GIRI, BISHWA KIRAN. Ph.D. Developmental genetics basis of life history variation in *Arabidopsis lyrata*. (2022)
Directed by Dr. David L. Remington. 220 pp.

Organisms differ in resource allocation and life-history strategies – an adaptive process that has reproduced great diversity of life on earth. Functional tradeoffs between growth and reproduction are an important determinant of lifetime fitness in iteroparous organisms, with optima varying by the environment. However, the developmental genetics context of the life-history tradeoff problem has been poorly studied.

*Arabidopsis lyrata*, a relative of the annual *A. thaliana*, provides an excellent model to study life-history tradeoffs' developmental and genetic basis, given its wide climatic distribution and life-history variation. Past research suggests that variation in apical dominance could be an essential aspect of life-history tradeoffs between populations. Auxin transport and signaling constitute major factors affecting apical dominance. Therefore, the primary objective of my study was to test the hypothesis that regulation of auxin transport underlies life-history variation in *A. lyrata*, specifically between two highly divergent populations, from Mayodan (North Carolina, USA) and Spiterstulen (Norway).

My first objective was to test the effects of auxin transport on life-history traits in *A. lyrata*, which showed mild evidence of variation consistent with the actual differences between the populations. My next objective was to identify cis-regulatory variation in genes within major life-history QTL mapped in a previous study using allele-specific expression (ASE) analyses in F1 hybrids. The result showed significant differences in ASE of *PIN3*, which encodes a major auxin transport regulator. Overall, this research advances our understanding of life-history variation's developmental and genetic basis and supports the hypothesis that developmental variation in early life stages can be a key mechanism governing plant life-history tradeoffs.

# DEVELOPMENTAL GENETICS BASIS OF LIFE HISTORY VARIATION IN *ARABIDOPSIS LYRATA*

by

Bishwa Kiran Giri

A Dissertation

Submitted to

the Faculty of The Graduate School at

The University of North Carolina at Greensboro

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Greensboro

2022

Approved by

Dr. David L. Remington
Committee Chair

# DEDICATION

*My education to this highest level would not have been possible without family, friends, and mentors I have in my life. This dissertation is dedicated to the people who have supported me though out this journey: my beautiful wife Priyanka Shrestha, my parents (Balu and Banita), Mom's Sister (Rabindra), Uncle (Kamal, Ritu), brothers (Pankaj, Paras, Prasanna) and cousins, my first friends in the USA and their family (Matt, Aram and their family).*

*Thanks to you all !!!*

APPROVAL PAGE

This dissertation written by Bishwa Kiran Giri has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. David L. Remington

Committee Members

Dr. Olav Rueppell

Dr. Malcolm D. Schug

Dr. Gloria K. Muday

May 3, 2022

Date of Acceptance by Committee

April 26, 2022

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to thank my dissertation advisor Dr. David L. Remington, who graciously accepted me into his research lab and provided an opportunity into the field of evolution and genomics and helped me become an independent thinker. I want to thank my committee members Dr. Olav Rueppell, Dr. Malcolm Schug, and Dr. Gloria, for their guidance and support during my research. Thanks to Yasmin Simkhada, Hannah Fernandez, John Pimiento, Jennifer Figueroa, Jessica, and other undergraduate students in the research lab who helped with plant care and sample preparations. Thank you, Dr. David Battigelli and Megan Corum, for your help with sample storage and the use of the research space during DNA and mRNA extraction. Thanks to Dr. Tomkiel for his encouragement and helpful discussion during my teaching and research. Thanks to Dr. Martin Jones (author of the book "Python for biologists"), who helped me start my career as a python developer. Thanks to Bhuwan Aryal, who helped refine and sharpen my coding skills. Thanks to Dr. Matt Marshall for being my buddy and engaging in scientific discussions. Thanks to Matt Miller for providing me with reading and training materials on R and Python. A big thanks to the UNCG Biology department for supporting my research and graduate career and providing me with the teaching and research assistantships, tuition stipends, and support that helped me complete this advanced degree. Thanks to Dr. Steimle and the Graduate School UNCG for supporting me when I did not feel well and had to take a break from my study. Thanks to Dr. Outi Savolainen. Dr. Päivi H. Leinonen and Dr. Tiina Mattila for sharing genomics data. Thanks to Dr. Detlef Weigel and his lab members for providing phased haplotype data. Thanks to Dr. Michael Love (author of DESEQ2), Gary Churchill (The Jackson Laboratory), and Dr. Alexander Dobin (author of rnaSTAR) for providing insights and comments on the questions I asked (about sequence alignment and ASE

analyses). Thanks to several individuals on StackOverflow, BioStars, and GitHub community for providing QA (Questions and Answers) support and feedbacks on the questions I asked and answers I wrote, during my research. And, finally, thanks to all I have missed inadvertently to acknowledge, and those who have been helpful and have brought joy and thoughtfulness during my career at UNCG, either inside or outside the UNCG school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF PROTOCOLS

LIST OF EXAMPLES

LIST OF FIGURES

# CHAPTER I: INTRODUCTION AND BACKGROUND

## Life History Evolution

Organisms differ in their life history strategies which has significant implications for adaptation and diversity among life forms. Life histories are a combination of traits that incorporate multiple developmental, reproductive, and survival characteristics. Phenotypic evolution results from selective advantage or neutral processes (Kimura, 1984, 1991), or both. The evolution of contrasting life histories can be shaped by both stochastic (drift) and deterministic factors (natural selection) (Wright, 1932) during and after population separation.

Resource allocation tradeoffs are central to phenotypic variation, and its fitness implications drive differences in key life-history traits between populations and taxa. Resource allocation tradeoffs emerge from the idea that organisms have limited resources and must allocate them to different functions essential for adaptation. The optimal balance of resource allocation to different functions in terms of fitness is often specific to particular environmental conditions. Natural selection favors the genotypes with the highest fitness in a particular environment and directs the population toward fitness peaks in its adaptive landscape (Fisher, 1958; Orr, 2005). The most basic life history tradeoff is current reproduction vs. maintenance and growth (Roff & Fairbairn, 2007; Stearns, 1992; van Noordwijk & de Jong, 1986). At their extremes, life histories are classified as either (1) semelparous – population or species with a single burst of reproduction that is significant and fatal, e.g., Pacific salmon (*Oncorhynchus spp.*), most annual plants; or (2) iteroparous – population/species with multiple reproductive events, e.g., virtually all birds, reptiles and mammals, most perennial plants. These categories have a broad but not definite overlap with r vs. K strategists (Young, 2010). However, it is more informative to view life histories as a quantitative continuum (Thomas et al., 2000).

Key questions about life history variation are as follows: Why is having more offspring not always selected (Charnov et al., 1973; Cole, 1954; van Noordwijk & de Jong, 1986)? Why is semelparity not always the best strategy?

Several studies have attempted to answer these questions (Brown & Venable, 1986; Charlesworth, 1971; Cohen, 1966; Cole, 1954), proposing evolutionary models and modes of selection for variation in life history. However, a direct empirical test for each model remains limited. The demographic model proposed by Cole (1954) remains widely accepted and fits well with the observed diversity of life history in nature (Charnov & Schaffer, 1973). Cole (1954) states that semelparous individuals achieve equal fitness to iteroparous individuals if they produce just one more offspring in exchange for their mortality; however, the paradox is that the iteroparous organisms seem to abundant in nature. Thus, factors affecting juvenile and adult survival must favor a balance between reproduction vs. continued survival and maintenance (Charnov & Schaffer, 1973).

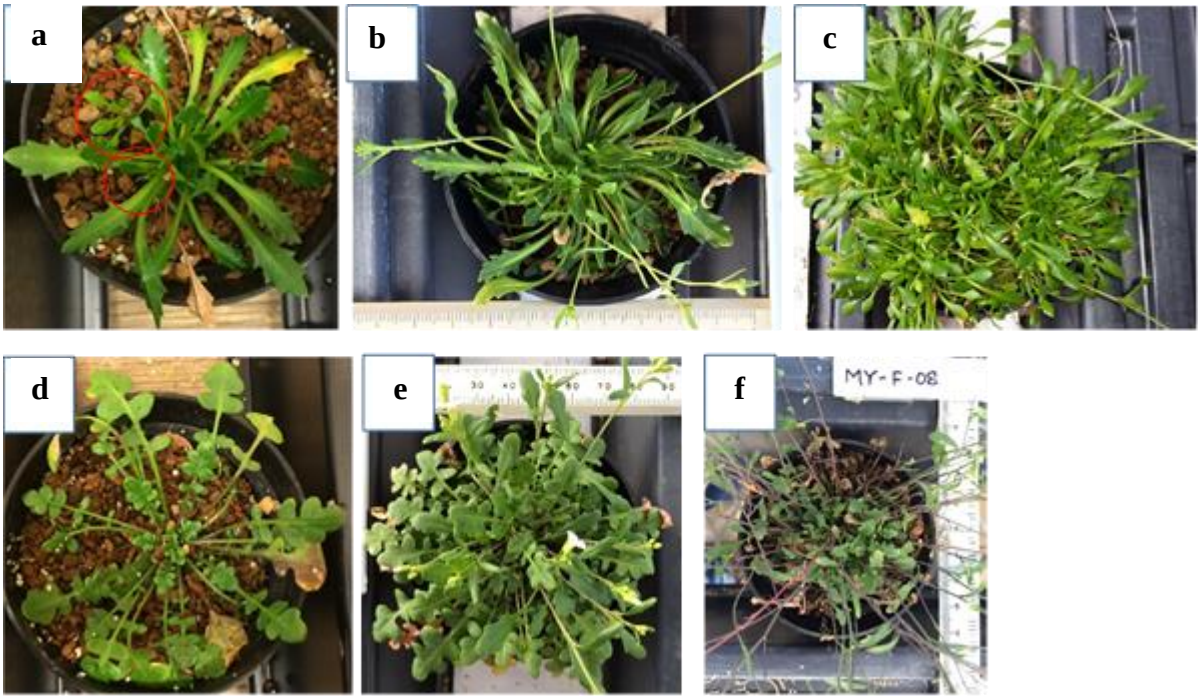
The underlying proximate and ultimate basis for life history are still unknown (Flatt & Heyland, 2011). The theoretical basis for life-history evolution is based on the principle of optimality; that one or more required resources (nutritional, temporal, the fate of the tissue or meristem) is limited, and reproduction (both quantitative and qualitative) vs. probability of survival are negatively correlated. Therefore, the organism should use the limited resource pool available for its growth during early development vs. during fecundity within a single year or at different years over a lifetime of an organism for adaptation.

Energetic tradeoffs, if nonexistent, would lead to the evolution of a "Darwinian demon" characterized by immortality, unlimited reproducibility, and survivability (Law, 1979), for which every selection strategy leads to increased fitness. In a simple sense, the nature of resource

allocation (timing, output, and bouts) then emerges from consequences of the cost of reproduction at a specific age, cost induced by selection on survivability and reproducibility at different ages of the organism (Bell, 1980; Charnov & Schaffer, 1973; Cole, 1954; Obeso, 2002; Williams, 1966). Therefore, an increase in fitness might include the promotion of one trait or intermediate values of two intrinsically constrained traits. However, observed positive vs. negative correlation between life-history traits could also be consequences of variation in acquisition vs. allocation of resources (van Noordwijk & de Jong, 1986). This raises additional questions in life-history analysis: What is the nature of resource allocation tradeoffs at the organismal level? Is the tradeoff physiological, meristematic, or hormonal? Is the tradeoff a single gene phenomenon or a functional consequence of multiple genes? How do genotype differences translate to phenotypic variation in resource allocation and fitness?

Most of the analysis of resource allocation tradeoffs has focused on genetic and phenotypic variance-covariance matrices and the identification of selective factors inducing reproductive costs. At the genetic level, reproductive costs are primarily associated with genes for resistance to disease and herbivory (Heidel et al., 2004; Holeski et al., 2010; Tian et al., 2003), which induce the cost on reproductive fitness emerging from functional constraints and also fit well with energy limitation theory. These genes affect some aspects of adaptive evolution and a certain level of variation in reproductive output but do not fully explain the evolution of semelparity vs. iteroparity.

**Figure 1.1 Figure showing Life history differences between Mayodan and Spiterstulen *A. lyrata* populations**




*Note:* Fig (a) is an early stage Spiterstulen plants; the upper red circle shows rhizomatous shoots emerging from the main plant, the lower red circle show the lateral shoot developing from the base of the main plant shoot. The later shoot are distinguishable because they showing patterns of growth that show leaf orientation outside the regular pin-wheel appearance. (b) The Spiterstulen plant during middle stage of development, and (c) during the reproductive period. The amount of lateral vegetative shoot growth, apical dominance is very obvious. (d) A Mayodan plant during the early development and (e) during the early reproductive stage and (f) around the end or reproductive season. The amount of inflorescences in the Mayodan plant is very high compared to the Spiterstulen plant. The Spiterstulen and Mayodan plants shown are not the same individuals though.

**Figure 1.2 Conceptual Representation of Limited Resource Availability and Allocation to Different Functions**

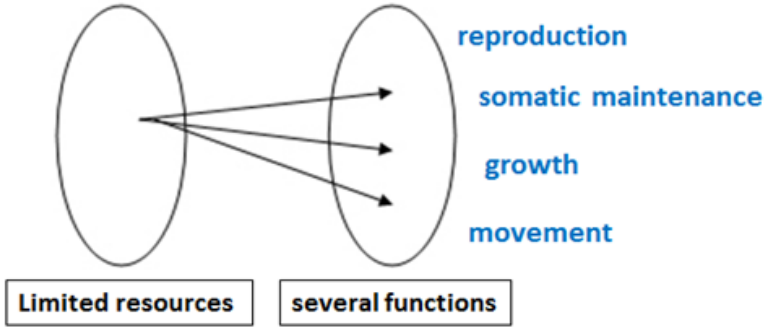
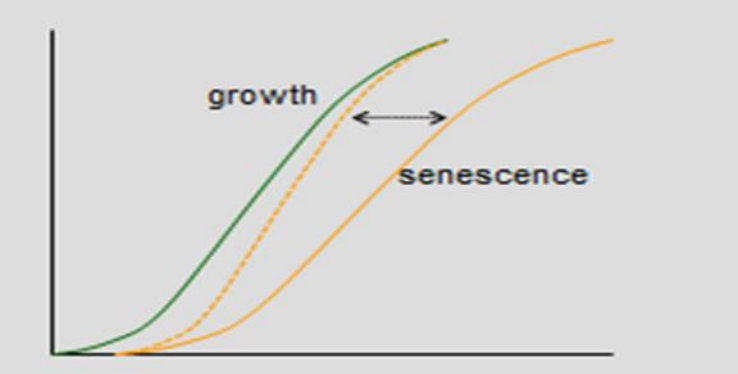



**Figure 1.3 Turnover (Production Vs. Consumption)**




*Note:* Death results when senescence catches up with growth. The green curve shows growth trend and yellow curve shows rate of senescence. The dotted yellow curves shows a case when the rate of senescence increases and catches up with the growth rate. Source: Based on Thomas (2013) New Phytol. 197: 696–711. Modified version: (Remington et al., 2015).

One essential life-history trait in plants is flowering time, which is equivalent to the mating season in animals. The genetics and ecology of flowering have been extensively studied in semelparous *A. thaliana* (Baker et al., 2005; Callahan et al., 2005; McKay et al., 2003; Munné-Bosch, 2008; Scarcelli et al., 2007; Thomas, 2004; Wilczek et al., 2009), and novel genes are still being explored. However, reproductive tradeoffs in this model occur in the limited sense

of the age of reproduction vs. size and reproductive output. For example, Thomas et al. (2000) suggests that annuality vs. perenniality in plants results from the balance between forwarding apical growth against following tissue death **(Figure 1.3)**. In flowering plants, this involves the production of new shoots vs. their more-or-less irreversible transition to reproduction.

**Resource Allocation Tradeoff Is an Integrated Complex Phenotype**

Developmental morphogenesis is characterized by changes in physical and chemical states over time. It is vital to realize that these developmental translations follow a general trend – a phenotype later in development is an output of a complex network of fine phenotypes that originate earlier in development. The optimum fitness strategy and level an organism gains is an output of the collection of a complex network of a cascade of developmental traits and the underlying genetic variation and environmental interaction rather than one factor affecting trait variations and fitness in a mendelian fashion. (Thomas et al., 2000).

Resource-allocation is an emergent property of many sub-phenotypes like number of inflorescences (reproductive shoots) vs number of vegetative shoots, flowers per inflorescence, siliques per inflorescence, number of seeds per silique, and the number of viable seeds. These sub-phenotypes grow out of a developmental cascade and are shaped by environmentally specific tradeoffs. Observed complexity in resource allocation results from integrating these sub-phenotypes from both different and same life cycle stages. Therefore, resource allocation – a highly emphasized but coarsely characterized life-history trait- can be considered an "integrated complex phenotype," which stands on the top of several adaptive features that provide basis for its existence. Resource allocation tradeoffs are fundamental in life-history analysis because they represent a developmental continuum where multiple components of fitness are eventually integrated. Understanding how genetic variation and environment shape the trajectory of

developmental cascades and their fitness consequences provides a holistic approach to understanding life-history evolution.

### *Arabidopsis lyrata* as a Model Organism for Understanding Perenniality

*Arabidopsis lyrata* provides an ideal system for understanding perenniality due to its wide climatic range, adaptive variation in life history, and availability of extensive genomic resources. *A. lyrata* (L.) O'Kane and Al-Shehbaz, an outcrossing perennial, separated from its annual relative *A. thaliana* about 10 million years ago (Hu et al., 2011). *A. lyrata* has a wide circumpolar but highly fragmented population distribution throughout the Northern hemisphere and grows primarily in low competition habitats. Each population experiences contrasting climate and ecology ranging from warm temperate to subarctic and alpine regimes (Clauss & Koch, 2006; Leinonen et al., 2009; Mitchell-Olds & Schmitt, 2006). Climatic variation, including edaphic and other conditions at the local site, shape the adaptive strategy and life history of *A. lyrata* populations (Schmickl et al., 2010). *A. lyrata* is distributed into two main gene pools (recognized subspecies): *A. lyrata* ssp. *lyrata* spreads across N. America from the great lakes region and southern Appalachian mountains to adjacent foothills, and *A. lyrata* ssp. *petraea* is found across northern Eurasia and Alaska. The Eurasian lineage contains higher genetic diversity suggesting that *A. lyrata* originated there. The Eurasian lineage probably expanded northward into Scandinavia and the British Isles and then eastward across Siberia and then into N. America through Beringia (a landmass that once connected Asia and North America during several glaciation periods and now exists as the Bering Strait). The expanding population faced isolation by climatic changes during the LGM (last glacial maximum). Surviving populations later contributed to the most recent expansion of *A. lyrata* lineage from unglaciated parts of the eastern Austrian Alps, arctic Eurasia, including Amphi-pacific Beringia (Koch &

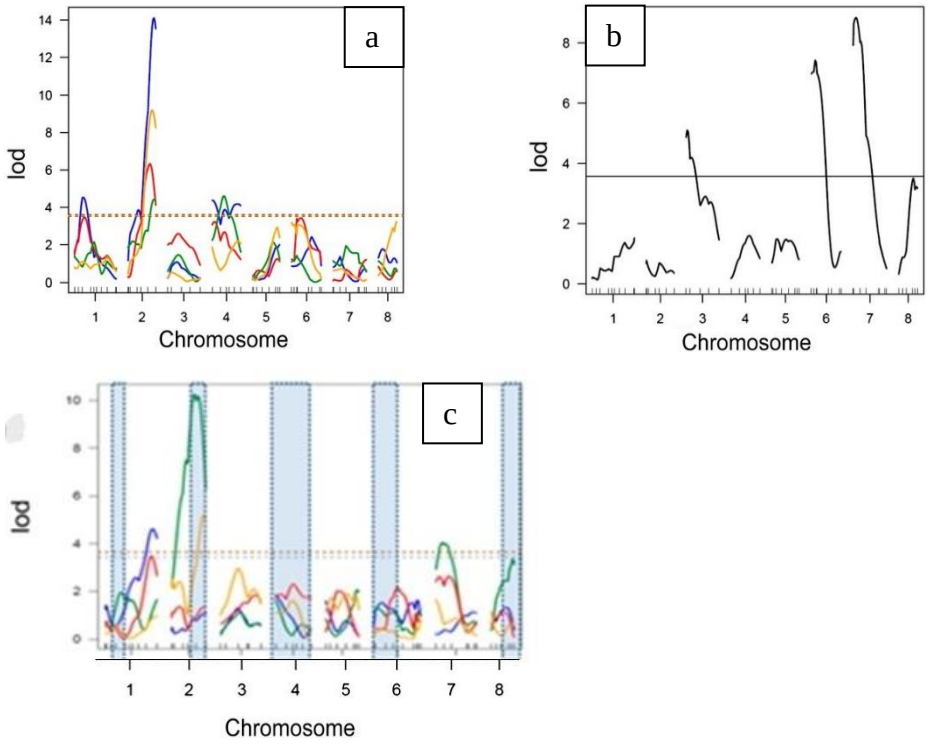
Matschinger, 2007; Ross-Ibarra et al., 2008; Schmickl et al., 2010). Following glacial expansion and retreat episodes, surviving refugia populations have now spread into their respective continents and structured the genetic diversity of the *A. lyrata* lineage (Schmickl et al., 2010). The North American lineage (*A. lyrata* ssp *lyrata*) represents a derived lineage state that experienced severe genetic bottlenecks during the expansion process, resulting in lower genetic diversity (Mattila et al., 2017; Pyhäjärvi et al., 2012; Ross-Ibarra et al., 2008). Local adaptation and contrasting variation in adaptive traits and vast geographic and climatic differences at their native sites make the populations of *A. lyrata* an excellent system for investigating adaptive variation, gene flow, and life history evolution.

### *Arabidopsis lyrata* Populations With Contrasting Resource Allocation Patterns Provide a Good Model for Genetic Analysis of Life-History

Several studies have revealed substantial phenotypic differentiation for adaptive and life history characteristics between populations of *A. lyrata* (Karkkäinen et al., 2004; Leinonen et al., 2009, 2011, 2012; Remington et al., 2013; Riihimäki & Savolainen, 2004; Sandring et al., 2007; Sandring & Ågren, 2009; Turner et al., 2010; Vergeer & Kunin, 2011) including variation in flowering time and flower morphology (Riihimäki et al., 2005; Riihimäki & Savolainen, 2004; Sandring et al., 2007). The *A. lyrata* populations we have been working with have broader relevance in testing the underlying genetic and functional basis of perenniality because they are native to sites that represent opposite extremes of the environmental variation (in terms of both the temperature and photoperiod) across the range of *A. lyrata* distribution and also occupy opposite extremes of the resource allocation continuum . One of our study populations is native to an alpine valley in Spiterstulen Norway (*A. lyrata* ssp. *petraea*) in Europe (61◦ 38_N, 8◦ 24_E,1106 m.a.s.l.) and experiences a shorter growing season with lower annual mean

temperature (0.87$^0$C; Lom, Norway, 10-year average; Norwegian Meteorological Institute). Another population is local to Mayodan NC (*A. lyrata* ssp. *lyrata*), near the southern range limits of *A. lyrata* in the United States (36◦25_ N, 79◦58_ W, 225 m.a.s.l.) and faces a longer growing season (14.5$^0$C; Greensboro, NC, 30-year average; U.S. National Weather Service). Mean annual precipitation in NC is 1092 mm and only 461 mm at the Norway site, but the growing season in North Carolina faces periodic summer drought. Life histories between these two populations vary broadly, which includes flowering time, length of flowering, number of reproductive shoots, siliques per shoot, and number of flowers (Leinonen et al., 2012; Remington et al., 2013) and is reflected in the pattern of apical dominance and shoot architecture these two populations exhibit. Based on microsatellite data, our study populations are highly differentiated (Fst = 0.668) (Muller et al., 2007) with an estimated divergence time between North American and Central European populations of about 260,000 years (Mattila et al., 2017; Toivainen et al., 2014).

**Figure 1.4 QTL Mapping Results for the North Carolina Field Site (a & b) And For The Norway Field Site (c)**




*Note:* Fig a LOD profiles for vegetative and reproductive traits: spring diameter (blue), reproductive shoots (red), siliques per shoot (green), and net reproductive season diameter growth (orange). Fig b LOD profiles for square root transformed flowering dates (Leinonen et al. 2013). Horizontal lines represent genome-wide $P = 0.05$ significance thresholds. Fig c LOD profiles for vegetative and reproductive traits in Norway environment. Source: Remington et al. 2013.

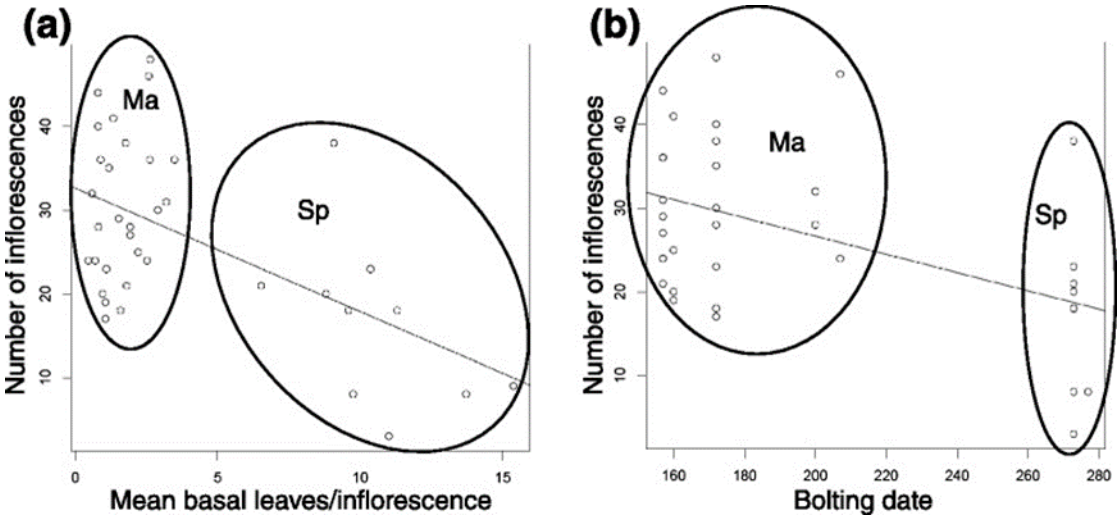
**Life History Differences in *A. lyrata* Populations Are Related to Developmental QTLs**

Reciprocal transplant study of Mayodan and Spiterstulen populations have shown evidence of local adaptation characterized by higher survival of Spiterstulen alleles when grown in Norway and higher reproductive output by Mayodan alleles when grown in North Carolina, at the expense of vegetative growth at several resource allocation QTLs (LG or chromosomes 1, 2, 4, 8) **(Figure 1.4, a and c)**. Importantly, these QTLs were largely independent of flowering time QTLs **(Figure 1.4, b).** More importantly, these resource allocation QTLs and not flowering time QTLs were the primary loci contributing to local adaptation (Leinonen et al., 2012; Remington et al., 2013).

