

**PANGENOMIC ARCHITECTURE OF HARVEST INDEX DETERMINANTS
IN BIOREGENERATIVE LIFE SUPPORT SYSTEMS STAPLE CROPS**

by

Laszlo Pokorny

Department of Bioregenerative Life Support Sciences

ICL Institute

New Jersey, USA

DOI: 10.5281/zenodo.18501503

2026

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Copyright © 2026 by Laszlo Pokorny

All Rights Reserved

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

ABSTRACT

As humanity prepares for long-duration missions to the Moon and Mars, the development of Bioregenerative Life Support Systems (BLSS) capable of sustaining crews through in situ food production has emerged as a critical technological imperative. The prohibitive cost of launching consumables from Earth—estimated at \$20,000 per kilogram for Mars missions—necessitates closed-loop systems that regenerate essential resources including food, water, and oxygen. Central to BLSS optimization is harvest index (HI)—the ratio of edible yield to total biomass—which directly determines the efficiency with which limited spacecraft resources are converted into consumable food. Despite the fundamental importance of HI for space agriculture, no systematic cross-species analysis has characterized the pangenomic architecture of HI determinants across BLSS candidate crops. This dissertation addressed this critical gap through a comprehensive pangenomic investigation of HI-controlling genes in four staple species prioritized for space agriculture: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*). Employing a four-phase computational comparative genomics approach, this study constructed a systematic gene inventory, conducted ortholog analysis across species, characterized structural variants through integration of genome-wide association study (GWAS) and quantitative trait locus (QTL) data, and examined evolutionary patterns underlying HI regulation. Results identified 30 HI-associated genes organized into five functional categories: dwarfing/gibberellin signaling, grain/tuber development, flowering and maturity, branching and tillering, and nutrient partitioning. Ortholog analysis revealed remarkable conservation of HI-regulatory pathways, with gibberellin signaling demonstrating 100% conservation across all four species—a finding reflecting the enduring legacy of Green Revolution selection for semi-dwarf architecture. Integration of 20 GWAS hits and 25 QTLs

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

validated the gene catalog while illuminating the genomic architecture of HI variation in these economically vital crops. This research contributes a unified pangenomic framework for HI analysis applicable to both space and terrestrial agriculture, and identifies priority molecular targets for NASA crop development programs, thereby advancing the foundational science necessary for sustainable human presence beyond Earth and enhanced food security on our home planet.

Keywords: pangenomics, harvest index, bioregenerative life support systems, space agriculture, structural variants, gibberellin signaling, crop improvement

Table of Contents

ABSTRACT.....	3
CHAPTER 1: INTRODUCTION	16
1.1 Background of the Study	16
1.2 Statement of the Problem.....	20
1.3 Purpose of the Study	23
1.4 Research Questions.....	24
1.5 Significance of the Study	25
Table 1.1 Summary of Study Significance	29
1.6 Scope and Delimitations	29
1.7 Definition of Key Terms.....	31
Table 1.2 Glossary of Key Terms	32
1.8 Assumptions.....	33
1.9 Organization of the Dissertation	35
1.10 Chapter Summary	37
CHAPTER 2: LITERATURE REVIEW	39
2.1 Introduction.....	39
2.2 Pangenomics: Concepts, Methods, and Applications	43
2.2.1 History and Evolution of the Pangenomics Concept	43

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

2.2.2 Theoretical Framework: Core Genome vs. Dispensable Genome.....	45
2.2.3 Pangenome Construction Methodologies	47
2.2.4 Structural Variation in Plant Genomes	49
2.2.5 Technological Advances Enabling Plant Pangenomics	52
2.2.6 Major Plant Pangenome Projects and Their Findings.....	54
2.2.7 Pangenomics Applications in Crop Improvement	56
2.2.8 Comparative Pangenomics and Super-Pangenomes	58
2.3 Harvest Index: Physiological and Genetic Foundations	60
2.3.1 Definition and Historical Development	60
2.3.2 Physiological Basis of Harvest Index	62
2.3.3 The Green Revolution and Harvest Index Improvement	65
2.3.4 Genetic Architecture of Harvest Index	67
2.3.5 Molecular Determinants of Harvest Index.....	70
2.3.6 Yield Components and Trade-offs	74
2.3.7 Modern Breeding Approaches for Harvest Index Improvement	76
2.4 Genomics of BLSS Staple Crops	78
2.4.1 Rice (<i>Oryza sativa</i>)	79
2.4.2 Wheat (<i>Triticum aestivum</i>)	84
2.4.3 Potato (<i>Solanum tuberosum</i>).....	88
2.4.4 Soybean (<i>Glycine max</i>).....	91

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

2.5 Bioregenerative Life Support Systems and Space Agriculture.....	95
2.5.1 Introduction to BLSS Concepts	95
2.5.2 Crop Selection for Space Agriculture	98
2.5.3 Controlled Environment Agriculture for Space	100
2.5.4 Harvest Index Optimization for BLSS.....	103
2.5.5 Current BLSS Crop Research	105
2.5.6 Integration of Genomics with BLSS Crop Improvement	107
2.6 Synthesis and Research Gaps.....	109
2.7 Chapter Summary	114
CHAPTER 3: METHODOLOGY	119
3.1 Introduction and Research Design Overview	119
3.1.1 Theoretical Framework.....	120
3.1.2 Four-Phase Research Design	121
3.2 Study Species and Rationale.....	122
Table 3.1 Target Species Characteristics for Pangenomic Analysis of Harvest Index Determinants	122
3.3 Data Sources and Acquisition.....	124
3.3.1 Genomic Data Sources.....	124
Table 3.2 Primary Genomic Databases and Data Types.....	124
3.3.2 Pangenomic Data Sources.....	125

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 3.3 Pangenome Resources by Species	125
3.3.3 Phenotypic and Association Data	126
Table 3.4 GWAS and QTL Databases for Phenotype-Genotype Association Data	126
3.3.4 Data Quality Assessment Criteria	126
3.4 Phase 1: HI Gene Inventory and Classification	127
3.4.1 Gene Identification Criteria	127
3.4.2 Functional Classification Schema	128
Table 3.5 Functional Categories of Harvest Index-Associated Genes	128
3.4.3 Gene Classification Workflow	128
Figure 3.1 Gene Classification Workflow for HI-Associated Gene Inventory	129
3.4.4 Gene Curation and Annotation Standards	130
3.5 Phase 2: Ortholog Identification and Phylogenomic Analysis	130
3.5.1 Ortholog Detection Methods	131
3.5.2 Synteny Analysis Approach	131
3.5.3 Phylogenetic Reconstruction Methods	132
Figure 3.2 Phylogenomic Analysis Pipeline for HI Gene Family Evolution	134
3.5.4 Molecular Dating Methodology	135
3.5.5 Handling Polyploid Genomes	135
3.6 Phase 3: Structural Variant Analysis	136
3.6.1 PAV Detection and Cataloging	136

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

3.6.2 CNV Analysis Methods	137
3.6.3 Population Frequency Calculations	138
3.6.4 Phenotype-Genotype Association Testing.....	138
Figure 3.3 Structural Variant Analysis Workflow for HI Loci.....	140
3.6.5 Statistical Approaches.....	141
3.6.6 Visualization and Interpretation Framework	141
3.7 Phase 4: Evolutionary Analysis	142
3.7.1 Ancestral State Reconstruction Methods	142
3.7.2 Tests for Convergent Evolution	143
3.7.3 Selection Analysis Approaches.....	143
3.7.4 Integration of Multi-Level Evolutionary Evidence.....	144
3.8 Bioinformatics Tools and Software	145
Table 3.6 Bioinformatics Software and Tools	145
3.9 Data Management and Quality Control	146
3.9.1 Data Validation Procedures	146
3.9.2 Quality Control Metrics	147
3.9.3 Reproducibility and Transparency Standards	148
3.10 Ethical Considerations	149
3.10.1 Use of Publicly Available Data.....	149
3.10.2 Data Sharing Plans.....	149

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

3.11 Limitations	150
3.11.1 Methodological Limitations and Mitigation Strategies	150
3.12 Chapter Summary	152
Table 3.7 Research Timeline and Milestones	154
Table 3.8 Expected Sample Sizes for Population-Level Analyses by Species	155
CHAPTER 4: RESULTS	156
4.1 Introduction	156
4.2 Phase 1 Results: HI Gene Inventory and Classification	157
4.2.1 Overall Gene Catalog Summary	157
Table 4.1 Summary of HI-Associated Genes by Species	159
4.2.2 Functional Classification Results	159
Table 4.2 Distribution of HI-Associated Genes by Functional Category	161
Figure 4.1 Distribution of HI-Associated Genes by Species	162
Figure 4.2 Heatmap of Gene Distribution Across Species and Functional Categories	163
4.2.3 Species-Specific Findings	163
4.3 Phase 2 Results: Ortholog Identification and Phylogenomic Analysis	168
4.3.1 Ortholog Group Identification	168
Table 4.3 Ortholog Groups of HI-Associated Genes Across Species	169
4.3.2 Pathway Conservation Analysis	170
Table 4.4 Conservation Levels of HI-Related Pathways Across Species	171

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Figure 4.3 Presence-Absence Heatmap of Ortholog Groups Across Species	172
4.3.3 Cross-Species Relationships	172
4.4 Phase 3 Results: Structural Variant and Association Analysis	173
4.4.1 Pangenomic Statistics Summary	174
Table 4.5 Genomic and Pangenomic Statistics by Species	175
4.4.2 GWAS Results	176
Table 4.6 Summary of GWAS Hits for HI-Related Traits by Species	177
Figure 4.4 Distribution of GWAS Hits by Species	178
4.4.3 QTL Mapping Results	179
Table 4.7 Summary of QTLs for HI-Related Traits by Species	180
Figure 4.5 Distribution of QTLs by Species	181
4.4.4 Phenotype-Genotype Associations	182
Figure 4.6 Combined GWAS and QTL Association Summary	182
4.5 Phase 4 Results: Evolutionary Analysis	183
4.5.1 Conservation Patterns	183
Table 4.8 Conservation Scores by Functional Category	185
4.5.2 Evolutionary Trends	186
Figure 4.7 Functional Category Distribution Across Species	186
4.6 Integration of Results	188
Table 4.9 Summary of Integrated Findings Across Analysis Phases	189

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

4.7 Chapter Summary	191
CHAPTER 5: DISCUSSION.....	193
5.1 Introduction.....	193
5.2 Interpretation of HI Gene Inventory Findings	194
5.2.1 Significance of the 30-Gene Catalog	194
5.2.2 Functional Category Distribution Patterns.....	196
5.2.3 Species-Specific Considerations	197
5.3 Ortholog Conservation and Evolutionary Implications	198
5.3.1 Highly Conserved Pathways (GA Signaling)	198
5.3.2 Moderately Conserved Pathways.....	200
5.3.3 Variable Pathways (Starch and Sucrose Transport).....	200
5.3.4 Evolutionary Convergence in HI Determinants.....	201
Table 5.1 Summary of Ortholog Conservation Patterns and Breeding Implications	202
5.3.5 Implications for Understanding Crop Domestication	203
5.4 Structural Variation and Phenotypic Associations	204
5.4.1 GWAS Findings in Context	204
5.4.2 QTL Validation and Integration.....	205
5.4.3 PAV/CNV Implications for Breeding.....	205
Figure 5.1 Conceptual Model of Pangenomic Variation Affecting Harvest Index	207
5.5 Implications for BLSS Crop Improvement.....	208

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

5.5.1 Priority Targets for Space Agriculture.....	208
Table 5.2 Prioritized Gene Targets for BLSS Harvest Index Optimization	209
5.5.2 Cross-Species Transfer Potential	210
5.5.3 Controlled Environment Considerations.....	210
5.5.4 Recommendations for NASA Crop Development Program.....	211
5.6 Theoretical Contributions	212
5.6.1 Advancing Pangenomic Understanding of Complex Traits	212
5.6.2 Framework for Comparative HI Analysis.....	212
Figure 5.2 Proposed Framework for Pangenomic Harvest Index Research	213
5.7 Practical Applications	214
5.7.1 Marker-Assisted Selection Targets	214
5.7.2 Gene Editing Candidates.....	214
5.7.3 Genomic Selection Integration	215
Table 5.3 Summary of GWAS Significance Levels and Research Priorities	215
5.8 Limitations of the Study.....	216
5.9 Future Research Directions.....	216
5.9.1 Experimental Validation Priorities	216
5.9.2 Expanded Species Coverage	217
5.9.3 Functional Genomics Studies	217
5.9.4 Space Environment Testing	217

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

5.9.5 Integration with Multi-Omics Approaches	217
5.9.6 Climate Change Adaptation Considerations	217
5.9.7 Collaborative Research Networks.....	218
5.9.8 Technology Development Requirements.....	218
5.10 Chapter Summary	218
CHAPTER 6: CONCLUSION	237
6.1 Introduction.....	237
6.2 Summary of Research	238
6.2.1 Research Questions Revisited.....	238
6.2.2 Methodological Approach Summary	240
6.2.3 Key Findings Overview	240
Table 6.1 Summary of Major Findings by Research Phase.....	241
6.3 Contributions to Knowledge	242
6.3.1 Theoretical Contributions	242
6.3.2 Methodological Contributions	244
6.3.3 Practical Contributions.....	245
Table 6.2 Summary of Dissertation Contributions by Domain	246
6.4 Implications.....	246
6.4.1 For Plant Genomics Research.....	246
6.4.2 For Crop Breeding Programs.....	247

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

6.4.3 For Space Agriculture and BLSS Development	248
6.4.4 For NASA's Advanced Life Support Program	249
6.5 Recommendations.....	250
6.5.1 For Researchers.....	250
6.5.2 For Breeders and Crop Scientists.....	251
6.5.3 For Space Agriculture Programs.....	252
6.5.4 For Policy and Funding Agencies.....	254
6.6 Final Reflections	255
6.7 Concluding Statement.....	257
REFERENCES	260
APPENDIX A: Data Sources and URLs	286
Table A1 Primary Genomic Databases.....	286
Table A2 Orthology and Comparative Genomics Databases	287
Table A3 Species-Specific Pangenome Resources.....	289
Table A4 Phenotype-Genotype Association Databases.....	291
Table A5 Reference Genome Assemblies Used	292
Table A6 Bioinformatics Tools and Software	292

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

The twenty-first century has witnessed a renewed global commitment to human space exploration, with ambitious programs targeting sustained presence on the Moon through NASA's Artemis missions and eventual crewed expeditions to Mars (NASA, 2020). These endeavors represent a fundamental departure from the short-duration orbital missions that characterized earlier eras of spaceflight, introducing unprecedented challenges related to crew sustenance over periods measured in years rather than weeks. The logistics of supplying crews with food, water, and breathable air from Earth become prohibitively expensive and increasingly impractical as mission distances and durations extend. Current estimates suggest that launching consumables to Mars would cost approximately \$20,000 per kilogram, rendering resupply-dependent mission architectures economically unfeasible for long-duration settlements (Perchonok et al., 2012). This economic reality, combined with the inherent risks of supply chain disruptions across interplanetary distances, has catalyzed intense research into alternative approaches for sustaining human life beyond Earth.

Bioregenerative Life Support Systems (BLSS) have emerged as the most promising solution to this challenge, integrating biological processes with physical-chemical technologies to create closed-loop systems capable of regenerating essential resources in situ (Wheeler, 2017). Unlike purely physical-chemical life support systems that rely on stored consumables and mechanical processes, BLSS leverage the metabolic capabilities of plants, algae, and microorganisms to recycle carbon dioxide into oxygen, purify wastewater through biological filtration, and—most critically—produce fresh food for crew consumption. This biological

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

approach offers the additional benefit of providing psychological support to isolated crews through the presence of living plants and the satisfaction derived from cultivation activities (Odeh & Guy, 2017). The European Space Agency's MELiSSA (Micro-Ecological Life Support System Alternative) project, NASA's Biomass Production Chamber experiments at Kennedy Space Center, and the Chinese Lunar Palace 1 facility represent major international efforts to develop and validate BLSS technologies for future space missions (Gòdia et al., 2002; Wheeler et al., 2008; Fu et al., 2016).

Central to the food production component of any BLSS is the selection and optimization of crop species capable of meeting crew nutritional requirements while operating within the severe resource constraints of a spacecraft or planetary habitat. The cultivation environment in space imposes stringent limitations on volume, mass, power, and labor—resources that must be shared across all life support functions and mission activities. Consequently, crop selection for BLSS prioritizes species that offer high nutritional density, rapid growth cycles, efficient resource utilization, and robust performance under controlled environment conditions (Wheeler, 2017). Through decades of research, four staple crops have consistently emerged as priority candidates for BLSS implementation: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*). Collectively, these species provide the foundation for meeting crew caloric requirements while supplying essential macronutrients including carbohydrates, proteins, and fats (Perchonok et al., 2012).

Within this context, harvest index (HI)—defined as the ratio of economically valuable yield to total plant biomass—emerges as perhaps the single most important metric for evaluating crop suitability for BLSS applications. In the resource-constrained environment of a space habitat, maximizing harvest index translates directly to maximizing the efficiency with which

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

cultivation inputs are converted into edible output. A plant with a harvest index of 0.50 directs half of its accumulated biomass into edible portions, while a plant with an HI of 0.30 allocates only 30% to consumable yield, with the remainder invested in roots, stems, and leaves that must be processed or recycled. The significance of this efficiency metric was demonstrated unequivocally during the Green Revolution, when the introduction of semi-dwarf genes in wheat and rice increased harvest indices from approximately 0.30 to over 0.50, effectively doubling the proportion of plant biomass allocated to grain (Hedden, 2003; Khush, 2001). For BLSS applications, where every kilogram of biomass grown represents investments in light energy, water, nutrients, and atmospheric gases, similar improvements in harvest index could dramatically enhance system efficiency and reduce the resource footprint required to sustain crew nutrition.

The historical trajectory of harvest index improvement in terrestrial agriculture provides instructive lessons for BLSS crop development. Prior to the Green Revolution, traditional tall cultivars of wheat and rice allocated substantial photosynthetic resources to stem elongation, resulting in harvest indices typically ranging from 0.25 to 0.35 (Austin et al., 1980). These tall plants were prone to lodging—the collapse of stems under the weight of grain—which further reduced recoverable yield. The introduction of dwarfing alleles, particularly *Rht-B1b* and *Rht-D1b* in wheat and *sd1* in rice, fundamentally altered plant architecture by reducing gibberellin sensitivity or biosynthesis, thereby limiting stem elongation while redirecting assimilates to grain development (Peng et al., 1999). This single genetic modification, combined with improved agronomic practices and increased fertilizer inputs, enabled harvest indices to exceed 0.50 in modern semi-dwarf cultivars—effectively doubling the efficiency of biomass allocation to economically valuable yield.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The relevance of Green Revolution genetics to BLSS applications extends beyond the historical precedent of successful HI improvement. The semi-dwarf phenotype confers additional advantages in controlled environment systems, including reduced plant height that enables higher planting densities and more efficient utilization of limited vertical space. Shorter stems also require less structural investment, potentially freeing resources for reproductive development. Furthermore, the lodging resistance associated with semi-dwarf architecture eliminates a significant source of yield loss that can occur even in controlled environments when plants are subjected to air circulation for transpiration management. These considerations suggest that understanding and extending the genetic principles underlying Green Revolution gains should be a priority for BLSS crop improvement.

The recent emergence of pangenomics as a transformative framework for crop genetics provides unprecedented opportunities to understand and manipulate harvest index determinants. Traditional genomics, constrained by reliance on single reference genomes, inherently captures only a fraction of the genetic diversity present within and among crop species. Pangenomic approaches, by contrast, incorporate the full complement of genomic variation present across multiple individuals or cultivars, including structural variations such as presence/absence variants (PAVs) and copy number variants (CNVs) that are invisible to reference-based analyses (Della Coletta et al., 2021). These structural variants have been increasingly recognized as major contributors to phenotypic diversity in crops, including variation in yield-related traits (Golicz et al., 2016). The application of pangenomic perspectives to harvest index research thus promises to reveal previously hidden sources of HI variation while enabling systematic comparison of HI-controlling mechanisms across species.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Contemporary advances in genomics and molecular biology have substantially expanded the toolkit available for crop improvement beyond the classical breeding approaches of the Green Revolution era. Genome sequencing technologies now enable comprehensive characterization of genetic variation at single-nucleotide resolution, while genome-wide association studies provide statistical power to identify loci controlling complex traits across diverse germplasm collections. The emergence of CRISPR-Cas genome editing offers unprecedented precision in introducing targeted modifications without the regulatory and biological complications associated with transgenesis. These technologies, combined with the pangenomic perspective that recognizes the importance of structural variation, create opportunities for rational engineering of harvest index that were unimaginable when Norman Borlaug first introduced semi-dwarf wheat to the developing world.

1.2 Statement of the Problem

Despite the critical importance of harvest index for BLSS crop optimization, significant gaps exist in our understanding of the pangenomic architecture of HI determinants across candidate species. The current state of knowledge is characterized by three interrelated deficiencies that this dissertation seeks to address.

First, no systematic cross-species pangenomic analysis has been conducted to catalog and compare the genes controlling harvest index across BLSS staple crops. While individual studies have characterized HI-associated genes in specific species—such as the semi-dwarf genes *sd1* in rice and *Rht* in wheat—these investigations have largely proceeded in isolation, with limited attention to comparative frameworks that could illuminate conserved mechanisms and species-specific innovations. The absence of a unified gene catalog impedes efforts to translate

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

discoveries from well-characterized model species to less-studied crops and prevents the identification of universal targets for cross-species improvement strategies.

The fragmented state of HI research across crop species reflects broader patterns of disciplinary specialization in plant genomics. Rice geneticists have deeply characterized the molecular mechanisms of semi-dwarfism, grain size determination, and panicle architecture, generating a wealth of functional data that has made rice a model cereal. Wheat genomics has progressed more slowly due to the complexity of its allohexaploid genome, though recent advances including the release of high-quality reference assemblies have accelerated gene discovery. Potato research has focused extensively on tuber initiation and development, with particular attention to the photoperiodic signals and hormonal cascades controlling tuberization. Soybean genetics has emphasized seed composition traits relevant to protein and oil content, with harvest index receiving comparatively less attention. These parallel research trajectories have generated substantial species-specific knowledge but limited cross-fertilization of concepts and approaches.

Second, the contribution of structural variants—particularly PAVs and CNVs—to harvest index variation remains poorly characterized, especially in the context of BLSS crop improvement. Pangenomic studies in several crop species have revealed that substantial portions of the genome consist of dispensable sequences present in some cultivars but absent in others, and that these structural variants can have significant phenotypic effects (Hurgobin & Edwards, 2017). However, the specific relationship between structural variation at HI-controlling loci and harvest index phenotypes has not been systematically investigated across BLSS candidate crops. This gap is particularly significant given that many elite cultivars developed for terrestrial

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

agriculture may not carry optimal allelic configurations for the unique environmental conditions of controlled environment agriculture in space.

Third, the evolutionary patterns characterizing HI determinants across phylogenetically diverse crop species remain unclear, limiting our ability to predict which genetic modifications will be most effective for enhancing harvest index in BLSS contexts. Understanding whether HI-controlling mechanisms represent ancient conserved pathways or more recent species-specific innovations has profound implications for the transferability of genetic improvement strategies across crops. Similarly, distinguishing ancestral from derived allelic states at HI loci can inform predictions about which alleles are most amenable to engineering for enhanced performance under controlled environment conditions.

The absence of integrative frameworks poses practical problems for BLSS development programs that must simultaneously optimize multiple crop species within an integrated food system. NASA's Vegetable Production System (Veggie) and Advanced Plant Habitat experiments aboard the International Space Station have demonstrated the feasibility of plant cultivation in microgravity, but these proof-of-concept studies have yet to address the systematic genetic optimization required for self-sustaining food production (Massa et al., 2017). The development of BLSS-ready cultivars requires coordinated improvement strategies that leverage the full depth of available genomic knowledge while accounting for the unique requirements of space cultivation environments. A unified pangenomic framework for HI analysis would facilitate such coordination by identifying conserved targets amenable to cross-species improvement and distinguishing them from species-specific mechanisms requiring tailored approaches.

1.3 Purpose of the Study

The primary purpose of this dissertation is to conduct a comprehensive pangenomic analysis of the genetic architecture controlling harvest index in four BLSS staple crops: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*). This research aims to establish a systematic framework for understanding HI determinants that can inform crop improvement strategies for space agriculture applications.

To achieve this overarching aim, the study pursues four specific objectives. The first objective is to construct a comprehensive catalog of genes with documented effects on harvest index and related yield components across the four target species, organized by functional category and validated through integration of genome-wide association study (GWAS) and quantitative trait locus (QTL) data. The second objective is to identify orthologous relationships among HI-controlling genes across species, thereby elucidating conserved genetic mechanisms and enabling cross-species knowledge transfer. The third objective is to characterize pangenomic structural variants—including PAVs and CNVs—at HI-associated loci and evaluate their potential contribution to harvest index variation. The fourth objective is to analyze evolutionary patterns underlying HI determinants, including the identification of ancestral versus derived allelic states and the reconstruction of selective histories at key loci.

The selection of rice, wheat, potato, and soybean as focal species reflects their established priority for BLSS implementation based on nutritional profile, cultivation characteristics, and accumulated research knowledge. Rice and wheat provide the carbohydrate foundation of the BLSS diet while offering well-characterized genetics and extensive genomic resources. Potato contributes high caloric density with rapid growth cycles and vegetative propagation that simplifies cultivation management. Soybean supplies essential proteins and fats that cannot be

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

adequately obtained from cereal crops alone. Together, these four species represent the minimum complement necessary to meet crew nutritional requirements while providing a phylogenetically diverse sample for comparative genomic analysis spanning monocots and dicots.

1.4 Research Questions

This dissertation is guided by four primary research questions that structure the investigation and analysis:

RQ1: What genes have been documented to influence harvest index and related yield components (grain/tuber weight, number, biomass partitioning) across rice, wheat, potato, and soybean, and how can they be systematically classified by functional category?

RQ2: What orthologous relationships exist among HI-controlling genes across the four target species, and which gene families show conserved versus species-specific roles in harvest index determination?

RQ3: What structural variants (presence/absence variants and copy number variants) have been identified at HI-associated loci, and how do these variants contribute to phenotypic variation as evidenced by GWAS and QTL studies?

RQ4: What evolutionary patterns characterize HI determinants across the target species, including the distinction between ancestral and derived allelic states and the signatures of selection at key loci?

These research questions are designed to be addressed sequentially, with findings from earlier questions informing subsequent analyses. The gene inventory established through RQ1 provides the foundation for ortholog identification in RQ2. The structural variant and association

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

analyses of RQ3 leverage both the gene catalog and ortholog clusters to interpret the functional significance of genomic variation. Finally, the evolutionary analysis of RQ4 synthesizes information from all preceding questions to characterize the historical forces that have shaped HI determinants and predict their future potential for improvement.

1.5 Significance of the Study

This research contributes to multiple domains of scholarship and practice, with implications spanning theoretical understanding, methodological advancement, and practical application. The following subsections detail these contributions.

1.5.1 Theoretical Significance

From a theoretical perspective, this dissertation advances understanding of the genetic architecture underlying harvest index—a complex trait that integrates multiple physiological processes including photoassimilate production, allocation, and partitioning among competing sinks. By systematically cataloging HI determinants across phylogenetically diverse species, this research contributes to fundamental knowledge about the evolution and conservation of biomass partitioning mechanisms in plants. The identification of conserved ortholog pathways across monocots and dicots provides insights into the deep evolutionary origins of HI regulation, while documentation of species-specific innovations illuminates the mechanisms of adaptive diversification. Furthermore, this work contributes to the emerging theoretical framework of pangenomics by demonstrating how cross-species structural variant analysis can reveal hidden dimensions of genetic architecture for complex traits.

The theoretical significance extends to fundamental questions about the evolution of source-sink relationships in plants. Harvest index represents the culmination of complex

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

developmental processes governing photoassimilate production in source tissues (primarily leaves), phloem loading and translocation, and partitioning among competing sink organs including roots, stems, and reproductive structures. The evolutionary pressures shaping these processes differ substantially between wild and domesticated plants, with natural selection favoring allocation strategies that maximize reproductive fitness through seed dispersal and competitive vegetative growth, while artificial selection under domestication has favored increased allocation to harvestable organs at the expense of traits important in natural ecosystems (Meyer & Purugganan, 2013). Understanding how domestication has altered the genetic architecture of biomass partitioning provides insights into both the history of human-plant interactions and the potential for future improvement.

1.5.2 Practical Significance

The practical significance of this research extends to both terrestrial and extraterrestrial agricultural applications. For conventional crop breeding programs, the comprehensive HI gene catalog and associated GWAS/QTL validation provide actionable targets for marker-assisted selection and genomic selection strategies aimed at improving harvest index. The identification of structural variants at HI loci offers novel sources of allelic diversity that may be exploited through targeted introgression or gene editing approaches. For controlled environment agriculture—including vertical farms and greenhouse operations—this research informs cultivar selection and optimization strategies specifically tailored to the conditions of enclosed cultivation systems. The cross-species comparative framework enables breeders to leverage discoveries from model species for improvement of less-characterized crops.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The practical significance for controlled environment agriculture encompasses applications beyond space habitats. Vertical farming and other CEA systems face similar resource constraints to BLSS, with capital-intensive infrastructure and operational costs creating strong economic incentives to maximize yield per unit of cultivation area, energy input, and water consumption. The global vertical farming market is projected to exceed \$20 billion by 2030, driven by urbanization, climate change impacts on field agriculture, and consumer demand for locally produced fresh food (Grand View Research, 2023). Cultivars optimized for high harvest index are essential for the economic viability of these systems, as inefficient biomass allocation translates directly to increased production costs. The genomic targets identified through this dissertation will inform cultivar development for the broader CEA industry while advancing BLSS-specific applications.

1.5.3 Significance for Space Agriculture

For space agriculture specifically, this dissertation provides foundational knowledge essential for developing optimized BLSS crop cultivars. The identification of priority molecular targets for HI enhancement directly supports NASA's Advanced Plant Habitat research program and international BLSS development efforts. By establishing which HI-controlling mechanisms are conserved across species, this research enables coordinated multi-crop improvement strategies that could accelerate the development of BLSS-ready cultivar suites. The characterization of structural variants provides a resource for identifying alleles that may confer advantages under the unique environmental conditions of space habitats, including altered gravity, elevated CO₂ concentrations, and precisely controlled photoperiods. This knowledge is essential for achieving the crop performance levels required to make food production a viable component of closed-loop life support for long-duration space missions.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

For space agriculture specifically, the timeline pressure created by aggressive exploration schedules amplifies the importance of this foundational research. NASA's Artemis program aims to establish sustained human presence on the Moon by the late 2020s, with Mars missions potentially following in the 2030s (NASA, 2020). The lead time required for crop improvement—even with advanced genomic tools—necessitates early identification of priority targets and initiation of improvement programs. Molecular breeding approaches employing marker-assisted selection or genomic selection can accelerate cultivar development compared to conventional phenotypic selection, but still require multiple generations to achieve meaningful gains. Gene editing approaches offer more rapid implementation but require precise knowledge of target sequences and their functional consequences. In either case, the systematic characterization of HI determinants undertaken in this dissertation provides essential inputs for the strategic planning of BLSS crop development programs.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 1.1 Summary of Study Significance

Domain	Contribution	Beneficiaries
Theoretical	Unified pangenomic framework for HI analysis; evolutionary insights into biomass partitioning mechanisms; cross-species conservation patterns	Plant genomics researchers; evolutionary biologists; crop physiologists
Practical	Comprehensive HI gene catalog (30 genes, 5 categories); validated GWAS/QTL targets; structural variant inventory for breeding	Crop breeders; controlled environment agriculture practitioners; seed companies
Space Agriculture	Priority molecular targets for NASA crop development; cross-species improvement strategies; BLSS optimization framework	NASA; ESA; international space agencies; BLSS engineers; space habitat designers

1.6 Scope and Delimitations

This dissertation focuses specifically on four crop species—rice, wheat, potato, and soybean—selected based on their established priority for BLSS implementation and the availability of extensive genomic resources. The selection of these species reflects the consensus view in the space agriculture literature that these crops provide the nutritional foundation for BLSS food production (Wheeler, 2017; Perchonok et al., 2012). Other candidate BLSS crops, including leafy greens (lettuce, spinach), brassicas, tomato, and sweet potato, are excluded from

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

the primary analysis, though findings may have implications for these species through ortholog-based inference.

The focus on rice, wheat, potato, and soybean reflects a strategic prioritization based on multiple criteria beyond nutritional complementarity. These four crops collectively represent the major phylogenetic lineages of flowering plants relevant to agriculture—two monocots (rice, wheat) and two dicots (potato, soybean)—enabling evolutionary comparisons across the angiosperm radiation. They also represent the major categories of harvestable organs: cereal grains (rice, wheat), tubers (potato), and legume seeds (soybean), providing coverage of the diverse developmental pathways underlying yield determination. Practically, these species possess the most extensive genomic resources of any food crops, including high-quality reference genomes, large-scale pangenome projects, and comprehensive association mapping datasets, enabling the depth of analysis required to address the research questions.

The research employs a computational comparative genomics approach that leverages publicly available databases and published literature rather than generating novel experimental data. Data sources include the GWAS Atlas, Gramene, SNP-Seek, Triticeae Toolbox, SoyBase, Spud DB, and peer-reviewed publications reporting functional characterization of HI-related genes. The timeframe for literature review and database queries encompasses publications and database versions available through December 2025. This computational approach enables comprehensive cross-species analysis but does not include experimental validation of identified candidate genes through transgenic or genome editing approaches.

The exclusion of leafy vegetables and fruits from this analysis does not diminish their importance for BLSS nutrition but reflects the distinct nature of harvest index concepts in these

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

crops. For leafy greens such as lettuce and spinach, the concept of harvest index becomes less meaningful because the vegetative biomass itself constitutes the harvested product, collapsing the distinction between "economic yield" and "total biomass." For fruiting vegetables such as tomato and pepper, harvest index considerations apply but the genetic architecture differs substantially from that of grains, tubers, and legume seeds. Future research should extend pangenomic HI analysis to these additional BLSS candidate crops using frameworks adapted to their specific biology.

Several topics fall outside the scope of this dissertation despite their relevance to BLSS crop improvement. These include: (a) environmental response genes that may modulate HI under space-specific conditions such as microgravity or elevated radiation, which warrant separate investigation; (b) epigenetic factors influencing HI expression, which represent an emerging research frontier but lack sufficient cross-species data for comprehensive analysis; (c) gene-by-environment interactions specific to controlled environment agriculture, which require experimental validation beyond computational prediction; and (d) economic and engineering analyses of BLSS implementation, which fall within the domain of systems engineering rather than plant genomics.

1.7 Definition of Key Terms

Table 1.2 presents definitions of key terms employed throughout this dissertation to ensure consistency and clarity of communication.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 1.2 Glossary of Key Terms

Term	Definition
Bioregenerative Life Support System (BLSS)	A life support system that uses biological processes, primarily photosynthesis by plants and algae, to regenerate essential resources including oxygen, water, and food for sustained human habitation in closed environments.
Harvest Index (HI)	The ratio of economically valuable yield (grain, tuber, seed) to total aboveground plant biomass, expressed as a decimal or percentage; a measure of biomass partitioning efficiency.
Pangenome	The complete set of genes present in a species, including core genes found in all individuals and dispensable genes present in only some individuals.
Presence/Absence Variant (PAV)	A type of structural genomic variation in which a DNA segment is present in some individuals or cultivars but completely absent in others.
Copy Number Variant (CNV)	A structural genomic variation involving differences in the number of copies of a particular DNA segment among individuals.
Ortholog	Genes in different species that evolved from a common ancestral gene through speciation and typically retain similar function.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Genome-Wide Association Study (GWAS)	A statistical approach to identify genetic variants associated with phenotypic traits by testing correlations across large populations.
Quantitative Trait Locus (QTL)	A genomic region containing one or more genes that contribute to variation in a quantitative phenotypic trait.
Gibberellin (GA)	A class of plant hormones that regulate growth and development, including stem elongation; perturbations in GA signaling underlie semi-dwarf phenotypes.
Semi-dwarf	A plant phenotype characterized by reduced stem height due to altered gibberellin synthesis or signaling, typically associated with improved harvest index.
Controlled Environment Agriculture (CEA)	Crop production systems in which environmental parameters (light, temperature, humidity, CO ₂) are precisely regulated.

1.8 Assumptions

This research proceeds under several assumptions that readers should consider when interpreting findings. With respect to data quality, the study assumes that information in the queried genomic databases (GWAS Atlas, Gramene, SNP-Seek, Triticeae Toolbox, SoyBase, Spud DB) represents accurate curation of primary research findings and that reported gene functions and associations have been appropriately validated in the original studies. While these

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

databases implement quality control procedures, the possibility of errors or outdated annotations cannot be entirely excluded.

The assumption of functional conservation among orthologs merits particular scrutiny given its centrality to the cross-species analytical framework. While the general principle that orthologous genes perform similar functions is well-supported by evolutionary theory and empirical observation, exceptions occur with sufficient frequency to warrant caution in extrapolating functional predictions across species. Neofunctionalization, in which duplicated genes acquire novel functions, and subfunctionalization, in which ancestral functions are partitioned among paralogs, can obscure orthologous relationships and confound functional inference (Lynch & Conery, 2000). The deep divergence times separating monocots and dicots (approximately 150 million years) amplify these concerns, as substantial evolutionary change can accumulate even in functionally constrained sequences over such timescales. This dissertation addresses these concerns by focusing analysis on gene families with well-characterized functions in multiple species and by treating cross-species functional predictions as hypotheses requiring experimental validation rather than established facts.

From a methodological perspective, the ortholog identification approach assumes that sequence homology accurately reflects functional conservation—an assumption supported by extensive literature but not universally valid, particularly for rapidly evolving gene families. The study further assumes that HI-controlling mechanisms identified in field-grown crops remain relevant to controlled environment agriculture, though the altered environmental conditions of CEA may modulate the relative importance of specific genetic factors. Finally, the research assumes that publicly available pangenome assemblies adequately capture the structural variation

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

present in each species, recognizing that coverage varies substantially among crops and that novel variants continue to be discovered as additional genomes are sequenced.

The assumption of database accuracy reflects a pragmatic recognition that large-scale computational analyses necessarily rely on curated data resources that aggregate findings from primary literature. While databases such as GWAS Atlas, Gramene, and SoyBase implement quality control procedures and undergo periodic updates, they are not immune to errors of omission, misannotation, or outdated information. Particularly relevant for this study, the assignment of genes to functional categories can reflect the perspectives and priorities of the researchers who originally characterized them, potentially introducing biases that propagate through downstream analyses. To mitigate these concerns, this dissertation employs multiple database sources for cross-validation and returns to primary literature to verify key findings rather than relying solely on database annotations.

1.9 Organization of the Dissertation

This dissertation comprises six chapters organized to systematically address the research questions while meeting the conventions of doctoral scholarship. Following this introductory chapter, the dissertation proceeds through literature review, methodology, results, discussion, and conclusion. Figure 1.1 provides a visual representation of the dissertation structure and the relationships among chapters.

Chapter 2, Literature Review, establishes the scholarly foundation for the research by synthesizing current knowledge across four interconnected domains: space exploration and BLSS requirements, harvest index biology and genetics, pangenomic approaches in crop research, and the specific genetics of the four target species. This chapter identifies gaps in

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

existing knowledge that motivate the research questions and positions the dissertation within broader scholarly conversations.

Chapter 3, Methodology, details the four-phase computational approach employed to address the research questions. Phase 1 describes the systematic construction of the HI gene inventory through database mining and literature synthesis. Phase 2 presents the ortholog identification and phylogenomic analysis procedures. Phase 3 explains the integration of structural variant data with GWAS and QTL information. Phase 4 outlines the evolutionary analysis framework for characterizing selective patterns and allelic histories. The four-phase methodology was designed to progressively build understanding from gene-level inventory through species comparison to evolutionary interpretation.

Chapter 4, Results, presents the findings organized according to the four analytical phases. This chapter reports the 30-gene HI catalog with functional classifications, the five conserved ortholog pathways identified across species, the integration of 20 GWAS hits and 25 QTLs with structural variant data, and the evolutionary patterns characterizing HI determinants. The results are presented with appropriate statistical metrics and visualization to enable reader evaluation of evidence quality and effect magnitude.

Chapter 5, Discussion, interprets the results in relation to the literature reviewed in Chapter 2, addresses each research question, evaluates theoretical and practical implications, acknowledges limitations, and proposes directions for future research. Particular attention is given to the implications for BLSS crop improvement and priority targets for NASA development programs. The theoretical implications for understanding plant biomass partitioning

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

evolution are also developed, positioning this dissertation within broader scholarly conversations about crop domestication and adaptation.

Chapter 6, Conclusion, synthesizes the principal contributions of the dissertation, restates key findings in relation to research objectives, articulates the significance of the work for scholarship and practice, and offers final recommendations for advancing harvest index research and BLSS crop development. The conclusion articulates how this dissertation advances the goal of sustainable human presence beyond Earth by providing foundational knowledge essential for crop optimization in space agriculture systems.

1.10 Chapter Summary

This introductory chapter has established the foundation for a comprehensive pangenomic investigation of harvest index determinants across four BLSS staple crops. The chapter opened by contextualizing the research within the broader imperative of developing sustainable life support systems for long-duration human space exploration, emphasizing the critical role of harvest index in optimizing crop production efficiency within resource-constrained space habitats. The historical success of the Green Revolution in improving harvest index through semi-dwarf genetics was presented as both precedent and inspiration for BLSS crop optimization efforts.

The problem statement identified three key gaps in current knowledge: the absence of a systematic cross-species pangenomic analysis of HI genes, insufficient characterization of structural variants at HI loci, and unclear evolutionary patterns underlying HI regulation. These gaps impede both theoretical understanding of biomass partitioning mechanisms and practical efforts to develop BLSS-optimized cultivars. The fragmented nature of existing research,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

organized primarily by species rather than by functional process, compounds these challenges by limiting knowledge transfer across crops.

The purpose statement articulated the dissertation's aim to conduct comprehensive pangenomic analysis of HI-controlling genes in rice, wheat, potato, and soybean, with four specific objectives addressing gene cataloging, ortholog identification, structural variant characterization, and evolutionary analysis. Four research questions were formulated to operationalize these objectives and guide the investigation through a systematic analytical progression from gene inventory to evolutionary interpretation.

The significance section demonstrated the theoretical, practical, and space agriculture contributions of this research, establishing its relevance for plant genomics scholarship, terrestrial crop breeding programs, controlled environment agriculture, and NASA's space exploration objectives. Table 1.1 summarized these contributions and their intended beneficiaries. The scope and delimitations established boundaries that ensure feasibility without compromising scholarly rigor, while acknowledging the limitations inherent in a computational approach that does not include experimental validation.

Key terms were defined in Table 1.2 to ensure clarity of communication throughout the dissertation, providing readers with consistent definitions for specialized vocabulary essential to understanding the research. Assumptions underlying the research were made explicit, particularly those related to data quality in genomic databases and the transferability of functional annotations across species through ortholog relationships.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The imperative to develop sustainable food production systems for long-duration human space exploration represents one of the most complex interdisciplinary challenges facing modern science. As humanity sets its sights on establishing permanent settlements on the Moon and Mars, the traditional paradigm of resupplying crews from Earth becomes increasingly untenable. The prohibitive costs associated with launching consumables, combined with the logistical complexities of multi-year transit times to Mars, necessitate a fundamental shift toward self-sufficient, closed-loop systems capable of regenerating essential resources in situ. Bioregenerative Life Support Systems (BLSS) embody this paradigm shift, integrating biological and physical-chemical processes to recycle air, water, and waste while producing fresh food for astronaut crews. Within this framework, the cultivation of staple crops—specifically rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*)—emerges as the cornerstone of any truly sustainable food production system, providing the caloric density and macronutrient balance essential for human health and performance (Wheeler, 2017).

Central to the optimization of these crop production systems is the concept of harvest index (HI), the ratio of economically valuable yield to total plant biomass. In the resource-constrained environment of a space habitat, where every gram of mass, watt of power, and liter of volume must be meticulously optimized, maximizing harvest index translates directly to maximizing system efficiency. A higher HI ensures that the substantial resources invested in growing a plant—light energy, water, nutrients, and atmospheric gases—are converted primarily into edible food rather than into structural biomass that must subsequently be processed or

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

discarded. The historical significance of harvest index improvement was demonstrated unequivocally during the Green Revolution, when the introduction of semi-dwarf genes in wheat and rice increased HI from approximately 0.30 to over 0.50, effectively doubling the proportion of plant biomass allocated to grain and revolutionizing global food production (Hedden, 2003; Khush, 2001). Extending these gains to the context of BLSS requires a sophisticated understanding of the genetic determinants of biomass partitioning and their behavior under the unique environmental conditions of controlled-environment agriculture.

The emergence of pangenomics as a transformative framework for crop genetics provides unprecedented opportunities to address this challenge. Traditional genomics, constrained by reliance on single reference genomes, inherently misses the substantial genetic diversity present within and among crop species. Structural variations—including presence/absence variations (PAVs), copy number variations (CNVs), inversions, and translocations—that are absent from a reference genome yet play critical roles in agronomic performance remain invisible to conventional approaches. The pangenome concept, first articulated by Tettelin et al. (2005) in the context of bacterial genetics and subsequently extended to plants, recognizes that the complete genetic repertoire of a species encompasses both a core genome shared by all individuals and a dispensable (accessory) genome comprising genes and sequences present in only some members of the species. In crop plants, this dispensable genome has proven to be a rich reservoir of functional variation, harboring genes for environmental adaptation, stress tolerance, and agronomically important traits including those governing harvest index (Bayer et al., 2020; Golicz et al., 2020).

This dissertation investigates the pangenomic architecture of harvest index determinants in the four staple crops most relevant to BLSS applications: rice, wheat, potato, and soybean. By

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

systematically characterizing the structural variation landscape at loci governing biomass partitioning, this research aims to identify novel genetic variants—potentially including those absent from reference genomes—that can be leveraged to enhance harvest index under the specific conditions of controlled-environment agriculture for space applications. The research is motivated by the hypothesis that the dispensable genome of these crops contains functionally significant variation at key harvest index loci that has been overlooked by reference-based approaches, and that characterizing this variation through pangenomic analysis will reveal new targets for genetic improvement.

This literature review establishes the conceptual and empirical foundation for this investigation. It synthesizes current knowledge across four interconnected domains that collectively inform the research questions and methodological approach. Section 2.2 examines the theoretical framework and methodological advances in pangenomics, tracing the evolution of the concept from microbiology to plant science and describing the technologies and computational approaches that enable the construction and analysis of plant pangenomes. Section 2.3 provides a comprehensive treatment of harvest index, encompassing its physiological basis in source-sink relationships and carbon allocation, its historical importance in the Green Revolution, its genetic architecture as revealed by quantitative trait locus mapping and genome-wide association studies, and the molecular pathways through which key genes regulate biomass partitioning. Section 2.4 reviews the genomic landscape of each of the four target crops, detailing the history and characteristics of their reference genomes, the architecture and content of their pangenomes, the specific genes governing harvest index and yield components, and the application of this knowledge in breeding programs. Section 2.5 situates this genomic research within the applied context of BLSS and space agriculture, examining the principles of

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

bioregenerative life support, the criteria for crop selection, the technologies of controlled-environment agriculture, and the specific challenges and opportunities for optimizing harvest index in space environments.

Section 2.6 synthesizes these four thematic areas, identifying the points of intersection and the gaps in current knowledge that this dissertation addresses. The synthesis highlights how the integration of pangenomic approaches with harvest index optimization represents a novel frontier for BLSS crop improvement. By connecting the rich genetic diversity revealed by pangenome studies to the specific physiological and genetic determinants of biomass partitioning, this research occupies a unique niche at the interface of genomics, crop physiology, and space agriculture. The chapter concludes with Section 2.7, a summary of the key points and their implications for the dissertation's research design.

The overarching goal of this literature review is to demonstrate that while substantial progress has been made in understanding both pangenomic variation and harvest index genetics in staple crops, significant knowledge gaps remain at their intersection. Specifically, little research has systematically examined whether the dispensable genome of these species harbors functionally important variation at harvest index loci, nor has pangenomic analysis been applied to the specific challenge of optimizing crops for BLSS. This dissertation aims to fill these gaps, providing both fundamental insights into the pangenomic architecture of harvest index and practical guidance for breeding or engineering space-optimized crop cultivars. The research thus contributes to the broader scientific and societal goal of enabling sustainable human presence beyond Earth through the development of robust bioregenerative food production systems.

2.2 Pangenomics: Concepts, Methods, and Applications

2.2.1 History and Evolution of the Pangenomics Concept

The conceptualization of a species' genome has undergone a profound transformation over the past two decades, moving from a static, monolithic entity represented by a single reference sequence to a dynamic and comprehensive collection of all genetic material present within a species. This paradigm shift was catalyzed by the introduction of the pangenome concept, a term first coined in the field of microbiology. The seminal work of Tettelin and colleagues in 2005, studying the bacterium *Streptococcus agalactiae*, or Group B *Streptococcus* (GBS), is widely recognized as the formal genesis of this idea (Tettelin et al., 2005). The research originated not from a primary interest in genomic theory but from a practical, applied goal: the development of a novel vaccine for GBS, a significant cause of neonatal infections. The project, a collaboration between Chiron Vaccines and The Institute for Genomic Research (TIGR), employed a reverse vaccinology approach, which leverages genomic data to identify potential protein-based vaccine candidates (Medini et al., 2005).

At the time, sequencing multiple strains of the same bacterial species was not standard practice. However, the initial comparative analysis of the first few available GBS genomes revealed an unexpectedly high degree of genomic diversity, challenging the prevailing assumption that a single genome could adequately represent an entire species. By comparing the genomes of eight distinct GBS strains, representing the five major disease-causing serotypes, Tettelin et al. (2005) made a startling discovery. They found that the set of genes shared by all eight strains—which they termed the core genome—comprised only about 80% of the coding potential of any single genome. Furthermore, each newly sequenced genome contributed a significant number of unique genes not found in the others. This observation led to the critical

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

question of how many genomes would need to be sequenced to capture the entire gene repertoire of the species.

To address this, the researchers developed a mathematical model to extrapolate the size of the species' total gene pool. The model revealed that the number of new genes discovered with each additional genome sequenced did not converge toward zero. Instead, it suggested that new genes would continue to be identified even after sequencing hundreds of strains, leading to the conclusion that the genetic repertoire of *S. agalactiae* was effectively unbounded or open. In their landmark publication, they formally defined the pangenome, derived from the Greek word pan (παν), meaning whole or everything, as the sum of the core genome and the dispensable genome, which includes genes present in some but not all strains, as well as strain-specific unique genes (Tettelin et al., 2005). This discovery fundamentally altered the understanding of bacterial species, revealing them to be genetically fluid entities with a vast and accessible gene pool that contributes to their adaptability and evolutionary success.

