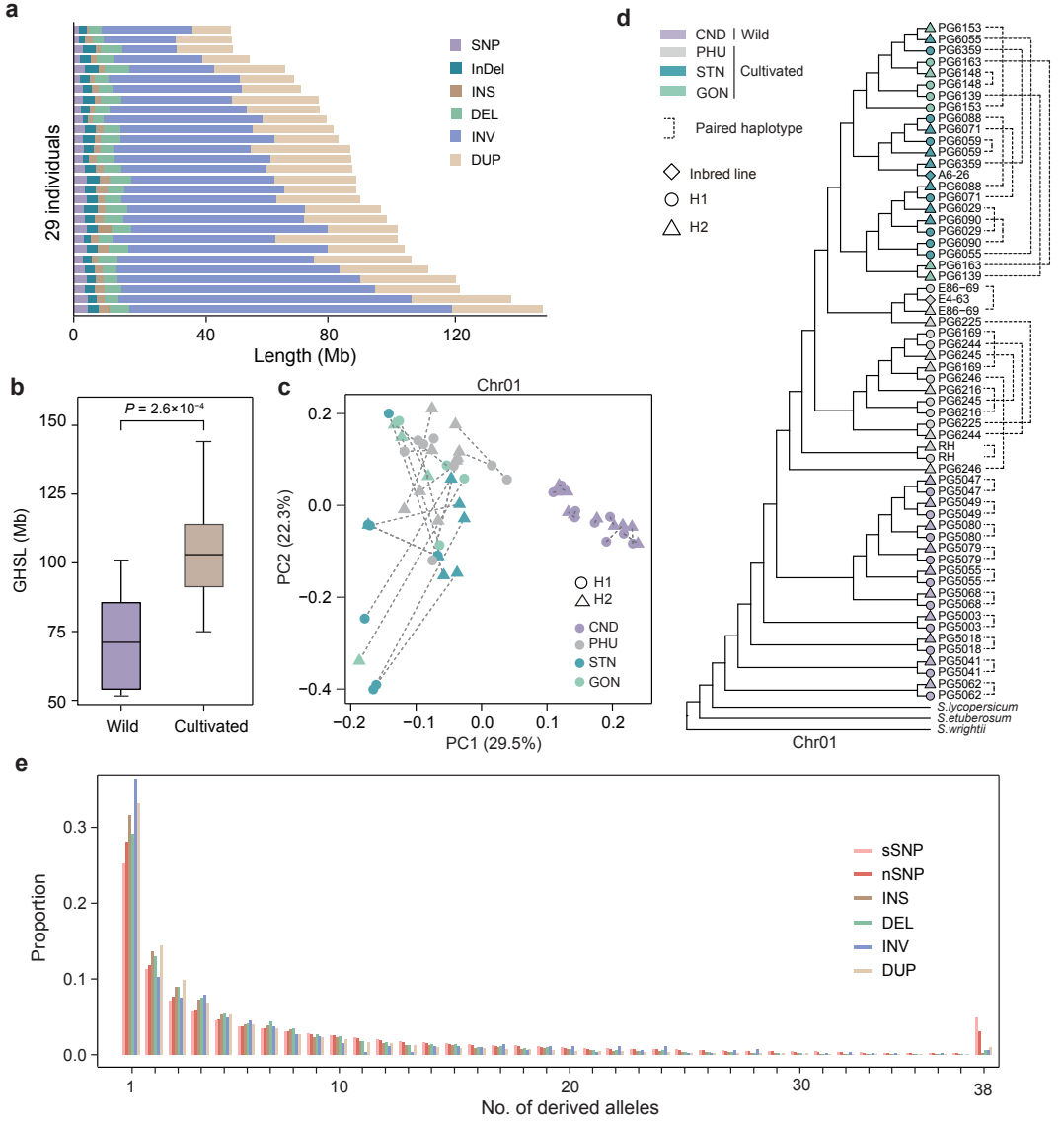


Figure 16: Potato domestication enhanced haplotype divergence.

a, Genomic length affected by haplotype-specific (heterozygous) variants by comparing haplotype 1 (H1) to haplotype 2 (H2) across 29 diploid potato accessions. b, Genome Heterozygosity by Sequence Length (GHSL) in wild and cultivated potatoes. The boxes represent 75% and 25% quartiles, the central line indicates the median and the whiskers extend to 1.5 times the interquartile range; $P = 2.6 \times 10^{-4}$, two-tailed Student's t-test, $n = 19$

(cultivated) and 10 (wild). c, PCA of chromosome 1 haplotypes, with dotted lines connecting paired haplotypes in heterozygous diploid potatoes. d, The phylogeny of 63 haplotypes for chromosome 1 was constructed using the maximum-likelihood method based on 1,074 single-copy genes. The dashed lines (right) connect the paired haplotypes in each diploid genome. RH: RH89-039-16. e, Unfolded site frequency spectrum of four types of SVs and non-synonymous single nucleotide polymorphisms (nSNPs) compared to putatively neutral synonymous SNPs (sSNPs) across cultivated potatoes. *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND).

We found that haplotype divergence in cultivated potatoes, as quantified by GHSL, is 1.46 times greater than that observed across wild potatoes (~105.2 Mb versus ~72.1 Mb; **Figure 16b**; $P = 2.6 \times 10^{-4}$, two-sided Student's t-test). Notably, two haplotype-specific inversions are quite prevalent in cultivated potatoes (**Figure 17** and **Figure 18**). One inversion has previously been reported and is associated with a locus regulating yellow tuber flesh (DMv6.1, chr03: 42.9-48.7 Mb) (Tang et al., 2022). The other inversion (DMv6.1, chr10: 52.7-59.1 Mb) contains 661 genes including *StPIN7* (Soltu.DM.10G026500.1) and *StPTB6* (Soltu.DM.10G026670.1), which may be

associated with auxin transport (Roumeliotis et al., 2013) and tuberization (Cho et al., 2015), respectively.

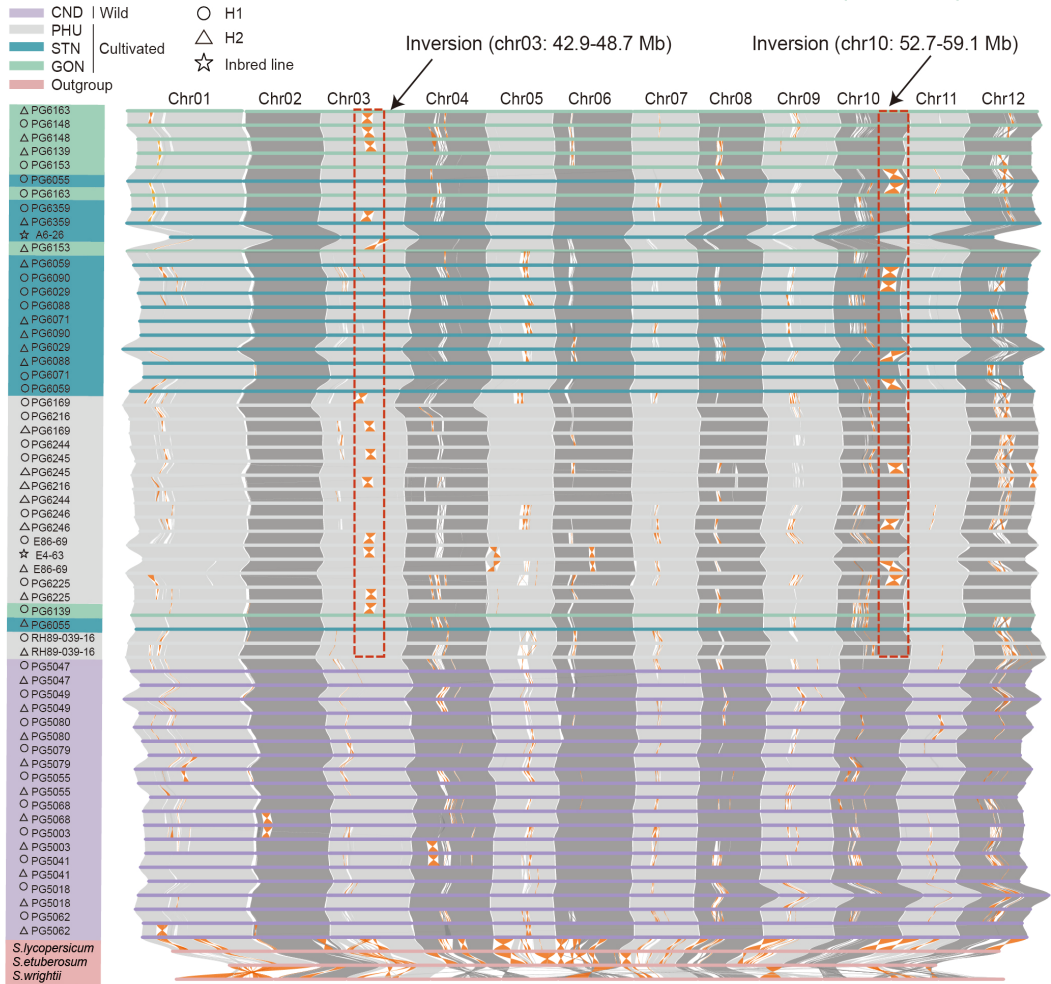


Figure 17: Different chromosomes are distinguished by alternating light gray and dark gray.

Two heterozygous inversions prevalent in cultivated potatoes are marked with dashed-line red rectangles. *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND).

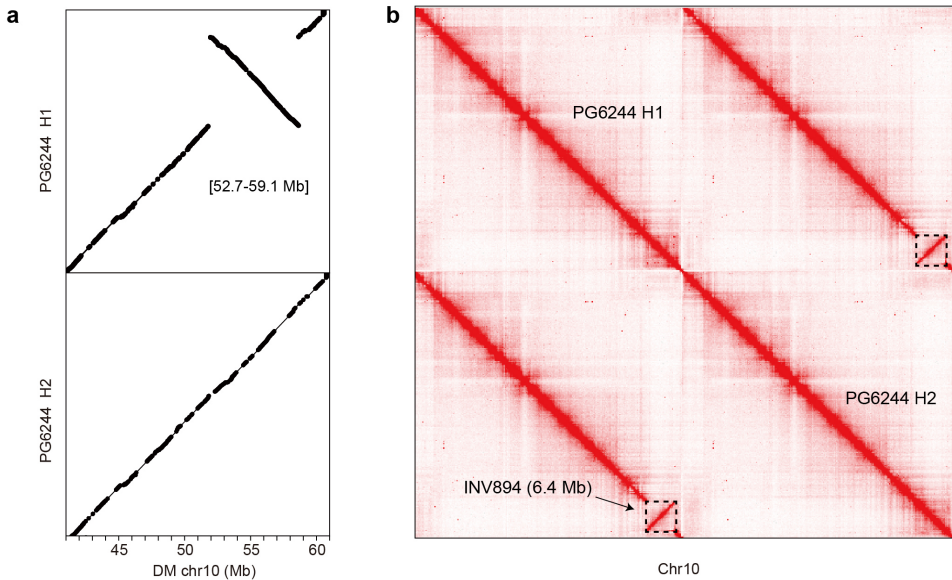


Figure 18: An example of a heterozygous inversion in the PG6244 genome assembly.

a, Synteny plot for two haplotypes aligned to the DM genome. The haplotype H1 of PG6244 exhibits a 6.4-Mb inversion (INV894, chr10: 52.7-59.1 Mb) on chromosome 10 but PG6244 H2 shows synteny with the DMv6.1 genome. b, The Hi-C map for the assembly of the two PG6244 haplotypes for chromosome 10. An inversion signal between the two haplotypes is marked by the black rectangles.

To explore the extent of genome heterozygosity and haplotype divergence during potato domestication, we conducted PCA separately for each chromosome and found that the two haplotypes of each wild potato form “nearest neighbors”, whereas the two haplotypes of each cultivated potato exhibit a more dispersed pattern (**Figure 16c**). We constructed haplotype trees for each chromosome pair based on single-copy genes, using three additional *Solanum* species as outgroups (**Figure 16d**). Analogous to the PCA-based patterns, the two haplotypes of each cultivated potato typically reveal more pronounced genetic divergence than those of each wild potato. We conducted additional validation through pairwise Jaccard similarity analysis and alignment for split windows, both of which indicate that multiple hybridization events have impacted the haplotype landscape in cultivated potatoes (**Figure 19**). Molecular evidence in maize hybrids suggests that the length of nonsyntenic regions between parental lines is strongly correlated with levels of heterosis (Wang et al., 2023). Our findings of increased genome heterozygosity and haplotype divergence in cultivated potatoes are similar to those in grapevine, another clonal crop (Xiao et al., 2023). The enhanced genome heterozygosity in cultivated potatoes, as compared to wild potatoes,

results in reduced homozygous deleterious burden (**Figure 20a,b**), indicating that potato domestication included an exploration of heterosis.