Mayodan alleles in both environments showed a larger number of inflorescences and higher flowering propensity, suggesting that intrinsic genetic components (Mayodan alleles) that influence life history are robust to environmental changes and drive plant development more toward reproductive investment. Mayodan alleles also resulted in a much greater loss of vegetative rosette diameter during the reproductive season than Spiterstulen alleles. In other words, they contributed to growth vs. reproduction tradeoffs. Also, Mayodan alleles at the LG2 QTL region delayed lateral shoot development before flowering, indicating greater apical dominance by Mayodan alleles. The LG2 primarily showed the most extensive effects (77% of parental mean difference in the number of fruits per inflorescence) of all chromosomes in North Carolina **(Figure 1.4, a)** as well as antagonistic tradeoffs in both the environments (Leinonen et al. 2012; Remington et al. 2013) making it an interesting candidate for further analysis. However, whether these resource allocation QTLs, especially LG2, harbors single or multiple genetic components with tradeoff functions has not been tested. This LG2 QTL region is syntenic to *A. thaliana* (At1g68060–74600) genome, which contains the *A. lyrata* orthologs of

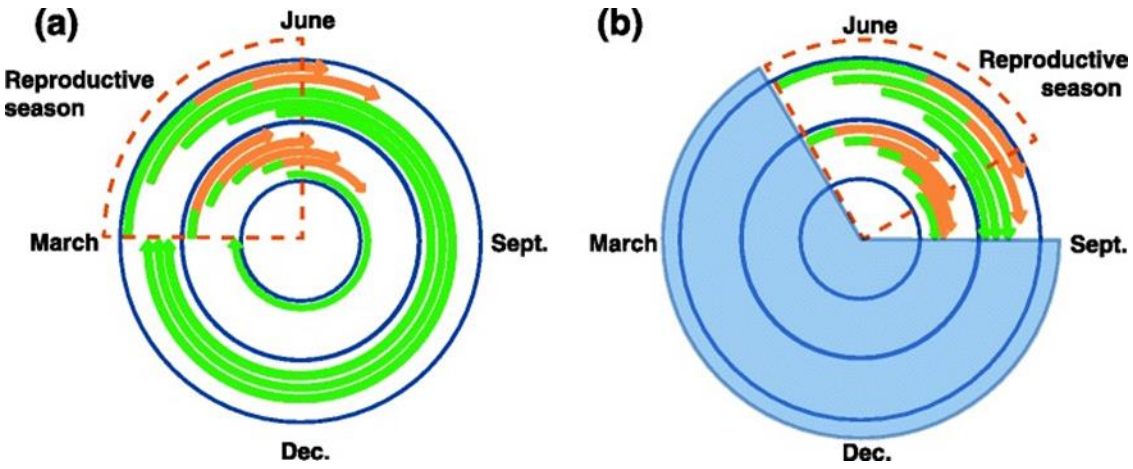
genes encoding two major auxin efflux carriers, *PIN1* (At1g73590) and *PIN3* (At1g70940), a recently identified regulator of intracellular auxin homeostasis, *PILS2* (At1g71090) (Barbez et al., 2012), and *BRC2*, a homolog of *tb1* gene from maize which has dramatic effects on shoot architecture and sex determination in the tip of maize (Doebley et al., 1995, 1997).

**Figure 1.5 (a) Scatterplots Showing the Regression of the Number of Inflorescences on the Mean Number of Basal Leaves per Inflorescence, and (b) the Day of First Bolting in the Same Plants**




*Note:* Dotted in the ellipses show the phenotypic distributions (number of inflorescences) in Spiterstulen (Sp) and Mayodan (Ma) plants. Source:(Remington et al., 2015).

**Figure 1.6 A Proposed Phenology Model Showing Variation in Growing Seasons in Two Extreme Environments of *A. lyrata* Habitat**




*Note:* (a. North Carolina environment; b. Norway environment. Source: (Remington et al., 2015).

A growth chamber study comparing development in Mayodan vs. Spiterstulen plants in 2015 showed that individual lateral shoots became more rapidly reproductive on Mayodan plants than on Spiterstulen, which explains the higher propensity of Mayodan towards reproduction (Remington et al., 2015). Overall, this and several other studies suggest that local adaptation in perennials, including *A. lyrata*, appears to have much more to do with flowering transitions on individual shoots, not on the date of first flowering as in annuals.

### *Arabidopsis lyrata* Growth and Development

After germination, *A. lyrata* undergoes vegetative development producing highly compressed shoots which appear as a pinwheel. The lateral vegetative shoots have meristems in their axils for further vegetative or reproductive tissue development during the reproductive episode. During reproductive season individuals from Mayodan populations produce long, slender, and large inflorescences, followed by leaves. In contrast, Spiterstulen individuals have fewer inflorescences, notably thicker in diameter. Spiterstulen plants also make many lateral

vegetative shoots before bolting, indicating that Mayodan individuals exert greater apical dominance than Spiterstulen individuals.

On one hand, increased apical dominance can potentially reduce branching, thereby reducing the number of meristems available for both vegetative reproductive growth during the reproductive episode. On the other hand, the observed greater apical dominance in Mayodan plants is associated with more reproduction. We conjecture that the timing of lateral shoot development becomes more compressed under strong apical dominance in Mayodan individuals, with new shoots switching over to reproductive much more rapidly after they initiate (Remington et al., 2015). Similar relationships between the timing of lateral shoot initiation and reproductive output were found in *Arabis alpina* (Wang et al., 2009) and *Erysimum capitatum* (Kim & Donohue, 2013). This information from several studies suggests that the developmental genetic basis of life-history tradeoffs we discover in *A. lyrata* may be more broadly prevalent in perennial plants.

**Auxin as a Candidate in Life-History Evolution**

Shoot architecture is strongly influenced by auxin dynamics (J. Friml, 2003; Petrásek et al., 2009) and is controlled by auxin transport proteins and other interacting proteins and genes. Genes involved in auxin biosynthesis, homeostasis, and gradient maintenance include *AUX1, PIN1 - 7, TAA1, YUCCA, GH3, PGP, TPL,* and *TIR1*. Homology, gene function, and genetic pathway for these candidates have been conserved between distant monocots and eudicots relatives (Gallavotti, 2013). *PIN (PIN FORMED)* genes have been extensively studied in *A. thaliana* (J. Friml, 2003; Petrásek et al., 2009), which derives its name from the pin-like phenotype in *PIN1* mutants (Okada et al., 1991). Different *PIN* proteins are asymmetrically localized in plants parts which direct auxin transport and gradient. This variation in auxin

gradient shapes developmental aspects like organogenesis, morphogenesis, and meristem patterning. (Gälweiler et al., 1998; Prusinkiewicz et al., 2009; Vieten et al., 2005). Mutations affecting the nature of apical dominance can have a pervasive effect on patterns of shoot development and not just shoot initiation (Barbez et al., 2012; J. J. Friml et al., 2002; Gälweiler et al., 1998; Prusinkiewicz et al., 2009), which could potentially translate into phenotypic variation in resource allocation tradeoffs. Auxin transport is also highly sensitive to environmental cues. Individuals with the same genetic background can produce different transport responses in other growth conditions (Lewis & Muday, 2009), suggesting an ecological influence on auxin dynamics. Therefore, variation in auxin transport between separated populations is a likely candidate that could explain the environmental and evolutionary basis of shoot architecture variation and ultimately life history variation. And *PIN1* and *PIN3* genes encode auxin transport proteins and are located in the largest-effect QTL region on LG2 of *A. lyrata*.

### *TB1* as a Candidate in Life-History Evolution

Another molecular candidate that has significant effects on branching is tb1 (teosinte branched 1), first identified in maize (Doebley et al., 1995) and its homolog PCF1 and PCF2 in rice (Choi et al., 2012) and in Arabidopsis BRC1 (BRANCHED 1) (Aguilar-Martínez et al., 2007; Poza-Carrión et al., 2007) and BRC2. These genes and orthologs are widely found in flowering plants, including A. thaliana, sorghum (Kebrom et al., 2006), and tomato (Martı´n-Trillo et al., 2011). A good example of change in apical dominance and its influence on shoot architecture can be observed in domesticated maize vs. its wild-type relative teosinte. This phenotypic change has been traced to tb1 (teosinte branched1) regulation, which affects several traits like the number and length of internodes and sex of the inflorescences on the tip of the

primary lateral branches in maize (Doebley et al., 1995, 1997). The pattern of tb1 expression and morphology of tb1 mutant plants show that tb1 acts to repress the growth of axillary organs and enable the formation of female inflorescences (Doebley et al., 1997). Both tb1 and BRC1 contain a highly conserved TCP domain (Aguilar-Martínez et al., 2007), facilitating DNA-protein dimerization and cell division (Cubas et al., 1999; Poza-Carrión et al., 2007). Transcription factors containing the TCP domain are highly regulated in growing flower primordia in A. thaliana (Cubas et al. 1999), confirming its role in organ elongation and meristem growth. BRC1 is also found to interact with endogenous auxin signal pathway, and carotenoid derivative produced in roots (Aguilar-Martínez et al., 2007; Niwa et al., 2013) and promotes bud dormancy in response to shading (González-Grandío et al., 2013). TCP regulation also responds variably to environmental stimuli like planting density and influences the number of branches. BRC1 is also found to interact with flowering locus (FLC) to regulate floral transition in axillary meristems (Niwa et al., 2013). PEP1 (perpetual flowering) in Arabis alpina (Wang et al., 2009), an ortholog of FLC (flowering locus), has pleiotropic effects on shoot branching, the number of meristems allocated to flowering, and return to the vegetative state. This highlights the role of BRC1 in shoot architecture, possibly by its interaction with auxin signals and other candidates involved in apical dominance. One of the significant reasons genes like tb1 homologs could impact resource allocation is that extensive branching is positively correlated with the number of reproductive structures. This suggests that tb1 like genes may influence the nature of meristem (reproductive vs. vegetative) and determines reproductive output by a functional mechanism that might be conserved in distant eudicots relatives. Importantly, BRC2 has more minor effects and different regulations than BRC1 in A. thaliana (Aguilar-Martínez et al., 2007) in short-cycling lab accessions. However, we speculate that role of BRC2 could be greater in perennial plants that

exhibit more prolonged vegetative states. Notably, the BRC2 gene also lies in the QTL LG2 related to local adaptation in *Arabidopsis lyrata*.

**Why Are Genes Involved in Apical Dominance Important?**

The evolution of similar traits in distinct lineages often involves mutations in the same gene (a phenomenon called "gene reuse") (Martin & Orgogozo, 2013), representing genetic hotspots of evolution. These de-novo mutations primarily include gain-of-function events at orthologous sites and produce similar phenotypic variation, representing important evolutionary mechanisms shaping adaptive evolution. Furthermore, repeated phenotypic evolution involves mutations at cis-regulatory sites more often than changes in coding regions (Stern & Orgogozo, 2008), which can promote rapid phenotypic response by altering the level of gene expression. The TB1 gene and its orthologs in other plant lineages represent a potential genetic hotspot where a similar phenotypic response may have evolved when demanded by the forces of selection. Also, PIN genes and their homolog are highly conserved in the plant kingdom in regulating diverse plant functions, suggesting PIN genes as a potential candidate for shaping plant life history.

**Genomic, Transcriptome, and Phenotypic Databases of *Arabidopsis thaliana* and *Arabidopsis lyrata* Provide Valuable Resources for Genetic Analysis of Perenniality**

*Arabidopsis thaliana*, a functional genetic model, has been very helpful in identifying key genes for several phenotypic traits in plants. Its adoption has helped establish sophisticated genetic tools, molecular techniques, and transcriptome databases for plant biology in recent decades http://plants.ensembl.org/*Arabidopsis_lyrata*/Info/Index, http://www.phytozome.net/, http://genome.jgi-psf.org/Araly1/Araly1.home.html. Despite the benefits of single candidate gene analyses, which can pinpoint specific life-history functions, lab models may be limited in

elucidating the novel aspects of life-history evolution as the organisms do not fully experience the constraints induced by the environment (Anderson et al., 2011). Also, genes conferring fitness advantage in one particular environment can have negative fitness consequences in another set of conditions (Heidel et al., 2004). Additionally, *A. thaliana* does not fit as an appropriate model for analyzing the genetic basis of perenniality due to its strictly annual nature. The study of emerging model systems in the natural environment will instead overcome these limitations and identify novel loci and genes that are equally important to life history analysis (Anderson et al., 2011). Apart from being perennial, the overall developmental aspects of *A. lyrata* are highly similar to that of *A. thaliana* (Grbic & Bleecker, 2000), which makes the phenotypic analysis, molecular assays, and genetic tools developed in *A. thaliana* system directly applicable to *A. lyrata*.

Studying the genetic basis of perenniality in *A. lyrata* benefits from an extensive genetic and genome database available for both species at http://www.*Arabidopsis*.org/. Also, the comparative genetic and syntenic maps between these sister species (Hu et al., 2011) and transcriptome database have been developed and available at (http://www.plantgdb.org/), and a transcription factor database is available at http://planttfdb.cbi.pku.edu.cn/help_datasrc.php. These databases provide an essential resource for comparative genetics and functional analyses to test our questions and hypothesis regarding perenniality in the *A. lyrata* model. Moreover, sequence data of the whole genome sequence of several new populations of *A. lyrata*, including the ones involved in our study, are now available to assist further in-depth analysis of the signature of selection, genetic differentiation, and adaptive polymorphisms that have evolved between populations (Mattila et al., 2017).

## The Rationale for This Research

The fundamental reason for this research is to understand the developmental genetic basis for QTLs affecting adaptive life-history variation in *A. lyrata*, which could provide novel insights into the traits underlying adaptive evolution in perennial plants. This is comparable in importance to the role of flowering time variation and the underlying genetics in annual plants. As discussed earlier, the genomic, transcriptomic, and phenotypic database for *A. thaliana* and *A. lyrata* provides a valuable resource and foundation to pinpoint the ultimate and proximate bases of life-history variation, including candidate developmental mechanisms and candidate genes underlying important QTL regions.

## Dissertation Goals

### Dissertation Goal 1; Chapter II

Determine whether differences in apical dominance and shoot architecture observed in the Mayodan and Spiterstulen populations can be explained by variation in the rate of auxin transport.

I tested variation in auxin transport rates in the plants that belonged to Mayodan and Spiterstulen populations grown in a controlled environment. To quantify the variation in transport rate, I used radiolabeled 3H-IAA (a synthetic auxin) in the inflorescence shoots and tested differences in radioactivity.

I applied NPA (an auxin inhibitor) to the plants from the Mayodan population to test for its effects on life-history traits. First, I measured several traits over the growing season at three different time points. The traits data I collected are diameter, changes in diameter, observed lateral shoot rating, and the number of inflorescences. Then I compared the NPA treated plants with non-treated control.

**Dissertation Goal 2; Chapter III**

Develop algorithms and tools for phasing and assigning haplotypes in outcrossing populations.

I developed three different methods to help with phasing haplotypes for unphased genotype and read-backed-phased genetic variants data. The three tools/algorithms are Phase-Extender, Phase-Stitcher, and ShortVariantPhaser and are designed to handle three different types of data structure generated in concurrent variant genotyping.

I also tested one of the tools (Phase-Extender) against the haplotype phasing tool ShapeIT and found that Phase-Extender can phase variants on par with ShapeIT using a small number of samples. The tools also provide a more controlled approach to haplotype phasing. It has the potential to be a good utility when phasing genomes that do not have a large number of reference haplotypes or in the situation when a small number of sample cohorts are only to be used for haplotype phasing.

**Dissertation Goal 3; Chapter IV**

Identify candidate genes underlying a key life-history QTL region by evaluating quantitative variation in expression of alleles from Mayodan and Spiterstulen genomes.

I established some F1 hybrid samples by crossing parental populations, Mayodan and Spiterstulen. First, I extracted total mRNA from the whole shoot of these F1s during the late vegetative stage. Then, the allele-specific expression in each F1 was quantified by measuring read counts by aligning the read competitively against a personalized diploid genome.

The statistical test of variation in allele expression was done using the DESEQ2 (Love et al., 2014) package.

# CHAPTER II: AUXIN TRANSPORT INHIBITION IN *ARABIDOPSIS LYRATA* AFFECTS SHOOT ARCHITECTURE AND LIFE-HISTORY TRAITS


## Abstract

Variation in apical dominance influences shoots architecture differences in plants. The two populations of *A. lyrata* (Mayodan and Spiterstulen) show two opposing spectra of life history properties, with Mayodan showing higher apical dominance and investment of meristems to reproductive shoots and tissues. In this research, we tested if differences in auxin transport explained those variations in apical dominance between two study populations. We further tested if reducing apical dominance in Mayodan altered its life-history traits (lateral shoot development and quantity of inflorescence) consistent with one seen in the Spiterstulen population. Results show mild evidence of higher auxin transport in Mayodan individuals (*F-statistic 4.082, P-value = 0.053*). Inhibition of auxin transport in Mayodan using NPA reduced apical dominance and increased the production of lateral vegetative shoots and showed a trend toward fewer inflorescences.


## Introduction

### Plant Hormones

Plant hormones are a class of chemical compounds that occur naturally in the plant or are artificially synthesized. Plant hormones are effective even at small concentrations and influence different physiological processes, mainly growth, differentiation, and development.

The idea that some chemical substances regulate plant growth and development was introduced long before discovering plant hormones. Sachs hypothesized that the chemical substances associated with plant growth might have distinct movement patterns throughout the plant (Enders & Strader, 2015). Around the same time, Charles Darwin and Francis Darwin were

experimenting with the effects of light and gravity on the growth of grass coleoptiles. Boysen-Jensen and Paál further improved the idea of plant growth regulators that helped to develop Cholodny–Went hypothesis (Thimann, 1988). This hypothesis suggests that auxin is the plant growth regulator asymmetrically distributed in the plant and results in tropism in root and shoot governed by the stimulus of light and gravity (Enders & Strader, 2015; Trewavas, 1992). This idea was further explored and documented by Fritz Went (Thimann, 1988). These early investigations in plant hormones involved dissecting the role of auxin in tropic growth responses, how auxin controls cell elongation, and appropriately reorients growth of plant organs in response to environmental stimuli.

The growth-promoting agents in plants are also known as phytohormones. Now, almost a century of research has resulted in the discovery of several phytohormones, including auxins, ethylene (ET), cytokinins (CK), gibberellins (GA), abscisic acid (ABA), brassinosteroids (BRs), jasmonic acid (JA), salicylic acid (SA), and the recently identified strigolactones (SLs) (Checker et al., 2018). These phytohormones work independently or respond appropriately against environmental or developmental signaling. The first plant hormone discovered and the major one, auxin, causes a growth response far from its synthesis site and qualifies as a chemical messenger (Wang et al., 2015). Auxin plays a crucial role in plant growth and developmental processes like tropism, embryogenesis, and organogenesis (Davies, 2010).

**Polar Auxin Transport**

One of the most studied and predominantly found plant hormones is indole-3-acetic acid (IAA), which helps determine the formation of primary and lateral apices, differentiation of vascular tissue, regulation of root system architecture, and embryo development (Casimiro et al., 2003; Kondhare et al., 2021). IAA in shoots exhibits a phenomenon called "polar auxin

transport," i.e., it moves unidirectionally from the apex to the base of the shoot. However, IAA transport in roots shows two distinct polarities. First, IAA moves from the root apex towards the base (Meuwly & Pilet, 1991). On the other hand, in the central cylinder of the roots, IAA move acropetally towards the root apex (Jones, 1998; Kerk & Feldman, 1995; Tsurumi & Ohwaki, 1978).

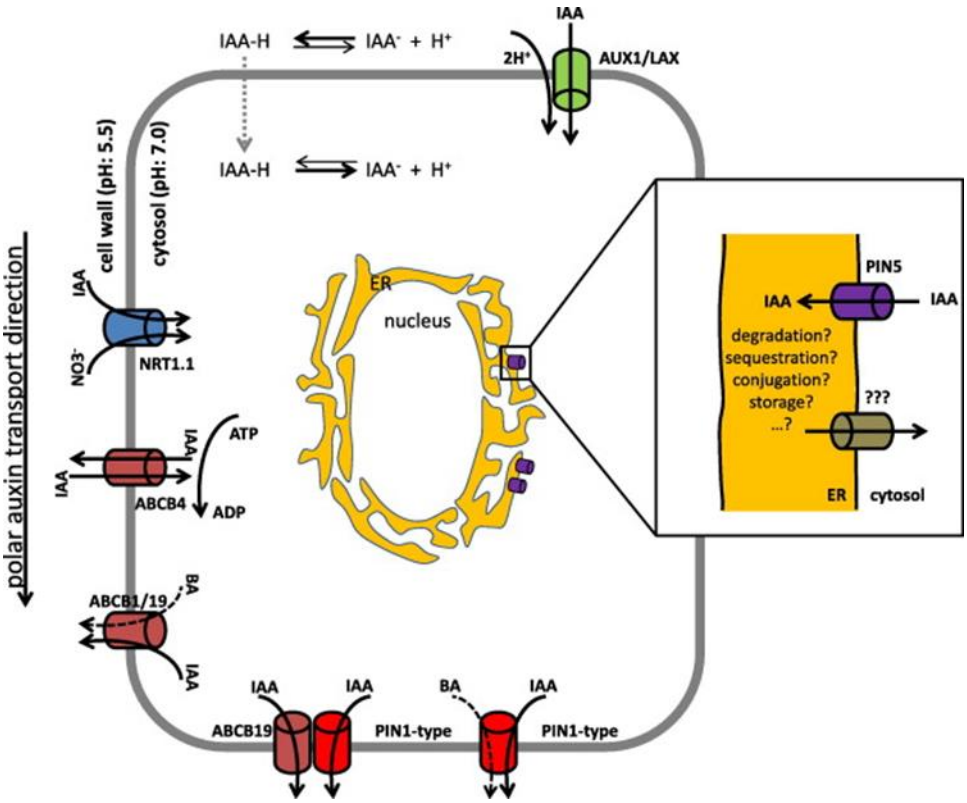
Another naturally occurring auxin is indole-3-butyric acid (IBA), which helps develop an adventitious root system. Some synthesized auxins are 2,4-D (2,4-dichlorophenoxyacetic acid) and NAA (1-naphthaleneacetic acid). These synthetic auxins are extensively used in plant tissue culturing to determine and manipulate plants' growth responses (Flasiński & Hac-Wydro, 2014). Auxin's polar movement through developing tissue and established concentration gradient triggers its mechanism of action. Moreover, this information also helps determine and analyze the positional information for the spatial modulation of gene expression patterns (Casimiro et al., 2003).

Most investigations into the role of IAA in plants are done in *Arabidopsis thaliana*, a biological model for molecular and genetics research in plants. In *Arabidopsis,* both polarities have been detected depending on specific physiological processes (A. M. Rashotte et al., 2000; Reed et al., 1998). Shoot basipetal movement of IAA confers apical dominance of the main shoot, promoting upward growth of the plant and reducing bushiness. On the other hand, the acropetal move of IAA from the shoot into the root has been shown to control the roots' lateral growth (Reed et al., 1998). However, basipetal movement of IAA is required for gravity response (A. M. Rashotte et al., 2000) and has also been suggested to affect the initial cell divisions during lateral root initiation (Casimiro et al., 2001).

The directional transport of auxin is mainly attributed to the family of auxin efflux carriers (the PIN proteins). The acronym "PIN" derives its name from the pin-like phenotype in *pin1* mutants (Okada et al., 1991). Most of the information about *PIN* (*PIN FORMED*) genes, their function, and their role in triggering developments have been generated through extensive genetic research in *Arabidopsis thaliana* (Friml, 2003; Petrásek et al., 2009). The unique location of PIN proteins in the cell membrane is responsible for maintaining the property of polarity in auxin transport. In addition, PIN proteins can also re-localize in response to the changes in auxin concentration gradients, which frequently happens during organ initiation (Leyser, 2005, 2009; Paciorek et al., 2006; Waldie & Leyser, 2018; Wisniewska et al., 2006).

Extensive research on *Arabidopsis* has led to the discovery of eight PIN proteins. These 8 proteins are divided into two types based on their molecular weights. One type is a long-looped *PIN* (*AtPIN* 1, 2, 3, 4, and 7), and the other is a short-looped *PIN* (*AtPIN* 5 and 8). The cell-to-cell transport of auxin is facilitated by the long-looped *PINs* present in the plasma membrane (Ganguly et al., 2010; Vieten et al., 2005). Several phosphorylation sites are present in the long-looped PIN-HLs, necessary for *PIN* polar trafficking. These conserved phosphorylation motifs are absent from short-looped *PINs* (Ganguly et al., 2010). Small-looped *PINs* are mainly located in the endoplasmic reticulum or plasma membrane and regulate cytosolic auxin homeostasis.

**Figure 2.1 Polar Auxin Transport – Chemiosmotic Hypothesis**




*Note:* Protonated auxin (IAA-H) readily diffuses into the cell, while non-protonated auxin ($IAA^-$) needs additional proton symporter AUX1/LAX transporters. Several PIN proteins mediate the direction of intercellular polar auxin transport. Source: (Löfke et al., 2013).

However, the functioning of short-looped *PINs* has received little attention compared to long-looped *PINs* (Ding et al., 2012; Mravec et al., 2009). Moreover, PIN6 (long-looped PIN) is uniquely present in the ER, unlike other long-looped PINs (Mravec et al., 2009). These different PIN proteins localize asymmetrically to specific faces of the plasma membrane in different parts of the plant and direct auxin transport and gradient (Friml, 2003; Petrásek et al., 2009). This process helps developmental processes like organogenesis, morphogenesis, and meristem patterning (Gälweiler et al., 1998; Prusinkiewicz et al., 2009; Vieten et al., 2005). The cellular mechanism that drives the dynamics of auxin transport is clathrin-mediated endocytosis and the

recycling of PINs which can induce rapid changes in cell polarity (Kitakura et al., 2011). Auxin itself inhibits this recycling, resulting in an accumulation of PIN proteins, specifically at the plasma membrane opposite the source, promoting its own efflux (Kleine-Vehn et al., 2011).

Environmental variation and growth conditions can also influence changes in auxin dynamics in the same genetic background and produce different transport responses (Lewis & Muday, 2009). This environmental-based variation in auxin transport and shoot architecture can then translate into variation in the numbers of reproductive shoots/meristems produced, which affect the number of flowers, and, ultimately, the reproductive output and any observable differences in life history. Therefore, exploring the variation in auxin transport between the populations that show contrasting life-history patterns could be an excellent approach to understanding the underlying environmental and evolutionary basis of life-history variation.

**Auxin Transport Inhibitors**

Auxin transport inhibitors (ATIs) are the pharmacological tools that researchers have been using for decades to understand the mechanisms underlying polar auxin transport and its impact on plant growth and development (Dhonukshe et al., 2008). Some of the ATIs are as follows: 1-naphthylphthalamic acid (NPA), 2-carboxyphenyl-3-phenylpro- pane-l,2-dione (CPD), 2,3,5-triiodobenzoic acid (TIBA), methyl-2-chIoro-9-hydroxyfluorene-9-carboxylate (CFM) (Shi et al., 2006) and 2-(1-pyrenoyl) benzoic acid (PBA) (Snyder, 1949).The exogenous application of these inhibitors alters the pattern of auxin distribution and interferes with plant development.

One of these auxin transport inhibitors, naphthylphthalamic acid (NPA), interferes with the directional auxin flow and critically affects plant growth. The application of NPA in the seedlings affects lateral root formation (Muday & Haworth, 1994; Reed et al., 1998). The

phenotypes of the mutants of the PIN family of auxin transporters (i.e., *barren stalk1 (ba1) and barren inflorescence2 (bif2)* mutants) resemble the NPA-treated plants (Wu & McSteen, 2007), which suggests that both the genes are somehow involved in a similar auxin transport pathway that NPA interferes. Moreover, recessive mutations in *TIR3* (Transport Inhibitor Response) genes show reduced polar auxin transport, followed by morphological abnormalities like short siliques, pedicles, roots, and inflorescences. It shows that the absence of TIR3 protein (i.e., NBP proteins) leads to the improper localization and distribution of IAA (Ruegger et al., 1997), thus affecting life-history traits.

**Auxin (Apical Dominance) as a Candidate in the Evolution of Life-History Tradeoffs**

Auxin coordinates shoot architecture patterning in plants by managing the development aspects of plants – quantitative variation in shoot, root, flowers, seeds, fruit, stem elongation, tissue differentiation. Auxin dynamics control organogenesis, morphogenesis, and meristem patterning at the molecular and developmental levels (Gälweiler et al., 1998; Prusinkiewicz et al., 2009; Vieten et al., 2005). These developmental level components are again essential contributors to shoot architecture at the organismal level (Friml, 2003; Petrásek et al., 2009; Remington et al., 2013). Therefore, it seems intuitive that variation in shoot architecture could be one of the key processes affecting resource allocation tradeoffs.