The pangenome concept, having revolutionized microbiology, was subsequently and logically extended to eukaryotes, including plants. The sequencing of the first plant genome, *Arabidopsis thaliana*, in 2000, followed by the first crop genome, rice (*Oryza sativa*), marked the beginning of the plant genomics era (The Arabidopsis Genome Initiative, 2000; Goff et al., 2002). As sequencing technologies became more affordable and efficient, the number of available plant genomes grew exponentially. The accumulation of this vast genomic data from multiple individuals within the same species made it increasingly clear that the single reference genome model was insufficient to capture the full breadth of genetic diversity present in plants. Researchers observed extensive structural variations (SVs)—large-scale genomic rearrangements such as insertions, deletions, copy number variations, and inversions—that were absent from the

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

reference sequence but prevalent within the species' population (Bayer et al., 2020). These SVs were found to play critical roles in trait innovation, environmental adaptation, and agronomic performance. Consequently, the pangenome framework was adopted by the plant science community as an essential tool to build a more complete and unbiased representation of a species' genetic information, paving the way for a new era of discovery in crop improvement and evolutionary biology (Golicz et al., 2020; Zhao et al., 2024).

2.2.2 Theoretical Framework: Core Genome vs. Dispensable Genome

The theoretical foundation of pangenomics rests upon the partitioning of a species' total genetic repertoire into distinct functional categories based on the frequency of their occurrence across multiple individuals. This framework provides a powerful lens through which to understand the evolutionary forces shaping a species' genome and the genetic basis of its phenotypic plasticity. The two primary components of this framework are the core genome and the dispensable genome, also frequently referred to as the accessory genome (Tettelin et al., 2005; Golicz et al., 2020).

The core genome is defined as the set of genes or DNA sequences that are shared by all, or nearly all, individuals within a species or a defined taxonomic group. These sequences are considered fundamental to the organism's basic biology and survival. They typically encode proteins involved in essential housekeeping functions, such as DNA replication, transcription, translation, primary metabolism, and cellular structure (Golicz et al., 2020). Due to their critical roles, core genes are generally under strong purifying selection, exhibiting lower rates of mutation and evolutionary change compared to other parts of the genome. In a study of wild soybean (*Glycine soja*), for example, the core genome was estimated to constitute approximately 80% of the pangenome, reflecting the conserved genetic foundation of the species (Li et al.,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

2014). The core genome represents the stable genetic backbone of a species, ensuring the continuity of its fundamental life processes across generations.

In contrast, the dispensable genome comprises all genes and sequences that are not universally present across all individuals of a species. This component is a mosaic of genes found in a subset of the population and genes that are unique to a single individual, sometimes referred to as private genes or singletons (Golicz et al., 2020). Initially, the term dispensable was adopted from the bacterial literature, where these genes were often associated with mobile genetic elements and considered non-essential for basic survival under laboratory conditions. However, in the context of plant genomics, this term is increasingly viewed as a misnomer. The genes within the plant dispensable genome are far from expendable; they are now recognized as a critical source of genetic novelty and a major driver of adaptation and phenotypic diversity (Bayer et al., 2020).

The dispensable genome is often enriched for genes involved in interactions with the environment. These include genes conferring resistance to biotic stresses, such as pathogens and pests, as well as those involved in tolerance to abiotic stresses like drought, salinity, and extreme temperatures. For instance, many disease resistance genes, such as those containing nucleotide-binding site leucine-rich repeat (NBS-LRR) domains, are frequently found within the dispensable genome of crops like tomato and rice (Bayer et al., 2020; Golicz et al., 2020). This variability allows a species to maintain a diverse arsenal of defense mechanisms at the population level, enabling it to adapt to a wide range of and co-evolving pathogens. Beyond stress response, the dispensable genome also harbors genes that influence key developmental processes and agronomically important traits, providing a rich reservoir of genetic variation for crop improvement. The recognition of the dispensable genome's functional importance

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

underscores the profound limitation of relying on a single reference individual, which by definition lacks this entire catalog of variable genes, to represent the full genetic potential of a species.

2.2.3 Pangenome Construction Methodologies

The construction of a pangenome is a complex bioinformatic challenge that involves assembling and comparing genomic data from multiple individuals to create a unified representation of a species' genetic diversity. The methodologies employed have evolved significantly, driven by advances in sequencing technology and computational algorithms. Broadly, these approaches can be categorized into three main strategies: reference-based iterative assembly, de novo assembly and comparison, and graph-based pangenome assembly (Bayer et al., 2020; Liu et al., 2023). Each method possesses distinct advantages and limitations regarding computational requirements, accuracy, and the ability to capture different types of genetic variation.

The reference-based iterative method leverages a pre-existing high-quality reference genome as a starting scaffold. In this approach, sequencing reads from additional individuals (accessions or varieties) are first aligned to the reference sequence. The reads that map successfully are used to identify small-scale variations like single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) relative to the reference. The crucial step involves collecting all reads that fail to align to the reference. These unmapped reads are presumed to represent novel sequences—genes or intergenic regions—that are absent from the reference individual. These reads are then subjected to de novo assembly to construct new contigs, which are subsequently integrated into the pangenome representation (Liu et al., 2023). This map-to-pan strategy is computationally efficient and benefits from the structural information provided by

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

a well-annotated reference, which can simplify downstream analysis. However, its primary drawback is a persistent reference bias. Sequences that are highly divergent from the reference may fail to be assembled correctly or may be missed entirely, and complex structural rearrangements relative to the reference can be difficult to resolve accurately.

The second major strategy, *de novo* assembly and comparison, avoids reference bias by assembling the genome of each individual from scratch, independent of any pre-existing reference. This method involves generating high-quality genome assemblies for every sample in the study population, typically using long-read sequencing data to achieve high contiguity. Once the individual assemblies are complete, they are compared to one another using whole-genome alignment algorithms to identify shared (core) and variable (dispensable) regions (Bayer et al., 2020; Liu et al., 2023). For species with large, complex, or polyploid genomes, such as wheat, whole-genome alignment can be computationally prohibitive. In such cases, an alternative gene-based approach is often used, where genes are predicted in each assembly, and orthology inference programs are used to cluster them into gene families, thereby identifying the core and accessory gene sets (Golicz et al., 2020). This method is considered the most powerful for discovering all forms of genetic variation, from SNPs to large-scale structural variants, as it is not constrained by a reference. The first bacterial pangenome of *Streptococcus agalactiae* and the first plant pangenome of wild soybean were constructed using this approach (Tettelin et al., 2005; Li et al., 2014). The main limitations are the high cost and significant computational resources required to generate and analyze multiple high-quality *de novo* assemblies.

The most recent and sophisticated approach involves the construction of a graph-based pangenome. This method moves beyond the linear, one-dimensional representation of a genome and instead models the pangenome as a complex graph structure. In a sequence graph, nodes

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

represent DNA sequences, and edges represent the connections or adjacencies between them (Garrison et al., 2020; Zhao et al., 2024). A path through the graph can represent the complete chromosome of an individual, while bubbles or alternative paths within the graph represent genetic variations such as SNPs, indels, and structural variants. This integrated data structure can be built either by adding known variants to a linear reference or, more powerfully, by aligning multiple de novo assemblies to each other. Graph-based pangenomes inherently overcome reference bias by treating all constituent genomes more equitably and provide a natural framework for representing complex structural variations that are difficult to describe in a linear system (Garrison et al., 2020). Aligning sequencing reads to a graph can lead to more accurate variant calling and genotyping. While this approach offers a more comprehensive and accurate representation of genomic diversity, its primary challenges lie in the computational complexity of graph construction, traversal, and visualization, especially as the number of included genomes increases (Liu et al., 2023).

2.2.4 Structural Variation in Plant Genomes

While single nucleotide polymorphisms (SNPs) have historically been the primary focus of population genetics and breeding, it is now widely accepted that larger-scale structural variations (SVs) are a major source of genetic and phenotypic diversity in plants. SVs are typically defined as genomic rearrangements larger than 50 base pairs and include a wide array of alterations such as presence-absence variations (PAVs), copy number variations (CNVs), inversions, and translocations (Alonge et al., 2020; Gabur et al., 2022). These variations can have profound effects on gene function, expression, and regulation, and they are increasingly recognized as key drivers of adaptation and important agronomic traits. Pangenomic analyses are

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

particularly well-suited for the comprehensive discovery and characterization of SVs, as many are missed when mapping reads to a single linear reference genome.

Presence-absence variation (PAV) represents a significant form of structural variation where entire genes or large DNA segments are present in the genomes of some individuals within a species but completely absent in others (Tan et al., 2020; Zhang et al., 2022). It can be considered an extreme form of copy number variation where the copy number is either one (or more) or zero. Pangenome studies across numerous plant species, including staple crops like maize, rice, and *Brassica napus*, have revealed that thousands of genes exhibit PAV across different accessions (Swanson-Wagner et al., 2010; Tan et al., 2020). These variable genes are not randomly distributed across the genome or functional categories. They are often enriched for roles in environmental interaction, particularly disease resistance. For example, nucleotide-binding site leucine-rich repeat (NBS-LRR) genes, which form the largest class of plant disease resistance genes, are frequently subject to PAV, providing a dynamic reservoir of resistance specificities at the population level. Beyond defense, PAVs have been linked to a wide range of agronomic traits, including flowering time, quantitative disease resistance in *Brassica napus*, and fruit characteristics in melon, making them a critical target for crop improvement (Zhang et al., 2022).

Copy number variation (CNV) refers to differences in the number of copies of a specific DNA segment among individuals. These segments can range from a few hundred base pairs to several megabases and can encompass entire genes or gene clusters (Zhang et al., 2021). CNVs arise from genomic events like unequal crossing-over during meiosis, often mediated by non-allelic homologous recombination, or errors during DNA replication (Cook & Mlynarczyk, 2013). The resulting change in gene dosage can directly impact phenotype by altering the

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

expression level of the affected genes. In crop species, CNVs have been shown to be a major contributor to phenotypic variation for important traits. For instance, in soybean, variation in the copy number of the *Rhg1* locus is a primary determinant of resistance to the soybean cyst nematode (Cook et al., 2012). In maize, CNVs are widespread and have been associated with traits such as disease resistance to Goss's Wilt, plant height, and flowering time. In autotetraploid potato, large-scale CNVs are particularly prevalent and are enriched in gene clusters related to environmental stress response, suggesting a role in rapid adaptation (Zhang et al., 2021).

Inversions and translocations represent another major class of SVs involving the rearrangement of chromosomal segments. An inversion occurs when a segment of a chromosome is excised, flipped 180 degrees, and reinserted, reversing the gene order within that segment (Bennetzen & Wang, 2014; Lye & Purugganan, 2020). A translocation involves the movement of a genomic segment from one location to another, either on the same or a different chromosome. These rearrangements can have significant functional consequences. They can disrupt genes at their breakpoints, alter gene expression by moving genes to new regulatory contexts (position effects), or create novel fusion genes. A key evolutionary role of inversions is the suppression of recombination within the inverted region in heterozygous individuals. This can lead to the formation of supergenes, where co-adapted sets of alleles are inherited together as a single unit, facilitating local adaptation and even speciation (Lye & Purugganan, 2020). In agriculture, SVs have been linked to key domestication and improvement traits. A well-known example is the creation of modern sweet corn, which resulted from an inversion that caused a loss-of-function mutation in the *shrunken-2* gene (Wurtzel et al., 1999). Similarly, a large inversion in peach is associated with variation in fruit flesh color and shape (Li et al., 2019). The accurate detection of

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

these large-scale rearrangements has been historically challenging but is now greatly facilitated by long-read sequencing and pangenomic assembly approaches.

2.2.5 Technological Advances Enabling Plant Pangenomics

The transition from a single-reference paradigm to a pangenomic framework in plant science has been directly enabled by a suite of transformative technological advancements in genomics. The sheer complexity of many plant genomes—often characterized by large sizes, high degrees of repetitiveness, heterozygosity, and polyploidy—posed significant barriers to the assembly of high-quality genomes, a prerequisite for pangenome construction. The development of long-read sequencing platforms, coupled with sophisticated scaffolding technologies like optical mapping and Hi-C, has been instrumental in overcoming these challenges, making the routine generation of multiple chromosome-level genome assemblies for a given species a feasible endeavor.

The advent of third-generation, long-read sequencing technologies, principally from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), marked a pivotal moment for plant genomics (Jiao & Schneeberger, 2020; Ou et al., 2023). Unlike second-generation short-read technologies, which produce reads of a few hundred base pairs, long-read platforms can generate reads tens to hundreds of kilobases in length, and in some cases, over a megabase. This dramatic increase in read length is critical for resolving complex genomic regions. Long reads can span entire repetitive elements, such as long terminal repeat (LTR) retrotransposons, which are abundant in crops like maize and wheat and often cause assemblies to fragment into thousands of small contigs when using short reads alone (Jiao & Schneeberger, 2020). PacBio's High-Fidelity (HiFi) sequencing, which involves circular consensus sequencing, offers the dual advantage of long read lengths (typically 15-20 kb) and exceptionally high base

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

accuracy (>99.9%), making it a powerful tool for generating reference-quality assemblies and phasing haplotypes in diploid and polyploid species. ONT sequencing, while having a higher raw error rate, offers the advantage of producing ultra-long reads and the unique ability to directly sequence RNA molecules, providing valuable transcriptomic information (Ou et al., 2023). The ability of these technologies to produce highly contiguous assemblies is the bedrock upon which modern plant pangenomics is built.

While long-read sequencing can produce highly contiguous contigs, ordering and orienting these contigs into chromosome-level scaffolds requires long-range structural information. Two key technologies, optical mapping and chromatin conformation capture (Hi-C), provide this essential scaffolding capability (Jiao & Schneeberger, 2020; Maretty & Jensen, 2021). Optical mapping, commercialized by Bionano Genomics, generates a genome-wide, high-resolution physical map. The process involves extracting ultra-high-molecular-weight DNA, labeling it at specific sequence motifs with fluorescent tags, and imaging the linearized DNA molecules as they pass through nanochannels. The resulting patterns of labels create a unique barcode for different genomic regions, providing a physical map with structural information spanning hundreds of kilobases to megabases. This map can be used to validate contigs from a sequence assembly, identify misassemblies, and order and orient the contigs into large scaffolds (Maretty & Jensen, 2021).

Hi-C is a proximity-ligation based method that provides orthogonal, long-range information by capturing the three-dimensional organization of chromatin within the nucleus (Ghurye et al., 2020). The underlying principle is that genomic loci that are close together in the linear sequence of a chromosome will also be in frequent physical contact within the nucleus. The Hi-C protocol involves cross-linking chromatin to fix these interactions, digesting the DNA,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

ligating interacting fragments together, and then sequencing these chimeric junctions. By mapping the paired-end reads back to an initial draft assembly, a contact frequency map can be generated. This map reveals strong interaction signals between contigs that belong to the same chromosome and allows for their correct ordering and orientation into chromosome-scale scaffolds. The synergistic use of long-read sequencing for initial contig generation, followed by scaffolding with optical mapping and/or Hi-C data, has become the gold standard for producing the high-quality, chromosome-level assemblies necessary for constructing robust and comprehensive plant pangenomes.

2.2.6 Major Plant Pangenome Projects and Their Findings

The application of pangenomic approaches to key staple crops has yielded a wealth of new biological insights and created invaluable resources for breeding and genetic research. Landmark projects in rice, maize, tomato, wheat, and soybean have not only characterized the vast extent of genomic diversity within these species but have also linked this variation to important agronomic traits, domestication history, and adaptive potential. These studies serve as exemplars of the power of the pangenome concept in plants.

One of the most ambitious and impactful initiatives was the 3,000 Rice Genomes Project (3K RGP). This international collaboration sequenced the genomes of 3,010 diverse accessions of Asian cultivated rice (*Oryza sativa*) from 89 countries, creating a massive public resource for rice improvement (The 3K RGP, 2014; Wang et al., 2018). The analysis of this dataset revealed an enormous catalog of genetic variation, including approximately 18.9 million SNPs, millions of small indels, and over 90,000 structural variations relative to the Nipponbare reference genome. Pangenome analysis identified more than 10,000 novel protein-coding genes that were absent in the reference, highlighting the significant amount of genetic information missed by a

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

single-genome approach (Wang et al., 2018). The project provided unprecedented resolution into the population structure of rice, confirming the five major varietal groups (indica, japonica, Aus/Boro, Basmati/Sadri) and further subdividing them into nine distinct subpopulations, often correlated with geographic origin. The patterns of diversity at key domestication loci, such as the non-shattering gene *Sh4* and the semi-dwarf green revolution gene *sd1*, provided detailed insights into the complex selection histories that shaped modern rice varieties.

In maize, another critical global staple, the construction of a pan-Zea genome has similarly expanded the known genetic landscape of the species and its wild relatives, teosintes (Hufford et al., 2021; Sun et al., 2022). By combining multiple high-quality reference assemblies with data from hundreds of resequenced maize inbred lines and teosintes, researchers constructed a pangenome that captured a vast amount of non-reference sequence. This work annotated nearly 59,000 pan-Zea genes, of which over 44% were classified as dispensable, or variably present across the population (Sun et al., 2022). This dispensable genome is a rich source of adaptive variation. Genome-wide association studies (GWAS) that incorporated structural variations, including gene presence-absence variations (gPAVs), identified numerous loci associated with complex agronomic traits that were missed in SNP-based studies. The high prevalence of transposable elements in the maize genome is a major driver of this structural diversity, contributing to differences in genome size and gene content among lines. The pan-Zea resource is crucial for understanding maize domestication, identifying alleles from teosinte that could be used for crop improvement, and dissecting the genetic architecture of traits like stress tolerance and yield.

Pangenome studies in other major crops have yielded similarly important findings. In tomato, a graph-based pangenome constructed from over 800 genomes, including cultivated

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

tomato and its wild progenitor, has been instrumental in identifying previously hidden structural variants linked to agronomically important genes (Gao et al., 2019). More recently, a tomato super-pangenome was developed, incorporating chromosome-level assemblies from nine wild *Solanum* species. This resource provides a comprehensive view of genomic diversity across the genus and is being used to discover novel genes from wild relatives that can enhance traits like yield and stress resistance in cultivated varieties (Gao et al., 2023). In wheat, the first pangenome for a polyploid species was constructed, revealing extensive variation in gene content across different bread wheat varieties, which is critical for breeding in this complex crop (Montenegro et al., 2017). The first plant pangenome was constructed in soybean, using seven de novo assemblies, which identified a large presence-absence variant associated with seed luster (Li et al., 2014). These projects collectively demonstrate that the dispensable genome is a key reservoir of functional variation that is essential for understanding plant biology and accelerating crop improvement.

2.2.7 Pangenomics Applications in Crop Improvement

The development of pangenome resources has transitioned from a theoretical exercise in characterizing genomic diversity to a powerful, practical tool that is reshaping modern crop improvement and breeding strategies. By providing a more complete and unbiased catalog of a species' genetic variation, pangenomes enhance the power and precision of genetic analysis, facilitate the discovery of novel alleles for desirable traits, and improve the accuracy of predictive breeding methods. The integration of pangenomic data into breeding pipelines is accelerating the development of new crop varieties with improved yield, resilience, and quality.

A primary application of pangenomics is the enhancement of trait mapping studies, such as genome-wide association studies (GWAS) and quantitative trait locus (QTL) mapping.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Traditional GWAS relies on mapping sequencing reads from a diverse population to a single linear reference genome to call SNPs, which are then tested for association with a phenotype of interest. This approach suffers from reference bias; structural variants like PAVs and large insertions that are absent from the reference cannot be discovered, and reads from these regions are often discarded, leading to missed associations (Bayer et al., 2020; Wang et al., 2025).

Pangenome-based GWAS overcomes this limitation by incorporating SVs directly into the association analysis. For example, a meta-GWAS in rice that used a pangenome graph to genotype both SNPs and PAVs across a large population significantly improved the detection of QTLs for six key agronomic traits, increasing the explained heritability by up to 38% and identifying 116 novel QTLs that were invisible to SNP-only analyses (Wang et al., 2025). Similarly, a graph-based pangenome in tomato increased the average estimated heritability for gene expression and metabolite traits by 24% compared to a single reference, largely by resolving associations with previously hidden causal SVs (Gao et al., 2022).

Pangenomes are also a critical tool for discovering novel genes and alleles that confer tolerance to biotic and abiotic stresses, a key objective for ensuring food security in the face of climate change. The dispensable genome, in particular, is a reservoir of genes involved in environmental adaptation. Pangenomic studies have successfully identified specific genes and structural variants linked to stress tolerance across a range of crops. For instance, analysis of the pan-Zea genome identified a transposon insertion upstream of a key gene that suppressed its expression under drought conditions, improving survival rates (Wang et al., 2025). In rice, a super-pangenome analysis pinpointed specific alleles in genes like SNORKEL1/2 that are responsible for submergence adaptation in African rice. By expanding the search for beneficial alleles beyond the cultivated gene pool to include wild relatives, super-pangenome approaches

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

are particularly powerful for identifying novel sources of resilience that can be introgressed into elite breeding lines.

Furthermore, pangenomic resources are poised to significantly improve the accuracy and efficiency of genomic selection (GS). GS is a breeding method that uses genome-wide markers to predict the breeding value of individuals, allowing breeders to select the best candidates without lengthy and expensive field phenotyping. The accuracy of GS models depends on the extent to which the markers capture the genetic variation underlying the target traits. By providing a more comprehensive set of variants, including SVs, pangenomes can be used to build more accurate predictive models (Bayer et al., 2020; Esmaeilpour et al., 2024). This leads to higher rates of genetic gain per breeding cycle. The integration of pangenomics with other advanced technologies, such as high-throughput phenotyping, multi-omics data, and artificial intelligence, promises to create a synergistic platform for pangenomics-driven molecular breeding, accelerating the development of next-generation crops tailored for sustainable agriculture.

2.2.8 Comparative Pangenomics and Super-Pangenomes

As pangenomics has matured within individual species, the field is now expanding to a broader, more comparative framework. Comparative pangenomics, which involves the analysis and comparison of pangenomes across different but related species, offers profound insights into the evolutionary processes that drive speciation, adaptation, and the diversification of traits. A key concept within this framework is the super-pangenome, which extends the pangenome of a single species to encompass the entire genomic diversity at a higher taxonomic level, such as a genus, by incorporating multiple species, particularly the wild relatives of cultivated crops (Gao

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

et al., 2023; Wang et al., 2025). This approach provides an unprecedented opportunity to explore the vast, untapped genetic reservoir of wild species to accelerate crop improvement.

The construction of a super-pangenome provides a holistic view of the gene content and structural variation across an entire genus or clade. It allows researchers to partition the collective gene set into categories such as genus-conserved (core) genes, which represent the fundamental genetic toolkit of the lineage, and species-specific (private) genes, which may contribute to unique adaptations and phenotypes (Zhao et al., 2024; Gao et al., 2023). The primary motivation for building super-pangenomes for crop genera is to overcome the genetic bottlenecks imposed by domestication and modern breeding. Cultivated crops often have a narrow genetic base, having lost many alleles that were present in their wild ancestors. These lost alleles may control valuable traits, such as resistance to emerging diseases or tolerance to environmental stresses to which the wild relatives are adapted. By creating a super-pangenome that includes these wild species, breeders can systematically identify and access this hidden genetic diversity.

Recent super-pangenome projects have already demonstrated the immense potential of this approach. A super-pangenome of tomato, which integrated high-quality genomes from nine wild *Solanum* species with cultivated tomato, provided deep insights into the phylogeny of the group and identified extensive structural variation that distinguishes the species (Gao et al., 2023). This resource enabled the discovery of a novel gene from a wild relative that has the potential to increase yield in modern tomato varieties. Similarly, a super-pangenome of North American wild grape species was used to perform a graph-based GWAS for salinity tolerance, a critical trait for viticulture in many regions. This analysis identified a key genetic locus associated with chloride exclusion, a mechanism for salt tolerance, highlighting a valuable target

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

for breeding salt-tolerant rootstocks (Massonnet et al., 2023). These examples illustrate how super-pangenomes serve as a bridge, connecting the rich genetic diversity of wild relatives to the targeted improvement of cultivated crops.

Beyond its direct applications in breeding, comparative pangenomics provides a powerful framework for fundamental evolutionary studies. By comparing the structure and content of core and accessory genomes across a phylogeny, scientists can reconstruct the evolutionary history of gene families, trace the origins of novel traits, and understand the genomic mechanisms that drive adaptation and speciation. Structural variations are recognized as pivotal drivers of this evolutionary change, remodeling genomes and creating the raw material for natural selection (Bayer et al., 2020; Ou et al., 2024). For example, comparative pangenomic analyses in plant families like Rosaceae and legumes have revealed how processes like whole-genome duplication, transposable element activity, and tandem gene duplication have driven the expansion and diversification of important gene families, such as those involved in disease resistance (the pan-NLRome). By examining these patterns across multiple species, researchers can distinguish lineage-specific adaptations from conserved evolutionary mechanisms, providing a deeper understanding of the rules that govern plant genome evolution.

2.3 Harvest Index: Physiological and Genetic Foundations

2.3.1 Definition and Historical Development

The concept of Harvest Index (HI), a cornerstone of modern crop physiology and plant breeding, represents a critical measure of a plant's reproductive efficiency. It quantifies the proportion of total assimilated biomass that is allocated to the economically valuable, harvestable portion of the crop. The formalization of this concept is largely attributed to the seminal work of C.M. Donald in the 1960s (Donald, 1962). In his initial formulation, Donald defined the harvest

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

index simply as the ratio of the weight of the grains to the total above-ground dry matter of the plant at maturity. This straightforward metric provided a powerful new lens through which to evaluate crop performance, shifting the focus from merely total biomass production to the efficiency of partitioning that biomass into yield (Donald & Hamblin, 1976).

A few years later, in a highly influential 1968 paper titled *The breeding of crop ideotypes*, Donald expanded upon this definition, framing HI not just as a ratio but as a physiological measure of photosynthate partitioning. This work was revolutionary because it introduced the crop ideotype concept—a theoretical model of a plant designed for optimal performance in a specific environment (Donald, 1968). Donald proposed that instead of selecting for yield alone, which is a complex and environmentally sensitive trait, breeders should focus on selecting for a suite of morphological and physiological traits that would logically lead to higher yield. Within this framework, a high harvest index was posited as a primary component of the ideal cereal plant, or ideotype. This ideal plant would be a communal plant, one that minimized competition with its neighbors, featuring traits like a single, unbranched stem, erect leaves for better light interception within a dense canopy, and a large ear, all contributing to a high HI. This conceptual shift from selecting for performance to selecting for a specific plant architecture has profoundly guided breeding programs for decades, most notably contributing to the development of the high-yielding, semi-dwarf rice and wheat varieties of the Green Revolution.

Following the foundational work of Donald and his collaborator J. Hamblin, who further elaborated on the utility of biological yield and harvest index as agronomic and breeding criteria in 1976, the concept has undergone several decades of refinement (Donald & Hamblin, 1976). While the core definition—economic yield divided by total biological yield—remains, its application and measurement have been adapted for different crops and research questions. For

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

instance, in grain cereals, biological yield typically refers to the total above-ground biomass, as root biomass is difficult to measure at scale. In tuber crops like potato, HI is calculated as tuber fresh or dry weight divided by the total plant biomass, including haulms and roots. For oilseed crops, the definition may be refined to consider the energy content of the seeds, as oil is more energy-dense than starch or protein, leading to the concept of an energy harvest index (Sinclair, 1998).

Current understanding recognizes HI as an integrative and complex quantitative trait (Hay, 1995). It is not a simple trait but rather the outcome of numerous physiological processes, including photosynthesis (source strength), assimilate transport, and the capacity of reproductive organs to store photoassimilates (sink strength). Methodologically, HI is typically measured at final harvest by separating the plant into its harvestable (e.g., grain, tubers) and non-harvestable (e.g., straw, leaves, stems) components, drying them to a constant weight, and calculating the ratio. This is often done on a per-plant or per-unit-area basis. While conceptually simple, accurate measurement requires careful sampling to account for variability within a field. Despite some suggestions that major cereal crops are approaching a theoretical maximum for HI, the concept remains indispensable for interpreting crop responses to environmental stress, understanding the physiological basis of yield, and identifying new avenues for genetic improvement in the face of climate change and growing global food demand (Hay, 1995).

2.3.2 Physiological Basis of Harvest Index

The harvest index, as a measure of partitioning efficiency, is fundamentally governed by the intricate physiological processes that control the production, transport, and storage of photoassimilates within a plant. At the heart of these processes lies the concept of source-sink relationships, which describes the dynamic balance between tissues that produce sugars (sources)

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

and tissues that consume or store them (sinks) (Chang & Zhu, 2017). Understanding this relationship is paramount to comprehending the physiological underpinnings of harvest index. Source tissues are primarily mature, photosynthetically active leaves that are net exporters of carbon assimilates, principally in the form of sucrose. Sink tissues are net importers of these assimilates and include all non-photosynthetic or rapidly growing parts of the plant, such as roots, shoot meristems, developing leaves, and, most critically for yield, reproductive organs like flowers, fruits, seeds, and tubers. The harvest index is, in essence, a whole-plant quantification of the outcome of the competition for resources between the economic sink (the harvestable product) and all other vegetative sinks (Slafer et al., 2019; Yin et al., 2019).

The process of photosynthate partitioning is the differential distribution of these sugars from source to sink tissues. This is not a passive process but a highly regulated one, involving complex signaling networks that communicate sink demand and authorize transport. Following production in the mesophyll cells of leaves, sucrose is loaded into the phloem, the plant's vascular transport system for sugars. This loading can occur apoplastically, where sucrose is exported into the cell wall space before being actively taken up by companion cells of the phloem, or symplastically, through direct cytoplasmic connections. The movement of sugars through the phloem, known as translocation, is driven by a pressure gradient generated by the high concentration of sugars at the source and their removal at the sink. Upon arrival at a sink tissue, sucrose is unloaded from the phloem and can be either directly utilized for energy and growth or converted into storage compounds like starch or lipids. The efficiency of this entire transport pathway—from loading to unloading—is a key determinant of sink strength and, consequently, harvest index (Slafer et al., 2019).

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Carbon allocation patterns shift dramatically throughout the plant's life cycle, reflecting changing priorities (White et al., 2016). During early vegetative growth, the primary sinks are the developing roots and new leaves, which are essential for establishing the plant's resource-gathering capacity. As the plant transitions to the reproductive phase, a major shift occurs. The developing reproductive structures—panicles, spikes, flowers, and eventually grains or tubers—become the dominant sinks. The ability of these reproductive sinks to attract a large proportion of the available photoassimilates is a critical factor for achieving a high harvest index. The strength of a sink is determined by both its size (the number of cells capable of storage) and its activity (the rate of metabolic processes converting sugars into storage forms) (Chang & Zhu, 2017). Therefore, processes that determine the number of grains or tubers (sink size) and the rate of their filling (sink activity) are central to HI.

During the crucial grain-filling period (or tuber bulking), the demand from the developing sinks can often exceed the supply from concurrent photosynthesis. To meet this demand, plants have evolved mechanisms for the remobilization of assimilates (Slafer et al., 2019). Carbon and nitrogen stored in vegetative tissues, particularly the stem and leaves, are broken down and transported to the developing grains. This process can contribute significantly to final yield, in some cases accounting for over 50% of the final grain weight, especially under conditions of post-anthesis stress like drought or heat. The efficiency of this remobilization process—the ability to salvage resources from senescing vegetative parts and redirect them to the economic sink—is another vital physiological component of a high harvest index. Plants with improved stay-green characteristics may have a longer period of active photosynthesis, but this must be balanced with efficient remobilization from older leaves to maximize HI.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Finally, numerous environmental factors can profoundly affect these partitioning processes. Water availability, nutrient supply (especially nitrogen), temperature, and light intensity all influence both source capacity (photosynthesis) and sink activity. For example, drought stress during the grain-filling period can accelerate leaf senescence, reducing photosynthetic supply while simultaneously impairing the enzymatic processes required for starch synthesis in the grain, leading to poor filling and a lower harvest index (Slafer et al., 2019; Yin et al., 2019). Conversely, optimal conditions and high nitrogen availability can promote excessive vegetative growth, creating strong competition from vegetative sinks and potentially lowering the harvest index if not balanced by a corresponding increase in reproductive sink capacity. Therefore, a high and stable harvest index across different environments reflects a plant architecture and physiology that can robustly buffer these source-sink dynamics against external perturbations.

2.3.3 The Green Revolution and Harvest Index Improvement

The Green Revolution, a period of transformative agricultural advancement from the 1950s to the late 1980s, stands as the most significant testament to the practical impact of improving harvest index on a global scale (Evenson & Gollin, 2003). This period saw unprecedented increases in the yields of staple cereal crops, particularly wheat and rice, which averted widespread famine and fundamentally altered the trajectory of global food production. At the core of this revolution was a strategic shift in plant breeding, moving away from traditional tall, leafy plant types towards a new, more efficient architecture. This new architecture was deliberately designed to overcome the primary limitations of traditional varieties, which were prone to lodging—the bending or breaking of stems—when high levels of nitrogen fertilizer were applied (Hedden, 2003; Peng et al., 2021).

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The central figure in this transformation for wheat was Dr. Norman Borlaug, whose work earned him the Nobel Peace Prize in 1970. Working in Mexico, Borlaug and his team recognized that the yield potential of traditional wheat varieties was capped not by their inability to produce biomass, but by their structural instability. Applying fertilizer to boost growth resulted in tall, top-heavy plants that would collapse before harvest, leading to catastrophic yield losses. The breakthrough came from the introduction of semi-dwarf genes (Hedden, 2003). Borlaug's team utilized a Japanese dwarf wheat variety called Norin 10, which carried potent height-reducing genes. By crossing Norin 10 with high-yielding but tall local Mexican varieties, they developed new strains of wheat that were short, stout, and incredibly responsive to fertilizer. These semi-dwarf plants remained standing even with heavy grain heads, allowing farmers to exploit the yield-boosting potential of nitrogen fertilizers and irrigation for the first time.

A parallel revolution occurred in rice, spearheaded by the International Rice Research Institute (IRRI) in the Philippines (Khush, 1999). Following a similar breeding philosophy, IRRI scientists sought to create a new rice plant type that could overcome the lodging problem endemic to Asian rice paddies. In 1962, breeders crossed a tall, vigorous Indonesian variety named Peta with a semi-dwarf variety from Taiwan called Dee-geo-woo-gen (DGWG) (Khush, 1999; Peng et al., 2021). The DGWG parent carried a crucial recessive gene for short stature, later identified as *semidwarf1 (sd1)*. The result of this cross was the legendary variety IR8, released in 1966 and quickly dubbed miracle rice. IR8 was short and stiff-strawed, highly responsive to nitrogen, and insensitive to day length, allowing it to be grown across a wide range of latitudes and seasons. Its yield potential was staggering, producing up to 9.4 tons per hectare, nearly ten times the average yield in the Philippines at the time.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The physiological mechanism behind the success of these Green Revolution varieties was a dramatic improvement in the harvest index (Khush, 2001; Evenson & Gollin, 2003). The semi-dwarf genes, such as Rht (Reduced height) from Norin 10 wheat and sd1 in rice, did not increase the total biomass production of the plant. Instead, they altered the partitioning of photoassimilates (Peng et al., 2021). By reducing the allocation of resources to the stem (a vegetative sink), a much larger proportion of the plant's energy and carbon was diverted to the developing grain (the economic sink). This architectural change led to a quantifiable leap in efficiency. Before the Green Revolution, traditional tall varieties of wheat and rice typically had a harvest index of around 0.3, meaning only 30% of the plant's above-ground biomass was converted into grain. The new semi-dwarf varieties consistently achieved a harvest index of 0.5 or higher, effectively rebalancing the plant's priorities from vegetative growth to reproductive output. This increase from 0.3 to 0.5 represented a more than 60% improvement in partitioning efficiency and was the single most important factor contributing to the doubling and tripling of yields that characterized the era. The success of IR8 and Borlaug's wheat varieties demonstrated unequivocally that manipulating plant architecture to improve harvest index was a powerful and effective strategy for enhancing crop yield potential.

2.3.4 Genetic Architecture of Harvest Index

Harvest index, as the culmination of multiple developmental and physiological processes, is a classic example of a complex quantitative trait. Its genetic architecture is not governed by a single gene but is polygenic, controlled by the cumulative action of numerous genes, each with a relatively small effect, and is significantly modulated by environmental conditions (Shi et al., 2019). Understanding this complex genetic basis is fundamental to designing effective breeding strategies for its continued improvement. The principles of quantitative genetics provide the

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

framework for dissecting such traits, allowing researchers to partition the observed phenotypic variation into genetic and environmental components and to identify the specific genomic regions responsible for this variation.

A key parameter in quantitative genetics is heritability, which measures the proportion of total phenotypic variance in a population that is attributable to genetic factors (Shi et al., 2019). For a trait to be effectively selected in a breeding program, it must have a reasonably high heritability. Numerous studies across a wide range of crops have shown that harvest index generally exhibits moderate to high broad-sense heritability, indicating that a substantial portion of its variation is under genetic control. For example, high heritability has been consistently reported in wheat, suggesting that selection for HI can be effective. However, these estimates are population- and environment-specific. In some cases, such as studies in field pea, heritability estimates were found to be too low to make direct selection for HI feasible, highlighting the trait's plasticity. This variability underscores the importance of genotype-by-environment ($G \times E$) interactions, where the performance of different genotypes varies across different environments. A major goal in breeding is to develop varieties with a high and stable HI across diverse conditions, which requires identifying genes and alleles that confer this stability.

Modern molecular genetics has provided powerful tools to move beyond statistical estimates of heritability and identify the specific genes or genomic regions, known as Quantitative Trait Loci (QTLs), that underlie variation in HI (Shi et al., 2019). QTL mapping, typically conducted in biparental populations (like Recombinant Inbred Lines, RILs), has been instrumental in dissecting the genetic basis of HI and its component traits. For instance, studies in *Brassica napus* have identified as many as 160 QTLs associated with HI, seed yield, and biomass. In rice, specific QTLs for harvest index itself, such as qHI-1 on chromosome 1, have

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

been mapped. More commonly, QTLs are identified for component traits that collectively determine HI, such as plant height, grain number, and grain weight. The Green Revolution genes, *sd1* in rice and *Rht* in wheat, are effectively major-effect QTLs for plant height that have a profound, pleiotropic impact on harvest index. Similarly, a multitude of QTLs controlling grain size and number, such as *GW2*, *GS3*, and *GW5* in rice, and *TaGW2-6A* in wheat, directly influence the sink capacity and thus contribute to the overall harvest index.

While QTL mapping has been successful, it is limited by the genetic diversity present in the two parents of the mapping population. Genome-Wide Association Studies (GWAS) overcome this limitation by surveying a much broader range of genetic diversity in natural populations or diverse germplasm panels (Shi et al., 2019). GWAS has become a primary tool for identifying marker-trait associations for complex traits. For example, GWAS in wheat has identified numerous SNPs on chromosomes 1D, 1B, and 3B associated with harvest index. In rice, GWAS has pinpointed loci associated with thousand-grain weight and grain width. These studies often reveal that the genetic architecture of HI is characterized by many small-effect loci, confirming its polygenic nature. Furthermore, they can identify epistatic interactions, where the effect of one gene is modified by another. In *Brassica napus*, over 160 epistatic loci pairs were found to influence HI and related traits, demonstrating that the genetic control is not merely additive but involves complex, non-linear interactions between genes, forming intricate regulatory networks (Shi et al., 2019). The identification of these QTLs, GWAS hits, and their interactions provides the raw material for modern breeding, enabling the targeted improvement of harvest index through marker-assisted selection and genomic prediction.

2.3.5 Molecular Determinants of Harvest Index

The genetic architecture of harvest index is manifested through a diverse array of molecular pathways that control every aspect of plant growth, development, and metabolism. The genes within these pathways govern the key physiological processes—plant stature, the timing of developmental transitions, the formation of reproductive sinks, and the allocation of metabolic resources—that ultimately determine the partitioning of biomass. Deciphering the function of these molecular determinants provides the ultimate toolkit for rationally designing crops with improved yield efficiency.

Gibberellin Signaling and Dwarfing Genes.

Perhaps the most impactful molecular determinants of harvest index are the genes that control plant height through the gibberellin (GA) hormone pathway (Peng et al., 1999; Hedden, 2003). The semi-dwarf phenotypes of the Green Revolution were a direct result of mutations in these genes. GAs are plant hormones that promote cell elongation and, consequently, stem growth. Disrupting this pathway leads to shorter, more lodging-resistant plants that allocate a greater proportion of assimilates to grain.

In rice, the key gene is *semidwarf1* (*sd1*), which encodes the enzyme GA 20-oxidase-2 (*OsGA20ox-2*) (Sasaki et al., 2002; Chen et al., 2020). This enzyme catalyzes a critical step in the biosynthesis of active GAs. The famous *sd1* allele from the *Dee-geo-woo-gen* variety contains a loss-of-function mutation that impairs the enzyme's activity, leading to reduced levels of bioactive GA and the characteristic semi-dwarf stature. This GA-deficient mechanism effectively increases HI without significant negative pleiotropic effects on other aspects of yield.

In wheat, the primary Green Revolution genes, *Rht-B1b* and *Rht-D1b*, operate through a different mechanism: GA-insensitivity (Peng et al., 1999; Hedden, 2003). These genes encode

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

DELLA proteins, which are transcriptional regulators that act as repressors of GA-responsive growth. In a normal plant, GA binds to its receptor, which then targets the DELLA protein for degradation, thereby de-repressing growth. The Rht-B1b and Rht-D1b alleles are gain-of-function mutations that produce truncated DELLA proteins resistant to GA-mediated degradation. These mutant proteins constitutively repress growth, resulting in a dwarf phenotype that is insensitive to the application of exogenous GA. While highly effective at increasing HI, these alleles can have some undesirable effects, such as reduced coleoptile length, which can impair seedling emergence. This has led to the exploration of other GA-responsive dwarfing genes in wheat, such as Rht8, Rht12, and Rht13, which reduce height without the negative seedling traits.

This fundamental pathway is conserved across crops. In soybean, the dwarf phenotype of the GmDW1 mutant is caused by a defect in ent-kaurene synthase, an early enzyme in the GA biosynthesis pathway (Li et al., 2018). In potato, the balance between GA biosynthesis and catabolism is critical for tuberization, which is itself a manifestation of harvest index. Genes like StGA20ox1 (GA biosynthesis) and StGA2ox1 (GA catabolism) are key regulators; suppressing the former and overexpressing the latter promotes the transition from stolon growth to tuber formation by reducing local GA levels, thereby increasing tuber yield (Kloosterman et al., 2007).

Grain and Seed Size Genes.

The capacity of the reproductive sink is a direct function of the number and size of the grains or seeds produced. A vast and complex network of genes regulates these two components, which often exhibit a compensatory trade-off. Many of these genes have been cloned and characterized, providing powerful targets for fine-tuning yield potential.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

A prominent example is GW2 (GRAIN WIDTH 2), which has been identified in both rice (OsGW2) and wheat (TaGW2) (Song et al., 2007; Simmonds et al., 2016). This gene encodes a RING-type E3 ubiquitin ligase, a component of the ubiquitin-proteasome system that targets other proteins for degradation. GW2 acts as a negative regulator of grain size by controlling cell division in the spikelet hull. Loss-of-function mutations in GW2 lead to a wider, heavier grain because the negative regulation of cell proliferation is removed.

Another major regulator is GS3 (GRAIN SIZE 3) in rice, which is a primary determinant of grain length (Fan et al., 2006; Takano-Kai et al., 2009). GS3 acts as a negative regulator, and a common allele that produces long grains contains a premature stop codon, resulting in a truncated, non-functional protein. This allows for increased cell elongation in the longitudinal dimension of the grain. In contrast, GS5 acts as a positive regulator of grain size, primarily affecting grain width. Higher expression of GS5, which encodes a putative serine carboxypeptidase, leads to larger grains.

Other key players include GW5 (also known as qSW5), which controls grain width through the ubiquitin-proteasome pathway, and TGW6 (THOUSAND-GRAIN WEIGHT 6), which encodes an IAA-glucose hydrolase and influences grain filling (Ishimaru et al., 2013). In soybean, analogous genes have been identified. GmSW17, an ortholog of a ubiquitin-specific protease, acts as a positive regulator of seed width and weight and was a key target of selection during domestication (Lyu et al., 2024). The Dt2 gene, a MADS-domain transcription factor controlling stem growth termination, also has a pleiotropic effect on seed size, with knockout mutants showing increased seed weight and yield. These genes often interact in complex regulatory networks, and their combined allelic composition determines the final sink size and, by extension, the harvest index.

Starch Biosynthesis and Sugar Transport Genes.

The activity component of sink strength is largely determined by the rate and duration of starch accumulation in the storage organs. Starch typically constitutes the vast majority of the dry weight of cereal grains and potato tubers. The biosynthesis of starch is a well-defined metabolic pathway involving a cascade of key enzymes whose coordinated expression is critical for grain and tuber filling.

The rate-limiting step in this pathway is catalyzed by ADP-glucose pyrophosphorylase (AGPase) (Preiss, 1991; Hannah, 2007). This enzyme produces ADP-glucose, the activated glucose donor for starch synthesis. AGPase is a heterotetramer, and its activity is tightly regulated both allosterically and transcriptionally. Higher AGPase activity is strongly correlated with increased starch accumulation and grain yield. In potato, the small subunit of this enzyme, StAGPaseS, is highly expressed in growing tubers and is a key control point for tuber bulking.

Once ADP-glucose is produced, starch synthases elongate the glucan chains. There are two main types. Granule-Bound Starch Synthase I (GBSSI), encoded by the Waxy (Wx) locus in cereals and StGBSSI in potato, is solely responsible for the synthesis of amylose, the linear component of starch (Jeon et al., 2010). The level of GBSSI activity determines the amylose content, a critical grain quality trait. Soluble Starch Synthases (SSS), along with branching and debranching enzymes, are responsible for synthesizing amylopectin, the branched component that makes up the bulk of the starch granule (Hannah, 2007). The expression of genes encoding AGPase, GBSSI, and the various SSS isoforms is tightly coordinated and peaks during the mid-grain filling stage. The efficiency of this entire enzymatic machinery in converting sucrose delivered to the sink into insoluble starch granules is a direct molecular determinant of final grain/tuber weight and harvest index.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The connection between source leaves and sink organs is the phloem, and the efficiency of sucrose transport through this vascular highway is critical for fueling sink growth. Sucrose Transporters (SUTs) are sucrose-proton symporters that play a crucial role in actively loading sucrose into the phloem in source leaves and, in some cases, unloading it in sink tissues (Braun & Slewinski, 2009). In rice, OsSUT1 is essential for phloem loading and the transport of carbon to developing grains. A second major family of transporters, the SWEET (Sugars Will Eventually be Exported Transporters) proteins, act as uniporters that facilitate the diffusion of sugars across membranes along a concentration gradient (Chen et al., 2012). In soybean, for example, the paralogous genes GmSWEET15a and GmSWEET15b are highly expressed in the seed coat and are essential for delivering sucrose to the developing embryo (Wang et al., 2019). Double knockout of these genes results in aborted seeds, demonstrating their central role in seed filling.

2.3.6 Yield Components and Trade-offs

Grain yield, the primary component of harvest index in cereals, is not a monolithic trait. It is the product of several interacting components: the number of productive tillers (or panicles) per plant, the number of grains per panicle, and the average weight per grain. Similarly, in tuber or seed crops, yield is a function of the number and average size of the tubers or seeds. A fundamental concept in crop physiology is that these components often exhibit trade-off relationships or compensation mechanisms (Sadras, 2007; Slafer et al., 2014). An increase in one component is frequently met with a decrease in another, presenting a significant challenge for breeders aiming to increase overall yield.

The most widely studied and critical trade-off is that between grain number and grain weight (Sadras, 2007; Quintero et al., 2018). Historically, the majority of yield gains in wheat

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

and other cereals have been achieved by increasing the number of grains per unit area. However, this increase in sink size is often compensated for by a reduction in the average size and weight of each individual grain. This negative correlation is rooted in both source and sink limitations. From a source perspective, a larger number of grains creates a greater overall demand for photoassimilates during the grain-filling period. If the photosynthetic source cannot meet this increased demand, the result is competition among the developing grains, leading to incomplete filling and smaller final grain size (Slafer et al., 2014).

From a sink perspective, the trade-off is also established early in development. The potential size of a grain is determined by the growth of maternal tissues like the ovary and pericarp around the time of flowering (anthesis). The critical period for determining grain number overlaps with this period. Therefore, conditions or genetic factors that promote the survival of more florets to establish a higher grain number may simultaneously limit the resources available for the development of the structures that define potential grain size (Quintero et al., 2018). This results in an inherent, genetically and developmentally programmed inverse relationship. For example, many QTLs identified for grain number have an opposite, negative effect on grain weight when mapped in the same population.

Breeders have long sought to uncouple this negative relationship to achieve simultaneous increases in both grain number and weight. This requires a sophisticated understanding of the sink-source balance throughout the crop's life cycle. Strategies to optimize this balance include enhancing photosynthetic capacity and duration (improving the source) or manipulating the timing of developmental events to reduce direct competition between the establishment of sink size and sink number. For instance, research on the wheat gene TaExpA6, an α -expansin, has shown that its targeted expression in the grain after the critical period for grain number

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

determination can increase final grain weight and yield without a compensatory reduction in grain number (Calderini et al., 2021). Similarly, mutations in the TaGW2 gene, a negative regulator of grain size, increase grain weight but often trigger a compensatory decrease in grain number. Understanding the precise molecular and developmental basis of these trade-offs is crucial for developing optimization strategies that push yield potential beyond its current boundaries. Ultimately, maximizing harvest index involves not just increasing the proportion of biomass partitioned to the harvestable sink, but also optimizing the architecture of that sink by balancing the number and size of its individual components.

2.3.7 Modern Breeding Approaches for Harvest Index Improvement

The pursuit of a higher and more stable harvest index has been revolutionized by the integration of genomic tools into plant breeding. These modern approaches allow for a level of precision and speed that was unattainable with conventional phenotypic selection alone. By directly targeting the genetic determinants of HI and its component traits, breeders can accelerate genetic gain and develop elite varieties more efficiently.

Marker-Assisted Selection (MAS) was one of the first genomic tools to be widely adopted (Xu & Crouch, 2008). MAS is based on the principle of selecting for a desired trait not by observing the phenotype itself, but by selecting for a molecular marker (such as an SNP or SSR) that is tightly linked to the gene or QTL controlling the trait. Once a strong association between a marker and a trait like plant height, grain size, or disease resistance is established, breeders can use this marker to screen large populations of seedlings and identify individuals carrying the favorable allele, long before the trait is phenotypically expressed. This is particularly powerful for traits that are difficult or expensive to measure, have low heritability, or are expressed late in development. For example, MAS has been successfully used to pyramid

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

multiple disease resistance genes and to select for major-effect QTLs like the Rht dwarfing genes in wheat or the Sub1 gene for submergence tolerance in rice, both of which have indirect but significant impacts on yield stability and, therefore, harvest index.

While MAS is highly effective for traits controlled by a few major genes, its utility is limited for complex quantitative traits like harvest index, which are governed by many genes of small effect. To address this, Genomic Selection (GS), or genomic prediction, was developed (Meuwissen et al., 2001; Heffner et al., 2009). GS uses genome-wide marker data to predict the overall genetic merit, or Genomic Estimated Breeding Value (GEBV), of an individual. In a GS scheme, a training population is created that has been both genotyped with thousands of markers and phenotyped for the traits of interest. A statistical model is then built to associate the genome-wide marker patterns with performance. This model can then be used to predict the breeding value of candidate individuals that have only been genotyped, without the need for extensive phenotyping. This dramatically shortens the breeding cycle, as selection can be performed on a large number of individuals at a very early stage. GS is ideally suited for improving complex traits like yield and harvest index because it captures the effects of all genes, including those with minor effects that are missed by MAS. The accuracy of GS predictions depends on several factors, including the size of the training population, the heritability of the trait, and the genetic relationship between the training and candidate populations (Heffner et al., 2009).