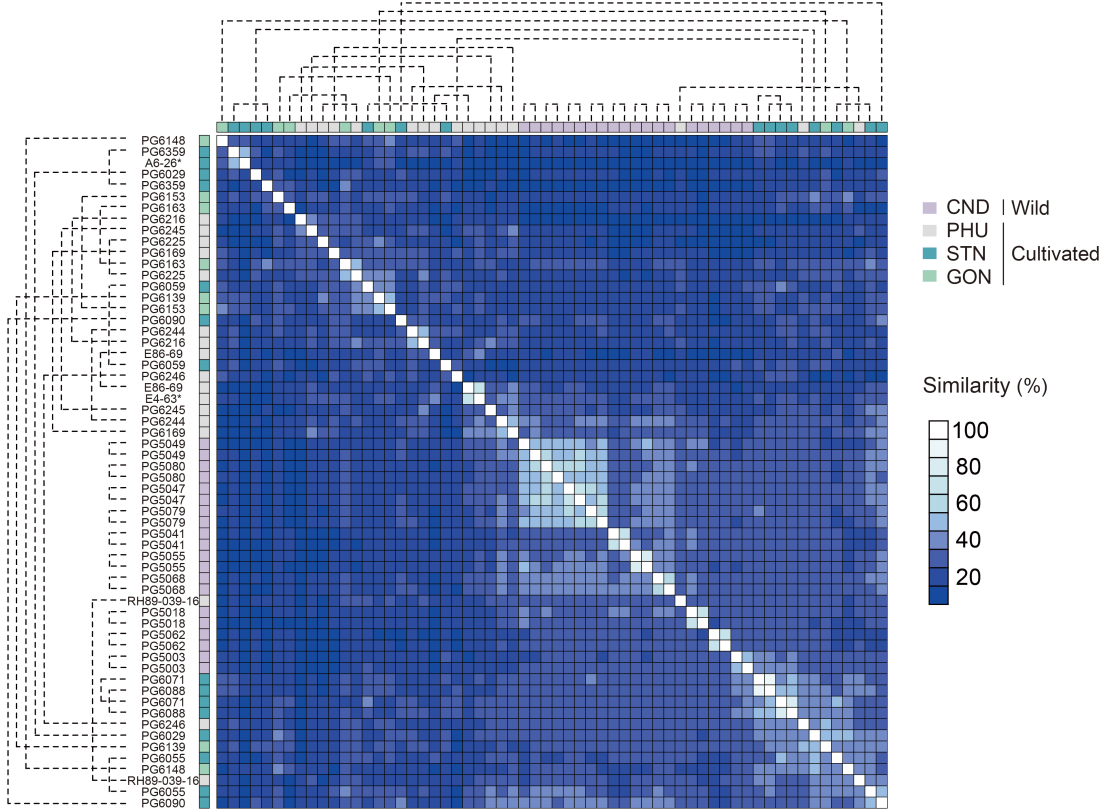


Figure 19: Pairwise Jaccard similarity analysis of 60 potato haplotypes.

The heatmap indicates the correlation among 60 potato haplotypes. Paired haplotypes of each accession are indicated at the top and left with dashed lines. The rectangle colors represent different potato groups (see legend in upper-right corner). *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND).

2.4 Fate of deleterious SVs (dSVs) under breeding

To characterize the fitness effects of SVs in potato cultivars, we computed the unfolded site frequency spectrum (SFS) based on SVs and SNPs from domesticated potatoes (**Figure 16e**). We found that SVs are over-represented for singletons and other minor-allele frequency classes, compared to both synonymous SNPs (sSNPs)

and even nonsynonymous SNPs (nSNPs) at minor frequencies, consistent with equivalent observations in grapevine genomes (Zhou et al., 2019). Cultivated potatoes harbor ~ 2.7 times more heterozygous SVs (average of 20,613 per individual) than homozygous SVs (average of 7,561) (**Figure 20a**). Additionally, 39.4% of the heterozygous SVs are located within either the gene bodies or in putative promoter regions (**Table 5**). These findings suggest that SVs are more likely to be strongly deleterious than SNPs and typically present in heterozygous state with relatively low frequencies across diploid potato cultivars.

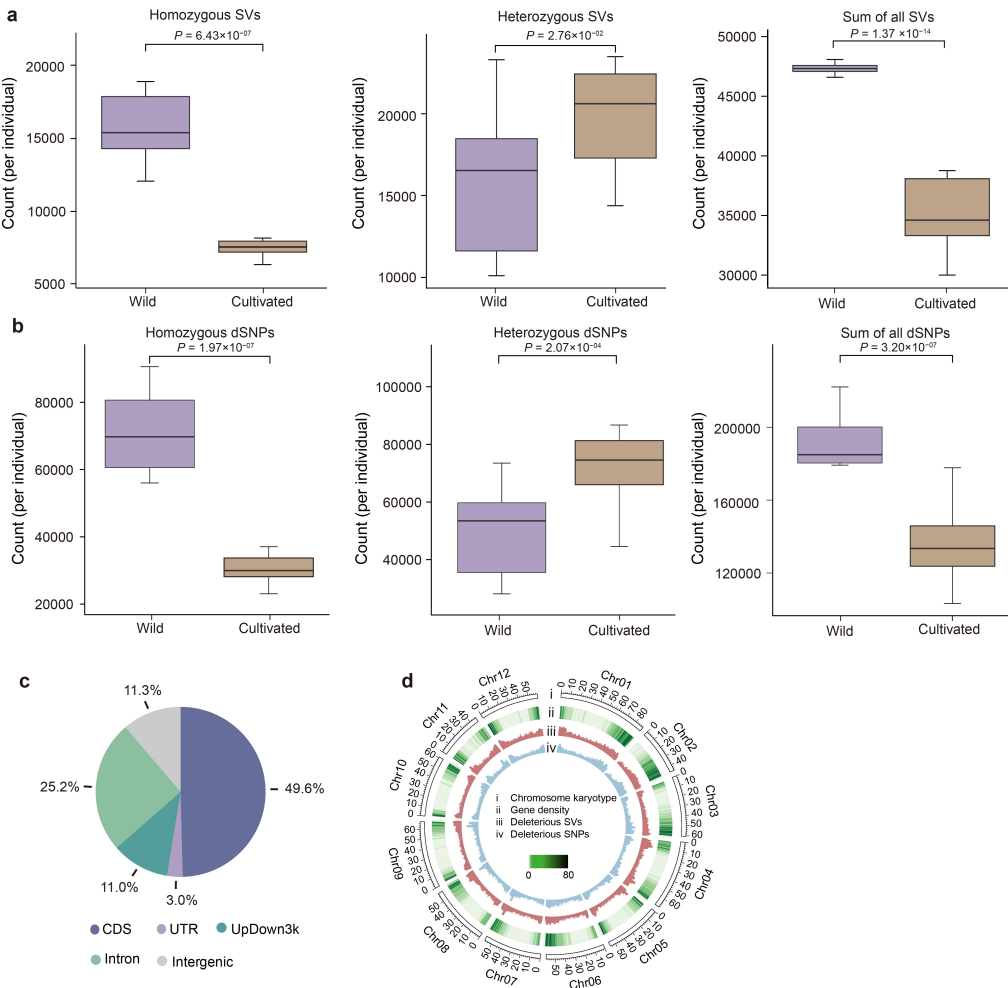


Figure 20: The characteristics of structural variations (SVs), deleterious SNPs (dSNPs) and deleterious SVs (dSVs) in the potato genome.

a, The number of homozygous SVs, heterozygous SVs and the sum of all SVs (twice the number of homozygous SVs plus the number of heterozygous SVs) across our panel of

potato genomes. b, The number of homozygous dSNPs, heterozygous dSNPs and the sum of all dSNPs across our panel of potato genomes. For a and b, the boxes represent 75% and 25% quartiles, the central line indicates the median and the whiskers extend to 1.5 times the interquartile range. P values are provided in each panel representing comparisons between wild ($n = 10$) and cultivated potato samples ($n = 19$), using two-tailed Student's t-tests. c, Genome architecture of dSVs across the potato genome. d, The distribution of dSVs and dSNPs across the potato genome.

Table 5: The distribution of heterozygous SVs in the potato genome

Chromosome	Gene body and putative promoter regions				Downstream (3K)	Intergenic	No. of heterozygous SVs
	CDS	Intron	Upstream (3K)	Sum			
01	1,024	2,354	1,345	4,723	1,287	5,586	11,596
02	564	1,460	870	2,894	906	2,413	6,213
03	587	1,518	897	3,002	898	3,046	6,946
04	818	1,789	977	3,584	1,013	5,235	9,832
05	510	1,219	668	2,397	687	3,917	7,001
06	644	1,484	851	2,979	881	3,491	7,351
07	490	1,199	740	2,429	673	3,526	6,628
08	591	1,429	727	2,747	777	3,728	7,252
09	578	1,289	744	2,611	714	4,272	7,597
10	521	1,169	678	2,368	678	3,988	7,034
11	582	1,422	654	2,658	696	2,400	5,754
12	693	1,491	833	3,017	800	2,773	6,590
Proportion	8.5%	19.8%	11.1%	39.4%	11.1%	49.4%	100.0%

The presence of deleterious SVs (dSVs) in heterozygous state could lead to strong inbreeding depression upon forced inbreeding, thus hindering the development of inbred lines, with important implications for hybrid potato breeding (Zhang et al., 2021; Zhang et al., 2019 and Bethke et al., 2022). Inferred purifying selection against SVs is stronger than against SNPs, implying that SV frequencies can be useful for an initial assessment of putative selection pressure (Mezmouk et al., 2014 and Li et al., 2010). Recently, deep phylogenomic analyses of 92 species in the family Solanaceae assessed genome-wide evolutionary constraints (with highly constrained regions amounting to 31.7 Mb under the threshold $GERP \geq 2.0$, including nonsynonymous, synonymous and noncoding sites), forming the basis for identifying and quantifying dSNPs at genome-wide scales in potato (Wu et al., 2023). Analogous to dSNP identification, we developed an unweighted approach to infer putative dSVs in the potato genome based on the joint criteria of evolutionarily derived state, low allele frequency, and being within gene-coding or evolutionarily constrained regions (**Figure 21a**, Methods). In total, we characterized 19,625 dSVs, 50.4% of which are located within non-coding regions (**Table 6**). For example, we identified a 51.1-kb

deleterious deletion occurring in only one haplotype of cultivar PG6090. This deletion leads to truncated proteins or gene loss and is associated with three genes (**Figure 22**).

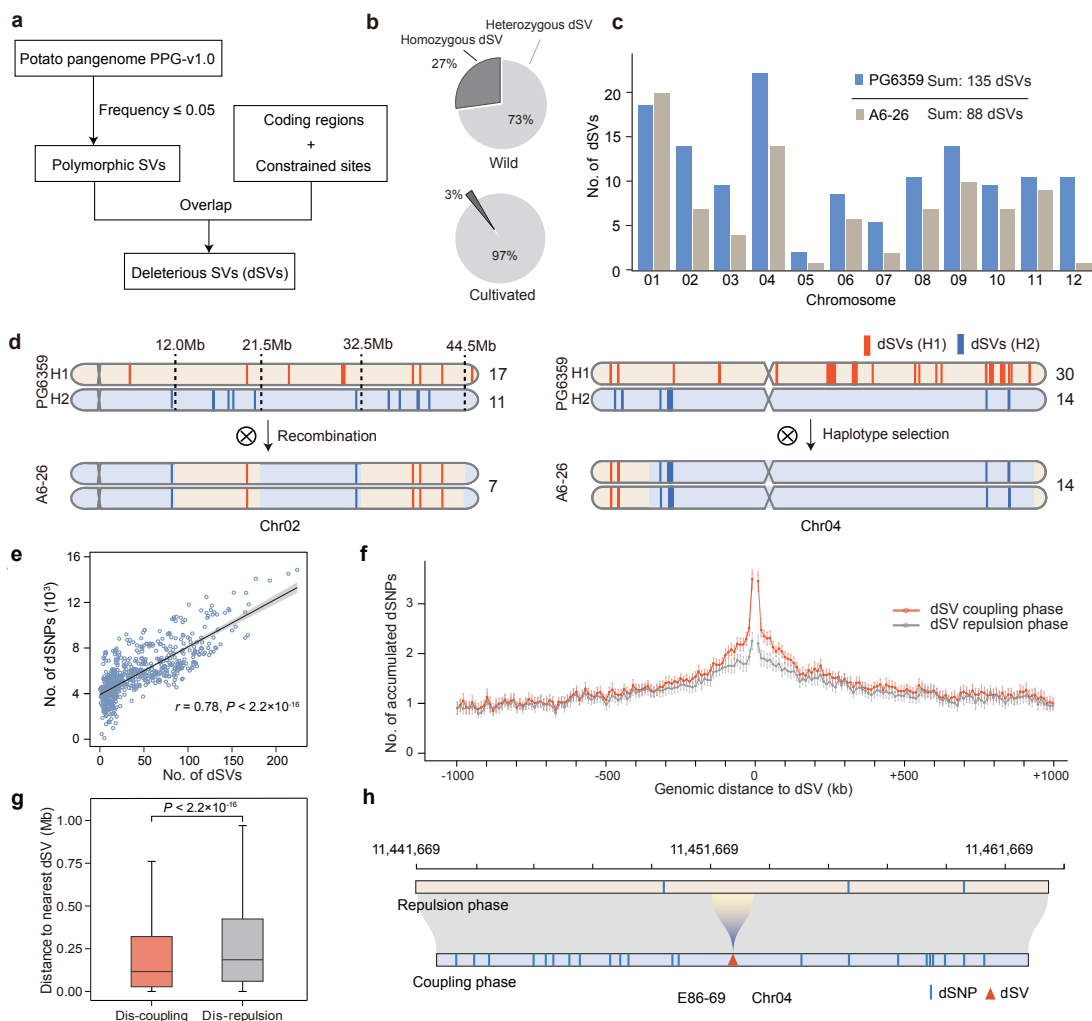


Figure 21: Fate of deleterious SVs (dSVs) in domestication and inbred-line development.

a, The pipeline for identifying dSVs includes the joint criteria of derived state of low-frequency SVs ($p \leq 0.05$) in gene-coding and/or evolutionarily constrained regions. b, The percentage of homozygous and heterozygous dSVs in wild and cultivated potato populations. c, The number of dSVs per chromosome in the founder PG6359 (average of the two haplotypes) and its derivative inbred line A6-26. d, A schematic diagram of recombination events in chromosome 2 (left) and haplotype selection in chromosome 4 (right). The potential recombination events are marked with dashed lines. Deleterious variants from the

H1 and H2 haplotypes are shown in different colors. e, Pearson correlation coefficient ($r = 0.78$) and P value (2.6×10^{-16}) between numbers of dSNPs and dSVs across all chromosomes ($n = 696$). The shaded area represents the bootstrapped 95% confidence interval. f, Distribution of the number of dSNPs flanking focal dSVs for 10-kb non-overlapping windows in the dSV coupling phase (orange) and the dSV repulsion phase (grey). The points represent means of 58 haplotypes and error bars represent 95% confidence intervals. g, Average distance of a focal dSV to the nearest dSV in coupling phase (same haplotype; Dis-coupling) and to the nearest dSV in syntenic regions of the other haplotype (Dis-repulsion). Only distances < 1 Mb are included. The boxes represent 75% and 25% quartiles, the central line indicates the median and the whiskers extend to 1.5 times the interquartile range; $P < 2.2 \times 10^{-16}$, two-tailed Student's t-test, $n = 13,717$ (Dis-coupling) and 11,679 (Dis-repulsion). h, A zoomed-in view of biased dSNP accumulation (blue bars) around a heterozygous dSV (red triangle) within a 20-kb segment of chromosome 4 in E86-69 H2.