Treatment of plants using polar auxin transport inhibitors has shown that auxin transport is essential for leaf initiation for vegetative development and initiation of flowers primordia for reproductive development (Okada et al., 1991; Reinhardt et al., 2000, 2003). Application of auxin transport inhibitors on maize inflorescences later in development alters the phenotype of spikelets by producing single instead of paired spikelets (Wu & McSteen, 2007). Andropogoneae, which includes more than 1000 grasses (maize, sorghum, sugarcane), contain

paired spikelets, a vital feature of the sorghum tribe, while all other grasses bear single spikelets. This alteration of apical dominance and change in spikelet number after NPA application asserts the role of auxin transport in the evolution of inflorescence architecture (Wu & McSteen, 2007). In another example, alteration of apical dominance by applying endogenous Gibberellic acid reduced the number of inflorescences in 'Afterglow' Bougainvillea (Chng & Moore, 2020). These results suggest that apical dominance provides a basis for shoot architecture variation as an adaptive response and is crucial in altering life history strategy. However, how apical dominance and auxin transport mechanistically shape plant life history strategy remains to be understood in more clarity – especially the development pathways apical dominance drives, so a plant evolves towards one approach vs. another.

Research by Remington et al. (2013, 2015) showed that the diverged populations of *A. lyrata* (Mayodan from NC and Spiterstulen from Norway), which are adapted to contrasting environments, show contrasting life-history trait differences and variation in apical dominance. These differences show up as variation in the investment of plant meristems to reproductive vs. vegetative tissues (Remington et al., 2015) during early development. Analyses by (Leinonen et al., 2013; Remington et al., 2013) indicated that a QTL region on chromosome 2 (LG2) has the most prominent effects on this adaptive variation between the two populations. This region contains two major genes, *PIN1* and *PIN3,* which encode auxin transport proteins and are vital to plant morphological development affecting shoot development and allocating meristems to a particular fate.

We are particularly interested in gaining insights on the following questions. First, is there variation in auxin transport between the two populations of *A. lyrata*? Second, would changing the auxin transport using transport inhibitors like NPA alter the life-history traits?

My first objective is to test whether these study populations differ in auxin transport. We predict that quantitative measurement of auxin transport in the inflorescences of both populations using radiolabeled auxin ($^{3}$H-IAA: indole-3-acetic acid) will demonstrate a higher rate of auxin transport in Mayodan individuals (which display greater apical dominance compared to Spiterstulen). My next objective was to analyze the effects of auxin transport inhibition on shoot architecture and reproductive output in Mayodan populations. For this second objective, I specifically predicated that inhibiting auxin transport in the Mayodan populations of *A. lyrata* using the pharmacological auxin transport inhibitor 1-N-Naphthylphthalamic acid (NPA) will reduce apical dominance in Mayodan individuals. I predict that it will also induce phenotypic changes in developmental patterns, mainly in shoot architecture, which could be observed through changes in rosette diameter. I predict that it will also affect the degree of lateral shoot development, which is inversely related to apical dominance. Overall, I expect inhibition of auxin transport will make the Mayodan genotypes more Spiterstulen-like in shoot development and reproductive output. Even though this component of our research doesn't provide direct answers to our question, any similarities between phenotypic changes caused by chemical inhibition of auxin transport and life history differences between populations would indicate that the underlying genes involve auxin transport and signaling.

## Methods

### Plant Material

In this study, I used *A. lyrata* seeds from two different populations. The first population is from Mayodan, North Carolina, USA (36°25′ N, 79°58′ W, 225 m.a.s.l.), while the second population is from Spiterstulen, Norway (61° 38'N, 8° 24'E,1106 m.a.s.l.). The Mayodan seed originated from open-pollinated maternal families, collected in the field in 2010. Spiterstulen

seed consisted of four unrelated full-sib families resulting from crossings between plants grown from seed gathered in the field. The Spiterstulen seeds were obtained from Outi Savolainen (University of Oulu, Finland).

## Experiment I: Variation in Apical Dominance Between Mayodan and Spiterstulen

### *Overview*

*A. lyrata* has a highly compressed vegetative shoot, and analysis of auxin transport in the vegetative shoot is problematic. However, the nature of auxin transport (a measure of apical dominance) can be easily analyzed in the inflorescences shoot right after reproductive transition using the methods developed for *A. thaliana* (Lewis & Muday, 2009; Okada et al., 1991). It involves applying synthetic radiolabeled 3H-IAA (indole-3-acetic acid) at the apical end of the inflorescence and counting the levels of radioactivity at the basal end of the inflorescence.

### *Growing condition*

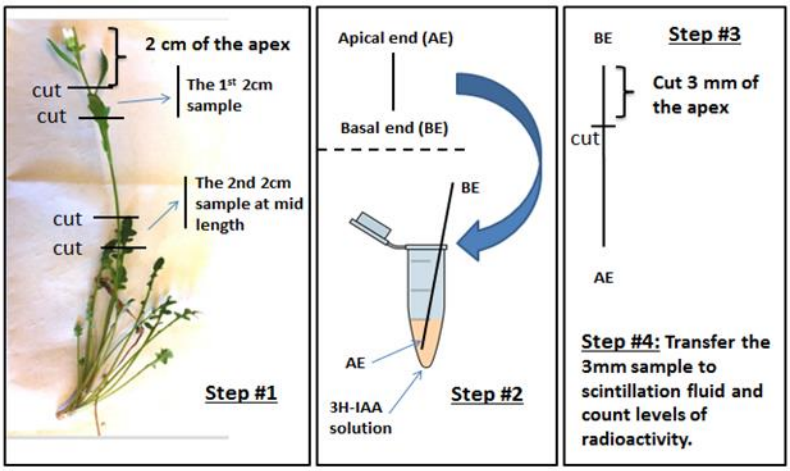
In February, we grew both *A. lyrata* populations in EGC growth chambers for auxin transport assay, simulating seasonal light hours (8/16:D/N for growing season, 12/12:D/N for flowering season) 2013. Temperature and photoperiod conditions were regulated accordingly to mimic seasonal winter and early spring growth patterns and flowering initiation. Plants were germinated in plastic inserts filled with Fafard germinating mix. After the seedlings grew about two true leaves, the lids of the plastic inserts were removed, and seedlings were watered three times a week, fertilized bi-weekly (1.25 mL L-1), and regularly monitored for growth. After two months, the plants were transplanted to cups filled with all-sport turface (baked clay grains), which helped retain moisture and nutrient. The germination mix-fritted clay combination was designed to imitate typical A. lyrata growth habitats in North Carolina, where plants usually thrive. The plants were then regularly monitored for signs of bolting, which generally

occur after 2-3 months of germination in the lab environment under 12 hours light/dark cycle. Plant locations in the growth chamber were rotated regularly. When plants bolted and further developed, we sampled the inflorescences for auxin transport assays. The earliest flowering date was June 20, 2013.

***Data collection***

I sampled three different individuals from each population. Six different inflorescence samples (3.5-4 cm inflorescence stem at mid-point) were taken from each individual to determine auxin transport. 4 ul of 3H-IAA was used at 25 Ci/mmol in 1 ml of agar (1.25% wt/vol) to yield a 100 nMol solution. A small amount, i.e., 5ul of the solution prepared at 100 nMol concentration, was then transferred to the bottom of the 1.5 mL centrifuge tube in 36 centrifuge tubes.

**Figure 2.2 Evaluating the 3H-IAA Auxin Flow in *A. lyrata* Inflorescence**




*Note:* A sample of inflorescence tissue is clipped, and the tissue is transferred to the radiolabeled 3H-IAA solution with an apical end in touch with the solution. After a few hours, the distal part of the tissue is sampled and transferred to a scintillation counter to measure radioactivity which provides a measure of auxin transport. Note that the samples aren't drawn to

scale, and tissue clipping shows at 2 cm, but in our Experiment, we only sampled one tissue section (4 cm inflorescence stem around mid-point along the stem).

To measure auxin transport, I clipped 4 cm of inflorescence tissue from around the mid-point along the sampled inflorescence stem. I then placed the sampled tissue in a $^{3}$H-IAA solution with the apical surface contacting the radiolabeled auxin and allowed the transport to continue for 6 hours. The apical 2 cm part of the sample touching the auxin droplet was clipped and discarded safely in a radioactive hazard collector. Next, the remaining 2 cm tissue from the basal end (where 3H-IAA is concentrated) was transferred to scintillation vials containing 3 ml of scintillation fluid for radioactivity quantification. This process is illustrated in Figure 2.2, and instructions for transport assay in inflorescence were derived from the protocol (Lewis & Muday, 2009; Okada et al., 1991); Box 1 and Procedure 9 B (Lewis & Muday, 2009) with some modifications, see, **Supplementary Materials S2.A - Protocol #1**. DPM (Disintegrations per minute) values were measured for quantitative investigation of auxin transport. The data gathered in this Experiment can be accessed at the following link - https://github.com/everestial/TestOfApicalDominanceInArabidopsisLyrata ).

***Statistical Analyses***

The data consists of 36 observations, with 18 tissue samples in each Mayodan and Spiterstulen population. Each population consisted of 3 different individuals, from which six inflorescence tissue were sampled for quantifying IAA transport. One observation is omitted from the analysis due to being a possible outlier as its DPM value is <4% of the next lowest value in the table and barely above background, indicating a likely setup error.

I used the R (v. 4.0.5) programming language for statistical analysis. First, I analyzed the effects of population on H3-IAA transport in the inflorescence shoots using classical pairwise-

test and non-parametric Wilcoxon Test. I also focused on the differences between individual groups. Then, for multiple groups analysis, we used ANOVA to check the hypotheses about the contrast within all groups and Tukey post hoc tests to investigate differences within each pair of groups. Finally, I performed an nested ANOVA analysis for additional insights.

## Experiment II: Effects of Auxin Inhibitor NPA on Life-History Traits of *A. lyrata* (Mayodan Population)

### *Overview*

I performed an auxin transport inhibition assay on individuals from the Mayodan population of *A. lyrata* using the pharmacological auxin transport inhibitor 1-N-naphthylphthalamic acid (NPA). Polar transport of auxin IAA in the inflorescences is mainly mediated by PIN1 protein (Okada et al., 1991), which are localized in vascular parenchyma (Blakeslee et al., 2007; Gälweiler et al., 1998; Steinmann et al., 1999), and the effects of NPA reduces the transport of IAA to the background levels (Lewis & Muday, 2009; Okada et al., 1991;  a M. Rashotte et al., 2001).

### *Growing condition and treatment assignment*

For auxin inhibition assay, I grew *A. lyrata* belonging to the Mayodan population in an EGC growth chamber using the same seed and growing materials and fertilization materials described in ***Experiment I***. After 30 days of germination, plants were potted in cups (on 8/4/2017) and grown under seasonal light hours (9hr, 20C / 15hr, 15C: D/N). NPA was prepared according to the protocol described in, **Supplementary Materials S2.A, Protocol #2**, at 10uM concentration. The experiment consisted of 60 plants from 9 different families representing an unequal number of individuals within each family (for more details, the Data-Sheet is available at the following https://github.com/everestial/TestOfApicalDominanceInArabidopsisLyrata). On

the 72nd day after the seeds were sown, individuals from each family were randomly assigned to one of the treatment groups, (1) Water application using a spray, (2) DMSO application using a spray, (3) NPA application using a spray, and (4) NPA application using drop on the tip of the apical meristem. The treatment was applied weekly. The settings were updated after 10 days of treatment to (12hr, 20C / 12hr, 15C: D/N) on 9/25/2017). The NPA concentration was increased to 20uM on 10/4/2017 because the plant biomass grew, and higher NPA doses may be required to block IAA transport effectively. In mid-December, the NPA treatment was stopped as NPA-treated plants showed extremely malformed development and were dying, suggesting that auxin transport inhibition was beginning to have toxic effects.

***Data Collection***

Plants were monitored every two days once the treatment began, and data were collected every week. The data collection involved taking top stock images of each plant weekly using an iPhone camera with the plant label and a ruler (inches) on the side (Figure 2.3). The collected photo data were analyzed at the end of the Experiment to gather data on the diameter (in mm), diameter changes (in mm), bolting (date of bolting), inflorescences (count on a particular date), and estimate of apical dominance using the degree of lateral vegetative branching in the rosette. The protocol for quantifying lateral vegetative branching is based on Remington (2015). However, our ratings ranged from 0 representing no visible lateral vegetive shoots or buds to 4, representing a rosette structure dominated completely by lateral shoots, with 4 being the highest. The rating method stays the same in these two studies except for the range, where Remington (2015) uses a range of 1-5.

For statistical analyses, only images and data from 3 specific months were chosen that represented equally distributed periods (Sept 18, 2017; Dec 18, 2017; and Mar 3, 2018). The data

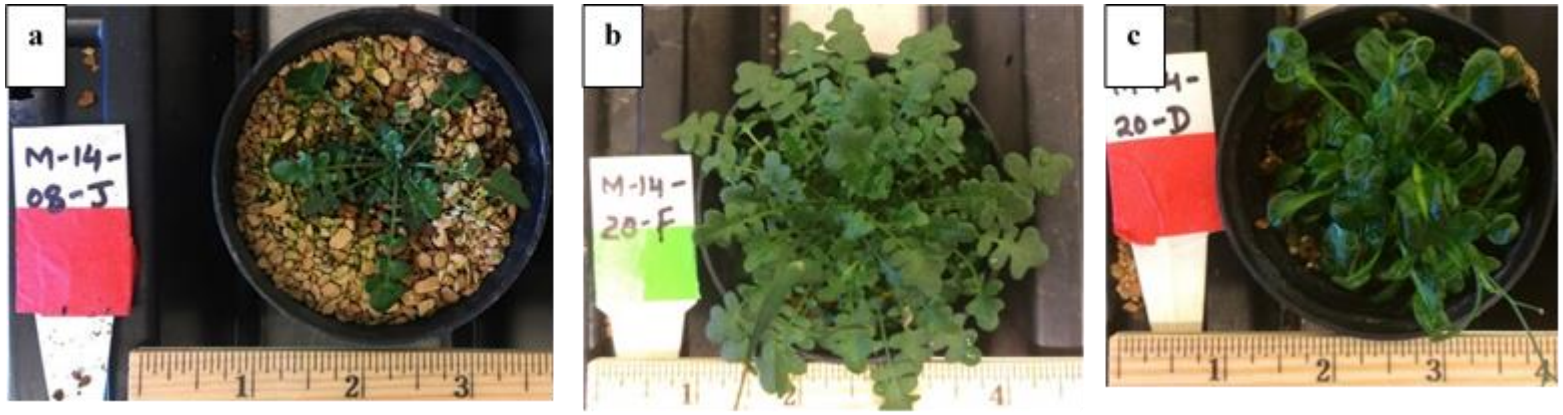
gathered in this Experiment can be accessed through the given link. (https://github.com/everestial/TestOfApicalDominanceInArabidopsisLyrata ).

***Statistical Analyses***

I used the R (v. 4.0.5) programming language for statistical analysis. The initial treatment consisted of 4 treatment groups – a) Spray with water, b) Spray with DMSO, c) Spray with NPA, d) Apply NPA using drip on the apical meristem. Due to the small sample size within each group, we narrowed the "treatments" into two groups as "treatment-level," (1) NPA treated group containing treatments using NPA spray and NPA applicator on the tip of the meristem, and (2) Control group containing treatments using DMSO and Water.

I tested for the effects of NPA treatment on diameter using mixed models using function *lmer* from *lme4 (v. 1.1-26)* package treating family, observable period and their interaction as a fixed effect, and plant-id as a random effect. In addition, I used a mixed model to test for treatment effects on the number of inflorescence shoots and lateral shoot rating.

**Figure 2.3 *A. lyrata* from Mayodan populations observed in different auxin inhibitor treatment groups**




*Note:* (a) A typical individual (Id: M-14-08-J) of *A. lyrata* from the Mayodan population at the beginning of the treatment. (b) A sample (Id: M-14-20-F) was treated with control treatment after 3 months. (c) A sample (Id: M-14-20-D) was treated with NPA. The pictures were taken for each individual from a top view using iPhone 5s camera weekly for each individual. The given picture of the plant (Id: M-14-08-J) was assigned to the treatment group "NPA application using spray". Since it is at the beginning of the treatment, there are no signs of effect by NPA inhibition of IAA. Another plant (Id: M-14-20-D) was also assigned to the treatment group "NPA application using spray" and showed results of NPA after a few months (during the month of December 2017).

## Results

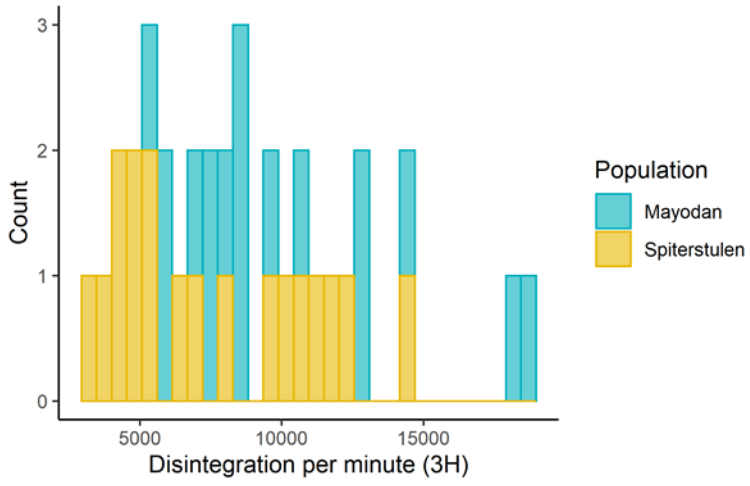
### Experiment I: Variation in Apical Dominance Between Mayodan and Spiterstulen

#### *Variation in Auxin Transport Between Mayodan and Spiterstulen*

The distribution of auxin transport for Mayodan plants does not pass the Shapiro normality test (see, **Supplementary Materials S2.B,** Table S2.B1, Figure S2.B1), while the distribution for the Spiterstulen group can roughly be treated as normal. At the same time, Levene's test (see, **Supplementary Materials S2.B,** Table S2.B2, *P*-value = 0.990) indicates that the two groups' values are homogenous.

Average disintegration per minute in Mayodan plants exceeds Spiterstulen by about 2400 DPM. However, the difference in the median is only 1663.15 DPM (see Table 2.1). As the histogram shows, a couple of samples show around 20000 DPM among the Mayodan population, leading to this group's high gap between mean and median.

**Figure 2.4 Histogram of DPM Values for Two Populations, From Experiment-I**

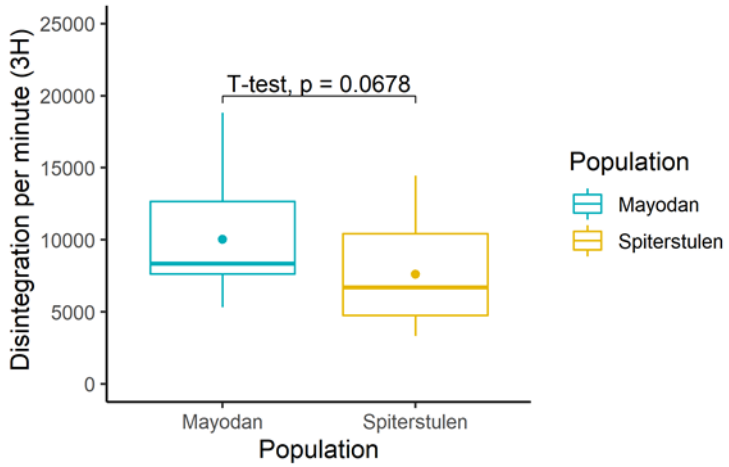



*Note:* The X-axis shows the DPM values, while the y-axis indicates the number of Mayodan and Spiterstulen with those DPM levels.

**Table 2.1 Summary Statistics of DPM for Populations My and Sp, From Experiment-I**

| Population | N | Mean | Std. Dev. | IQR | %25 Q | %50 Q | %75 Q |
|---|---|---|---|---|---|---|---|
| **Mayodan** | 17 | 10022.045 | 4062.547 | 5036.40 | 7634.050 | 8356.310 | 12670.45 |
| **Spiterstulen** | 18 | 7603.797 | 3458.120 | 5681.84 | 4756.292 | 6693.175 | 10438.13 |

**Figure 2.5 Box Plots Showing DPM Values for Populations My and Sp**




*Note:* Y-axis Represents DPM Values, From Experiment-I

A paired t-test (Figure 2.5) suggests marginal evidence for the significant difference in auxin transport between the two populations on the 10% confidence level ($P$=0.0678). However, since assumptions of normality are violated for these groups, a non-parametric Wilcoxon test was also conducted (see Table 2.2) with almost the same conclusion ($P = 0.0616$).

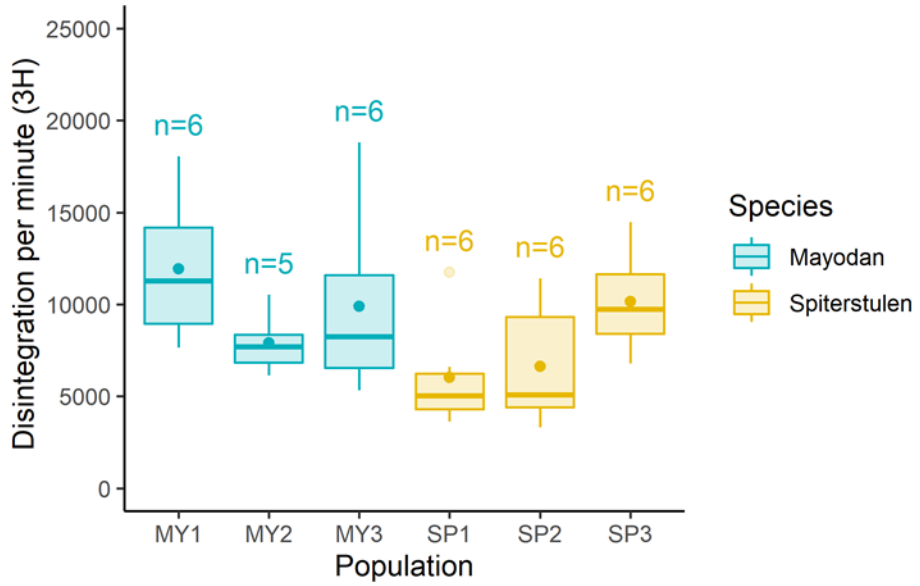
**Table 2.2 Wilcoxon-Test for Comparison of DPM Between My and Sp Individuals, From Experiment-I**

| Variable | $\mu_{Mayodan}$ | $\mu_{Spiterstulen}$ | $N_{Mayodan}$ | $N_{Spiterstulen}$ | W-Statistic | $P$ |
|---|---|---|---|---|---|---|
| **dpm3H** | 10022.045 | 7603.797 | 17 | 18 | 210 | 0.0616 |

**Table 2.3 Summary Statistics of DPM for Each Individual Plant in Both My and Sp Populations, From Experiment-I**

| Population | Indiv. | Mean | Std. dev. | IQR | %25 Q | %50 Q | %75 Q |
|---|---|---|---|---|---|---|---|
| **Mayodan** | MY1 | 11924.39 | 4000.068 | 5247.30 | 8941.14 | 11270.45 | 14188.448 |
| **Mayodan** | MY2 | 7902.842 | 1687.814 | 1532.45 | 6823.86 | 7690.50 | 8356.310 |
| **Mayodan** | MY3 | 9885.695 | 5067.361 | 5028.37 | 6558.53 | 8241.13 | 11586.910 |
| **Spiterstulen** | SP1 | 6023.430 | 2987.290 | 1918.62 | 4312.28 | 5019.43 | 6230.908 |
| **Spiterstulen** | SP2 | 6629.805 | 3459.220 | 4903.93 | 4404.55 | 5088.97 | 9308.487 |
| **Spiterstulen** | SP3 | 10158.15 | 2793.555 | 3254.92 | 8390.90 | 9743.83 | 11645.835 |

**Figure 2.6 Box Plots Showing DPM Values for Each Individual in Both Population (My and Sp) From Experiment-I**

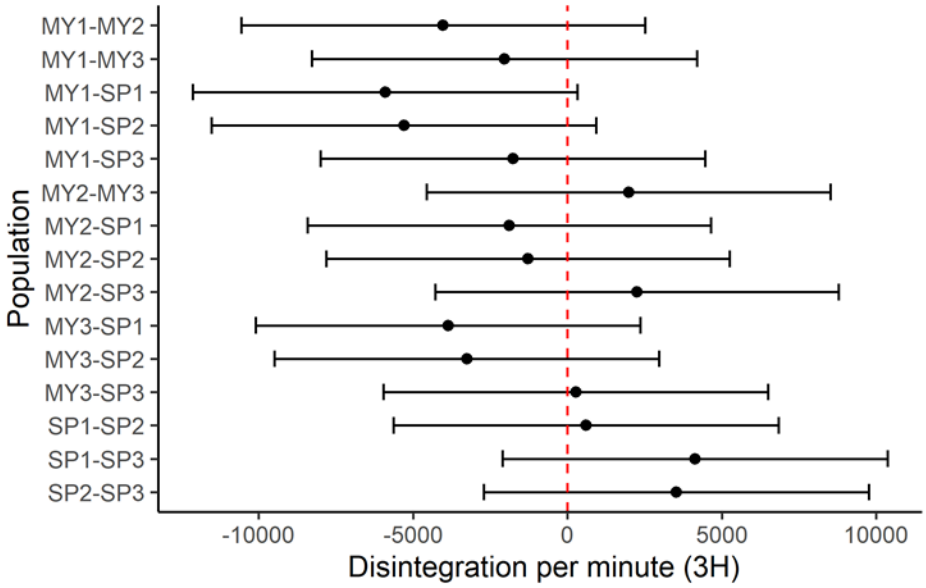



*Note*: X-axis Represents DPM Values, From Experiment-I. Box Plots Are Displayed for Each Individual for Both Populations.

***Tukey post hoc test***

Tukey post hoc test (see Figure 2.7) is conducted to determine whether any pair of groups show a significant difference in means. Since all the confidence intervals include 0, none of these differences can be considered significant. Only pairs MY1-SP1 and MY1-SP2 are close to not capturing the zero.

**Figure 2.7 Tukey Post Hoc Test of Significance Between Each Individual in Both My and Sp Populations for Observed DPM Values**




The classical pairwise t-test shows a similar picture (see **Supplementary Materials S2.D**, Table S2.D2). The difference in means for groups MY1 – SP2 and MY1 – SP3 is significant on the 5% confidence level. However, none of the differences are significant after the *P*-value is adjusted (via the FDR method).

***Nested Analysis***

To account for the individual-level effects on population differences, we conducted a nested ANOVA, which shows that the population-level effects are marginally significant. In contrast, the effects from individuals were not significant (see Table 2.4).

**Table 2.4 ANOVA Test Statistics for Population, Including Individual-Level Effects on Observed DPM Values, From Experiment-I**

| Effect | *F* | *P-value* |
|---|---|---|
| **Population** | 4.082 | 0.053 |
| **Population:Individual** | 2.078 | 0.110 |

**Experiment II: Effects of Auxin Inhibitor NPA on Life-History Traits of *A. lyrata* (Mayodan Population)**

Overall, data contained observations from 59 individual plants from September 2017, December 2017, and March 2018 (Table 2.5). After removing observations containing missing values (as the plants died during the research period), the dataset resulted in overall 152 observations.

**Table 2.5 The Number of Observations for Experiment-II for Each Period From 2017 to 2018**

| Month | N | NA | N - NA |
|---|---|---|---|
| **September 2017** | 59 | 1 | 58 |
| **December 2017** | 59 | 9 | 50 |
| **March 2018** | 59 | 15 | 44 |
| **Sum** | 177 | 25 | 152 |

***Effects of NPA Treatment on Diameter***

Plants in both treatment groups had no significant difference in mean diameter and showed similar data distribution before/when the treatment was started in September 2017 (see Figure 2.8). As time progressed, plants treated with NPA showed a smaller average diameter

growth with a relatively higher deviation by December 2017. However, between December 2017 and March 2018, after the NPA treatment was discontinued, the NPA-treated plants increased growth rate, resulting in a slightly higher average diameter (see Figure 2.8). Individuals in the control group (treated with water or DMSO) demonstrated a rapid increase in diameter during the first few months, followed by a smaller change during the second stage. Those treated with NPA experienced a steady acceleration of diameter growth (see Figure 2.9).