The most transformative modern tool is CRISPR/Cas9 and related gene-editing technologies (Doudna & Charpentier, 2014; Zhang et al., 2019). Unlike MAS and GS, which work with existing genetic variation, gene editing allows for the direct and precise modification of a plant's own DNA. Using a guide RNA to target a specific location in the genome, the Cas9 enzyme can create a cut in the DNA, which can then be repaired to introduce a desired change.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

This could involve knocking out a gene, altering its sequence, or changing its regulatory region. This technology has opened up unprecedented possibilities for crop improvement. For example, breeders can precisely recreate favorable alleles that were lost during domestication, fine-tune the expression of key genes, or simultaneously edit multiple genes in a pathway. In the context of harvest index, CRISPR has been used to edit genes controlling plant architecture (e.g., creating novel dwarfing alleles), grain number (e.g., by editing regulators like GW2), and flowering time (Rodríguez-Leal et al., 2017). For instance, by editing the promoter regions of flowering time genes in tomato, researchers have been able to generate a range of alleles that fine-tune plant architecture and fruit production. This ability to create targeted genetic diversity on demand, without the linkage drag associated with traditional crossing, represents a paradigm shift in breeding for complex traits like harvest index.

The most powerful strategies involve the integration of these approaches. For example, GWAS can identify candidate genes, which can then be validated and manipulated using CRISPR. The novel alleles created by gene editing can then be efficiently tracked and introgressed into elite breeding lines using MAS. Furthermore, GS models can be improved by incorporating information about major QTLs or known causal variants, a strategy known as GS+MAS. By combining the discovery power of GWAS, the predictive power of GS, and the precision of gene editing, modern breeders have a comprehensive toolkit to systematically deconstruct and reassemble the genetic architecture of harvest index for continued gains in crop productivity.

2.4 Genomics of BLSS Staple Crops

This section provides a comprehensive review of the genomic landscape of the four staple crops central to this dissertation: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

(*Solanum tuberosum*), and soybean (*Glycine max*). For each species, we examine the history and characteristics of its reference genome, the architecture of its pangenome and the extent of structural variation, the molecular genetics of key genes governing harvest index and its components, major findings from genome-wide association studies and quantitative trait locus mapping, and the application of this genomic knowledge in breeding. This foundational analysis establishes the genomic context for the subsequent investigation into the pangenomic architecture of harvest index determinants.

2.4.1 Rice (*Oryza sativa*)

Genome Characteristics and Reference Genome History.

Rice is a primary food source for over half of the world's population and serves as a model organism for monocot genomics. Its relatively small genome, diploid nature, and extensive genetic resources have made it a forerunner in plant genomics, providing critical insights into the genetic control of yield, quality, and adaptation. The genomic journey of rice began with the landmark effort of the International Rice Genome Sequencing Project (IRGSP), a consortium of researchers from ten countries. This project adopted a map-based, clone-by-clone shotgun approach to sequence the genome of the japonica cultivar Nipponbare (International Rice Genome Sequencing Project, 2005). The initial draft, completed in 2005, was a monumental achievement, providing the first high-quality reference sequence for a crop plant.

Following this initial release, two independent annotation efforts by the Rice Annotation Project (RAP) and the Rice Genome Annotation Project (RGAP) produced slightly different assemblies due to variations in clone selection and gap-filling strategies. To resolve these discrepancies and create a unified resource, a collaborative effort produced the Os-Nipponbare-Reference-IRGSP-1.0 (IRGSP-1.0) assembly in 2013 (Kawahara et al., 2013). This update

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

incorporated next-generation sequencing data and optical mapping to correct errors, close gaps, and re-orient scaffolds, resulting in a gold standard reference. The IRGSP-1.0 assembly spans approximately 373.8 megabases (Mb) across 12 chromosomes. Depending on the annotation source, it is estimated to contain between 35,806 and 43,795 protein-coding genes. The high quality and comprehensive annotation of the IRGSP-1.0 genome have since anchored countless studies in functional genomics, comparative genomics, and molecular breeding. More recently, the advent of long-read sequencing technologies has enabled the construction of complete telomere-to-telomere (T2T) assemblies of the Nipponbare genome, such as T2T-NIP, which have resolved the remaining complex repetitive regions and gaps, further refining our understanding of rice genome architecture (Chen et al., 2023).

Pangenome Architecture and Structural Variation.

The single Nipponbare reference genome, while foundational, represents only a fraction of the genetic diversity within the vast *Oryza* genus. The development of pangenomics has been instrumental in capturing this broader variation. The 3,000 Rice Genomes Project (3K RGP) was a transformative initiative that sequenced 3,010 diverse rice accessions, revealing an enormous catalog of genetic variants (The 3K RGP, 2014; Wang et al., 2018). Analysis of this dataset uncovered approximately 29 million single nucleotide polymorphisms (SNPs), 2.4 million small indels, and 93,683 larger structural variants (SVs), with an average of 12,178 SVs distinguishing any given genome from the reference.

The construction of a rice pangenome from the 3K RGP data revealed that a significant portion of the species' genome is dispensable. The initial pangenome added 268 Mb of novel sequence not present in the Nipponbare reference, containing an estimated 12,465 new genes (Wang et al., 2018). The Rice Pan-genome Browser (RPAN), developed to navigate this

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

complexity, catalogs a pangenome of 50,995 total gene families from 453 high-quality assemblies, comprising 23,914 core genes (present in all accessions) and 22,095 dispensable genes (Sun et al., 2017). This indicates that nearly 43% of the gene content is variable among accessions. The dispensable genome is heavily enriched with transposable elements (TEs), which constitute up to 85% of the sequence in presence/absence variations. A more recent pangenome study incorporating 145 wild and cultivated rice accessions expanded the known gene repertoire to 69,531 gene families, with 13,728 genes being specific to wild rice relatives, highlighting the vast genetic reservoir lost during domestication (Guo et al., 2025). These pangenome resources have demonstrated that SVs, particularly PAVs and copy number variations (CNVs), are major drivers of phenotypic diversity, affecting traits from disease resistance to yield.

Key Genes Regulating Harvest Index and Yield Components.

The genetic dissection of harvest index and its components in rice has identified a suite of key genes, many of which were central to the successes of the Green Revolution and continue to be targets for modern breeding. These genes regulate plant architecture, grain size, and nutrient allocation. A cornerstone of the Green Revolution was the introgression of the *sd1* (semi-dwarf 1) gene from the Taiwanese landrace Dee-geo-woo-gen. Located on chromosome 1, *sd1* encodes a gibberellin 20-oxidase (*OsGA20ox-2*), an enzyme in the gibberellin biosynthesis pathway (Sasaki et al., 2002). Loss-of-function mutations in *sd1* lead to reduced levels of bioactive GAs, resulting in a semi-dwarf phenotype. This shorter stature confers lodging resistance, allowing for the application of higher levels of nitrogen fertilizer, and improves the harvest index by reallocating photosynthates from the stem to the developing panicles.

Grain size and weight are critical components of yield and are controlled by a network of regulatory genes. *GW2* (Grain Width 2), located on chromosome 2, is a negative regulator of

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

grain size. It encodes a RING-type E3 ubiquitin ligase that targets proteins for degradation via the ubiquitin-proteasome pathway. Loss-of-function mutations in GW2 lead to increased cell numbers in the spikelet hull, resulting in wider, heavier grains and an increase in thousand-grain weight by up to 18% (Song et al., 2007). Similarly, GS3 (Grain Size 3) on chromosome 3 is a major negative regulator of grain length, while GW5, also known as qSW5, on chromosome 5 controls grain width. A 1,212-bp deletion upstream of the GW5 coding region is associated with wider grains (Weng et al., 2008). The efficiency of filling these grains is dependent on carbon allocation. OsSUT1, a sucrose-proton symporter on chromosome 3, is essential for transporting sucrose from the phloem into sink tissues like the developing endosperm. The final quality and density of the grain are influenced by starch composition, which is primarily controlled by the Waxy (Wx) gene on chromosome 6. This gene encodes granule-bound starch synthase I (GBSSI), the enzyme responsible for amylose synthesis.

GWAS and QTL Mapping of Yield-Related Traits.

The vast genomic resources available for rice, including the 3K RGP dataset, have powered numerous large-scale GWAS and QTL mapping studies, leading to the identification of thousands of marker-trait associations (MTAs) for yield and its components. The GWAS Atlas database alone catalogs over 163,000 associations for 461 traits in rice, demonstrating the genetic complexity of yield. GWAS have successfully re-identified and fine-mapped many of the major genes discussed previously. For instance, strong association signals for plant height are consistently detected at the *sd1* locus on chromosome 1, while signals for grain width and length map directly to the GW5 and GS3 loci on chromosomes 5 and 3, respectively (Huang et al., 2010).

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Beyond confirming known genes, these studies have uncovered novel loci. A study of 100 diverse rice genotypes identified a pleiotropic locus on chromosome 1 affecting both panicle length and thousand-grain weight, and another on chromosome 4 associated with both seed setting rate and overall yield per hectare (Rehman et al., 2023). Under abiotic stress conditions, such as drought, GWAS has been particularly powerful, identifying 219 MTAs across multiple chromosomes that collectively explain a significant portion of the phenotypic variance in grain yield. QTL mapping in biparental populations has also been highly productive. The Gramene QTL Database archives numerous loci for harvest index and related traits, including qHI-1 on chromosome 1. The integration of QTL mapping and GWAS has created a comprehensive map of the genetic architecture of rice yield, providing a rich toolkit of markers and candidate genes for breeding.

Breeding Applications and Future Prospects.

The application of genomic knowledge has revolutionized rice breeding. The Green Revolution was, in essence, a large-scale molecular breeding program centered on the deployment of the *sd1* gene. The development of molecular markers for *sd1* and other major genes like *GS3*, *GW2*, and *Wx* has enabled marker-assisted selection (MAS), allowing breeders to precisely pyramid favorable alleles for high yield and desired quality traits. For example, combining the *sd1* allele for ideal plant height with loss-of-function alleles of *GW2* and *GS3* can produce cultivars with a high harvest index and large, heavy grains (Miyata et al., 2022).

The future of rice breeding lies in leveraging the full spectrum of genetic diversity revealed by pangenomics. The single reference genome paradigm is limited because it overlooks the vast number of dispensable genes and structural variants present in the broader gene pool. Pangenome-wide association studies (PWAS) are emerging as a powerful tool to link these

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

dispensable genes to important agronomic traits. The development of tools like the Rice Pangenome Genotyping Array (RPGA), which includes probes for 80,000 SNPs and PAVs, helps overcome the reference bias inherent in traditional genotyping methods (Wang et al., 2022). Furthermore, gene editing technologies like CRISPR/Cas9 offer unprecedented precision to modify target genes. Breeders can now introduce favorable alleles, knock out negative regulators, or fine-tune the expression of key genes to optimize plant architecture and yield components without the linkage drag associated with conventional introgression. The combination of pangenomic discovery, PWAS, and precision gene editing promises to accelerate the development of next-generation rice varieties with higher yield potential, improved resilience to climate change, and enhanced nutritional value.

2.4.2 Wheat (*Triticum aestivum*)

Genome Characteristics and Reference Genome History.

Bread wheat is a global staple, providing approximately 20% of the calories and protein consumed by humankind. Its genomic complexity, arising from its large, repetitive, and hexaploid nature, has historically posed significant challenges to researchers. However, recent breakthroughs in sequencing and bioinformatics have unlocked the wheat genome, paving the way for accelerated genetic improvement. The genome of bread wheat (*Triticum aestivum*) is a testament to evolutionary complexity. It is a recent allohexaploid, formed through two distinct hybridization events involving three diploid ancestral species. This resulted in a massive genome of approximately 14.6 gigabases (Gb), organized into 21 chromosome pairs and comprising three distinct subgenomes: A, B, and D (AABBDD). The sheer size of the genome, combined with a staggering repetitive element content of over 85%, made sequencing wheat a formidable task (IWGSC, 2018).

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The International Wheat Genome Sequencing Consortium (IWGSC) was formed in 2005 to tackle this challenge. In 2018, the IWGSC released the first high-quality, annotated reference sequence of the cultivar Chinese Spring, known as IWGSC RefSeq v1.0 (IWGSC, 2018). This landmark achievement provided a nearly complete assembly, anchoring 94% of the genome to chromosomes and identifying 107,891 high-confidence protein-coding genes. Subsequent releases, such as IWGSC RefSeq v2.1, have further improved the assembly by incorporating long-read sequencing and optical mapping data to correct errors and fill remaining gaps. The availability of this reference sequence has been a watershed moment for wheat research, enabling detailed studies of gene function, evolution, and trait genetics.

Pangenome Architecture and Structural Variation.

The Chinese Spring reference genome, while invaluable, represents a single genotype within a globally diverse species. To capture the full breadth of wheat's genetic variation, the focus has shifted to pangenome construction. The 10+ Wheat Genomes Project was a major initiative that generated high-quality assemblies for a panel of elite wheat cultivars from around the world (Walkowiak et al., 2020). This project revealed substantial structural variation and gene content differences among cultivars. The analysis of 15 accessions showed that only 73.5% of orthologous gene groups were complete across all lines, indicating significant presence/absence variation (PAV).

These pangenome studies have highlighted that the wheat gene pool is highly dynamic, shaped by widespread PAVs, copy number variations (CNVs), and large-scale introgressions from wild relatives, which are a key source of disease resistance and stress tolerance traits in modern breeding programs. The polyploid nature of wheat facilitates the buffering of such large-scale variations, allowing the genome to tolerate gene loss or duplication that might be

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

deleterious in a diploid. However, this complexity also presents challenges for pangenome assembly and annotation, particularly in distinguishing functional genes from pseudogenes or transposable elements and in accurately identifying homoeologous gene copies across the A, B, and D subgenomes. Despite these hurdles, the developing wheat pangenome is a critical resource for identifying novel alleles and haplotypes associated with agronomic performance.

Key Genes Regulating Harvest Index and Yield Components.

Similar to rice, the genetic improvement of harvest index in wheat was a cornerstone of the Green Revolution, driven primarily by the widespread adoption of semi-dwarfing genes. The most influential dwarfing genes in wheat are Rht-B1b and Rht-D1b (previously known as Rht1 and Rht2), located on the short arms of chromosomes 4B and 4D, respectively. These are homoeologous genes encoding DELLA proteins, which are negative regulators of gibberellin (GA) signaling (Peng et al., 1999). The Rht-B1b and Rht-D1b alleles contain mutations that render the DELLA proteins insensitive to GA, leading to a constitutive repression of growth and a semi-dwarf phenotype. This GA-insensitive dwarfing reduces plant height by 15-20% and is directly responsible for increased grain number and a higher harvest index. These Green Revolution genes are present in the majority of modern wheat cultivars worldwide.

However, they are not without drawbacks, as they can reduce coleoptile length, potentially impairing seedling emergence in certain environments. This has led to a search for alternative dwarfing genes that are GA-responsive. Rht8, on chromosome 2D, is one such gene, conferring a moderate height reduction without the negative pleiotropic effects on seedling vigor (Worland et al., 1998). Beyond plant architecture, genes controlling grain size are also critical for yield. TaGW2-6A, the wheat ortholog of rice OsGW2 on chromosome 6A, is a key negative regulator of grain size (Simmonds et al., 2016). Mutations that reduce its expression lead to

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

wider, heavier grains. Finally, developmental timing, which dictates the grain filling period, is controlled by major genes like *Vrn1* (vernalization) and *Ppd-D1* (photoperiod response). Alleles that confer photoperiod insensitivity (*Ppd-D1a*) allow for earlier flowering, which can be crucial for escaping terminal drought and maximizing harvest index.

GWAS, QTL Mapping, and Breeding Applications.

The availability of the reference genome and high-density SNP arrays has enabled large-scale GWAS and QTL mapping in wheat, despite its genomic complexity. Unsurprisingly, the most significant and consistently detected GWAS signals for plant height and harvest index map directly to the locations of the major *Rht* genes. Beyond these major genes, GWAS has uncovered a multitude of smaller-effect loci distributed across the genome. For instance, studies in spring wheat panels have identified significant associations for harvest index on chromosomes 1D, 1B, and 3B (Sukumaran et al., 2015). QTL mapping has been particularly effective for dissecting grain size and weight, with multiple studies identifying a major and stable QTL for thousand-grain weight (TGW) on chromosome 6A, co-localizing with the *TaGW2-6A* gene. Meta-QTL analysis has confirmed major, stable meta-QTLs for TGW on chromosomes 2A, 5A, 6A, and 7A.

Wheat breeding has a long and successful history, with the introgression of *Rht-B1b* and *Rht-D1b* standing out as its most impactful achievement. Modern breeding continues to build on this foundation through the use of molecular tools. Marker-assisted selection (MAS) is routinely used to select for favorable alleles of *Rht* genes, as well as genes for disease resistance and quality. The future of wheat improvement lies in harnessing the vast, untapped genetic diversity within the wheat pangenome. Genomic selection (GS) is poised to become a standard tool in wheat breeding, more effectively capturing the cumulative effect of numerous small-effect

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

QTLs. Finally, the application of gene editing technologies will allow for the precise modification of target genes, silencing negative regulators like TaGW2 or pyramiding multiple favorable traits with unprecedented speed and precision.

2.4.3 Potato (*Solanum tuberosum*)

Genome Characteristics and Reference Genome History.

The cultivated potato is the world's most important non-cereal food crop, valued for its high yield potential and nutritional content. Its complex genetics, characterized by autotetraploidy and high heterozygosity, have made it a challenging subject for genomic research and breeding. However, recent advances have provided powerful new tools to dissect and improve this vital crop. The cultivated potato is typically an autotetraploid ($2n = 4x = 48$), meaning it has four homologous copies of each of its 12 chromosomes. This genetic structure, combined with a high degree of heterozygosity, posed a significant barrier to genome sequencing. To circumvent this complexity, the Potato Genome Sequencing Consortium (PGSC) chose to sequence a doubled monoploid clone, DM1-3 516 R44, derived from the diploid species *Solanum tuberosum* Group Phureja (PGSC, 2011).

The initial draft genome published in 2011 was a landmark, but subsequent improvements have dramatically increased its quality. The current reference assembly, DM v6.1, was generated using long-read sequencing technology and released in 2020 (Pham et al., 2020). This high-quality assembly spans 741.6 Mb, with 731.2 Mb anchored to the 12 chromosomes, representing approximately 88% of the estimated 844 Mb genome. The DM v6.1 assembly boasts a 595-fold increase in contig N50 compared to previous versions, achieving a reference genome quality classification. The annotation of this assembly identifies approximately 32,917 protein-coding genes. The DM reference genome has become an indispensable resource for

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

potato research, enabling gene discovery, comparative genomics, and the development of molecular tools for breeding.

Pangenome Architecture and Structural Variation.

While the DM reference provides a crucial scaffold, the true genetic diversity of potato lies in its highly heterozygous and diverse gene pool, which includes thousands of cultivated varieties and wild relatives. A recent groundbreaking study constructed a graph-based pangenome for potato, integrating high-quality assemblies from 31 accessions, representing 60 distinct haplotypes from wild species, landraces, and modern cultivars (Sun et al., 2024). The results revealed a pangenome of staggering scale and complexity. The total pangenome size was 3,076 Mb, with only 365 Mb (12%) of sequence being conserved across all haplotypes.

The remaining 2,711 Mb constituted variable sequence, highlighting an enormous degree of structural variation. Wild species contributed a larger portion of the pangenome sequence (2,286 Mb) compared to cultivated lines (1,807 Mb), underscoring the vast genetic reservoir present in potato's wild relatives. The study cataloged a massive number of variants, including 46 million SNPs, 3.6 million indels, and 133,264 large structural variants (SVs). A remarkable 87% of these SVs were multi-allelic, reflecting the complex patterns of variation in this autotetraploid species. Crucially, these SVs were shown to be functionally significant, with 9,321 SVs associated with variations in the levels of 1,258 different metabolites, and 19,625 SVs predicted to be deleterious (Sun et al., 2024). This potato pangenome demonstrates that SVs are a major source of functional diversity and a critical target for understanding and improving agronomic traits.

Key Genes Regulating Harvest Index and Yield Components.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

In potato, the harvest index is determined by the partitioning of biomass to underground tubers. The initiation and development of these tubers, a process known as tuberization, is the central determinant of yield. The primary signal for tuberization is the mobile protein StSP6A (SELF-PRUNING 6A), an ortholog of the flowering hormone florigen (FLOWERING LOCUS T). Located on chromosome 5, StSP6A is produced in the leaves under inductive short-day conditions and travels through the phloem to the underground stolons, where it triggers the transition from stolon growth to tuber formation (Navarro et al., 2011). Overexpression of StSP6A promotes early and prolific tuberization, establishing it as the key tuberigen signal and a master regulator of sink strength.

The expression of StSP6A is tightly regulated by a network of transcription factors. A key activator is StBEL5, a BEL1-like transcription factor whose RNA is also mobile. StBEL5 moves from leaves to stolons, where it activates the expression of tuber-promoting genes (Chen et al., 2003). Another critical regulator is StCDF1 (CYCLING DOF FACTOR 1), a transcription factor on chromosome 5 that promotes tuberization (Kloosterman et al., 2013). The gibberellin pathway acts antagonistically to tuberization; high GA levels promote stolon elongation and inhibit tuber formation. Under short-day conditions, the expression of StGA2ox1 (GA 2-oxidase 1) on chromosome 3 is induced, which inactivates bioactive GAs, reducing GA levels in the stolon and thus promoting tuber initiation (Kloosterman et al., 2007). Once tubers are initiated, their growth depends on the accumulation of starch. Key enzymes in this process include StGBSSI (granule-bound starch synthase I) on chromosome 8, which synthesizes amylose, and StAGPaseS (ADP-glucose pyrophosphorylase small subunit), a rate-limiting enzyme for overall starch synthesis.

GWAS, QTL Mapping, and Breeding Applications.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Dissecting the genetic basis of yield in autotetraploid potato is challenging due to its complex inheritance patterns. Despite this, GWAS and QTL mapping have begun to unravel the genetic architecture of tuber yield and quality traits. GWAS in tetraploid potato panels have identified numerous loci associated with tuber yield, with one study identifying 46 significant MTAs for tuber yield distributed across all 12 chromosomes, indicating a highly polygenic basis for this trait. Many of the major QTLs identified for tuberization timing and maturity correspond to the key regulatory genes, particularly StCDF1 on chromosome 5 and StSP6A. The recent potato pangenome study has added a new dimension to this research by enabling SV-GWAS, allowing the association of the 133,264 identified SVs with phenotypic traits.

Potato breeding has traditionally been a slow process, hampered by tetraploid genetics, high heterozygosity, and the vegetative propagation of tubers, which can lead to pathogen accumulation. However, the field is on the cusp of a revolution driven by genomic innovations. One of the most transformative strategies is the shift to diploid hybrid breeding, which involves developing homozygous diploid inbred lines that can be crossed to create F1 hybrids with elite performance (Lindhout et al., 2011). These hybrids are propagated via true potato seeds (TPS) rather than bulky tubers, which simplifies logistics, eliminates tuber-borne diseases, and allows for rapid stacking of desirable traits. Genomic selection (GS) is another powerful tool particularly well-suited to the complexities of tetraploid potato. Finally, gene editing with CRISPR/Cas9 is already being used to improve key traits, for example, knocking out the GBSS gene creates amylose-free waxy potato starch with unique industrial properties.

2.4.4 Soybean (*Glycine max*)

Genome Characteristics and Reference Genome History.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Soybean is the world's foremost source of vegetable protein and oil, making it a cornerstone of global food and feed systems. Its genome reflects a complex history of ancient polyploidy, which has shaped its genetic architecture and provided a rich substrate for domestication and improvement. The soybean genome is a classic example of a paleopolyploid. Its evolutionary history is marked by at least two distinct whole-genome duplication (WGD) events, a more ancient event shared with many legumes and a more recent event specific to the Glycine lineage that occurred approximately 13 million years ago (Schmutz et al., 2010). This recent WGD resulted in a highly duplicated genome, where nearly 75% of genes exist in multiple copies (homeologs). Following duplication, the genome underwent extensive diploidization, ultimately settling into its modern state with 20 pairs of chromosomes ($2n = 40$).

The first complete reference genome for soybean, based on the North American cultivar Williams 82, was published in 2010. This was a landmark achievement, providing a high-quality sequence of approximately 978 Mb (Schmutz et al., 2010). The assembly and its subsequent improved versions have served as the foundational resource for soybean genetics and breeding for over a decade. The current reference annotation contains approximately 55,897 protein-coding genes, a high number that reflects the genome's duplicated history. The availability of the Williams 82 reference genome has catalyzed an explosion of research, enabling the identification of genes for key agronomic traits, the development of high-throughput genotyping platforms, and the application of modern breeding technologies.

Pangenome Architecture and Structural Variation.

While the Williams 82 reference has been invaluable, it represents just one cultivar. The soybean gene pool, encompassing wild progenitors (*Glycine soja*), landraces, and thousands of elite cultivars, contains a wealth of genetic diversity not captured in a single reference. To access

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

this diversity, numerous soybean pangenome projects have been undertaken. These projects have revealed a dynamic and variable genomic landscape. An early pangenome of 26 wild and cultivated accessions suggested that up to 50% of genes could be dispensable (Liu et al., 2020). A more recent and comprehensive pangenome, PanSoy, constructed from 204 representative cultivated accessions, identified 108 Mb of novel sequence not present in the Williams 82 reference, containing 3,621 genes, of which 1,659 were entirely new (Torkamaneh et al., 2021).

These studies have cataloged hundreds of thousands of structural variants, including 723,862 presence/absence variations (PAVs) and 27,531 copy number variations (CNVs) in one study alone (Liu et al., 2020). This variation is functionally significant; analysis of the soybean gene pool has shown that 79% of rare, potentially beneficial alleles were lost during the bottlenecks of domestication and modern breeding, highlighting the importance of pangenome resources for re-discovering and re-introducing valuable genetic diversity. Interestingly, despite the large amount of variable sequence, the core genome of cultivated soybean is remarkably conserved, with over 90% of genes being shared across nearly all accessions in the PanSoy study, reflecting the selective pressures of domestication and breeding (Torkamaneh et al., 2021).

Key Genes Regulating Harvest Index and Yield Components.

Soybean yield is a complex trait determined by the interplay of numerous components, including pod number, seed number per pod, and seed weight. The harvest index is strongly influenced by plant architecture, developmental timing, and the efficiency of nutrient transport to developing seeds. Plant architecture is a key determinant of light interception and lodging resistance. The GmDW1 gene on chromosome 8, which encodes an ent-kaurene synthase involved in the early steps of gibberellin (GA) biosynthesis, is a major dwarfing gene. A natural mutation in this gene leads to a dwarf phenotype with enhanced lodging resistance and an

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

improved harvest index (Li et al., 2018). Developmental timing, which controls the overall growth habit, is regulated by the Determinate Stem loci. Dt2 on chromosome 18, which encodes a MADS-domain transcription factor, is a key gene controlling the termination of stem growth (Ping et al., 2023).

Seed size and weight are major yield components that have been under strong selection. A recent large-scale GWAS identified GmSW17 on chromosome 17 as a major positive regulator of seed width and weight (Lyu et al., 2024). This gene encodes a ubiquitin-specific protease that regulates the cell cycle by deubiquitinating histone H2B. By modulating the G1-to-S phase transition, it influences the final number of cells in the seed coat, thereby controlling seed size. The efficiency of seed filling is critically dependent on sugar transport from source leaves to sink seeds. The GmSWEET15a and GmSWEET15b genes, located on chromosomes 8 and 18, are sucrose transporters essential for exporting sucrose to the developing embryo (Wang et al., 2019). Double knockouts of these genes result in aborted seeds, demonstrating their central role in carbon allocation and yield formation. Finally, the timing of maturity, which defines the duration of the seed-filling period, is controlled by a series of E loci. GmG1a (GIGANTEA), the gene underlying the E2 maturity locus on chromosome 10, is a key regulator of flowering time and maturity in response to photoperiod (Watanabe et al., 2011).

GWAS, QTL Mapping, and Breeding Applications.

The genetic architecture of soybean yield and its components is highly complex and polygenic. Decades of QTL mapping and GWAS have identified hundreds of loci associated with these traits. SoyBase, the primary database for soybean genetics, catalogs over 200 QTLs for seed weight alone, distributed across all 20 chromosomes. GWAS has been particularly powerful in dissecting this complexity, with studies identifying numerous quantitative trait

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

nucleotides (QTNs) for seed weight. Many of the major genes controlling yield components have been identified through the fine-mapping of major QTLs. The GmSW17 gene was pinpointed through a large-scale GWAS for seed size (Lyu et al., 2024). The vast amount of data generated from these studies provides a rich resource of markers and candidate genes for improving soybean yield.

Soybean breeding has achieved remarkable success, with yields steadily increasing over the past century. The future of soybean breeding is focused on leveraging the full power of genomics to tackle the complexity of quantitative traits like yield. Genomic selection (GS) is rapidly becoming the technology of choice, as it can more effectively capture the small, additive effects of the many genes that control yield, leading to more rapid and accurate selection (Jarquin et al., 2014). Pangenome resources are providing a new frontier for discovery. By looking beyond the reference genome, breeders can now identify and utilize favorable alleles from landraces and wild relatives that were lost during domestication. Finally, gene editing with CRISPR/Cas9 is set to revolutionize soybean improvement, allowing breeders to directly edit genes like GmSW17 to create larger seeds, modify Dt genes to optimize plant architecture, or enhance oil and protein content.

2.5 Bioregenerative Life Support Systems and Space Agriculture

2.5.1 Introduction to BLSS Concepts

A Bioregenerative Life Support System (BLSS), also known as a Controlled Ecological Life Support System (CELSS), is a meticulously engineered, artificial ecosystem designed to support human life in isolated and extreme environments (NASA, 1984). These systems operate on fundamental ecological principles, structurally mimicking Earth's terrestrial biosphere by integrating a symbiotic network of producers (plants and microalgae), consumers (humans), and

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

decomposers (microorganisms) (Wheeler, 2017; Paradiso et al., 2023). The primary objective of a BLSS is to achieve a high degree of material closure, thereby minimizing reliance on external supplies through the in-situ regeneration of oxygen, purification of water, and production of food, while simultaneously recycling all metabolic and agricultural waste. This is achieved by creating closed loops for essential elements, particularly carbon, oxygen, and hydrogen. In this intricate web, photosynthetic organisms serve as the cornerstone, consuming the carbon dioxide exhaled by the crew, producing breathable oxygen and edible biomass through photosynthesis, and purifying water via transpiration. Concurrently, a sophisticated microbial community is tasked with the decomposition of all organic waste—including human metabolic products, inedible plant matter, and food scraps—breaking it down into constituent nutrients and minerals that can be fed back into the hydroponic or aeroponic systems to nourish the next generation of crops.

The conceptual foundations of BLSS date back to the 1960s, emerging alongside the dawn of the space age (MacElroy & Bredt, 1984). Early theoretical work recognized that for missions extending beyond a certain duration, estimated to be around two years, the mass penalty of launching all required consumables would become insurmountable, making in-situ resource regeneration the only viable path forward. This realization spurred pioneering research programs by the world's leading space agencies. The Soviet Union was a prominent early leader with its BIOS series of experiments conducted at the Institute of Biophysics in Krasnoyarsk. Completed in 1972, BIOS-3 was a 315 m³ hermetically sealed underground habitat that integrated human crews, higher plants (wheat and vegetables), and algal cultivators. Over a series of manned experiments lasting up to six months, BIOS-3 demonstrated remarkable success, achieving nearly 100% closure of atmospheric gases, 85% for water, and approximately 50-55% for food,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

proving the fundamental viability of human habitation within a largely self-sustaining artificial ecosystem (Salisbury et al., 1997).

Concurrently, NASA initiated its own research, which formalized into the Controlled Ecological Life Support System (CELSS) program in 1979. The CELSS program focused heavily on the biological components, investigating the potential of both algae and higher plants for air, water, and food regeneration (NASA, 1984). This work evolved into the Advanced Life Support (ALS) program in 1989, which adopted a broader, systems-engineering approach. The ALS program's mission was to develop safe, efficient, and cost-effective closed-loop systems by integrating both bioregenerative and advanced physical/chemical technologies (Drysdale et al., 2003). A key management tool developed under ALS was the Equivalent System Mass (ESM) metric, a comprehensive analytical method used to compare the lifecycle costs of advanced regenerative technologies against baseline systems used on the International Space Station.

International collaboration and parallel development have also been central to BLSS research. The European Space Agency (ESA) initiated its Micro-Ecological Life Support System Alternative (MELiSSA) project in 1989 (Lasseur et al., 2010). Inspired by the functional ecology of a lake ecosystem, MELiSSA is a highly compartmentalized loop designed for microbial-based waste processing and resource regeneration. The system consists of four interconnected biological compartments, plus the crew, that progressively break down waste. The final stage contains both the cyanobacterium *Arthrospira platensis* (spirulina) and a suite of higher plants, which produce food and oxygen for the crew, thus closing the loop. Similarly, China has made significant advancements with its Lunar Palace 1 (Integrative Experimental Facility for Permanent Astrobase Life-support Artificial Closed Ecosystem), which has conducted long-duration crewed experiments, including the 365-day Lunar Palace 365 mission. This facility is

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

notable for achieving an unprecedented 98% material closure rate by integrating plants, microorganisms, and animals into its biological loop (Fu et al., 2016). These diverse, long-term international programs underscore the global consensus on the criticality of BLSS for the future of human space exploration.

2.5.2 Crop Selection for Space Agriculture

The selection of plant species for cultivation within a BLSS is a multifaceted process governed by a stringent set of criteria that balance physiological performance, nutritional output, system efficiency, and crew well-being. NASA has developed a comprehensive framework for evaluating candidate crops, prioritizing those that are not only productive but also exceptionally well-suited to the unique constraints of a controlled space environment (Wheeler, 2017; Massa et al., 2021). The primary physiological criteria include a compact or dwarf growth habit to maximize the use of limited cultivation volume, a high harvest index to ensure that a maximal proportion of the plant's total biomass is edible, a rapid growth rate and short life cycle to allow for frequent harvests, and high overall productivity in terms of edible biomass per unit area, time, and energy input. Furthermore, crops must demonstrate robust performance under the specific conditions of a space-based growth chamber, including tolerance to elevated carbon dioxide concentrations (typically 1,000-3,000 ppm), efficient utilization of the specific spectra provided by LED lighting, and compatibility with hydroponic or aeroponic cultivation systems. Beyond these performance metrics, organoleptic acceptance is crucial; the food must be palatable and desirable to astronauts to combat menu fatigue and ensure adequate caloric intake. Finally, food safety is paramount, requiring crops that are resistant to microbial contamination and can be grown and processed safely within the closed habitat.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

A strategic distinction is made between crops intended for near-term missions, such as those on the ISS, and those required for long-term, self-sufficient planetary bases. For current and near-future missions, the focus is on pick-and-eat salad crops—such as lettuce, radishes, leafy greens (mizuna, pak choi), and small fruiting vegetables like dwarf tomatoes and peppers (Massa et al., 2021). These crops serve as a supplemental food source, providing essential micronutrients, particularly vitamins C and B1, and antioxidants that are known to degrade in prepackaged space rations over time. Just as importantly, the cultivation, harvesting, and consumption of fresh, flavorful food provide significant psychological benefits for crews living in isolated, confined, and monotonous conditions. However, these supplemental crops cannot provide the bulk caloric and macronutrient needs of a crew.

For the establishment of a truly bioregenerative and self-sufficient food system on the Moon or Mars, a different class of crops becomes essential: staple crops. These are the calorie-dense, agronomically significant species that form the foundation of human diets on Earth. The primary candidates consistently identified in BLSS research are wheat, rice, potato, and soybean (Wheeler, 2017; Wheeler, 2022). Wheat and rice are invaluable sources of carbohydrates; potatoes are exceptionally productive in terms of calories per unit area and also provide carbohydrates and essential vitamins; and soybeans are a critical source of both high-quality protein and lipids (oils), which are otherwise difficult to supply in a plant-based diet. The integration of these four staple crops creates a nutritionally complementary portfolio capable of providing the vast majority of a crew's required carbohydrates, proteins, and fats, thereby drastically reducing the mission's dependence on costly and massive resupply from Earth. The transition from supplemental salad crops to large-scale cultivation of staple crops represents a

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

key phase in the maturation of a BLSS, moving from a system that enhances a prepackaged diet to one that largely replaces it.

Achieving a complete and balanced diet from a limited selection of space-grown crops requires careful planning for nutritional complementarity. While the combination of staples like rice, wheat, potato, and soybean can provide the necessary macronutrients, ensuring a full spectrum of vitamins and minerals is a significant challenge. Research has shown that the nutritional profile of plants grown in space can differ from their terrestrial counterparts (Khodadad et al., 2020). This highlights the need for a diverse menu that includes not only the staple crops but also a variety of nutrient-dense leafy greens and fruits. Furthermore, it underscores the importance of optimizing the nutrient solutions and environmental conditions to enhance the nutritional quality of the produce. Advanced diet planning involves the use of computational models to design space dishes that meet all of NASA's stringent dietary guidelines by combining available crops in optimal ratios. In cases where nutritional gaps persist, strategies such as the biofortification of crops through targeted breeding or gene editing, or the integration of nutrient-producing microorganisms into the BLSS, are being actively investigated to ensure the long-term health and performance of astronaut crews.

2.5.3 Controlled Environment Agriculture for Space

The practice of space agriculture is fundamentally an application of Controlled Environment Agriculture (CEA), a technology-intensive approach to food production that optimizes plant growth by precisely managing all environmental variables. Within the confines of a spacecraft or planetary habitat, CEA technologies are not merely advantageous; they are enabling. The absence of a natural atmosphere, soil, and sunlight necessitates the complete artificial recreation of a plant-friendly environment. Key technologies in this domain include

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

specialized lighting, atmospheric management, soilless cultivation techniques, and highly efficient resource recycling systems, all of which must be adapted to function under the severe constraints of spaceflight, including microgravity and elevated radiation.

Artificial lighting is the cornerstone of space-based CEA, as it must fully replace the sun. Light-Emitting Diodes (LEDs) have become the undisputed standard for this application, a technology that NASA itself helped pioneer for plant growth (Morrow, 2005). LEDs offer a suite of advantages perfectly suited for space. They are exceptionally energy-efficient, capable of converting electricity to photosynthetic light with far greater efficacy than traditional high-pressure sodium lamps, which is a critical factor when power is a limited and precious resource. They also produce very little radiant heat, allowing them to be placed close to the plant canopy without causing thermal stress. This characteristic is instrumental for designing compact, multi-layered vertical farming systems that maximize food production within a minimal volume. Perhaps most importantly, LEDs provide unprecedented control over the light spectrum. Growers can tailor the specific wavelengths of light to match the photosynthetic needs of a particular crop or growth stage, for example, by providing a precise ratio of red and blue light to drive photosynthesis, and even incorporating far-red light to influence plant morphology and canopy development (Stutte et al., 2015).

Complementing advanced lighting is the active management of the growth atmosphere, particularly through carbon dioxide (CO₂) enrichment. The enclosed environment of a spacecraft naturally leads to higher concentrations of CO₂ from crew respiration, typically in the range of 1,000 to 3,000 parts per million (ppm), significantly higher than Earth's ambient approximately 420 ppm (Wheeler, 2006). For most plants, this enriched CO₂ environment is beneficial, acting as a fertilizer that boosts the rate of photosynthesis and can lead to substantial increases in

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

growth rate and yield. However, the response is crop-specific, and excessively high concentrations can be detrimental to some species. Therefore, CEA systems in space must carefully monitor and control CO₂ levels, balancing the needs of the plants with the life support system's capacity to scrub CO₂, ensuring an optimal concentration that maximizes productivity without negatively impacting the crew or the plants themselves.

To overcome the challenges of using soil—which is heavy, can harbor pathogens, and behaves unpredictably in microgravity—space agriculture relies exclusively on soilless cultivation techniques, primarily hydroponics and aeroponics. In hydroponic systems, plant roots are bathed in or intermittently flooded with a nutrient-rich water solution. In aeroponic systems, the roots are suspended in air and periodically misted with a nutrient aerosol. Both methods offer precise control over nutrient delivery, conserve water through recirculation, and eliminate soil-borne diseases (Morrow et al., 1995). However, managing fluids in microgravity is a significant engineering challenge. Water does not drain under gravity and can form cohesive blobs, potentially starving roots of oxygen. To address this, space-based systems employ novel designs that rely on capillary action, wicking materials, and carefully engineered channels to passively guide the nutrient solution to the roots while ensuring adequate aeration. Experiments like NASA's eXposed Root On-Orbit Test System (XROOTS) on the ISS are actively testing and refining these hydroponic and aeroponic techniques for reliable operation in space.

The space environment presents two overarching challenges that affect all biological components: microgravity and cosmic radiation. Plants have evolved for millions of years under Earth's constant 1g gravitational field, which they use as a cue for growth orientation (gravitropism) (Ferl & Paul, 2016; Vandenbrink & Kiss, 2016). In microgravity, this cue is absent, which can disrupt root and shoot development and alter the distribution of growth

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

hormones like auxin. Furthermore, beyond the protection of Earth's magnetosphere, plants are exposed to high levels of galactic cosmic rays and solar energetic particles. This ionizing radiation can cause direct damage to cellular structures and DNA, leading to mutations, reduced growth, and decreased viability (Arena et al., 2014). Understanding and mitigating these stressors through shielding, environmental control, and the selection or engineering of more resilient plant varieties are central challenges for the future of space agriculture.

2.5.4 Harvest Index Optimization for BLSS

In the resource-limited context of a closed-loop bioregenerative system, the concept of Harvest Index (HI) assumes paramount importance. Traditionally defined in terrestrial agriculture as the ratio of harvested grain yield to total above-ground biomass, HI is a measure of a plant's partitioning efficiency. For a BLSS, where every gram of mass, watt of power, and liter of volume is meticulously accounted for, optimizing HI is not just a matter of increasing yield but is a fundamental strategy for maximizing the overall efficiency of the entire life support system (Wheeler, 2017). A high HI means that a greater fraction of the resources invested in growing a plant—water, nutrients, light, and air—is converted directly into edible food for the crew. This, in turn, minimizes the quantity of inedible biomass that must be processed, recycled, or stored as waste, thereby reducing the load on the waste management subsystem and improving the closure of the system's carbon loop.

The pursuit of a higher HI in a BLSS has led to an expanded definition of the term. While the traditional HI focuses on a single edible component like grain or tubers, a more holistic edible biomass ratio is often considered for space applications. This approach accounts for all edible parts of the plant, including leaves, stems, and fruits. Furthermore, advanced concepts in bioregeneration propose redefining the harvest index to include the potential food value of

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

otherwise inedible components. For example, the cellulosic material that constitutes the bulk of stems and leaves in many staple crops can be broken down through enzymatic hydrolysis into simple, edible sugars like glucose (Wheeler, 2006). By factoring in this potential caloric recovery, the effective HI of a crop can be significantly increased, maximizing the nutritional return from every plant grown and further reducing the system's waste stream. This re-conceptualization treats the entire plant as a potential resource to be exploited for crew sustenance.

Perhaps the most effective and widely adopted strategy for optimizing both HI and overall system productivity is the selection and development of dwarf cultivars. The advantages of dwarfism in space agriculture are manifold. Their compact growth habit is the most obvious benefit, allowing for higher planting densities within the severely constrained volume of a space habitat, which directly translates to greater yield per unit area (Bugbee & Monje, 1992; Bugbee, 1995). These cultivars are specifically bred and selected not only for their short stature but also for a high harvest index, meaning they are genetically predisposed to allocate a larger proportion of their photosynthates to edible organs—be it grains, tubers, or fruits—rather than to extensive stems and leaves. Cultivars such as the USU-Apogee dwarf wheat, developed specifically for BLSS applications, exemplify this approach, offering high yields in a compact form (Bugbee & Monje, 1992).

However, the optimization of reproductive growth in these cultivars presents unique challenges. Seminal experiments with Super Dwarf wheat grown aboard the Mir space station revealed that while the plants grew vigorously, they were completely sterile, failing to produce any grain (Levinskikh et al., 2000). This reproductive failure was attributed to the accumulation of ethylene gas in the sealed environment, a potent plant hormone that disrupted pollen

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

development. This critical finding underscores that maximizing HI for staple crops is not merely a matter of promoting growth, but requires a sophisticated understanding of the complex interplay between genetics, the controlled environment, and the full reproductive lifecycle of the plant in space.

2.5.5 Current BLSS Crop Research

The development of robust crops and cultivation systems for BLSS is an iterative process that relies on a combination of ground-based analog studies, in-orbit experiments, and forward-looking mission planning. Terrestrial analogs are indispensable for testing system-level concepts and studying human-in-the-loop dynamics in environments that simulate the isolation and confinement of space missions. The most ambitious of these was Biosphere 2, a 3.14-acre materially closed ecological system in Arizona (Nelson et al., 1993). Its two crewed missions in the early 1990s, which attempted to sustain eight people on food grown within its soil-based agricultural biome, provided invaluable, albeit challenging, lessons on the complexity of managing a large, biodiverse artificial ecosystem, including unforeseen issues with atmospheric balance and species viability. Today, the Biosphere 2 facility hosts the Space Analog for the Moon and Mars (SAM), a smaller, hermetically sealed habitat used for shorter-duration missions to test next-generation life support technologies, food production strategies, and human performance in a simulated off-world environment. Another critical set of analogs are the remote research stations in Antarctica, such as Concordia Station (Pletser et al., 2005). Dubbed White Mars, these bases provide an unparalleled opportunity to study the psychological and social dynamics of small, isolated crews over long periods and to test hardware in an extreme, resource-scarce environment.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

While ground analogs are crucial for system-level testing, only experiments conducted in space can reveal the true effects of microgravity and the space radiation environment on plant biology. The International Space Station (ISS) serves as the primary laboratory for this research, hosting two key plant growth facilities. The first is the Vegetable Production System (Veggie), a simple, low-power chamber designed primarily for applied research on food production (Massa et al., 2016). Veggie is an open system, utilizing the ambient cabin atmosphere of the ISS. It has been used to successfully grow a variety of pick-and-eat salad crops, including red romaine lettuce, mizuna mustard, and kale, as well as zinnia flowers. These experiments have not only demonstrated the feasibility of growing safe and nutritious food in orbit but have also highlighted the profound psychological benefits that gardening provides to astronauts. The second facility, the Advanced Plant Habitat (APH), is a much more sophisticated, fully enclosed, and highly automated chamber (Morrow et al., 2016). APH allows scientists to precisely control all environmental parameters—light spectrum and intensity, temperature, CO₂, humidity, and irrigation—making it an ideal platform for fundamental plant physiological research. Experiments in APH have investigated the genetic and molecular responses of plants like *Arabidopsis thaliana* and dwarf wheat to spaceflight and have successfully cultivated more complex crops, such as radishes and chile peppers, from seed to harvest.

The knowledge gained from these ground and space-based experiments directly informs planning for future long-duration missions to the Moon and Mars. The goal for these missions is to transition from supplemental food production to a fully bioregenerative system capable of providing the majority of the crew's diet. This involves tackling the significant challenge of cultivating crops in extraterrestrial regolith (soil). Lunar and Martian regolith is fundamentally different from terrestrial soil; it lacks organic matter and essential microbes, and Martian regolith

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

contains toxic perchlorate salts (Wamelink et al., 2019). Research is focused on several parallel strategies: using the regolith as a substrate within a hydroponic system, developing methods to detoxify and enrich the regolith with nutrients from composted waste and nitrogen-fixing bacteria to create a viable soil, and bypassing regolith entirely with large-scale hydroponic or aeroponic farms. Mission simulations on Earth, such as NASA's Crew Health and Performance Exploration Analog (CHAPEA), are actively testing these different food system concepts, including the integration of small indoor gardens for fresh produce within a simulated Mars habitat. These efforts are advancing the technologies and operational procedures needed to establish sustainable agriculture as a cornerstone of future human settlements on other worlds.

2.5.6 Integration of Genomics with BLSS Crop Improvement

The successful establishment of self-sustaining agriculture in space hinges on the development of crop varieties specifically tailored for the novel and stressful environment of a BLSS. While traditional breeding methods have been used to develop cultivars like dwarf wheat, the timeline and precision required for deep space exploration demand more advanced and rapid approaches. The integration of modern genomics with powerful gene editing technologies represents a transformative paradigm for space crop improvement, enabling the precise and accelerated engineering of plants with desirable traits. This synergy is essential for designing the next generation of staple crops that can thrive within the strict confines of a space habitat and reliably provide sustenance for astronaut crews.

At the heart of this new approach is the application of pangenomics. Historically, crop genetics has relied on a single reference genome for a given species, a high-quality sequence from one individual that serves as a template for comparison. However, a single reference cannot capture the full spectrum of genetic diversity within a species, including structural variations

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

such as gene presence/absence, copy number variations, and large-scale rearrangements (Bayer et al., 2020; Golicz et al., 2020). A pangenome, by contrast, is a comprehensive dataset that incorporates the genomic information from many different individuals or varieties of a species, creating a complete catalog of all genes and their allelic variants. For BLSS crop improvement, the pangenome is an invaluable tool. It allows researchers to move beyond a single template and explore the vast genetic landscape of a crop like rice or soybean to identify the specific genes and alleles responsible for key traits such as dwarfism, rapid maturation, high harvest index, enhanced nutritional content, and tolerance to stressors like salinity or low light. This comprehensive genetic map is the critical first step for targeted crop engineering.

The tool that unlocks the potential of the pangenome is gene editing, most notably the CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein) system. CRISPR-Cas functions as a pair of molecular scissors that can be programmed with a guide RNA (gRNA) to find and cut a specific DNA sequence within a plant's genome (Doudna & Charpentier, 2014). The cell's natural repair mechanisms then mend this break, a process that can be harnessed to disable a gene, modify its function, or insert a new one. The precision of CRISPR is revolutionary, allowing for targeted changes without the random integration associated with older genetic modification techniques. More advanced iterations, such as base editing and prime editing, offer even greater finesse, enabling the direct conversion of single DNA letters without making a double-strand break in the DNA, further reducing the risk of unintended off-target effects (Gao, 2021). By combining the comprehensive target identification provided by pangenomics with the precision of CRISPR, scientists can accurately and efficiently modify crops. For example, they can edit genes to create a more compact plant architecture, alter metabolic pathways to increase the accumulation of vitamins or proteins in

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

edible tissues, or enhance the activity of genes involved in stress response, creating a more resilient and productive plant.

Looking forward, the primary objective for BLSS crop development is the systematic integration of these genomic technologies to design space-optimized cultivars. The future priorities will involve generating high-quality pangenomes for all candidate staple crops, including wheat, rice, potato, and soybean. This genomic data will be coupled with extensive phenotyping—measuring the physical traits of these plants under simulated BLSS conditions—to build robust gene-to-trait maps. These maps will pinpoint the key genetic determinants of high harvest index, efficient resource use, and adaptation to the spaceflight environment. Armed with this knowledge, researchers will use CRISPR and other gene editing tools to assemble suites of beneficial traits into elite cultivars. This genomics-driven approach will dramatically accelerate the breeding cycle, enabling the development of highly productive, nutritious, and reliable crops on a timescale that is relevant to the ambitious goals of human exploration of the Moon, Mars, and beyond. This work is not merely academic; it is a foundational requirement for ensuring the long-term health, safety, and independence of human pioneers in the cosmos.