Table 6: The distribution of dSVs in the potato genome.

Chromosome	Coding regions	Non-coding regions				No. of dSVs
		UTR	UpDown3k	Intros	Intergenic	
01	1,266	77	310	644	274	2,571
02	756	76	255	468	213	1,768
03	713	52	221	477	218	1,681
04	1,118	64	190	501	245	2,118
05	694	42	158	366	170	1,430
06	704	49	183	399	175	1,510
07	625	41	151	378	196	1,391
08	747	45	172	379	165	1,508
09	765	35	126	307	163	1,396
10	677	35	123	298	118	1,251
11	754	35	129	428	153	1,499
12	916	31	136	295	124	1,502
Sum	9,735	582	2,154	4,940	2,214	19,625

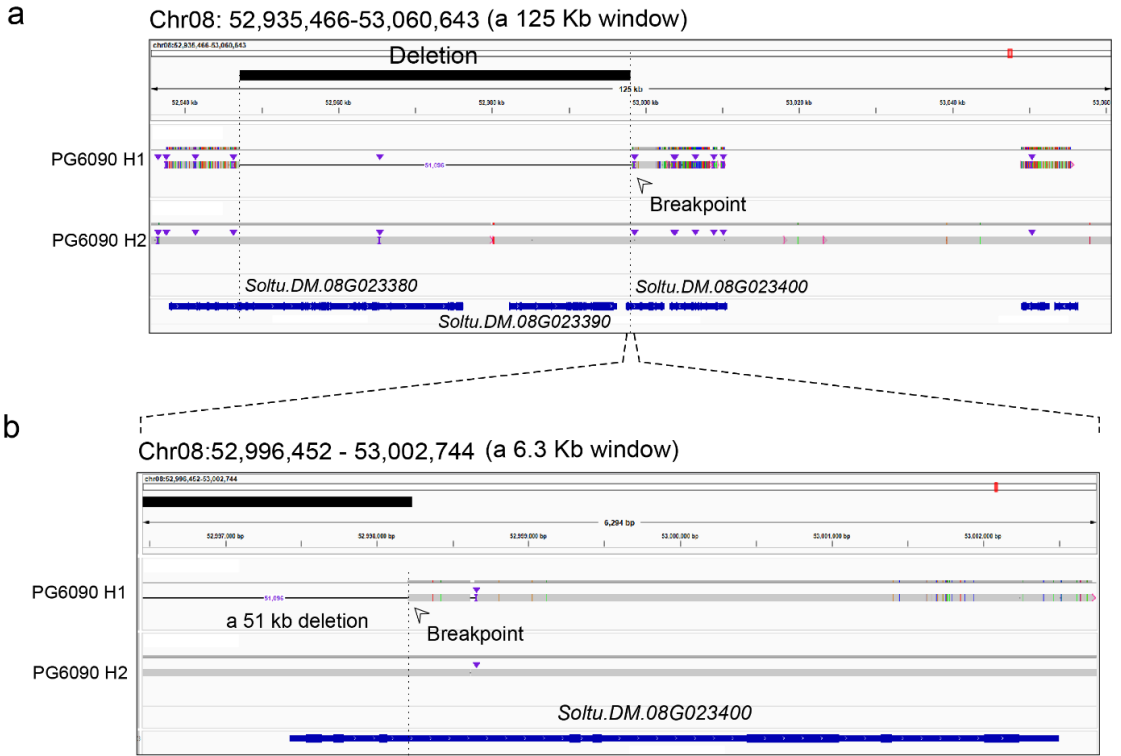


Figure 22: A 51 kb deleterious deletion eliminated three genes in potato cultivar PG6090 genome.

- a, A heterozygous 51 kb deletion is present in the cultivar PG6090 haplotype H1, involving the loss of function of three genes: *Soltu.DM.08G023380*, *Soltu.DM.08G023390*, and *Soltu.DM.08G023400*. b, The zoomed-in deletion and gene structure (*Soltu.DM.08G023400*). *Soltu.DM.08G023400* is a highly conserved single-copy gene encoding a member of endoribonucleases involved in RNA biosynthesis. Haplotype 1: H1; Haplotype 2: H2.

Excluding the inbred lines, we found that each diploid potato individual harbors an average of 843 dSVs, comprising 23.1 Mb of genomic sequence. Our analyses revealed that 73% of dSVs are in heterozygous state across wild potatoes, while this proportion increases to 97% in cultivated lines (**Figure 21b**). These patterns indicate that dSVs tend to be sheltered in heterozygous state, potentially largely avoiding negative selection on them. Hence, these heterozygous dSVs require attention from breeders when choosing suitable materials for breeding.

Purging dSVs is crucial for developing elite inbred lines in hybrid potato breeding. We established that the available inbred lines, A6-26 (88) and E4-63 (303), carry fewer dSVs compared to the average haplotypes from the two parents, PG6359 (135) and E86-69 (351), respectively (**Figure 21c**). To investigate the process of purging dSVs at the haplotype level, we constructed a genome-wide recombination map based on inbred lines and their respective parental haplotypes (**Figure 23**). Our results track at least 16 recombination events in one of the inbred lines across five generations of self-fertilization. For example, the two haplotypes from accession PG6359 carry 17 and 11 dSVs on chromosome 2; recombination events decreased the number of dSVs to seven in the derivative inbred line A6-26. For chromosome 4 of PG6359 H2, which carries fewer dSVs than PG6359 H1 (14 vs. 30), 85.2% are retained in the inbred line A6-26 (chr04: 7.5-59.5 Mb; 14 dSVs) (**Figure 21d**).

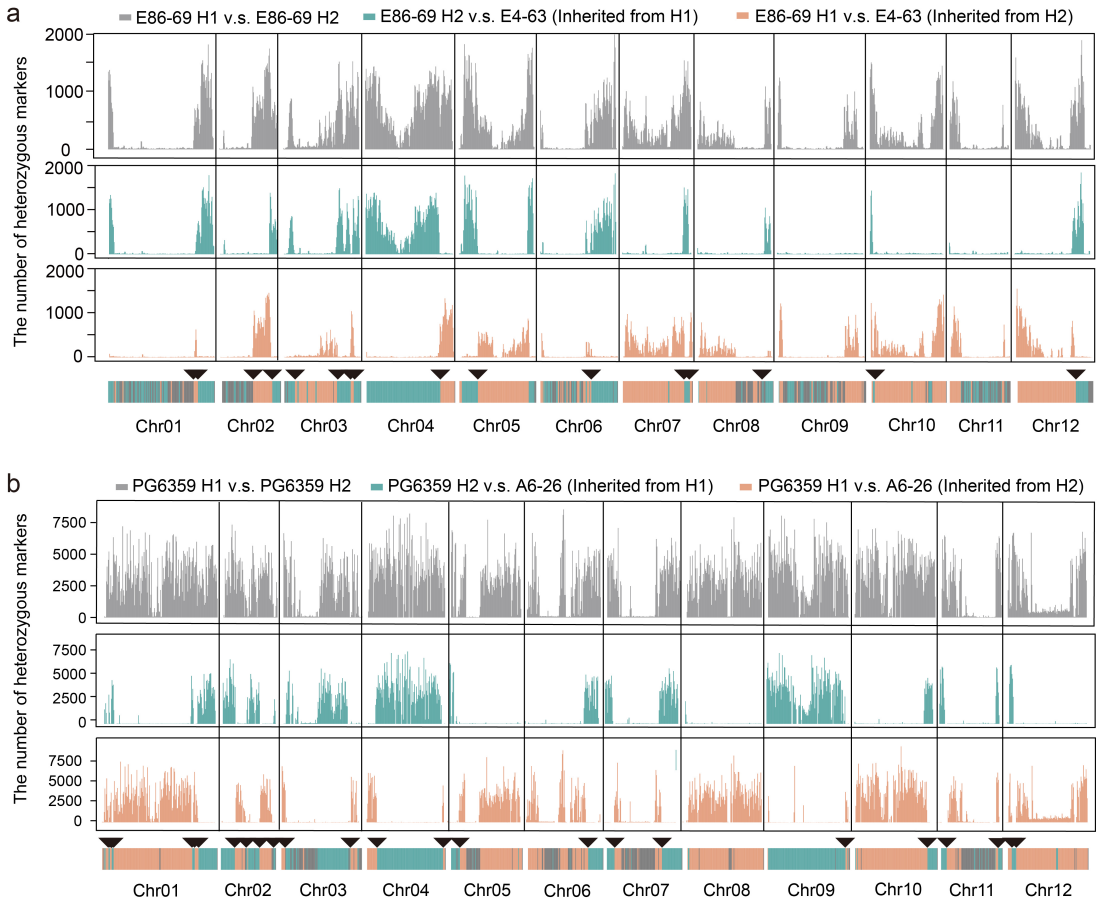


Figure 23: The recombination map of two inbred lines.

a, Recombination map of the parent E86-69 and its inbred offspring E4-63. The x-axis represents the chromosomes. The y-axis represents the number of heterozygous markers (SNPs and InDels). Top: heterozygosity was identified by comparing E86-69 haplotype 1 (H1) and E86-69 Haplotype 2 (H2); Middle: heterozygosity was identified by comparing E86-69 H2 and E4-63; Bottom: Heterozygosity was identified by comparing E86-69 H1 and E4-63. The recombination map and events (solid triangles) were inferred based on the inheritance pattern of heterozygous variants between the parental (E86-69 H1 and H2) and their inbred offspring haplotypes (E4-63). b, Analogous recombination map of the parental line PG6359 and its inbred offspring A6-26.

2.5 The ‘broken-window’ effect of dSVs

Based on the genome-wide map of deleterious variants, we discovered a significant positive correlation between the numbers of dSVs and dSNPs per haplotype ($r=0.78$, $P < 2.2 \times 10^{-16}$, Pearson correlation coefficient, **Figure 21e**). In diploid genomes, we define the coupling phase as the haplotype containing a focal dSV, while the repulsion phase refers to the other haplotype. To assess the potential impact of dSVs on dSNPs, we compared the number of dSNPs surrounding dSVs in the coupling phase to the number of dSNPs in the syntenic region of the repulsion phase. We partitioned the genome into intervals based on dSV distribution across both haplotypes and calculated statistical significance to identify impacted genomic regions (**Figure 24a**; see Methods). Our results suggest that dSNPs are significantly more frequently present in dSV coupling phase than in the dSV repulsion phase (**Figure 21f** and **Figure 24b**). This accumulation signal of dSNPs tends to be more pronounced in close proximity to dSVs.

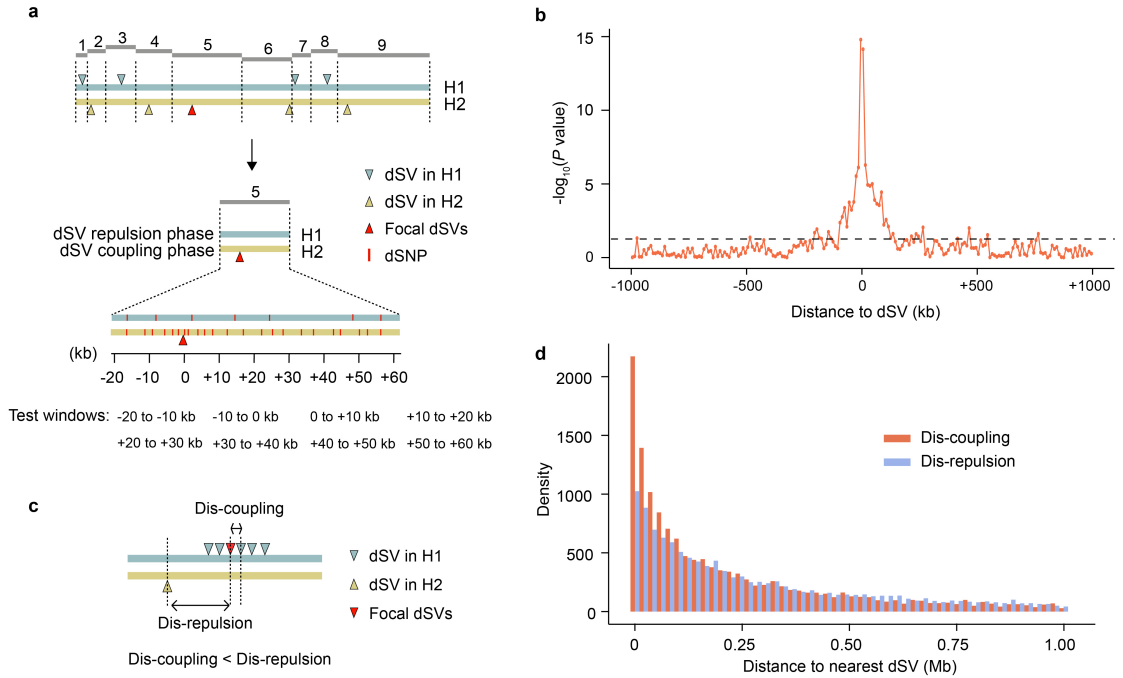


Figure 24: Statistical significance of deleterious SNP (dSNP) enrichment in the deleterious SV (dSV) coupling phase.

a, Diagram depicting our dSNP accumulation assessment. For calculation purposes, chromosomes were segmented into fragments based on the midpoint between adjacent dSVs (irrespective whether in coupling or repulsion phase). Statistical tests were performed on each 10-kb non-overlapping window within each fragment. b, Statistical significance for the number of dSNPs flanking dSVs using the 10-kb non-overlapping windows between dSV coupling and repulsion phases. The significance threshold of $P = 0.05$ (two-sided Student's t-test) is indicated. c, Diagram illustrating the calculation of the distance of a focal dSV to the nearest dSV in coupling phase (Dis-coupling) and to the nearest dSV in repulsion phase (Dis-repulsion). The focal dSV is highlighted in red. d, The frequency distribution of Dis-coupling and Dis-repulsion in potato genomes.