Despite having a different behavior in the growth process, the usage of NPA does not make a significant difference in overall diameter at the last stage of observation. These observations are supported by the Student's t-test (Table 2.6 and Table 2.7) and the mixed effect model estimates for diameter. In addition, the diameter coefficients at different periods are positive, meaning that the plants tend to grow over time. However, coefficients for treatment level are not significant except for their interaction with the December 2017 time period.

**Figure 2.8 Box Plots for Observations From Experiment-II for Observed Rosette Diameter by Treatment Level and Period**

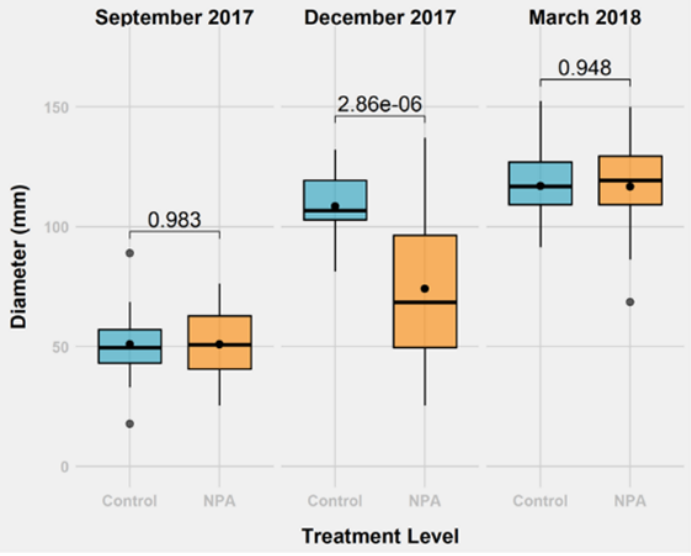


**Figure 2.9 Rosette Diameters Change by Treatment Level and Time From Experiment II**

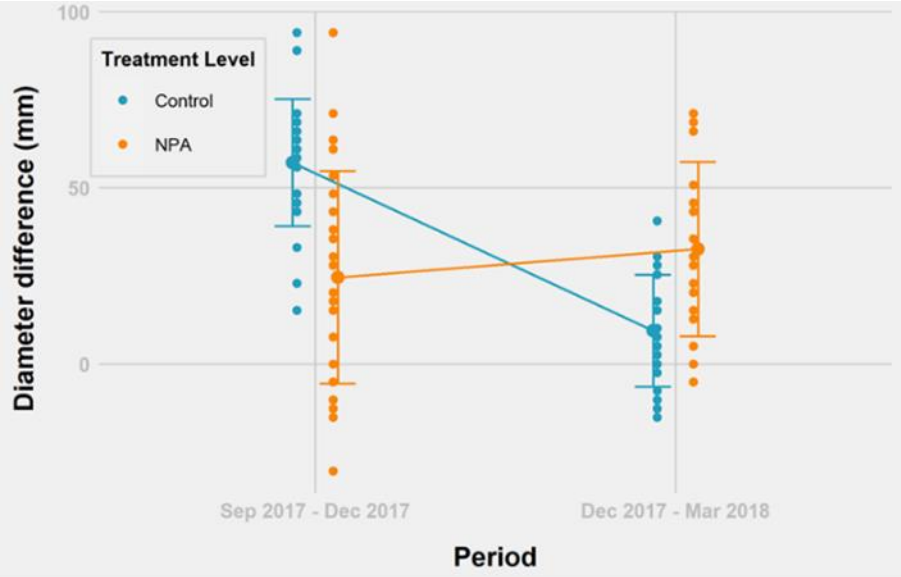



**Table 2.6 T-tests: For the Test of Difference in Diameter by Treatment Level for Each Time Period, From Experiment-II**

| Stage | Prediction | $\mu_{Control}$ | $\mu_{NPA}$ | $t$ | $P-value$ |
|---|---|---|---|---|---|
| September 2017 | $\mu_{Control} - \mu_{NPA} = 0$ | 50.89 | 50.97 | -0.022 | 0.98 |
| December 2017 | $\mu_{Control} - \mu_{NPA} > 0$ | 108.45 | 74.13 | 5.485 | < 0.01 |
| March 2018 | $\mu_{Control} - \mu_{NPA} = 0$ | 117.06 | 116.72 | 0.066 | 0.95 |

**Table 2.7 T-tests: For the Test of Changes in Diameter by Treatment Levels Between Two Consecutive Periods, From Experiment-II**

| Stage | Prediction | $\mu_{Control}$ | $\mu_{NPA}$ | $t$ | $P$ |
|---|---|---|---|---|---|
| September 2017 – December 2017 | $\mu_{Control} - \mu_{NPA} = 0$ | 57.20 | 24.62 | 4.643 | < 0.01 |
| December 2017 – March 2018 | $\mu_{Control} - \mu_{NPA} < 0$ | 9.47 | 32.66 | -3.636 | < 0.01 |

**Table 2.8 Mixed Models for Diameter by Treatment Level Across Different Periods, From Experiment-II**

| | **dia mm** | | |
|---|---|---|---|
| *Predictors* | *Estimates* | *CI* | *P* |
| (Intercept) | 50.89 | 44.07 – 57.71 | **<0.001** |
| Treatment Level [NPA] | -0.10 | -9.57 – 9.37 | 0.984 |
| Date [December 2017] | 57.35 | 48.19 – 66.51 | **<0.001** |
| Date [March 2018] | 65.91 | 56.75 – 75.07 | **<0.001** |
| Treatment Level [NPA] * Date [December 2017] | -33.95 | -46.54 – -21.37 | **<0.001** |
| Treatment Level [NPA] * Date [March 2018] | -1.54 | -14.63 – 11.55 | 0.817 |
| **Random Effects** | | | |
| $\sigma^2$ | 266.16 | | |
| $\tau_{00\ id}$ | 66.91 | | |
| ICC | 0.20 | | |
| $N_{\ id}$ | 59 | | |
| Observations | 152 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.713 / 0.770 | | |

Although NPA treatment does not seem to contribute to the final diameter of plants, it seems to affect the pattern of growth. The growth rate for plants treated with water outpaced the growth of the plants treated with NPA during the period in which NPA was applied (Sept-Dec). The mixed-effects model (Table 2.9) for diameter change suggests that plants with no NPA treatment on average grow 57.21 mm compared with 24.62 mm for plants with NPA. However, plants under no treatment show only about 9.5 mm growth, while using NPA causes a 32.66 mm increase after the NPA treatment was discontinued.

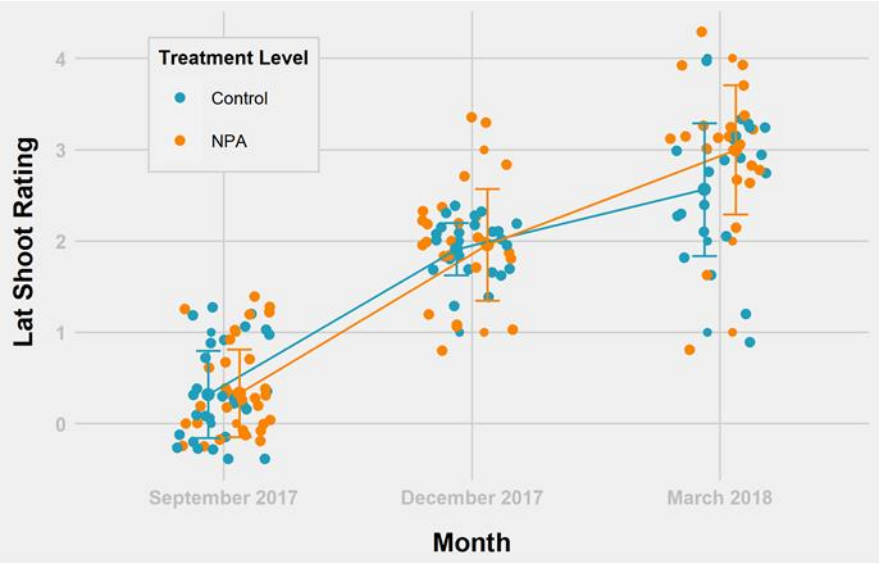
**Table 2.9 Mixed Models for Diameter by Treatment Level Between Two Consecutive Time Periods**

| | **dia diff** | | |
|---|---|---|---|
| *Predictors* | *Estimates* | *CI* | *P* |
| (Intercept) | 57.21 | 47.57 – 66.84 | **<0.001** |
| Treatment Level [NPA] | -32.59 | -45.81 – -19.36 | **<0.001** |
| Period [Dec 2017 – Mar 2018] | -47.74 | -61.52 – -33.96 | **<0.001** |
| Treatment Level [NPA] * Period [Dec 2017 – Mar 2018] | 55.78 | 36.44 – 75.11 | **<0.001** |
| **Random Effects** | | | |
| $\sigma^2$ | 540.34 | | |
| $\tau_{00\ id}$ | 0.00 | | |
| $N_{id}$ | 50 | | |
| Observations | 92 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.355 / NA | | |

***Effects of NPA Treatment on Lateral Shoot Rating***

The visual scoring of the extent of lateral shoot development initially increased at a similar rate for both treatment groups. However, the NPA treatment group generally had a higher number of plants with a higher lateral shoot rating for most of the growth period. For example, the means were similar in December, but the NPA group had a higher mean rating in March. This suggests a change in apical dominance during the early stage of the plants during NPA treatment.

**Figure 2.10 Lateral Shoot Rating by Growth Stage (or Time) From Experiment-II**




*Note:* The data points are "jittered" around the X-axis timepoints and the Y-axis integer lateral shoot ratings.

**Table 2.10 Observed Lateral Shoot Rating Values for Each Treatment Group at Different Times From Experiment II**

| Treatment | Lateral Shoots Rating Scale | | | | |
|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** |
| | ***September 2017*** | | | | |
| **Control** | **19** | **9** | **0** | **0** | **0** |
| **NPA** | **20** | **10** | **0** | **0** | **0** |
| | ***December 2017*** | | | | |
| **Control** | **0** | **2** | **21** | **0** | **0** |
| **NPA** | **0** | **5** | **16** | **4** | **0** |
| | ***March 2018*** | | | | |
| **Control** | **0** | **2** | **7** | **13** | **1** |
| **NPA** | **0** | **1** | **2** | **14** | **4** |

The difference between the two groups is not dramatic for earlier periods. However, the results from a Wilcoxon test provide favorable evidence that plants treated with NPA have higher lateral shoot ratings during the period of March (*P*-value for March = 0.0378, see Table 2.11).

**Table 2.11 Wilcoxon-Tests for Lateral Shoot Rating Between Control and NPA Treatment Group, From Experiment II. Tests Are Shown for Three Different Time Periods**

| Stage | Prediction | $\mu_{Control}$ | $\mu_{NPA}$ | $w$ | $P-value$ |
|---|---|---|---|---|---|
| **September 2017** | $\mu_{Control} - \mu_{NPA} = 0$ | **0.321** | **0.333** | **415** | **0.931** |
| **December 2017** | $\mu_{Control} - \mu_{NPA} = 0$ | **1.913** | **2.00** | **278** | **0.8** |
| **March 2018** | $\mu_{Control} - \mu_{NPA} = 0$ | **2.565** | **3.00** | **164** | **0.0378** |

***Effects of NPA Treatment on Inflorescence Number***

In September and December, the average number of reproductive shoots in the NPA-treated and Control groups is very similar, which is a period when the meristems have not yet committed to reproductive shoots. Wilcoxon test shows no significant difference in the number of inflorescences between the control and NPA treated group for each period (see Table 2.13). The confidence intervals are large, which does not let us assume the real difference in means to exist. However, specifically during March, the mean number of inflorescences was fewer in the NPA-treated group, though the difference is not significant.

Nevertheless, the results show some negative correlation between lateral shoot rating vs. the number of inflorescences, suggesting NPA might have played a role in altering resource allocation by changing apical dominance.

**Figure 2.11 Number of Reproductive Shoots by Months for Each Treatment Level**

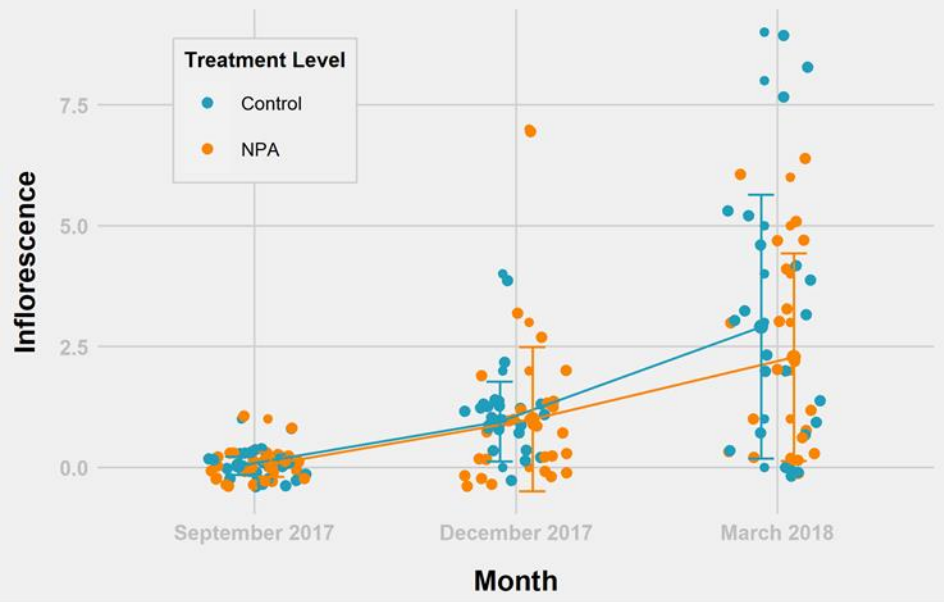



**Table 2.12 The Number of Plants by the Number of Inflorescences Across Treatment Groups and Periods**

| Treatment | Number of inflorescences | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | *September 2017* | | | | | | | | | |
| Control | 27 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| NPA | 28 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *December 2017* | | | | | | | | | |
| Control | 5 | 16 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| NPA | 12 | 10 | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 0 |
| | *March 2018* | | | | | | | | | |
| Control | 5 | 4 | 3 | 3 | 2 | 3 | 0 | 0 | 2 | 1 |
| NPA | 6 | 4 | 2 | 3 | 1 | 3 | 2 | 0 | 0 | 0 |

**Table 2.13 Wilcoxon-Tests for the Number of Inflorescences Between Control and NPA Treatment Groups. Tests Are Shown for Three Different Periods**

| Stage | Prediction | $\mu_{Control}$ | $\mu_{NPA}$ | $w$ | $P-value$ |
|---|---|---|---|---|---|
| **September 2017** | $\mu_{Control} - \mu_{NPA} = 0$ | **0.0357** | **0.0667** | **407** | **0.612** |
| **December 2017** | $\mu_{Control} - \mu_{NPA} > 0$ | **0.9565** | **1.00** | **351** | **0.39** |
| **March 2018** | $\mu_{Control} - \mu_{NPA} = 0$ | **2.913** | **2.286** | **267** | **0.551** |

The mixed-effects model estimates show that the treatment level does not significantly affect the number of inflorescences (reproductive shoots).

## Discussion

### Experiment I

We predicted that analysis of $^{3}$H-IAA transport would show a higher level of auxin transport in Mayodan individuals. We found weak evidence (*F-statistic* = 4.082, *P*=0.0527) of variation in the strength of auxin transport between populations, with Mayodan individuals showing higher auxin transport. However, the results might have been confounded by variation in the diameter of inflorescences as Spiterstulen individuals mostly have a thicker diameter which is visually apparent. However, we did not measure inflorescence diameters, preventing using diameter measurements as covariates in this analysis. We would expect the amount of transport to increase with diameter, so the actual differences in the transport rate might have been more significant if those were taken into account.

Future studies measuring transport differences with larger biological and technical replicates and the measurements of the diameter of the inflorescences (as a covariate) can provide more clarity on this issue. In addition, further analyses of transport differences can be done at population levels, using auxin pulse-chase assay or further investigating variations in auxin transport rate in individuals with contrasting homozygous genotypes in the QTL regions.

This test would help us test the role of specific QTLs in auxin transport and potentially on life-history tradeoffs.

**Experiment II**

It is known that the response of the different tissues to IAA transport inhibitor is not uniform (Negi et al., 2008; Lewis and Muday, 2009). We found weak evidence indicating that auxin transport inhibition affects life-history traits (lateral shoot rating and the number of inflorescences). While the evidence is not strong, it points to the direction as predicted and observed in Experiment I; altering apical dominance using NPA inhibitor increased the emergence of lateral vegetative shoots and reduced the number of inflorescences. The plant treated with NPA also had high mortality, thus reducing the test's statistical power. This mortality could be due to altered auxin dynamics causing direct toxic effects of NPA or the ecological consequences of the tradeoff or NPA affecting some developmental pathways. In the NPA treatment group, we observed a delay in the apparent effects of transport inhibition on lateral shoot rating and the number of inflorescences, with effects showing up three months after NPA treatment was discontinued. This is consistent with the idea that variation in traits such as apical dominance in early development can cascade through later developmental stages, changing the entire trajectory of life history. If the NPA treatment could have been continued without any adverse effects based on the dose of NPA, the measured differences in life history traits between the two groups might have become greater.

Also, the last set of measurements was taken early in the reproductive period. If we had continued the measurements during the the reproductive period, we would expect that the number of inflorescences might have increased over time for the control group and decreased for the NPA treated individuals. And, for the lateral vegetative shoots, the application of NPA could

have reduced apical dominance and allowed more lateral shoots to grow. These lateral shoots could have turned into lateral vegetative shoots at the later stage of plant development, even after the effects of NPA wore off. The results conclude that similar experiments are required with larger sample size and more controls to attest that variation in apical dominance in early life stages could be one of the fundamental mechanisms driving life-history variation and tradeoffs.

I also emphasize that even though none of the differences in either the auxin transport or auxin inhibition assays were statistically significant other than diameter growth during the month of December, and lateral shoot rating during the month of March, all the trends were in the predicted direction. This provides evidence, although tentative, that genetic variation affecting auxin transport could underlie adaptive variation in life history in *A. lyrata*.

Future studies involving optimized doses of NPA treatment can provide a more robust estimate of this question. Additionally, genetic insertion of My alleles (for auxin transport-related genes) on the Spiterstulen genotype background and vice-versa can test whether auxin transport genes result in life-history variation.

## Supplementary Materials: Chapter II

### Abbreviations

| | |
|---|---|
| 2,4-D | 2,4-dichlorophenoxyacetic acid |
| $^{3}$H-IAA | radiolabeled auxin |
| ABA | abscisic acid |
| ABCB transporters | ATP-binding cassette transporters |
| ANOVA | Analysis of variance |
| approx. | Approximately |
| ATIs | Auxin transport inhibitors |
| *ba1* | barren stalk1 |
| *bif2* | barren inflorescence2 |
| BRs | brassinosteroids |
| CFM | methyl-2-chIoro-9-hydroxyfluorene-9-carboxylate |
| CK | cytokinins |
| cm | centimeter |
| CPD | 2-carboxyphenyl-3-phenylpro- pane-l,2-dione |
| d | days |
| d.p.m. | disintegrations per minute |
| dd$H_2O$ | double-distilled water |
| DMSO | Dimethyl sulfoxide |
| DPM | Disintegrations per minute |
| EGC | Environmental Growth Chamber |
| ET | ethylene |
| FDR | false discovery rate |
| fmol | femtomole |
| GA | gibberellins |
| IAA | indole-3-acetic acid |
| IBA | indole-3-butyric acid |
| IQR | interquartile range |
| JA | jasmonic acid |
| MES | 2-(n-morpholino)-ethanesulfonic acid |
| mg | milligram |
| mg/mL | milligram per milliliter |
| ml | milliliter |
| mm | millimeter |
| mmol | milimole |
| MY | Mayodan |
| MY1, MY2, MY3 | Mayodan individual 1, 2 and 3 |
| NAA | 1-naphthaleneacetic acid |
| NBP proteins | nonamer-binding protein |

| | |
|---|---|
| nMol | number of moles |
| NPA | 1-naphthylphthalamic acid |
| PBA | 2-(1-pyrenoyl) benzoic acid |
| pH | potential of hydrogen |
| PIN | pin-like phenotype |
| PIN-HL | PIN-Hydrophilic Loop |
| QQ-plot | quantile-quantile |
| *QTL* | quantitative trait locus |
| SA | salicylic acid |
| SLs | strigolactones |
| SP | Spiterstulen |
| SP1, SP2, SP3 | Spiterstulen individual 1, 2 and 3 |
| TIBA | 2,3,5-triiodobenzoic acid |
| *TIR3* | Transport Inhibitor Response |
| uM | micrometer |
| wt/vol | Mass by Volume |

**Supplementary Materials S2.A: Protocols for Auxin Transport and Auxin Transport Inhibition Assay**

**Protocol # 1. for auxin transport assay**

This method was used to quantify the basipetal transport of auxin in inflorescence tissue samples. The original protocol (Lewis and Muday, 2009) was modified to fit our experimental needs. The radioactive substance's medium can be prepared either as agar or liquid. However, necessary measures should be taken if the solution is aqueous as it promotes wicking and diffusion and affects the measurements.

- The grown plants with healthy inflorescences were prepared as described in the experiment 1A.
- The 100 nM 3H-IAA was prepared by adding approximately 2–4 ml of 3H-IAA at 20–50 Ci mmol1 to 1 ml of 0.05% MES, pH 5.5–5.7.

- About 20 ml of radioactive IAA solution was dispensed into the 0.5-ml microcentrifuge tubes.
- The inflorescence stem was cut at 2 cm and 4.5 cm from the apex using microscissors.
- The apical part of the samples (cut segment of the inflorescences) were dipped into the IAA solution for 6 hours.

A 2 cm section of the sample was again excised using micro scissors from the non-submerged end of the segment (basipetal end of the tissue) and transferred to a scintillation vial to measure 3H d.p.m.

**Protocol # 2. for Auxin inhibition assay**

The NPA stock solution was purchased from https://www.chemservice.com/n-1-naphthylphthalamic-acid-solution-s-12507t1-1ml.html , with concentration of 100 ug/ml in T-butylmethyl Ether solution (**Part #:**S-12507T1-1ML, **CAS:** 132-66-1).

Step A: Prepare a concentrated and dilute NPA stock.

1) Dissolve 100 mg of NPA in 5 mL DMSO to get a 2 % w/v (i.e. 20 mg/mL) solution of NPA.

2) Prepare a 10x dilution of the NPA stock solution by combining the appropriate quantity of ddH2O; for example, if you want to create a 20 mL dilute NPA stock, combine 2 mL stock solution with 18 mL ddH20. It gives a 2 mg/mL dilution.

3) Wrap the container in aluminum foil and place it in the freezer to keep the concentrated (2%) stock safe.

Step B: Prepare the NPA treatment and control solutions.

1) Make two 500 mL dd-H20 media bottles.

2) Add 728 ul of 10x dilution NPA stock solution to one of the ddH2O stocks using sterile pipette tips.

3) Add 728 ul of DMSO to the other flask of cooled medium for preparing the control treatments. **NOTE:** Make sure to keep the NPA control and treatment solutions in the dark and cool (4 °C) environment before and after the use.

Step C: Bi-weekly application of treatment and control solutions

**NOTE:** Step-C can vary according to the experimental requirements.
1) Divide the plants into four categories based on their treatment assignment: a) Spray with water , b) Spray with DMSO, c) Spray with NPA, d) Apply NPA using drip on the apical meristem.

2) Apply respective treatments weekly. **NOTE:** Remove the plants from the GC (growth chamber) while applying treatments to avoid the treatment residues from spreading to other groups of plants.

3) Apply treatments bi-weekly; be sure to follow all necessary safety precautions.

- Spray the crown of the plants in group-A with water.
- Spray the crowns of the plants in group-B with DMSO solution in a way that it sufficiently wets the crown.
- Spray the crowns of the plants in group-C with NPA solution in a way that it sufficiently wets the crown.
- For the plants in group-D apply the NPA solution (0.5-1.0 mL) on the apical meristem (tip of the plant's main shoot apex) using an applicator.

**Protocol # 3. Lateral shoot rating system.**

Unlike the original ratings that use a scale of 1-5, this study uses a 0-4, with no apparent distinction in ratings between the two studies, except for their rating values (derived from source (Remington et al., 2015)).

Lateral vegetative shoot ratings:

> 0 – All visible rosette leaves are primary leaves (on main stem, not emerging from lateral shoots). All newer leaves (not fully elongated yet) are attached above the older, fully-elongated leaves. Primary shoot apex is obvious and dominant, and the leaves extend horizontally from it.
>
> 1 – Some leaves emerging from lateral shoots are visible but are much smaller than fully-elongated primary leaves. Some newer leaves are obviously attached below larger leaves on main stem. Primary shoot apex is obvious and still clearly dominant over lateral vegetative shoots.
>
> 2 – Leaves from lateral shoots are apparent, and some may be difficult to distinguish from primary leaves. The primary shoot apex is still apparent but is losing its dominance, and some lateral shoots are nearly as vigorous as the main shoot. The rosette is beginning to acquire a bushy form, with many leaves in a vertical orientation.
>
> 3 – Many lateral shoot leaves are nearly as large as the primary leaves. The primary and lateral shoot apices are becoming difficult to distinguish, though larger primary leaves produced earlier may still be apparent on the lower part of the plant. The rosette has a bushy form, with leaves extending at all angles.
>
> 4 – The primary and lateral shoots can no longer be distinguished. All fully-elongated leaves are relatively compact. The rosette has a dense cushiony appearance, with leaves extending at all angles.