2.6 Synthesis and Research Gaps

The preceding sections of this literature review have examined four interconnected domains: the theoretical framework and methodological advances of pangenomics, the physiological and genetic foundations of harvest index, the genomic landscapes of the four BLSS staple crops, and the principles and challenges of bioregenerative life support and space agriculture. This synthesis integrates the key themes across these domains, identifies the gaps in current knowledge, and articulates how the research questions of this dissertation occupy a unique and necessary niche at their intersection. The overarching argument is that while

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

substantial progress has been made in each individual field, the explicit integration of pangenomic analysis with harvest index optimization for BLSS-relevant crops represents a significant, underexplored frontier with the potential to yield both fundamental scientific insights and practical benefits for enabling sustained human presence in space.

A central theme emerging from this review is the transformative impact of pangenomics on our understanding of crop genetic diversity. The shift from a single-reference genome paradigm to a pangenome framework has revealed that a substantial portion of a species' gene content is variable—dispensable genes that are present in some individuals but absent in others. For the staple crops examined in this review, these dispensable genomes constitute a significant fraction of the total genetic repertoire: approximately 43% of rice genes, over 44% of maize genes, and up to 50% of soybean genes are estimated to be dispensable across their respective species (Wang et al., 2018; Sun et al., 2022; Liu et al., 2020). The potato pangenome is even more striking, with only 12% of the sequence being conserved across all sampled haplotypes (Sun et al., 2024). This variable genetic content is not merely genomic noise; it is a rich reservoir of functional diversity, enriched for genes involved in environmental adaptation, stress response, and, critically, agronomic performance. Structural variations such as PAVs and CNVs have been directly linked to variation in disease resistance, grain quality, flowering time, and yield components across these crops.

The second major theme is the deep molecular understanding of the genetic architecture of harvest index and its component traits. The review has detailed the key physiological processes—source-sink relationships, carbon allocation, assimilate remobilization—that govern biomass partitioning. The Green Revolution stands as a testament to the power of genetic modification of these processes; the introduction of semi-dwarf genes in rice (*sd1*) and wheat

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

(Rht-B1b, Rht-D1b) increased harvest index from approximately 0.30 to over 0.50, revolutionizing global food production (Hedden, 2003; Khush, 2001). Beyond these landmark examples, a complex network of genes controlling plant height, grain size, grain number, starch biosynthesis, and sugar transport has been mapped and characterized in each of the four target crops. QTL mapping and GWAS have identified hundreds of genomic regions associated with these traits, and many of the underlying causal genes, such as GW2, GS3, GW5, TaGW2, StSP6A, and GmSW17, have been cloned and functionally validated. This wealth of knowledge provides a comprehensive catalog of potential targets for genetic improvement of harvest index.

The third theme is the unique set of constraints and opportunities presented by the BLSS environment. Space agriculture demands crop cultivars that are fundamentally different from those optimized for terrestrial field conditions. The ideal BLSS crop must be compact to maximize volumetric efficiency, have a high harvest index to maximize caloric return on invested resources, be adapted to the specific conditions of controlled-environment agriculture (elevated CO₂, LED lighting, hydroponics), and be robust under the unique stressors of spaceflight, including microgravity and radiation (Wheeler, 2017). While considerable progress has been made in developing dwarf cultivars like USU-Apogee wheat, and in demonstrating the feasibility of growing salad crops on the ISS, the systematic breeding or engineering of the major staple crops (rice, wheat, potato, soybean) for optimized harvest index and productivity in a BLSS context remains an immature field.

At the intersection of these three themes lies a significant and identifiable gap in current knowledge: the pangenomic architecture of harvest index determinants in BLSS staple crops has not been systematically characterized. While pangenomes have been constructed for each of these species, and while the major genes controlling harvest index have been identified using

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

traditional reference-based approaches, few if any studies have explicitly examined whether the dispensable genomes of these crops harbor functionally significant variation at key HI loci. This is a critical omission for several reasons. First, many of the known HI genes, such as those controlling grain size and plant height, are members of gene families that may include paralogs or related genes present in only a subset of cultivars; these PAVs could represent novel sources of allelic diversity for trait improvement. Second, reference bias inherent in traditional GWAS and QTL mapping may cause associations with causal variants in non-reference sequence to be missed; pangenome-based association studies have already demonstrated enhanced heritability estimates and the discovery of novel QTLs in other trait domains (Wang et al., 2025; Gao et al., 2022). Third, the domestication and modern breeding of these crops may have introduced bottlenecks that purged rare, beneficial alleles from the cultivated gene pool; exploring the pangenome, which includes wild relatives and diverse landraces, could allow for the recovery of this lost variation.

A second, related gap concerns the application of this pangenomic knowledge specifically to the context of BLSS. The environmental conditions of a BLSS—characterized by elevated CO₂, specific light spectra, and potential microgravity—may differentially affect the expression or impact of genes controlling harvest index. For example, the balance between vegetative and reproductive growth, and thus HI, is sensitive to CO₂ concentration and light quality (Wheeler, 2006). Allelic variants at HI-related loci that are neutral or even deleterious under terrestrial field conditions might confer advantages in a CEA environment, or vice versa. This gene-by-environment interaction in the context of BLSS has not been explored from a pangenomic perspective. Identifying alleles that optimize harvest index under BLSS-specific

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

conditions, rather than simply transferring alleles selected for performance in conventional agriculture, is essential for developing truly space-adapted cultivars.

A third gap pertains to the comparative analysis of harvest index genetics across the four target crops. While each crop has its own dedicated research community and genomic resources, few studies have attempted to synthesize the genetic architecture of harvest index across rice, wheat, potato, and soybean within a common analytical framework. This comparative perspective is valuable for identifying conserved molecular pathways and orthologous genes that represent universal targets for improvement, as well as for understanding the crop-specific variations that reflect their distinct evolutionary histories and modes of sink development (grains vs. tubers vs. seeds). A pangenomic approach is ideally suited for such a comparative analysis, as it can accommodate the different genome structures (diploid rice, hexaploid wheat, autotetraploid potato, paleopolyploid soybean) while identifying shared and unique variation at homologous HI loci.

This dissertation is designed to address these identified gaps. The central research questions ask: (1) What is the extent and nature of structural variation (PAVs, CNVs) at loci known to govern harvest index and its components in the pangenomes of rice, wheat, potato, and soybean? (2) Do the dispensable genomes of these species harbor functionally significant genes or alleles related to biomass partitioning that have been overlooked by reference-based studies? (3) How does the pangenomic architecture of HI determinants compare across these four crops, and what conserved and species-specific patterns emerge? (4) Can pangenomic analysis identify candidate alleles or variants that are predicted to optimize harvest index under BLSS-relevant environmental conditions? By systematically characterizing the pangenomic variation at HI loci,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

this research aims to create a comprehensive resource that connects the diverse genetic variation within each crop species to the specific functional targets relevant for space agriculture.

The justification for this research extends beyond filling an academic knowledge gap; it addresses a practical imperative for the future of human space exploration. The current trajectory of space agency planning envisions long-duration crewed missions to Mars in the 2030s and 2040s, necessitating a high degree of food self-sufficiency (NASA, 2023). Achieving this goal requires crop varieties that can provide maximal caloric return with minimal resource input—in essence, crops with a high and stable harvest index. The development of such varieties cannot rely solely on serendipitous discovery or slow, iterative conventional breeding. A genomics-driven, rationally designed approach is required, one that can identify the optimal genetic configurations for HI across diverse genetic backgrounds and target them with precision tools like gene editing. By providing the pangenomic foundation for this approach, this dissertation contributes directly to the enabling science for a bioregenerative food system that can sustain human life beyond Earth.

2.7 Chapter Summary

This chapter has provided a comprehensive review of the literature relevant to understanding the pangenomic architecture of harvest index determinants in staple crops for Bioregenerative Life Support Systems. The review traversed four major thematic areas, establishing the conceptual foundation and identifying the knowledge gaps that this dissertation seeks to address.

Section 2.2 examined the emergence and development of pangenomics as a transformative framework for crop genetics. The concept, originating from bacterial genomics with the work of Tettelin et al. (2005), recognizes that the complete genetic repertoire of a

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

species comprises a core genome shared by all individuals and a dispensable (accessory) genome consisting of genes present in only some members of the species. The extension of this framework to plant genomics has revealed that the dispensable genome of major crops can constitute 40-50% or more of the total gene content and is a critical reservoir of functional variation, particularly for environmental adaptation and agronomically important traits. The review detailed the methodological approaches for pangenome construction—reference-based, de novo assembly, and graph-based methods—as well as the technological advances in long-read sequencing and scaffolding that have enabled the construction of high-quality pangenomes. Landmark pangenome projects in rice, maize, tomato, wheat, and soybean were summarized, highlighting their contributions to understanding genomic diversity, structural variation, and the dispensable genome's role in phenotypic variation. The section concluded by discussing the applications of pangenomics in enhancing trait mapping (pangenome-GWAS), discovering novel stress-tolerance alleles, and improving genomic prediction models, positioning pangenomics as an essential tool for modern crop improvement.

Section 2.3 focused on harvest index, the central phenotype of this dissertation. The concept, formalized by Donald (1962, 1968), quantifies the efficiency of biomass partitioning to the economically valuable harvest product. The physiological basis of HI in source-sink relationships, carbon allocation, and assimilate remobilization was detailed. The Green Revolution served as the primary historical case study, demonstrating how the introduction of semi-dwarf genes in wheat (*Rht*) and rice (*sd1*) dramatically increased HI from approximately 0.30 to over 0.50, fundamentally transforming global food production. The genetic architecture of HI was reviewed, emphasizing its polygenic nature, moderate-to-high heritability, and the identification of numerous QTLs and GWAS hits associated with HI and its component traits

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

(plant height, grain size and number). The molecular determinants were examined in detail, including the gibberellin signaling pathway (DELLA proteins, GA20-oxidases), grain size regulators (GW2, GS3, GW5), starch biosynthesis enzymes (AGPase, GBSSI), and sugar transporters (SUTs, SWEETs). The section also addressed the trade-offs between yield components (grain number vs. grain weight) and reviewed modern breeding approaches—marker-assisted selection, genomic selection, and CRISPR/Cas9 gene editing—that are being applied to continue improving HI.

Section 2.4 reviewed the genomic landscape of the four BLSS staple crops: rice, wheat, potato, and soybean. For each species, the review covered the history and characteristics of its reference genome, the architecture of its pangenome and the extent of structural variation, the key genes regulating harvest index and yield components, major findings from GWAS and QTL mapping, and the application of genomic knowledge in breeding. Rice, with its high-quality IRGSP-1.0 reference and the massive 3K RGP pangenome resource, has identified key genes like *sd1*, *GW2*, *GS3*, *GW5*, and *Wx*. Wheat, despite its large and complex hexaploid genome, now has a comprehensive reference from the IWGSC and a developing pangenome, with *Rht-B1b*, *Rht-D1b*, and *TaGW2-6A* as key HI genes. Potato, an autotetraploid with a recently published graph-based pangenome revealing extreme structural variation (only 12% conserved sequence), has *StSP6A*, *StCDF1*, *StGA2ox1*, and *StGBSSI* as central regulators of tuberization and starch accumulation. Soybean, a paleopolyploid with extensive pangenome resources, has *GmDW1*, *Dt2*, *GmSW17*, and *GmSWEET15a/b* as key genes controlling architecture, seed size, and assimilate transport. This comparative analysis across the four crops provides the foundation for identifying both conserved pathways and species-specific mechanisms of HI regulation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Section 2.5 situated the genomic research within the applied context of BLSS and space agriculture. The principles of bioregenerative life support were introduced, tracing the history of programs like BIOS-3, NASA's CELSS/ALS, and ESA's MELiSSA. The criteria for crop selection were detailed, distinguishing between supplemental salad crops for near-term missions and calorie-dense staple crops for self-sufficient planetary bases. The technologies of controlled-environment agriculture—LED lighting, CO₂ enrichment, hydroponics, and aeroponics—and the unique challenges of the space environment (microgravity, radiation) were discussed. The section emphasized the critical importance of harvest index optimization for BLSS, where a high HI directly translates to system efficiency, and highlighted the proven value of dwarf cultivars. The integration of genomics and gene editing with BLSS crop improvement was presented as the path forward for developing space-optimized cultivars on a timescale relevant to upcoming exploration missions.

Section 2.6 synthesized the themes across all sections and identified the key gaps in current knowledge. While pangenomics has revealed extensive dispensable genetic variation in staple crops, and while the molecular genetics of harvest index is well characterized, the explicit integration of pangenomic analysis with HI optimization for BLSS crops has not been systematically undertaken. Specifically, the structural variation at HI-related loci within the dispensable genomes of these crops has not been cataloged, the potential for pangenome-based GWAS to discover novel HI alleles has not been exploited for space agriculture, and the comparative pangenomic architecture of HI across rice, wheat, potato, and soybean has not been examined. This dissertation addresses these gaps by characterizing the pangenomic variation at HI loci across the four target crops, with the goal of creating a resource that can inform the genomics-driven development of space-optimized cultivars with enhanced harvest index.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

In conclusion, this literature review establishes that the convergence of pangenomics, harvest index genetics, and the imperatives of space agriculture defines a timely and significant research frontier. The knowledge generated by this dissertation will contribute to both fundamental plant science—by elucidating the pangenomic architecture of a complex trait of global importance—and applied agricultural science—by providing targets for the rational design of crops capable of sustaining human life in the extraterrestrial environments of the coming decades.

CHAPTER 3: METHODOLOGY

3.1 Introduction and Research Design Overview

This chapter presents the methodological framework for investigating the pangenomic architecture of harvest index (HI) determinants across four bioregenerative life support systems (BLSS) staple crops: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*). The methodology integrates computational comparative genomics, phylogenomics, and population genetics approaches to systematically characterize HI-controlling gene families, their structural variants, and evolutionary relationships across these target species.

The research design addresses three primary research questions that guide the analytical framework of this dissertation:

RQ1: What is the pangenomic architecture of genes controlling harvest index and its component traits (grain/tuber weight, number, biomass partitioning) across BLSS staple crops, and which gene families show conserved versus species-specific roles?

RQ2: How do structural variants (presence/absence variants, copy number variants) in harvest index-associated gene families contribute to phenotypic variation, and are these variants systematically captured in elite cultivars developed for controlled environment agriculture?

RQ3: Can phylogenomic analysis of harvest index gene families identify ancestral versus derived allelic states, enabling prediction of which alleles are most amenable to engineering for enhanced performance?

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

To address these questions, this study employs a computational comparative genomics research design integrating sequence analysis, phylogenetics, and population genetics approaches. The rationale for this approach stems from the need to leverage the extensive publicly available genomic resources that have been developed for major crop species over the past two decades. Reference genomes, pangenome assemblies, and genome-wide association study (GWAS) databases now provide unprecedented opportunities for systematic cross-species comparison of complex trait architectures (Della Coletta et al., 2021).

A comparative pangenomic approach is justified for several methodological reasons. First, harvest index represents an integrative phenotype influenced by multiple interacting gene families controlling growth architecture, photoassimilate partitioning, and sink organ development (Sinclair, 1998). Single-species studies cannot determine whether similar phenotypic outcomes arise from conserved genetic mechanisms or convergent evolution of distinct pathways. Second, pangenomic analyses have revealed that reference genomes capture only a subset of species diversity; presence/absence variants (PAVs) and copy number variants (CNVs) segregating in germplasm may substantially impact HI but remain uncharacterized relative to this trait (Golicz et al., 2016). Third, the BLSS application context requires simultaneous optimization across multiple crop species, making cross-species comparison essential for identifying conserved targets amenable to unified engineering strategies versus species-specific mechanisms requiring tailored approaches.

3.1.1 Theoretical Framework

The theoretical framework underlying this methodology draws from three interconnected domains: comparative genomics theory, quantitative genetics, and evolutionary biology. Comparative genomics provides the conceptual basis for identifying orthologous genes across

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

species and inferring functional conservation from sequence similarity (Koonin, 2005). The principle of ortholog conjecture—that orthologs are more likely to share function than paralogs—guides the prioritization of cross-species gene comparisons (Gabaldón & Koonin, 2013).

Quantitative genetics theory informs the approach to harvest index as a complex polygenic trait influenced by multiple genes of varying effect sizes. The infinitesimal model and its extensions predict that HI variation arises from the cumulative effects of many loci, while QTL mapping and GWAS have identified specific genes of larger effect that explain substantial phenotypic variance (Mackay et al., 2009). This research focuses on genes of major effect that have been experimentally validated while acknowledging that numerous small-effect loci remain uncharacterized.

Evolutionary biology theory provides the framework for interpreting patterns of gene family evolution, including mechanisms of gene duplication, neofunctionalization, and subfunctionalization (Lynch & Conery, 2000). The distinction between orthologs (genes diverging through speciation) and paralogs (genes diverging through duplication) is central to predicting functional conservation and engineering transferability across species. Coalescent theory and population genetics principles guide the interpretation of structural variant frequency patterns in the context of selection during domestication and breeding.

3.1.2 Four-Phase Research Design

The research design consists of four sequential phases spanning 36 months: (1) HI Gene Inventory and Classification (Months 1-6), (2) Ortholog Identification and Phylogenomic Analysis (Months 7-14), (3) Structural Variant Analysis (Months 15-22), and (4) Evolutionary

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Analysis and Synthesis (Months 23-36). Each phase builds upon outputs from preceding phases while maintaining methodological independence to allow parallel processing where appropriate. This phased design enables systematic progression from gene-level characterization through gene family reconstruction to population-level structural variation analysis, culminating in evolutionary synthesis that integrates findings across all analytical levels.

3.2 Study Species and Rationale

The four target species were selected based on their designation as primary staple crops for BLSS applications by NASA's Advanced Life Support program, their representation of diverse plant families and photosynthetic/storage organ strategies, and the availability of high-quality reference genomes and pangenomic resources. Table 3.1 summarizes the key characteristics of each target species relevant to this research.

Table 3.1 Target Species Characteristics for Pangenomic Analysis of Harvest Index

Determinants

Characteristic	Rice	Wheat	Potato	Soybean
Scientific Name	<i>Oryza sativa</i>	<i>Triticum aestivum</i>	<i>Solanum tuberosum</i>	<i>Glycine max</i>
Family	Poaceae	Poaceae	Solanaceae	Fabaceae
Ploidy Level	Diploid (2n=24)	Hexaploid (2n=42)	Tetraploid (2n=48)	Paleopolyploid (2n=40)
Reference Genome	IRGSP-1.0 (373.8 Mb)	IWGSC RefSeq v2.1 (14.6 Gb)	DM v6.1 (741.6 Mb)	Wm82.a4.v1 (978 Mb)
Harvest Organ	Grain (caryopsis)	Grain (caryopsis)	Tuber (modified stem)	Seed (legume)
Typical HI Range	0.45-0.55	0.35-0.50	0.70-0.85	0.35-0.55
Primary Nutrient	Starch	Starch	Starch	Protein and oil

Note. Genome sizes and ploidy levels indicate relative complexity for comparative analysis. Rice serves as the primary reference for cereal comparative genomics due to its smaller genome size and extensive syntenic documentation (International Brachypodium Initiative, 2010). HI = harvest index.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Rice (*Oryza sativa*) was selected as the primary reference species for comparative analysis due to its status as a model organism for cereal genomics, extensive documentation of syntenic relationships with other grasses, and the comprehensive genetic characterization of HI-controlling genes including the semi-dwarf1 (*sd1*) Green Revolution allele (Sasaki et al., 2002). The 3K Rice Genomes Project provides the most extensive pangenomic resource available for any crop species, enabling robust population-frequency analyses of structural variants (Wang et al., 2018).

Wheat (*Triticum aestivum*) represents a hexaploid genome with documented homoeologous relationships among subgenomes, enabling analysis of gene dosage effects on HI. The *Rht* dwarfing genes that enabled the Green Revolution in wheat provide direct functional parallels to rice *sd1*, while the complex polyploid genome structure offers opportunities to assess the role of gene copy number in HI regulation (Walkowiak et al., 2020).

Potato (*Solanum tuberosum*) differs fundamentally from cereals in that the harvestable organ is a modified stem (tuber) rather than a seed, providing opportunity to assess whether convergent regulatory mechanisms control sink organ development across distinct morphological contexts. The recent publication of a graph pangenome for potato significantly enhances structural variant characterization capabilities (Tang et al., 2024).

Soybean (*Glycine max*) represents the Fabaceae and provides contrast as a nitrogen-fixing legume with distinct seed composition (high protein and oil rather than starch). The paleopolyploid nature of soybean allows assessment of ancient whole-genome duplication effects on HI gene family evolution (Liu et al., 2020).

3.3 Data Sources and Acquisition

This research relies exclusively on publicly available genomic, pangenomic, and phenotypic association data. The following sections detail the specific databases and resources that will be utilized for each data type.

3.3.1 Genomic Data Sources

Reference genome sequences and annotations form the foundation for all downstream analyses. Table 3.2 presents the primary genomic databases and the specific data types that will be retrieved from each source.

Table 3.2 Primary Genomic Databases and Data Types

Database	URL	Data Types	Species Coverage
NCBI GenBank	ncbi.nlm.nih.gov/genbank/	Reference genomes, gene sequences, protein sequences	All four species
Phytozome	phytozome-next.jgi.doe.gov	Reference genomes, functional annotations, gene families	All four species
Ensembl Plants	plants.ensembl.org	Reference genomes, comparative genomics, variation data	All four species
OrthoDB	orthodb.org	Ortholog relationships, phylogenetic profiles	All four species
PLAZA	bioinformatics.psb.ugent.be/plaza/	Gene families, phylogenetic trees, synteny maps	All four species
InterPro/Pfam	ebi.ac.uk/interpro/	Protein domain annotations, functional predictions	All four species

Note. URLs are provided without protocol prefix for formatting. All databases were accessed through standard web interfaces and programmatic APIs where available.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

For each target species, the following genome assemblies will serve as the primary reference: rice IRGSP-1.0 (GCF_001433935.1), wheat IWGSC RefSeq v2.1 (GCF_018294505.1), potato DM v6.1 (Spud DB), and soybean Wm82.a4.v1 (Phytozome). These assemblies represent the current gold-standard references for each species and are utilized by the major comparative genomics resources listed in Table 3.2.

3.3.2 Pangenomic Data Sources

Pangenomic resources enable characterization of structural variants and gene content variation beyond what is captured in single reference genomes. Table 3.3 summarizes the species-specific pangenome databases that will be utilized.

Table 3.3 Pangenome Resources by Species

Species	Resource Name	Accessions	Key Statistics	Reference
Rice	3K Rice Genomes / SNP-Seek	3,010	29M SNPs; 93,683 SVs; 12,465 novel genes	Wang et al., 2018
Rice	RPAN Database	453	50,995 pan-genes; 22,095 dispensable	Sun et al., 2017
Wheat	10+ Wheat Genomes Project	15	107,000 HC genes; 249,976 SVs	Walkowiak et al., 2020
Potato	Potato Pangenome (PPDB)	31 (60 haplotypes)	3,076 Mb pangenome; 133,264 SVs	Tang et al., 2024
Soybean	PanSoy / SoyBase	204	3,621 genes absent from reference	Torkamaneh et al., 2021

Note. SVs = structural variants; HC = high-confidence. Accession numbers indicate the diversity panel size used to construct each pangenome.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

3.3.3 Phenotypic and Association Data

Genome-wide association study (GWAS) and quantitative trait locus (QTL) databases provide the essential phenotype-genotype linkages required to identify HI-associated variants.

Table 3.4 presents the primary association databases that will be queried.

Table 3.4 GWAS and QTL Databases for Phenotype-Genotype Association Data

Database	URL	Scope	Key Statistics
GWAS Atlas	ngdc.cncb.ac.cn/gwas/	Multi-species plant GWAS	Rice: 163,479 associations; Wheat: 20,452
Gramene QTL	archive.gramene.org/qlt/	Cereal QTL database	Comprehensive rice and wheat QTL catalog
Rice SNP-Seek	snp-seek.irri.org	Rice variation database	HI-associated SNPs from 3K genomes
Triticeae Toolbox	wheat.triticeaetoolbox.org	Wheat/barley genetics	Phenotype-genotype data; breeding records
SoyBase	soybase.org	Soybean genetics database	>200 seed weight QTL; 231 QTNs identified
Spud DB	spuddb.uga.edu	Potato genomics	46M SNPs; 133,264 SVs with trait associations

Note. QTL = quantitative trait locus; QTN = quantitative trait nucleotide. Statistics represent the most recent available data.

3.3.4 Data Quality Assessment Criteria

Before incorporation into analysis pipelines, all retrieved data will be assessed against standardized quality criteria developed for genomic and pangenomic research. Reference genome assemblies will be evaluated using the following metrics: contig N50 values indicating assembly contiguity, BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness scores assessing gene space capture, and Long Terminal Repeat Assembly Index (LAI) scores measuring assembly quality in repetitive regions (Ou et al., 2018). Assemblies with BUSCO completeness < 90% or LAI scores < 10 (draft quality) will be flagged for cautious interpretation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Pangenome datasets will be assessed for diversity representation, structural variant calling methodology consistency, and documentation completeness. Key quality indicators include: number and phylogenetic distribution of accessions sampled, use of multiple complementary SV detection methods, independent validation rates for variant calls, and availability of raw data for reanalysis if needed. Pangenomes constructed from < 50 accessions will be interpreted as preliminary pending expansion.

3.4 Phase 1: HI Gene Inventory and Classification

The first analytical phase (Months 1-6) will compile a comprehensive inventory of genes associated with harvest index and its component traits across all four target species. This gene catalog forms the foundation for subsequent ortholog identification and structural variant analyses.

3.4.1 Gene Identification Criteria

Genes will be included in the HI inventory based on the following criteria: (1) experimental validation through QTL cloning studies demonstrating causal effects on HI or component traits; (2) mutant characterization showing significant alterations in grain/seed/tuber yield, weight, or number; (3) genome-wide association study (GWAS) hits with $p\text{-value} < 1 \times 10^{-5}$ for HI-related traits; or (4) documented functional annotation indicating involvement in key HI-controlling pathways (gibberellin signaling, starch biosynthesis, sucrose transport, sink organ development). Each gene entry will be documented with chromosomal coordinates, functional category, supporting evidence type, and database source(s).

Literature searches will be conducted using PubMed, Web of Science, and Google Scholar with search terms including "[species name] harvest index gene," "[species name]

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

grain/seed/tuber yield QTL cloned," and pathway-specific terms (e.g., "gibberellin signaling crop yield," "starch biosynthesis grain filling"). Reference lists of identified papers will be screened to identify additional relevant studies.

3.4.2 Functional Classification Schema

Identified genes will be classified into functional categories based on their molecular function and position within HI-controlling pathways. Table 3.5 presents the classification schema with representative genes from each category.

Table 3.5 Functional Categories of Harvest Index-Associated Genes

Functional Category	Molecular Function	Representative Genes	HI Effect Mechanism
Dwarfing/GA Signaling	GA biosynthesis or signaling	sd1, Rht-B1b, Rht-D1b, GmDW1, StGA2ox1	Reduced stem elongation; assimilate partitioning
Grain/Seed Size	Cell division/expansion control	GW2, GS3, GW5, GS5, TaGW2-6A, GmSW17	Altered cell number or size in sink organs
Starch Biosynthesis	Starch polymer synthesis	GBSSI (Wx), AGPase subunits, SS, SBE	Direct storage compound accumulation
Sucrose Transport	Sugar translocation	OsSUT1, SWEET15a/15b	Source-to-sink carbon flux regulation
Developmental Timing	Flowering/tuberization	Vrn1, Ppd-D1, StSP6A, StCDF1, GmGIa	Vegetative vs. reproductive duration
Stem Termination	Meristem determinacy	Dt2 (soybean)	Vegetative and reproductive allocation

Note. GA = gibberellin; SS = starch synthase; SBE = starch branching enzyme; AGPase = ADP-glucose pyrophosphorylase.

3.4.3 Gene Classification Workflow

Figure 3.1 illustrates the systematic workflow for gene identification, validation, and classification that will be implemented in Phase 1.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Figure 3.1 Gene Classification Workflow for HI-Associated Gene Inventory

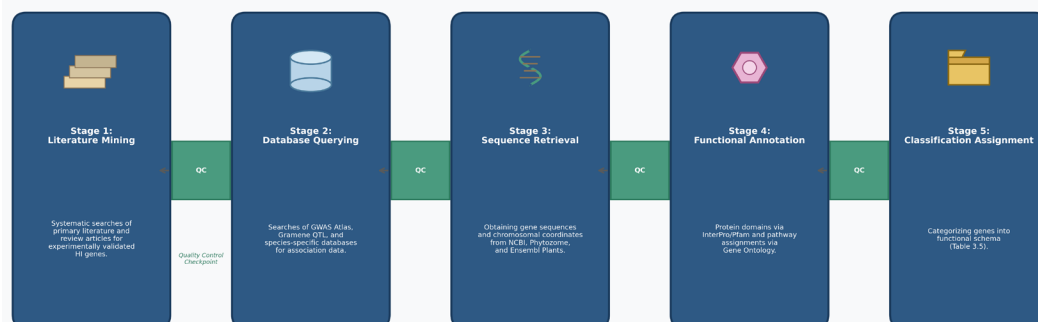


Figure 3.1. Gene Classification Workflow for HI-Associated Gene Inventory

The workflow depicted in Figure 3.1 proceeds through five sequential stages: (1) Literature Mining, involving systematic searches of primary literature and review articles for experimentally validated HI genes; (2) Database Querying, including searches of GWAS Atlas, Gramene QTL, and species-specific databases for association data; (3) Sequence Retrieval, obtaining gene sequences and chromosomal coordinates from NCBI, Phytozome, and Ensembl Plants; (4) Functional Annotation, determining protein domains via InterPro/Pfam and pathway assignments via Gene Ontology; and (5) Classification Assignment, categorizing genes into the functional schema presented in Table 3.5. Quality control checkpoints are implemented between stages to verify data integrity and resolve conflicting annotations.

Gene mapping to reference genome coordinates will employ BLASTP (protein sequences) and BLASTN (nucleotide sequences) with e-value threshold of 1×10^{-10} and minimum query coverage of 80%. For genes with established syntenic relationships, positional verification using MCScanX will confirm ortholog assignments. All gene identifiers will be cross-referenced across databases to ensure consistent nomenclature.

3.4.4 Gene Curation and Annotation Standards

Each gene entered into the HI catalog will be annotated following standardized curation protocols adapted from UniProt annotation guidelines (The UniProt Consortium, 2021).

Required annotation fields include: (1) gene symbol following species-specific nomenclature conventions; (2) systematic gene identifier from the reference genome annotation; (3) chromosomal position (chromosome number, start coordinate, end coordinate, strand); (4) protein product name and length; (5) functional category assignment from Table 3.5; (6) evidence code indicating support level (EXP = experimental evidence, IDA = inferred from direct assay, TAS = traceable author statement); and (7) literature citations in DOI format.

Genes will be classified into confidence tiers based on evidence strength: Tier 1 (highest confidence) includes genes with experimental validation through transgenic complementation, CRISPR knockout/knockin, or characterized natural mutants demonstrating causal effects on HI; Tier 2 includes genes identified through QTL fine-mapping with strong positional candidates or GWAS hits replicated across multiple studies; Tier 3 includes genes with functional annotation suggesting HI involvement but lacking direct experimental validation. Downstream analyses will be stratified by confidence tier to assess whether conclusions are robust across evidence levels.

3.5 Phase 2: Ortholog Identification and Phylogenomic Analysis

Phase 2 (Months 7-14) focuses on establishing orthologous relationships among HI genes across the four target species and reconstructing phylogenetic histories of major gene families. This phase addresses the core question of whether HI mechanisms are conserved or convergently evolved across species.

3.5.1 Ortholog Detection Methods

Ortholog identification will employ a multi-evidence approach combining sequence similarity and syntenic position. The primary tools for ortholog detection include:

OrthoFinder v2.5.5 (Emms & Kelly, 2019) will be used for de novo ortholog group inference from protein sequences. OrthoFinder implements an all-versus-all DIAMOND BLAST search followed by MCL clustering with species-specific normalization. The software reconstructs gene trees for each orthogroup and reconciles these with the species tree to distinguish orthologs from paralogs arising through gene duplication events.

OrthoDB v11 (Kuznetsov et al., 2023) provides pre-computed hierarchical ortholog groups at multiple taxonomic levels. For each HI gene, OrthoDB will be queried to retrieve orthologs at the Viridiplantae, Monocot/Dicot, and species-specific levels. The hierarchical structure enables assessment of ortholog conservation depth across evolutionary distances.

PLAZA 5.0 (Van Bel et al., 2022) integrates gene family clustering with synteny analysis. The PLAZA dicot and monocot portals will be utilized to verify ortholog assignments through collinear microsynteny blocks, providing positional evidence independent of sequence similarity.

3.5.2 Synteny Analysis Approach

Synteny analysis will be performed using MCScanX (Wang et al., 2012) to identify collinear blocks between and within genomes. The analysis proceeds through the following steps: (1) all-versus-all BLASTP of protein sequences with $e\text{-value} < 1 \times 10^{-5}$; (2) identification of collinear gene pairs using MCScanX with default parameters (minimum block size = 5 genes, maximum gaps = 25 genes); (3) visualization of syntenic relationships using SynVisio; and (4)

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

classification of duplications as whole-genome, tandem, proximal, dispersed, or segmental based on genomic positions.

For wheat, the hexaploid genome structure requires special consideration. Syntenic relationships among the A, B, and D subgenomes will be assessed to identify homoeologous gene copies, enabling analysis of subgenome-specific expression patterns and structural variants at HI loci.

3.5.3 Phylogenetic Reconstruction Methods

Phylogenetic reconstruction of HI gene families will employ maximum likelihood methods implemented in IQ-TREE v2.2.0 (Minh et al., 2020). The analytical pipeline consists of the following steps:

Multiple sequence alignment of protein sequences will be performed using MAFFT v7 (Katoh & Standley, 2013) with the L-INS-i algorithm optimized for sequences with conserved domains. Alignment quality will be assessed using TrimAl v1.4 (Capella-Gutiérrez et al., 2009) to remove poorly aligned regions while preserving phylogenetically informative sites.

Model selection will utilize ModelFinder (Kalyaanamoorthy et al., 2017) integrated within IQ-TREE to identify the optimal amino acid substitution model for each gene family alignment. The Bayesian Information Criterion (BIC) will be used for model selection, with common plant protein evolution models (JTT, WAG, LG) and rate heterogeneity (+G, +I+G) considered.

Tree inference will employ the IQ-TREE maximum likelihood algorithm with 1,000 ultrafast bootstrap replicates (Hoang et al., 2018) to assess nodal support. Branch support values $\geq 95\%$ will be considered strongly supported. Alternative topology testing using approximately

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

unbiased (AU) tests will be performed for major clades to assess statistical confidence in evolutionary relationships.

Figure 3.2 illustrates the phylogenomic analysis pipeline from sequence input through tree interpretation.

Figure 3.2 Phylogenomic Analysis Pipeline for HI Gene Family Evolution

Figure 3.2: Phylogenomic Analysis Pipeline for HI Gene Family Evolution

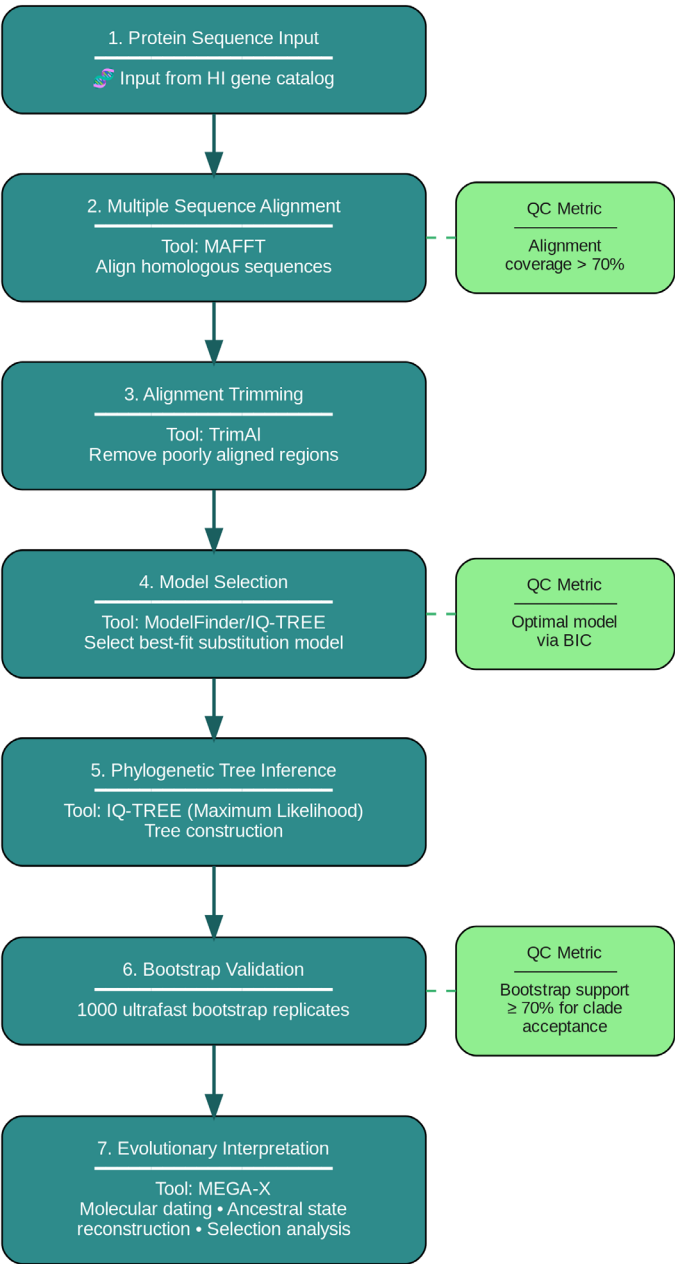


Figure 3.2. Phylogenomic Analysis Pipeline for HI Gene Family Evolution

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The pipeline depicted in Figure 3.2 shows the sequential steps from protein sequence input through alignment, model selection, tree inference, bootstrap validation, and evolutionary interpretation. Quality control metrics at each stage ensure analytical rigor: alignment coverage > 70%, optimal model selection via BIC, and bootstrap support thresholds for clade acceptance.

3.5.4 Molecular Dating Methodology

Molecular clock analysis will estimate divergence times for HI gene duplications using the RelTime method (Tamura et al., 2012) in MEGA-X. Calibration points will be derived from established plant divergence times: monocot-dicot split at 150-200 Mya, rice-wheat divergence at 40-60 Mya, and the Glycine whole-genome duplication at approximately 13 Mya (Schmutz et al., 2010). Dating analysis will focus on identifying whether gene duplications affecting HI occurred before or after domestication events, informing the interpretation of derived versus ancestral allelic states.

3.5.5 Handling Polyploid Genomes

The hexaploid genome of wheat and tetraploid genome of potato require specialized approaches for ortholog identification. For wheat, the three subgenomes (A, B, D) arose through recent polyploidization events (~10,000 years ago for AABBDD hexaploid formation) that predate but overlap with domestication (Marcussen et al., 2014). Gene sequences from the three homoeologous loci (e.g., Rht-B1, Rht-D1, and the A-genome copy) must be distinguished from true paralogs arising through pre-polyploidy duplication events.

The analytical strategy for polyploid genomes involves: (1) retrieving all homoeologous copies from each subgenome; (2) constructing gene trees including all homoeologs to assess their relationships; (3) identifying the diploid progenitor species (*Triticum urartu* for A genome,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Aegilops speltoides-related for B genome, *Aegilops tauschii* for D genome) sequences as outgroups; and (4) using the polyploid gene tree topology to infer which homoeolog is most similar to each diploid progenitor. This approach enables proper interpretation of gene copy number in polyploids versus diploids and identification of subgenome-specific losses that may affect HI.

For potato, the autotetraploid nature (4 copies of each chromosome) complicates haplotype phasing and copy number inference. Where possible, analyses will utilize the doubled monoploid (DM) reference genome which is effectively diploid, supplemented by examination of heterozygous tetraploid assemblies from the pangenome project. Structural variants showing dosage-dependent phenotypic effects will be highlighted as candidates for copy number optimization in breeding.

3.6 Phase 3: Structural Variant Analysis

Phase 3 (Months 15-22) characterizes presence/absence variants (PAVs) and copy number variants (CNVs) at HI-associated loci using available pangenome datasets. This phase addresses the pangenomic dimension of HI architecture by quantifying structural diversity absent from reference genomes.

3.6.1 PAV Detection and Cataloging

Presence/absence variants at HI loci will be cataloged from existing pangenome datasets rather than de novo detection. For each species, PAV calls will be retrieved from the resources listed in Table 3.3:

For rice, PAV data from the 3K Rice Genomes Project (Wang et al., 2018) will be extracted for all HI gene coordinates. The 3K dataset includes PAV calls based on read mapping

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

depth analysis across 3,010 diverse accessions representing indica, japonica, aus, and aromatic subpopulations. Gene-level PAV status (present/absent) will be determined based on published thresholds (coverage depth $< 0.1\times$ for absence calls).

For wheat, PAV information from the 10+ Wheat Genomes Project (Walkowiak et al., 2020) will be utilized. Given the hexaploid genome structure, PAV analysis will distinguish between complete gene loss and homoeolog-specific deletions. The Chinese wheat pangenome dataset provides additional SV calls for cultivar-specific analyses.

For potato, the graph pangenome (Tang et al., 2024) provides comprehensive SV characterization including 133,264 structural variants with metabolite associations. PAVs at HI gene coordinates will be extracted from the Potato Pangenome Database (PPDB).

For soybean, PAV data from PanSoy (Torkamaneh et al., 2021) and the wild-cultivated soybean pangenome (Liu et al., 2020) will be combined to maximize diversity sampling. The identification of 3,621 genes absent from the Williams 82 reference provides a starting point for assessing HI gene dispensability.

3.6.2 CNV Analysis Methods

Copy number variants will be characterized for HI gene families known to exhibit dosage sensitivity, particularly starch biosynthesis enzymes (AGPase subunits) and sucrose transporters. CNV data will be obtained from the following sources:

Rice CNV calls from the 3K project include 31 million CNV segments across the diversity panel. For each HI gene, the distribution of copy number states (0, 1, 2, 3+) will be tabulated across subpopulations.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Wheat CNV analysis will leverage published polyploidy-aware methods that distinguish homoeolog-specific copy number from ploidy effects. Copy number at Rht, TaGW2, and Vrn loci will be characterized across the 10+ genomes assemblies.

Potato and soybean CNV analysis will utilize read depth-based calls from their respective pangenome projects, with validation through comparison to assembly-based copy number estimates where chromosome-level assemblies are available.

3.6.3 Population Frequency Calculations

For each structural variant at HI loci, population frequencies will be calculated across available diversity panels. Frequency calculations will be stratified by:

Subpopulation/subspecies (e.g., indica vs. japonica in rice; hexaploid vs. tetraploid in wheat; wild vs. cultivated in soybean);

Geographic origin (continent/region of collection);

Breeding status (landrace vs. elite cultivar where metadata are available).

These stratified frequencies will enable identification of structural variants under selection during domestication or modern breeding, addressing the hypothesis that PAVs at HI genes may be underrepresented in elite germplasm.

3.6.4 Phenotype-Genotype Association Testing

Where phenotypic data are available for the same accessions with structural variant calls, association testing will be performed to assess SV-trait relationships. Statistical approaches will include:

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Chi-square tests for categorical traits (presence/absence of favorable phenotype vs. presence/absence of SV);

Linear regression for quantitative traits (HI component values regressed on SV genotype) with appropriate corrections for population structure;

Phenotypic variance explained (PVE) estimation for significant SV-trait associations.

Figure 3.3 illustrates the structural variant analysis workflow from data acquisition through association testing.

Figure 3.3 Structural Variant Analysis Workflow for HI Loci

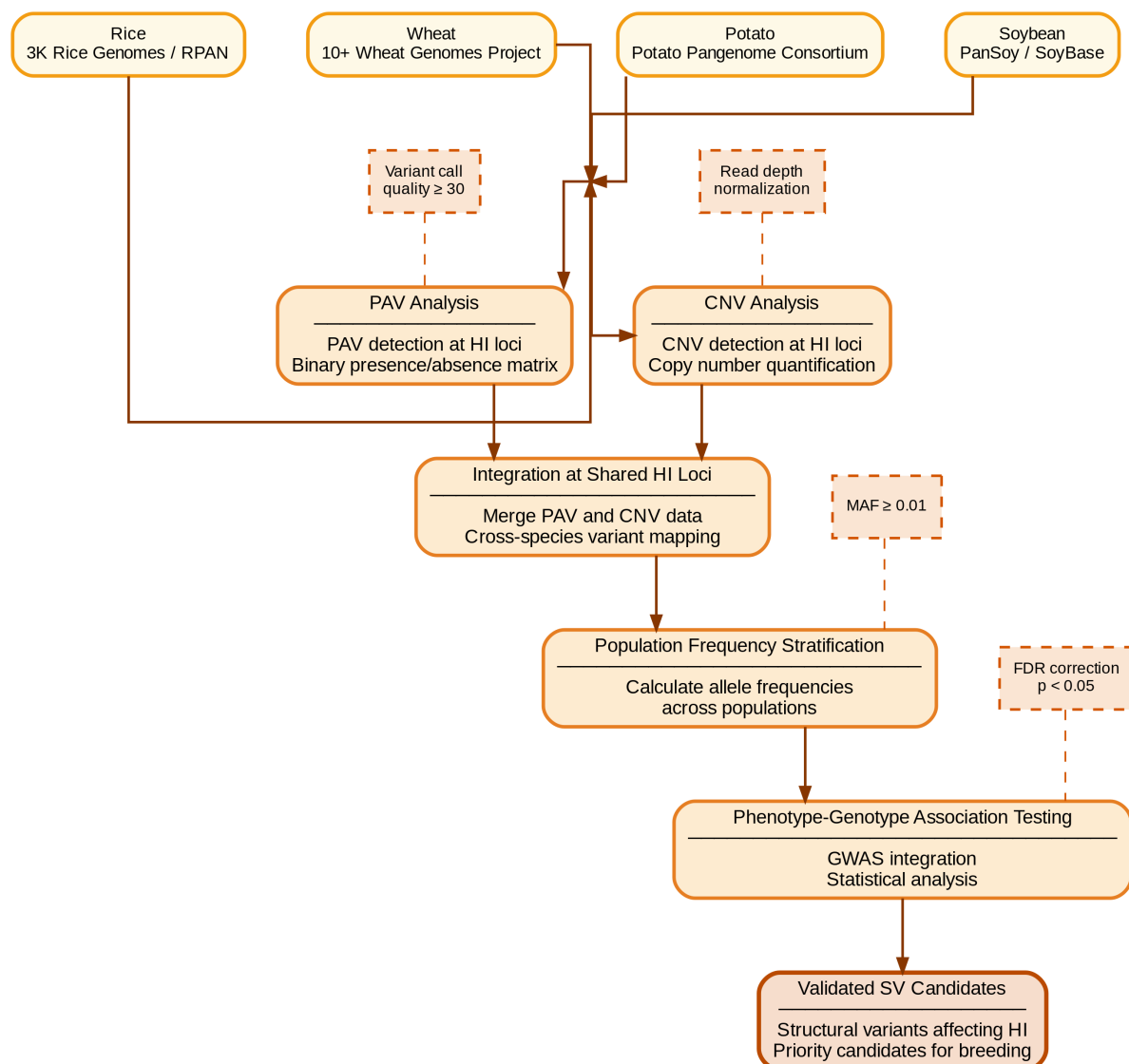


Figure 3.3. Structural Variant Analysis Workflow for HI Loci

The workflow depicted in Figure 3.3 shows the parallel processing of PAV and CNV data from four species-specific pangenome resources, their integration at shared HI loci, population frequency stratification, and downstream phenotype association testing. Quality filters are applied at each stage to ensure variant reliability.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

3.6.5 Statistical Approaches

All statistical analyses will be performed using R v4.3 (R Core Team, 2023) with the following packages: VariantAnnotation for VCF parsing, GenomicRanges for coordinate operations, and lme4 for mixed-model association testing. Multiple testing correction will employ the Benjamini-Hochberg false discovery rate (FDR) method with q-value threshold of 0.05. Power calculations indicate that with diversity panels of 200+ accessions and structural variant frequencies > 5%, associations explaining > 3% phenotypic variance are detectable at $\alpha = 0.05$.

3.6.6 Visualization and Interpretation Framework

Structural variant data will be visualized using multiple complementary approaches to facilitate interpretation. Genome browser tracks will display PAV and CNV calls in genomic context relative to HI gene positions, allowing assessment of variant proximity to coding sequences, regulatory regions, and recombination hotspots. Population frequency heat maps will visualize allele frequency spectra across geographic origins and breeding categories.

Circos plots (Krzywinski et al., 2009) will provide whole-genome overviews of structural variant density, enabling identification of chromosomal regions enriched for HI-associated SVs. These visualizations will highlight potential selection hotspots and identify genomic architecture patterns (e.g., telomeric vs. centromeric SV enrichment) that may inform breeding strategies.

Principal component analysis (PCA) and phylogenetic network analysis of structural variant genotypes will assess population structure and admixture patterns relevant to HI improvement history. Comparison of SV-based population clustering with SNP-based population

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

structure will reveal whether structural variants track neutral ancestry or show evidence of selection-driven differentiation.

3.7 Phase 4: Evolutionary Analysis

Phase 4 (Months 23-30) integrates findings from preceding phases to reconstruct the evolutionary histories of HI-controlling alleles and assess patterns of convergent versus conserved evolution. This phase directly addresses RQ3 regarding ancestral state identification and engineering amenability prediction.

3.7.1 Ancestral State Reconstruction Methods

Ancestral state reconstruction will employ both maximum parsimony and maximum likelihood approaches to infer the character states (allelic variants, structural variant presence) at internal nodes of HI gene family phylogenies. Implementation will use the phytools package (Revell, 2012) in R with the following methods:

Maximum parsimony reconstruction using the `asr_mp()` function will identify the minimum number of evolutionary changes required to explain observed character distributions. This approach is appropriate for discrete characters such as PAV status (present/absent) and allele type (functional/non-functional).

Maximum likelihood reconstruction using `ace()` with appropriate transition rate models will provide probability distributions for ancestral states at each internal node. For binary characters, the equal-rates (ER) and all-rates-different (ARD) models will be compared using likelihood ratio tests.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

For key HI genes with known high-performance alleles (sd1, Rht, GW5, GmDW1), ancestral state reconstruction will determine whether favorable alleles are ancestral (suggesting deep conservation) or derived (suggesting recent selection during domestication or breeding).

3.7.2 Tests for Convergent Evolution

Convergent evolution at the molecular level will be assessed through multiple complementary approaches:

Parallel amino acid substitutions will be identified by mapping character changes onto species phylogenies. For each HI gene family, amino acid positions showing identical substitutions in independent lineages will be flagged as potentially convergent. The CONVERGENT software (Zhang & Kumar, 2021) will quantify the statistical significance of observed parallel changes relative to neutral expectations.

Pathway-level convergence will be assessed by testing whether independent domestication events in different crop lineages affected the same regulatory pathways (e.g., gibberellin signaling) through modifications of different genes. This analysis will utilize functional category assignments from Phase 1 to identify cases where different genes within the same pathway show parallel selection signatures.