Furthermore, we investigated whether dSVs form clusters by calculating two types of distances: the distance between the focal dSV to the nearest dSV in coupling phase (same haplotype; Dis-coupling), and its distance to the nearest dSV in repulsion phase, (Dis-repulsion) (**Figure 24c**). Dis-repulsion is significantly larger than Dis-coupling (**Figure 21g**, 264.1 kb versus 202.1 kb, $P < 2.2 \times 10^{-16}$, two-sided Student's t-test), suggesting that dSVs tend to form clusters within the same haplotype (**Figure 24d**). We provide an illustrative example of this phenomenon using the diploid potato line E86-69.

The presence of dSVs, enhancing the occurrence of dSVs and dSNPs in the same phase, is analogous to a sociological hypothesis called the ‘broken-window’ effect (Harcourt et al., 1998). Therefore, we refer to this pattern as the ‘broken window’ effect of dSVs. We validated this pattern by examining local assemblies, which cover the 10-kb regions flanking dSVs with an average coverage of 19 reads, as well as six long reads spanning the flanking regions. Theoretical considerations as well as empirical evidence suggest that SVs affect recombination rates of closely linked genomic regions (Morgan et al., 2017; Porubsky et al., 2022 and Rowan et al., 2019). Those associations may result from insufficient purging around the dSV due to reduced recombination, while the dSV repulsion phase tends to remain functional, thus preventing strong selection on deleterious recessive variants (Zhang et al., 2019 and Ramu et al., 2017), leading to gradual differences in deleterious burden between the two haplotypes (**Figure 21h**). The ‘broken window’ effect reveals that dSVs and dSNPs are not randomly distributed across potato genomes, as they tend to form clusters in coupling phase. Such deleterious clusters are important features of potato genomes, and they are easy to spot in our phased pangenome. We found that many deleterious variants were not efficiently purged during our previous development of two inbred lines (A6-26 and E4-63) (Zhang et al., 2021). Line A6-26 carries 88 dSVs and 61,017 dSNPs, whereas E4-63 harbors 303 dSVs and 71,200 dSNPs. These deleterious variants in inbred lines dramatically reduce their fitness (Zhang et al., 2019) and thus need to be purged as a foundation for future potato breeding.

2.6 Genome design of ideal haplotypes

Plant breeders adopted the concept of ideal plant architecture (IPA) to guide breeding of superior varieties by combining multiple desirable traits (Jiao et al., 2010). Similar to the IPA, we propose the ideal potato haplotypes (IPHs) strategy to guide breeders to develop a suite of inbred lines that are as close as possible to the ideal genotype (see Methods).

Based on PPG-v1.0, the starting point of computationally designed IPHs-v1.0 are two heterotic groups; nine cultivars from the STN and GON groups with the inbred line A6-26 (Heterotic Group A), and eight cultivars from the PHU group with the inbred line E4-63 (Heterotic Group E) (**Figure 25a,b**). In principle, recombination events can reduce genetic burden in simulated haplotype combinations (**Figure 25c,d**). Based on the recombination map of two inbred lines (**Figure 23**), we assumed no more than four recombination events per chromosome in our IPHs. The consideration of recombination breakpoints takes into account the recombination coldspots in certain regions, such as heterochromatic centromeric regions and inversions. We estimated the distribution of dSVs and dSNPs under various haplotype donors and recombination events to identify the optimal combinations of ideal potato haplotypes, IPHs-A and IPHs-E.

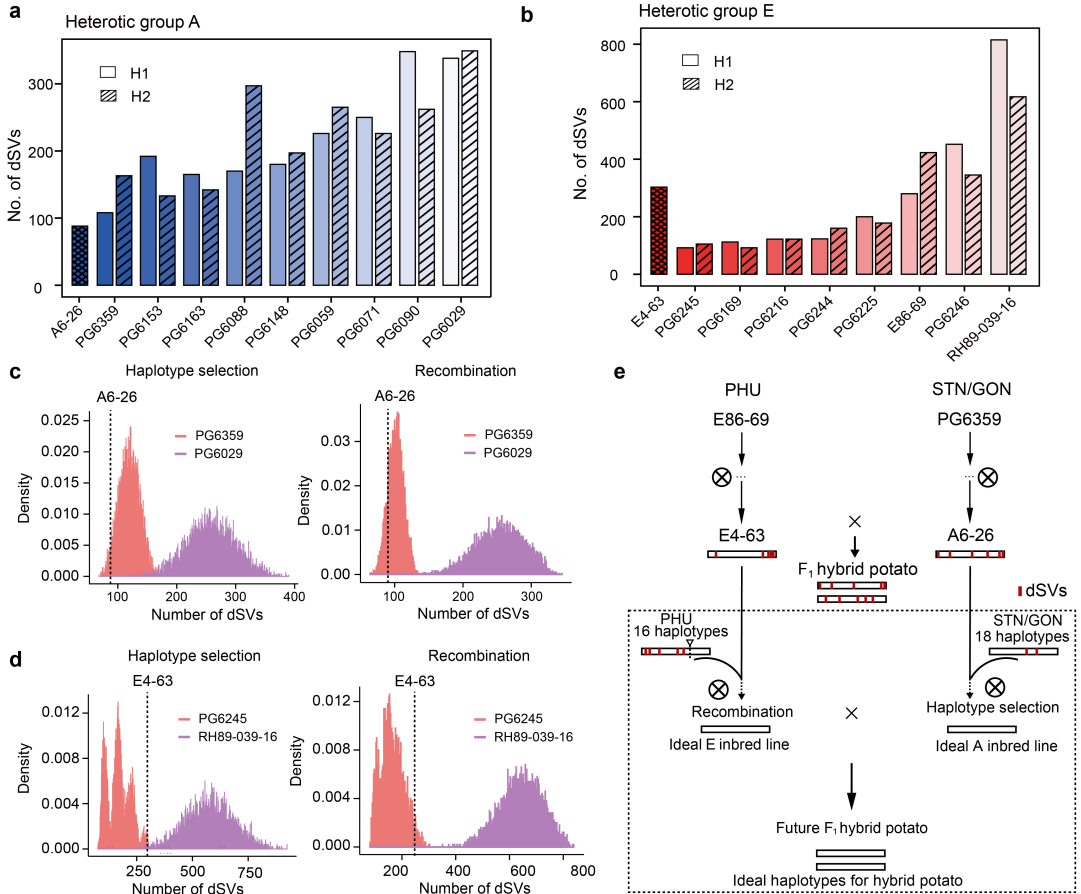


Figure 25: The schematic diagram of ideal potato haplotypes (IPHs) for inbred lines.

a, Four cultivated group are used for the simulation: *S. tuberosum* Group Phureja, PHU; *S. tuberosum* Group Stenotomum, STN; *S. tuberosum* Group Goniocalyx, GON), *S. candolleianum* (CND). The number of deleterious structural variations (dSVs) in the GON/STN group for inferring the ideal A inbred line. b, Similarly, the number of dSVs in the PHU group for designing the ideal E inbred line. c, Validating the power of recombination (simulated single recombination event) and haplotype selection in eliminating dSVs. We simulated selfing experiments and the distribution of dSVs in offspring. Two accessions (the optimal and worst-case scenarios) from heterotic group A, PG6359 and PG6029, are displayed. d. Similarly, two accessions (the optimal and worst-case scenarios) from heterotic group E, PG6245 and RH89-039-16. The simulated progenies show that recombination could dramatically reduce the number of dSVs at the haplotype level. e, Schematic diagram indicating the generation of ideal potato haplotypes (IPHs) based on two

heterotic groups including inbred lines E4-63 and A6-26. This approach prioritizes the minimization of the dSV count as its central objective. Recombination and haplotype selection could reduce the number of dSVs during potato recurrent selection.

We found that ideally, all dSVs could be removed from IPHs-A and IPHs-E and the number of dSNPs could be reduced to 41,262 and 35,360, implying a reduction of 32.4% (A6-26) and 50.3% (E4-63) compared to the previous inbred lines (**Figure 26a,b**). Haplotype distances predict that hypothetical F1 hybrid genomes between IPHs-A and -E would carry only 10,830 homozygous dSNPs (**Figure 26c**), implying a reduction of 54.5% compared to the previous F1 hybrid potato (Zhang et al., 2021). By integrating the genetic linkage map (Jiang et al., 2023), we found that achieving IPHs-A chr01 would require at least five generations of recurrent selection from hybrid progenies, and that one of the designed breakpoints is located in a low-recombination region (near chr01: 44.3 Mb), requiring a population size of at least 1,321 individuals to obtain at least one copy of IPHs-A chromosome 1. It is worth noting that relying solely on natural recombination is unlikely to be sufficient. New technologies such as targeted recombination, gene editing, and synthetic biology to precisely eliminate deleterious variants may provide future pathways for approaching IPHs (Hayes et al., 2024; Filler-Hayut et al., 2021 and Schmidt et al., 2020).

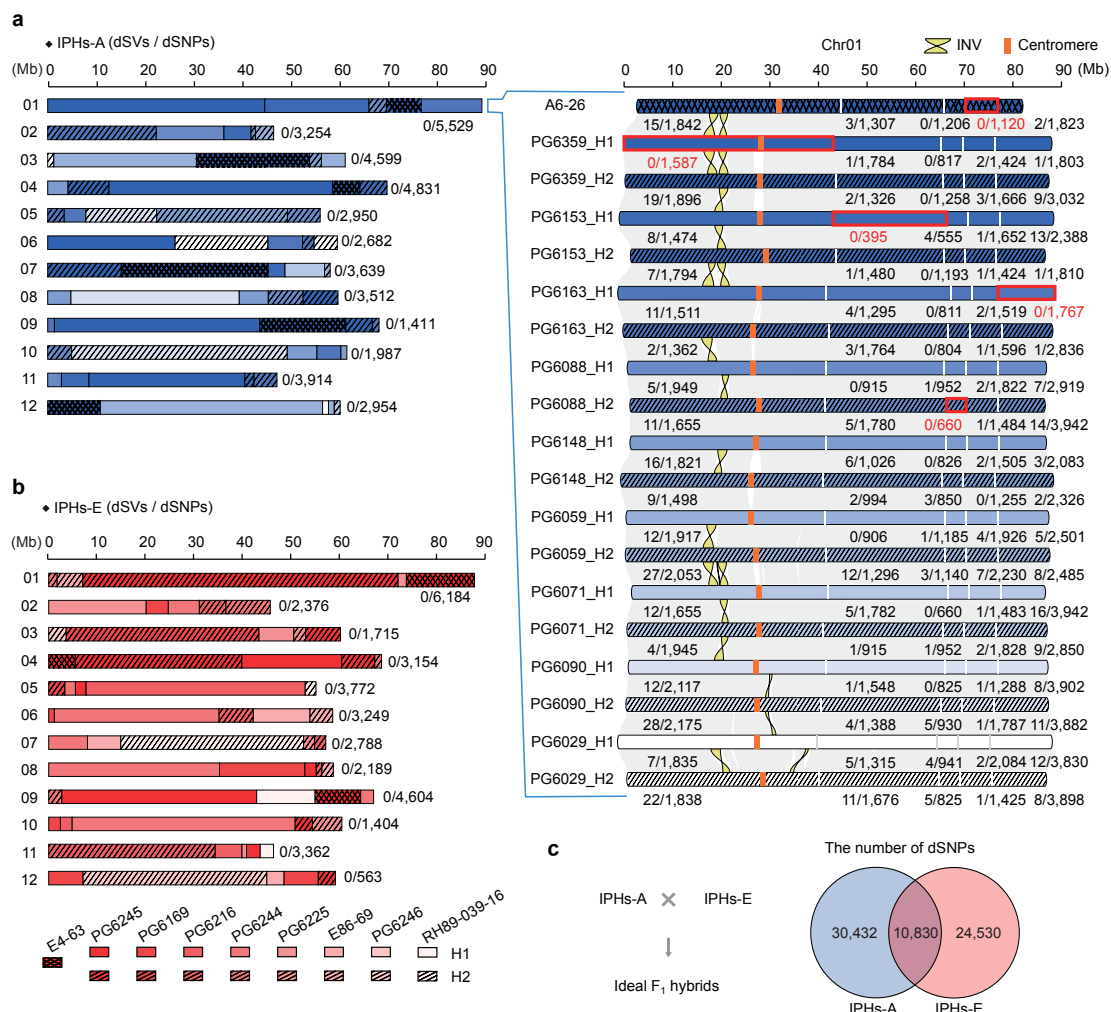


Figure 26: Ideal potato haplotypes (IPHs) designed from the phased potato pangenome graph (PPG).

a, Chimeric map of the IPHs-A genome and the most ideal graph path of 12 chromosomes from heterotic A group (left). The numbers of deleterious SVs (dSVs) and deleterious SNPs (dSNPs) for each chromosome are indicated. Haplotypes for chromosome 1 are shown in detail (right), and numbers of dSVs and dSNPs for each haplotype are represented by five parts derived from the ideal designed recombination events. The chimeric graph path of IPHs-A 'ideal' chromosome 1 is highlighted with red frames. Inversions and centromeres are colored in genomic collinear blocks and chromosomes, respectively. b, Similarly, the chimeric map of the IPHs-E genome and the ideal chimeric graph paths from heterotic E

group are presented. c, The number of dSNPs of IPHs-A and IPHs-E, the overlap indicating the number of homozygous dSNPs in the inferred ideal F1 hybrid generated from IPHs-A and IPHs-E.