**Supplementary Materials S2.B: Checking for Normality and Endogeneity (Experiment #1)**

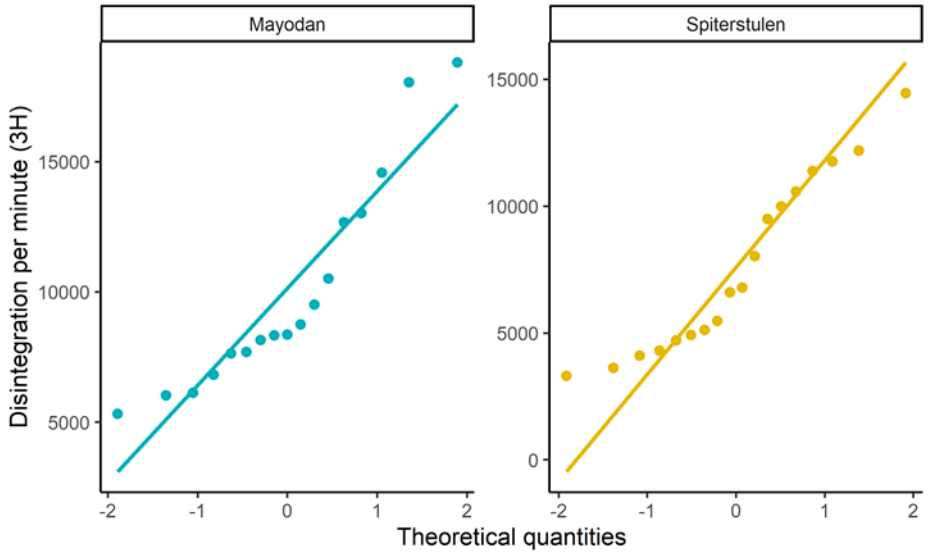
**Table S2.B1 Shapiro Test for Observed DPM Values in Individuals of Each My and Sp Population**

| **Population** | **Variable** | **Statistic** | ***P*-value** |
|---|---|---|---|
| **Mayodan** | dpm3H | 0.875 | 0.027 |
| **Spiterstulen** | dpm3H | 0.917 | 0.116 |

**Table S2.B2 Levene's Test Statistic for normality of DPM Values**

| **Statistic** | ***P*-value** |
|---|---|
| **0.0001666** | 0.990 |

**Figure S2.B1: QQ-Plot – Checking for Normality and Endogeneity for Population-Level DPM Values**



**Supplementary Materials S2.C: Checking for Normality and Endogeneity for Each Individual. (Experiment #1)**

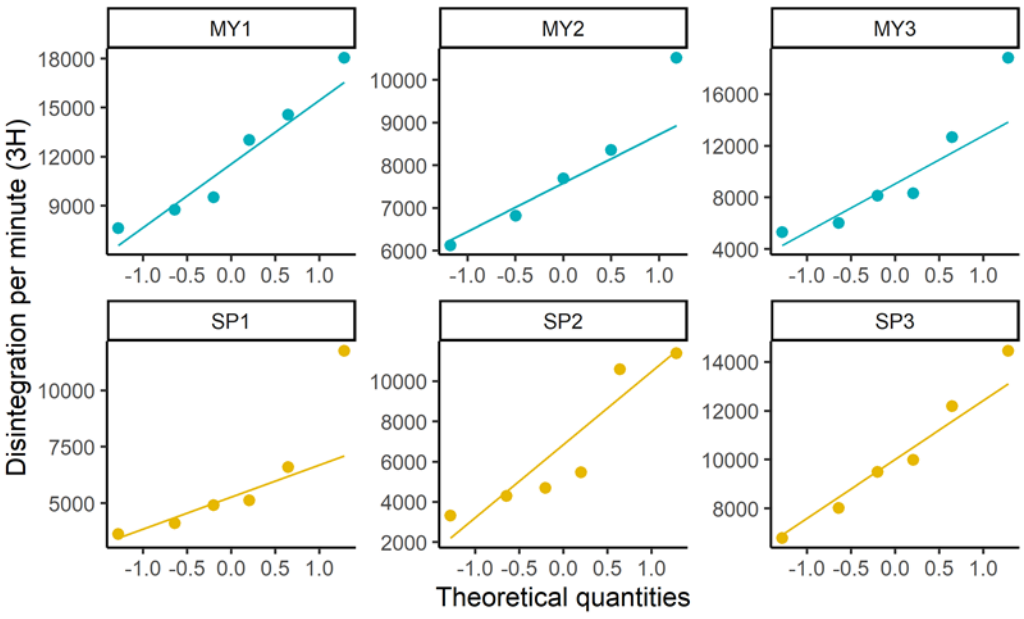
**Table S2.C1 Shapiro Test for observed DPM values in individuals of population**

| Population | Variable | Statistic | *P*-value |
|---|---|---|---|
| **MY1** | dpm3H | 0.933 | 0.604 |
| **MY2** | dpm3H | 0.948 | 0.722 |
| **MY3** | dpm3H | 0.866 | 0.212 |
| **SP1** | dpm3H | 0.791 | 0.049 |
| **SP2** | dpm3H | 0.826 | 0.099 |
| **SP3** | dpm3H | 0.968 | 0.880 |

**Table S2.C2 Levene's test statistic for normality of DPM values.**

| Statistic | *P*-value |
|---|---|
| **0.6746619** | 0.646 |

**Figure S2.C1: QQ-plot for DPM (3H) (y-axis) of each population**



**Supplementary Materials S2.D: Pairwise Comparison Between Individuals of Both the Populations**

**Table S2.D1 Tukey Test Results between individuals of both the populations**

| Term | Group1 | Group2 | Estimate | CI lower | CI upper | *P*-value (adj.) |
|---|---|---|---|---|---|---|
| **individual** | MY1 | MY2 | -4021.55 | -10554.78 | 2511.6750 | 0.436 |
| **individual** | MY1 | MY3 | -2038.7 | -8267.89 | 4190.4884 | 0.915 |
| **individual** | MY1 | SP1 | -5900.97 | -12130.16 | 328.2234 | **0.0713** |
| **individual** | MY1 | SP2 | -5294.59 | -11523.78 | 934.5984 | **0.132** |
| **individual** | MY1 | SP3 | -1766.24 | -7995.43 | 4462.9484 | 0.952 |
| **individual** | MY2 | MY3 | 1982.85 | -4550.38 | 8516.0827 | 0.937 |
| **individual** | MY2 | SP1 | -1879.41 | -8412.64 | 4653.8177 | 0.949 |
| **individual** | MY2 | SP2 | -1273.04 | -7806.27 | 5260.1927 | 0.991 |
| **individual** | MY2 | SP3 | 2255.31 | -4277.92 | 8788.5427 | 0.896 |
| **individual** | MY3 | SP1 | -3862.27 | -10091.46 | 2366.9251 | 0.428 |
| **individual** | MY3 | SP2 | -3255.89 | -9485.08 | 2973.3001 | 0.609 |
| **individual** | MY3 | SP3 | 272.46 | -5956.73 | 6501.6501 | 1 |
| **individual** | SP1 | SP2 | 606.38 | -5622.82 | 6835.5651 | 1 |
| **individual** | SP1 | SP3 | 4134.73 | -2094.47 | 10363.9151 | 0.354 |
| **individual** | SP2 | SP3 | 3528.35 | -2700.84 | 9757.5401 | 0.526 |

**Table S2.D2 Pairwise Test results between individuals of both the populations**

| Variable | Group1 | Group2 | n1 | n2 | *P*-value | Significance | *P*-value (adj.) | Significance (adj.) |
|---|---|---|---|---|---|---|---|---|
| **dpm3H** | MY1 | MY2 | 6 | 5 | 0.0707 | ns | 0.212 | ns |
| **dpm3H** | MY1 | MY3 | 6 | 6 | 0.327 | ns | 0.493 | ns |
| **dpm3H** | MY2 | MY3 | 5 | 6 | 0.362 | ns | 0.493 | ns |
| **dpm3H** | MY1 | SP1 | 6 | 6 | **0.00726** | ** | 0.109 | ns |
| **dpm3H** | MY2 | SP1 | 5 | 6 | 0.388 | ns | 0.493 | ns |
| **dpm3H** | MY3 | SP1 | 6 | 6 | 0.0688 | ns | 0.212 | ns |
| **dpm3H** | MY1 | SP2 | 6 | 6 | **0.0148** | * | 0.111 | ns |
| **dpm3H** | MY2 | SP2 | 5 | 6 | 0.557 | ns | 0.643 | ns |
| **dpm3H** | MY3 | SP2 | 6 | 6 | 0.122 | ns | 0.261 | ns |
| **dpm3H** | SP1 | SP2 | 6 | 6 | 0.769 | ns | 0.824 | ns |
| **dpm3H** | MY1 | SP3 | 6 | 6 | 0.394 | ns | 0.493 | ns |
| **dpm3H** | MY2 | SP3 | 5 | 6 | 0.301 | ns | 0.493 | ns |
| **dpm3H** | MY3 | SP3 | 6 | 6 | 0.895 | ns | 0.895 | ns |
| **dpm3H** | SP1 | SP3 | 6 | 6 | 0.0523 | ns | 0.212 | ns |
| **dpm3H** | SP2 | SP3 | 6 | 6 | 0.0949 | ns | 0.237 | ns |

**Supplementary Materials S2.E: Data, Codes, and R-scripts (Experiment #1)**

All the source code and data for these analyses are hosted on this GitHub repo.

https://github.com/everestial/TestOfApicalDominanceInArabidopsisLyrata

**NOTE: Supplementary Materials for Experiment 2 begins.**

**Supplementary Materials S2.F Checking for Normality and Endogeneity (Experiment #2)**

**Table S2.F1 Shapiro Test for normality of Diameter (mm) in each Treatment Level**

| Treatment Level | Date | Variable | Statistic | *P*-value | N |
|---|---|---|---|---|---|
| **Control** | September 2017 | Diameter (mm) | 0.958 | 0.317 | 28 |
| **Control** | December 2017 | Diameter (mm) | 0.954 | 0.352 | 23 |
| **Control** | March 2018 | Diameter (mm) | 0.970 | 0.689 | 23 |
| **NPA** | September 2017 | Diameter (mm) | 0.968 | 0.505 | 30 |
| **NPA** | December 2017 | Diameter (mm) | 0.976 | 0.770 | 27 |
| **NPA** | March 2018 | Diameter (mm) | 0.960 | 0.515 | 21 |

**Table S2.F2 Shapiro Test for normality of Diameter difference (mm) in each Treatment Level**

| Treatment Level | Date | Variable | Statistic | *P*-value | N |
|---|---|---|---|---|---|
| **Control** | Sep 2017 – Dec 2017 | Diameter difference (mm) | 0.929 | 0.105 | 23 |
| **Control** | Dec 2017 – Mar 2018 | Diameter difference (mm) | 0.957 | 0.425 | 23 |
| **NPA** | Sep 2017 – Dec 2017 | Diameter difference (mm) | 0.984 | 0.944 | 27 |
| **NPA** | Dec 2017 – Mar 2018 | Diameter difference (mm) | 0.929 | 0.129 | 21 |

**Table S2.F3 Shapiro Test for normality of Lateral Shoots ratings in each Treatment Level**

| Treatment Level | Date | Variable | Statistic | *P*-value | N |
|---|---|---|---|---|---|
| **Control** | September 2017 | Lateral Shoots | 0.590 | 0.000 | 28 |
| **Control** | December 2017 | Lateral Shoots | 0.324 | 0.000 | 23 |
| **Control** | March 2018 | Lateral Shoots | 0.808 | 0.000 | 23 |
| **NPA** | September 2017 | Lateral Shoots | 0.59 | 0.000 | 30 |
| **NPA** | December 2017 | Lateral Shoots | 0.770 | 0.000 | 27 |
| **NPA** | March 2018 | Lateral Shoots | 0.762 | 0.000 | 21 |

**Table S2.F4 Shapiro Test for normality Inflorescence number in each Treatment Level**

| Treatment Level | Date | Variable | Statistic | *P*-value | N |
|---|---|---|---|---|---|
| **Control** | September 2017 | Inflorescence | 0.188 | 0.000 | 28 |
| **Control** | December 2017 | Inflorescence | 0.639 | 0.000 | 23 |
| **Control** | March 2018 | Inflorescence | 0.885 | 0.013 | 23 |
| **NPA** | September 2017 | Inflorescence | 0.275 | 0.000 | 30 |
| **NPA** | December 2017 | Inflorescence | 0.655 | 0.000 | 27 |
| **NPA** | March 2018 | Inflorescence | 0.872 | 0.010 | 21 |

**Figure S2.D1 QQ-plot for rosette diameter (mm) (y-axis) of each Treatment Level (Control and NPA)**

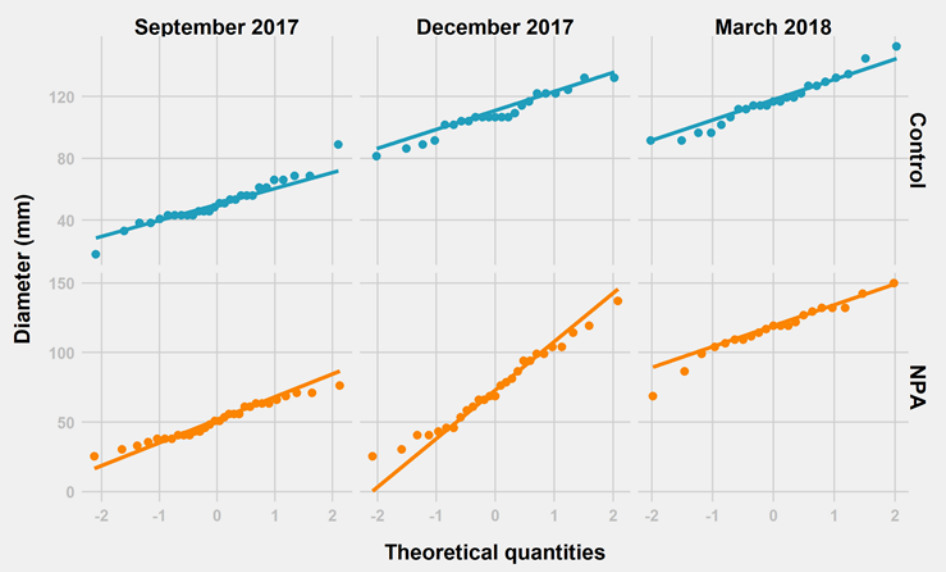



**Figure S2.D2 QQ-plot for observed diameter difference (mm) in each Treatment Level**

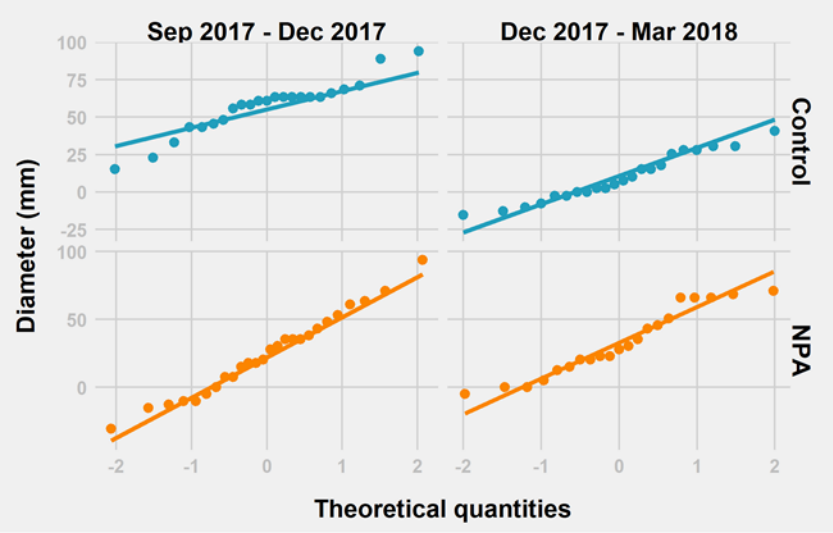


**Figure S2.D3 QQ-plot for observed Lateral Shoots rating in each Treatment Level**

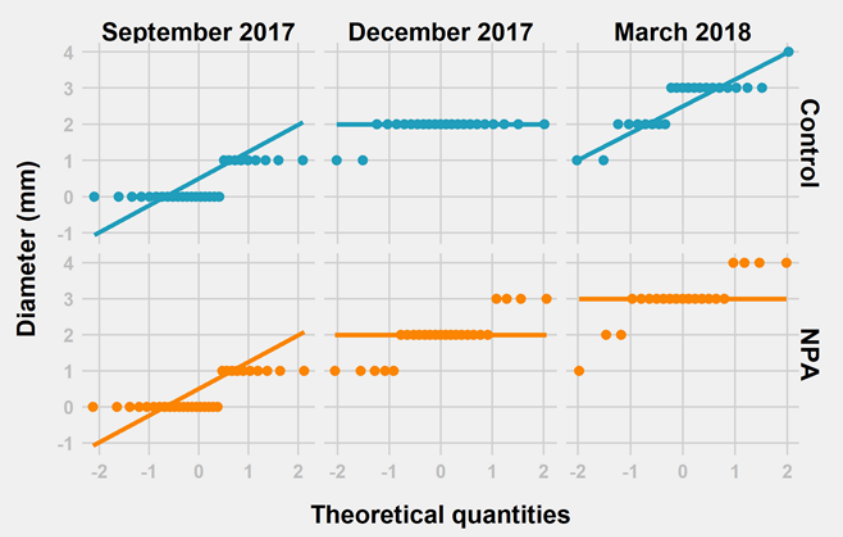
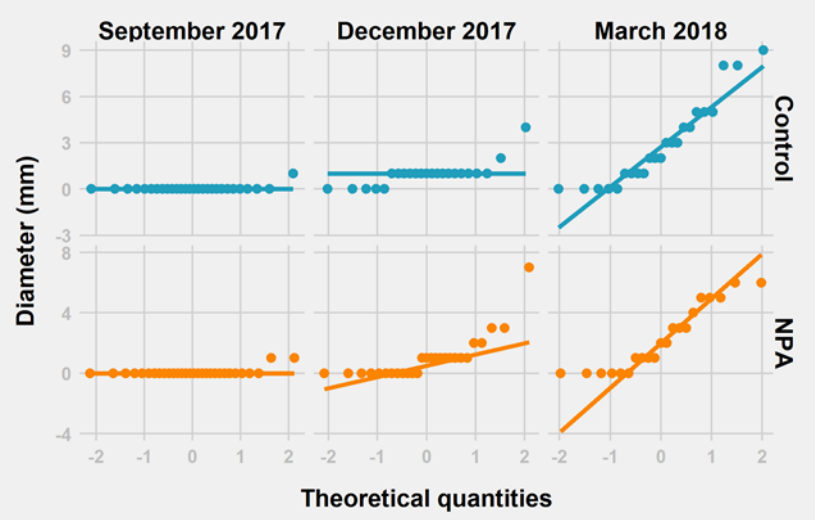



**Figure S2.D4 QQ-plot for the observed number of inflorescence in each Treatment Level**




**Supplementary Materials S2.G: Data, Codes, and R-scripts (Experiment #2)**

All the source code and data for these analyses are hosted on this GitHub repo.

https://github.com/everestial/TestOfApicalDominanceInArabidopsisLyrata

# CHAPTER III: PHASING OF INDIVIDUAL GENOMES USING READ-BACKED-PHASED HAPLOTYPE AND MARKOV CHAIN MAXIMUM LIKELIHOOD ESTIMATION


## Abstract

Haplotype phasing is the second most crucial issue in genomics after sequence alignment. A polyploid genome generates not appropriately arranged variants in a haplotype that forms a whole chromosome homolog. If the organism is diploid or polyploid, the called variants need to be scrutinized, so they are assigned to the proper homolog they originated from. Unfortunately, it isn't easy to separate a pair of chromosomes with current technology, and we often get the two haplotypes mixed.

Phased haplotypes provide better estimates while doing analyses related to genetics, phylogenetics, and dissecting genotype-trait association. If the organism is diploid or polyploid, the called variants need to be scrutinized and assigned to the proper homolog they originated from. While there are different approaches to haplotype phasing, a statistical model is the most applied methods to phasing haplotypes.

Most of the sequencing tools are challenging to customize during haplotype phasing. They also require a large number of reference haplotype samples. Therefore, we built a group of haplotype phasing tools/algorithms: Phase-Extender, Phase-Stitcher, and ShortVariantPhaser. These tools aim to fulfill the following major requirements - be able to customize parameters during haplotype phasing and run haplotype phasing using genotype containing unphased and read-backed-phased VCF data. Phasing using read-backed-phased haplotype blocks reduces the number of required reference panels. Additionally, this tool helps phase a cohort of samples where the partially phased genotypes information can help the samples phase each other.

We expect this tool to be beneficial for those organisms that do not have a substantial number of reference panels. While comparing results with other tools, Phase-Extender performs on par with ShapeIT using only 10 read-backed-phased samples. In comparison, ShapeIT uses 1010 samples to achieve the same level of phasing accuracy.

# Introduction

## General Introduction to Modern Genomics

The revolutionary advances in genome sequencing technology have, in turn, required advances in bioinformatics tools, including tools for sequence alignments, variant calling, and individual to population-level analyses of the variants. These tools help identify the genetic basis of traits and diseases and also help us apply the developed methods to the genomics and transcriptomics analyses in other biological models.

## Introduction to Haplotype Phasing

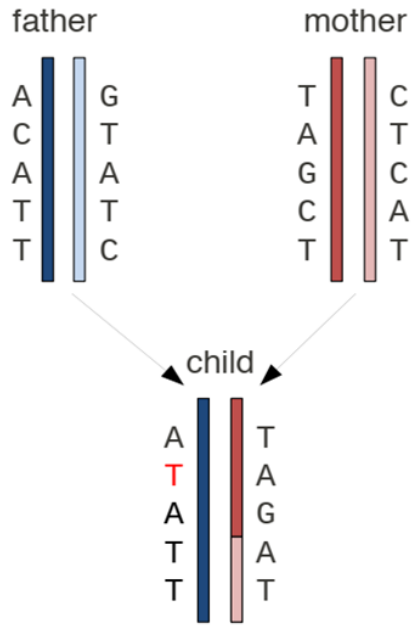
"Haplotype phasing" is a significant discipline in bioinformatics that deals with putting the called variants into their original order on homologous chromosomes in a polyploid genome, with the earliest known method developed by Clark (1990). The technique for phasing haplotypes has grown over time and has been reformulated to incorporate new information gained from the increased sample size, increased length of the sequence reads, and partially phased variants, read-backed-phased (RBP) haplotypes, haplotypes from long reads. It also has seen improvements in statistical methods.

A haplotype is the linearly arranged genotypes (by position) in a phased state, linked along a chromosome. A phased haplotype can also be called a single "haploid genotype" and, in such instances, can contain multiple SNP genotypes (Consortium et al., 2005). Technically, a haplotype represents strongly correlated adjacent genotypes in linkage disequilibrium (Daly et

al., 2001; Reich et al., 2001). Theoretically, a whole chromosome can be considered a haplotype. Therefore, a haplotype corresponds to a segment in the genome that provides necessary information for testing questions related to ancestry, demography, and association between genetic variation and phenotype or disease.

The set of genetic variants received exclusively from the father are considered one haplotype, and the variations solely received from the mother are considered another haplotype of a particular chromosome. "Haplotype phasing" involves resolving this original state or order of inherited genetic variation, i.e., which specific variations came from which parent (Figure 3.1).

**Figure 3.1 Evolution of Haplotypes.**




*Note:* A child inherits one chromosome from the father and one from the mother. In this example, there is a mutation at the second site of the paternal chromosome (C → T). There is also recombination on the maternal chromosome between the third and fourth site. Source:(Lo, 2014).

**Importance of Haplotype Phasing**

Haplotypes provide a holistic and detailed insight into the organism's genome and are much more informative than genotypes. In the polyploid genome, haplotypes refer to combined genetic variants on each homolog, highlighting the collective variation between the homologs. These unique variations in each homolog (aka chromosomal haplotypes) are vital in dissecting several evolutionary and molecular mechanisms related to allele-specific genetic and epigenetic changes (Adey et al., 2013; Shendure & Aiden, 2012; Snyder et al., 2015). They are also vital in association studies, detecting positive selection (Sabeti et al., 2002, 2007), understanding gene function, identifying recombination hotspots and recombination rates, and studying regions of the genome that are functionally related. Knowledge of chromosomal haplotype can also improve the accuracy of the analyses that depend on a diploid genome, such as competitive alignment of sequence reads to identify allele-specific expression (Pendleton et al., 2015; Seo et al., 2016). They also help identify a somatic mutation in heterogeneous cell populations (Loh et al., 2018; Nik-Zainal et al., 2012) and single-cell genomes (Zhang & Pellman, 2015), identifying associations between particular genes and disease. However, the associations we can detect with haplotypes cannot always be detected only using unphased genotypes (see Figure 3.2). (Atwell et al., 2011; Giakountis et al., 2009).

**Figure 3.2 Haplotypes Help With Detecting Associations. With Only Genotype Data, We Cannot Establish the Association Between an Individual With the Disease and Their Genotype.**

| Sample | Genotype | Haplotype |
|---|---|---|
| Disease | {(G,C), (G,A)}<br>{(0,1), (0,1)} | 'GG' and 'CA'<br>'00' and '11' |
| Disease | {(G,C), (G,G)}<br>{(0,1), (0,0)} | 'GG' and 'CG'<br>'00' and '10' |
| No Disease | {(G,C), (G,A)}<br>{(0,1), (0,1)} | 'GA' and 'CG'<br>'01' and '10' |

*Note:* But if haplotypes are known, we can detect an association of disease with haplotype 'GG' ('00'). Source: (Lo, 2014)

**Approaches to Haplotype Phasing**

Haplotype inference of polyploid genome requires knowledge of DNA sequence on each homologous chromosome. The most common method for phasing haplotype is analyzing the genotypes of related individuals (Browning & Browning, 2011; Kong et al., 2008; Loh, Palamara, et al., 2016). For example, the haplotype of an individual can be prepared by learning the genotypes of the parents or inferred from the genotypes of several related individuals (siblings) who provide "surrogate parent genotypes." Another method is applying statistical modeling of recombination based on genotype frequencies in a population (Loh, Danecek, et al., 2016). Statistical methods aim to estimate the probability of recombination between adjacent polymorphic sites from a panel of reference haplotypes and produce the haplotype of an individual as the most likely configuration based on the estimated recombination probabilities. Statistical phasing makes highly accurate local haplotypes, but long-range haplotypes are only possible for closely related individuals or huge sample cohorts. It also has limited power in identifying low-frequency haplotypes and cannot phase *de novo* mutations or rare variants.

More accurate haplotype information can be directly obtained by sequencing or genotyping chromosomal fragments that have homologs separated (Browning & Browning, 2011; Snyder et al., 2015). In this experimental separation method, the DNA sequences of each homolog are tagged before genotyping or sequencing. Experimental phasing is done at the single chromosome level (Fan et al., 2011; Ma et al., 2010; Porubsky et al., 2017; Yang et al., 2011; Zhang et al., 2015), which can produce whole-chromosome phased haplotype of an individual genome with no switch errors. But even in such cases, the haplotype is often incomplete due to uneven sequence coverage caused by amplification bias in the generation of sequencing libraries from a single chromosome. In addition, single chromosomes are difficult to isolate, and the process is highly laborious and not scalable.

**Rationale**

This tool described below was born out of the requirement to have a diploid phased genome for ASE (allele-specific expression) analyses (Chapter 4). A. lyrata is an outcrossing model with high heterozygosity, affecting the mappability of the sequence reads from two haplotypes when aligned on a single haploid reference genome. For RNAseq reads, these differences in mappability can lead to biased outcomes as analyses based on RNAseq mainly depend upon the count of reads aligned to the genes. Adding to that problem, our RNAseq reads are from F1 hybrids of plants from two different populations (Mayodan, NC, USA; and Spiterstulen, Norway). Given that the reference genome of A. lyrata has most of its variants from the North American population, ASE analyses required extra caution. We had to identify the sources of RNA sequence reads in F1 hybrids of a cross between parents from genetically diverse parental populations, and genome and transcriptome sequences were not available from the parents. This is especially important for RNAseq data analyses where the counts of reads

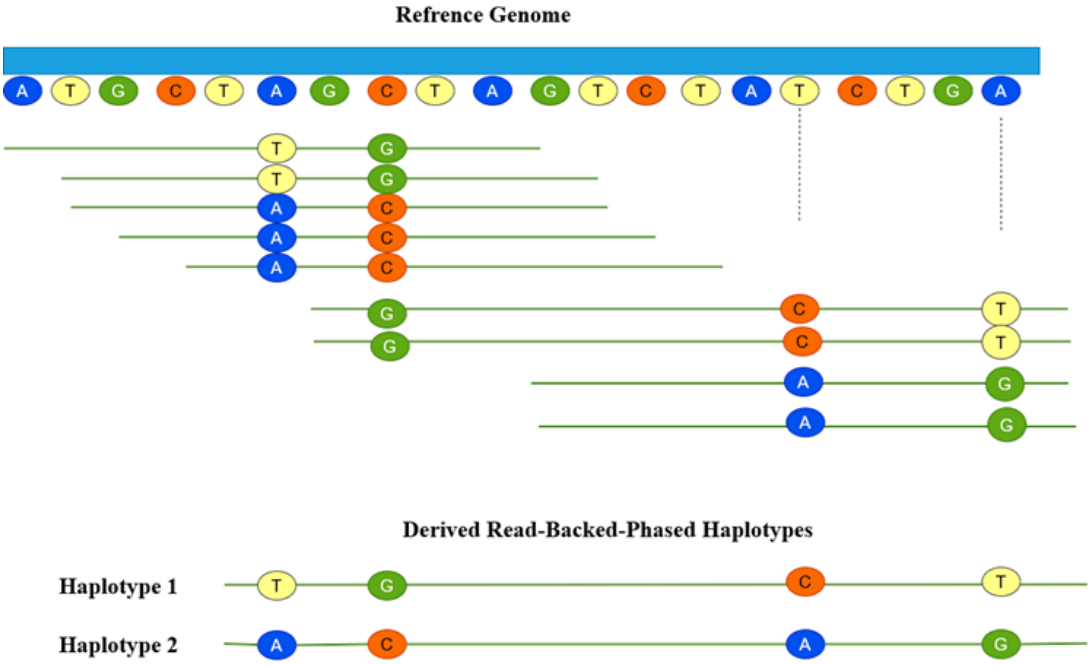
aligned have some significance. Another inherent problem was that A. lyrata does not have a population-specific genome published. The current genome (Hu et al., 2011) is closer to North American populations, including Mayodan.