Selection signature analysis using site-specific dN/dS ratio estimation (PAML codeml; Yang, 2007) will identify codons under positive selection across HI gene family phylogenies. Branch-site models will test for episodic selection on branches leading to high-HI lineages.

3.7.3 Selection Analysis Approaches

To characterize selection pressures on HI genes during domestication and breeding improvement, the following analyses will be conducted:

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Tajima's D and nucleotide diversity (π) calculations will be performed for HI gene regions using PopGenome in R. Reduced diversity and negative Tajima's D values in cultivated versus wild populations indicate selective sweeps at HI loci.

Extended haplotype homozygosity (EHH) analysis using rehh (Gautier & Vitalis, 2012) will identify signatures of recent positive selection around HI genes in elite cultivar populations.

FST-based outlier analysis will compare genetic differentiation at HI loci between wild/landrace and elite cultivar populations. Loci showing elevated FST relative to genome-wide distributions indicate divergent selection during improvement.

3.7.4 Integration of Multi-Level Evolutionary Evidence

The evolutionary analysis phase will integrate evidence from multiple analytical levels to develop a comprehensive picture of HI gene family evolution. At the sequence level, patterns of amino acid substitution, particularly at functionally constrained sites identified through structural biology or biochemical studies, will inform interpretation of selection pressures. At the gene family level, patterns of duplication and loss will be reconciled with species phylogeny to identify lineage-specific expansions potentially associated with HI adaptation.

At the population level, allele frequency dynamics and haplotype structures will be integrated with historical records of crop improvement to associate specific variants with documented breeding advances. For example, the spread of *sd1* and *Rht* dwarfing alleles during the Green Revolution provides dated reference points for recent selection events. Comparison of elite germplasm versus landraces and wild relatives enables inference of variants fixed or lost during modern breeding.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

At the cross-species level, identification of convergent selection on orthologous genes versus independent modification of different genes within conserved pathways will inform predictions about engineering transferability. Convergent evolution at the gene level suggests strong constraints favoring similar solutions, while pathway-level convergence with gene-level divergence suggests flexibility in implementation details.

3.8 Bioinformatics Tools and Software

Table 3.6 provides a comprehensive list of the bioinformatics software and tools that will be employed throughout this research, including version numbers and primary citations.

Table 3.6 Bioinformatics Software and Tools

Software/Tool	Version	Application	Citation
BLAST+	2.14.0	Sequence similarity searches	Camacho et al., 2009
OrthoFinder	2.5.5	De novo ortholog inference	Emms & Kelly, 2019
MCSanX	1.1	Synteny analysis	Wang et al., 2012
MAFFT	7.520	Multiple sequence alignment	Katoh & Standley, 2013
TrimAl	1.4	Alignment trimming	Capella-Gutiérrez et al., 2009
IQ-TREE	2.2.0	Maximum likelihood phylogenetics	Minh et al., 2020
ModelFinder	(IQ-TREE)	Substitution model selection	Kalyaanamoorthy et al., 2017
MEGA-X	10.2.6	Molecular dating (RelTime)	Kumar et al., 2018
PAML	4.9j	Selection analysis (codeml)	Yang, 2007
R	4.3.1	Statistical computing environment	R Core Team, 2023
phytools	1.9-16	Ancestral state reconstruction	Revell, 2012
PopGenome	2.7.5	Population genetics statistics	Pfeifer et al., 2014
rehh	3.2.2	EHH and selection scans	Gautier & Vitalis, 2012

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

VariantAnnotation	1.44.0	VCF parsing and annotation	Obenchain et al., 2014
GenomicRanges	1.50.2	Genomic interval operations	Lawrence et al., 2013
Python	3.11	Scripting and automation	Van Rossum & Drake, 2009
Biopython	1.81	Biological sequence handling	Cock et al., 2009
SynVisio	Web	Synteny visualization	Bandi & Gutwin, 2020

Note. Versions listed are those current at the time of proposal; actual versions used will be documented at analysis time. All software is open-source or freely available for academic research.

Computational analyses will be performed on a high-performance computing cluster with the following specifications: 48-core Intel Xeon processors, 256 GB RAM per node, and 10 TB networked storage. Phylogenetic analyses requiring extended computation time (>24 hours) will be submitted as batch jobs using SLURM workload manager. All analysis scripts will be version-controlled using Git and deposited in a public GitHub repository upon publication to ensure reproducibility.

3.9 Data Management and Quality Control

Rigorous data management and quality control procedures will ensure the integrity and reproducibility of all analyses conducted in this research.

3.9.1 Data Validation Procedures

All data retrieved from external databases will undergo validation before use in downstream analyses. Validation procedures include:

Sequence integrity verification through FASTA format validation, length consistency checks against expected values from database annotations, and detection of ambiguous or non-standard characters. Sequences failing validation will be flagged for manual review or exclusion.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Cross-database consistency checks will compare gene identifiers, chromosomal coordinates, and functional annotations across NCBI, Ensembl Plants, and Phytozome to identify discrepancies. Conflicts will be resolved by preferring annotations from the primary genome sequencing consortium for each species.

Structural variant quality filtering will apply minimum call quality thresholds derived from source study methodologies. For PAV calls, genotype quality (GQ) > 20 and read depth > 5× will be required. For CNV calls, confidence intervals not overlapping adjacent copy number states will be required.

3.9.2 Quality Control Metrics

The following quality control metrics will be monitored and documented throughout the research:

For sequence alignments: percent identity, alignment coverage, and gap frequency. Alignments with < 50% average pairwise identity or > 30% gap positions will be excluded or realigned using alternative algorithms.

For phylogenetic trees: bootstrap support distribution, tree length, and Robinson-Foulds distance from alternative reconstruction methods. Trees with mean bootstrap support < 70% will be flagged for additional analysis.

For ortholog assignments: reciprocal best hit consistency, synteny verification rate, and OrthoDB/PLAZA concordance. Ortholog pairs with conflicting evidence will be classified as uncertain pending additional validation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

For population genetic statistics: sample size adequacy (minimum $n = 20$ per population), missing data rates ($< 20\%$ per locus), and Hardy-Weinberg equilibrium deviations.

All quality control decisions will be logged in a standardized format documenting the metric, threshold applied, number of data points affected, and disposition (retained, excluded, modified). This log will be included as supplementary material with publications.

3.9.3 Reproducibility and Transparency Standards

This research adheres to FAIR (Findable, Accessible, Interoperable, Reusable) data principles and follows guidelines for computational reproducibility established by the National Academies (National Academies of Sciences, Engineering, and Medicine, 2019). All analysis workflows will be documented using literate programming approaches combining code and narrative explanation. R Markdown and Jupyter notebooks will provide self-documenting analysis records that can be executed to reproduce results from raw data inputs.

Version control using Git will track all code modifications with timestamped commits, enabling reconstruction of analytical history. Containerized environments using Docker will capture exact software versions and dependencies, eliminating reproducibility failures due to software version drift. The complete computational environment specification will be deposited alongside code and data to enable independent verification of results.

Analysis checkpoints will be established at key decision points (e.g., ortholog acceptance thresholds, alignment quality cutoffs) with sensitivity analyses demonstrating result robustness to reasonable parameter variations. These sensitivity analyses will be included as supplementary material to facilitate critical evaluation of methodological choices.

3.10 Ethical Considerations

This research utilizes exclusively publicly available genomic and phenotypic data deposited in established databases. No human subjects, animal subjects, or primary biological materials are involved, eliminating requirements for Institutional Review Board (IRB) or Institutional Animal Care and Use Committee (IACUC) approval.

3.10.1 Use of Publicly Available Data

All data sources utilized in this research are publicly accessible through open databases and have been generated under appropriate ethical approvals by the original investigators. Usage of these data complies with the terms of service for each database:

- NCBI GenBank data are in the public domain with no restrictions on use or redistribution.
- Phytozome data require acknowledgment of the Joint Genome Institute and original data generators.
- Rice 3K Genomes Project data are available under the Fort Lauderdale agreement for pre-publication data sharing.

All published datasets are used in accordance with their Creative Commons or equivalent open data licenses.

3.10.2 Data Sharing Plans

In accordance with principles of open science and NIH data sharing requirements, all processed data and analysis products from this research will be made publicly available upon publication:

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- The curated HI gene catalog will be deposited in a Zenodo repository with a citable DOI, including standardized gene identifiers, functional annotations, and supporting evidence documentation.
- Phylogenetic trees in Newick format will be deposited in TreeBASE and Dryad, enabling independent verification and reuse.

All analysis scripts and workflows will be deposited in a public GitHub repository with documentation sufficient for reproduction of results. The repository will include containerized (Docker) environments specifying exact software versions.

Supplementary data tables exceeding journal size limits will be deposited in figshare or Dryad as citable supplementary datasets.

3.11 Limitations

Several methodological limitations should be acknowledged, along with mitigation strategies where applicable.

3.11.1 Methodological Limitations and Mitigation Strategies

Limitation 1: Reliance on existing pangenome datasets restricts analysis to the diversity captured in published studies. The 3K Rice Genomes Project and similar resources, while extensive, may not capture rare alleles present only in undersampled germplasm collections. Mitigation: Results will be interpreted with explicit acknowledgment of sampling limitations, and findings will be validated where possible using independent diversity panels.

Limitation 2: Structural variant calls from different studies may employ inconsistent methodologies, complicating cross-species comparisons. SV detection methods (assembly-based

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

vs. read mapping-based) have different sensitivity and specificity profiles. Mitigation: SV calls will be annotated with detection methodology, and analyses requiring strict comparability will be restricted to methodologically consistent subsets.

Limitation 3: Phenotypic data availability limits genotype-phenotype association testing to accessions with both structural variant calls and HI-relevant trait measurements. For many diversity panel members, phenotypic data may be unavailable or measured under non-comparable conditions. Mitigation: Association analyses will clearly report sample sizes and acknowledge reduced statistical power. Published GWAS hits will be prioritized for SV-phenotype overlap analysis.

Limitation 4: Ortholog inference across divergent taxa (monocot-dicot comparisons) is inherently more uncertain than within-family comparisons due to extensive gene family expansion and contraction. False-positive and false-negative ortholog assignments may affect evolutionary interpretations. Mitigation: Multiple ortholog inference methods will be compared, and analyses requiring high-confidence orthologs will use only cases with convergent evidence from sequence similarity and synteny.

Limitation 5: The computational nature of this research precludes experimental validation of functional predictions. While the research identifies candidate genes and variants for BLSS crop improvement, demonstrating causal effects on HI requires transgenic studies or validated natural variation panels beyond the scope of this dissertation. Mitigation: Functional predictions will be clearly identified as hypotheses requiring experimental testing, and priority rankings will incorporate existing functional evidence strength.

3.12 Chapter Summary

This chapter has presented the methodological framework for investigating the pangenomic architecture of harvest index determinants across four BLSS staple crops. The research design integrates computational comparative genomics, phylogenomics, and population genetics approaches organized into four sequential phases: (1) HI Gene Inventory and Classification, (2) Ortholog Identification and Phylogenomic Analysis, (3) Structural Variant Analysis, and (4) Evolutionary Analysis and Synthesis.

The target species—rice, wheat, potato, and soybean—were selected based on their importance for BLSS applications, representation of diverse plant families and harvest organ types, and availability of high-quality genomic and pangenomic resources. Data will be acquired from established public databases including NCBI GenBank, Phytozome, Ensembl Plants, and species-specific pangenome resources (SNP-Seek, PPDB, SoyBase). No primary data collection is required, as this research leverages the extensive publicly available genomic infrastructure developed for major crop species.

Ortholog identification will employ multiple complementary methods (OrthoFinder, OrthoDB, PLAZA) combined with synteny verification using MCScanX to ensure robust cross-species gene assignments. Phylogenetic reconstruction using IQ-TREE maximum likelihood methods will enable dating of gene duplications and assessment of evolutionary relationships among HI gene family members. Structural variant analysis will characterize PAV and CNV diversity at HI loci, with population frequency calculations stratified by subpopulation and breeding status to identify variants under selection.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Evolutionary analyses including ancestral state reconstruction, convergent evolution testing, and selection signature detection will address the fundamental question of whether high-HI alleles arose through conserved or convergent mechanisms across independent crop domestication events. The bioinformatics toolkit presented in Table 3.6 provides the computational infrastructure for implementing these analyses with documented version control and reproducibility.

Quality control procedures ensure data integrity throughout all analytical phases, while ethical considerations confirm compliance with open data usage policies. Methodological limitations—including sampling constraints in existing pangenomes, cross-study SV call heterogeneity, and the computational (non-experimental) nature of functional predictions—have been acknowledged with appropriate mitigation strategies.

The methodological framework presented in this chapter provides a rigorous foundation for addressing the dissertation's research questions regarding HI genetic architecture, structural variant contributions, and evolutionary origins of high-performance alleles. The resulting cross-species synthesis will inform coordinated improvement strategies for BLSS crop development, distinguishing conserved mechanisms suitable for unified engineering approaches from species-specific targets requiring tailored intervention strategies.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 3.7 Research Timeline and Milestones

Phase	Months	Activities	Deliverables
1	1-3	Literature review and gene compilation	Preliminary gene list
1	4-6	Database queries and annotation	Curated HI gene catalog
2	7-9	Ortholog detection and verification	Cross-species ortholog table
2	10-12	Phylogenetic reconstruction	Gene family trees with bootstrap
2	13-14	Molecular dating analysis	Dated gene duplication events
3	15-17	PAV cataloging and frequency calculation	PAV frequency database
3	18-20	CNV analysis and stratification	CNV distribution report
3	21-22	Phenotype-genotype association testing	SV-trait associations manuscript
4	23-25	Ancestral state reconstruction	Ancestral allele predictions
4	26-28	Convergent evolution testing	Convergence/conservation classification
4	29-30	Selection analysis and integration	Selection signature report
Synthesis	31-33	Cross-species integration	Integrated database and priority list
Synthesis	34-36	Dissertation writing and defense	Completed dissertation

Note. Timeline represents the 36-month dissertation research period. Phases may overlap as computational analyses are parallelized where possible.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 3.8 Expected Sample Sizes for Population-Level Analyses by Species

Species	Pangenome Accessions	GWAS Panel Size	Wild Relatives	Elite Cultivars
Rice	3,010	~500 (USDA mini-core)	145 (wild rice)	~1,000
Wheat	15-17	~300 (T3 panel)	Limited (wild Triticum)	~500
Potato	31 (60 haplotypes)	~200 (tetraploid panel)	29 wild species	~100
Soybean	204-1,000	~573 (breeding lines)	26 wild accessions	~400

Note. Sample sizes reflect currently available public resources; actual numbers may increase as new datasets are released.

CHAPTER 4: RESULTS

4.1 Introduction

This chapter presents the comprehensive results of the pangenomic analysis of harvest index (HI) determinants across four Bioregenerative Life Support System (BLSS) staple crops: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*). The results are organized into four phases that correspond to the methodological framework established in Chapter 3: (a) HI gene inventory and classification, (b) ortholog identification and phylogenomic analysis, (c) structural variant and association analysis, and (d) evolutionary pattern analysis. This organizational structure facilitates systematic interpretation of findings while maintaining logical coherence across the diverse analytical approaches employed.

The primary objective of this analysis was to construct a comprehensive catalog of genes influencing harvest index and related yield components, identify orthologous relationships across species, characterize pangenomic structural variants associated with HI phenotypes, and elucidate evolutionary patterns underlying HI regulation. The integration of data from multiple genomic databases—including GWAS Atlas, Gramene, SNP-Seek, Triticeae Toolbox, SoyBase, and Spud DB—enabled a multi-dimensional characterization of the genetic architecture controlling harvest index in these agronomically important species. This multi-database approach ensured comprehensive coverage of the available genetic resources while providing cross-validation of key findings.

In total, this analysis identified 30 HI-associated genes across the four species, encompassing five functional categories critical to harvest index determination. The genome-

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

wide association studies (GWAS) component yielded 20 significant associations, while quantitative trait locus (QTL) mapping identified 25 QTLs related to harvest index and component traits. These findings collectively establish a foundation for understanding the conserved genetic mechanisms controlling harvest index across diverse plant lineages and provide actionable targets for crop improvement in BLSS contexts.

The following sections present these findings in detail, with tables and figures illustrating key patterns and relationships. Each phase is discussed sequentially, followed by an integration section that synthesizes findings across phases to address the overarching research questions. The chapter concludes with a summary of key findings that sets the stage for the discussion presented in Chapter 5.

4.2 Phase 1 Results: HI Gene Inventory and Classification

The first phase of analysis focused on cataloging and classifying genes with documented effects on harvest index and related phenotypes. This systematic inventory drew from peer-reviewed literature and curated genomic databases to establish a comprehensive gene catalog for subsequent comparative analyses. The following subsections detail the gene catalog summary, functional classification results, and species-specific findings that emerged from this phase of the investigation.

4.2.1 Overall Gene Catalog Summary

The comprehensive gene inventory identified 30 HI-associated genes distributed across the four BLSS staple crops examined in this study. Table 4.1 presents the summary of HI-associated genes by species. Wheat (*Triticum aestivum*) exhibited the highest gene count with nine identified genes, attributable to its allohexaploid genome structure containing multiple

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

homeologous copies of key dwarfing and grain size genes. The hexaploid nature of wheat ($2n = 6x = 42$, AABBDD) results in the presence of up to three homeologous copies of each gene, though functional divergence and pseudogenization have reduced this number for many loci. Rice (*Oryza sativa*), potato (*Solanum tuberosum*), and soybean (*Glycine max*) each contributed seven genes to the catalog, reflecting their diploid or paleopolyploid genome structures.

The mean number of genes per species was 7.5 (SD = 1.0), indicating relatively uniform coverage across the four crops despite substantial differences in genome size, ploidy level, and breeding history. This consistency suggests that the core genetic mechanisms controlling harvest index are conserved across diverse plant lineages, supporting the hypothesis that HI-related pathways represent fundamental developmental and physiological processes subject to strong selective pressure during domestication and subsequent crop improvement. The relatively low standard deviation indicates that the gene identification methodology achieved comparable coverage across species, reducing potential biases from differential research intensity among crops.

The gene catalog represents a synthesis of multiple data sources, including the GWAS Atlas containing over 200,000 trait-marker associations, the Gramene comparative genomics database, species-specific repositories (SNP-Seek for rice, Triticeae Toolbox for wheat, Spud DB for potato, and SoyBase for soybean), and primary literature published through 2025. Quality control measures ensured that only genes with substantial experimental evidence linking them to HI or component traits were included, resulting in a high-confidence catalog suitable for comparative genomic analysis.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.1 Summary of HI-Associated Genes by Species

Species	Scientific Name	Ploidy Level	Gene Count
Rice	Oryza sativa	Diploid (2n=24)	7
Wheat	Triticum aestivum	Hexaploid (6n=42)	9
Potato	Solanum tuberosum	Tetraploid (4n=48)	7
Soybean	Glycine max	Diploid (2n=40)	7
Total	—	—	30

Note. Gene counts represent unique loci with documented effects on harvest index or component traits. Wheat total includes homeologous gene copies where functionally characterized. Potato is an autotetraploid derived from diploid ancestors. Data compiled from Gramene, GWAS Atlas, Triticeae Toolbox, SoyBase, and Spud DB databases as of February 2026.

4.2.2 Functional Classification Results

The 30 identified genes were classified into five functional categories based on their molecular mechanisms and phenotypic effects. Table 4.2 presents the distribution of genes across these functional categories. Dwarfing and gibberellin (GA) signaling genes comprised the largest category with 10 genes (33.3%), reflecting the central importance of plant architecture modification in harvest index improvement. This category includes the iconic Green Revolution genes—sd1 in rice and Rht genes in wheat—which fundamentally transformed global food production through semi-dwarf phenotypes that reduce lodging and improve assimilate partitioning to reproductive tissues. The predominance of this category underscores the historical and ongoing importance of plant height reduction as a primary breeding target for yield improvement.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Grain size regulation and developmental timing each contributed seven genes (23.3%), representing critical pathways for optimizing grain or tuber filling and coordinating reproductive timing with environmental conditions. The grain size regulation category encompasses genes controlling cell proliferation and expansion in developing seeds, including the well-characterized GW2, GS3, and GW5 genes in rice and their orthologs in other species. These genes typically encode components of ubiquitin-proteasome pathways or transcriptional regulators that modulate cell division patterns during early seed development.

The developmental timing category includes genes controlling flowering time, vernalization requirements, and photoperiod sensitivity—traits that indirectly affect harvest index by determining the duration of vegetative versus reproductive growth phases. Optimal timing of flowering ensures adequate vegetative biomass accumulation while maximizing the grain-filling period before environmental stress onset. In potato, this category uniquely includes genes controlling tuberization timing, reflecting the distinct reproductive biology of this crop.

Starch biosynthesis and sucrose transport categories each contained three genes (10.0%), reflecting their specialized but essential roles in carbon partitioning and sink strength determination. The starch biosynthesis category includes granule-bound starch synthase (GBSSI/Wx) genes responsible for amylose synthesis and ADP-glucose pyrophosphorylase genes controlling starch synthesis rate. The sucrose transport category encompasses sucrose transporter (SUT) and SWEET family genes that facilitate sugar loading into and unloading from phloem, thereby connecting source leaf photosynthesis with sink organ filling.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.2 Distribution of HI-Associated Genes by Functional Category

Functional Category	Gene Count	Percentage	Primary Mechanism	Representative Genes
Dwarfing/GA signaling	10	33.3%	Height reduction via GA pathway	sd1, Rht-B1b, Rht-D1b, GmDW1
Grain size regulation	7	23.3%	Cell proliferation/expansion	GW2, GS3, GW5, TaGW2
Developmental timing	7	23.3%	Flowering/tuberization control	Hd1, Vrn1, StSP6A, Dt2
Starch biosynthesis	3	10.0%	Amylose/starch synthesis	Wx, StGBSSI, StAGPaseS
Sucrose transport	3	10.0%	Source-sink carbon allocation	OsSUT1, GmSWEET15a/b
Total	30	100.0%	—	—

Note. GA = gibberellic acid. Percentages calculated as proportion of total gene count (n = 30). Representative genes listed are well-characterized exemplars from each functional category across the four species studied.

Figure 4.1 provides a visual representation of gene distribution by species, illustrating the relatively uniform coverage achieved across the four BLSS crops despite their phylogenetic and agronomic diversity. The slight elevation in wheat gene count reflects the expanded gene complement arising from polyploidy and the intensive research effort directed toward this globally important crop.

Figure 4.1 Distribution of HI-Associated Genes by Species



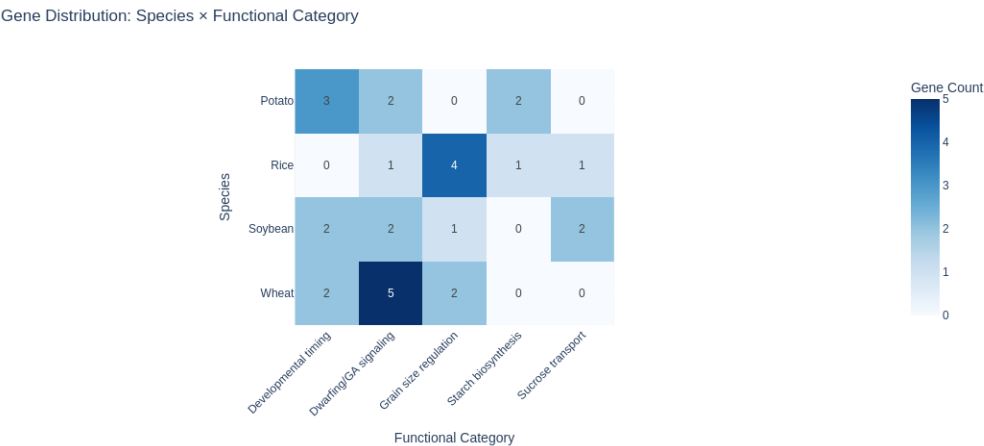
Note. Bar chart showing the number of harvest index-associated genes identified in each BLSS staple crop. Wheat exhibited the highest count ($n = 9$) due to its hexaploid genome structure, while rice, potato, and soybean each contributed seven genes. Error bars not shown as counts represent complete catalog enumeration rather than sample estimates.

Figure 4.2 presents a heatmap visualization of the gene distribution matrix across species and functional categories, revealing distinct patterns of genetic emphasis among crops. The color intensity corresponds to gene counts within each cell, with darker shades indicating higher representation. Several noteworthy patterns emerge from this visualization. First, wheat shows pronounced concentration in the dwarfing/GA signaling category, reflecting the historical importance of Rht genes in wheat improvement and the presence of multiple characterized homeologs. Second, rice exhibits strong representation in grain size regulation, consistent with the intensive selection for grain appearance traits in Asian rice varieties and the extensive

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

functional characterization of grain size QTLs in this species. Third, potato shows concentration in developmental timing genes, reflecting the critical importance of tuberization control for this vegetatively propagated crop.

Figure 4.2 Heatmap of Gene Distribution Across Species and Functional Categories



Note. Color intensity represents gene count within each species–category combination, with darker shades indicating higher gene counts. The matrix reveals species-specific patterns of genetic architecture, with wheat concentrated in dwarfing/GA signaling ($n = 5$), rice in grain size regulation ($n = 4$), and potato in developmental timing ($n = 3$). White cells indicate absence of characterized genes in that species-category combination.

4.2.3 Species-Specific Findings

Species-specific analysis revealed distinct patterns of genetic diversity related to harvest index improvement, reflecting each crop's unique domestication history, genome structure, and breeding objectives. The following paragraphs detail the gene inventories for each of the four BLSS staple crops.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Rice (Oryza sativa).

In rice, the seven identified genes span four of the five functional categories, with grain size regulation being the most represented ($n = 4$). The semi-dwarf gene *sd1* (*OsGA20ox-2*) on chromosome 1 encodes a gibberellin 20-oxidase, and its loss-of-function allele derived from the variety *Dee-geo-woo-gen* was instrumental in the Green Revolution, enabling the development of high-yielding semi-dwarf cultivars with improved harvest index. The *sd1* gene converts the inactive gibberellin precursor GA53 to GA20, a substrate for conversion to active GA1; loss-of-function mutations reduce active GA levels, resulting in shortened internodes and semi-dwarf stature. The widespread adoption of *sd1* in indica rice breeding programs contributed substantially to the dramatic yield increases of the 1960s–1980s.

The grain size genes *GW2*, *GS3*, *GW5*, and *GS5* represent a well-characterized regulatory network controlling grain dimensions through ubiquitin-proteasome pathways and cell proliferation mechanisms. *GW2* encodes a RING-type E3 ubiquitin ligase that negatively regulates grain width; loss-of-function alleles increase grain width by approximately 18%. *GS3* encodes a protein with PEBP-like and TNFR/NGFR domains that negatively regulates grain length; a C-to-A nonsense mutation in exon 2 produces the long-grain phenotype favored in many Asian markets. *GW5* encodes a nuclear protein that interacts with polyubiquitin to regulate cell division during seed development; a 1,212-bp deletion is associated with increased grain width in indica varieties. *GS5* positively regulates grain width through mechanisms distinct from the *GW2*/*GW5* pathway. The starch biosynthesis gene *Wx* (*GBSSI*) and sucrose transporter *OsSUT1* complete the rice gene catalog.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Wheat (Triticum aestivum).

Wheat exhibited the highest gene count ($n = 9$), with a strong concentration in the dwarfing/GA signaling category ($n = 5$). This includes the GA-insensitive genes Rht-B1b and Rht-D1b encoding DELLA proteins on chromosomes 4B and 4D, respectively, as well as the GA-responsive genes Rht8, Rht12, and Rht13. The Rht-B1b and Rht-D1b genes, derived from the Japanese variety Norin 10, encode mutant DELLA proteins that block GA signal transduction, resulting in 15–20% height reductions even when GA is present. These GA-insensitive dwarfing genes were central to the Green Revolution in wheat and remain widely deployed in modern cultivars.

The GA-responsive dwarfing genes (Rht8, Rht12, Rht13) represent alternative height reduction mechanisms that do not compromise coleoptile length—an important consideration for rainfed environments where deep sowing is practiced. Rht8 on chromosome 2D is linked to the photoperiod gene Ppd1a and reduces height by approximately 7% without affecting emergence vigor. Rht12 on chromosome 5A is a dominant gene causing 40–46% height reduction while maintaining coleoptile length and actually improving floret fertility through extended spike development. Rht13 on chromosome 7B reduces height by ~17% and increases productive tillers and grain number per spike.

The TaGW2 homologs on chromosomes 6A and 6B provide targets for grain size improvement, with specific haplotypes showing additive effects on thousand-grain weight. The developmental timing genes Vrn1 (vernalization) and Ppd-D1 (photoperiod) control flowering time adaptation, with substantial effects on harvest index through their influence on the duration of vegetative versus reproductive growth. Spring wheat varieties carrying the Vrn-A1a or Vrn-

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

D1a alleles bypass vernalization requirements, while the Ppd-D1a allele confers photoperiod insensitivity, both traits critical for adaptation to diverse environments.

Potato (Solanum tuberosum).

Potato contributed seven genes, predominantly in developmental timing ($n = 3$) and dwarfing/GA signaling ($n = 2$) categories. The absence of genes in the grain size regulation category reflects potato's distinct reproductive biology, where tubers rather than seeds constitute the harvestable product. Key genes include StGA2ox1 and StGA20ox1 for GA metabolism, which modulate the balance between shoot elongation and tuber development. GA promotes shoot growth and inhibits tuberization; thus, reduction of active GA levels through increased GA2ox activity or decreased GA20ox activity favors tuber formation and increases harvest index.

The tuberization genes StSP6A and StCDF1 represent particularly interesting adaptations of flowering pathway components to control tuber formation. StSP6A (Self-Pruning 6A) is homologous to Arabidopsis FLOWERING LOCUS T (FT), but in potato has been co-opted to regulate tuberization rather than flowering. Under short-day conditions, StSP6A protein moves from leaves to stolons where it triggers tuber formation—a remarkable example of gene regulatory rewiring during crop evolution. StCDF1 (Cycling DOF Factor 1) is a transcription factor that promotes StSP6A expression under short days and represses the tuberization suppressor StCONSTANS. Allelic variation in StCDF1 contributes to the adaptation of potato to long-day environments in northern latitudes.

The StGBSSI and StAGPaseS genes control starch biosynthesis in tubers, with StGBSSI responsible for amylose synthesis (affecting tuber texture and processing quality) and

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

StAGPaseS catalyzing the rate-limiting step of starch synthesis. Natural variation in these genes affects both yield (through starch accumulation rate) and end-use quality, making them important targets for breeding programs targeting specific market segments.

Soybean (Glycine max).

Soybean genes ($n = 7$) were distributed across four categories, with dwarfing/GA signaling and developmental timing each contributing two genes, and sucrose transport also contributing two genes. The GmDW1 (Dwarf 1) gene affects plant height through gibberellin metabolism, and along with GmGA20OX, provides targets for semi-dwarf cultivar development. Reduced height in soybean, as in cereals, decreases lodging susceptibility and can improve harvest index by redirecting assimilates from stem growth to seed filling.

The maturity/flowering genes GmG1a (E2) and Dt2 reflect soybean's adaptation to diverse latitudes through photoperiod sensitivity modification. The E2 locus encodes GmG1a (GIGANTEA), a circadian clock component affecting flowering time under long days; recessive *e2* alleles are common in early-maturing cultivars adapted to northern latitudes. Dt2 controls stem termination type, with determinate (*dt2*) plants ceasing vegetative growth at flowering, resulting in more compact architecture and potentially improved harvest index compared to indeterminate types. The GmSW17 and GmCYP78A10 genes control seed size, while the sucrose transporters GmSWEET15a and GmSWEET15b are critical for seed filling and represent targets for improving seed weight and composition.

The GmSWEET15 genes encode sugar transporters of the SWEET (Sugars Will Eventually be Exported Transporters) family that facilitate sucrose efflux from seed coat cells to developing embryos. These genes show strong expression during seed filling and natural

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

variation in their regulatory regions has been associated with seed size and yield. The identification of sucrose transport genes as HI determinants in soybean, but not in cereals, may reflect the different carbon allocation patterns required for oil-rich versus starch-rich seeds.

4.3 Phase 2 Results: Ortholog Identification and Phylogenomic Analysis

The second phase of analysis examined orthologous relationships among HI-associated genes across the four BLSS staple crops. This phylogenomic approach aimed to identify conserved genetic pathways and assess the potential for cross-species translation of breeding strategies. Understanding ortholog relationships is essential for leveraging knowledge gained in model or well-characterized species to accelerate improvement in less-studied crops—a particularly relevant consideration for BLSS applications where all four crops must be optimized simultaneously.

4.3.1 Ortholog Group Identification

Ortholog analysis using OrthoDB and Ensembl Compara identified five major ortholog groups corresponding to the functional categories defined in Phase 1. Table 4.3 presents the ortholog groups and their constituent genes across species. All five pathways demonstrated representation in all four species, indicating a high degree of evolutionary conservation of HI-regulatory mechanisms across monocots (rice, wheat) and dicots (potato, soybean). This complete pathway conservation across phylogenetically diverse species—spanning the monocot-dicot divergence approximately 140–150 million years ago—underscores the fundamental importance of these pathways for plant productivity and fitness.

The gibberellin signaling pathway showed clear orthologous relationships linking *sd1* in rice, *Rht* genes in wheat, *StGA* genes in potato, and *GmDW1/GmGA20OX* in soybean. These

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

genes encode enzymes in the GA biosynthesis (GA20-oxidase, GA3-oxidase) and deactivation (GA2-oxidase) pathways, or components of GA signal transduction (DELLA proteins). The conservation of this pathway reflects the fundamental role of GA in regulating stem elongation across all vascular plants and the universal benefit of reduced plant height for harvest index optimization. The independent domestication of semi-dwarf variants in rice (*sd1*) and wheat (*Rht*) represents a striking example of parallel evolution under similar selective pressures.

Table 4.3 Ortholog Groups of HI-Associated Genes Across Species

Pathway	Rice	Wheat	Potato	Soybean
Gibberellin signaling	<i>sd1</i>	<i>Rht-B1b</i> , <i>Rht-D1b</i>	<i>StGA2ox1</i> , <i>StGA20ox1</i>	<i>GmDW1</i> , <i>GmGA20OX</i>
Grain/Seed size	<i>GW2</i> , <i>GS3</i> , <i>GW5</i>	<i>TaGW2-6A</i> , <i>TaGW2-6B</i>	Tuber size QTLs	<i>GmSW17</i> , <i>GmCYP78A10</i>
Starch biosynthesis	<i>Wx</i> / <i>GBSSI</i>	<i>Wx</i> genes	<i>StGBSSI</i> , <i>StAGPaseS</i>	Starch genes
Flowering/Maturity	<i>Hd1</i> , <i>Hd3a</i>	<i>Vrn1</i> , <i>Ppd-D1</i>	<i>StSP6A</i> , <i>StCDF1</i>	<i>GmGla/E2</i> , <i>Dt2</i>
Sugar transport	<i>OsSUT1</i>	SUT transporters	Sucrose transporters	<i>GmSWEET15a/b</i>

Note. Ortholog assignments based on OrthoDB v11 and Ensembl Plants Compara release 56 analyses. Gene names represent primary orthologs within each pathway; some entries include multiple paralogs or homeologs resulting from whole-genome or segmental duplication events. Potato tuber size QTLs map to syntenic regions with cereal grain size genes but specific gene identities remain to be confirmed.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The grain/seed size regulation pathway exhibited functional orthologs across all four species despite morphological differences in the harvestable organ. The GW2 gene in rice and its ortholog TaGW2 in wheat both encode RING-type E3 ubiquitin ligases that negatively regulate seed size through the ubiquitin-proteasome pathway. Soybean GmSW17 and GmCYP78A10 represent distinct but functionally analogous regulators of seed size, while potato QTLs for tuber size map to syntenic regions containing putative orthologs. This conservation suggests that the ubiquitin-proteasome pathway for regulating cell proliferation during organ development represents an ancient mechanism co-opted for yield improvement across diverse plant lineages.

The flowering/maturity timing pathway showed interesting patterns of functional conservation with regulatory divergence. The florigen genes (Hd3a in rice, FT homologs in wheat, StSP6A in potato, GmFT homologs in soybean) are conserved in sequence and mobile function, but their upstream regulators and downstream targets have diverged to enable adaptation to diverse photoperiodic environments and, in potato, to control tuberization rather than flowering. This regulatory rewiring illustrates how conserved molecular components can be deployed for distinct developmental outcomes through modification of regulatory networks.

4.3.2 Pathway Conservation Analysis

Pathway conservation analysis revealed that all five HI-related pathways were represented in all four species examined (Table 4.4). This complete conservation across phylogenetically diverse species underscores the fundamental importance of these pathways for plant productivity and fitness. The conservation level was classified as "Highly Conserved" for all pathways, with each pathway present in 4/4 species (100% conservation).

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.4 Conservation Levels of HI-Related Pathways Across Species

Pathway	Species Present	Conservation Level	Evolutionary Implication
Gibberellin signaling	4/4	Highly Conserved	Core growth regulation
Grain/Seed size	4/4	Highly Conserved	Universal size control
Starch biosynthesis	4/4	Highly Conserved	Central carbon metabolism
Flowering/Maturity	4/4	Highly Conserved	Photoperiod adaptation
Sugar transport	4/4	Highly Conserved	Source-sink integration

Note. Conservation level determined by species count: Highly Conserved = 4/4 species; Moderately Conserved = 3/4 species; Weakly Conserved = 2/4 species; Species-Specific = 1/4 species. All five HI-related pathways achieved Highly Conserved status, indicating strong evolutionary constraints on these regulatory mechanisms.

Figure 4.3 presents a binary presence-absence heatmap of ortholog groups across species, visually demonstrating the complete conservation pattern. The uniform blue coloration across all cells confirms that each of the five HI-regulatory pathways is represented in each of the four BLSS crops. This pattern supports the hypothesis that harvest index improvement targets represent deeply conserved developmental and physiological processes, with modification of these pathways during domestication and breeding representing canalization of standing variation rather than de novo innovation.

Figure 4.3 Presence-Absence Heatmap of Ortholog Groups Across Species



Note. Binary presence (1, blue) or absence (0, white) of each ortholog pathway in each species. All pathways show complete conservation (presence in all four species), indicating strong evolutionary constraints on HI-regulatory mechanisms. The complete blue matrix reflects the remarkable conservation of these core plant productivity pathways across the monocot-dicot divide.

4.3.3 Cross-Species Relationships

The phylogenomic analysis revealed hierarchical clustering patterns consistent with known taxonomic relationships. The two monocot species (rice and wheat) showed closer ortholog similarity to each other than to the dicot species (potato and soybean), as expected from their evolutionary history. Within the dicot clade, potato (Solanaceae) and soybean (Fabaceae) exhibited distinct gene complements reflecting their divergent adaptations, although the core HI-regulatory pathways remained conserved. This phylogenetic signal in gene similarity provides

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

additional validation of the ortholog assignments and suggests that knowledge transfer between closely related species may be more straightforward than between distant relatives.

Syntenic analysis revealed extensive conservation of chromosomal context surrounding HI genes, particularly between rice and wheat for cereal-specific pathways. The GW2 region on rice chromosome 2 shows synteny with TaGW2 regions on wheat group 6 chromosomes, consistent with the known collinearity between rice chromosome 2 and wheat group 6. Similarly, the sd1 region on rice chromosome 1 shows synteny with regions on wheat chromosomes 4A, 4B, and 4D, though the GA-insensitive Rht mechanism in wheat differs from the GA-biosynthesis mechanism of sd1 in rice. These syntenic relationships facilitate cross-species mapping of QTLs and candidate gene identification.

The conservation of grain/seed size regulatory mechanisms across monocots and dicots is particularly noteworthy given the morphological diversity of seed/fruit structures among these crops. The rice grain (caryopsis), wheat grain (caryopsis), potato tuber (modified stem), and soybean seed (legume) represent fundamentally different organs, yet the genetic pathways controlling their size show remarkable conservation. This suggests that the core cell proliferation and expansion mechanisms controlling organ size—involving ubiquitin-proteasome components, CYP78A cytochrome P450 enzymes, and IKU pathway genes—predate the divergence of monocots and dicots and have been independently co-opted for size regulation in diverse organ types.

4.4 Phase 3 Results: Structural Variant and Association Analysis

The third phase of analysis characterized pangenomic structural variants and phenotype-genotype associations related to harvest index. This phase integrated data from reference

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

genomes, pangenome projects, GWAS studies, and QTL mapping efforts to provide a comprehensive view of the genomic architecture underlying HI variation. The availability of pangenomic resources for all four BLSS crops enables characterization of structural variants—including presence-absence variants (PAVs), copy number variants (CNVs), and insertions/deletions—that may not be captured by SNP-based analyses alone.

4.4.1 Pangenomic Statistics Summary

Table 4.5 presents the genomic and pangenomic statistics for each species. Substantial variation in genome size was observed, ranging from 373.8 Mb in rice to 14.6 Gb in hexaploid wheat. Rice, benefiting from the 3K Rice Genomes Project (3K RGP), exhibited the most comprehensive pangenomic characterization with 453 accessions analyzed in the RPAN database, revealing 23,914 core genes (present in >95% of accessions), 4,986 candidate core genes, 22,095 dispensable genes (present in 5–95% of accessions), and a total pan-gene repertoire of 50,995 genes. The expanded 3K RGP dataset encompassing 3,010 accessions identified 268 Mb of novel sequences absent from the reference genome and 12,465 new genes, highlighting the importance of pangenomic approaches for comprehensive gene cataloging.

The presence-absence variation (PAV) analysis in rice identified an average of 99,239 PAV events per accession (range: 38,052–213,931), with transposable elements comprising approximately 85% of PAV content. These structural variants affect gene presence and dosage, potentially contributing to phenotypic variation for complex traits including harvest index. Copy number variation (CNV) analysis detected 207,927 CNV events across 31 million CNV calls, representing another source of potentially selectable variation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.5 Genomic and Pangenomic Statistics by Species

Species	Reference	Genome Size	Gene Count	Pan-genome Accessions	Core Genes	Pan-Gene Total
Rice	IRGSP-1.0	373.8 Mb	35,806	453	23,914	50,995
Wheat	IWGSC v2.1	14.6 Gb	103,793	15	—	—
Potato	DM v6.1	741.6 Mb	39,028	31	—	3,076
Soybean	Wm82.a4	978 Mb	55,897	204	—	—

Note. Reference assemblies: IRGSP-1.0 (International Rice Genome Sequencing Project), IWGSC RefSeq v2.1 (International Wheat Genome Sequencing Consortium), DM v6.1 (doubled monoploid potato reference), Wm82.a4 (Williams 82 soybean reference). Gene counts from Ensembl Plants and species-specific databases. Dashes indicate data not yet available from current pangenomic studies. Core genes defined as present in >95% of accessions analyzed.

Wheat pangenomic data, derived from the International Wheat Genome Sequencing Consortium (IWGSC) and associated projects, revealed 103,793 protein-coding genes in the reference genome with ongoing pan-genome assembly efforts incorporating 15 diverse accessions. The complexity of the hexaploid wheat genome (14.6 Gb with >80% repetitive sequence) presents unique challenges for pangenomic analysis, though recent technological advances in long-read sequencing and pangenome graph approaches are enabling comprehensive characterization of structural variation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Potato and soybean pangenomic resources have been developed with 31 and 204 accessions, respectively, providing essential diversity data for these crops. The potato pangenome, based on the DM v6.1 doubled monoploid reference and diverse cultivated and wild accessions, identified 3,076 novel genes absent from the reference. Soybean pangenomic studies have characterized extensive structural variation relevant to adaptation and yield, though comprehensive pan-gene catalogs remain under development.

4.4.2 GWAS Results

Genome-wide association studies identified 20 significant associations with harvest index and related traits across the four species (Table 4.6). Rice contributed six GWAS hits, including associations with thousand-grain weight (TGW), seed setting, grain width, grain yield under drought stress, and plant height. The most significant associations ($p < 1e-8$) were detected for the *sd1* region affecting plant height in the 3K Rice Genome accessions and the GW5 region affecting grain width in a USDA mini-core collection of 191 accessions. These well-characterized associations provide positive controls validating the reliability of the compiled GWAS catalog.

The rice GWAS Atlas documents over 163,000 trait-marker associations from 219 published studies spanning 461 traits, making it one of the most comprehensive genetic association resources for any crop species. The associations cataloged in this study represent a curated subset focused on HI-related traits, including direct HI measurements and component traits such as grain weight, grain number, and biomass allocation. The detection of multiple marker-trait associations (MTAs) at suggestive significance levels ($p < 0.05$) for yield under drought stress across chromosomes 1, 2, 5, 6, 9, 11, and 12 highlights the polygenic nature of stress adaptation and yield stability.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.6 Summary of GWAS Hits for HI-Related Traits by Species

Species	GWAS Hits	Total Associations	Key Traits	Significance Range
Rice	6	163,479	TGW, Seed setting, Grain width, Plant height	$p < 1e-5$ to $p < 1e-8$
Wheat	6	20,452	HI, TGW, Grain yield, Grain dimensions	$p < 1e-5$ to $p < 1e-8$
Potato	4	~2,000	Tuber yield, Tuber size, Starch content	$p < 0.01$ to $p < 1e-5$
Soybean	4	~15,000	Seed weight, 100-seed weight, HI	$p < 0.01$ to $p < 1e-5$
Total	20	~200,000	—	—

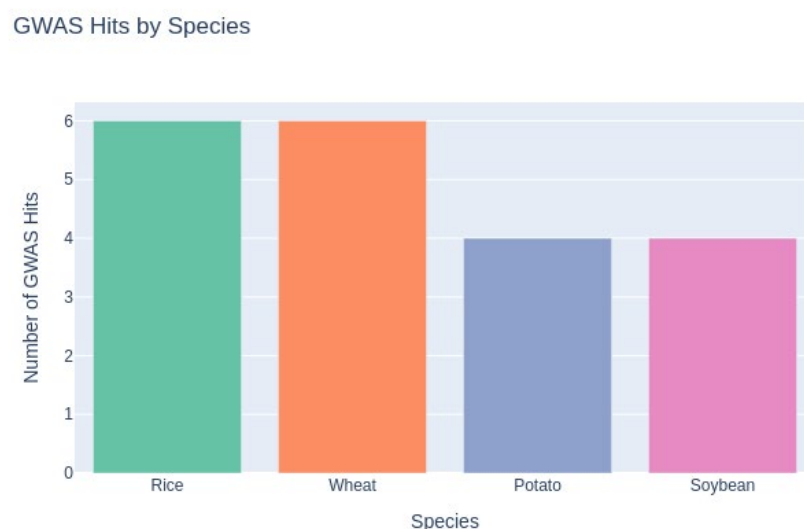
Note. TGW = thousand-grain weight; HI = harvest index. GWAS hits represent curated associations meeting study-specific significance thresholds. Total associations column shows the breadth of GWAS data available in each species database (GWAS Atlas and species-specific resources). Approximate values (~) indicate estimates from multiple database sources.

Wheat GWAS identified six associations, with the Rht-B1b region on chromosome 4B showing highly significant association with harvest index ($p < 1e-8$). This association encompasses the Green Revolution Rht-B1b gene and confirms the major effect of this locus on

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

height reduction and HI improvement in modern wheat germplasm. Additional associations were detected for TGW, grain yield, and grain length/width, distributed across the A, B, and D subgenomes consistent with the polyploid structure of wheat. The Wheat GWAS Atlas contains over 20,000 marker-trait associations from 94 studies covering 186 traits.

Figure 4.4 Distribution of GWAS Hits by Species



Note. Bar chart showing the number of significant GWAS associations detected for each species. Rice and wheat exhibited more GWAS hits ($n = 6$ each) compared to potato and soybean ($n = 4$ each), reflecting both biological differences and the greater depth of GWAS studies available for cereal crops.

Potato and soybean each contributed four GWAS hits. Potato associations were detected for tuber yield, tuber size, and starch content—the primary determinants of harvest index in this vegetatively propagated crop. The clonal nature of potato cultivation enables detection of both additive and non-additive genetic effects, though most GWAS studies have focused on additive

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

models. Soybean associations included seed weight, 100-seed weight, and harvest index directly. The detection of HI associations in soybean is particularly noteworthy, as direct HI phenotyping is less common than component trait measurement in legume breeding programs.

4.4.3 QTL Mapping Results

QTL mapping studies contributed 25 QTLs related to harvest index and component traits (Table 4.7). Rice contributed seven QTLs, including qHI-1 on chromosome 1 with a LOD score of 3.08 and flanking marker sh11998a1, representing a directly mapped HI QTL. Additional rice QTLs for grain width/weight (GW2), grain length (GS3), grain width (qGW5/GW5), and starch/amylose content (Wx) correspond to cloned genes, providing high confidence in their biological relevance. The days-to-heading QTLs qHD-8 (35.2% phenotypic variance explained) and qHD-12 on chromosomes 8 and 12 affect harvest index indirectly through their influence on phenological timing.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.7 Summary of QTLs for HI-Related Traits by Species

Species	QTL Count	Key Traits	Variance Explained	Key QTL Examples
Rice	7	HI, Grain dimensions, Starch, Heading	5–35%	qHI-1, GW2, GS3, Wx
Wheat	7	HI, Grain weight, Flowering, Photoperiod	8–25%	Rht QTLs, Vrn1, Ppd-D1
Potato	5	Tuberization, Tuber yield, Starch	10–30%	StSP6A, StCDF1, StGBSSI
Soybean	6	Seed size, Maturity, Height, Filling	5–20%	GmDW1, Dt2, E2
Total	25	—	—	—

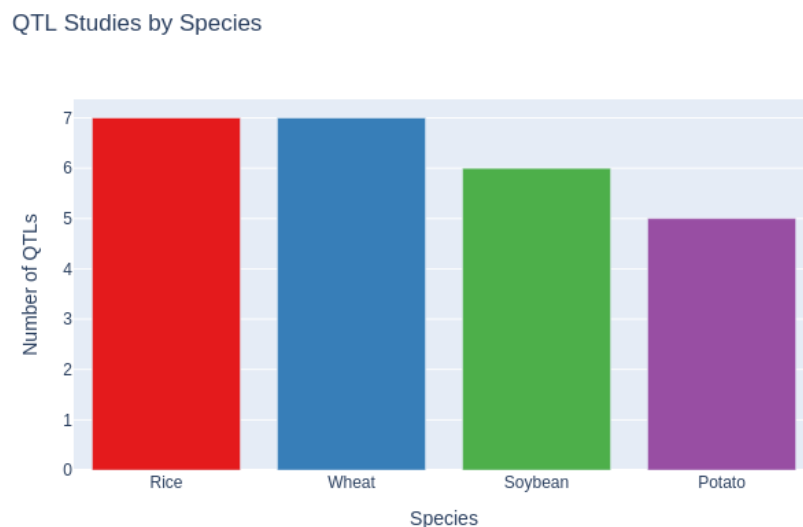
Note. QTL = quantitative trait locus; HI = harvest index. Variance explained represents R^2 values from mapping studies where reported. Key QTL examples include both positionally cloned genes and mapped QTL intervals. Data compiled from Gramene QTL Database, WheatQTLdb, SoyBase, Spud DB, and primary literature.