3. Discussion

In this study, we developed a graph-based phased potato pangenome reference comprising 60 haplotype sequences, reducing reference bias and identifying more variants that remain unaccounted by reads-based methods (Eizenga et al., 2020 and Rakocovic et al., 2019). Our phased pangenome provides an example of a typical clonally propagated crop with a highly heterozygous genome. The integration of pedigree information and high-accuracy ultra-long reads can further improve phase accuracy in highly repetitive regions.

Consistent with observations in grapevine (Zhou et al., 2019), SVs are over-represented among singletons and other minor-allele frequency classes, suggesting that SVs are under stronger purifying selection. SVs can be associated with agronomic traits (Alonge et al., 2020; Leffler et al., 2017 and Yan et al., 2023), and by disrupting gene structure and cis-regulatory elements, dSVs can exert strong effects. We classified 14% of all SVs to be putative dSVs. However, inference of dSVs shows lower efficacy in certain cases due to overlapping SVs and ambiguous ancestral state. Therefore, quantifying their fitness penalties will be crucial to guide breeding practices in the future.

Asexually reproducing crops tend to accumulate more heterozygous deleterious mutations (Zhou et al., 2019 and Ramu et al., 2017). The ‘broken-window’ effect of dSVs may be reinforced due to very limited recombination under long-term clonal propagation. After one haplotype is hit by a major-effect deleterious mutation, it seems less likely that deleterious mutations accumulate in the dSV repulsion phase, or else the affected individual might not survive to reproduce. This inference underscores the importance of preserving at least one functional copy of genes to maintain genetic robustness (Conrad et al., 2010 and Xue et al., 2013). Eradicating dSVs is therefore a new focus in the development of inbred lines to generate diverse hybrid potato combinations, a principle that should also be relevant for other clonally propagated crops.

To develop inbred lines, it is more favorable to start from haplotypes of diploid landraces with a lower deleterious burden (Wu et al., 2023). The relatively high deleterious burden per haplotype in a dihaploid generated from tetraploid varieties poses formidable obstacles for this step. Therefore, we plan to develop the first generation of inbred lines from diploid landraces concurrently with inducing dihaploids from elite tetraploid varieties. The IPHs strategy provides a blueprint for the combination of haplotypes to reduce the negative impact of deleterious variants. While achieving IPHs at the chromosome level through crossing is possible, it is challenging to combine the ideal chromosomes into one genome without additional recombination. Currently, the primary value of IPHs lies in their ability to outline an

optimal goal for haplotype recombination, allowing breeders to compare real haplotypes, including dSVs and dSNPs, against an idealized standard, thus facilitating refinement of breeding strategies.

4. Materials and Methods

4.1 Sequencing Materials

We selected 31 potato accessions based on phylogenetic trees generated by previous studies. Of these accessions, 10 are from *S. candolleianum* and 21 are from three groups of diploid potato cultivars. Among these potato cultivars, nine accessions represent *S. tuberosum* Group Phureja (including inbred line E4-63), eight accessions represent *S. tuberosum* Group Stenotomum (including inbred line A6-26), and four are from *S. tuberosum* Group Goniocalyx. Previous high-fidelity (HiFi) data are available for 21 accessions and previous Hi-C reads are available for five accessions (PRJNA754534). Additionally, we newly sequenced 10 accessions using HiFi and 25 accessions using Hi-C in this study. HiFi reads for 10 accessions were generated using the ccs program version 6.4.0 (<https://github.com/PacificBiosciences/ccs>) and subreads obtained from the Pacific Biosciences Sequel II platform, which were then converted to FASTQ format by SAMtools (v1.17) (Li et al., 2009). In total, we generated 318.49 Gb of HiFi data, ranging from 27.72 to 35.32 Gb per sample. For Hi-C libraries, DNA was extracted from in vitro seedlings and digested with the restriction enzyme MboI, using previously described Hi-C library preparation protocols (Burton et al., 2013 and Kaplan et al., 2013). A total of 2.70 Tb of Hi-C data were generated based on the Illumina HiSeq platform. To facilitate genome annotation, we used RNA-seq data from six plant tissues (including roots, stems, leaves, stolon, tubers and flowers) from a previous study (Tang et al., 2022).

4.2 De novo genome assembly of 60 haplotypes

For genome size and heterozygosity estimation, we used Jellyfish (v2.3.0) (Marcais et al., 2011) to obtain a frequency distribution of the k-mers and estimated the histograms by GenomeScope (v2.0) (Rannallo-Benavidez et al., 2020). We then assembled haplotype-resolved assemblies with the HiFi reads and Hi-C reads using hifiasm (<https://github.com/chhylp123/hifiasm>) (v0.16) (Cheng et al., 2021) with default parameters. Subsequently, we aligned haplotypes using the Juicer pipeline and then generated Hi-C maps for the 3D-DNA pipeline (v201008) (Dudchenko et al., 2017) with parameter “-q 0”. The assembly from two haplotypes was scaffolded and ordered using the Hi-C data-based rough scaffold and compared with the reference (accession DM1-3 516 R44) (Pham et al., 2020). Two sets of pseudo-chromosomes were constructed for each accession, and this workflow was applied to all 29 diploid potato accessions, resulting in 58 haplotypes. For the two inbred lines E4-63 and A6-26, we generated haploid assemblies. The Hi-C format file was visualized using the Juicebox program (v2.16.00) (Durand et al., 2016) and misassembled contigs were manually curated. Finally, each haplotype was integrated into the 3D-DNA pipeline based on the run-asm-pipeline-post-review.sh script. To ensure the accuracy of the

assemblies, we filtered out contigs smaller than 50 kb and used chromosome-based haplotypes for subsequent analyses.

4.3 Evaluation of the genome assemblies

The contig N50 statistic was calculated using assembly-stats (<https://github.com/sanger-pathogens/assembly-stats>). The Benchmarking Universal Single-Copy Orthologs (BUSCOs) program (v5.4.4) (Simao et al., 2015) was used to evaluate assembly quality with the embryophyte_odb10 protein database. Both switch and Hamming errors were calculated using two VCF files produced by the pipeline calc_switchErr (https://github.com/tangerzhang/calc_switchErr), based on the ‘compare’ function of WhatsHap (v1.1) (Martin et al., 2017). Genome completeness was estimated via the KAT (v2.4.2) (Mapleson et al., 2017) ‘comp’ command. Reliability of our assemblies was estimated by Flagger, a reads-mapping-based pipeline optimized for diploid assemblies (Liao et al., 2023). A detailed workflow and commands can be found in the GitHub (<https://github.com/Chenglin20170390/Haplotype-diversity>).

4.4 Annotation of repetitive elements

Different repeat elements are identified based on 60 haplotypes, including transposable elements, segmental duplications, tandem repeats and satellite repeats. To identify transposable elements, we employed the Extensive de novo TE Annotator (EDTA) (v2.1.0) (Ou et al., 2019), including long terminal repeat retrotransposons (LTR), DNA transposons with terminal inverted repeat (TIR) sequences, and helitron transposons. For the remaining transposon elements, RepeatModeler (Flynn et al., 2020) was used to search for a second round. To search for large segmental duplications (>1 kb), we employed Asgart (v2.4.0) (Delehelle et al., 2018) using a kmer-based method with the parameter “-CRSV”. Tandem repeats were identified using Tandem Repeats Finder (v4.09.1) (Benson et al., 1999) with default parameters. Satellite repeats were identified using Satellite Repeat Finder (SRF, commit e54ca8c, <https://github.com/lh3/srf>) with HiFi reads.

4.5 Prediction of protein-coding genes

A comprehensive strategy consisting of transcript evidence, ab initio prediction, and homology alignment was applied for gene prediction. First, we aligned RNA-seq reads to assembled haplotypes using HISAT2 (v2.2.1) (Kim et al., 2015) with the “-dta” parameter and then assembled by StringTie (v2.2.1) (Pertea et al., 2015) with the “-rf” parameter. Subsequently, we used the BRAKER2 (v2.1.5) (Bruna et al., 2021) program to train the ab initio prediction model from AUGUSTUS (v3.4.0) (Stanke et al., 2006) (<https://github.com/Gaius-Augustus/Augustus>) and collected high-quality RNA-Seq hints using the Hidden Markov Model (HMM) from GeneMark-ET (v3.67) (Lukashin et al., 19898) with the parameter “--nucleanup --softmasking”. To improve gene structure prediction, we performed a homology search using a curated plant protein dataset downloaded from the UniProt Swiss-Prot database (<https://www.uniprot.org/downloads>). We merged these homologous proteins with previously published peptides from tomato (The tomato genome consortium, 2012) and potato and eliminated potential redundancy using the CD-HIT-est (v4.6.8) (Li et

al., 2006) program with default parameters. The MAKER2 (v3.01.03) (Holt et al., 2011) program was used to combine the homology search, expression evidence and ab initio prediction through two rounds. Finally, we used the Mikado (v2.3.4) (Venturini et al., 2018) program to identify the representative set of transcripts from transcript assemblies, before those were fed to the PASA pipeline (v2.5.1) (Haas et al., 2003) to update gene structures.

To perform functional gene annotation, we utilized the InterProScan (v5.34-73.0) (Jones et al., 2014) program, which identifies potential protein domains and Gene Ontology (GO) terms based on sequence signatures.

4.6 Construction of the potato pangenome

The Minigraph-Cactus (MC) pipeline²⁷ and PanGenome Graph Builder (PGGB) (Garrison et al., 2024) were used to construct a pseudo-phased pangenome with all 61 haplotypes based on the whole genome alignment (including the DMv6.1 reference genome). For the PGGB, we estimated the divergence of each chromosome with mash distances and confirmed chromosome community with wfmash (Marco-Sola et al., 2021) mapping. Then, we used “pggb -s 10000 -n 61 -p 90 -k 47 -P asm20 -O 0.001” to build each chromosome graph. We visualized the graph 1D layout and estimated presence and absence ratios to the DMv6.1 reference in 100-kb sliding windows using ODGI (Guarracino et al., 2022). The small variants and SVs were detected by vg deconstruct from snarls, and we only kept top-level and <1-Mb variants with vcfbub. For the MC pipeline, we assigned DMv6.1 as the guide for the paths, and progressively aligned the 60 haplotypes to it. We used the cactus-pangenome script with parameters “--gfa full --gbz full --vcfReference DMv6.1” to generate complete workflows and execute commands. The generated Graph Fragment Assembly (GFA) format graph was used for edge, node and coverage statistics and subgraph generation from a BED input.

4.7 Pangenome size and growth

To fully capture the genome diversity of our potato populations, we used the Panacus (Parmigiani et al., 2024) to assess pangenome size and growth ratio. It estimates pangenome openness directly by applying the binomial formula. We calculated cumulative bases based on quorum (minimum fraction of haplotypes sharing a graph feature after haplotypes are sequentially added to the growth histogram), and proportion of conserved ($\geq 90\%$ of haplotypes) and variable sequences ($< 90\%$ of haplotypes) in the pangenome graphs. These statistics were obtained using Panacus hist and growth scripts with parameters “-c bp -l 1,1,1 -q 0,0.1,0.5,0.9 -S”, and we selected different haplotype paths for group-specific statistics.

4.8 Phylogeny and syntenic analysis

To build haplotype phylogenetic trees at the chromosome level, annotated protein sequences from the 60 newly assembled haplotypes and three published genomes (outgroups: *S. wrightii* (Wu et al., 2023); *S. etuberosum* (Tang et al., 2022) and *S. lycopersicum*) (Zhou et al., 2020) were aligned to produce all-versus-all alignments using Diamond (v0.9.21) (Buchfink et al., 2015). The gene families were inferred

using the OrthoFinder (v2.5.4) (Emms et al., 2019) program, which employs the Markov cluster (MCL) algorithm. Separately for each chromosome pair, all single-copy orthologous protein sequences were merged into a single FASTA file, which was then fed into IQ-TREE (v2.0.6) (Nguyen et al., 2015) using the maximum-likelihood method. Each chromosome tree was visualized with the function `ggtree` (v3.0.2) (Yu et al., 2017) in R (v.4.2.0). Synteny plots and pan-genome annotations were generated by the R package Genespace (v0.9) (Lovell et al., 2022).