This limitation regularly occurs in samples from many natural populations, including humans. Therefore, we developed an in-house tool that provides a comprehensive and more controlled approach to haplotype phasing and helps with phasing when very few samples are utilizing the read-backed-phased haplotypes.

**Read-backed-phased Haplotypes**

RBP (read-backed-phased) haplotypes are generally short, phased haplotype block generated by sequence reads that overlap two or more heterozygous variants. Due to the increase in the size of PE reads, RBP haplotypes are an increasingly common occurrence in current VCF databases. However, as discussed in the previous section, whole-genome level haplotype generation is still a challenge. Figure 3.3 shows a sequence read spanning two heterozygous sites. If there are multiple reads spanning multiple heterozygous sites, a large RBP haplotype can be prepared. Individual genome or transcriptome sequence reads provide phase information for variants on the same read, and overlapping reads sharing the same variants allow this to be extended.

**Figure 3.3 RBP Variants in the Sequence Reads Aligned to the Reference Sequence.**




*Note:* The former part of the alignment reads contains two heterozygous sites (T, A), (G, C). This read-level information helps phase those two sites. The reads in the later part of the alignment support the phasing of 3 heterozygous sites. If all the reads are taken into account, we can have 5 variants phased together because a few reads overlap at the heterozygous site (G, C).

**Objective**

Our objective is to create an algorithm and pipeline that helps to

- Phase consecutive read-backed-phased haplotypes in an individual diploid genome.
- Phase interspersed non-phased variants in an individual diploid genome between the two RBP haplotypes.

**Proposed Haplotype Phasing Method**

The primary intent of the proposed method is to take in partially phased haplotype information from several diploid samples (see Table 3.1 for more details) and apply one or more of the following:

- join two consecutive blocks,
- assign the haplotypes from an RBP block to a specific population if distinct populations are set in the input data,
- create short haplotypes using non-RBP variants, the unphased ones interspersed between two RBP blocks.

During the phasing process, other samples in the cohort that have variants (phased haplotype blocks and unphased genotypes) with similar properties provide the statistical power to solve haplotype phasing for the particular individual diploid variants. We expect this pipeline to be beneficial primarily to the community working with emerging models and organisms that are a part of a genomics study but do not have the necessary haplotype reference panels.

**Data Requirements**

The data required for this tool and pipeline can be prepared using:

- Phaser (Castel et al., 2016) – This application uses the aligned BAM or SAM file and VCF file to produce RBP VCF. The haplotype block size obtained using this method may range from 2 to higher. Generally, on average, 10 to 50 variants are phased within the block, but it mainly depends upon the size of the Single End or Paired End reads.
- VCF-Simplify (Giri, unpublished, https://github.com/everestial/VCF-Simplify ) – This application takes in the RBP VCF and simplifies the variants into a table format for further phasing. Other similar tools can be used to prepare the haplotype data in the required format.

## Algorithms

We provide three classes of algorithms:

- PhaseExtender – This algorithm/tool joins two consecutive RBP haplotype blocks in a proper configuration using likelihood estimates derived from a population of samples and computed using Markov chains.
- PhaseStitcher – This algorithm/tool reads a RBP haplotype from the F1 diploid hybrid and assigns each haplotype (in the short haplotype block) to a respective parental population given variant information of the two distinct populations are provided.
- ShortVariantPhaser – This algorithm/tool phases the non-RBP interspersed variants between two RBP blocks.

In Table 3.1, a typical haplotype dataset prepared using Phaser and VCF-Simplify is shown. The RBP VCF is simplified into a table, and numeric genotypes are converted into IUPAC bases. The data represents standard partially phased haplotype data prepared nowadays due to the increased size of PE reads and the availability of several tools to do RBP. In this dataset, **CHROM** represents the chromosome number. **POS** represents the genomic position. The example data contains four samples (**S1** to **S4**), **S*_PI** represents the index of a block, and **S*_PG** represents the phased genotype at that genomic position. The state of the genotype (phased or unphased) depends upon the **S*_PI**; if **PI** values are the same, the two consecutive genotypes are considered phased. Two blocks with different **PI** values are not in a phased state, e.g., the haplotype blocks belonging to **PI=3** and **PI=4** in sample **S1_PI** are not in a phased state with each other. Furthermore, any genotypes with "/" or "." in **PG** are not in a phased state with consecutive genotypes (above or below).

The objective of **phaseExtender** is to solve the phase state between two consecutive blocks; for example, **PI=3** and **PI=4** in sample **S1** using information from other samples. Similarly, consecutive blocks in a different sample can also be extended using data from other samples.

The objective of **phaseStitcher** is to assign haplotype (left and right) from a particular block to a specific population if samples in the data were categorized into two distinct populations. For example, if sample **S1** was a hybrid of two populations, **A** and **B**, this method helps determine if the left haplotype of S1 belongs to population **A** or **B**. All the haplotypes assigned to a particular population are then joined to create a genome-wide haplotype.

The objective of **shortVariantPhaser** is to solve phasing in interspersed variants where RBP was absent. For example, in sample **S3**, the interspersed variants with PI= "." are phased, and the index (**PI**) is updated with a new and unique value**.** The new phased block is then joined with consecutive blocks using **phaseExtender** or **phaseStitcher**.

**Table 3.1 A Typical Haplotype Blocks Produced by Readbackedphasing**

| CHROM | POS | S1_PI | S1_PG | S2_PI | S2_PG | S3_PI | S3_PG | S4_PI | S4_PG |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 3 | A\|T | 4 | A\|T | 2 | A\|T | 7 | T\|A |
| 2 | 11 | 3 | T\|G | 4 | T\|G | 2 | T\|G | 7 | G\|T |
| 2 | 16 | 3 | G\|C | 5 | G\|C | 2 | G\|C | 7 | C\|G |
| 2 | 26 | 3 | C\|A | 5 | A\|A | . | A/C | 7 | A\|C |
| 2 | 32 | 4 | G\|C | 5 | G\|C | . | G/C | 7 | C\|G |
| 2 | 48 | 4 | T\|G | 5 | T\|G | . | G/C | 7 | G\|T |
| 2 | 59 | 4 | A\|T | 5 | A\|T | 3 | A\|T | 7 | T\|C |
| 2 | 81 | 4 | C\|A | 5 | C\|A | 3 | C\|A | 7 | A\|C |
| 2 | 95 | . | T/C | 5 | C\|T | 3 | C\|G | 8 | C\|T |
| 2 | 99 | . | C/A | 5 | A\|C | 3 | A\|C | 8 | A\|C |

*Note:* The CHROM represent the chromosome number, POS represents the position of the allele in the genome. The other headers with *_PI indicate the block index of the phased haplotype block, and *_PG indicates the diploid genotype at that that OS. Two PI with same name and number indicates that the genotypes are in phased state.

**Algorithm #1: Phase-Extender**

We developed an algorithm/method to phase RBP haplotypes in an individual diploid genome by applying Markov chains between the adjacent blocks.

***Overview and Objective***

Consider a chromosome from a diploid organism (with $y$ heterozygous sites) split into $z$ Readback-Phased haplotype blocks of a random size such that:

$$L_{z_x} \in (1, y-1)$$

Each RBP block is represented by two haplotype strings in a diploid organism,

$$z = \{h, \bar{h}\}$$

Therefore, using the intuitive notation $H_z = \{h_z, \bar{h}_z\}$, a chromosome-wide haplotype is,

$$\mathrm{CW}_{\mathrm{hap}} = \{H_1\}, \{H_2\}, \ldots, \{H_z\}$$

However, unlike traditional haplotype phasing, where the genotype at each genomic position has to be phased, we solve the whole phased state by solving the phase state between two consecutive haplotype blocks at one time.

Given two adjacent RBP haplotypes, $H_1 = (h_1, \bar{h}_1)$ and $H_2 = (h_2, \bar{h}_2)$, the objective is to identify the most likely haplotype, $H = (h, \bar{h})$, in a sample ($S_i$) conditional upon the given reference haplotypes in other samples $h_{S=1}^{n}$. Analytically,

$$H = \arg\max_{H} \mathbb{P}(H|H_1, H_2|h_{S=1}^{n})$$

***Algorithm***

The most likely haplotype state, $H = \{h, \bar{h}\}$, given two consecutive haplotypes,

$$\begin{aligned} \text{Block 1: } H_1 &= \{h_1, \bar{h_1}\} \\ \text{Block 2: } H_2 &= \{h_2, \bar{h_2}\} \end{aligned}$$

can be,

$$H = \begin{cases} \{h_1 h_{2,} \overline{h}1\, \overline{h}_{2,}\} - \text{if phased in } parallel \\ \{h_1 \overline{h}_{2,} \overline{h_1} h_{2,}\} - \text{if phased in } alternate \end{cases}$$

To solve the phase state, we build a Markov chain from each site in Block 1 to each site in Block 2 and compute the likelihood of each possible configuration. If Block 1 had $m$ sites (indexed as $i$) and Block 2 has $n$ sites (indexed as $j$), and if "ATGC" represents the four possible nucleotide bases represented by $D$ as:

D: A=1, T=2, G=3, C=4

Then the maximum likelihood estimates of the configuration $L_1$ (parallel) and $L_2$ (alternate) are:

$$\begin{cases} L_1 = \prod_{k=1=l}^{2} \left\{ \prod_{i=1}^{m} \left[ \frac{N(D_i)}{2s} \prod_{j=1}^{n} \left( \frac{\sum_{f=1}^{s} N(D_{ij})}{N(D_i)} \right) \right] \right\} \\ L_2 = \prod_{k=1\neq l}^{2} \left\{ \prod_{i=1}^{m} \left[ \frac{N(D_i)}{2s} \prod_{j=1}^{n} \left( \frac{\sum_{f=1}^{s} N(D_{ij})}{N(D_i)} \right) \right] \right\} \end{cases}$$

where we compactly define $N(D_i)$ to be the number of allele $D$ at the $i$th position across the entire sample, and $\sum_{f=1}^{s} N\left(D_{ij}\right)$ as the sum of allele $D$ at the $i$th position to D at the $j$th position across the entire sample.

We then calculate the likelihood ratio *R* to join two haplotypes in either configuration.

$$R = \frac{L_1}{L_2}$$

So, if $\log_2^{(R)}$ is positive, "parallel configuration" is the more likely configuration; else "alternate configuration" is more likely. We can also set a threshold value for *R* that can be set for assigning phases. We provide a detailed process and likelihood estimation using examples in **Supplementary Materials S3.A**.

**Algorithm #2: Phase-Stitcher**

We developed this algorithm/method to phase RBP haplotypes in the F1 hybrid by haplotype segregation.

***Overview and Objective***

Consider a diploid hybrid sample $S_i$, with haplotype pairs $(h, \bar{h})$ and reference haplotypes from two populations, population A and population B, both with the number of haplotypes $m$ and $n$:

$$h^m_{S_a=1}, h^n_{S_b=1}$$

The objective is to assign haplotypes $H \circ (h, \bar{h})$ from a diploid phased (or RBP) F1 hybrid to the respective parental population. Analytically,

$$H \circ (h, \bar{h}) = \arg\max_H \mathbb{P}(h \in \{A, B\}, \bar{h} \in \{A, B\} | h^m_{S_a=1}, h^n_{S_b=1})$$

***Algorithm***

Given the reference haplotypes from two populations ($S_a$ and $S_b$) representing the hybrid, we can solve the phase assignment of each haplotype in $\{h, \bar{h}\}$ to either **population A** or **population B**, by estimating the likelihood of each haplotype conditional on it belonging to each population using Markov chain models.

$$h \in A, \bar{h} \in B$$
$$\bar{h} \in A, h \in B$$

To prepare a genome-wide haplotype, the haplotypes assigned to a particular population can be strung together. If the two haplotypes in a RBP haplotype block were $h$ and $\bar{h}$ and if populations A and B have alleles in $m$ sites (indexed as $i$), and if "ATGC" represents the four possible nucleotide bases represented by $D$ as,

D: A=1, T=2, G=3, C=4

the likelihood estimates of haplotype assignment are then,

$$\begin{cases} L_{h\in A} = \prod_{i=1}^{l} \frac{N_a(D_i)}{2S_a} \frac{\sum_{f=1}^{S_a} N_a\,(D_{ik})}{N_a(D_i)}, i < k \leq l \\ L_{h\in B} = \prod_{i=1}^{l} \frac{N_b(D_i)}{2S_b} \frac{\sum_{f=1}^{S_b} N_b\,(D_{ik})}{N_b(D_i)}, i < k \leq l \\ L_{\bar{h}\in A} = \prod_{i=1}^{l} \frac{N_a(D_i)}{2S_a} \frac{\sum_{f=1}^{S_a} N_a\,(D_{ik})}{N_a(D_i)}, i < k \leq l \\ L_{\bar{h}\in B} = \prod_{i=1}^{l} \frac{N_b(D_i)}{2S} \frac{\sum_{f=1}^{S_b} N_b\,(D_{ik})}{N_b(D_i)}, i < k \leq l \end{cases}$$

Where we intuitively define $N_a(D_i)$ to be the allele D at the $i$th position in population A, and $S_a$ to be the number of haplotypes in population A. Additionally, we define $\sum_{f=1}^{S_a} N_a\,(D_{ik})$ as the transition matrix counts for alleles from the $i$th to the $k$th position across the population. This is done to specify three scenarios: if the phase is known, $D_{ik} = 1$; if the phase is unknown, $D_{ik} = 0.5$; and if the phase does not occur, $D_{ik} = 0$.

We then calculate the likelihood ratio *R* to assign haplotype to the population they most likely belong to.

Likelihood ratio that left haplotype belongs to Population A vs Population B.

$$Odds_L = \frac{P_{L_{Pa}}}{P_{L_{Pb}}}$$

Likelihood ratio that right haplotype belongs to Population A vs Population B.

$$Odds_R = \frac{P_{R_{Pa}}}{P_{R_{Pb}}}$$

Odds that the left haplotype blongs to Population A and the right haplotype to Population B vs. the alternative assignment.

$$R = \frac{Odds_L}{Odds_R}$$

So, if $\log_2^{(R)}$ is positive, "left haplotype" is the more likely to belong to population A; else, "right haplotype" is more likely belong to population A. We can also set a threshold value for *R* that can be set for assigning haplotype to population. We provide a detailed process and likelihood estimation using examples in **Supplementary Materials S3.B**.

**Algorithm #3: Short-Variant-Phaser**

An algorithm/method developed to phase variants interspersed between RBP haplotypes in a diploid genome based on the HapHedge data structure and algorithm explained in Eagle2 (Loh, Danecek, et al., 2016).

***Objective***

Consider a panel of diploids (pairs of haplotypes) containing $n$ variant sites $S_k$ , such that $k \in (1, n)$. Our objective is to determine the most likely orientation of a given (unphased) target diploid $H = (h, \bar{h})$; that is, on a per-site basis, we determine if switching to the "alternative" orientation is more probable than staying in a predefined "parallel" orientation. Analytically, we can say our objective is to determine, for each site $S_k$,

$$H_k = \arg\ \mathrm{m}\ \ \mathbb{P}\left(H_k = \left(h_k, \bar{h}_k\right) \mid H_{k-1} = \left(h_{k-1}, \bar{h}_{k-1}\right)\right)$$

This quantity is calculated using the Positional Burrows-Wheeler Transform (PBWT) to determine the transition matrix. Positional Burrows-Wheeler Transform (PBWT) provides an appropriate data structure for bi-allelic data as it supports a run-length compressed representation of aligned haplotype data (Durbin, 2014). PBWT approaches haplotype matching using suffix array ideas. This compresses with run-length encoding on large data sets by more than a factor of a hundred smaller than using gzip on the raw data. With the increasing sample sizes, more multi-

allelic sites are expected to be observed. Hence, there is a necessity to handle multi-allelic genotype data(Naseri et al., 2019). In addition, PBWT data structure has been used for genotype imputation (Rubinacci et al., 2020).

***Algorithm***

We use a shorthand notation for the transformation of the variant site $S_k$. Let $S'_k$ denote the PBWT of the site $S_k$ such that,

$$\text{PBWT}\ (S_k): S_k \mapsto S'_k$$

allows us to determine the transition matrix $T_k$ between sites $S'_k$ and $S_{k+1}$. More specifically,

$$T_k = \begin{bmatrix} \frac{1}{c(0)} & \frac{1}{c(1)} \end{bmatrix} \cdot \begin{bmatrix} c(0_t 0) & c(0_t 1) \\ c(1_t 0) & c(1_t 1) \end{bmatrix}$$

Where, for $i, j \in \{0,1\}$, we define an operation $c(i)$ to count the number of transitions from the allele encoded $i$ in the site $S'_k$, and $c(i_t j)$ to count the number of transitions from $i$ to $j$ between sites $S'_k$ to $S_{k+1}$. Our problem described in the above section then reduces to determining the greatest elements in each row of $T_k$. **Supplementary Materials S3.C** provides a detailed process and likelihood estimation for this method**.**

## Application Test, Results, and Usage

This section deals with testing the haplotype phasing application Phase-Extender and compares the phase quality of the haplotype by comparing it with the known dataset belonging to a sample of human genomes, NA12891. This test and result can also be used as a test case for doing recursive phase-extension with RBP data derived using a phaser or other sources. A more extensive version of the test and tutorials are available at https://github.com/everestial/TestSwitchErrors.

**Test Data And Parameters**

I used phased genomes (HapMap3 release #2, NCBI build b36) released in 2009, b36 (Altshuler et al., 2012; Consortium et al., 2005; Howie et al., 2009) for testing phasing quality by phase-Extender. Data source - https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference, https://www.internationalgenome.org/data-portal/sample/NA12891

I set sample NA12891 as the target sample for phasing and other samples in the cohort as the reference samples. To emulate RBP haplotype blocks, we split each phased genome into fragments of varying lengths. The fragments were prepared by breaking the whole genome into blocks containing a random number of variants between 3 and 12 per block and flipping the haplotype position (left to right and vice-versa) randomly. Other samples in the cohort were also randomly fragmented and flipped, so it represented a typical dataset for haplotype phase extension. Overall, we prepared two sets of data, Set-A containing 10 samples including sample NA12891 and Set-B containing 25 samples including NA12891. The simulated RBP blocks (each containing 3 to 12 variants) were prepared assuming Poisson-like distribution. A Python script for this simulation is available as file "makeHapFile.py" in the source repo.

Phasing accuracy was measured using the proportion of adjacent haplotypes that were phased incorrectly with each other (switch errors) against the known dataset for sample NA12891.

***Phase extension was run using the following parameters.***

***Set-A:***

***Iteration 01***

$ item= ‘NA12891’

$ python3 –m phase-extender --input SetA/simulated_RBphasedHaplotype_SetA.txt --SOI ${item} --output SetA/phased_${item}_SetA_run01 --numHets 25 --lods 5 --writeLOD yes --hapStats yes --addMissingSites no

***Iteration 02***

$ python3 –m phase-extender --input SetA_02/phaseExtendedHaplotype_SetA_02.txt --SOI ${item} --output SetA_02/phased_${item}_SetA_run02 --numHets 40 --lods 1 --writeLOD yes --hapStats yes --addMissingSites no

***Set-B:***

***Iteration 01***

$ python3 –m phase-extender --input SetB/simulated_RBphasedHaplotype_SetB.txt --SOI {item} --output SetB/phased_${item}_SetB_run01 --numHets 25 --lods 5 --writeLOD yes --hapStats yes --addMissingSites no

***Iteration 02***

$ python3 –m phase-extender --input SetB_02/phaseExtendedHaplotype_SetB_02.txt --SOI {item} --output SetB_02/phased_${item}_SetB_run02 --numHets 40 --lods 1 -- writeLOD yes --hapStats yes --addMissingSites no

Note that haplotypes may not phase genome-wide in just two iteration and may require further iteration. The phasing process can be fine-tuned by adjusting parameters such as "lods", "numHets", "useSample", "snpTh".

***A short description of each of the arguments of phase extender are as follows:***

It denotes the number of heterozygotes. The maximum number of heterozygote SNPs is used from each consecutive block to compute the maximum likelihood estimate of each configuration between two blocks. The default value is 40.

**lods** :- log2 of Odds cut off threshold. The cutoff threshold used to extend consecutive haplotype blocks. The default value is set at (2^5 = 32 times likely). Two consecutive blocks will be joined in parallel configuration if computed log2(likelihood) > lods threshold

**useSample** :- Samples to use in the given input haplotype file (plus reference haplotype) to compute transition matrix. By default all the samples in the reference haplotype file and input file will be used.

**writeLOD** :- It is a *Boolean* argument which writes the calculated LODs between two consecutive haplotype blocks when processing phase extension to the output file if true is passed. Note: the 'lods-score' are printed regardless of if the consecutive blocks are joined or not.

hapStats :- Prepare descriptive statistics and histogram of the haplotype size distribution of the input haplotype file vs. extended haplotype for the sample of interest. By default, the program doesn't print haplotype statistics.

**AddMissingSites** :- It directs programs to include the non-phased and missing genotype data from the input haplotype file to the final phase-extended output file.

## Results and Discussion

The phase-extension using phase extender showed an increase in the number of variants/blocks phased with each iteration for both data sets (See, histogram, figure 3.3, 3.4, 3.5 both Set-A and figure 3.8, 3.9, 3.10 for Set-B). As expected, the switch-error (SE) was lower for Set-B (iteration 01, adjusted SE = 0.026; iteration 02, adjusted SE = 0.0144) compared to Set-A (iteration 01, adjusted SE = 0.046; iteration 02, adjusted SE = 0.017483), see Table 3.3 and Table 3.4. The phasing accuracy was also comparable to the one done by ShapeIT (SE = 0.03) (Zhan, 2017).

**Comparison Of Phase-Extender With Shape-IT**

We compared the results from Phase-Extender with results from SHAPEIT, a fast and accurate tool for the estimation of haplotypes (aka phasing) from genotype or sequencing data. It is a new computational algorithm to infer haplotypes under the genetic model of coalescence with recombination in Phasev2.1 (Delaneau et al., 2008). It uses binary trees to represent the sets of candidate haplotypes for each individual, making it faster than other alternatives. It can also be used in other HMM-based inferencing software from various fields.

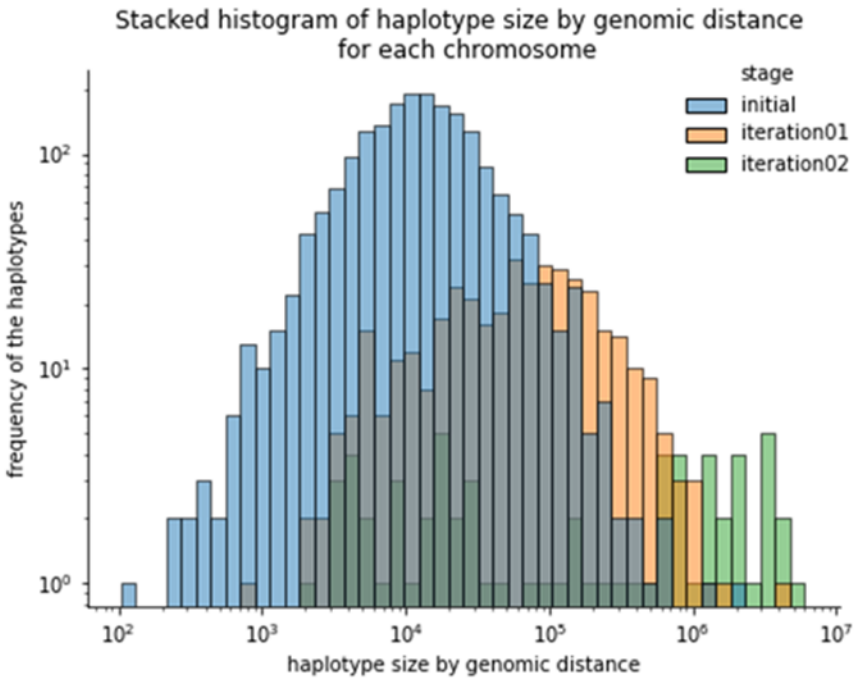
It is important to distinguish that ShapeIT application phased completely unphased NA12891 sample using 1010 phased reference samples producing SER=0.03, while phase-Extender produced adjusted SER= 0.032 with only 10 RBP samples. The SER for set-B was even lower (adj. SER = 0.00982). This result shows that phase-extension using RBP haplotypes provide a good method for phasing haplotypes.

Phase-Extender has not been tested for speed against other tools like ShapeIT, beagle given it's written in python, but the convenience and accuracy it would provide with small data makes this application an excellent utility for phasing haplotype. The most significant advantage is that the application can use few sequenced samples that are related to each other and can help phase each other. Overall, we expect Phase-Extender to be a tool of choice when phasing genome with a limited number of reference panels or in experiments where cohort of sequenced samples can help phase each other. Another benefit of phase-Extender is that it provides finer controls over haplotype phasing when accuracy and reiterative process is important during phasing process.

**Results: from phase-Extender on Set-A Dataset**

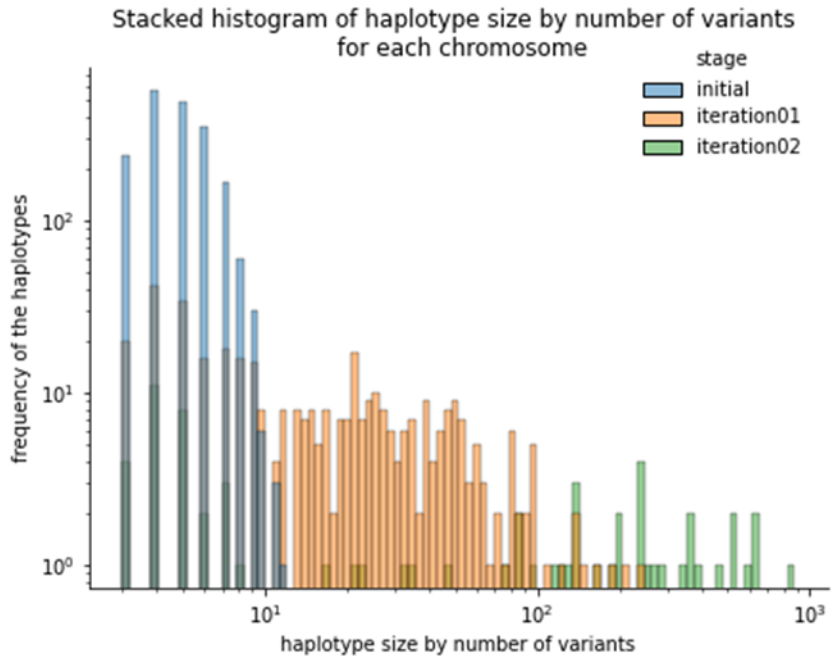
Table 3.2 shows the results of phase extension for set-A data after 2 iterations. The initial data set has 9666 variants distributed in 1933 RBP blocks. The block with minimum variants has 3 variants and block with maximum variants has 12 variants. The block covering the shortest genomic distance covers 104 bp and longest one covers 2057927 (bp – base pairs). The first iteration of phase extension joined several RBP blocks reducing the number of haplotypes to 391 and increasing the maximum number of variants to 237 variants in the new haplotype block. Phase extension also increased the genomic distance covered. Subsequently iteration-02 further increased haplotype phase extension.