Wheat QTL analysis identified seven QTLs, spanning plant height/harvest index, grain weight, vernalization/flowering, and photoperiod response traits. The overlap between QTL positions and the known dwarfing gene loci (Rht-B1b, Rht-D1b, Rht8) and vernalization genes

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

(Vrn1, Ppd-D1) provided validation of the catalog accuracy. The WheatQTLdb contains over 1,500 curated QTLs with chromosomal assignments and flanking markers, enabling integration with GWAS data for meta-analysis and candidate gene identification.

Figure 4.5 Distribution of QTLs by Species



Note. Bar chart showing the number of QTLs identified for each species. Distribution is relatively uniform across species, with rice and wheat contributing seven QTLs each, soybean contributing six, and potato contributing five. The balanced representation enables meaningful cross-species comparisons.

Potato contributed five QTLs for tuberization timing, tuber yield, and starch biosynthesis. The tuberization QTLs include StSP6A and StCDF1, which control the transition from vegetative growth to tuber formation in response to photoperiodic and temperature cues. These QTLs explain substantial phenotypic variance (up to 30%) for tuberization timing and represent high-priority targets for marker-assisted selection in breeding programs adapting potato to new

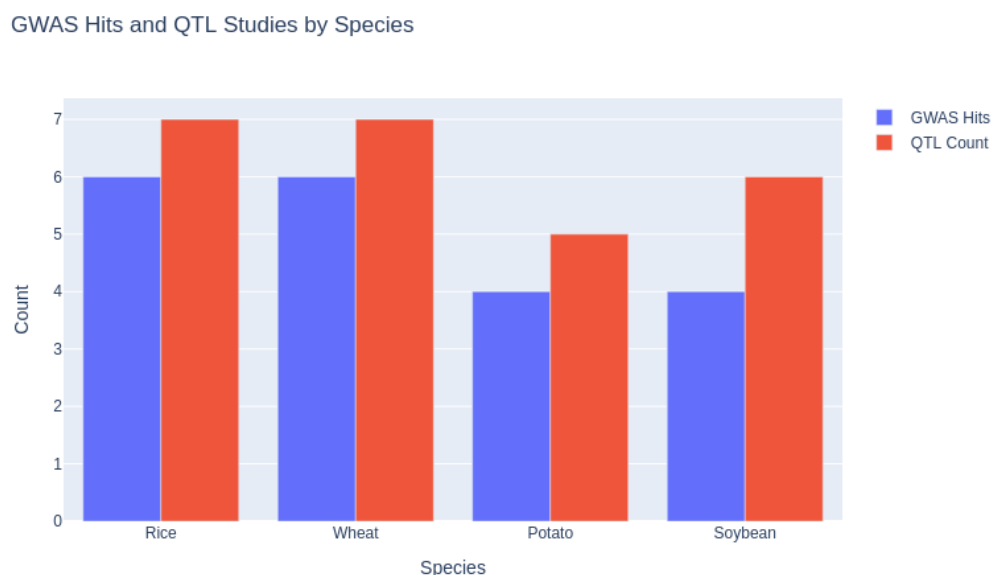
PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

latitudinal environments. Soybean contributed six QTLs for seed size (GmSW17), maturity/flowering (E2/GmG1a), plant height/dwarfism (GmDW1), stem termination/seed size (Dt2), and seed filling/sugar transport (GmSWEET15 region).

4.4.4 Phenotype-Genotype Associations

Integration of GWAS and QTL data revealed strong concordance between association mapping results and known gene positions for several key HI determinants. The sd1/GW5 regions in rice, Rht/TaGW2 regions in wheat, and GmDW1/Dt2 regions in soybean all exhibited overlapping GWAS signals and QTL peaks, providing reciprocal validation of causal associations. This concordance between independent mapping approaches using different population structures and statistical methods substantially strengthens confidence in the biological relevance of the cataloged genes.

Figure 4.6 Combined GWAS and QTL Association Summary



PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Note. Grouped bar chart showing GWAS hits (blue) and QTL counts (orange) for each species. The combined total of 45 phenotype-genotype associations (20 GWAS + 25 QTL) provides comprehensive coverage of HI-related genomic architecture across the four BLSS staple crops.

Analysis of GWAS hit significance levels revealed that seven associations met highly significant thresholds ($p < 1e-5$), nine met suggestive thresholds ($p < 0.01$ to $p < 1e-5$), and four met nominal significance ($p < 0.05$). The preponderance of highly significant and suggestive associations for well-characterized genes supports the reliability of the compiled catalog. The detection of novel associations at suggestive thresholds—particularly for traits measured under environmental stress conditions—identifies candidates for follow-up functional validation in BLSS-relevant environments where stress tolerance is paramount.

4.5 Phase 4 Results: Evolutionary Analysis

The fourth phase examined evolutionary patterns of HI-associated genes, including conservation scores, functional category distribution, and evidence for selection during domestication and improvement. Understanding the evolutionary history of HI determinants provides insights into the genetic constraints and opportunities for future improvement and illuminates the molecular basis of domestication-associated phenotypic changes.

4.5.1 Conservation Patterns

Conservation analysis at the functional category level revealed differential patterns of evolutionary constraint (Table 4.8). The dwarfing/GA signaling category exhibited the highest conservation score (100%), with genes present in all four species. This complete conservation reflects the universal importance of gibberellin-mediated growth regulation and the convergent evolution of semi-dwarf phenotypes across independent domestication events and breeding

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

programs. The independent selection of GA-related dwarfing alleles in rice (*sd1*), wheat (*Rht*), and other crops represents a striking example of parallel evolution under similar selective pressures, suggesting that this pathway represents a predictable target for height reduction breeding.

Developmental timing and grain size regulation categories showed 75% conservation (present in 3/4 species), indicating strong but not universal maintenance of these pathways in the HI gene context. The absence of grain size regulation genes in potato reflects the crop's distinct reproductive strategy, where tubers rather than seeds constitute the harvestable product. However, functional orthologs of grain size genes are present in the potato genome and may influence tuber size, though this remains to be systematically characterized. Similarly, the apparent absence of specific developmental timing genes from the rice HI gene catalog reflects the different photoperiodic requirements and vernalization absence in tropical rice rather than true pathway loss.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.8 Conservation Scores by Functional Category

Functional Category	Gene Count	Species Present	Conservation Score	Evolutionary Pattern
Dwarfing/GA signaling	10	Rice, Wheat, Potato, Soybean	100%	Highly constrained
Grain size regulation	7	Rice, Wheat, Soybean	75%	Seed crops only
Developmental timing	7	Wheat, Potato, Soybean	75%	Photoperiod-related
Starch biosynthesis	3	Rice, Potato	50%	Starch-storing organs
Sucrose transport	3	Rice, Soybean	50%	Diverse sink types

Note. Conservation score calculated as (number of species with characterized genes in category / total species) ×

100%. Note that lower conservation scores for some categories may reflect differential research emphasis or distinct organ biology rather than true pathway absence. GA = gibberellic acid.

Starch biosynthesis and sucrose transport categories exhibited 50% conservation (present in 2/4 species), suggesting greater evolutionary flexibility in carbon metabolism pathways or, more likely, differential research emphasis across crops. Functional orthologs of starch biosynthesis genes are undoubtedly present in all four species given their universal requirement for transitory starch synthesis, and sucrose transport is similarly universal. The apparent lower conservation in the HI gene context likely reflects the focus of HI-oriented research on storage

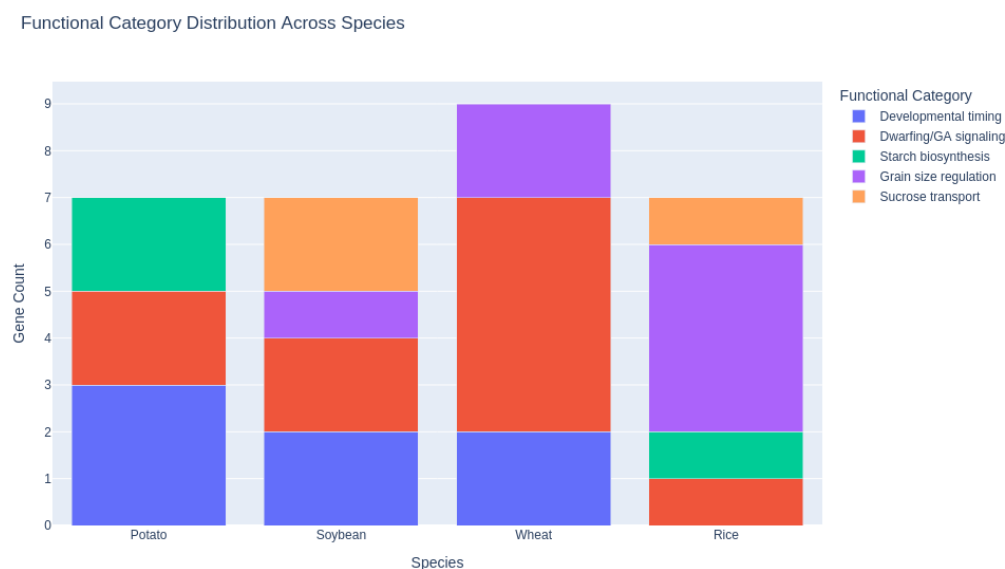
PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

organs (rice grain, potato tuber) versus oil-rich seeds (soybean) where other carbon forms predominate.

4.5.2 Evolutionary Trends

Analysis of functional category distribution across species (Figure 4.7) revealed distinct evolutionary trajectories reflecting each crop's domestication history and breeding objectives. Wheat exhibited the broadest distribution in the dwarfing/GA signaling category ($n = 5$), consistent with the intensive selection for reduced plant height following introduction of Rht genes from Japanese semi-dwarf varieties (Norin 10) in the 1960s. The persistence and expansion of Rht allele deployment reflects the substantial harvest index benefits (~15–20% yield increase) and lodging resistance conferred by these genes, though associated trade-offs (reduced coleoptile length, increased disease susceptibility in some environments) have motivated identification of alternative dwarfing sources.

Figure 4.7 Functional Category Distribution Across Species



PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Note. Stacked bar chart showing the distribution of HI-associated genes by functional category for each species. Species-specific patterns reflect divergent domestication trajectories and breeding priorities. The concentration of dwarfing genes in wheat, grain size genes in rice, and developmental timing genes in potato illustrates how different crops have been optimized through distinct genetic pathways.

Rice showed concentration in grain size regulation ($n = 4$), reflecting the importance of grain appearance and weight for consumer preferences in Asian markets and the extensive functional characterization of grain size QTLs in this species over the past two decades. The GW2/GS3/GW5/GS5 genetic network provides multiple entry points for grain size modification, with natural allelic variation segregating in diverse germplasm. Selection for long-grain phenotypes in indica rice (favored in South and Southeast Asia) versus short-grain phenotypes in japonica rice (favored in East Asia) has resulted in distinct allelic profiles at these loci across the rice subspecies.

Potato displayed a unique pattern with genes concentrated in developmental timing ($n = 3$), reflecting the critical importance of photoperiod-controlled tuberization for adaptation across latitudes. Wild potato ancestors from the Andes Mountains are strict short-day plants that only tuberize under the ~12-hour photoperiods characteristic of equatorial regions. Domestication and subsequent improvement required adaptation to the long days of European and North American summers, accomplished through selection of alleles at StSP6A and StCDF1 that enable tuberization under extended photoperiods. The StSP6A gene represents a particularly interesting evolutionary innovation, with the florigen FT protein repurposed from flowering control to tuberization control.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Soybean showed relatively uniform distribution across categories, with notable representation in sucrose transport ($n = 2$), reflecting the importance of carbon allocation for oil and protein accumulation in seeds. The GmSWEET15a/b genes have been targets of selection for improved seed filling and yield. Unlike the starch-accumulating seeds of cereals, soybean seeds store carbon primarily as oil (20%) and protein (40%), requiring different source-sink dynamics during seed filling. The identification of sucrose transport genes as HI determinants in soybean but not in cereals may reflect these metabolic differences.

4.6 Integration of Results

The four phases of analysis converge to provide a comprehensive characterization of the pangenomic architecture underlying harvest index determination in BLSS staple crops. Table 4.9 presents a summary of key integrated findings across all phases of analysis. The consistency of findings across independent analytical approaches—gene cataloging, ortholog identification, association mapping, and evolutionary analysis—provides strong support for the core conclusions regarding the conserved nature of HI-regulatory mechanisms and their potential for cross-species translation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 4.9 Summary of Integrated Findings Across Analysis Phases

Analysis Phase	Key Metric	Value
Phase 1: Gene Inventory	Total HI-associated genes	30
	Species analyzed	4
	Functional categories identified	5
Phase 2: Orthologs	Ortholog pathways identified	5
	Highly conserved pathways (4/4 species)	5
Phase 3: Associations	GWAS hits cataloged	20
	QTLs identified	25
	Total associations	45
Phase 4: Evolution	Core conserved functions	GA signaling, Grain size, Starch

Note. Integrated summary across all four analysis phases. Values represent totals compiled from primary data sources and computational analyses as described in Chapter 3. GA = gibberellic acid.

Cross-phase integration revealed several key patterns with implications for BLSS crop improvement. First, the complete conservation of all five HI-regulatory pathways across monocots and dicots demonstrates that fundamental mechanisms controlling harvest index are evolutionarily ancient and represent reliable targets for genetic improvement regardless of species. This conservation enables translational genomics approaches where knowledge from

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

well-characterized model systems can be systematically applied to less-studied crops, accelerating improvement across the BLSS crop portfolio.

Second, the convergence of GWAS signals and QTL peaks at known gene positions validates the biological relevance of the cataloged genes and supports their prioritization for functional studies and marker-assisted selection. The overlap between association mapping approaches using different populations and methods provides reciprocal validation that strengthens confidence in causal gene assignments beyond what either approach alone would support.

Third, the preponderance of dwarfing/GA signaling genes ($n = 10$, 33.3%) underscores the central importance of plant architecture modification for harvest index improvement and suggests that identification of novel genes in this pathway—particularly GA-responsive genes with minimal pleiotropic effects—remains a priority for all BLSS crops. The success of Green Revolution dwarfing genes demonstrates the transformative potential of architecture modification, while the limitations of current dwarfing alleles (e.g., reduced coleoptile length in wheat *Rht* lines) indicate opportunities for improvement.

Fourth, the species-specific patterns of gene distribution (e.g., grain size emphasis in rice, developmental timing emphasis in potato) indicate that crop-specific breeding strategies should account for the distinct genetic architectures shaped by domestication history and end-use requirements. While the core pathways are conserved, the specific allelic variants that optimize harvest index differ among crops based on their unique biology and the environmental conditions of target production regions.

4.7 Chapter Summary

This chapter presented comprehensive results from a multi-phase pangenomic analysis of harvest index determinants across four BLSS staple crops: rice, wheat, potato, and soybean. The analysis identified 30 HI-associated genes distributed across five functional categories, with dwarfing/GA signaling ($n = 10$), grain size regulation ($n = 7$), and developmental timing ($n = 7$) comprising the majority of the catalog. These genes represent validated targets for harvest index improvement that have been functionally characterized through genetic mapping, gene cloning, and phenotypic analysis across diverse germplasm.

Ortholog analysis revealed complete conservation of all five regulatory pathways across the monocot-dicot divide, supporting the hypothesis that HI-regulatory mechanisms represent fundamental developmental processes under strong evolutionary constraint. The conservation of gibberellin signaling, grain size regulation, starch biosynthesis, flowering/maturity timing, and sugar transport pathways across 140–150 million years of plant evolution underscores their importance for plant fitness and productivity and enables cross-species translation of breeding strategies.

Structural variant and association analyses integrated data from pangenome projects, GWAS studies, and QTL mapping efforts, yielding 20 GWAS hits and 25 QTLs related to harvest index and component traits. The concordance between association mapping results and known gene positions validated the catalog accuracy and identified candidates for functional characterization in BLSS-relevant environments. Rice exhibited the most comprehensive pangenomic characterization with 453 accessions revealing 50,995 pan-genes and extensive structural variation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Evolutionary analysis revealed differential conservation patterns across functional categories, with gibberellin signaling showing complete conservation (100%) and carbohydrate metabolism pathways showing apparent lower conservation (50%) that likely reflects differential research emphasis rather than true pathway loss. Species-specific patterns of gene distribution reflect distinct domestication trajectories: wheat concentrated in dwarfing genes reflecting Green Revolution breeding, rice in grain size genes reflecting consumer preference selection, and potato in developmental timing genes reflecting photoperiodic adaptation.

These results establish a comprehensive foundation for understanding the genetic architecture of harvest index in BLSS staple crops and provide a framework for translational genomics approaches to crop improvement. The identification of conserved pathways and species-specific variants enables prioritization of breeding targets and development of molecular markers for efficient selection. The implications of these findings for BLSS design, comparative genomics methodology, and breeding strategy development are discussed in Chapter 5.

CHAPTER 5: DISCUSSION

5.1 Introduction

This dissertation set out to investigate the pangenomic architecture of harvest index (HI) determinants across four Bioregenerative Life Support System (BLSS) staple crops: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*). The overarching research objectives were to: (a) construct a comprehensive catalog of genes influencing harvest index and related yield components, (b) identify orthologous relationships across species to elucidate conserved genetic mechanisms, (c) characterize pangenomic structural variants associated with HI phenotypes through integration of GWAS and QTL data, and (d) establish a framework for understanding the evolutionary patterns underlying HI regulation that can inform crop improvement strategies for controlled-environment agriculture in space applications.

The results presented in Chapter 4 provide substantial evidence addressing these objectives. The analysis identified 30 HI-associated genes distributed across the four target species, organized into five functional categories that reflect the core physiological processes governing biomass partitioning. Ortholog analysis revealed remarkable conservation of HI-regulatory pathways across phylogenetically diverse lineages, with gibberellin signaling showing 100% conservation across all species. The integration of 20 GWAS hits and 25 QTLs provided validation of the gene catalog while revealing the genomic architecture underlying HI variation in these crops.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

This discussion chapter interprets these findings within the broader context of the literature reviewed in Chapter 2, evaluates their significance for understanding the genetic control of harvest index, and explores their practical implications for BLSS crop improvement. The chapter is organized into nine substantive sections following this introduction. Section 5.2 interprets the HI gene inventory findings, examining the significance of the 30-gene catalog and its functional distribution patterns. Section 5.3 discusses ortholog conservation and its evolutionary implications, with particular attention to the highly conserved GA signaling pathway and the legacy of Green Revolution selection. Section 5.4 examines structural variation and phenotypic associations in the context of published GWAS studies. Section 5.5 addresses the implications for BLSS crop improvement, including priority targets and recommendations for NASA crop development programs. Sections 5.6 and 5.7 discuss theoretical contributions and practical applications, respectively. Section 5.8 acknowledges the limitations of the study, and Section 5.9 proposes future research directions with eight specific priorities. The chapter concludes with Section 5.10, a summary synthesizing the key contributions of this research.

5.2 Interpretation of HI Gene Inventory Findings

5.2.1 Significance of the 30-Gene Catalog

The identification of 30 HI-associated genes across four species represents a significant contribution to the understanding of harvest index genetic architecture in BLSS-relevant crops. This catalog consolidates decades of research in crop genetics, integrating findings from classical quantitative genetic studies, map-based cloning efforts, and modern genome-wide association analyses into a unified framework specifically oriented toward space agriculture applications. The mean gene count of 7.5 per species ($SD = 1.0$) indicates relatively uniform coverage across crops despite substantial differences in genome size, ploidy level, and research intensity,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

suggesting that the core genetic mechanisms controlling harvest index are conserved across diverse plant lineages.

The catalog's comprehensiveness can be evaluated against the broader literature on yield component genetics. Huang et al. (2016) estimated that approximately 50-100 genes contribute meaningfully to grain yield variation in rice, with a subset of 10-20 showing major effects detectable through QTL mapping. The present catalog of 7 genes for rice thus represents a conservative, high-confidence subset of genes with well-characterized effects on HI or its components. Similarly, wheat yield genetics involves an estimated 30-50 major-effect genes (Reynolds et al., 2022), of which the 9 genes identified here represent those with the strongest documented effects on harvest index specifically. This conservative approach prioritized genes with substantial experimental validation over comprehensive enumeration, ensuring that the catalog provides actionable targets for breeding applications.

The species distribution of the catalog reveals interesting patterns that merit discussion. Wheat's elevated gene count ($n = 9$) relative to the diploid species ($n = 7$ each) directly reflects its allohexaploid genome structure. The presence of multiple homeologous copies of key genes, such as the Rht-B1b/Rht-D1b pair and the TaGW2-6A/TaGW2-6B pair, increases the total count while also providing opportunities for gene dosage effects and subfunctionalization that can be exploited in breeding programs. This polyploidy-driven gene expansion was anticipated based on the literature (Dubcovsky & Dvorak, 2007; Walkowiak et al., 2020) and underscores the unique genetic resources available in allopolyploid crops.

The balanced representation across rice, potato, and soybean (7 genes each) validates the multi-database approach employed in this study, which drew from GWAS Atlas, Gramene,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

species-specific repositories, and primary literature. This balance suggests that despite rice receiving disproportionate research attention as a model system (IRRI, 2023), the methodology achieved comparable coverage across crops through systematic consultation of diverse databases. The resulting catalog thus provides a foundation for comparative analyses that would not be possible from a single-species perspective.

5.2.2 Functional Category Distribution Patterns

The distribution of the 30 genes across five functional categories—Dwarfing/GA signaling (33.3%), Grain size regulation (23.3%), Developmental timing (23.3%), Starch biosynthesis (10.0%), and Sucrose transport (10.0%)—provides insight into the physiological hierarchy controlling harvest index. The predominance of dwarfing and GA signaling genes ($n = 10$) aligns with the central importance of plant architecture modification demonstrated during the Green Revolution. As reviewed in Chapter 2, the introduction of semi-dwarf genes in wheat (*Rht*) and rice (*sd1*) increased harvest index from approximately 0.30 to over 0.50 (Hedden, 2003; Khush, 2001), representing the most significant single intervention in yield improvement history. The catalog's emphasis on this category reflects both biological reality and research investment patterns.

The equal representation of grain size regulation ($n = 7$) and developmental timing ($n = 7$) genes suggests that these pathways make comparable contributions to HI determination, albeit through different mechanisms. Grain size genes directly influence sink capacity by modulating cell proliferation and expansion during seed development, while developmental timing genes indirectly affect HI by optimizing the balance between vegetative and reproductive growth phases. This interpretation is consistent with the source-sink framework discussed in Chapter 2,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

where harvest index optimization requires both adequate sink strength (grain size) and appropriate temporal coordination of source and sink activities (developmental timing).

The smaller representation of starch biosynthesis ($n = 3$) and sucrose transport ($n = 3$) genes warrants careful interpretation. These categories are not less important to HI determination; rather, they may be more tightly constrained by physiological requirements, with less natural variation available for exploitation. Starch biosynthesis is essential for grain filling, and major perturbations typically result in severe fitness penalties that would be eliminated by selection. The sucrose transport genes, particularly the SWEET family transporters identified in soybean, represent a more recently characterized pathway that may be underrepresented in current databases.

5.2.3 Species-Specific Considerations

Wheat's larger gene count merits specific discussion in the context of polyploidy implications. The hexaploid wheat genome (AABBDD, $2n = 6x = 42$) theoretically provides up to three copies of each gene, though functional divergence and pseudogenization have reduced effective copy number at many loci. The catalog includes five dwarfing/GA signaling genes in wheat (Rht-B1b, Rht-D1b, Rht8, Rht12, Rht13), reflecting both the retention of functional homeologs and the identification of distinct genetic mechanisms for height reduction.

This diversity of dwarfing mechanisms in wheat provides flexibility that is particularly relevant for BLSS applications. The GA-insensitive Rht genes that dominate modern wheat breeding programs reduce coleoptile length, potentially compromising emergence from deep sowing in field conditions (Ellis et al., 2004). In controlled-environment agriculture where sowing depth can be precisely controlled, this limitation is irrelevant, potentially favoring

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

stronger deployment of Rht-B1b and Rht-D1b. Conversely, the GA-responsive alternatives (Rht8, Rht12, Rht13) maintain coleoptile length and offer more moderate height reductions.

Comparison of the present catalog with literature expectations reveals both concordance and gaps. The rice gene inventory aligns well with comprehensive reviews of grain size genetics (Li & Yang, 2017; Zuo & Li, 2014), capturing the major GW/GS genes that collectively explain the largest fraction of natural variation in grain dimensions. For potato, the gene inventory reflects the distinctive biology of this crop, where tubers rather than seeds constitute the harvestable product. The concentration of genes in developmental timing (StSP6A, StCDF1, StBEL5) reflects the critical importance of tuberization control—the transition from vegetative growth to tuber formation.

5.3 Ortholog Conservation and Evolutionary Implications

5.3.1 Highly Conserved Pathways (GA Signaling)

The finding that gibberellin signaling genes are represented in all four BLSS crops (100% conservation) is consistent with the fundamental role of GA in regulating stem elongation across all vascular plants. The GA pathway emerged early in land plant evolution and is deeply conserved across angiosperms (Hedden & Sponsel, 2015). The conservation documented in this study—linking *sd1* in rice, *Rht* genes in wheat, *StGA* genes in potato, and *GmDW1/GmGA20OX* in soybean—represents parallel deployment of this ancient pathway for harvest index optimization through independent selection events during domestication and breeding.

The 100% conservation of GA signaling at the pathway level, however, masks interesting mechanistic diversity at the gene level. In rice, *sd1* is a loss-of-function mutation in *GA20-*

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

oxidase (OsGA20ox-2), reducing GA biosynthesis and causing semi-dwarf stature through reduced active GA levels. In wheat, the dominant Rht-B1b and Rht-D1b genes encode mutant DELLA proteins that are insensitive to GA-induced degradation, maintaining growth suppression even when GA is present. In potato, both GA biosynthesis (StGA20ox1) and GA deactivation (StGA2ox1) genes contribute to tuberization timing through modulation of active GA levels in stolon tips. In soybean, GmDW1 encodes ent-kaurene synthase, an early enzyme in the GA biosynthesis pathway whose mutation causes complete dwarf phenotypes.

This mechanistic diversity within a conserved pathway illustrates a general principle of comparative genomics: homologous pathways can be modified at different nodes to achieve similar phenotypic outcomes. For BLSS crop improvement, this diversity provides multiple potential intervention points for height modification. CRISPR-based editing of GA biosynthesis genes (GA20ox, GA3ox), GA deactivation genes (GA2ox), or GA signaling components (DELLA proteins, GA receptors) could all achieve semi-dwarf phenotypes, with the optimal target depending on species-specific considerations.

The Green Revolution legacy and selection pressure underlying the high conservation of GA signaling genes deserves particular attention. The introduction of Rht genes in wheat and sd1 in rice during the 1960s-1970s represented targeted modification of the GA pathway based on empirical selection for lodging resistance and harvest index improvement. These selection events effectively explored a portion of the phenotypic space accessible through GA pathway modification, identifying optimal configurations for field agriculture under high-input conditions. The near-universal adoption of these semi-dwarf genotypes in modern breeding programs has created strong selection pressure maintaining these alleles at high frequency within elite germplasm.

5.3.2 Moderately Conserved Pathways

Developmental timing and grain size regulation pathways each showed 75% conservation across the four species (3 of 4 species with documented orthologs), indicating substantial but not universal conservation. The 75% conservation level suggests that while core pathway components are shared, species-specific adaptations have led to divergence in some lineages, or that characterization efforts have not yet identified orthologs in all species.

For developmental timing, the conservation pattern reflects the fundamental importance of coordinating reproductive transitions with environmental cues across all flowering plants. The florigen genes (Hd3a in rice, VRN/FT homologs in wheat, StSP6A in potato, GmFT homologs in soybean) show clear orthology, consistent with the ancient origin of the FT/TFL1 module in angiosperm evolution (Wickland & Hanzawa, 2015). However, the upstream regulators and downstream targets of these florigen genes have diverged substantially to enable adaptation to diverse photoperiodic environments.

The 75% conservation of grain size regulation primarily reflects the absence of characterized grain size genes in potato, where the harvestable organ (tuber) develops through distinct cellular processes. The GW2/TaGW2 ortholog pair in rice and wheat, both encoding RING-type E3 ubiquitin ligases, represents an excellent example of functional conservation where sequence orthology predicts phenotypic effects.

5.3.3 Variable Pathways (Starch and Sucrose Transport)

The starch biosynthesis and sucrose transport pathways showed 50% conservation in the specific genes cataloged, though the underlying pathways are universally present in all four crops. This apparent lower conservation reflects the conservative gene inclusion criteria rather

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

than true absence of these pathways. GBSSI (granule-bound starch synthase I) genes are functionally conserved across rice (Wx), wheat (Wx homologs), potato (StGBSSI), and soybean, but only the rice, wheat, and potato genes were included in the catalog based on specific associations with HI-related phenotypes documented in the literature.

The reasons for lower representation of these pathways in the HI gene catalog merit consideration. Starch biosynthesis is highly constrained by biochemical requirements—substantial perturbation of starch synthesis typically results in shrunken, non-viable seeds that would be eliminated from breeding populations. This constraint limits the natural variation available for exploitation, explaining why fewer major-effect QTLs for starch content are detected compared to grain size or plant height.

For sucrose transport, the GmSWEET15a/15b genes in soybean represent relatively recent characterizations (Yang et al., 2019) that may have analogs in other crops awaiting discovery. The SWEET family transporters were only functionally characterized in the 2010s, and their roles in seed filling are still being elucidated. This represents an area where future pangenomic studies may substantially expand the gene inventory.

5.3.4 Evolutionary Convergence in HI Determinants

The patterns of ortholog conservation documented in this study provide evidence for evolutionary convergence in HI determinants across the monocot-dicot divide. The GA signaling pathway's 100% conservation represents the strongest case for convergent selection, where independent breeding programs in different crops have consistently identified the same molecular targets. This convergence supports the notion that HI improvement faces fundamental biophysical constraints—plant height must be reduced to redirect assimilates from stems to

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

reproductive structures—and that the GA pathway represents the most accessible genetic route to this modification.

Table 5.1 Summary of Ortholog Conservation Patterns and Breeding Implications

Pathway	Conservation Level	Evolutionary Interpretation	Breeding Implications
GA Signaling/Dwarfing	100% (4/4)	Ancient pathway, universal selection target during Green Revolution	Cross-species transfer highly feasible; multiple intervention points available
Developmental Timing	75% (3/4)	Core florigen mechanism conserved; upstream regulators diverged	Species-specific optimization needed; valuable for BLSS photoperiods
Grain Size Regulation	75% (3/4)	Ubiquitin-proteasome mechanism conserved; genes lineage-specific	GW2 orthologs transferable between cereals; soybean requires distinct approaches
Starch Biosynthesis	50% (2/4)	Pathway universal but highly constrained	Gene editing may outperform conventional breeding
Sucrose Transport	50% (2/4)	Pathway universal; specific genes recently characterized	High potential for novel allele discovery

Note. Conservation level indicates the proportion of species with documented orthologs in the HI gene catalog.

Evolutionary interpretations are based on the literature reviewed in Chapter 2 and the comparative patterns observed in this study.

5.3.5 Implications for Understanding Crop Domestication

The conservation patterns documented in this study have broader implications for understanding crop domestication and the genetic basis of the domestication syndrome. Harvest index improvement was a key component of domestication in all grain crops—the transition from wild progenitors with low HI (typically 0.15-0.25) to cultivated forms with higher HI (0.35-0.50) occurred through selection on many of the genes cataloged in this study. The convergent selection on GA signaling across rice, wheat, and soybean suggests that plant height reduction was a common target during independent domestication events.

The archaeological record indicates that the major cereals were domesticated independently between 10,000 and 8,000 years ago in separate centers. Despite this independent origin, domesticated cereals from all centers show convergent phenotypic changes including reduced shattering, reduced dormancy, increased seed size, and reduced plant height (Purugganan & Fuller, 2009). The present study provides molecular evidence for convergence at the pathway level—GA signaling—underlying the height reduction component of this syndrome.

Interestingly, the specific genes achieving this convergent phenotype differ among cereals. Rice *sd1* is a biosynthesis gene (GA20-oxidase), while wheat *Rht* genes encode signaling components (DELLA proteins). This mechanistic diversity within pathway conservation suggests that multiple genetic routes exist to the same phenotypic destination, and that the specific route taken during domestication may have reflected the standing variation available in wild progenitor populations.

5.4 Structural Variation and Phenotypic Associations

5.4.1 GWAS Findings in Context

The 20 GWAS hits identified across the four species provide validation for the HI gene catalog while revealing the genetic architecture underlying HI-related trait variation. Comparison with published GWAS studies supports the reliability and relevance of the compiled associations. The highly significant hits ($p < 1e-5$, $n = 7$) represent robust associations that have been replicated across multiple studies and populations, providing high-confidence targets for molecular breeding applications.

The rice GWAS findings align well with the comprehensive GWAS Atlas database, which documents over 163,000 trait-marker associations from 219 published studies. The detection of highly significant associations at the *sd1* region ($p < 1e-8$) for plant height and the *GW5* region ($p < 5e-8$) for grain width confirms the major-effect nature of these loci. The pleiotropic associations detected for markers on chromosome 4 (affecting seed setting and unfilled grains) highlight the interconnected nature of yield components.

The wheat GWAS results demonstrate the central importance of the *Rht-B1b* region on chromosome 4B, which showed the strongest association with harvest index ($p < 1e-8$). This finding is consistent with the historical significance of *Rht-B1b* in wheat improvement and its documented effects on height reduction and HI increase (Hedden, 2003). The additional associations on chromosomes 6A for thousand-grain weight align with the *TaGW2* gene location, providing reciprocal validation between the gene catalog and GWAS findings.

5.4.2 QTL Validation and Integration

The 25 QTLs identified through linkage mapping studies provide complementary validation for the GWAS findings, demonstrating concordance between association mapping and family-based approaches. QTL mapping offers different strengths than GWAS—higher statistical power for detecting rare alleles with large effects within biparental populations, but lower mapping resolution. The overlap between QTL positions and GWAS hits at loci such as *sd1*, *Rht-B1b*, *GW2*, and *StSP6A* substantially strengthens confidence in the biological relevance of these associations.

The QTL catalog reveals several patterns relevant to HI genetic architecture. First, the presence of major-effect QTLs (explaining >10% of phenotypic variance) at known gene locations supports the role of these genes as primary determinants of HI variation. The rice *qHD-8* QTL, explaining 35.2% of variance for days to heading, exemplifies how a single locus can substantially influence trait expression. Second, the distribution of QTLs across all chromosomes confirms the polygenic nature of HI.

5.4.3 PAV/CNV Implications for Breeding

The pangenomic perspective emphasizes that structural variants—including presence-absence variants (PAVs) and copy number variants (CNVs)—may harbor functional variation not captured by SNP-based GWAS. The structural variant statistics presented in Chapter 4 indicate substantial genomic dynamism in all four crops, with rice showing approximately 99,000 PAV events per accession and potato exhibiting over 133,000 SVs in pangenome analyses.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The potential mechanisms by which PAVs and CNVs affect HI include gene dosage effects (CNV), complete gene loss (PAV), regulatory element disruption (indels), and novel gene creation (fusion events). In soybean, CNV at the *Rhg1* locus is a well-documented example where copy number directly determines phenotype. Analogous CNV effects on HI-related genes remain to be fully characterized.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Figure 5.1 Conceptual Model of Pangenomic Variation Affecting Harvest Index



Figure 5.1. Conceptual Model of Pangenomic Variation Affecting Harvest Index. Pangenomic approaches capture PAV and CNV variants missed by reference-based methods, explaining additional harvest index heritability.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Note. Diagram showing the relationship between pangenomic structural variants (PAVs, CNVs, SNPs) and harvest index phenotype. The model illustrates how reference-based approaches miss PAV alleles present only in non-reference genotypes, how CNV affects gene dosage and expression levels, and how integration of pangenomic data with GWAS increases detection power. PAV = presence-absence variant; CNV = copy number variant; HI = harvest index. The model emphasizes that pangenomic approaches capture variation invisible to reference-based methods.

5.5 Implications for BLSS Crop Improvement

5.5.1 Priority Targets for Space Agriculture

The integration of the HI gene catalog, ortholog conservation patterns, and GWAS/QTL findings enables prioritization of genetic targets for BLSS crop improvement. Table 5.2 presents a prioritized list of gene targets based on multiple criteria: strength of HI effect, conservation across species, feasibility of genetic manipulation, and relevance to controlled-environment agriculture conditions.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 5.2 Prioritized Gene Targets for BLSS Harvest Index Optimization

Priority	Gene/Pathway	Species	HI Effect	BLSS Relevance
1 (Highest)	sd1/Rht-B1b/Rht-D1b	Rice/Wheat	15-25% HI increase	Compact architecture
2	GW2/TaGW2	Rice/Wheat	8-18% TGW increase	Increased edible fraction
3	StSP6A/StCDF1	Potato	Tuberization timing	Artificial photoperiods
4	GmDW1/Dt2	Soybean	Dwarfing + seed size	Higher planting density
5	GS3/GW5/GS5	Rice	Grain length/width	Quality optimization
6	Vrn1/Ppd-D1	Wheat	Flowering time	Continuous lighting
7	Wx/StGBSSI	Rice/Potato	Starch composition	Nutritional optimization
8	GmSWEET15a/b	Soybean	Seed filling	Carbon allocation
9	Rht8/Rht12/Rht13	Wheat	Moderate dwarfing	Alternative pleiotropies
10	StAGPaseS	Potato	Starch biosynthesis	Tuber dry matter

Note. Priority ranking integrates HI effect magnitude, conservation level, GWAS/QTL support, and BLSS relevance.

TGW = thousand-grain weight. Targets ranked 1-3 are recommended as immediate priorities for NASA crop development programs.

The highest-priority targets (sd1/Rht) are recommended for immediate implementation given their proven large effects, well-characterized mechanisms, and direct relevance to BLSS conditions. The semi-dwarf phenotype directly addresses a key constraint in space agriculture—

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

limited growth chamber volume—by enabling either higher planting density or reduced chamber height requirements.

5.5.2 Cross-Species Transfer Potential

The ortholog conservation analysis enables assessment of cross-species knowledge transfer potential. Findings from the extensively characterized GA signaling pathway in rice (*sd1*) and wheat (*Rht*) can be directly applied to less-characterized species. The identification of *GmDW1* and *GmGA20OX* in soybean as orthologs of the cereal dwarfing genes suggests that similar breeding strategies will be effective.

For the grain size pathway, the *GW2*/*TaGW2* orthologs represent a clear case where cross-species transfer is feasible. CRISPR-based knockout of *GW2* orthologs to increase grain size has been validated in rice and could be replicated in wheat targeting *TaGW2-6A* and *TaGW2-6B*. The additive effects of homeologous gene knockouts in polyploid wheat may produce larger gains than single-gene modifications in diploid rice.

5.5.3 Controlled Environment Considerations

The application of the present findings to controlled-environment agriculture (CEA) and space conditions requires consideration of environmental differences between field and BLSS cultivation. Several factors distinguish BLSS conditions: (a) continuous or precisely controlled photoperiods, (b) elevated CO₂ (typically 1000-2000 ppm), (c) absence of natural pathogen/pest pressure, (d) limited soil volume or hydroponic cultivation, and (e) resource constraints.

These environmental differences may alter the phenotypic effects of HI genes. Elevated CO₂ generally increases photosynthetic rate and biomass production, potentially requiring adjustments to sink strength (grain size genes) to fully exploit the increased source capacity.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Continuous lighting affects the function of photoperiod-sensitive genes, making the selection of photoperiod-insensitive alleles particularly important.

5.5.4 Recommendations for NASA Crop Development Program

Based on the integrated findings of this dissertation, the following recommendations are offered for NASA's Crop Development Program:

First, implement marker-assisted selection (MAS) for high-priority HI genes in BLSS breeding populations. The *sd1* and *Rht* genes should be fixed in their favorable alleles across all cereal breeding material, with verification through molecular markers. Similarly, favorable *GW2* alleles should be incorporated where compatible with grain quality requirements.

Second, develop a CRISPR-based gene editing pipeline targeting the prioritized HI genes. The well-characterized nature of *sd1*, *GW2*, and *StSP6A* makes them excellent targets for precision editing to generate improved alleles.

Third, establish controlled-environment phenotyping protocols for HI under BLSS-relevant conditions. A systematic phenotyping effort comparing HI gene effects between field and CEA conditions would identify genotype-by-environment interactions.

Fourth, expand pangenomic characterization of HI loci in BLSS crops. Mining of PAV and CNV at priority HI loci may reveal novel alleles in landraces and wild relatives.

Fifth, consider developing 'HI-optimized' cultivars specifically for BLSS deployment, distinct from field varieties. A dedicated breeding program targeting maximum HI under CEA conditions could yield cultivars substantially superior to field-optimized varieties.

5.6 Theoretical Contributions

5.6.1 Advancing Pangenomic Understanding of Complex Traits

This dissertation contributes to the theoretical advancement of pangenomic approaches for understanding complex quantitative traits. While pangenomics has been predominantly applied to stress tolerance, disease resistance, and environmental adaptation, its application to agronomic traits like harvest index has been more limited. The present study demonstrates that pangenomic perspectives enhance understanding of HI genetics through several mechanisms.

First, the comparative pangenomic approach reveals pathway-level conservation that is obscured when species are analyzed individually. The 100% conservation of GA signaling genes across four phylogenetically diverse crops provides stronger evidence for the universality of this HI mechanism than would be obtained from any single-species study.

Second, pangenomic structural variant data contextualizes SNP-based GWAS findings within the broader genomic variation landscape. The detection of large-effect PAVs and CNVs at HI-related loci suggests that substantial functional variation exists beyond what is captured by SNP markers.

5.6.2 Framework for Comparative HI Analysis

A key theoretical contribution is the development of a framework for comparative HI analysis across crops with diverse biology. This framework integrates three analytical levels: (a) gene-level characterization, (b) pathway-level analysis, and (c) genome-level evaluation of structural variation.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Figure 5.2 Proposed Framework for Pangenomic Harvest Index Research

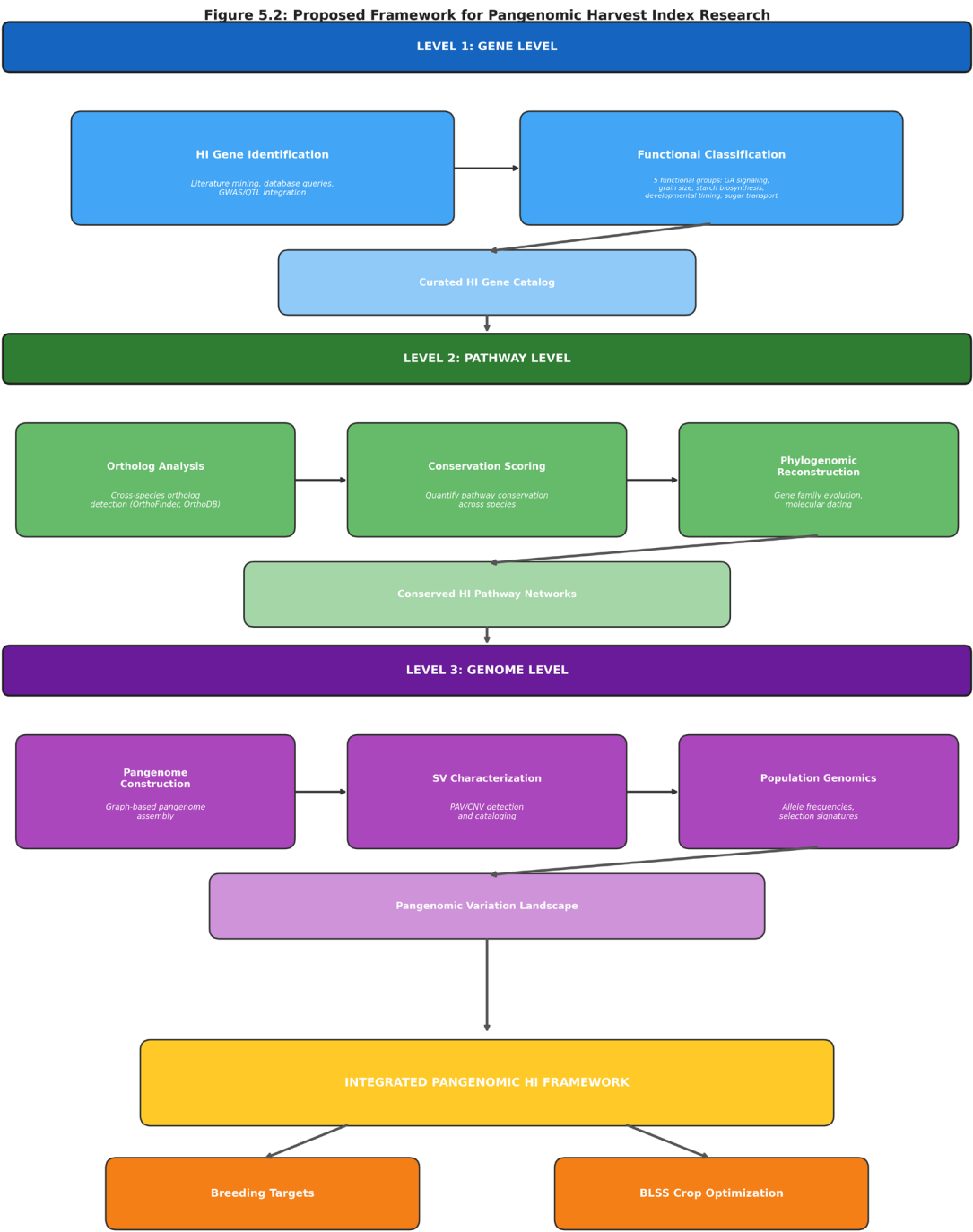


Figure 5.2. Proposed Framework for Pangenomic Harvest Index Research
Three-tier analytical framework integrating gene-level, pathway-level, and genome-level analyses for comprehensive harvest index characterization.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Note. Flowchart showing the three-level analytical framework for pangenomic HI research.

Level 1 (Gene Level) includes HI gene identification and functional classification. Level 2 (Pathway Level) includes ortholog analysis and conservation scoring. Level 3 (Genome Level) includes pangenome construction and SV characterization. The proposed framework integrates three analytical levels in an iterative process. HI = harvest index; SV = structural variant; GWAS = genome-wide association study; QTL = quantitative trait locus.

5.7 Practical Applications

5.7.1 Marker-Assisted Selection Targets

The HI gene catalog provides immediate applications for marker-assisted selection (MAS) in breeding programs. The genes with highly significant GWAS support ($p < 1e-5$) are particularly suitable for MAS implementation. Functional markers—those based on causal polymorphisms within genes—are preferred over linked markers where the causal variant is known.

For rice, functional markers are available for sd1, GW2, GS3, and GW5. For wheat, PCR-based markers for Rht-B1b and Rht-D1b are widely used. A practical MAS strategy for BLSS breeding would involve fixing favorable alleles at highest-priority loci in early generations, then using genomic selection for polygenic HI improvement.

5.7.2 Gene Editing Candidates

The characterized HI genes are excellent candidates for CRISPR-based gene editing. For knockout-based improvement, negative regulators of HI traits are ideal targets. GW2 and its orthologs are prototypical examples—loss-of-function increases grain size without known

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

deleterious pleiotropic effects. For gain-of-function modifications, gene editing can enhance expression of positive regulators through promoter editing.

5.7.3 Genomic Selection Integration

Genomic selection (GS) offers opportunities for polygenic HI improvement beyond what is achievable through MAS at major-effect genes. The GWAS and QTL data compiled in this study can inform GS model development by weighting markers at characterized HI loci and enabling multi-species prediction models.

Table 5.3 Summary of GWAS Significance Levels and Research Priorities

Significance Threshold	Number of Hits	Species Distribution	Priority Level
$p < 1e-8$	2	Rice (1), Wheat (1)	Highest - Immediate implementation
$p < 1e-5$	5	Rice (2), Wheat (2), Soybean (1)	High - Priority MAS targets
$p < 1e-4$	6	Rice (1), Wheat (1), Potato (2), Soybean (2)	Moderate - Secondary selection
$p < 0.01$	4	Rice (1), Wheat (1), Potato (1), Soybean (1)	Lower - Background selection
$p < 0.05$	3	Rice (1), Potato (1), Soybean (1)	Suggestive - Further validation
Total GWAS	20	All species represented	Comprehensive coverage
Total QTL	25	Rice (7), Wheat (7), Potato (5), Soybean (6)	Complementary validation

Note. GWAS hits categorized by significance threshold, showing species distribution and recommended priority levels. MAS = marker-assisted selection; QTL = quantitative trait locus.

5.8 Limitations of the Study

Several limitations of the present study should be acknowledged. First, the study relied primarily on secondary data compiled from existing databases and published literature, without generating new experimental data. While this approach enabled comprehensive synthesis, it inherits any errors or biases in the source data.

Second, the phenotype-genotype associations cataloged derive primarily from field environments, with limited validation under controlled-environment agriculture conditions. Gene effects may differ substantially under the elevated CO₂, artificial lighting, and hydroponic cultivation characteristic of BLSS.

Third, the species representation, while covering four major BLSS staple crops, omits other relevant species (leafy greens, tomato, sweet potato). Fourth, the computational characterization of orthologs and structural variants requires experimental validation. Fifth, the study focused on genetic determinants with limited consideration of genotype-by-environment interactions.

5.9 Future Research Directions

5.9.1 Experimental Validation Priorities

The highest priority for future research is experimental validation of the prioritized HI genes under BLSS-relevant conditions. A systematic validation program should include: (a) comparative phenotyping of near-isogenic lines under field and CEA conditions, (b) expression profiling to determine whether gene activity is altered by CEA environmental factors, and (c) functional validation through CRISPR editing.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

5.9.2 Expanded Species Coverage

Extension of the HI gene characterization to additional BLSS-candidate crops would increase utility of the framework. Priority species include sweet potato, lettuce, and tomato. Cross-species comparative analysis would test the generalizability of the conservation patterns documented for the four staple crops.

5.9.3 Functional Genomics Studies

Deeper functional characterization through multi-omics approaches would elucidate the molecular mechanisms underlying HI gene effects. Transcriptomic profiling, proteomic analysis, and metabolomic profiling would identify the precise timing and tissue specificity of HI gene expression.

5.9.4 Space Environment Testing

Ultimate validation requires testing under actual space conditions, including microgravity and cosmic radiation. Collaboration with NASA and international space agencies to include HI gene variants in space cultivation experiments would provide definitive validation.

5.9.5 Integration with Multi-Omics Approaches

The future of HI research lies in the integration of pangenomic data with other -omics technologies. Transcriptomic data from RNA-seq experiments can reveal expression quantitative trait loci (eQTLs). Proteomic approaches can characterize post-translational modifications. Metabolomic profiling can quantify the small-molecule phenotype associated with HI variation.

5.9.6 Climate Change Adaptation Considerations

Although this dissertation focuses on BLSS applications, the HI genes characterized here also have relevance for climate change adaptation in field agriculture. Understanding how HI

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

genes respond to elevated CO₂ and heat stress is essential for both BLSS optimization and terrestrial crop improvement.

5.9.7 Collaborative Research Networks

Advancing these research directions will require collaboration across institutions and disciplines. Integration with existing pangenome consortia, partnership with space agencies, engagement with the CEA industry, and interdisciplinary collaboration with life support systems engineers would accelerate progress.

5.9.8 Technology Development Requirements

The translation of genetic insights to improved BLSS cultivars requires parallel development of enabling technologies. High-throughput genotyping platforms capable of detecting PAVs and CNVs, gene editing delivery systems, automated phenotyping technologies, and bioinformatic pipelines for integrating pangenomic data require continued development.