4.9 Identification of Single Nucleotide Polymorphisms (SNPs), InDels and SVs

Assembly-based calling was used to detect SVs with a minimum length of 50 bp. Genome sequences were aligned to the DMv6.1 reference genome to produce alignment BAM files using Minimap2 (v2.17) (Li et al., 2021) with “-ax asm5”. SyRI (v1.5) (Goel et al., 2019) was then utilized to call genome-wide variants and generate assemblies-based VCFs. We kept variants of INS, DEL, INV, and DUP from the output of SyRI as the individual SV dataset. The accuracy of the SV dataset was validated by randomly selecting and manually verifying 100 SVs longer than 100 kb using Hi-C maps. Finally, we merged the individual SVs based on 80% overlap using SURVIVOR with the parameters “0.8 1 1 -1 -1 -1” to produce the final population SVs file. For SNPs and InDels, we merged the output of assembly-based variation calling from SyRI using BCFtools (v1.13) (Danecek et al., 2021).

4.10 Haplotype diversity and genome heterozygosity by sequence (GHSL) statistic

The length of the haplotype-specific (heterozygous) variants for each diploid accession was calculated by summing all variants present in only one haplotype. In detail, we aligned the two haplotypes of one accession to each other using Minimap2 (v 2.17), identified variants by SyRI (v1.5), and then calculated the number of haplotype-specific SNPs and the length of haplotype-specific variants (InDels and SVs). The Genome Heterozygosity by Sequence Length (GHSL) approach considers reference bias and regions potentially impacted by haplotype-specific variants (<https://github.com/Chenglin20170390/Haplotype-diversity>). Principal component analysis (PCA) based on haplotypes (12 chromosomes) was performed using Plink (v1.90) (Purcell et al., 2007) on the VCF file with the parameter “--pca 5”; the results were visualized using the R package ggplot2 (Wickham et al., 2011). Pairwise Jaccard similarities between haplotypes were estimated using BEDTools (v2.30.0) with the “jaccard” command.

4.11 SV dynamics in potato domestication

To calculate the unfolded site frequency spectrum (SFS), we analyzed 19 cultivated potato accessions (excluding the two inbred lines) from VCFs. The genome of *Solanum lycopersicum* (SL 5.0) (Zhou et al., 2022) was used as an outgroup to infer the ancestral states of genotypes. We distinguish different types of SNPs (synonymous and non-synonymous) and SVs (deletions, insertions, inversions, and duplications) and counted the number of haplotypes carrying the derived variants. For each accession, we also calculated the incidences (numbers) of heterozygous and

homozygous SVs and the sums of all SVs (twice the number of homozygous SVs plus the number of heterozygous SVs).

4.12 Identification of TEMRs

SV breakpoints with additional ± 100 -bp flanking sequences were used to identify repeat elements (transposable elements, segmental duplications, and tandem repeats) associated with SVs. To detect the potentially causal relationship between TE movement and SV formation, the documented genomic position of a focal SV and overlapping TEs determine the following categories: ‘No TE SV’ for SVs without any TE overlap, ‘Incomplete TE SVs’ for SVs overlapping with a single TE with $<95\%$ coverage, ‘Single TE SVs’ for SVs overlapping with a single TE with coverage $\geq 95\%$, and ‘Multi TE SVs’ for SVs overlapping with multiple TEs with coverage $>95\%$. Since no single TE SV events were considered, insertions were excluded. To quantify the formation of SVs (deletions, inversions and duplications) by ectopic recombination via TEMRs, SV breakpoints (± 100 -bp) overlapping with the identical class of TEs were calculated. To assess the distribution of repeats in SV breakpoints, we randomly simulated SVs using the shuffle function of BEDTools with the number and length of SVs set according to the potato SV variation map. Two-sided Student’s t-tests were used to test the significance of the proportions between observed and simulated TE-related SVs. To infer the insertion times of TEMRs, we extracted insertion times from the pass.list of EDTA and kept intact TE sequences overlapping with the breakpoints of SVs. Sequence homology of TEMRs was calculated using 200-bp flanking sequences based on global pairwise sequence alignment from Needle (https://github.com/nanjiangshu/my_needle).

4.13 Haplotype recombination in inbred lines

The fifth-generation inbred lines A6-26 and E4-63 were generated from the diploid cultivars E86-69 and PG6359, respectively. To evaluate haplotype recombination events in each inbred line, we utilized two different methods: heterozygosity analysis and haplotype-specific k-mer analysis. Genome-wide heterozygous peaks were identified by aligning the sequences of the haplotypes. The different heterozygous peaks were compared between haplotypes from inbred lines and their founders. The regions inherited from the founders’ haplotypes in the alignment would be in homozygous state. These switch signals from homozygosity to heterozygosity were used to characterize putative recombination events. Haplotype-specific k-mers were calculated by the number of specific k-mers from the inbred line based on non-overlapping 500-kb regions and compared with parental haplotypes using the Meryl program (available at <https://github.com/marbl/meryl>).

4.14 Identification of dSVs and dSNPs

Putative dSVs were identified based on three criteria: (1) ancestral versus derived state of SVs. *Solanum lycopersicum* (SL 5.0) (Zhou et al., 2022) was used as the outgroup for ancestral state inference, as phylogenetic analyses have placed this species in a clade relatively close to *Solanum* section *Petota* (Tang et al., 2022). (2) SVs with frequencies below 0.05 in our 60 haplotypes were considered, as most deleterious variants occur at low population frequencies (Chun et al., 2009; Mezouk

et al., 2014 and Li et al., 2010). (3) SVs in coding regions (CDS) that may damage proteins' function or SVs in the 92 Solanaceae evolutionarily constrained regions are considered to be deleterious (Wu et al., 2023). We thus identified putative dSVs as being derived-state SVs with low-frequency (<0.05) that overlap with CDS regions or evolutionarily constrained regions; we implemented this with the BEDTools 'overlap' command to infer putative dSVs. Putative dSNPs were identified by showing GERP values >2.75 and overlapping with evolutionarily constrained regions (Wu et al., 2023).

4.15 Estimation of the 'broken window' effect

The Pearson correlation coefficient was used to examine the correlation between the number of dSNPs and dSVs per chromosome. Focal dSVs located in haplotype 1 (H1) indicates H1 was considered to be the dSV coupling phase and corresponding syntenic regions of H2 were considered to be the dSV repulsion phase. We calculated the distance of the focal dSV to the nearest dSV in coupling phase and to the nearest dSV in repulsion phase. To exclude the potential concurrent influence of dSVs on both coupling and repulsion phases, we segmented chromosomes into fragments based on the midpoint between adjacent dSVs (irrespective whether in coupling or in repulsion phase). Regions that may be affected by the 'broken window' effect were estimated based on 10-kb non-overlapping sliding windows. For each sliding window, we calculated the number of dSNPs flanking each dSV in the coupling and repulsion phases. To assess dSNP enrichment, for each sliding window we performed a two-tailed Student's t-test for the numbers of associated dSNPs in coupling and repulsion phase, respectively (regions significantly affected by the 'broken-window' effect of dSVs). The depth of reads near the dSV is calculated using the SAMtools (v1.17) 'depth' command. Finally, we employed SafFire (<https://github.com/mrvollger/SafFire>) to visualize schematic representations of dSVs and dSNPs within haplotypes.

4.16 The definition and construction of ideal potato haplotypes (IPHs).

IPHs offer a strategy to design ideal haplotypes in potato inbred lines with minimized dSNPs and without dSVs. The central focus is constructing chimeric haplotypes in inbred lines A and E from two heterotic groups (heterotic group A and heterotic group E, respectively). According to the haplotype phylogenetic trees and the F1 hybrid potato generated from inbred lines E4-63 and A6-26, heterotic group A comprises nine cultivars (PG6359, PG6153, PG6163, PG6088, PG6148, PG6059, PG6071, PG6090 and PG6029 from the STN and GON groups) as well as inbred line A6-26; these lines are considered as the source material to design the ideal inbred line A (IPHs-A). Correspondingly, eight cultivars (PG6245, PG6169, PG6216, PG6244, PG6225, E86-69, PG6246, and RH89-039-16) and inbred line E4-63 from heterotic group E were used to design the ideal inbred line E (IPHs-E).

In principle, recombination can yield new mosaic-type haplotypes with reduced deleterious burden in both the A and E heterotic groups. Purging dSVs is the primary goal in constructing ideal inbred lines. Therefore, we split each chromosome into 50

bins. We conducted simulations to estimate the distribution of dSVs and dSNPs under various haplotype donors and recombination events for scanning the optimal combinations. We assumed that alternative haplotypes have an equal probability of inheritance after recombination. The number of deleterious variants for each chromosome was estimated using the following formula:

$$IPHsburden = \min_{1 \ll j \ll C_{50}^{n-1}} \left[\sum_{i=1}^n (P_i^j + p_i^j) \right]$$

where n represents the number of recombination fragments. The parameter j indicates recombination events in each haplotype (50 bins). The P indicates the number of dSVs (dSV burden) and p indicates the number of dSNPs (dSNP burden). Combinations with IPHs burdens smaller than those of the current A6-26 and E4-63 inbred lines are considered ideal haplotypes. The most ideal haplotype of heterotic A and E groups is the chimeric haplotype with the minimum value of IPHs burden.

In this study we assumed $n=5$ and, due to the ‘broken window’ effect, prioritized haplotypes (combinations) with $P = 0$. We collected combinations that should eliminate dSVs across the genome as an initial IPHs dataset. Subsequently, we calculated the distribution of dSNPs based on the initial IPHs dataset and excluded recombination events in centromeres and inversions. Finally, the mosaic haplotypes with lowest the number of dSNPs emerged as the final, most promising IPHs-A and IPHs-E, respectively. To describe the sheltering of heterozygous dSVs and dSNPs in F1 hybrids generated by crossing IPHs-A and IPHs-E, we calculated the number of heterozygous and homozygous dSNPs in these hypothetical hybrids. For each chromosome from IPHs-A and IPHs-E, we also simulated a hypothetical recurrent selection scenario with two breeding rounds comprising five haplotype donors (S1-S5). The first round entailed two recombination events in two pairs of haplotypes (S1-S2 and S3-S4). The S5 haplotype was introduced as a homologous haplotype. The second round entailed two recombination events to form the final mosaic haplotype with five fragments. Finally, we integrated a genetic linkage map into in silico breeding simulations for estimating the required population size for each generation of the hypothetical recurrent selection scenario.

4.17 Design an ideal potato chromosome and aggregating 12 ideal chromosomes.

Our simulations are based on the recombination rate information obtained from the genetic population, with the minimum population size being described by the formula:

$$\text{Minimum population size } n \geq \frac{\log(P \text{ value})}{\log(1 - \max P)}$$

Here, P value is tested at significance levels of 0.05, 0.01, and 0.001, with $\max P = P(\text{non-recombination}) \times P(\text{breakpoints}) \times P(\text{haplotype selection})$. P value 0.05 indicates 95% probability of obtaining at least one individual carrying the ideal recombination. Our findings suggested that the simulated cross experiments are feasible, as the majority of cross population sizes are less than 2,000. Moreover, we

explored simulations for creating an ideal individual containing all 12 ideal chromosomes.

The breeding design aims to assemble 12 pre-constructed chromosomes into a single plant through recurrent selection, forming an ideal genome without recombination of ideal chromosomes during the process. Therefore, we simulated recurrent hybridization to assemble 12 pre-constructed IPH chromosomes into a single plant across multiple rounds of hybridization. The impotence is non-recombination and maintenance of ideal chromosome combinations. We integrated 6 ideal chromosomes through three rounds of hybridization and construct an individual containing all 12 ideal chromosomes through one round of hybridization. During recurrent selection, segments L where ideal chromosomes do not recombine and the required population size n follow the functions and probability distributions outlined in the legend, guided by the formula:

$$\text{Minimum population size } n \geq \frac{\log(P \text{ value})}{\log(1-(1-i*0.01)^L)}$$

Here, P value ranges from 0.05 to 0.001, n represents the minimal population size, i denotes the recombination rate, L indicates the genomic length without recombination. For each simulated generation, we also considered the P (haplotype selection) = $\lceil 0.5 \rceil^N$, with N representing the number of selected chromosomes. The transition from a heterozygous to a homozygous genome encompassed 12 chromosomes (N=12). Population size under recurrent selection is related to recombination rate and aggregated genome length

5. References

- Alonge, M. *et al.* Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161. e23 (2020).
- Balachandran, P. *et al.* Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115 (2022).
- Bao, Z. *et al.* Genome architecture and tetrasomic inheritance of autotetraploid potato. *Mol. Plant* **15**, 1211–1226 (2022).
- Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
- Berdan, E. L., Blanckaert, A., Butlin, R. K. & Bank, C. Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLoS Genet.* **17**, e1009411 (2021).
- Bethke, P. C. *et al.* Diploid potatoes as a catalyst for change in the potato industry. *Am. J. Potato Res.* **99**, 337–357 (2022).
- Bozan, I. *et al.* Pangenome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. *Proc. Natl. Acad. Sci. USA* **120**, e2211117120 (2023).
- Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* **3**, lqaa108 (2021).
- Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
- Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
- Cho, S. K. *et al.* Polypyrimidine tract-binding proteins of potato mediate tuberization through an interaction with *StBEL5* RNA. *J. Exp. Bot.* **66**, 6835–6847 (2015).
- Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
- Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
- Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
- Delehelle, F., Cussat-Blanc, S., Alliot, J.-M., Luga, H. & Balaesque, P. ASGART: fast and parallel genome scale segmental duplications mapping. *Bioinformatics* **34**, 2708–2714 (2018).
- Domínguez, M. *et al.* The impact of transposable elements on tomato diversity. *Nat. Commun.* **11**, 4058 (2020).

- Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
- Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
- Eizenga, J. M. *et al.* Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
- Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
- Fang, Z. *et al.* Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics* **191**, 883–894 (2012).
- Filler-Hayut, S., Kniazev, K., Melamed-Bessudo, C. & Levy, A. A. Targeted inter-homologs recombination in *Arabidopsis* euchromatin and heterochromatin. *Int. J. Mol. Sci.* **22**, 12096 (2021).
- Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
- Freire, R. *et al.* Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. *G3* **11**, jkab330 (2021).
- Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, btac743 (2023).
- Garrison, E. *et al.* Building pangenome graphs. *Nat. Methods* **21**, 2008–2012 (2024).
- Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
- Gong, Z. *et al.* Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24**, 3559–3574 (2012).
- Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
- Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
- Harcourt, B. E. Reflecting on the subject: a critique of the social influence conception of deterrence, the broken windows theory, and order-maintenance policing New York style. *Mich. Law Rev.* **97**, 291 (1998).
- Hardigan, M. A. *et al.* Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc. Natl. Acad. Sci. USA* **114**, E9999–E10008 (2017).
- Hayes, B. J. *et al.* Potential approaches to create ultimate genotypes in crops and livestock. *Nat. Genet.* <https://doi.org/10.1038/s41588-024-01942-0> (2024).
- Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).

- Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
- Jansky, S. H. et al. Reinventing potato as a diploid inbred line-based crop. *Crop Sci.* **56**, 1412–1422 (2016).
- Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
- Jiao, Y. et al. Regulation of *OsSPL14* by OsmiR156 defines ideal plant architecture in rice. *Nat. Genet.* **42**, 541–544 (2010).
- Jiang, X. et al. Genomic features of meiotic crossovers in diploid potato. *Hortic. Res.* **10**, uhad079 (2023).
- Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
- Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
- Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
- Leffler, E. M. et al. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356**, eaam6393 (2017).
- Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
- Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Li, Y. et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
- Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
- Lian, Q. et al. Acquisition of deleterious mutations during potato polyploidization. *J. Integr. Plant Biol.* **61**, 7–11 (2019).
- Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
- Liu, Q., Zhou, Y., Morrell, P. L. & Gaut, B. S. Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* **34**, 908–924 (2017).
- Lovell, J. T. et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* **11**, e78526 (2022).
- Lukashin, A. V. & Borodovsky, M. GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
- Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
- Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

- Marco-Sola, S., Moure, J. C., Moreto, M. & Espinosa, A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**, 456–463 (2021).
- Marsden, C. D. *et al.* Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc. Natl. Acad. Sci. USA* **113**, 152–157 (2016).
- Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *BioRxiv*. <https://doi.org/10.1101/085050> (2016).
- Mezmouk, S. & Ross-Ibarra, J. The pattern and distribution of deleterious mutations in maize. *G3* **4**, 163–171 (2014).
- Morgan, A. P. *et al.* Structural variation shapes the landscape of recombination in mouse. *Genetics* **206**, 603–619 (2017).
- Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
- Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
- Parmigiani, L., Garrison, E., Stoye, J., Marschall, T., & Doerr, D. (2024). Panacus: fast and exact pangenome growth and core size estimation. *bioRxiv*, <https://www.biorxiv.org/content/10.1101/2024.06.11.598418v1> (2024).
- Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
- Pham, G. M. *et al.* Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* **9**, giaa100 (2020).
- Porubsky, D. *et al.* Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005 (2022).
- Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).
- Ramu, P. *et al.* Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
- Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
- Roumeliotis, E., Kloosterman, B., Oortwijn, M., Visser, R. G. & Bachem, C. W. The PIN family of proteins in potato and their putative role in tuberization. *Front. Plant Sci.* **4**, 524 (2013).
- Rowan, B. A. *et al.* An ultra high-density *Arabidopsis thaliana* crossover map that refines the influences of structural variation and epigenetic features. *Genetics* **213**, 771–787 (2019).

- Schreiber, M., Jayakodi, M., Stein, N. & Mascher, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* **25**, 563–577 (2024).
- Schmidt, C., Schindele, P. & Puchta, H. From gene editing to genome engineering: restructuring plant chromosomes via CRISPR/Cas. *Abiotech* **1**, 21–31 (2020).
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
- Spooner, D. M., Ghislain, M., Simon, R., Jansky, S. H. & Gavrilenko, T. Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. *Bot. Rev.* **80**, 283–383 (2014).
- Stokstad, E. The new potato. *Science* **363**, 574–577 (2019).
- Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
- Sun, H. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).
- Tang, D. *et al.* Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).
- The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
- The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
- Van Lieshout, N. *et al.* Solyntus, the new highly contiguous reference genome for potato (*Solanum tuberosum*). *G3* **10**, 3489–3495 (2020).
- Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, giy093 (2018).
- Wang, N. *et al.* Structural variation and parallel evolution of apomixis in citrus during domestication and diversification. *Natl. Sci. Rev.* **9**, nwac114 (2022).
- Wang, B. *et al.* *De novo* genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. *Nat. Genet.* **55**, 312–323 (2023).
- Wicker, T. *et al.* A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: recombination, rearrangements and repeats. *Plant J.* **41**, 184–194 (2005).
- Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).
- Wu, Y. *et al.* Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328.e15 (2023).
- Xiao, H. *et al.* Adaptive and maladaptive introgression in grapevine domestication. *Proc. Natl. Acad. Sci. USA* **120**, e2222041120 (2023).
- Xue, J. R. *et al.* The functional and evolutionary impacts of human-specific deletions in conserved elements. *Science* **380**, eabn2253 (2023).

- Yan, H. *et al.* Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nat. Genet.* **55**, 507–518 (2023).
- Ye, M. *et al.* Generation of self-compatible diploid potato by knockout of *S-RNase*. *Nat. Plants* **4**, 651–654 (2018).
- Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
- Zhang, C. *et al.* Genome design of hybrid potato. *Cell* **184**, 3873–3883 (2021).
- Zhang, C. *et al.* The genetic basis of inbreeding depression in potato. *Nat. Genet.* **51**, 374–378 (2019).
- Zhang, X. *et al.* Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* **53**, 1250–1259 (2021).
- Zhou, Y. *et al.* The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
- Zhou, Q. *et al.* Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **52**, 1018–1023 (2020).
- Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).

4

Chapter IV General conclusion and perspectives

1. General conclusion

This thesis contributes to a fundamental paradigm shift in potato breeding by demonstrating the power of phased pangenomes to dissect, quantify, and strategically eliminate deleterious genetic variation in diploid potato. Through the construction of a comprehensive graph-based pangenome from 60 haplotype-resolved genomes, we provide a foundational resource for genome-informed breeding in a clonally propagated and highly heterozygous crop. This work paves the way for the rational design of elite inbred lines, overcoming historical bottlenecks that have constrained potato improvement (Zhang et al., 2021; Wu et al., 2023).

Building upon this central achievement, we first examine how phased pangenomes provide an unprecedented lens through which to capture the full extent of genomic diversity in potato and beyond.

1.1 Phased pangenomes capture comprehensive genomic diversity

Since the first plant pangenome was published in soybean (Li et al., 2014), plant pangenome research has expanded to include key crop species such as rice (Zhao et al., 2018) and barley (Jayakodi et al., 2020; Jayakodi et al., 2024). Different pangenome studies have different focal points. These investigations have highlighted how core and variable gene sets contribute to agronomic traits, local adaptation or domestication (Schreiber et al., 2024). The cotton pangenome integrated both diploid and polyploid accessions with core genes, providing insights into genome convergence and divergence during polyploidization (Li et al., 2024). Similarly, our pangenome research also includes core and dispensable gene classification but places greater emphasis on how pangenome graphs can be utilized for crop breeding.

Beyond constructing a conventional pangenome, we advanced to generate a phased potato pangenome, which, to our knowledge, represents the first phased, graph-based pangenome in plants.

This milestone enables unprecedented resolution in capturing both allelic and structural variants across diverse diploid potato accessions. Compared to previous unphased or linear pangenomes in crops like rice (Zhao et al., 2018), maize (Lu et al., 2015 and Hufford et al., 2021), and soybean (Liu et al., 2020), our approach emphasizes haplotype resolution—a crucial advantage in a highly heterozygous and clonally propagated species like cassava (Prochnik et al., 2012), banana (Davey et al., 2013) and grapevine (Zhou et al., 2019). We also compared the performance of graph-based pangenomes versus linear reference-based variant calling. Our results clearly demonstrated the advantages of graph-based methods, which captured a broader spectrum of SNPs and InDels, especially in regions rich in TEs. Moreover, graph-based variant calling mitigated reference bias, enhanced read alignment quality, and allowed more precise detection of genomic variation (Liao et al., 2023 and Garrison et.al 2023).

This high-resolution genomic resource captures both allelic and structural variations with unprecedented accuracy, offering a comprehensive view of genome-wide diversity. Notably, phased genome/pangenomes are essential in understanding genomic architecture of polyploids, as they distinguish haplotypes (Sun et al., 2025; Bao et al., 2022, Healey et al., 2024 and He et al., 2021). Currently, only a limited number of polyploid pangenomes are available, constraining our ability to fully understand the complexity, evolution, and diversity of polyploid species. In this context, our phased pangenome provides a valuable case study that demonstrates the feasibility and importance of resolving haplotypes in polyploid genomes. It offers both a methodological reference and biological insights, potentially guiding future efforts in polyploid genome assembly, annotation, and evolutionary analysis.

Having established the value of phased pangenomes for exploring structural and allelic variation, we next turned our attention to one of the key drivers of such variation: transposable elements.

1.2 TEs are major drivers of genomic diversity

Leveraging the power of the phased pangenome, we had the chance to uncover several important findings. Our findings in potato, underscore transposable elements as major contributors to structural variation, accounting for over half of the SV events. This mirrors observations in tomato (Domínguez et al., 2020), where TE insertions polymorphisms (TIPs) derived from 337 TE families have been implicated in substantial phenotypic diversification, despite the crop's narrow genetic base. In both species, TEs serve as engines of genome evolution (Schnable et al., 2009; Walkowiak et al., 2020), generating genetic variability through insertions, deletions, and rearrangements, and preferentially inserting near genes involved in stress responses and development.

Yet, important distinctions emerge between the two systems. In tomato, the primary focus has been on genic TE insertions and their transcriptional consequences, with TIPs shown to alter gene expression and even generate novel transcript isoforms (Domínguez et al., 2020). In contrast, the potato genome—characterized by high heterozygosity—reveals TEs as dominant forces driving structural divergence between haplotypes. TE-induced SVs in potato tend to contribute more broadly to genome architecture and haplotypic differentiation.

Another notable difference lies in reproductive strategies. While tomato is mostly self-pollinating, potato is clonally propagated and highly heterozygous, potentially limiting the efficacy of purifying selection against deleterious or disruptive TE insertions. This might explain a greater tolerance of structural complexity and a higher burden of TE-driven variation in potato. Furthermore, plant-specific TE families and differential transpositional activity suggest that the evolutionary dynamics and functional outcomes of TE mobilization are shaped by both genome structure and species-specific selective pressures (Domínguez et al., 2020 and Dayi et al., 2024).

Together, these comparative insights reinforce the central role of TEs in genome evolution across *Solanaceae* species while emphasizing the importance of genome

context—such as ploidy, reproductive mode, and evolutionary history—in determining how TEs influence diversity, adaptation, and trait variation (Wicker et al., 2018; Bennetzen and Wang, 2014).

In addition to beneficial structural innovations, genome evolution in potato is also constrained by the persistence of deleterious variants—highlighting critical limitations of traditional breeding approaches.

1.3 Deleterious variants and the limits of conventional breeding

In our study, we also developed a robust pipeline for the identification of dSVs in potato. This pipeline integrates SV frequency (Zhou et al., 2019), sequence conservation (Wu et al., 2023), and gene coding region (Keightley et al., 2003), enabling reliable discrimination between neutral and deleterious variants. Our approach not only provides new insights into the identification and distribution of dSVs in potato, but also establishes a transferable framework that can be adapted to other crops with available genome assemblies.

We identified that dSNPs and dSVs are not random distributed but often clustered, particularly near each other, forming what we have termed the "broken window" effect (Harcourt, et al., 1998). This phenomenon suggests that a single deleterious mutation may trigger a cascade of linked deleterious variants, thereby increasing the haplotype burden. Consequently, traditional marker-assisted or genomic selection may be insufficient to purge these clusters, emphasizing the need for a haplotype-based breeding approach.

Comparative studies in other crops, such as rice and grapevine, have also revealed the presence of deleterious burdens (Liu et al., 2017; Zhou et al., 2019). For instance, in maize, it was observed that deleterious alleles often reside in low-recombination regions, making them more difficult to eliminate through conventional breeding methods (Lozano et al., 2021). Similarly, in rice, haplotypes harboring multiple weakly deleterious mutations (Liu et al., 2017). And in potato some deleterious variations (white seedling (ws1)) were found to persist due to linkage and selection on nearby beneficial traits (normal plant architecture (PA1)) (Zhou et al., 2020). These findings underscore the widespread nature of this issue across crops and reinforce the necessity for haplotype-aware selection strategies in modern breeding programs.

In potato, the challenge is further exacerbated by its tetraploid nature with high heterozygosity and limited historical recombination, which makes the accumulation and maintenance of deleterious haplotypes even more likely. Therefore, a shift toward haplotype-level variant analysis and breeding design is not only timely but essential for achieving long-term genetic improvement.

To overcome these constraints and realize the full potential of genomic information, we propose a shift toward precision breeding strategies tailored to complex, heterozygous genomes.

1.4 Precision breeding in highly heterozygous crops

Traditional breeding typically selects plants with desirable phenotypes to introduce beneficial alleles (Philip et al., 2011 and Reynolds et al., 2009). However, many beneficial alleles are often linked with deleterious variations—a phenomenon called "the cost of domestication" (Moyers et al., 2018). Such phenotype-based selection approaches are insufficient to break the linkage for highly heterozygous genomes, such as those of potato and cassava.