**Figure 3.4 Histogram Showing the Frequency of Haplotype by Size of the Haplotype (Measured As Genomic Distance)**



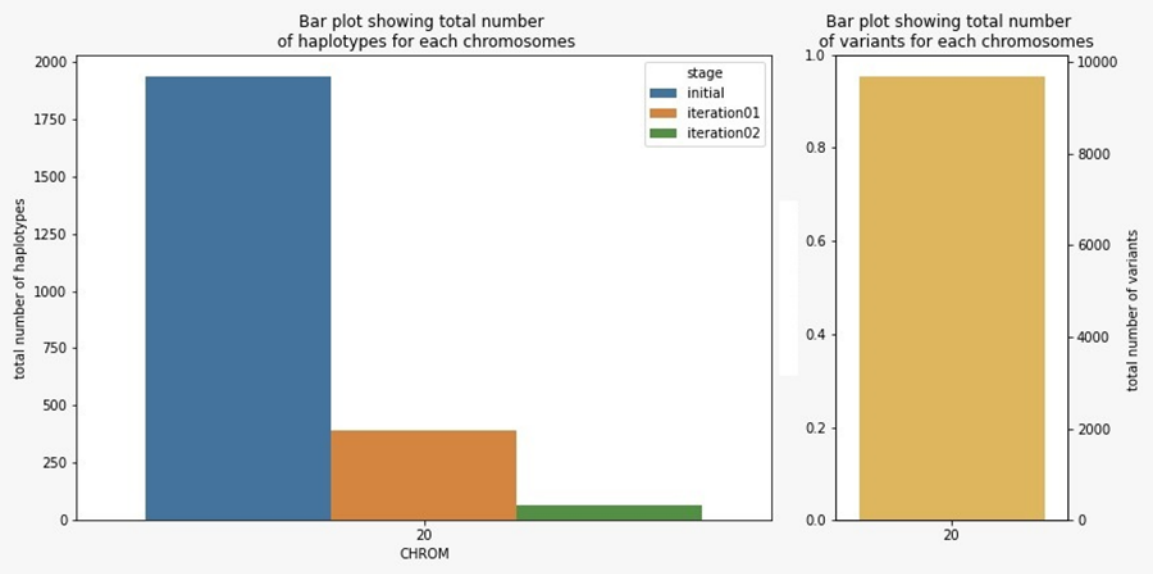
*Note:* With each iteration, the size of the haplotype increases with decrease in the number of haplotypes. With further iteration, we expect there to be a single haplotype block covering the whole genome.

**Figure 3.5 Histogram Showing Frequency of Haplotype by Size of the Haplotype (Measured As Number of Heterozygous Sites Within the Haplotype)**




*Note:* With each iteration the size of the haplotype increases with a decrease in the number of haplotypes. With further iteration, we expect a single haplotype block covering the whole genome.

**Figure 3.6 Bar Plot Showing Frequency of Haplotype in Each Iteration for Chromosome #20**

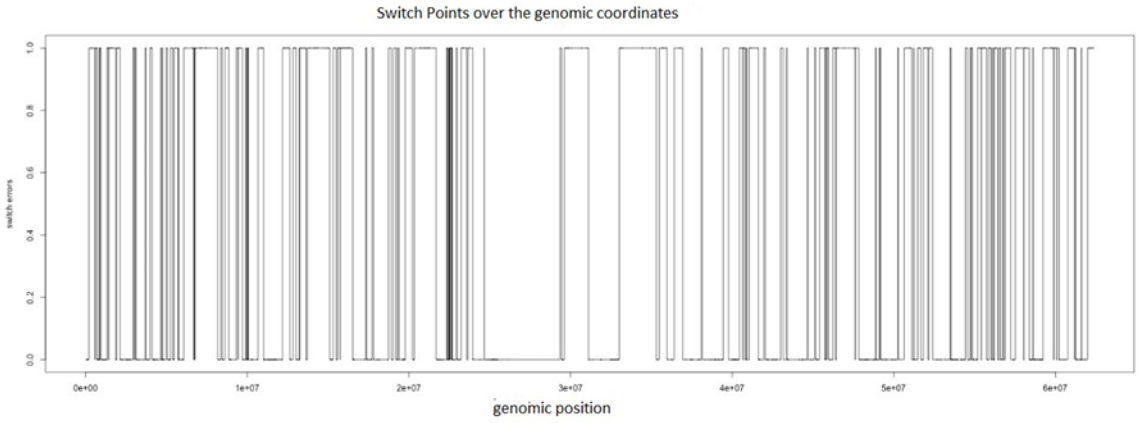



*Note:* Each iteration reduces the number of haplotypes, showing that smaller haplotype blocks are being extended into longer ones. With further iteration, we expect a single haplotype block covering the whole genome. Right subplot shows total numbers of variants present in chromosome.

**Table 3.1 Metrics of Changes in Haplotypes for Set-A Data**

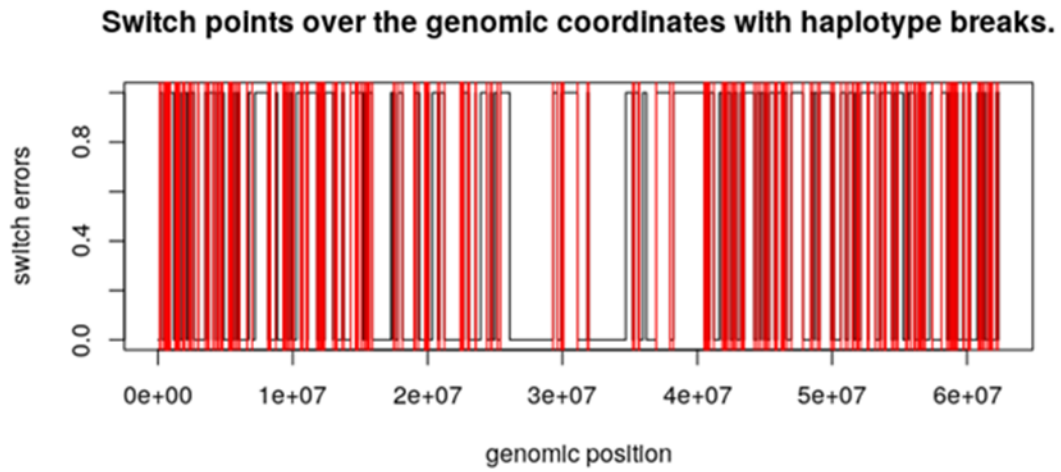
| Stage | No of variants | No of haplotypes | No of variants in (shortest, longest) block | Genomic distance in (shortest, longest) block | Switch Error Rate | Switch error rate (adjusted) |
|---|---|---|---|---|---|---|
| initial | 9666 | 1933 | 3 - 12 | 104 - 2057927 | - | - |
| iteration 1 | 9666 | 391 | 3 - 237 | 753 - 4386983 | 0.03062 | 0.046658 |
| iteration 2 | 9666 | 67 | 3 - 879 | 2245 - 6022477 | 0.01665 | 0.017483 |

**Figure 3.7 Switch Error Points After First Iteration of Phase Extension for Sample NA12891 Using Data Set-A**




*Note:* The switch error is calculated at 0.03062.

**Figure 3.8 Switch Error Overlayed With Haplotype Breaks for Sample NA12891 Using Data Set-A**



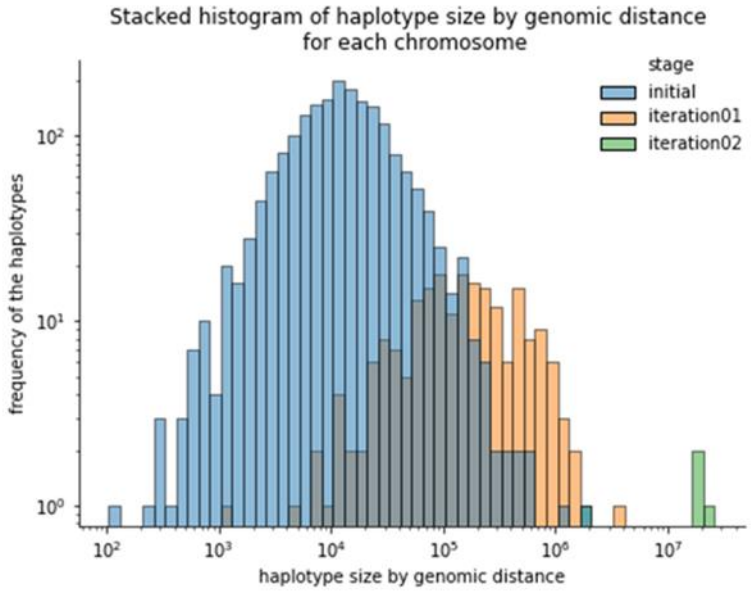

*Note:* The adjusted switch error is calculated at 0.04658 after accounting for haplotype breaks (two consecutive haplotypes which did not join).

**Results: from phase-Extender on Set-B Dataset**

Table 3.3 shows the results of phase extension for set-B data after 2 iterations. Initially the data has 9666 variants distributed in 1933 RBP blocks. The block with minimum variants has

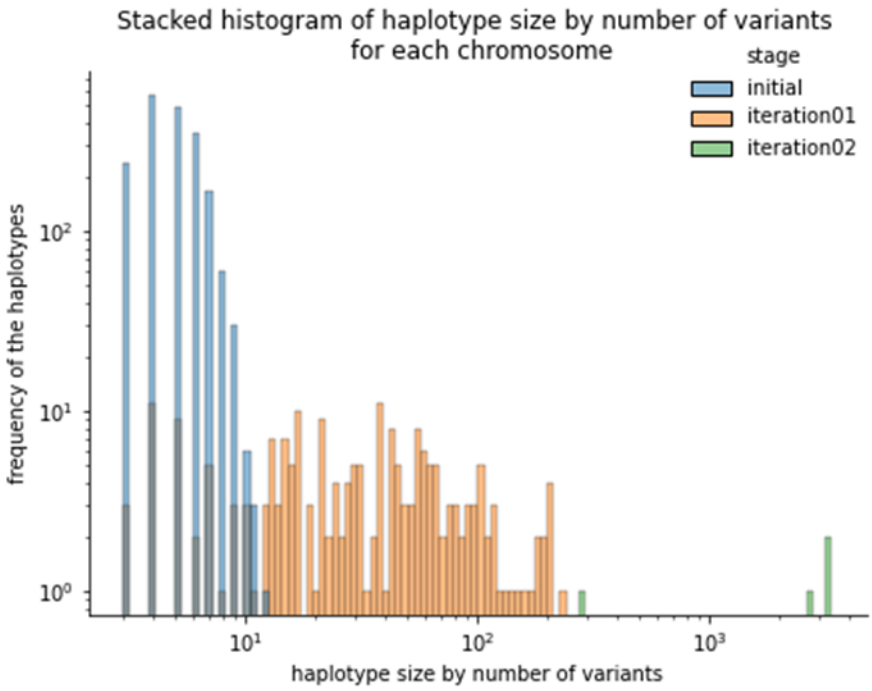
3 variants and the block with maximum variants has 12 variants. The block covering the shortest genomic distance covers 104 bp and the longest one covers 2057927 (bp – base pairs). The first iteration of phase extension joined several RBP blocks reducing the number of haplotypes to 391 and increasing the maximum number of variants to 237 in the new haplotype block. Phase extension also increases the genomic distance covered. Subsequently iteration-02 further increased haplotype phase extension.

**Figure 3.9 Histogram Showing Frequency of Haplotype by Size of the Haplotype (Measured As Genomic Distance)**



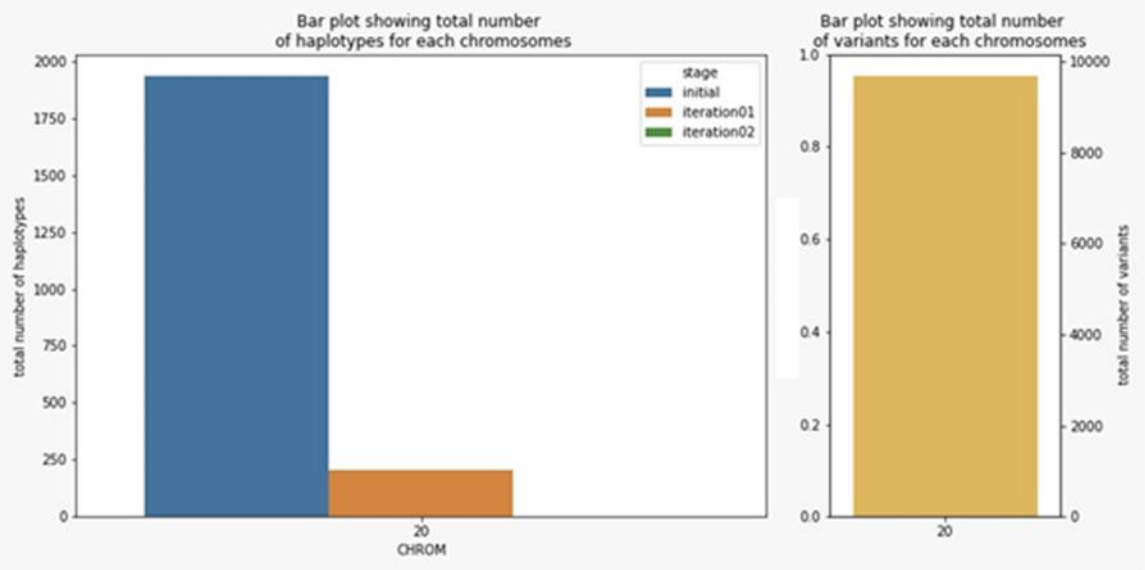

*Note:* With each iteration the size of the haplotype increases with decrease in the number of haplotypes. With further iteration, we expect a single haplotype block covering the whole genome.

**Figure 3.10 Histogram Showing Haplotype Frequency by Size of the Haplotype (Measured As Number of Heterozygous Sites Within the Haplotype)**




*Note:* With each iteration, the size of the haplotype increases with a decrease in the number of haplotypes. With further iteration, we expect a single haplotype block covering the whole genome.

**Figure 3.11 Bar Plot Showing the Frequency of Haplotype in Each Iteration for Chromosome**




*Note:* Note that each iteration reduces the haplotype frequency, suggesting smaller haplotype blocks are being extended into longer ones. We expect a single haplotype block covering the whole genome with further iteration.

**Table 3.2 Metrics of Changes in Haplotypes for Set-B Data**

| Stage | No of variants | No of haplotypes | No of variants in (shortest, longest) block | Genomic distance in (shortest, longest) block | Switch Error Rate | Switch error rate (adjusted) |
|---|---|---|---|---|---|---|
| initial | 9666 | 1933 | 3 - 12 | 104 - 2057927 | - | - |
| iteration 01 | 9666 | 207 | 3 - 232 | 1290 - 3442116 | 0.01572522 | 0.02607076 |
| iteration 02 | 9666 | 4 | 284 - 3386 | 2245 - 26087506 | 0.01417339 | 0.0144837575 |

## Application Repos

All three applications are available at

1. https://github.com/everestial/phase-extender
2. https://github.com/everestial/phase-stitcher
3. https://github.com/everestial/short-variant-phaser

A detailed description and test result is provided in **Supplementary Materials S3.D**. Due to time and resources constraints; only application **phase-Extender** has been tested in detail.

Web source for phasing on same sample NA12891 using ShapeIT.

1. https://gist.github.com/zhanxw/3c4e764cf1a3be6eb74c88dff08be3f4
2. https://portal.biohpc.swmed.edu/content/training/bioinformatics-nanocourses/gwas/zhan-phasing-workshop/

## Supplementary Materials: Chapter III

### Supplementary Materials S3.A: Phase-Extender in Detail

Example showing the detailed computation of haplotype phasing using Phase-Extender.

This section will provide a quantitative explanation of the Phase-Extender algorithm meant to prepare long-range haplotypes (and possibly genome-wide haplotype) by threading two adjacent RBP haplotypes. Phase-Extender applies the LD test between two consecutive blocks to estimate the possible configuration for phase extension.

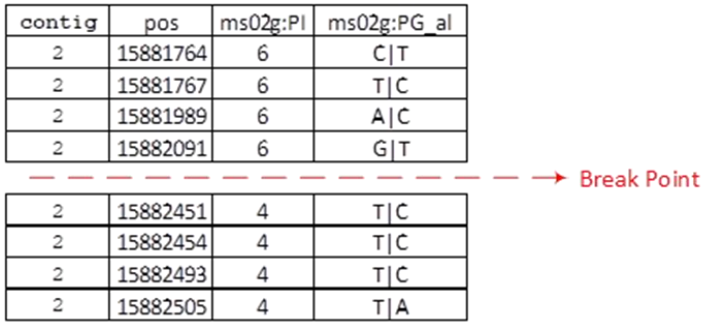
**Table S3.A1: A typical VCF file produced by phaser**

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | ms01e | ms02g | ms03g | ms04h | ms05h | ms06h |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 15881764 | . | T | C | 1253 | PASS | . | GT:PI:PG | 0/1:4:0\|1 | 0/1:6:0\|1 | 0/0:.:0/0 | 0/0:.:0/0 | . | . |
| 2 | 15881767 | . | C | T | 9683 | PASS | . | GT:PI:PG | 0/0:4:0\|0 | 0/1:6:0\|1 | 0/0:.:0/0 | 0/0:.:0/0 | . | . |
| 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 15882451 | . | T | C | 9683 | PASS | . | GT:PI:PG | 0/1:4:1\|0 | 0/1:4:0\|1 | 0/0:.:0/0 | 0/0:.:0/0 | . | . |
| 2 | 15882454 | . | T | C | 9683 | PASS | . | GT:PI:PG | 0/1:4:1\|0 | 0/1:4:0\|1 | 0/0:.:0/0 | 0/0:.:0/0 | . | . |
| 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

**Table S3.A2: A typical haplotype file produced from the VCF (not exact, though)**

| CHROM | POS | REF | all-alleles | ms01e:PI | ms01e:PG_al | ms02g:PI | ms02g:PG_al | ms03g:PI | ms03g:PG_al | ms04h:PI | ms04h:PG_al | ms05h:PI | ms05h:PG_al | ms06h:PI | ms06h:PG_al |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 15881764 | . | . | 4 | C\|T | 6 | C\|T | 7 | T\|T | 7 | T\|T | 7 | C\|T | 7 | C\|T |
| 2 | 15881767 | . | . | 4 | C\|C | 6 | T\|C | 7 | C\|C | 7 | C\|C | 7 | T\|C | 7 | C\|C |
| 2 | 15881989 | . | . | 4 | C\|C | 6 | A\|C | 7 | C\|C | 7 | C\|C | 7 | A\|T | 7 | A\|C |
| 2 | 15882091 | . | . | 4 | G\|T | 6 | G\|T | 7 | T\|A | 7 | A\|A | 7 | A\|T | 7 | A\|C |
| 2 | 15882451 | . | . | 4 | C\|T | 4 | T\|C | 7 | T\|T | 7 | T\|T | 7 | C\|T | 7 | C\|A |
| 2 | 15882454 | . | . | 4 | C\|T | 4 | T\|C | 7 | T\|T | 7 | T\|T | 7 | C\|T | 7 | C\|T |
| 2 | 15882493 | . | . | 4 | C\|T | 4 | T\|C | 7 | T\|T | 7 | T\|T | 7 | C\|T | 7 | C\|T |
| 2 | 15882505 | . | . | 4 | A\|T | 4 | T\|A | 7 | T\|T | 7 | T\|T | 7 | A\|C | 7 | A\|T |

**Figure S3.A1 Representing a breakpoint in the "sample – ms02g"**



| contig | pos | ms02g:PI | ms02g:PG_al |
|---|---|---|---|
| 2 | 15881764 | 6 | C\|T |
| 2 | 15881767 | 6 | T\|C |
| 2 | 15881989 | 6 | A\|C |
| 2 | 15882091 | 6 | G\|T |
| | | | Break Point |
| 2 | 15882451 | 4 | T\|C |
| 2 | 15882454 | 4 | T\|C |
| 2 | 15882493 | 4 | T\|C |
| 2 | 15882505 | 4 | T\|A |

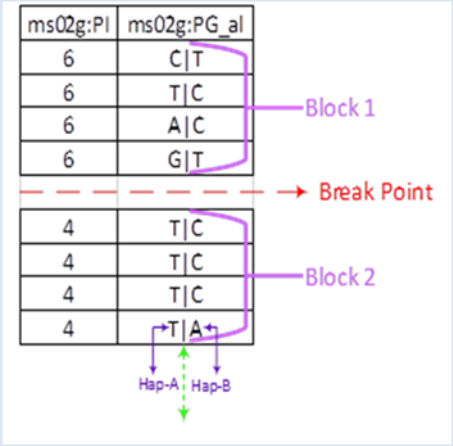
In the above haplotype, there is a breakpoint in sample **ms02g** at position 15882091-15882451. The RBP haplotypes are [C-T-A-G, T-C-C-T] at index PI=6, and [T-T-T-T and C-C-C-A] at index **PI=4**. We want to solve which phase from **PI=6** connects with **PI=4**. Given that all other samples have haplotype intact that bridges this breakpoint position, we can compute LD between the two blocks using the other samples and find the most likely configurations for joining the two haplotypes.

Looking at the data, the left block of PI-6 (C-T-A-G) is more likely to phase with the right block of PI-4 (C-C-C-A) generating C-T-A-G-C-C-C-A and T-C-C-T-T-T-T-T. Therefore, we use the first-order Markov chain to compute the likelihood estimates extend the haplotype in the most likely configuration.

**Steps:**

1. Prepare emission and transition probability matrix to estimate likelihoods.
2. Compute maximum likelihood estimate of each haplotype configuration.

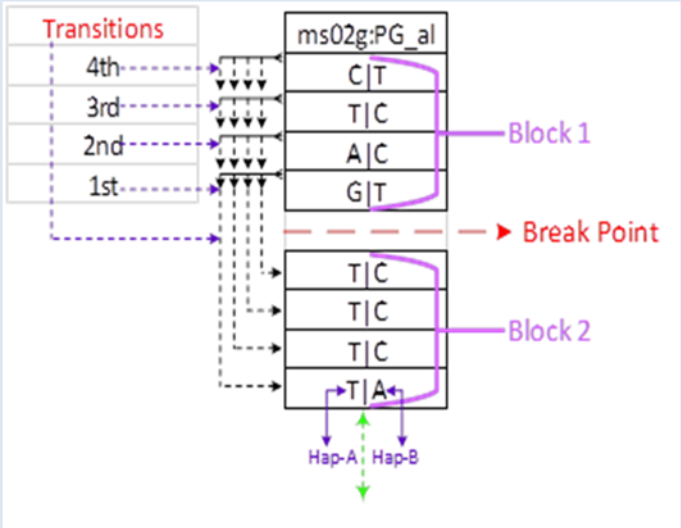
**Figure S3.A2 Two consecutive haplotype blocks**




*Note:* Top PI-6 is Block-1, and the bottom PI-4 is Block-2. The phased haplotype in the left is Hap-A and on the right is Hap-B.

**Figure S3.A3 Allele transition from all sites of former block-01 to all sites of later block-02**

**Identify possible phase configuration**




- Parallel Configuration:

  Block 1-HapA with Block 2-HapA, so B01-HapB with B02-HapB

- Alternate Configuration:

  Block 1-HapA with Block 2-HapB, so B 1-HapB with B 2-HapA

**Build Markov chains for likelihood estimation**

Start at two sites of the blocks closest to each other (here, the positions are 15882091 and 15882451). Starting at the nearest site maximizes the two blocks' LD (linkage disequilibrium) estimates under the condition the whole block is not used for LD estimates.

1. Then, estimate emission counts of each nucleotide (A, T, G, C); see Figure-A1. And convert the emission counts to emission probabilities; see Table S3.A3 and S3.A4.
2. Estimate transition counts from each nucleotide (A, T, G, C) of PI-6 to each nucleotide (A, T, G, C) of PI-4 for both haplotype configurations across all the samples. See Table 04-A. The observed transition is counted as "1" if the "PI value" match between the former and later nucleotide, else as "0.5". Transition counts are then converted to transition probabilities, see Table A6.
3. The likelihood estimates are maximized along the chain using either a max-sum or max-product approach.
4. Finally, log2Odds (of the maximum likelihood estimates) are computed to identify the possible haplotype configuration.

**Table S3.A3 Emission counts of each possible nucleotide**

| ↓ pos\allele → | A | T | G | C | total |
|---|---|---|---|---|---|
| 15882091 | 5 | 4 | 2 | 1 | 12 |
| 15882451 | 1 | 7 | 0 | 4 | 12 |

**Table S3.A4 Emission probabilities of each possible nucleotide**

| ↓pos\allele→ | A | T | G | C | total |
|---|---|---|---|---|---|
| 15882091 | 5/12 | 4/12 | 2/12 | 1/12 | 1 |
| 15882451 | 1/12 | 7/12 | 0 | 4/12 | 1 |

*Note:* Count the emission values for all the nucleotides (A, T, G, C) at each position of two consecutive blocks. Then convert the emission counts to emission probabilities.

**Calculation of transition counts and probability**

**Table S3.A5 Representation of transition matrix (counts)**

| | → to (pos 15882451) | | | | |
|---|---|---|---|---|---|
| **from↓** | A | T | G | C | total |
| A | 0 | 3 | 0 | 2 | 5 |
| T | 0 | 3.5 | 0 | 0.5 | 4 |
| G | 0 | 0.5 | 0 | 1.5 | 2 |
| C | 1 | 0 | 0 | 0 | 1 |

**Table S3.A6 Representation of transition matrix (probabilities)**

| | → to (pos 15882451) | | | | |
|---|---|---|---|---|---|
| **from↓** | A | T | G | C | total |
| A | 0 | 3/5 | 0 | 2/5 | 1 |
| T | 0 | 3.5/4 | 0 | 0.5/4 | 1 |
| G | 0 | 0.5/2 | 0 | 1.5/2 | 1 |
| C | 1/1 | 0 | 0 | 0 | 1 |

*Note:* Transitions are computed starting with the closest heterozygote sites between the two blocks. Therefore, the "1st" transition begins with the heterozygous sites of two blocks most close to each other. Likewise, the "2nd" transition begins from the 2nd nearest heterozygous site of the former block with the first heterozygous site of the later block. All other transitions are computed similarly.

**Figure S3.A4 "Emission probabilities" of nucleotides at position 15882091.**

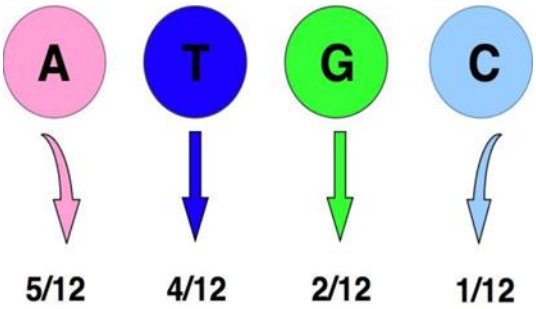
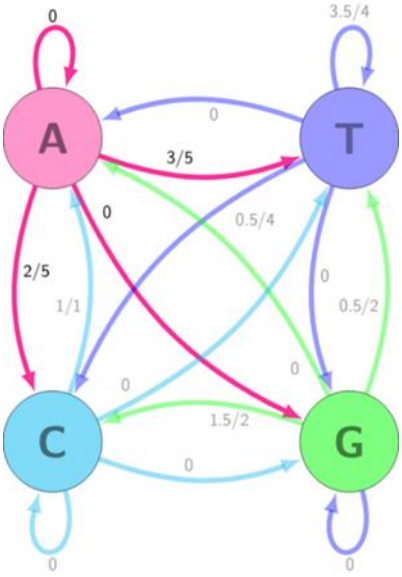



**Figure S3.A5 Transition probabilities of nucleotides to position 15882451 following emissions at position 15882091**




**Estimate maximum likelihood**

The maximum likelihood for a Markov chain is estimated as or "maximized product" of all the "p(emission) x p(transitions)" in the chain. Below, I show the maximum likelihood estimate for an example snippet using the max-product approach (using just two positions) and another example with two small haplotype blocks.