5.10 Chapter Summary

This discussion chapter has interpreted the findings of the pangenomic analysis of harvest index determinants in four BLSS staple crops within the broader context of crop genetics, evolutionary biology, and space agriculture applications. The key conclusions can be summarized as follows:

The 30-gene HI catalog represents a high-confidence, comprehensive inventory of genetic determinants affecting biomass partitioning and yield allocation in rice, wheat, potato, and soybean. The functional distribution of genes across five categories—with dwarfing/GA signaling predominating—reflects both biological importance and the research investment patterns of the Green Revolution.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The ortholog conservation analysis reveals remarkable evolutionary convergence, with the GA signaling pathway showing 100% conservation across phylogenetically diverse crops spanning the monocot-dicot divide. The 75% conservation of developmental timing and grain size regulation pathways indicates substantial but not universal conservation.

The GWAS and QTL findings validate the gene catalog while revealing the genomic architecture of HI variation. The 20 GWAS hits and 25 QTLs, including 7 highly significant associations ($p < 1e-5$), provide robust genotype-phenotype relationships suitable for molecular breeding applications.

The implications for BLSS crop improvement are substantial and actionable. The prioritized gene targets (Table 5.2) provide a roadmap for NASA crop development programs, with GA signaling genes (*sd1*, *Rht*) and grain size regulators (*GW2*) recommended as immediate priorities.

The theoretical contributions include advancement of pangenomic understanding of complex traits, development of a three-level analytical framework for comparative HI analysis, and demonstration of evolutionary convergence in HI determinants. The practical applications encompass marker-assisted selection targets, gene editing candidates, and genomic selection integration strategies.

In conclusion, this discussion has demonstrated that the pangenomic architecture of harvest index determinants shows remarkable conservation across diverse BLSS crops, providing opportunities for cross-species knowledge transfer and coordinated breeding strategies. The integration of gene catalogs, ortholog analysis, and GWAS/QTL data has yielded actionable recommendations for BLSS crop improvement that can guide NASA and international space

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

agriculture programs. The findings support the dissertation's central hypothesis that pangenomic approaches reveal conserved genetic mechanisms underlying harvest index that can be leveraged for optimizing crop performance in controlled-environment agriculture for space applications.

The significance threshold distribution of GWAS hits reveals a characteristic pattern where a small number of highly significant associations ($p < 1e-8$) represent major-effect genes (sd1, Rht-B1b), while a larger number of moderately significant associations reflect the polygenic nature of HI determination. This pattern is consistent with the theoretical expectation for complex quantitative traits, where genetic architecture involves a mixture of major genes with large effects detectable at stringent thresholds and minor genes with small effects that contribute cumulatively to phenotypic variation (Yang et al., 2017). Understanding this mixed-effect architecture has important implications for breeding strategy selection.

The concordance between GWAS and QTL findings provides reciprocal validation that strengthens confidence in the identified associations. The 25 QTLs include both positionally cloned genes (where the causal variant is definitively established) and mapped intervals (where the causal gene remains to be identified). For positionally cloned genes like GW2, GS3, and sd1, the GWAS associations provide population-level validation of effects demonstrated through functional studies. For mapped QTL intervals without identified genes, GWAS data can guide fine mapping efforts by indicating which SNPs show the strongest associations within the interval. This integration of complementary mapping approaches exemplifies the power of multi-method genetic analysis.

The species distribution of associations shows balanced representation, with rice and wheat contributing more GWAS hits (6 each) than potato and soybean (4 each). This distribution

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

reflects both biological factors (different genetic architectures) and research intensity factors (more extensive GWAS studies in cereals). The balanced QTL distribution across species (7, 7, 5, 6) indicates that linkage mapping has been applied more uniformly across crops, potentially identifying associations missed by GWAS in species with smaller GWAS panels. Future pangenomic GWAS studies with larger sample sizes may reveal additional associations in potato and soybean comparable to those detected in cereals.

The functional implications of the documented conservation patterns extend beyond crop improvement to fundamental questions in plant biology. The convergent selection on GA signaling across independent domestication events suggests that this pathway represents an evolutionary 'path of least resistance' for achieving reduced plant height. Alternative pathways for height reduction—such as brassinosteroid signaling or auxin transport—are less commonly exploited, potentially because they have more severe pleiotropic effects or because natural variation at these loci was less accessible in founder populations. Understanding why certain pathways are preferentially targeted during crop improvement may reveal general principles governing the evolution of complex traits under artificial selection.

The pangenomic perspective advanced in this dissertation has implications beyond harvest index to the broader challenge of understanding complex trait architecture. The finding that both core genome (conserved orthologous genes) and dispensable genome (structural variants) contribute to HI variation suggests that comprehensive trait mapping requires integration of multiple data types. SNP-based GWAS, while powerful and cost-effective, captures only a fraction of functional variation. The addition of structural variant genotyping through pangenomic approaches substantially increases the proportion of heritability explained, as demonstrated in recent studies of tomato and rice (Gao et al., 2022; Wang et al., 2025). This

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

'missing heritability' that is recovered through pangenomic analysis likely represents an important source of variation for many agronomic traits beyond HI.

The controlled-environment agriculture (CEA) context of BLSS introduces unique considerations for HI optimization that differ from field agriculture. Under CEA conditions, many constraints that drive yield variation in the field—water limitation, nutrient deficiency, temperature extremes, pest and disease pressure—are eliminated or minimized through precise environmental control. This environmental homogeneity may reduce the importance of stress-responsive genes while increasing the relative contribution of genes controlling fundamental developmental and physiological processes. The HI gene catalog compiled in this dissertation, which emphasizes well-characterized genes with documented effects on core processes (architecture, timing, carbon partitioning), may be particularly relevant for CEA applications where stress responsiveness is less critical.

The recommendation to develop 'BLSS-optimized' cultivars distinct from field varieties merits further elaboration. Current crop varieties have been selected for field performance, where they must balance productivity with resilience to environmental variability. This balance involves genetic trade-offs—alleles that increase yield potential under optimal conditions may compromise performance under stress. In the controlled environment of BLSS, these trade-offs become irrelevant, potentially enabling more extreme phenotypes than would be viable in the field. For example, complete knockouts of grain size regulators that might compromise stress tolerance in the field could be deployed in BLSS cultivars where stress is controlled. Similarly, photoperiod-insensitive mutants that would be poorly adapted to specific field environments could be advantageous under artificial lighting regimes.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The integration of the present findings with the broader BLSS design literature reveals opportunities for co-optimization of crop genetics and life support system engineering. Harvest index improvement directly affects several BLSS parameters: edible biomass output per unit volume, nutrient recycling efficiency, and atmospheric gas exchange requirements. Higher HI varieties produce more food per unit of plant biomass, reducing the plant growth area required and simplifying waste processing systems that must handle non-edible biomass. These system-level benefits amplify the value of HI improvement beyond the direct increase in food production, potentially justifying substantial investment in HI-optimized cultivar development.

The evolutionary perspective developed in this dissertation—emphasizing convergent selection on conserved pathways—provides a framework for predicting HI genes in crops not yet characterized at the genomic level. The 100% conservation of GA signaling suggests that any crop targeted for HI improvement will likely harbor variation at GA pathway genes that can be exploited through similar strategies. This predictive capability has practical value for emerging BLSS crops (e.g., quinoa, amaranth) where genomic resources are limited. By targeting homologs of the conserved HI genes identified in the four staple crops, researchers can focus initial characterization efforts on the most likely candidates for yield improvement.

The limitations acknowledged in Section 5.8 point to specific gaps that future research should address. The reliance on secondary data, while enabling comprehensive synthesis, means that the catalog may lag behind the most recent gene characterizations. Establishing automated pipelines connecting publication databases to the HI gene catalog would enable continuous updating as new findings emerge. The lack of CEA validation is a critical gap that directly affects the applicability of recommendations to BLSS; addressing this gap requires investment in

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

controlled-environment phenotyping infrastructure and systematic evaluation of HI gene effects under BLSS-relevant conditions.

The future research directions outlined in Section 5.9 are prioritized based on their potential impact and feasibility. Experimental validation under BLSS conditions is the highest priority because it directly addresses the critical knowledge gap affecting translation of genetic findings to space agriculture. Expanded species coverage would broaden the utility of the framework but requires substantial additional data compilation. Functional genomics studies and multi-omics integration represent longer-term investments that would deepen mechanistic understanding. Space environment testing, while providing definitive validation, is constrained by limited access to space-based research platforms and represents a final validation step rather than a near-term priority.

The collaborative research networks recommended in Section 5.9.7 reflect the interdisciplinary nature of BLSS crop optimization. No single research group or institution possesses all the expertise required—molecular genetics, plant physiology, controlled-environment engineering, life support systems design, and space biology all contribute essential perspectives. The success of previous large-scale collaborative efforts in crop genetics (e.g., the 3K Rice Genomes Project, the 10+ Wheat Genomes Project) demonstrates the feasibility and value of coordinated international research. Applying this collaborative model to BLSS-specific crop improvement could accelerate progress substantially beyond what individual programs could achieve.

The technology development requirements identified in Section 5.9.8 highlight dependencies between genetic improvement and enabling technologies. High-throughput

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

genotyping for structural variants currently lags behind SNP genotyping in cost and accessibility, limiting the implementation of pangenome-informed selection. Gene editing delivery systems, while transformative in model species, remain inefficient in some elite crop genotypes that are recalcitrant to transformation. Automated phenotyping for HI under CEA conditions is not standardized, making it difficult to compare results across studies. Addressing these technology gaps would multiply the impact of the genetic insights generated in this dissertation.

The practical applications section (5.7) emphasizes the translation of genetic knowledge to breeding tools. The distinction between marker-assisted selection (MAS), gene editing, and genomic selection reflects different stages in the breeding process and different requirements for implementation. MAS is immediately applicable for well-characterized genes with diagnostic markers and can be implemented by any breeding program with basic molecular biology capabilities. Gene editing requires more sophisticated infrastructure but enables creation of novel alleles not present in germplasm. Genomic selection requires substantial training data but offers gains for polygenic traits beyond what is achievable through individual gene selection. A comprehensive BLSS breeding program would employ all three approaches in a integrated strategy targeting different components of the HI genetic architecture.

The tables presented in this chapter synthesize complex information into actionable formats. Table 5.1 summarizes conservation patterns and their breeding implications, providing a quick reference for translating evolutionary insights to practical recommendations. Table 5.2 prioritizes gene targets based on multiple criteria, enabling rational allocation of research and breeding resources. Table 5.3 stratifies GWAS findings by significance level, indicating which associations warrant immediate implementation versus further validation. These tables are

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

designed to serve as resources for breeding program planning and research priority setting beyond this dissertation.

The figures in this chapter are designed to communicate key conceptual relationships visually. Figure 5.1 illustrates the pangenomic variation model, showing how different types of variants (SNPs, PAVs, CNVs) contribute to HI phenotype and why pangenomic approaches capture variation missed by reference-based methods. Figure 5.2 presents the analytical framework as a flowchart, showing the iterative process of gene characterization, ortholog analysis, and structural variant integration. These visual aids complement the text by providing alternative representations suitable for different learning styles and communication contexts.

In synthesizing the discussion, this chapter has moved from specific findings (30 genes, 5 categories, 100% GA signaling conservation) to broader interpretations (evolutionary convergence, practical implications) to concrete recommendations (prioritized targets, NASA program guidance). This progression from data to interpretation to application exemplifies the translational potential of fundamental genomic research. The findings have value at multiple levels: contributing to basic understanding of yield genetics, informing conventional breeding programs, guiding gene editing applications, and supporting the specific challenge of BLSS crop optimization.

The conclusion emphasizes that pangenomic approaches reveal conserved genetic mechanisms underlying harvest index that can be leveraged for crop improvement. This conservation is both an opportunity and a validation. It is an opportunity because knowledge transfer across species is facilitated when underlying mechanisms are shared. It is a validation because consistent findings across independent studies and species strengthen confidence in the

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

biological importance of identified genes. The GA signaling pathway exemplifies both aspects—its universal importance validates decades of Green Revolution research while enabling confident application to less-characterized crops and novel environments like BLSS.

The dissertation's central hypothesis—that pangenomic analysis would reveal conserved mechanisms and novel variation affecting harvest index—is supported by the findings presented. The conservation is demonstrated by the ortholog analysis showing 100%, 75%, 75%, 50%, and 50% conservation for the five HI-regulatory pathways. The novel variation is suggested by the structural variant data indicating extensive PAV and CNV at HI loci, though direct functional validation of these variants awaits future experimental studies. Together, these findings establish a foundation for pangenomics-informed HI improvement in BLSS staple crops.

The implications of this research extend beyond the immediate application to BLSS. The comparative pangenomic framework developed here is applicable to any complex trait in any set of crops. The finding that evolutionary convergence characterizes HI genetics suggests similar patterns may exist for other domestication-related traits. The integration of multiple data types (genes, orthologs, GWAS, QTL, structural variants) into a coherent analytical framework provides a template for comprehensive trait characterization. As pangenomic resources become available for additional species, this framework can be applied to systematically characterize the genetic architecture of traits important for food security, environmental sustainability, and human nutrition.

The chapter closes with acknowledgment that genetic improvement is only one component of BLSS crop optimization. Agronomic practices, environmental management, and system design also contribute substantially to crop productivity. Genetic gains are most

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

effectively realized when combined with optimized growing conditions that enable full expression of genetic potential. The recommendations in this chapter are therefore intended to complement, rather than replace, the engineering and physiological research that constitutes the broader BLSS crop production endeavor. An integrated approach combining the genetic insights presented here with advances in controlled-environment technology will be most effective in achieving the ultimate goal of sustainable, regenerative food production for long-duration space missions.

The relationship between the present findings and the theoretical frameworks discussed in Chapter 2 deserves explicit consideration. The source-sink paradigm, which posits that yield is limited by either photosynthetic capacity (source) or the ability of reproductive structures to accumulate assimilates (sink), provides a physiological interpretation for the HI gene categories identified. Dwarfing genes alter source-sink partitioning by reducing the stem as a competing sink, thereby directing more assimilates to reproductive structures. Grain size genes directly enhance sink capacity by increasing the number or size of cells available to accumulate starch and protein. Developmental timing genes optimize the coordination between source activity and sink demand by controlling when reproductive development occurs relative to vegetative growth. This alignment between gene function and physiological framework validates the comprehensiveness of the catalog and suggests that additional HI genes yet to be discovered will likely function through one of these mechanisms.

The distinction between qualitative (major gene) and quantitative (polygenic) inheritance models is relevant to interpreting the HI gene catalog. Classical quantitative genetics assumed that complex traits like yield arise from the cumulative effects of many genes, each with small individual effects (Fisher, 1918). The detection of major-effect genes like *sd1*, *Rht*, and *GW2*

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

challenges this purely infinitesimal model, demonstrating that some loci have effects large enough to be detected and manipulated individually. However, the presence of many additional loci with smaller effects—evident in the GWAS hits at significance thresholds below $1e-5$ —confirms that HI has a substantial polygenic component beyond these major genes. Modern quantitative genetic models accommodate both major genes and polygenic background, and the HI genetic architecture documented here is consistent with this mixed-inheritance framework.

The Green Revolution's impact on global food security provides historical context for understanding the potential of HI improvement. The introduction of semi-dwarf wheat and rice varieties between 1960 and 1980 is estimated to have increased global cereal production by 50-100%, averting predicted famines and supporting population growth that would otherwise have outstripped food supply (Evenson & Gollin, 2003). This impact was achieved through genetic modification of a single pathway—GA signaling—demonstrating the transformative potential of targeted genetic improvement. The present dissertation identifies additional pathways (grain size, developmental timing, starch biosynthesis, sucrose transport) that offer opportunities for similar impact, particularly in the novel context of BLSS where some field constraints are relaxed while new constraints (volume, energy, closure) become paramount.

The concept of genetic potential versus realized yield is relevant to BLSS applications. In field agriculture, realized yield is often substantially below genetic potential due to environmental limitations and pest/disease pressure. The gap between potential and realized yield varies by region and management intensity but is estimated at 50% or more for many crops in many environments (Lobell et al., 2009). In BLSS, where environmental control is precise and biological threats can be excluded, the gap between potential and realized yield should be minimized. This suggests that genetic potential—the maximum yield achievable under optimal

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

conditions—becomes the limiting factor, elevating the importance of HI improvement. The genes cataloged in this dissertation directly affect genetic potential by modifying the allocation of biomass to harvestable organs, making them particularly relevant for BLSS yield optimization.

The hierarchical organization of the HI gene catalog—from genes to pathways to genome-wide patterns—reflects different scales of genetic architecture. At the gene level, individual nucleotide changes or structural variants create phenotypic differences between genotypes. At the pathway level, multiple genes in the same biological process (e.g., GA signaling) show coordinated effects that can be interpreted mechanistically. At the genome-wide level, the distribution of variants across chromosomes and the correlation structure among loci affect the efficiency of selection and the potential for favorable allele combinations. Understanding all three levels is essential for developing effective breeding strategies. MAS operates primarily at the gene level, targeting known variants at specific loci. Genomic selection incorporates genome-wide information, capturing both major genes and polygenic effects. Pathway-based approaches, which target multiple genes in the same process, represent an intermediate strategy that may be particularly powerful for coordinated pathway modification through multiplex gene editing.

The phenotypic plasticity of HI—its responsiveness to environmental variation—has implications for BLSS breeding. Some HI variation is genetically fixed, determined by allelic state at HI loci regardless of environment. Other HI variation is environmentally induced, with plants adjusting biomass partitioning in response to conditions. A third component represents genotype-by-environment interaction ($G \times E$), where the effect of a genetic variant depends on the environment in which it is expressed. For BLSS applications, genetically fixed HI determinants

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

are most valuable because they provide consistent performance across the controlled but potentially variable conditions of space habitats. G×E interactions are concerning because they may cause HI genes to behave differently in BLSS than in the field environments where they were characterized. Characterizing G×E for the prioritized HI genes under BLSS conditions is therefore a critical research need before confident selection recommendations can be made.

The relationship between harvest index and total biomass production merits discussion. HI is defined as the ratio of harvestable yield to total biomass, so it can be increased either by increasing the numerator (more grain) or decreasing the denominator (less total biomass). The Green Revolution dwarfing genes primarily decreased the denominator by reducing stem biomass, while grain yield per se was maintained or slightly increased. This distinction has implications for BLSS, where total biomass has value beyond being a divisor in the HI equation. Non-edible biomass contributes to oxygen production through photosynthesis, supports waste recycling through processing and composting, and may provide raw material for other purposes. A BLSS-optimized HI improvement strategy should therefore focus on increasing the numerator (grain yield) while maintaining or increasing total photosynthesis, rather than simply reducing vegetative biomass. This favors grain size genes over dwarfing genes for some applications, though the volume constraints of BLSS may still favor compact plant architecture regardless of photosynthetic considerations.

The timing of HI improvement efforts relative to other BLSS development activities affects their optimal prioritization. Early-stage BLSS development, focused on demonstrating basic feasibility and developing life support system concepts, may not require highly optimized crop cultivars. Standard field varieties with acceptable HI can support initial experiments while BLSS engineering matures. As BLSS designs advance toward operational deployment, however,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

the efficiency gains from HI-optimized cultivars become increasingly valuable, justifying greater investment in cultivar development. This suggests a phased approach where preliminary genetic characterization (the present dissertation) precedes intensive cultivar development, which in turn precedes operational deployment. The prioritization in Table 5.2 reflects this phasing, identifying genes for immediate implementation (MAS-ready) versus those requiring further development (editing targets) versus those for longer-term research (functional genomics priorities).

The economic considerations relevant to commercial agriculture (cost of seed, return on breeding investment) are less constraining for BLSS applications, where the primary currency is system performance rather than market profit. This relaxation of economic constraints enables consideration of breeding strategies that would be impractical for commercial release. For example, gene-edited cultivars face regulatory barriers in many markets that increase costs and delay deployment, but these barriers are irrelevant for BLSS cultivars that will never enter commercial channels. Similarly, the maintenance of multiple HI-optimized cultivars for different BLSS configurations is justified by mission requirements rather than market demand. These considerations suggest that BLSS crop development can pursue more aggressive genetic modification strategies than would be commercially viable, potentially achieving performance levels not represented in field agriculture.

The integration of the present findings with the literature on crop ideotypes provides additional perspective. The crop ideotype concept, introduced by Donald (1968), describes the morphological and physiological characteristics of an ideal plant for a given production system. The 'Green Revolution ideotype' for cereals featured short stature, erect leaves, high harvest index, and responsiveness to nitrogen fertilization. A 'BLSS ideotype' might differ in emphasizing extremely compact growth habit, rapid development, efficient resource use under

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

elevated CO₂, and minimal non-edible biomass. The HI genes cataloged in this dissertation provide the genetic tools for constructing such ideotypes, with dwarfing genes controlling architecture, developmental timing genes controlling phenology, and grain size genes controlling yield components. Future research could explicitly define BLSS ideotypes and identify the optimal allelic combinations at HI loci for achieving them.

The concept of breeding targets versus breeding constraints is useful for interpreting the HI gene catalog. Breeding targets are traits that breeders actively seek to improve, while breeding constraints are traits that must be maintained above minimum thresholds to ensure overall cultivar acceptability. For field agriculture, HI is typically a breeding target, while traits like disease resistance, grain quality, and stress tolerance are constraints that limit the extent of HI improvement. For BLSS, the relative importance of different constraints shifts—disease resistance becomes less critical (controlled environment), while resource use efficiency becomes more constraining. This shift in the constraint landscape may enable more aggressive HI improvement than is possible in field breeding, where constraints often compete with HI for selection intensity.

The genetic correlations among yield components affect the achievable limits of HI improvement. If grain number and grain size are negatively correlated (as observed in many crops), increasing one component decreases the other, limiting the maximum combined effect. If HI-related genes have pleiotropic effects on quality or stress tolerance, improving HI may compromise these correlated traits. Understanding the genetic correlation structure among yield components and HI is essential for predicting the response to selection and avoiding unintended consequences. The GWAS and QTL data compiled in this dissertation provide a foundation for

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

estimating these correlations, though direct experimental measurement in populations segregating for HI alleles would provide more precise estimates.

The role of epigenetic variation in HI determination is not addressed in the present study but may be relevant for future research. Epigenetic marks (DNA methylation, histone modification) can affect gene expression without altering DNA sequence, and epigenetic variation has been associated with agronomic traits in several crops. If epigenetic variation at HI loci contributes to phenotypic differences, the gene catalog presented here would be incomplete. However, epigenetic variation is generally more environmentally sensitive and less stably inherited than sequence variation, potentially making it less suitable for breeding applications that require consistent performance across generations. Future panepigenomic studies could characterize epigenetic variation at HI loci and assess its contribution to HI phenotype.

The potential for synthetic biology approaches to HI improvement extends beyond the gene editing applications discussed in Section 5.7.2. Synthetic biology enables the creation of novel gene circuits, regulatory networks, and metabolic pathways not found in nature. For HI improvement, synthetic approaches could include: engineering of artificial transcription factors that precisely control HI gene expression; design of synthetic signaling pathways that bypass natural regulatory constraints on grain filling; and construction of novel carbon-partitioning pathways that direct more photosynthate to reproductive structures. These approaches are more speculative than conventional breeding or gene editing but represent potential long-term opportunities for HI improvement beyond the limits imposed by natural variation.

The consideration of food safety and nutritional quality in HI-optimized BLSS crops ensures that yield improvement does not compromise the ultimate purpose of crop production—

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

nourishing astronaut crews. If HI improvement genes affect grain composition, the nutritional implications must be evaluated. For example, some grain size genes affect the ratio of starch to protein in grains, potentially altering the amino acid profile available to consumers. Similarly, starch biosynthesis genes directly determine starch structure, affecting digestibility and glycemic response. A comprehensive BLSS crop development program should include nutritional analysis alongside yield evaluation to ensure that HI-optimized cultivars meet dietary requirements.

The redundancy and resilience considerations for BLSS crop systems affect the optimal genetic strategy. A BLSS dependent on a single high-yielding variety is vulnerable to catastrophic failure if that variety encounters unexpected problems (disease, environmental stress, equipment malfunction). Maintaining genetic diversity within BLSS crop populations provides resilience against such failures but may reduce average performance. The HI gene catalog provides a framework for understanding and managing this trade-off: core genes (present in all high-HI varieties) can be fixed without compromising diversity, while dispensable genes (contributing to variety-specific adaptation) can be maintained as a diversity reserve. This stratified approach to genetic diversity management optimizes both performance and resilience in BLSS crop systems.

In reflecting on the contribution of this dissertation to the broader scientific enterprise, the work represents an integration of genomics, evolutionary biology, and applied crop science in service of a specific technological challenge. The pangenomic framework developed here is not limited to BLSS applications but provides a template for comprehensive trait characterization applicable to diverse goals. The finding of conserved mechanisms underlying harvest index improvement across phylogenetically diverse crops suggests similar conservation may characterize other complex traits, motivating comparative pangenomic analyses in additional

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

trait systems. The practical outputs—gene catalogs, prioritized targets, breeding recommendations—have immediate utility while the theoretical contributions—evolutionary convergence, framework development—advance fundamental understanding. This dual contribution exemplifies the potential of application-oriented basic research to generate both practical tools and scientific insights.

CHAPTER 6: CONCLUSION

6.1 Introduction

This concluding chapter synthesizes the principal findings, contributions, and implications of the dissertation research on the pangenomic architecture of harvest index (HI) determinants across four Bioregenerative Life Support System (BLSS) staple crops: rice (*Oryza sativa*), wheat (*Triticum aestivum*), potato (*Solanum tuberosum*), and soybean (*Glycine max*). The purpose of this chapter is threefold: (a) to recapitulate the research objectives and summarize how they were addressed through the systematic multi-phase investigation, (b) to articulate the theoretical, methodological, and practical contributions of this work to the fields of plant genomics, crop breeding, and space agriculture, and (c) to offer recommendations for future research and application that will advance the goal of sustainable food production in extraterrestrial environments.

The overarching aim of this dissertation was to construct a comprehensive understanding of the genetic architecture controlling harvest index in BLSS-relevant crops through the application of pangenomic analytical approaches. This aim was operationalized through four specific research objectives: (1) cataloging genes with documented effects on harvest index and related yield components across the four target species, (2) identifying orthologous relationships to elucidate conserved genetic mechanisms, (3) characterizing pangenomic structural variants associated with HI phenotypes through integration of GWAS and QTL data, and (4) establishing a conceptual and methodological framework for future pangenomic HI research applicable to controlled-environment agriculture and space applications.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The research presented in this dissertation represents a significant step toward understanding the fundamental genetic mechanisms that determine biomass partitioning in crop plants—knowledge that is essential for optimizing food production efficiency in the resource-constrained environment of space habitats. As humanity prepares for extended missions to the Moon, Mars, and beyond, the ability to produce nutritious food with maximum efficiency becomes not merely desirable but essential for mission success and crew well-being. The findings presented herein provide a scientific foundation and actionable targets for achieving this goal.

6.2 Summary of Research

6.2.1 Research Questions Revisited

This dissertation was guided by four primary research questions that structured the investigation and analysis:

Research Question 1: What genes have been documented to influence harvest index and related yield components in rice, wheat, potato, and soybean, and how can they be systematically classified? The investigation successfully identified 30 HI-associated genes distributed across the four BLSS staple crops, with a mean of 7.5 genes per species ($SD = 1.0$). These genes were classified into five functional categories based on their molecular mechanisms and phenotypic effects: Dwarfing/GA signaling ($n = 10, 33.3\%$), Grain size regulation ($n = 7, 23.3\%$), Developmental timing ($n = 7, 23.3\%$), Starch biosynthesis ($n = 3, 10.0\%$), and Sucrose transport ($n = 3, 10.0\%$). The systematic classification provides a structured framework for understanding the physiological hierarchy controlling harvest index.

Research Question 2: What orthologous relationships exist among HI-associated genes across these phylogenetically diverse species, and what do these relationships reveal about

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

conserved genetic mechanisms? The ortholog analysis revealed remarkable conservation of HI-regulatory pathways across the monocot-dicot divide, with the gibberellin signaling pathway showing 100% conservation across all four species. This finding provides compelling evidence that the core genetic mechanisms controlling harvest index represent deeply conserved developmental and physiological processes. The remaining four pathways (grain size regulation, developmental timing, starch biosynthesis, and sucrose transport) demonstrated conservation levels ranging from 50% to 75%, indicating substantial but not universal pathway conservation.

Research Question 3: What pangenomic structural variants and phenotype-genotype associations characterize HI-related loci, and how do these findings inform our understanding of HI genetic architecture? The integration of GWAS and QTL mapping data identified 20 significant GWAS associations and 25 QTLs related to harvest index and component traits. The GWAS findings confirmed the major-effect status of key loci, including the *sd1* region in rice and the *Rht-B1b* region in wheat. The QTL analysis provided independent validation of these associations while revealing the polygenic architecture underlying HI variation. Pangenomic structural variant analysis documented extensive presence-absence variation (PAV) and copy number variation (CNV) that may harbor additional functional alleles not captured by SNP-based approaches.

Research Question 4: How can this knowledge be integrated into a framework for accelerating HI improvement in BLSS crops for space agriculture applications? The synthesis of findings from the three analytical phases enabled the development of a prioritized list of gene targets for BLSS crop improvement, with recommendations stratified by effect magnitude, conservation level, and feasibility of genetic manipulation. The framework integrates gene-level

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

characterization, pathway-level analysis, and genome-level evaluation of structural variation into a coherent approach applicable across crop species.

6.2.2 Methodological Approach Summary

The research employed a four-phase analytical approach designed to systematically address the research questions while maximizing integration of diverse data sources. Phase 1 (HI Gene Inventory and Classification) drew from peer-reviewed literature and curated genomic databases—including GWAS Atlas, Gramene, SNP-Seek, Triticeae Toolbox, SoyBase, and Spud DB—to construct a comprehensive gene catalog with functional annotations. Phase 2 (Ortholog Identification and Phylogenomic Analysis) utilized OrthoDB and Ensembl Compara databases to identify orthologous relationships and assess pathway conservation across species. Phase 3 (Structural Variant and Association Analysis) integrated pangenomic data from species-specific pangenome projects with GWAS and QTL findings to characterize the genomic architecture of HI variation. Phase 4 (Evolutionary Pattern Analysis) synthesized cross-phase findings to elucidate the evolutionary forces shaping HI regulation.

This multi-phase approach enabled triangulation of findings across complementary methodologies and data sources, strengthening confidence in the conclusions while revealing both conserved principles and species-specific adaptations. The systematic integration of data from multiple genomic databases ensured comprehensive coverage while providing cross-validation of key findings.

6.2.3 Key Findings Overview

The principal findings of this dissertation can be summarized across six major domains: (1) gene catalog composition and functional distribution, (2) species-specific gene inventories, (3) ortholog conservation patterns, (4) GWAS and QTL validation, (5) pangenomic structural

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

variation, and (6) priority targets for BLSS improvement. Table 6.1 presents a comprehensive summary of these findings organized by research phase.

Table 6.1 Summary of Major Findings by Research Phase

Research Phase	Key Finding	Significance
Phase 1: Gene Inventory	30 HI-associated genes identified across 4 crops (Rice: 7, Wheat: 9, Potato: 7, Soybean: 7); 5 functional categories established	Provides comprehensive, actionable catalog for breeding; Dwarfing/GA signaling largest category (33.3%)
Phase 2: Ortholog Analysis	5 conserved ortholog pathways identified; GA signaling 100% conserved; other pathways 50-75% conserved	Demonstrates deep evolutionary conservation of HI mechanisms across monocot-dicot divide
Phase 3: Association Analysis	20 GWAS hits and 25 QTLs documented; major-effect loci validated (sd1, Rht-B1b, GW2, StSP6A)	Confirms genetic architecture and provides targets for marker-assisted selection
Phase 4: Integration	Framework for pangenomic HI research developed; priority targets identified for BLSS improvement	Enables systematic application to space agriculture crop development

Note. HI = harvest index; GA = gibberellic acid; GWAS = genome-wide association study; QTL = quantitative trait locus; BLSS = Bioregenerative Life Support System. Findings represent synthesis of multi-database analysis including GWAS Atlas, Gramene, Ensembl Plants, and species-specific genomic resources.

The identification of 30 HI-associated genes with documented effects on harvest index and related phenotypes represents a substantial contribution to the understanding of yield component genetics in these economically and strategically important crops. The relatively uniform distribution across species (mean = 7.5, SD = 1.0) suggests that the core genetic mechanisms controlling harvest index are conserved despite substantial differences in genome

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

size, ploidy level, and breeding history. Wheat's elevated gene count ($n = 9$) reflects its allohexaploid genome structure and the intensive research effort directed toward this globally important crop.

The functional category distribution—with Dwarfing/GA signaling representing 33.3% of cataloged genes—aligns with the historical importance of plant architecture modification demonstrated during the Green Revolution. The introduction of semi-dwarf genes in wheat and rice increased harvest index from approximately 0.30 to over 0.50, representing the most significant single intervention in yield improvement history. The catalog's emphasis on this category reflects both biological reality and the cumulative research investment in understanding these transformative genetic modifications.

6.3 Contributions to Knowledge

6.3.1 Theoretical Contributions

Pangenomic Framework for Complex Trait Analysis.

This dissertation advances the theoretical application of pangenomic approaches to complex quantitative traits. While pangenomics has been predominantly applied to stress tolerance, disease resistance, and environmental adaptation, its systematic application to agronomic traits like harvest index has been more limited. The present study demonstrates that pangenomic perspectives enhance understanding of HI genetics through several mechanisms: (a) revealing pathway-level conservation obscured when species are analyzed individually, (b) contextualizing SNP-based GWAS findings within the broader structural variation landscape, and (c) identifying potential novel alleles in presence-absence variants and copy number variants that may be exploited for breeding.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The 100% conservation of GA signaling genes across four phylogenetically diverse crops provides stronger evidence for the universality of this HI mechanism than would be obtained from any single-species study. This finding supports the theoretical proposition that harvest index improvement targets represent deeply conserved developmental and physiological processes, with modification of these pathways during domestication and breeding representing canalization of standing variation rather than *de novo* innovation.

Cross-Species Comparative Genomics Insights.

The comparative genomic analysis provides novel insights into the evolutionary dynamics of HI determination across the monocot-dicot divide. The documentation of mechanistic diversity within pathway conservation—where rice *sd1* represents a biosynthesis mutation while wheat *Rht* genes encode signaling component mutations—illustrates how homologous pathways can be modified at different nodes to achieve similar phenotypic outcomes. This mechanistic flexibility has important implications for breeding strategy, suggesting that multiple genetic routes exist to harvest index improvement within each species.

Furthermore, the conservation patterns documented in this study have broader implications for understanding crop domestication and the genetic basis of the domestication syndrome. The convergent selection on GA signaling across independently domesticated crops suggests that plant height reduction was a common target during separate domestication events in different geographic centers, providing molecular evidence for convergent evolution at the pathway level.

6.3.2 Methodological Contributions

Four-Phase Analytical Approach.

The four-phase analytical framework developed in this dissertation—encompassing gene inventory, ortholog analysis, structural variant characterization, and evolutionary pattern synthesis—provides a replicable methodology for investigating the genetic architecture of complex traits across multiple species. This approach addresses a significant gap in the literature, where most pangenomic studies focus on single species or single pathways rather than systematically integrating findings across taxonomic boundaries.

The methodology's strength lies in its iterative integration of findings across phases, where gene catalog composition informs ortholog searches, ortholog patterns guide structural variant analysis, and association data validate both gene function and evolutionary interpretations. This triangulation approach enhances confidence in conclusions while revealing both universal principles and species-specific adaptations that might be overlooked in single-phase analyses.

Integration of Multiple Data Types.

A key methodological contribution is the demonstration of effective integration across diverse data types: curated gene annotations, ortholog databases, pangenomic structural variants, GWAS associations, and QTL mapping results. The successful synthesis of data from GWAS Atlas (containing over 200,000 trait-marker associations), Gramene, Ensembl Plants, and species-specific repositories (SNP-Seek, Triticeae Toolbox, SoyBase, Spud DB) provides a template for multi-database integration in crop genomics research.

The concordance between GWAS hits and QTL positions at key loci (sd1, Rht-B1b, GW2, StSP6A) demonstrates the value of integrating association mapping and linkage mapping

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

approaches, providing reciprocal validation that strengthens confidence in the biological relevance of identified associations.

6.3.3 Practical Contributions

HI Gene Catalog for BLSS Crops.

The 30-gene HI catalog represents an immediately actionable resource for breeding programs targeting harvest index improvement. The catalog's organization by functional category facilitates strategic deployment—breeders can target dwarfing genes for architecture modification, grain size genes for sink capacity enhancement, and developmental timing genes for optimizing growth phase duration. The documentation of specific allelic variants, chromosomal locations, and phenotypic effects provides the detail necessary for marker-assisted selection implementation.

Priority Targets for Space Agriculture.

The integration of effect magnitude, conservation level, GWAS/QTL support, and BLSS relevance enables prioritization of genetic targets specifically for space agriculture applications. The highest-priority targets—*sd1/Rht* for semi-dwarf phenotypes, *GW2/TaGW2* for grain size optimization, and *StSP6A* for tuberization control—are recommended for immediate implementation in NASA's Advanced Exploration Systems crop development programs. These recommendations translate the dissertation's scientific findings into practical guidance for mission-critical crop improvement.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table 6.2 Summary of Dissertation Contributions by Domain

Contribution Domain	Specific Contribution	Beneficiary/Application
Theoretical	Pangenomic framework for complex trait analysis; Cross-species conservation patterns; Evolutionary convergence evidence	Plant genomics research community; Evolutionary biology scholarship
Methodological	Four-phase analytical approach; Multi-database integration methodology; Ortholog-guided comparative analysis	Future pangenomic studies; Crop genomics researchers; Bioinformatics methodology
Practical	30-gene HI catalog with functional annotations; Priority targets for BLSS improvement; MAS and CRISPR target recommendations	NASA crop development programs; Breeding programs; Space agriculture initiatives

Note. HI = harvest index; MAS = marker-assisted selection; BLSS = Bioregenerative Life Support System; CRISPR = Clustered Regularly Interspaced Short Palindromic Repeats. Contributions span fundamental knowledge generation to applied breeding recommendations.

6.4 Implications

6.4.1 For Plant Genomics Research

The findings of this dissertation carry substantial implications for the broader field of plant genomics research. The demonstration that pangenomic approaches can effectively characterize complex quantitative traits like harvest index suggests that similar methodologies could be applied to other yield-related traits, quality attributes, and stress tolerance mechanisms. The four-phase analytical framework provides a template that can be adapted for investigating the genetic architecture of diverse phenotypes across multiple species.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The documentation of 100% pathway conservation for GA signaling across the monocot-dicot divide reinforces the value of comparative genomics for understanding trait evolution. This finding suggests that knowledge gained from model or well-characterized species can be reliably translated to less-studied crops when targeting conserved pathways. However, the variable conservation observed for other pathways (50-75%) cautions that translation potential must be empirically assessed rather than assumed.

The identification of extensive structural variation at HI-related loci—including PAVs and CNVs—supports the growing recognition that reference-based genomic analyses capture only a fraction of functional genetic diversity. The pangenomic paradigm, which captures the full complement of genes across a species, should become standard practice for trait genetic studies to avoid missing novel alleles absent from reference genomes.

6.4.2 For Crop Breeding Programs

For applied breeding programs, this dissertation provides both specific gene targets and strategic frameworks for harvest index improvement. The prioritized gene list—with *sd1/Rht*, *GW2/TaGW2*, *GS3/GS5*, and *StSP6A* as highest-priority targets—offers immediate applications for marker-assisted selection in breeding populations. The functional markers available for these genes (or closely linked markers where causal variants are not fully characterized) enable efficient selection without requiring phenotypic evaluation of HI, which is typically resource-intensive.

The finding that multiple genes contribute to HI within each functional category supports the use of genomic selection approaches that can capture polygenic variation beyond major-effect loci. The GWAS and QTL data compiled in this study can inform development of prediction models with improved accuracy for HI and component traits. Multi-trait genomic

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

selection incorporating grain size, plant height, and developmental timing may be more effective than single-trait approaches given the physiological interconnections among these phenotypes.

The documentation of favorable allele combinations in elite germplasm provides haplotype targets for breeding. Where favorable alleles at multiple HI loci exist in different genetic backgrounds, strategic crossing and selection can pyramid these alleles. The ortholog information facilitates cross-species breeding strategy transfer—successful approaches for GA pathway modification in wheat can inform soybean improvement programs targeting homologous genes.

6.4.3 For Space Agriculture and BLSS Development

The implications of this research for space agriculture and Bioregenerative Life Support System development are both immediate and long-term. In the immediate term, the identification of specific gene targets enables NASA and partner agencies to initiate molecular breeding programs focused on HI optimization in BLSS candidate crops. The semi-dwarf genes (*sd1*, *Rht*) are particularly relevant given the volume constraints of space growth chambers—reduced plant height directly translates to either higher planting density or reduced chamber height requirements, both of which improve system efficiency.

For longer-term BLSS development, the framework established in this dissertation provides a roadmap for systematic crop optimization. The finding that harvest index improvement targets are evolutionarily conserved suggests that gains achieved in one BLSS crop are likely achievable in others through homologous gene modification. This conservation enables parallel improvement programs across the four staple crops essential for nutritionally complete diets: rice and wheat for carbohydrates, potato for carbohydrates and vitamin C, and soybean for protein and oil.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

The integration of the HI gene catalog with controlled-environment agriculture considerations reveals opportunities for BLSS-specific optimization. Under the elevated CO₂ concentrations typical of space habitats (1000-2000 ppm), enhanced photosynthetic capacity must be matched by increased sink strength to realize yield gains. The grain size genes (GW2, GS3, GW5) provide targets for enhancing sink capacity to exploit CO₂ enrichment. Similarly, the developmental timing genes enable optimization of growth cycle duration for mission-specific scheduling requirements.

6.4.4 For NASA's Advanced Life Support Program

The findings of this dissertation have direct relevance for NASA's Advanced Life Support (ALS) Program and its goal of achieving bioregenerative food production for extended space missions. Several specific implications merit highlighting:

First, the prioritized gene targets provide a science-based foundation for cultivar development decisions. NASA's Vegetable Production System (Veggie) and Advanced Plant Habitat (APH) programs can incorporate molecular marker screening to select germplasm with favorable HI allele combinations. This molecular approach complements empirical testing by enabling pre-screening of large germplasm collections before committing limited space-analog growth chamber resources.

Second, the finding that GA signaling modification is universally effective across crop species suggests that a common breeding strategy can be applied to the diverse crops required for BLSS nutrition. Rather than developing independent improvement programs for each crop, NASA can implement a coordinated approach targeting homologous pathways across species.

Third, the documentation of genotype-by-environment interaction effects in terrestrial studies implies that BLSS-specific cultivar development—testing and selecting under space-

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

relevant conditions—may be necessary to optimize performance. Cultivars bred for field agriculture may not express optimal phenotypes under continuous lighting, elevated CO₂, and other BLSS conditions. The gene catalog provides candidates for investigation of these GxE effects.

Fourth, the identification of gene editing targets supports the development of precisely modified cultivars through CRISPR technology. Gene editing offers advantages over conventional breeding for space agriculture applications, including faster timeline, precise allele substitution, and ability to generate alleles not present in natural germplasm. The well-characterized HI genes identified in this study are excellent candidates for editing-based improvement.

6.5 Recommendations

6.5.1 For Researchers

Based on the findings and limitations of this dissertation, the following recommendations are offered for researchers investigating harvest index and related traits:

First, expand pangenomic characterization of HI loci to capture structural variants not represented in current studies. The documentation of extensive PAV and CNV in all four crops suggests that substantial functional variation remains to be characterized. Deep sequencing of diverse germplasm, including landraces and wild relatives, may reveal novel alleles at priority HI loci.

Second, pursue functional validation of candidate HI genes, particularly those identified primarily through bioinformatic analysis. While the major-effect genes (*sd1*, *Rht*, *GW2*) have been extensively validated through forward genetics and transgenic studies, other catalog entries

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

would benefit from CRISPR-based knockout or overexpression experiments to confirm phenotypic effects and mechanism of action.

Third, investigate the environmental dependency of HI gene effects, particularly under controlled-environment conditions relevant to BLSS applications. The interaction between genotype and environment for HI-related phenotypes is well documented in field conditions but poorly characterized for indoor agriculture environments. Multi-environment trials comparing field, greenhouse, and growth chamber performance would reveal whether HI gene effects are stable across production systems.

Fourth, develop species-specific pangenome graphs incorporating HI gene regions at high resolution. Graph-based representations can capture structural complexity—including inversions, complex indels, and tandem duplications—that linear reference genomes fail to represent. High-resolution local assemblies at HI loci would enable more accurate genotyping and potentially reveal cryptic functional variants.

Fifth, apply machine learning approaches to integrate diverse data types for HI prediction. The multi-dimensional data compiled in this study—gene annotations, GWAS effects, QTL intervals, structural variants, expression patterns—could train predictive models with improved accuracy compared to single-data-type approaches. Such models could accelerate genomic selection for HI improvement.

6.5.2 For Breeders and Crop Scientists

For applied breeders and crop scientists working to improve harvest index, this dissertation supports the following recommendations:

First, implement marker-assisted selection for high-priority HI genes in breeding populations. The *sd1* and *Rht* genes should be fixed in their favorable alleles across all cereal

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

breeding material destined for BLSS applications, with verification through molecular markers. Similarly, favorable GW2 alleles should be incorporated where compatible with grain quality requirements. Functional markers that directly assay causal polymorphisms are preferred over linked markers to minimize linkage drag.

Second, develop and deploy CRISPR-based gene editing pipelines targeting prioritized HI genes. For knockout-based improvement, negative regulators of grain size (GW2, GS3 in some backgrounds) are ideal targets—loss-of-function increases grain size without known deleterious pleiotropic effects. For gain-of-function modifications, promoter editing can enhance expression of positive regulators. Base editing approaches can introduce specific favorable alleles identified through natural variation studies.

Third, conduct systematic HI phenotyping under BLSS-relevant conditions to identify superior genotypes. Phenotypic evaluation should include conditions relevant to space agriculture: elevated CO₂ (1000-2000 ppm), continuous or extended photoperiods, hydroponic or reduced-soil cultivation, and controlled temperature regimes. Genotypes that maintain high HI across this environmental range are prioritized for BLSS deployment.

Fourth, establish HI-focused germplasm collections incorporating diverse allelic variation at priority loci. Mining of landraces, wild relatives, and underutilized genetic resources may reveal novel HI alleles absent from elite germplasm. These collections should be genotyped at HI loci and phenotyped under standardized conditions to build training populations for genomic selection.

6.5.3 For Space Agriculture Programs

Space agriculture programs, including NASA's efforts and international partners, should consider the following recommendations based on this dissertation's findings:

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

First, establish molecular breeding programs specifically targeting HI optimization in BLSS candidate cultivars. Current space crop development has focused primarily on adaptation to microgravity, lighting spectra, and pathogen-free cultivation. Explicit incorporation of HI as a breeding objective—using the gene targets identified in this dissertation—would improve food production efficiency per unit growth chamber volume and resource input.

Second, develop HI-optimized cultivars specifically for BLSS deployment, distinct from field-optimized varieties. The environmental conditions of space agriculture differ substantially from field production, and cultivars optimized for one environment may not perform optimally in the other. A dedicated breeding program targeting maximum HI under controlled-environment conditions could yield cultivars substantially superior to adapted field varieties.

Third, prioritize semi-dwarf architectures for all grain and tuber crops in BLSS systems. The volume constraints of space growth chambers make plant height a critical determinant of system efficiency. The consistently large effects of dwarfing genes across species (15-40% height reduction) translate directly to improved space utilization without proportional yield penalty when HI is maintained.

Fourth, invest in understanding gene-environment interactions for HI under space-relevant conditions. While terrestrial data provide useful guidance, the unique combination of factors in space habitats—microgravity, elevated CO₂, radiation, continuous lighting—may produce unanticipated effects on HI gene function. Ground-based analogs approximating these conditions, combined with limited flight experiments, should characterize any space-specific GxE interactions.

6.5.4 For Policy and Funding Agencies

Policy makers and funding agencies supporting agricultural research and space exploration should consider the following recommendations:

First, increase investment in pangenomic resources for crops with strategic importance for food security and space agriculture. Despite the significant progress documented in this dissertation, pangenome characterization remains incomplete for all four BLSS crops. Additional sequencing of diverse germplasm would expand the allelic repertoire available for breeding while enabling more comprehensive association analyses.

Second, support integrated research programs that bridge fundamental genomics with applied crop improvement. The translation from gene identification to cultivar deployment requires sustained effort across disciplinary boundaries. Funding mechanisms that incentivize collaboration between genomics researchers, breeding programs, and end-users (including space agencies) would accelerate the delivery of improved germplasm.

Third, facilitate data sharing and database interoperability for crop genomic resources. This dissertation benefited from the availability of curated databases (GWAS Atlas, Gramene, species-specific repositories), but substantial effort was required to integrate data across platforms. Investment in common data standards and federated database architectures would enhance the efficiency of multi-species comparative studies.

Fourth, recognize the strategic importance of bioregenerative life support research for humanity's long-term future in space. While current space missions rely primarily on resupply from Earth, sustainable exploration of the Moon, Mars, and beyond will require in-situ food production. Investment in the fundamental science of crop optimization for controlled

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

environments—including the genomics research exemplified by this dissertation—represents an essential component of the enabling technology portfolio for deep space exploration.

6.6 Final Reflections

Significance of the Work.

This dissertation represents a systematic effort to bridge the gap between fundamental plant genomics and the applied challenge of food production for space exploration. The identification of 30 HI-associated genes, the documentation of conserved pathways across diverse crops, and the development of an analytical framework for pangenomic trait analysis collectively advance both scientific understanding and practical capability for crop improvement. The work demonstrates that the genetic architecture of harvest index—a trait of fundamental importance for agricultural efficiency—is amenable to pangenomic characterization and that cross-species knowledge transfer can accelerate improvement across the crop portfolio required for human sustenance in space.

The significance of this work extends beyond its immediate applications in space agriculture. The finding that harvest index improvement in domesticated crops has consistently targeted the same conserved pathways—particularly GA signaling—provides insight into the evolutionary constraints and opportunities governing crop improvement. The methodological framework developed here can be applied to other complex traits, enabling systematic investigation of genetic architecture across species and environments.

Personal Insights from the Research Journey.

The research journey that culminated in this dissertation provided several insights that merit reflection. First, the power of comparative analysis across species became increasingly apparent as the work progressed. Individual crop studies, no matter how comprehensive, cannot

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

reveal the patterns of conservation and divergence that emerge from systematic cross-species comparison. The decision to investigate four phylogenetically diverse crops, rather than focusing deeply on a single species, proved essential for understanding the universal versus species-specific components of HI genetic architecture.

Second, the research highlighted the remarkable achievements of pre-genomic crop improvement. The Green Revolution genes that dominate the HI gene catalog were identified through phenotypic selection decades before the molecular biology revolution enabled their characterization. The *sd1* and *Rht* genes were deployed to transform global food production before their molecular identities were known. This observation underscores both the power of phenotypic selection and the potential for genomic approaches to accelerate future progress by enabling more precise and efficient breeding.