Therefore, we designed optimized haplotypes represent theoretical optima in the genomic landscape, serving not merely as descriptive models but as actionable targets for precision breeding. By providing a tangible genomic benchmark, the Ideal Potato Haplotypes offer breeders a reference against which current breeding materials and strategies can be quantitatively evaluated. This comparison enables the identification of suboptimal haplotypes and informs the rational design of breeding pipelines aimed at progressively converging toward these ideal genomic configurations. Ultimately, this paradigm has the potential to transform traditional breeding from largely empirical practices into data-driven, genome-informed engineering of elite cultivars.

Despite the advantages of phased pangenomes and ideal potato haplotypes, achieving these ideal genomic configurations through conventional breeding remains highly challenging. Recombination is often restricted in regions where deleterious variants are concentrated, such as centromeres and large inversions. Generating IPHs via natural or random recombination would require extremely large segregating populations to overcome these barriers. Additionally, the linkage between dSNPs and dSVs complicates the disentangling of favorable and unfavorable alleles.

Therefore, precision breeding technologies must be integrated into potato improvement programs. Recent developments in targeted recombination, CRISPR/Cas-based genome editing, and recombination suppression techniques provide promising avenues to bypass biological limitations (Rönspies et al., 2022; Grubben et al., 2024). In 2014, Gao et al. used transcription activator-like effector nucleases technology to simultaneously knock out the MLO genes in the A, B, and D genomes of hexaploid wheat (Wang et al., 2014), achieving broad-spectrum resistance to powdery mildew and setting a successful precedent for gene editing. Subsequently, gene editing technologies have been successfully applied to the breeding of species, including potato (Johansen et al., 2019), rapeseed (Li et al., 2022), sugarcane (Mall et al., 2024), strawberry (Vondracek et al., 2026), and banana (Tripathi et al., 2020), to quickly improve the vigor, yield, biotic and abiotic stress resistance, and herbicide resistance of cultivated varieties. These new technologies offer the potential to manipulate genome content and architecture, enabling more efficient generation of IPHs.

While our findings provide a robust foundation for improving diploid potato breeding, their implications extend well beyond the immediate scope of this study. In the following section, we outline broader perspectives and applications that can further advance breeding in clonally propagated and polyploid crops.

2. General future perspectives

Importantly, the current study lays a foundational framework for phased pangenome-based breeding not only in potato but also in other clonally propagated or polyploid crops such as cassava, sugarcane, grapevine, and banana (Secomandi et al., 2025). The graph-based structure of our pangenome is scalable, allowing the integration of thousands of future accessions and enabling advanced applications such as ancestry tracking, genotype imputation, and association mapping.

1. Extension to Tetraploid Genomes: Although our current study focuses on diploid potato breeding, tetraploid breeding has been the focus of efforts for over a century. As a result, there are already well-established cultivars, such as 'Russet Burbank', that represent valuable genetic resources. Exploring the diversity within these tetraploid materials could further enhance future breeding programs and facilitate the transfer of desirable traits into diploid lines. Therefore, extending the phased pangenome framework to tetraploid accessions will be vital to enhance trait discovery, dissect allelic dosage effects, and develop elite cultivars with optimal performance.

2. Integration with Targeted Breeding Technologies: To achieve IPHs in practice, advanced tools like targeted recombination and synthetic chromosomes must be adapted for use in potato. These approaches will help overcome recombination barriers and facilitate the fixation of favorable haplotypes.

3. Functional Validation of Deleterious Variants: Understanding the phenotypic and regulatory impacts of dSNPs and dSVs under different environmental contexts will be key to refining the IPH framework and improving selection accuracy.

4. Development of Breeder-Friendly Platforms: To realize the full potential of phased pangenomes, graph-aware tools and interfaces must be developed and integrated into routine breeding pipelines. This includes visualization platforms, decision-support systems, and genotype-to-phenotype predictive models.

5. Toward Telomere-to-Telomere (T2T) Assemblies: Achieving complete, gap-free genome assemblies will enhance variant detection and functional annotation (Li et al., 2023; Nurk et al., 2022 and Miga et al., 2020). This will require a combination of PacBio HiFi reads, ONT ultra-long reads, Hi-C, and trio-based strategies. Such assemblies are essential for a deeper understanding of genome architecture, including centromeres and telomeres (Cheng et al., 2022).

6. Maximizing Germplasm Potential: Our current phased potato pangenome includes 31 accessions, comprising 10 wild and 21 cultivated varieties. However, all of the wild accessions belong to the CND clade, highlighting the need to sample more diverse wild populations to better capture genetic variation and support future potato breeding efforts. The phased pangenome provides a roadmap for exploring and utilizing global germplasm collections (Liao et al., 2023). It supports pre-breeding by enabling the identification of rare alleles, prioritizing accessions with complementary haplotypes, and guiding strategic crosses.

7. Applications Beyond Potato: The methodologies developed here can be applied to other clonally propagated or polyploid species such as grapevine, banana, cassava, and even livestock, offering new opportunities for complex genome breeding.

Together, these future directions form an ambitious but feasible roadmap for next-generation plant breeding. They lead naturally into the broader impact and legacy of this thesis.

In conclusion, this work establishes a scalable and conceptually innovative framework for tackling the complexities of breeding clonally propagated crops. By shifting from traditional, linear reference-based approaches to graph-based, haplotype-resolved systems, we lay the groundwork for precision agriculture in species that have long resisted genetic simplification. The phased pangenome model not only elevates our understanding of the potato genome but also opens new avenues for translating complex genomic insights into practical outcomes. As global challenges to agriculture intensify, including climate change, pathogen evolution, and food security demands, the tools, data, and concepts presented in this thesis represent a forward-looking step toward building resilient and genetically optimized crops. This work is not an end, but a beginning—one that invites deeper exploration, broader collaboration, and continued innovation in the field of crop genomics and breeding.

3. References

- Bao, Z. et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Mol. Plant* **15**, 1211–1226 (2022).
- Bennetzen J L, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology*, **2014**, 65: 505-530.
- Cheng, H., Jarvis, E.D., Fedrigo, O. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* **40**, 1332–1335 (2022).
- Davey M W, Gudimella R, Harikrishna J A, et al. A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter-and intra-specific Musa hybrids. *BMC genomics*, **14**: 1-20 (2013).
- Dayi M. Diversity and evolution of transposable elements in the plant-parasitic nematodes. *BMC genomics*, **25**(1): 511(2024).
- Domínguez M, Dugas E, Benchouaia M, et al. The impact of transposable elements on tomato diversity. *Nature communications*, **11**(1): 4058 (2020).
- Garrison E, Sirén J, Novak A M, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, **36**(9): 875-879 (2018).
- Grubben J, Bijsterbosch G, Visser R G F, et al. Key determinants of CRISPR/Cas9 induced inversions in tomato. *bioRxiv* (2024). <https://doi.org/10.1101/2024.01.09.574821>.
- Harcourt, B. E. Reflecting on the subject: a critique of the social influence conception of deterrence, the broken windows theory, and order-maintenance policing New York style. *Mich. Law Rev.* **97**, 291 (1998).
- He S, Sun G, Geng X, et al. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nature Genetics*, 2021, **53**(6): 916-924.
- Healey A L, Garsmeur O, Lovell J T, et al. The complex polyploid genome architecture of sugarcane. *Nature*, **628**(8009): 804-810 (2024).
- Hu W, Ji C, Shi H, et al. Allele-defined genome reveals biallelic differentiation during cassava evolution. *Molecular Plant*, **14**(6): 851-854 (2021).
- Hufford M B, Seetharam A S, Woodhouse M R, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**(6555): 655-662 (2021).
- Jayakodi, M. et al. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).
- Jayakodi, M. et al. Structural variation in the pangenome of wild and domesticated barley. *Nature* **636**, 654–662 (2024).
- Johansen, I. E. et al. High efficacy full allelic CRISPR/Cas9 gene editing in tetraploid potato. *Sci. Rep.* **9**, 17715 (2019)
- Keightley P D, Gaffney D J. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proceedings of the National Academy of Sciences*, **100**(23): 13402-13406 (2003).

- Li, Y. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
- Li, J., Yu, X., Zhang, C., Li, N. & Zhao, J. The application of CRISPR/Cas technologies to Brassica crops: current progress and future perspectives. *aBIOTECH* **3**, 146–161 (2022).
- Li, H. and Durbin, R et al. Genome assembly in the telomere-to-telomere era. *Nature Reviews Genetics*, pp.1-13 (2024).
- Li, J. et al. Convergence and divergence of diploid and tetraploid cotton genomes. *Nat. Genet.* 1–12 (2024).
- Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
- Liu Y, Du H, Li P, et al. Pan-genome of wild and cultivated soybeans. *Cell*, **182**(1): 162-176. e13 (2020).
- Liu Q, Zhou Y, Morrell P L, et al. Deleterious variants in Asian rice and the potential cost of domestication. *Molecular Biology and Evolution*, **34**(4): 908-924 (2017).
- Lozano R, Gazave E, Dos Santos J P R, et al. Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nature Plants*, **7**(1): 17-24 (2021).
- Lu F, Romay M C, Glaubitz J C, et al. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature communications*, **6**(1): 6914 (2015).
- Miga K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020)
- Mall, A. K. et al. CRISPR/Cas-Mediated Genome Editing for Sugarcane Improvement. *Sugar Tech* 1–13 (2024).
- Moyers B T, Morrell P L, McKay J K. Genetic costs of domestication and improvement. *Journal of Heredity*, **109**(2): 103-116 (2018).
- Nurk S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
- Reynolds M, Manes Y, Izanloo A, et al. Phenotyping approaches for physiological breeding and gene discovery in wheat. *Annals of Applied Biology*, **155**(3): 309-320 (2009).
- Philip V M, Sokoloff G, Ackert-Bicknell C L, et al. Genetic analysis in the Collaborative Cross breeding population. *Genome research*, **21**(8): 1223-1238 (2011).
- Prochnik S, Marri P R, Desany B, et al. The cassava genome: current progress, future directions. *Tropical plant biology*, **5**: 88-94 (2012).
- Rönspies M, Schmidt C, Schindele P, et al. Massive crossover suppression by CRISPR–Cas-mediated plant chromosome engineering. *Nature Plants*, 2022, **8**(10): 1153-1159.
- Tripathi, L., Ntui, V. O. & Tripathi, J. N. CRISPR/Cas9-based genome editing of banana for disease resistance. *Curr. Opin. Plant Biol.* **56**, 118–126 (2020)
- Vondracek, K., Altpeter, F., Liu, T. & Lee, S. Advances in genomics and genome editing for improving strawberry (*Fragaria\anassa*). *Front. Genet.* **15**, 1382445 (2024)
- Walkowiak S, Gao L, Monat C, et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, **2020**, 588(7837): 277-283.

- Wang, Y. et al. Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nat. Biotechnol.* **32**, 947–951 (2014).
- Wicker T, Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 2018, **8**(12): 973-982.
- Wu, Y. *et al.* Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328.e15 (2023).
- Schnable P S, Ware D, Fulton R S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 2009, **326**(5956): 1112-1115.
- Schreiber M, Jayakodi M, Stein N, et al. Plant pangenomes for crop improvement, biodiversity and evolution. *Nature Reviews Genetics*, **25**(8): 563-577 (2024).
- Secomandi S, Gallo G R, Rossi R, et al. Pangenome graphs and their applications in biodiversity genomics. *Nature Genetics*: **57**, 13–26 (2025).
- Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).
- Zhao Q, Feng Q, Lu H, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature genetics*, **50**(2): 278-284 (2018).
- Zhang Y, Zhang D. Asexual and sexual reproductive strategies in clonal plants. *Frontiers of Biology in China*, **2**: 256-262 (2007).
- Zhang, C. *et al.* Genome design of hybrid potato. *Cell* **184**, 3873–3883 (2021).
- Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).

Appendix-Publications

1. **Lin Cheng**, Nan W, Z Bao, Q Zhou, *et al.*, Leveraging a phased pangenome for haplotype design of hybrid potato. **640**, 408–417 *Nature* (2025).
2. **Lin Cheng**, Zhigui B, Ludivine L, *et al.* Genome analyses and breeding of polyploid crops. *Nature Plant.* (2025, accepted).
3. Yanling Peng, N Wang, **Lin Cheng** *et al.*, The genomic and epigenomic landscapes of hemizygous genes across crops with contrasting reproductive systems. *Proceedings of the National Academy of Sciences* (2025).
4. Pei W, **Lin Cheng**, Jun P, L Ma, *et al.*, A 6.49-Mb inversion associated with the purple embryo spot trait in potato. *aBIOTECH*, (2025).
5. Guo M, Lian Q, Mei Y, **Lin Cheng**, *et al.*, Analyzes of pan-genome and resequencing atlas unveil the genetic basis of jujube domestication. *Nature Communications*, 2024, **15**(1): 9320.
6. Zhou Y, Zhang Z, Bao Z, **Lin Cheng**, *et al.*, Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, 2022, **606** (7914): 527-534.
7. Tang D, Jia Y, Zhang J, **Lin Cheng**, *et al.*, Genome evolution and diversity of wild and cultivated potatoes. *Nature*, 2022, **606** (7914): 535-541.
8. Bao Z, Li C, Li G, **Lin Cheng**, *et al.*, Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular plant*, 2022, **15**(7): 1211-1226.
9. **Cheng Lin**, Kim K W, Park Y J. Evidence for selection events during domestication by extensive mitochondrial genome analysis between japonica and indica in cultivated rice. *Scientific reports*, 2019, **9**(1): 10846.
10. **Cheng Lin**, Nam J, Chu S H, *et al.*, Signatures of differential selection in chloroplast genome between japonica and indica. *Rice*, 2019, **12**: 1-13.