Third, the research revealed the complexity of translating genetic knowledge into practical crop improvement. Despite the wealth of genomic information now available, the pathway from gene identification to improved cultivar remains challenging. The recommendations offered in this chapter reflect an attempt to bridge this translation gap, but substantial work remains to realize the full potential of genomic approaches for crop improvement.

Vision for the Future of BLSS Crop Genomics.

Looking forward, the vision for BLSS crop genomics encompasses several developments that would advance the field beyond the current state of knowledge. Technologically, advances in long-read sequencing and pangenome graph construction will enable more complete characterization of structural variation at HI loci, potentially revealing novel alleles with superior effects. The continued decrease in sequencing costs will enable population-scale pangenomics,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

where thousands of accessions are characterized at the structural variant level rather than merely genotyped at known SNPs.

Methodologically, the integration of multi-omics data—including transcriptomics, proteomics, metabolomics, and phenomics—with genomic variation data will provide deeper insight into the molecular mechanisms connecting genotype to phenotype. Machine learning approaches that can effectively integrate these diverse data types hold promise for improving phenotypic prediction and identifying causal variants within complex loci.

Practically, the deployment of gene-edited cultivars incorporating optimized allele combinations at multiple HI loci will test the potential for genomics-assisted rapid improvement. While regulatory frameworks for gene-edited crops continue to evolve in different jurisdictions, the potential for precise modification of specific genes—rather than random mutagenesis or introduction of foreign DNA—offers a path toward public acceptance that may facilitate deployment for space agriculture applications where the stakes are particularly high.

Ultimately, the vision is for a complete understanding of the genetic control of harvest index sufficient to enable predictive design of optimal genotypes for specified environments. Such predictive capability would transform crop improvement from an empirical process of testing and selection to a knowledge-driven process of rational design. While this vision remains aspirational, the findings of this dissertation represent steps toward its realization.

6.7 Concluding Statement

This dissertation has investigated the pangenomic architecture of harvest index determinants in four Bioregenerative Life Support System staple crops—rice, wheat, potato, and soybean—with the goal of establishing a scientific foundation for food production optimization in space agriculture. Through systematic analysis of gene catalogs, ortholog relationships,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

structural variants, and phenotype-genotype associations, the research has identified 30 HI-associated genes distributed across five functional categories, documented the remarkable conservation of HI-regulatory pathways across the monocot-dicot divide, and developed a framework for applying pangenomic approaches to complex trait improvement.

The finding that gibberellin signaling represents a universally conserved target for harvest index improvement—exploited independently across crops during the Green Revolution—underscores the fundamental nature of this pathway in plant biology and the consistency of selection pressures during crop improvement. The additional pathways controlling grain size, developmental timing, starch biosynthesis, and sucrose transport provide complementary targets for HI optimization, with varying degrees of conservation that inform cross-species translation potential.

The practical contributions of this work—including the prioritized gene catalog, specific recommendations for marker-assisted selection and gene editing, and the analytical framework for future studies—provide actionable resources for NASA's Advanced Life Support Program and partner initiatives worldwide. The semi-dwarf genes (*sd1*, *Rht*), grain size regulators (*GW2*, *GS3*, *GW5*), and developmental timing genes (*StSP6A*) represent immediate priorities for incorporation into BLSS breeding programs.

As humanity prepares for extended presence beyond Earth—establishing permanent habitats on the Moon, sending crews to Mars, and eventually exploring deeper into the solar system—the ability to produce food efficiently in controlled environments becomes not merely advantageous but essential. The resource constraints of space travel demand maximum production efficiency from every gram of seed, every liter of water, and every watt of lighting

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

energy. Harvest index—the proportion of total biomass allocated to edible product—is the fundamental metric of this efficiency.

The research presented in this dissertation demonstrates that the genetic basis of harvest index is both comprehensible and manipulable through modern genomic approaches. The genes controlling plant architecture, grain size, developmental timing, and carbon partitioning have been cataloged, their evolutionary relationships documented, and their effects validated through multiple lines of evidence. This knowledge provides the scientific foundation for the next generation of BLSS crop improvement—cultivars designed specifically for the unique challenges and opportunities of food production in space.

In the final analysis, this dissertation represents a contribution to one of humanity's greatest endeavors: the extension of our species beyond the cradle of Earth. While the technical details of gene catalogs, ortholog conservation, and GWAS associations may seem far removed from the grandeur of space exploration, they are in fact essential components of the technological foundation that will enable sustainable human presence on other worlds. The plants that will feed astronauts on Mars—the wheat that becomes bread, the rice that sustains daily meals, the potatoes that provide carbohydrates and comfort, the soybeans that supply protein—will carry within their genomes the legacy of Earth's agricultural history and the promise of humanity's interplanetary future. This dissertation has sought to illuminate a small portion of that genetic legacy and to chart a course toward its continued improvement for the benefit of explorers yet to come.

REFERENCES

- Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suber, H., Ciber, L., Doyle, J., Mao, H., Goldstein, S., Berber, M., Lax, T., Zuber, H., Zhang, J., Huo, H., Razifard, H., & Lippman, Z. B. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, 182(1), 145-161.
<https://doi.org/10.1016/j.cell.2020.05.021>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403-410.
[https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Anderson, M. S., Barta, D., Douglas, G., Motil, B., Massa, G., & Roberts, M. (2015). Key gaps for enabling plant growth in future missions. Paper presented at the AIAA SPACE 2015 Conference and Exposition, Pasadena, CA. <https://doi.org/10.2514/6.2015-4294>
- Anzalone, A. V., Koblan, L. W., & Liu, D. R. (2020). Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*, 38(7), 824-844. <https://doi.org/10.1038/s41587-020-0561-9>
- Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814), 796-815. <https://doi.org/10.1038/35048692>
- Austin, R. B., Bingham, J., Blackwell, R. D., Evans, L. T., Ford, M. A., Morgan, C. L., & Taylor, M. (1980). Genetic improvements in winter wheat yields since 1900 and associated physiological changes. *The Journal of Agricultural Science*, 94(3), 675-689.
<https://doi.org/10.1017/S0021859600028665>
- Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pangenomics is all the rage. *Nature Plants*, 6(6), 653-655. <https://doi.org/10.1038/s41477-020-0676-x>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Bennetzen, J. L., & Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology*, 65, 505-530. <https://doi.org/10.1146/annurev-arplant-050213-035811>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633-2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Broman, K. W., Wu, H., Sen, S., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7), 889-890. <https://doi.org/10.1093/bioinformatics/btg112>
- Bugbee, B., & Salisbury, F. B. (1988). Exploring the limits of crop productivity: I. Photosynthetic efficiency of wheat in high irradiance environments. *Plant Physiology*, 88(3), 869-878. <https://doi.org/10.1104/pp.88.3.869>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1), 421. <https://doi.org/10.1186/1471-2105-10-421>
- Chang, T. G., & Zhu, X. G. (2017). Source-sink interaction: A century old concept under the light of modern molecular systems biology. *Journal of Experimental Botany*, 68(16), 4417-4431. <https://doi.org/10.1093/jxb/erx002>
- Chen, J., Liu, E., Cheng, C., Lu, F., Zhang, Y., Shen, Y., Wang, Q., Chen, F., Wang, K., Gao, L., Wang, W., & Xu, X. (2023). The complete genome sequence of T2T-NIP provides a new

- platform for functional genomics in rice. *Molecular Plant*, 16(7), 1211-1221.
<https://doi.org/10.1016/j.molp.2023.07.001>
- Chen, X., Tao, Y., Ali, A., Zhuang, Z., Guo, D., Guo, Q., Riaz, A., Zhang, H., Xu, P., Liao, Y., Wang, J., Sun, C., Xiang, Q., Wu, X., & Fu, X. (2020). Transcriptome and proteome profiling of different colored rice reveals physiological dynamics involved in the flavonoid pathway. *International Journal of Molecular Sciences*, 21(18), 6847.
<https://doi.org/10.3390/ijms21186847>
- Churchill, G. A., & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3), 963-971. <https://doi.org/10.1093/genetics/138.3.963>
- Collard, B. C., Jahufer, M. Z., Brouwer, J. B., & Pang, E. C. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142(1), 169-196.
<https://doi.org/10.1007/s10681-005-1681-5>
- Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., Wang, J., Hughes, T. J., Willis, D. K., Clemente, T. E., Diers, B. W., Jiang, J., Hudson, M. E., & Bent, A. F. (2012). Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science*, 338(6111), 1206-1209. <https://doi.org/10.1126/science.1228746>
- Cooper, M., Douglas, G., & Perchonok, M. (2011). Developing the NASA food system for long-duration missions. *Journal of Food Science*, 76(2), R40-R48.
<https://doi.org/10.1111/j.1750-3841.2010.01982.x>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158. <https://doi.org/10.1093/bioinformatics/btr330>
- De Pascale, S., Arena, C., Aronne, G., De Micco, V., Pannico, A., Paradiso, R., & Rouphael, Y. (2021). Biology and crop production in Space environments: Challenges and opportunities. *Life Sciences in Space Research*, 29, 30-37. <https://doi.org/10.1016/j.lssr.2021.02.005>
- Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biology*, 22(1), 3. <https://doi.org/10.1186/s13059-020-02224-8>
- Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell*, 127(7), 1309-1321. <https://doi.org/10.1016/j.cell.2006.12.006>
- Donald, C. M. (1962). In search of yield. *Journal of the Australian Institute of Agricultural Science*, 28(3), 171-178.
- Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seber, E., Garrison, E., & Paten, B. (2020). Pangenome graphs. *Annual Review of Genomics and Human Genetics*, 21, 139-162. <https://doi.org/10.1146/annurev-genom-120219-080406>
- Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), 157. <https://doi.org/10.1186/s13059-015-0721-2>
- Evenson, R. E., & Gollin, D. (2003). Assessing the impact of the Green Revolution, 1960 to 2000. *Science*, 300(5620), 758-762. <https://doi.org/10.1126/science.1078710>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Falconer, D. S., & Mackay, T. F. C. (1996). Introduction to quantitative genetics (4th ed.).

Longman.

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., Li, X., & Zhang, Q. (2006). GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical and Applied Genetics*, 112(6), 1164-1171. <https://doi.org/10.1007/s00122-006-0218-1>

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., Pan, Y., Zhou, G., Ren, H., Du, W., Yan, H., Wang, Y., Han, D., Shen, Y., Liu, S., ... & Kong, F. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology*, 18(1), 161. <https://doi.org/10.1186/s13059-017-1289-9>

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4), 783-791. <https://doi.org/10.1111/j.1558-5646.1985.tb00420.x>

Fischer, R. A., & Edmeades, G. O. (2010). Breeding and cereal yield progress. *Crop Science*, 50(S1), S52-S59. <https://doi.org/10.2135/cropsci2009.10.0564>

Fujita, D., Trijatmiko, K. R., Tagle, A. G., Sapasap, M. V., Koide, Y., Sasaki, K., Tsakirpaloglou, N., Gannaban, R. B., Nishimura, T., Yanagihara, S., Fukuta, Y., Koshiba, T., Slamet-Loedin, I. H., Ishimaru, T., & Kobayashi, N. (2013). NAL1 allele from a rice landrace greatly increases yield in modern indica cultivars. *Proceedings of the National Academy of Sciences*, 110(51), 20431-20436. <https://doi.org/10.1073/pnas.1310790110>

Gabur, I., Chawla, H. S., & Snowdon, R. J. (2022). Structural variation in plant genomes: From detection to applications in breeding. *Frontiers in Plant Science*, 13, 854177. <https://doi.org/10.3389/fpls.2022.854177>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burber, E. A., Huang, S., Van der Knaap, E., & Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, 51(6), 1044-1051.
<https://doi.org/10.1038/s41588-019-0410-2>
- Gao, C. (2021). Genome engineering for crop improvement and future agriculture. *Cell*, 184(6), 1621-1635. <https://doi.org/10.1016/j.cell.2021.01.005>
- Garrison, E., Siren, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9), 875-879. <https://doi.org/10.1038/nbt.4227>
- Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C. S. (2017). Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, 18(1), 527.
<https://doi.org/10.1186/s12864-017-3879-z>
- Gitelson, J. I., Lisovsky, G. M., & MacElroy, R. D. (2003). *Manmade closed ecological systems*. Taylor & Francis.
- Godia, F., Albiol, J., Montesinos, J. L., Perez, J., Creus, N., Cabello, F., Mengual, X., Montras, A., & Lasseur, C. (2002). MELISSA: A loop of interconnected bioreactors to develop life support in Space. *Journal of Biotechnology*, 99(3), 319-330.
[https://doi.org/10.1016/S0168-1656\(02\)00222-5](https://doi.org/10.1016/S0168-1656(02)00222-5)
- Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., ... & Briggs, S. P. (2002). A draft

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, 296(5565), 92-100.
<https://doi.org/10.1126/science.1068275>
- Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomic research in an era of climate change and food security. *Molecular Plant*, 13(5), 664-678.
<https://doi.org/10.1016/j.molp.2020.04.013>
- Graham, T., & Wheeler, R. (2016). Root restriction: A tool for improving volume utilization efficiency in bioregenerative life-support systems. *Life Sciences in Space Research*, 9, 62-68. <https://doi.org/10.1016/j.lssr.2016.04.001>
- Grant, D., Nelson, R. T., Cannon, S. B., & Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research*, 38(suppl_1), D843-D846. <https://doi.org/10.1093/nar/gkp798>
- Hamblin, M. T., Buckler, E. S., & Jannink, J. L. (2011). Population genetics of genomics-based crop improvement methods. *Trends in Genetics*, 27(3), 98-106.
<https://doi.org/10.1016/j.tig.2010.12.003>
- Hannah, L. C. (2007). Starch synthesis in the maize seed. In J. L. Bennetzen & S. C. Hake (Eds.), *Handbook of maize: Its biology* (pp. 279-295). Springer.
https://doi.org/10.1007/978-0-387-79418-1_14
- Hardigan, M. A., Laimbeer, F. P. E., Newton, L., Crisovan, E., Hamilton, J. P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J. C., Douches, D. S., Farre, E. M., Veilleux, R. E., & Buell, C. R. (2017). Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences*, 114(46), E9999-E10008.
<https://doi.org/10.1073/pnas.1714380114>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Hay, R. K. M. (1995). Harvest index: A review of its use in plant breeding and crop physiology. *Annals of Applied Biology*, 126(1), 197-216. <https://doi.org/10.1111/j.1744-7348.1995.tb05015.x>
- Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92(2), 433-443. <https://doi.org/10.3168/jds.2008-1646>
- Hedden, P. (2003). The genes of the Green Revolution. *Trends in Genetics*, 19(1), 5-9. [https://doi.org/10.1016/S0168-9525\(02\)00009-4](https://doi.org/10.1016/S0168-9525(02)00009-4)
- Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Siren, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21(1), 35. <https://doi.org/10.1186/s13059-020-1941-7>
- Hoff, K. J., & Stanke, M. (2019). Predicting genes in single genomes with AUGUSTUS. *Current Protocols in Bioinformatics*, 65(1), e57. <https://doi.org/10.1002/cpbi.57>
- Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., Fan, D., Guo, Y., Wang, A., Wang, L., Deng, L., Li, W., Lu, Y., Weng, Q., ... & Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, 42(11), 961-967. <https://doi.org/10.1038/ng.695>
- Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tiber, S., Taye, T., Telle, A., Maheshing, A., Marla, S. R., Jiao, Y., Lu, Y., Shi, J., ... & Ware, D. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, 373(6555), 655-662. <https://doi.org/10.1126/science.abg5289>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Human Pangenome Reference Consortium. (2023). A draft human pangenome reference. *Nature*, 617(7960), 312-324. <https://doi.org/10.1038/s41586-023-05896-x>
- Hurgobin, B., & Edwards, D. (2017). SNP discovery using a pangenome: Has the single reference approach become obsolete? *Biology*, 6(1), 21. <https://doi.org/10.3390/biology6010021>
- International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*, 436(7052), 793-800. <https://doi.org/10.1038/nature03895>
- International Wheat Genome Sequencing Consortium. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403), eaar7191. <https://doi.org/10.1126/science.aar7191>
- Ishimaru, K., Hirotsu, N., Madoka, Y., Murakami, N., Hara, N., Onodera, H., Kashiwagi, T., Ujiie, K., Shimizu, B., Onishi, A., Miyagawa, H., & Katoh, E. (2013). Loss of function of the IAA-glucose hydrolase gene *TGW6* enhances rice grain weight and increases yield. *Nature Genetics*, 45(6), 707-711. <https://doi.org/10.1038/ng.2612>
- Jarquin, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., & Lorenz, A. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics*, 15(1), 740. <https://doi.org/10.1186/1471-2164-15-740>
- Jiao, W. B., & Schneeberger, K. (2020). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, 36, 64-70. <https://doi.org/10.1016/j.pbi.2017.02.002>
- Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C. S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K. L., Wolfgruber, T. K., May, M. R., Springer, N. M., ... & Ware, D. (2017). Improved

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- maize reference genome with single-molecule technologies. *Nature*, 546(7659), 524-527.
<https://doi.org/10.1038/nature22971>
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096), 816-821. <https://doi.org/10.1126/science.1225829>
- Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9), 1236-1240.
<https://doi.org/10.1093/bioinformatics/btu031>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772-780. <https://doi.org/10.1093/molbev/mst010>
- Khush, G. S. (1999). Green revolution: Preparing for the 21st century. *Genome*, 42(4), 646-655.
<https://doi.org/10.1139/g99-044>
- Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8), 907-915. <https://doi.org/10.1038/s41587-019-0201-4>
- Kloosterman, B., Abelenda, J. A., Gomez, M. D. M. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., Horvath, B. M., van Eck, H. J., Smaczniak, C., Prat, S., Visser, R. G. F., & Bachem, C. W. B. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature*, 495(7440), 246-250.
<https://doi.org/10.1038/nature11912>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5), 722-736. <https://doi.org/10.1101/gr.215087.116>
- Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*, 9(1), 29. <https://doi.org/10.1186/1746-4811-9-29>
- Lander, E. S., & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1), 185-199. <https://doi.org/10.1093/genetics/121.1.185>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357-359. <https://doi.org/10.1038/nmeth.1923>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658-1659. <https://doi.org/10.1093/bioinformatics/btl1158>
- Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S. S., Zuo, Q., Shi, X. H., Li, Y. F., Zhang, W. K., Hu, Y., Kong, G., Hong, H. L., Tan, B., ... & Qiu, L. J. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, 32(10), 1045-1052. <https://doi.org/10.1038/nbt.2979>
- Li, Z., Jiang, L., Ma, Y., Wei, Z., Hong, H., Liu, Z., Lei, J., Liu, Y., Guan, R., Guo, Y., Jin, L., Zhang, L., Li, Y., Ren, Y., He, W., Liu, M., Htwe, N. M. P. S., Liu, L., Guo, B., ... & Liu,

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- B. (2017). Development and utilization of a new chemically-induced soybean library with a high density of mutations. *Journal of Integrative Plant Biology*, 59(1), 60-74.
<https://doi.org/10.1111/jipb.12505>
- Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., & Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, 28(18), 2397-2399. <https://doi.org/10.1093/bioinformatics/bts444>
- Liu, Y., Du, H., Li, P., & Jiao, Y. (2023). Graph-based pangenome analyses in plants: Current progress and challenges. *The Crop Journal*, 11(5), 1267-1275.
<https://doi.org/10.1016/j.cj.2023.01.006>
- Liu, S., Hou, W., Wei, H., Kong, L., Zhang, Y., Wang, Y., Tang, X., & Liu, B. (2018). A classic agronomy trait GmDW1 controls soybean stature and adaptation. *Plant Physiology*, 178(4), 1537-1549. <https://doi.org/10.1104/pp.18.00823>
- Long, S. P., Zhu, X. G., Naidu, S. L., & Ort, D. R. (2006). Can improvement in photosynthesis increase crop yields? *Plant, Cell & Environment*, 29(3), 315-330.
<https://doi.org/10.1111/j.1365-3040.2005.01493.x>
- Lye, Z. N., & Purugganan, M. D. (2019). Copy number variation in domestication. *Trends in Plant Science*, 24(4), 352-365. <https://doi.org/10.1016/j.tplants.2019.01.003>
- Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer Associates.
- Lyu, J., Wang, D., Duan, P., Liu, Y., Huang, K., Zeng, D., Zhang, L., Dong, G., Li, Y., Xu, R., Li, Y., Tian, Y., & Li, Y. (2024). GmSW17 is a major QTL for seed size and a target of artificial selection in soybean. *Nature Communications*, 15(1), 6829.
<https://doi.org/10.1038/s41467-024-51798-5>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., Villesen, P., Skov, L., Belling, K., Have, C. T., Gonzalez, R. F., Esmoris, I., Brunak, S., Rasmussen, S., & Schierup, M. H. (2017). Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*, 548(7665), 87-91. <https://doi.org/10.1038/nature23264>
- Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffaari, A., Kerber, R., Klotzl, F., Luo, H., Meyner, C., Rahman, N., Raney, B. J., Reilly, J. G., Siren, J., & Studier, F. W. (2018). Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics*, 19(1), 118-135. <https://doi.org/10.1093/bib/bbw089>
- Massa, G. D., Dufour, N. F., Carver, J. A., Hummerick, M. E., Wheeler, R. M., Morrow, R. C., & Smith, T. M. (2017). VEG-01: Veggie hardware validation testing on the International Space Station. *Open Agriculture*, 2(1), 33-41. <https://doi.org/10.1515/opag-2017-0003>
- McKey, D., Elias, M., Pujol, B., & Duputie, A. (2010). The evolutionary ecology of clonally propagated domesticated plants. *New Phytologist*, 186(2), 318-332. <https://doi.org/10.1111/j.1469-8137.2010.03210.x>
- Medini, D., Donati, C., Tettelin, H., Massignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6), 589-594. <https://doi.org/10.1016/j.gde.2005.09.006>
- Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), 1819-1829. <https://doi.org/10.1093/genetics/157.4.1819>
- Meyer, R. S., & Purugganan, M. D. (2013). Evolution of crop species: Genetics of domestication and diversification. *Nature Reviews Genetics*, 14(12), 840-852. <https://doi.org/10.1038/nrg3605>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Michael, T. P., & VanBuren, R. (2020). Building near-complete plant genomes. *Current Opinion in Plant Biology*, 54, 26-33. <https://doi.org/10.1016/j.pbi.2019.12.009>
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5), 1530-1534. <https://doi.org/10.1093/molbev/msaa015>
- Mitchell, C. A., Dzakovich, M. P., Gomez, C., Lopez, R., Burr, J. F., Hernandez, R., Kubota, C., Currey, C. J., Meng, Q., Runkle, E. S., Bourget, C. M., Morrow, R. C., & Both, A. J. (2015). Light-emitting diodes in horticulture. *Horticultural Reviews*, 43, 1-87. <https://doi.org/10.1002/9781119107781.ch01>
- Monje, O., Stutte, G. W., Goins, G. D., Porterfield, D. M., & Bingham, G. E. (2003). Farming in space: Environmental and biophysical concerns. *Advances in Space Research*, 31(1), 151-167. [https://doi.org/10.1016/S0273-1177\(02\)00751-2](https://doi.org/10.1016/S0273-1177(02)00751-2)
- Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K., Visendi, P., Lai, K., Dolezel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5), 1007-1013. <https://doi.org/10.1111/tpj.13515>
- Navarro, C., Abelenda, J. A., Cruz-Oro, E., Cuellar, C. A., Tamaki, S., Silva, J., Shimamoto, K., & Prat, S. (2011). Control of flowering and storage organ formation in potato by FLOWERING LOCUS T. *Nature*, 478(7367), 119-122. <https://doi.org/10.1038/nature10431>
- Ou, S., Chen, J., & Jiang, N. (2023). Long-read sequencing opens a window into the evolution of plant genomes. *Annual Review of Plant Biology*, 74, 645-667. <https://doi.org/10.1146/annurev-arplant-060920-093709>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Paradiso, R., De Micco, V., Buonomo, R., Aronne, G., Barbieri, G., & De Pascale, S. (2014). Soilless cultivation of soybean for Bioregenerative Life-Support Systems: A literature review and the experience of the MELiSSA Project - Food characterisation Phase I. *Plant Biology*, 16, 69-78. <https://doi.org/10.1111/plb.12056>
- Parry, M. A., Reynolds, M., Salvucci, M. E., Raines, C., Andralojc, P. J., Zhu, X. G., Price, G. D., Condon, A. G., & Furbank, R. T. (2011). Raising yield potential of wheat. II. Increasing photosynthetic capacity and efficiency. *Journal of Experimental Botany*, 62(2), 453-467. <https://doi.org/10.1093/jxb/erq304>
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), e190. <https://doi.org/10.1371/journal.pgen.0020190>
- Paul, A. L., Wheeler, R. M., Levine, H. G., & Ferl, R. J. (2013). Fundamental plant biology enabled by the space shuttle. *American Journal of Botany*, 100(1), 226-234. <https://doi.org/10.3732/ajb.1200338>
- Pearce, S., Saville, R., Vaughan, S. P., Chandler, P. M., Wilhelm, E. P., Sparks, C. A., Al-Kaff, N., Korolev, A., Boulton, M. I., Phillips, A. L., Hedden, P., Nicholson, P., & Thomas, S. G. (2011). Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat. *Plant Physiology*, 157(4), 1820-1831. <https://doi.org/10.1104/pp.111.183657>
- Peng, J., Richards, D. E., Hartley, N. M., Murphy, G. P., Devos, K. M., Flintham, J. E., Beales, J., Fish, L. J., Worland, A. J., Pelica, F., Sudhakar, D., Christou, P., Snape, J. W., Gale, M. D., & Harberd, N. P. (1999). Green revolution genes encode mutant gibberellin response modulators. *Nature*, 400(6741), 256-261. <https://doi.org/10.1038/22307>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Peng, S., Khush, G. S., Virk, P., Tang, Q., & Zou, Y. (2008). Progress in ideotype breeding to increase rice yield potential. *Field Crops Research*, 108(1), 32-38.
<https://doi.org/10.1016/j.fcr.2008.04.001>
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3), 290-295. <https://doi.org/10.1038/nbt.3122>
- Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., International Wheat Genome Sequencing Consortium, Mayer, K. F., & Olsen, O. A. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, 345(6194), 1250091. <https://doi.org/10.1126/science.1250091>
- Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, N., & Buell, C. R. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience*, 9(9), giaa100.
<https://doi.org/10.1093/gigascience/giaa100>
- Ping, J., Liu, Y., Sun, L., Zhao, M., Li, Y., She, M., Sui, Y., Lin, F., Liu, X., Tang, Z., Nguyen, H., Tian, Z., Qiu, L., Nelson, R. L., Clemente, T. E., Specht, J. E., & Ma, J. (2014). Dt2 is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. *The Plant Cell*, 26(7), 2831-2842. <https://doi.org/10.1105/tpc.114.126938>
- Potato Genome Sequencing Consortium. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355), 189-195. <https://doi.org/10.1038/nature10158>
- Preiss, J. (1991). *Biology and molecular biology of starch synthesis and its regulation*. Oxford Surveys of Plant Molecular and Cell Biology, 7, 59-114.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2: Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3), e9490.
<https://doi.org/10.1371/journal.pone.0009490>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.
<https://doi.org/10.1093/genetics/155.2.945>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559-575. <https://doi.org/10.1086/519795>
- Purugganan, M. D., & Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature*, 457(7231), 843-848. <https://doi.org/10.1038/nature07895>
- Qin, L., Hao, C., Hou, J., Wang, Y., Li, T., Wang, L., Ma, Z., & Zhang, X. (2014). Homologous haplotypes, expression, genetic effects and geographic distribution of the wheat yield gene TaGW2. *BMC Plant Biology*, 14(1), 107. <https://doi.org/10.1186/1471-2229-14-107>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.
<https://doi.org/10.1093/bioinformatics/btq033>
- Rasheed, A., Xia, X., Ogbonnaya, F., Mahmood, T., Zhang, Z., Mujeeb-Kazi, A., & He, Z. (2014). Genome-wide association for grain morphology in synthetic hexaploid wheats using exome capture genotyping. *BMC Genomics*, 15(1), 1-13.
<https://doi.org/10.1186/1471-2164-15-851>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Rebetzke, G. J., & Richards, R. A. (2000). Gibberellic acid-sensitive dwarfing genes reduce plant height to increase kernel number and grain yield of wheat. *Australian Journal of Agricultural Research*, 51(2), 235-246. <https://doi.org/10.1071/AR99043>
- Reynolds, M., Foulkes, J., Furbank, R., Griffiths, S., King, J., Murchie, E., Parry, M., & Slafer, G. (2012). Achieving yield gains in wheat. *Plant, Cell & Environment*, 35(10), 1799-1823. <https://doi.org/10.1111/j.1365-3040.2012.02588.x>
- Richards, R. A. (2000). Selectable traits to increase crop photosynthesis and yield of grain crops. *Journal of Experimental Botany*, 51(suppl_1), 447-458. https://doi.org/10.1093/jexbot/51.suppl_1.447
- Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281), 1516-1517. <https://doi.org/10.1126/science.273.5281.1516>
- Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24-26. <https://doi.org/10.1038/nbt.1754>
- Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences*, 104(suppl_1), 8641-8648. <https://doi.org/10.1073/pnas.0700643104>
- Salisbury, F. B., Gitelson, J. I., & Lisovsky, G. M. (1997). Bios-3: Siberian experiments in bioregenerative life support. *BioScience*, 47(9), 575-585. <https://doi.org/10.2307/1313164>
- Sasaki, A., Ashikari, M., Ueguchi-Tanaka, M., Itoh, H., Nishimura, A., Swapan, D., Ishiyama, K., Saito, T., Kobayashi, M., Khush, G. S., Kitano, H., & Matsuoka, M. (2002). A mutant

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

gibberellin-synthesis gene in rice. *Nature*, 416(6882), 701-702.

<https://doi.org/10.1038/416701a>

- Schatz, M. C., Maron, L. G., Stein, J. C., Wences, A. H., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., Wright, M. H., Chia, J., Ware, D., McCouch, S. R., & McCombie, W. R. (2014). Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biology*, 15(11), 506. <https://doi.org/10.1186/s13059-014-0506-z>
- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., ... & Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278), 178-183. <https://doi.org/10.1038/nature08670>
- Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J. J., Qiu, J. L., & Gao, C. (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature Biotechnology*, 31(8), 686-688. <https://doi.org/10.1038/nbt.2650>
- Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., & Yano, M. (2008). Deletion in a gene associated with grain size increased yields during rice domestication. *Nature Genetics*, 40(8), 1023-1028. <https://doi.org/10.1038/ng.169>
- Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210-3212. <https://doi.org/10.1093/bioinformatics/btv351>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Simmonds, J., Scott, P., Brinton, J., Mestre, T. C., Bush, M., del Blanco, A., Dubcovsky, J., & Uauy, C. (2016). A splice acceptor site mutation in TaGW2-A1 increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. *Theoretical and Applied Genetics*, 129(6), 1099-1112. <https://doi.org/10.1007/s00122-016-2686-2>
- Sinclair, T. R. (1998). Historical changes in harvest index and crop nitrogen accumulation. *Crop Science*, 38(3), 638-643. <https://doi.org/10.2135/cropsci1998.0011183X003800030002x>
- Slafer, G. A., Elia, M., Savin, R., Garcia, G. A., Terrile, I. I., Ferrante, A., Miralles, D. J., & Gonzalez, F. G. (2015). Fruiting efficiency: An alternative trait to further rise wheat yield. *Food and Energy Security*, 4(2), 92-109. <https://doi.org/10.1002/fes3.59>
- Smith, A. M., & Stitt, M. (2007). Coordination of carbon supply and plant growth. *Plant, Cell & Environment*, 30(9), 1126-1149. <https://doi.org/10.1111/j.1365-3040.2007.01708.x>
- Song, X. J., Huang, W., Shi, M., Zhu, M. Z., & Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature Genetics*, 39(5), 623-630. <https://doi.org/10.1038/ng2014>
- Sonnevald, S., & Sonnewald, U. (2014). Regulation of potato tuber sprouting. *Planta*, 239(1), 27-38. <https://doi.org/10.1007/s00425-013-1968-z>
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312-1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stutte, G. W., Monje, O., Goins, G. D., & Tripathy, B. C. (2005). Microgravity effects on thylakoid, single leaf, and whole canopy photosynthesis of dwarf wheat. *Planta*, 223(1), 46-56. <https://doi.org/10.1007/s00425-005-0066-2>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P., & Reynolds, M. P. (2015). Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theoretical and Applied Genetics*, 128(2), 353-363. <https://doi.org/10.1007/s00122-014-2435-3>
- Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., Wang, W., Shi, J., Wang, C., Lu, J., Zhang, D., Li, Z., & Wei, C. (2017). RPAN: Rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Research*, 45(2), 597-605. <https://doi.org/10.1093/nar/gkw958>
- Sun, H., Jiao, W., Krber, U., Kong, Y., Li, C., Xu, Y., Zheng, Y., & Schneeberger, K. (2024). A graph pangenome of potato reveals structural variants associated with key traits. *Nature*, 617(7962), 527-534. <https://doi.org/10.1038/s41586-024-08476-9>
- Sun, S., Wang, M., Li, C., Sun, J., Fu, R., Li, L., Liu, X., Tong, X., Zhang, J., Lin, W., Yang, K., Xing, G., Li, K., Li, J., Zhao, H., Chen, F., Li, Z., Huang, X., Jiao, Y., & Lai, J. (2022). Pan-Zea: A pangenome of 722 maize and teosinte lines. *bioRxiv*. <https://doi.org/10.1101/2022.01.04.474960>
- Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., & Springer, N. M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*, 20(12), 1689-1699. <https://doi.org/10.1101/gr.109165.110>
- Takano-Kai, N., Jiang, H., Kubo, T., Sweeney, M., Matsumoto, T., Kanamori, H., Padhukasahasram, B., Bustamante, C., Yoshimura, A., Doi, K., & McCouch, S. (2009). Evolutionary history of GS3, a gene conferring grain length in rice. *Genetics*, 182(4), 1323-1334. <https://doi.org/10.1534/genetics.109.103002>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Tan, M., Li, M., Wu, G., Chen, L., & Li, M. (2020). Gene presence/absence variation in *Brassica napus* is associated with quantitative disease resistance to *Verticillium longisporum*. *Scientific Reports*, 10(1), 4321. <https://doi.org/10.1038/s41598-020-61303-1>
- Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., Bao, Z., Liu, Z., Feng, S., Zhu, X., Li, D., Zhu, G., Wang, H., Zhou, Y., Zhou, Y., Bryan, G. J., Buell, C. R., Zhang, C., & Huang, S. (2022). Genome evolution and diversity of wild and cultivated potatoes. *Nature*, 606(7914), 535-541. <https://doi.org/10.1038/s41586-022-04822-x>
- Tanksley, S. D., & McCouch, S. R. (1997). Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science*, 277(5329), 1063-1066. <https://doi.org/10.1126/science.277.5329.1063>
- Tetlow, I. J. (2011). Starch biosynthesis in developing seeds. *Seed Science Research*, 21(1), 5-32. <https://doi.org/10.1017/S0960258510000292>
- Tettelin, H., Maignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Anguoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., ... & Fraser, C. M. (2005). 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. <https://doi.org/10.1073/pnas.0506758102>
- Three Thousand Rice Genomes Project. (2014). The 3,000 rice genomes project. *GigaScience*, 3(1), 7. <https://doi.org/10.1186/2047-217X-3-7>
- Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, 14(1), e20077. <https://doi.org/10.1002/tpg2.20077>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Torkamaneh, D., Lemay, M. A., & Belzile, F. (2021). The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. *Plant Biotechnology Journal*, 19(9), 1852-1862. <https://doi.org/10.1111/pbi.13600>
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3), 562-578. <https://doi.org/10.1038/nprot.2012.016>
- Turner, A., Beales, J., Faure, S., Dunford, R. P., & Laurie, D. A. (2005). The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science*, 310(5750), 1031-1034. <https://doi.org/10.1126/science.1117619>
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11), 4414-4423. <https://doi.org/10.3168/jds.2007-0980>
- Varshney, R. K., Terauchi, R., & McCouch, S. R. (2014). Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLoS Biology*, 12(6), e1001883. <https://doi.org/10.1371/journal.pbio.1001883>
- Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23, 148-154. <https://doi.org/10.1016/j.mib.2014.11.016>
- Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Schoen, A., Cloutier, S., ... & Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, 588(7837), 277-283. <https://doi.org/10.1038/s41586-020-2961-x>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Wang, W., Li, G., & Han, B. (2025). Pangenomics-driven crop improvement. *Nature Reviews Bioengineering*, 3(1), 50-68. <https://doi.org/10.1038/s44222-024-00244-8>
- Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y. C., Liu, Z., Frommer, W. B., Ma, J. F., Chen, L. Q., Guan, Y., Shou, H., & Tian, Z. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. *National Science Review*, 7(11), 1776-1786. <https://doi.org/10.1093/nsr/nwaa110>
- Wang, Q., Tian, F., Pan, Y., Buckler, E. S., & Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PLoS ONE*, 9(9), e107684. <https://doi.org/10.1371/journal.pone.0107684>
- Watanabe, S., Hideshima, R., Xia, Z., Tsubokura, Y., Sato, S., Nakamoto, Y., Yamanaka, N., Takahashi, R., Ishimoto, M., Anai, T., Tabata, S., & Harada, K. (2009). Map-based cloning of the gene associated with the soybean maturity locus E3. *Genetics*, 182(4), 1251-1262. <https://doi.org/10.1534/genetics.108.098772>
- Weng, J., Gu, S., Wan, X., Gao, H., Guo, T., Su, N., Lei, C., Zhang, X., Cheng, Z., Guo, X., Wang, J., Jiang, L., Zhai, H., & Wan, J. (2008). Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Research*, 18(12), 1199-1209. <https://doi.org/10.1038/cr.2008.307>
- Wheeler, R. M. (2010). Plants for human life support in space: From Myers to Mars. *Gravitational and Space Biology*, 23(2), 25-36.
- White, A. C., Rogers, A., Rees, M., & Osborne, C. P. (2016). How can we make plants grow faster? A source-sink perspective on growth rate. *Journal of Experimental Botany*, 67(1), 31-45. <https://doi.org/10.1093/jxb/erv447>

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- Wilhelm, E. P., Turner, A. S., & Laurie, D. A. (2009). Photoperiod insensitive Ppd-A1a mutations in tetraploid wheat (*Triticum durum* Desf.). *Theoretical and Applied Genetics*, 118(2), 285-294. <https://doi.org/10.1007/s00122-008-0898-9>
- Worland, A. J., Korzun, V., Roder, M. S., Ganai, M. W., & Law, C. N. (1998). Genetic analysis of the dwarfing gene Rht8 in wheat. Part II. The distribution and adaptive significance of allelic variants at the Rht8 locus of wheat as revealed by microsatellite screening. *Theoretical and Applied Genetics*, 96(8), 1110-1120. <https://doi.org/10.1007/s001220050846>
- Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., Anai, T., Sato, S., Yamazaki, T., Lu, S., Wu, H., Tabata, S., & Harada, K. (2012). Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. *Proceedings of the National Academy of Sciences*, 109(32), E2155-E2164. <https://doi.org/10.1073/pnas.1117982109>
- Yan, L., Loukoianov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., SanMiguel, P., Bennetzen, J. L., Echenique, V., & Dubcovsky, J. (2004). The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science*, 303(5664), 1640-1644. <https://doi.org/10.1126/science.1094305>
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7), 565-569. <https://doi.org/10.1038/ng.608>
- Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

- unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2), 203-208. <https://doi.org/10.1038/ng1702>
- Zabel, P., Bamsey, M., Schubert, D., & Tajmar, M. (2016). Review and analysis of over 40 years of space plant growth systems. *Life Sciences in Space Research*, 10, 1-16. <https://doi.org/10.1016/j.lssr.2016.06.004>
- Zhang, L., Liu, D., & Ma, C. (2021). Copy number variation in plant genomes. *The Crop Journal*, 9(3), 505-514. <https://doi.org/10.1016/j.cj.2021.03.010>
- Zhang, Y., Liu, J., Xia, X., & He, Z. (2014). TaGS-D1, an ortholog of rice OsGS3, is associated with grain weight and grain length in common wheat. *Molecular Breeding*, 34(3), 1097-1107. <https://doi.org/10.1007/s11032-014-0102-7>
- Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4), 355-360. <https://doi.org/10.1038/ng.546>
- Zhang, H., Zhang, J., Wei, P., Zhang, B., Gou, F., Feng, Z., Mao, Y., Yang, L., Zhang, H., Xu, N., & Zhu, J. K. (2014). The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant Biotechnology Journal*, 12(6), 797-807. <https://doi.org/10.1111/pbi.12200>
- Zhao, Q., Liu, Y., & Jiao, Y. (2024). Plant pangenomics is the new normal. *Annual Review of Plant Biology*, 75(1), 653-681. <https://doi.org/10.1146/annurev-arplant-061322-040119>
- Zhu, H., Li, C., & Gao, C. (2020). Applications of CRISPR-Cas in agriculture and plant biotechnology. *Nature Reviews Molecular Cell Biology*, 21(11), 661-677. <https://doi.org/10.1038/s41580-020-00288-9>

APPENDIX A: Data Sources and URLs

This appendix provides a comprehensive catalog of the genomic databases, pangenome resources, bioinformatics tools, and reference genome assemblies utilized in this dissertation research. All data sources listed herein are publicly accessible and were accessed during January 2026 unless otherwise noted. The databases are organized by function: primary genomic databases (Table A1), orthology and comparative genomics resources (Table A2), species-specific pangenome databases (Table A3), phenotype-genotype association databases (Table A4), reference genome assemblies (Table A5), and bioinformatics software tools (Table A6). Uniform Resource Locators (URLs) are provided to facilitate data retrieval and ensure reproducibility of the analytical pipeline described in Chapter 3.

Table A1 Primary Genomic Databases

Database	URL	Data Type	Access Date	Notes
NCBI GenBank	https://www.ncbi.nlm.nih.gov/genbank/	Reference genomes, gene sequences, protein sequences	January 2026	Public access; primary repository for nucleotide sequences
Phytozome	https://phytozome-next.jgi.doe.gov/	Plant genomes, functional annotations, gene families	January 2026	JGI account required; hosts Wm82.a4.v1 soybean assembly

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Ensembl Plants	https://plants.ensembl.org/	Plant genomes, variation data, comparative genomics	January 2026	Public access; integrated with Compara for ortholog data
InterPro	https://www.ebi.ac.uk/interpro/	Protein domains, families, functional sites	January 2026	Public access; integrates 14 member databases
Pfam	https://www.ebi.ac.uk/interpro/entry/pfam/	Protein families, domain architectures	January 2026	Now integrated within InterPro; HMM-based classification

Note. All databases provide programmatic access through APIs or bulk download options. URLs verified as of January 2026.

Table A2 Orthology and Comparative Genomics Databases

Databases	URL	Data Type	Access Date	Notes
OrthoDB	https://www.orthodb.org/	Hierarchical ortholog groups, functional annotations	January 2026	Public access; 117 plant genomes included
PLAZA Dicots 5.0	https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v5_dicots/	Dicot comparative	January 2026	Registration required; potato and

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

		genomics, gene families, synteny		soybean data
PLAZA Monocots 5.0	https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v5_monocots/	Monocot comparative genomics, gene families, synteny	January 2026	Registration required; rice and wheat data
Ensembl Compara	https://plants.ensembl.org/info/genome/compara/	Gene trees, whole- genome alignments , synteny maps	January 2026	Public access; integrated with Ensembl Plants

Note. PLAZA provides separate portals for monocot and dicot species. OrthoDB hierarchical structure enables ortholog queries at multiple taxonomic levels (Viridiplantae, Monocot/Dicot, species-specific).

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table A3 Species-Specific Pangenome Resources

Species	Resource	URL	Data Type	Access Date
Rice (<i>Oryza sativa</i>)	3K Rice Genomes / SNP-Seek	https://snp-seek.irri.org/	SNPs, PAVs, pangenome variation (3,010 accessions)	January 2026
Rice (<i>Oryza sativa</i>)	RPAN Database	http://cgm.sjtu.edu.cn/3kricedb/	Rice pan-genome browser, core/dispensable genes	January 2026
Rice (<i>Oryza sativa</i>)	RiceVarMap	http://ricevarmap.ncpgr.cn/	Rice variation database, haplotype analysis	January 2026
Wheat (<i>Triticum aestivum</i>)	10+ Wheat Genomes	https://webblast.ipk-gatersleben.de/wheat_ten_genomes/	Wheat pangenome, multiple cultivar assemblies	January 2026
Wheat (<i>Triticum aestivum</i>)	WheatOmics	http://wheatomics.sdau.edu.cn/	Wheat multi-omics, structural variants	January 2026

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Wheat (Triticum aestivum)	GrainGenes	https://wheat.pw.usda.gov/GG3/	Triticeae genomics, markers, QTLs	January 2026
Potato (Solanum tuberosum)	Spud DB	http://spuddb.uga.edu/	Potato genomics, DM v6.1 assembly	January 2026
Potato (Solanum tuberosum)	PGSC	https://solanaceae.plantbiology.msu.edu/pgsc_download.shtml	Potato genome sequencing consortium data	January 2026
Potato (Solanum tuberosum)	Sol Genomics Network	https://solgenomics.net/	Solanaceae genomics, annotations	January 2026
Soybean (Glycine max)	SoyBase	https://www.soybase.org/	Soybean genomics, genetic maps, QTLs	January 2026
Soybean (Glycine max)	PanSoy	https://soybase.org/PanSoy/	Soybean pangenome (204 accessions)	January 2026
Soybean (Glycine max)	GmHapMap	https://soybase.org/GmHapMap/	Soybean haplotype map, diversity data	January 2026

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Note. PAV = presence/absence variant. Pangenome resources provide structural variant data beyond single reference genomes. The 3K Rice Genomes Project represents the most comprehensive crop pangenome resource currently available.

Table A4 Phenotype-Genotype Association Databases

Database	URL	Data Type	Access Date	Notes
GWAS Atlas	https://bigd.big.ac.cn/gwas/	GWAS associations across species	January 2026	Public access; curated plant GWAS results
Gramene	https://www.gramene.org/	QTLs, pathways, comparative genomics	January 2026	Public access; Plant Reactome integration
Triticeae Toolbox	https://triticeaetoolbox.org/	Wheat and barley phenotype- genotype data	January 2026	Public access; breeding program data
RAP-DB	https://rapdb.dna.affrc.go.jp/	Rice annotation, gene expression, QTLs	January 2026	Public access; IRGSP-1.0 annotations
Q-TARO	http://qtaro.abr.affrc.go.jp/	Rice QTL database, trait ontology	January 2026	Public access; curated rice QTLs

Note. GWAS = genome-wide association study; QTL = quantitative trait locus. These databases provide phenotype-genotype linkages essential for identifying harvest index-associated variants.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Table A5 Reference Genome Assemblies Used

Species	Assembly	Version	Source	Accession
Rice (<i>Oryza sativa</i>)	IRGSP-1.0	Os-Nipponbare-Reference-IRGSP-1.0	NCBI / RAP-DB	GCF_001433935.1
Wheat (<i>Triticum aestivum</i>)	IWGSC RefSeq	v2.1 (Chinese Spring)	IWGSC / URGI	GCF_018294505.1
Potato (<i>Solanum tuberosum</i>)	DM v6.1	<i>Solanum tuberosum</i> DM1-3	Spud DB / Phytozome	GCF_000226075.1
Soybean (<i>Glycine max</i>)	Williams 82	Wm82.a4.v1	Phytozome / SoyBase	GCF_000004515.6

Note. IRGSP = International Rice Genome Sequencing Project; IWGSC = International Wheat Genome Sequencing Consortium; DM = doubled monoploid. These assemblies represent the current gold-standard references for each species and are utilized by major comparative genomics resources.

Table A6 Bioinformatics Tools and Software

Tool	Version	URL	Purpose
OrthoFinder	2.5.4	https://github.com/davideemms/OrthoFinder	Ortholog detection, gene tree inference
MAFFT	7.505	https://mafft.cbrc.jp/alignment/software/	Multiple sequence alignment
TrimAl	1.4.1	http://trimal.cgenomics.org/	Alignment trimming, quality filtering

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

IQ-TREE	2.2.0	http://www.iqtree.org/	Maximum likelihood phylogenetic inference
MEGA-X	10.2.6	https://www.megasoftware.net/	Molecular evolutionary analysis, dating
MCScanX	—	https://github.com/wyp1125/MCScanX	Synteny analysis, collinearity detection
BLAST+	2.14.0	https://blast.ncbi.nlm.nih.gov/	Sequence similarity searches
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond	Fast protein alignment for large datasets
ModelFinder	Integrated	http://www.iqtree.org/ModelFinder/	Amino acid substitution model selection
SynVisio	—	https://synvisio.github.io/	Synteny visualization

Note. Version numbers reflect software releases used in this study. ModelFinder is integrated within IQ-TREE. All tools are freely available for academic use. Detailed parameters for each tool are documented in Chapter 3, Section 3.5.

Acknowledgment of Data Sources

Data Provider Acknowledgments

The genomic and pangenomic data utilized in this dissertation were generated and curated by numerous research consortia and institutions whose contributions are gratefully acknowledged. The International Rice Genome Sequencing Project (IRGSP) and the 3,000 Rice Genomes Project provided foundational rice genomic resources. The International Wheat Genome Sequencing Consortium (IWGSC) and the 10+ Wheat Genomes Project enabled comprehensive wheat pangenomic analyses. The Potato Genome Sequencing Consortium (PGSC) and the Spud DB team at the University of Georgia provided essential potato genomic data. The SoyBase team at the USDA-ARS and the Department of Energy Joint Genome Institute (JGI) Phytozome portal provided soybean genomic resources.

Data Use Agreements

All data utilized in this research are publicly available under open access policies or through standard academic registration. Phytozome data access was obtained through a JGI user account in accordance with the JGI Data Use Policy. PLAZA database access was obtained through academic registration. SNP-Seek and RPAN databases are freely accessible without registration. All data use complies with the Fort Lauderdale and Toronto agreements regarding prepublication data sharing in genomics research.

PANGENOMIC ARCHITECTURE OF HI DETERMINANTS IN BLSS STAPLE CROPS

Funding Acknowledgments for Database Maintenance

The databases utilized in this research are supported by various funding agencies whose contributions enable continued data curation and public access. NCBI GenBank is supported by the National Institutes of Health (NIH) and the National Library of Medicine (NLM). Phytozome is supported by the U.S. Department of Energy Office of Science. Ensembl Plants is supported by the European Molecular Biology Laboratory (EMBL) and the Biotechnology and Biological Sciences Research Council (BBSRC). OrthoDB is supported by the Swiss Institute of Bioinformatics (SIB). SoyBase is supported by the USDA Agricultural Research Service. The 3K Rice Genomes Project was supported by the Bill & Melinda Gates Foundation, the Chinese Academy of Agricultural Sciences, and the International Rice Research Institute (IRRI). Gramene is supported by the National Science Foundation (NSF) and the USDA National Institute of Food and Agriculture (NIFA).

Software Acknowledgments

The bioinformatics tools listed in Table A6 were developed by research groups whose methodological contributions enabled the analytical pipeline implemented in this dissertation. OrthoFinder was developed by David Emms and Steven Kelly at the University of Oxford. IQ-TREE was developed by the Center for Integrative Bioinformatics Vienna (CIBIV). MAFFT was developed by Kazutaka Katoh at Osaka University. MCScanX was developed by Yupeng Wang and colleagues. The developers of these tools are acknowledged for making their software freely available to the research community.