**Estimation of maximum likelihood with examples**

**Example 1. Computing Maximum Likelihood For Just Two Sites**

If we consider just two positions (15882091 & 15882451), we can estimate the maximum likelihood as:

**Table S3.A7 Example 1 Parallel configuration**

| Parallel configuration: | Alternate configuration: |
|---|---|
| G \| T | G \| T |
| T \| C | C \| T |
| $P_P = P(G) \times P(GtT) \times P(T) \times P(TtC)$ | $P_A = P(G) \times P(GtC) \times P(T) \times P(TtT)$ |
| $= \frac{2}{12} \times \frac{.5}{2} \times \frac{4}{12} \times \frac{.5}{4}$ | $= \frac{2}{12} \times \frac{1.5}{2} \times \frac{4}{12} \times \frac{3.5}{4}$ |
| $= .001736$ | $= 0.036458$ |
| $P_{P_{Avg}} = \frac{P_P}{2} = .000868$ | $P_{A_{Avg}} = \frac{P_A}{2} = .018229$ |

Note:

"$AtC$"$->$ represents "A" to "C" transition

**Estimate likelihood ratio(R):**

$$R = \frac{P_{P_{Avg}}}{P_{A_{Avg}}} = \frac{0.000868}{0.018229} = 0.047619$$

So, if $\log_2^{(R)}$ is positive, "parallel configuration" is the more likely configuration; else, "alternate configuration" is more likely. Here, in example-1 $\log_2^{(R)} = -4.392$, suggesting alternate-configuration is more likely.

**Example 2. Maximum Likelihood Using Data From The Whole Block**

We run a Markov chain from each site in the former block to each site in the later block. Only a certain number of sites may be used in a real dataset when blocks have many sites. This capability is also provided in the built application. We maximize the score by multiplying the likelihood of each "emission-transition" estimate.

**Likelihood estimates of Parallel Configuration (Block-1-Hap-A (C-T-A-G) with Block-2-Hap-A (T-T-T-T))**

$$
\begin{aligned}
P_P &= P_{HA} \times P_{HB} \\
P_{HA} &= P(G) \times \{P(GtT) \times P(GtT) \times P(GtT) \times P(GtT)\} \times P(A) \times \{P(AtT) \times P(AtT) \times P(AtT) \times P(AtT)\} \\
&\quad \times P(T) \times \{P(TtT) \times P(TtT) \times P(TtT) \times P(TtT)\} \times P(C) \times \{P(CtT) \times P(CtT) \times P(CtT) \times P(CtT)\} \\
&= \left(\frac{2}{12}\right) \times \left\{\frac{.5}{2} \times \frac{.5}{2} \times \frac{.5}{2} \times \frac{.5}{2}\right\} \times \left(\frac{3}{12}\right) \times \left\{\frac{.5}{3} \times \frac{.5}{3} \times \frac{.5}{3} \times \frac{.5}{3}\right\} \\
&\quad \times \left(\frac{2}{12}\right) \times \left\{\frac{.5}{2} \times \frac{.5}{2} \times \frac{.5}{2} \times \frac{.5}{2}\right\} \times \left(\frac{4}{12}\right) \times \left\{\frac{.5}{4} \times \frac{.5}{4} \times \frac{.5}{4} \times \frac{.5}{4}\right\} \\
&= 0.000651 \times 0.0001929 \times 0.000651 \times 0.00008138 \\
&= 6.6529E-15
\end{aligned}
$$

$$
\begin{aligned}
P_{HB} &= P(T) \times \{P(TtC) \times P(TtC) \times P(TtC) \times P(TtA)\} \times P(C) \times \{P(CtC) \times P(CtC) \times P(CtC) \times P(CtA)\} \\
&\quad \times P(C) \times \{P(CtC) \times P(CtC) \times P(CtC) \times P(CtA)\} \times P(T) \times \{P(TtC) \times P(TtC) \times P(TtC) \times P(TtA)\} \\
&= \left(\frac{4}{12}\right) \times \left\{\frac{.5}{4} \times \frac{.5}{4} \times \frac{.5}{4} \times \frac{.5}{4}\right\} \times \left(\frac{8}{12}\right) \times \left\{\frac{1.5}{8} \times \frac{1.5}{8} \times \frac{1.5}{8} \times \frac{1.5}{8}\right\} \\
&\quad \times \left(\frac{10}{12}\right) \times \left\{\frac{2.5}{10} \times \frac{2.5}{10} \times \frac{2.5}{10} \times \frac{2.5}{10}\right\} \times \left(\frac{8}{12}\right) \times \left\{\frac{.5}{8} \times \frac{.5}{8} \times \frac{.5}{8} \times \frac{.5}{8}\right\} \\
&= 0.00008138 \times 0.00082397 \times 0.003255 \times 0.0000101725 \\
&= 2.22E-15
\end{aligned}
$$

$$
P_{P_{Avg}} = P_{HA} \times P_{HB} = (6.6529E-15) \times (2.22E-15) = 1.477E-29
$$

**Likelihood estimates of Alternate Configuration(Block-1-Hap-A (C-T-A-G) with Block-2-Hap-B (C-C-C-A))**

$$
\begin{aligned}
P_A &= P_{HA} \times P_{HB} \\
P_{HA} &= P(G) \times \{P(GtC) \times P(GtC) \times P(GtC) \times P(GtA)\} \times P(A) \times \{P(AtC) \times P(AtC) \times P(AtC) \times P(AtA)\} \\
&\quad \times P(T) \times \{P(TtC) \times P(TtC) \times P(TtC) \times P(TtA)\} \times P(C) \times \{P(CtC) \times P(CtC) \times P(CtC) \times P(CtA)\} \\
&= \left(\frac{2}{12}\right) \times \left\{\frac{1.5}{2} \times \frac{1.5}{2} \times \frac{1.5}{2} \times \frac{1.5}{2}\right\} \times \left(\frac{3}{12}\right) \times \left\{\frac{2.5}{3} \times \frac{2.5}{3} \times \frac{2.5}{3} \times \frac{2.5}{3}\right\} \\
&\quad \times \left(\frac{2}{12}\right) \times \left\{\frac{1.5}{2} \times \frac{1.5}{2} \times \frac{1.5}{2} \times \frac{1.5}{2}\right\} \times \left(\frac{4}{12}\right) \times \left\{\frac{3.5}{4} \times \frac{3.5}{4} \times \frac{3.5}{4} \times \frac{3.5}{4}\right\} \\
&= 0.0527344 \times 0.1205633 \times 0.0527344 \times 0.195394 \\
&= 86.55E-5
\end{aligned}
$$

$$
\begin{aligned}
P_{HB} &= P(T) \times \{P(TtT) \times P(TtT) \times P(TtT) \times P(TtAT)\} \times P(C) \times \{P(CtT) \times P(CtT) \times P(CtT) \times P(CtT)\} \\
&\quad \times P(C) \times \{P(CtT) \times P(CtT) \times P(CtT) \times P(CtT)\} \times P(T) \times \{P(TtT) \times P(TtT) \times P(TtT) \times P(TtT)\} \\
&= \left(\frac{4}{12}\right)\left\{\frac{3.5}{4} \times \frac{3.5}{4} \times \frac{3.5}{4} \times \frac{2.5}{4}\right\} \times \left(\frac{8}{12}\right)\left\{\frac{5.5}{8} \times \frac{6.5}{8} \times \frac{6.5}{8} \times \frac{6.5}{8}\right\} \\
&\quad \times \left(\frac{10}{12}\right)\left\{\frac{6.5}{10} \times \frac{7.5}{10} \times \frac{7.5}{10} \times \frac{6.5}{10}\right\} \times \left(\frac{8}{12}\right)\left\{\frac{6.5}{8} \times \frac{7.5}{8} \times \frac{7.5}{8} \times \frac{6.5}{8}\right\} \\
&= 0.139567 \times 0.245839 \times 0.198047 \times 0.38681 \\
&= 2.628448E-03
\end{aligned}
$$

$$P_{A_{Avg}} = P_{HA} \times P_{HB} = 86.55E-5 \times 2.628448E-03 = 2.2745E-06$$

**Likelihood estimate of Parallel vs. Alternate configuration**

$$R = \frac{P_P}{P_A} = \frac{1.477E-29}{2.2745E-06} = 6.4937E-22$$

Here, In example-2 $\log_2^{(R)} = -70.38$ suggesting alternate configuration is more likely.

**Table S3.A8 Output Data from phase-Extender**

| contig | pos | ref | all-alleles | ms02g_PI | ms02g_PG_al |
|---|---|---|---|---|---|
| 2 | 15881764 | . | . | 6 | C\|T |
| 2 | 15881767 | . | . | 6 | T\|C |
| 2 | 15881989 | . | . | 6 | A\|C |
| 2 | 15882091 | . | . | 6 | G\|T |
| 2 | 15882451 | . | . | 6 | C\|T |
| 2 | 15882454 | . | . | 6 | C\|T |
| 2 | 15882493 | . | . | 6 | C\|T |
| 2 | 15882505 | . | . | 6 | A\|T |

**Supplementary Materials S3.B: Phase-Stitcher in Detail**

**Example showing a detailed computation of haplotype phasing using Phase-Stitcher.**

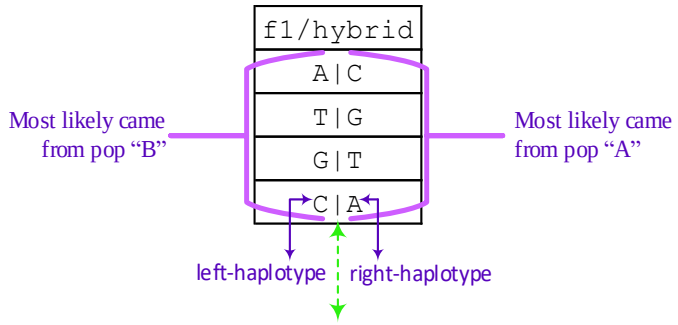
This section will provide a quantitative explanation of the phase-stitcher algorithm for preparing long-range haplotypes (and possibly genome-wide haplotype) in an F1 hybrid.

**Table S3.B1 A typical haplotype file containing data from F1 hybrid and two representative parental populations**

| **pos** | **f1/hybrid** | **a1** | **a2** | **a3** | **a4** | **a5** | **a6** | **b1** | **b2** | **b3** | **b4** | **b5** | **b6** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | A\|C | C\|C | T\|C | C\|G | C\|C | C\|C | C\|C | A\|A | A\|A | A\|A | T\|A | C\|A | A\|A |
| 17 | T\|G | G\|G | C\|G | G\|G | G\|G | G\|G | G\|G | T\|T | A\|T | T\|T | T\|G | T\|T | T\|T |
| 23 | G\|T | T\|T | T\|T | A\|C | T\|G | T\|T | T\|T | G\|G | G\|G | A\|G | G\|G | G\|G | G\|G |
| 37 | C\|A | A\|A | A\|A | A\|A | A\|A | A\|A | A\|A | C\|C | C\|A | C\|C | C\|C | C\|A | C\|A |

*Note*: this haplotype output does not match the VCF file shown in Table S3.A1 and Table S3.A2.

**Figure S3.B1 A haplotype representing a hybrid sample. Based on the data (Table B1) we know which haplotype came from either "A" or "B"**




Intuitively, we can tell that the left haplotype of the "F1" hybrid is more likely to have come from the population "B," and the right haplotype is more likely to have come from population "A," see data in **Table S3.B1**. To solve the phase assignment, we prepare the first order Markov-chain of emission and transition matrix to compute the likelihood estimates. We then take $log_2^{Odds}$ of the ratio of the computed likelihoods and assign each haplotype to the most likely population it came from. Several assigned haplotypes from F1 hybrids can then be joined to prepare a genome-wide haplotype.

**Steps:**

1. Prepare emission and transition probability matrix to estimate likelihoods.
2. Compute maximum likelihood and assign haplotypes to respective populations.

Identify possible haplotype assignment

- Left haplotype belong to population A vs B.
- Left haplotype belong to population A vs B.

**Figure S3.B2 Markov chains of allele transition matrix in simple case**

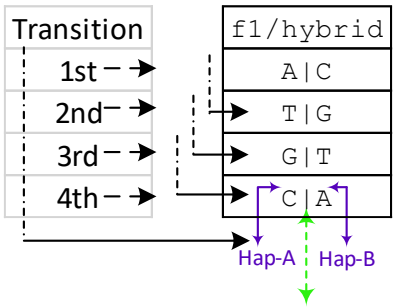
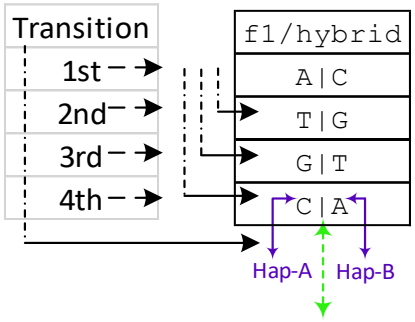



**Figure S3.B3 Markov chains of allele transition matrix in complex case**




**Build Markov chains for likelihood estimation**

- Estimate the emission counts of each nucleotide (A, T, G, C) at each position for each population; see **Table S3.B2** (for population A) and **Table S3.B3** (for population B). To prevent the likelihood estimates from turning out to zero due to non-observed alleles, all the alleles are given default starting counts of 0.25. The count is then added to the actual observed counts; see **Table S3.B4** and S3.**B5**. The emission counts are then converted to emission probabilities. See **Table S3.B6** and **Table S3.B7**
- Estimate transition counts from each nucleotide (A, T, G, C) in the former position to each nucleotide (A, T, G, C) in the later positions, see **Table S3.B8** and **Table S3.B9**. The observed transition is counted as "1" if the "PI value" match between the former and later nucleotide in that sample, else as "0.5".

- To prevent the likelihood estimates turning out to zero due to non-observed allele transition (or haplotypes), all the possible changes (or haplotypes) are given default starting counts of 1/16 (i.e., 0.0625). The count is then added to the actual observed counts; see **Table S3.B10** and **Table S3.B11**. Finally, the transition counts are converted to transition probabilities; see **Table S3.B12** and **Table S3.B13**.

**Method for computing transition values.**

- Transitions are computed from alleles in the position earlier in the haplotype block to alleles later in the block.
- In "simple-case" the transition is run only between two consecutive alleles starting at the beginning of the blocks. However, in the "complex-case" the transition values are calculated from each allele in earlier part of the block to each allele in later part of the block.

**Compute the maximum likelihood estimates for haplotype assignment**

- The likelihood estimates of each haplotype belonging to either population are maximized along the chain. Maximum likelihood for a Markov chain can be estimated as a "maximized product" of all the "p(emission) x p(transitions)" in the chain.
- Finally, $log_2^{Odds}$ ratio of the maximum likelihood estimates is taken to identify the possible haplotype assignment to the respective population.

**Estimation of maximum likelihood with examples**

**Example 3. Computing Maximum Likelihood For Just Two Sites**

If we consider just two positions (11 and 17), we can estimate the likelihood of haplotype assignment by computing "emission" and "transition" counts as shown in the tables below:

**Emission counts:**

- This emission count is computed for all the nucleotides (A, T, G, C) at all positions in the block. In Table S3.B4, only counts at positions 11 and 17 are shown.
- Add a default pseudo count of 0.25 to each allele (A, T, G, C) due to an unobserved allele in the path to avoid the maximum likelihood estimate turning out zero. If the alleles are observed, the observed count is added to the pseudo count. For example, since nucleotide "A" is not observed in population "A", at position 11, a pseudo count (i.e., count of 0.25) is added to the list of alleles.

**Table S3.B2 Emission counts of each possible nucleotide at position 11 and 17 in population "A"**

| Pop A↓/ nucleotide | A | T | G | C | Total |
|---|---|---|---|---|---|
| pos11 | 0 | 1 | 1 | 10 | 12 |
| pos17 | 0 | 0 | 11 | 1 | 12 |

**Table S3.B3 Emission counts of each possible nucleotide at position 11 and 17 in population "B"**

| Pop B↓/nucleotide ⟶ | A | T | G | C | Total |
|---|---|---|---|---|---|
| pos11 | 10.25 | 1.25 | 0.25 | 1.25 | 13 |
| pos17 | 1.25 | 10.25 | 1.25 | 0.25 | 13 |

**Table S3.B4 Emission counts of each nucleotide for population "A" after adding pseudo counts (0.25 count per allele)**

| Pop B↓/ nucleotide → | A | T | G | C | Total |
|---|---|---|---|---|---|
| pos11 | 10 | 1 | 0 | 1 | 12 |
| pos17 | 1 | 10 | 1 | 0 | 12 |

**Table S3.B5 Emission counts of each nucleotide for population "B" after adding pseudo counts (0.25 count per allele)**

| Pop A↓/nucleotide ⟶ | A | T | G | C | Total |
|---|---|---|---|---|---|
| pos11 | 0.25 | 1.25 | 1.25 | 10.25 | 13 |
| pos17 | 0.25 | 0.25 | 11.25 | 1.25 | 13 |

Step 01-B: Convert emission count to emission probabilities.

- This emission probability is computed based on counts adjusted with pseudo counts (i.e., Table S3.B3 and Table S3.B4).

**Table S3.B6 Emission probabilities of each possible nucleotide in population "A"**

| Pop A↓/nucleotide⟶ | A | T | G | C | Total |
|---|---|---|---|---|---|
| pos11 | 0.25/13 | 1.25/13 | 1.25/13 | 10.25/13 | 1 |
| pos17 | 0.25/13 | 0.25/13 | 11.25/12 | 1.25/13 | 1 |

**Table S3.B7 Emission probabilities of each possible nucleotide in population "B"**

| Pop B↓/nucleotide⟶ | A | T | G | C | Total |
|---|---|---|---|---|---|
| pos11 | 10.25/13 | 1.25/13 | 0.25/13 | 1.25/13 | 1 |
| pos17 | 1.25/13 | 10.25/13 | 1.25/13 | 0.25/13 | 1 |

Step 02-A: Compute the transition count.

- The transition matrix is prepared from nucleotides (A, T, G, C) at each earlier position to nucleotides (A, T, G, C) at each of the later positions. For example, in **Table S3.B8 and Table S3.B9**, only transition from alleles at positions 11 to 17 is shown.
- If "PI" values match between two blocks in a sample, the observed transition is counted as 1, else 0.5, because non-matching "PI" indicates that all possible configurations are likely.
- To avoid the maximum likelihood estimate turning out zero (0) due to unobserved transition in the path, a default pseudo count of 1/16 is attributed to each transition from {A, T, G, C} to {A, T, G, C}. If the alleles are observed, the observed count is added to

the pseudo count, e.g., a pseudo transition count (i.e., count of 1/16) is added to all other observed and unobserved transitions from nucleotide at position 11 to 17 in **Table S3.B10 and Table S3.B11**.

**Transition counts and probabilities for nucleotides from position 11 to 17**

**Table S3.B8 Representation of transition count in population "A"**

| from↓ | A | T | G | C | total |
|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 0 | 0 |
| T | 0 | 0 | 0 | 1 | 1 |
| G | 0 | 0 | 1 | 0 | 1 |
| C | 0 | 0 | 10 | 0 | 10 |

**Table S3.B9 Representation of transition count in population "B"**

| from↓ | A | T | G | C | total |
|---|---|---|---|---|---|
| A | 1 | 8 | 1 | 0 | 10 |
| T | 0 | 1 | 0 | 0 | 1 |
| G | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 1 | 0 | 0 | 1 |

*Note:* Transition counts and probabilities for nucleotides from positions 11 to 17 in both the populations after accounting for pseudo transition counts.

**Table S3.B10 Representation of transition counts (position 11 to 17) in population "A" after adding a pseudo count of 1/16.**

| From↓ / to -> | A | T | G | C | total |
|---|---|---|---|---|---|
| A | 1/16 | 1/16 | 1/16 | 1/16 | 0.25 |
| T | 1/16 | 1/16 | 1/16 | 17/16 | 1.25 |
| G | 1/16 | 1/16 | 17/16 | 1/16 | 1.25 |
| C | 1/16 | 1/16 | 161/16 | 1/16 | 10.25 |

**Table S3.B11 Representation of transition counts (position 11 to 17) in population "B" after adding a pseudo count of 1/16**

| from↓ / to -> | A | T | G | C | total |
|---|---|---|---|---|---|
| A | 17/16 | 129/16 | 17/16 | 1/16 | 10.25 |
| T | 1/16 | 17/16 | 1/16 | 1/16 | 1.25 |
| G | 1/16 | 1/16 | 1/16 | 1/16 | 0.25 |
| C | 1/16 | 17/16 | 1/16 | 1/16 | 1.25 |

Step 02-B: Compute the transition probabilities.

- Transition probabilities are computed based on counts adjusted with pseudo counts (i.e., Table B10 and Table B11).

**Table S3.B12 Representation of transition matrix(probabilities) in population "A" (position 11 to 17)**

| from↓ | A | T | G | C | total |
|---|---|---|---|---|---|
| A | 0.25 | 0.25 | 0.25 | 0.25 | 1 |
| T | 0.05 | 0.05 | 0.05 | 0.85 | 1 |
| G | 0.05 | 0.05 | 0.85 | 0.05 | 1 |
| C | 0.006 | 0.006 | 0.98 | 0.006 | 1 |

**Table S3.B13 Representation of transition matrix(probabilities) in population "B" (position 11 to 17)**

| from↓ | A | T | G | C | total |
|---|---|---|---|---|---|
| A | 0.103 | 0.787 | 0.103 | 0.006 | 1 |
| T | 0.05 | 0.85 | 0.05 | 0.05 | 1 |
| G | 0.25 | 0.25 | 0.25 | 0.25 | 1 |
| C | 0.05 | 0.85 | 0.05 | 0.05 | 1 |

Emission and transition probabilities for a 4-state Markov process can be represented with figures below:

**Figure S3.B4 "Emission probabilities" of nucleotides at position 11 for Pop A**

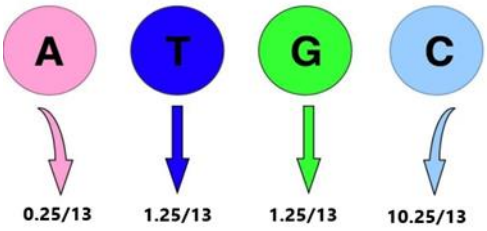



**Figure S3.B5 "Transition probabilities" of nucleotides from position 11 to position 17 for Pop A before applying pseudo count**

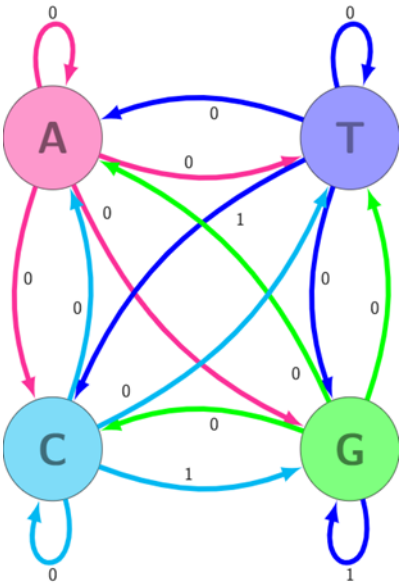
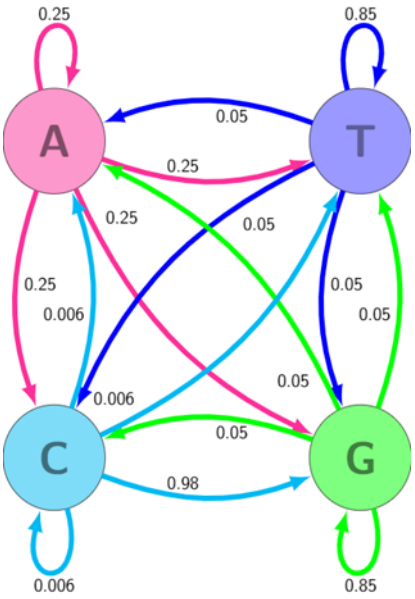


**Figure S3.B6 "Transition probabilities" of nucleotides from position 11 to position 17 for Pop A after applying pseudo count**




*NOTE:* Maximum likelihood estimation using alleles at sites 11 and 17 only

**Table S3.B14 For just two positions (11 & 17), we can estimate the likelihood as**

ml (lh is pop "a"):
A|C
T|G

$$P_{L_{Pa}} = P(A) \times P(AtT)$$
$$= \frac{0.25}{13} \times \frac{1}{16 * 0.25}$$
$$= 0.0048$$

ml (lh is pop "b"):
A|C
T|G

$$P_{L_{Pb}} = P(A) \times P(AtT)$$
$$= \frac{10.25}{13} \times \frac{129}{16 \times 10.25}$$
$$= 0.62$$

ml (rh is pop "a"):
A|C
T|G

$$P_{R_{Pa}} = P(C) \times P(CtG)$$
$$= \frac{10.25}{13} \times \frac{161}{16 \times 10.25}$$
$$= 0.77$$

ml (rh is pop "b"):
A|C
T|G

$$P_{R_{Pb}} = P(C) \times P(CtG)$$
$$= \frac{1.25}{13} \times \frac{1}{16 \times 1.25}$$
$$= 0.0048$$

*Note:* "$AtT$" → represents "A" to "T" transition

**Estimate Likelihood ratio(R):**

$$Odds_L = \frac{P_{L_{Pa}}}{P_{L_{Pb}}} = \frac{.0048}{.62} = .0077$$

$$Odds_R = \frac{P_{R_{Pa}}}{P_{R_{Pb}}} = \frac{0.77}{0.0048} = 160.41$$

$$R = \frac{Odds_L}{Odds_R} = \frac{.0077}{160.41} = .000048$$

Therefore, there is ample evidence that the left haplotype belongs to population "B", and right haplotype belongs to population "A".

**Example 4. Maximum Likelihood Estimation Using Alleles From The Whole Block.**

**Maximum likelihood estimates for left haplotype.**

$$Odds_L = \frac{P_{L_{Pa}}}{P_{L_{Pb}}}$$

$$P_{L_{Pa}} = P(A)\times P(AtT)\times P(A)\times P(AtG)\times P(A)\times P(AtC)\times P(T)\times P(TtG)\times P(T)\times P(TtC)\times P(G)\times P(GtC)$$

$$= \frac{0.25}{13}\times\frac{1}{16\times 0.25}\times\frac{0.25}{13}\times\frac{1}{16\times 0.25}\times\frac{0.25}{13}\times\frac{1}{16\times 0.25}\times\frac{0.25}{13}\times\frac{1}{16\times 0.25}\times\frac{0.25}{13}\times\frac{1}{16\times 0.25}\times\frac{1.25}{13}\times\frac{1}{16\times 1.25}$$

$$= 1.2348E-14$$

$$P_{L_{Pb}} = P(A)\times P(AtT)\times P(A)\times P(AtG)\times P(A)\times P(AtC)\times P(T)\times P(TtG)\times P(T)\times P(TtC)\times P(G)\times P(GtC)$$

$$= \frac{10.25}{13}\times\frac{129}{16\times 10.25}\times\frac{10.25}{13}\times\frac{145}{16\times 10.25}\times\frac{10.25}{13}\times\frac{113}{16\times 10.25}\times\frac{10.25}{13}\times\frac{145}{16\times 10.25}\times\frac{10.25}{13}\times\frac{113}{16\times 10.25}\times\frac{12.25}{13}\times\frac{129}{16\times 12.25}$$

$$= 0.055$$

$$Odds_L = \frac{P_{L_{Pa}}}{P_{L_{Pb}}} = \frac{1.2348E-14}{0.055} = 2.2452E-13$$

**Maximum likelihood estimates for right haplotype**

$$Odds_R = \frac{P_{R_{Pa}}}{P_{R_{Pb}}}$$

$$P_{R_{Pa}} = P(C)\times P(CtG)\times P(C)\times P(CtT)\times P(C)\times P(CtA)\times P(G)\times P(GtT)\times P(G)\times P(GtA)\times P(T)\times P(TtA)$$

$$= \frac{10.25}{13}\times\frac{161}{16\times10.25}\times\frac{10.25}{13}\times\frac{129}{16\times10.25}\times\frac{10.25}{13}\times\frac{161}{16\times10.25}\times\frac{10.25}{13}\times\frac{129}{16\times10.25}\times\frac{11.25}{13}\times\frac{177}{16\times11.25}\times\frac{9.25}{13}\times\frac{145}{16\times9.25}$$

$$= 0.1367$$

$$P_{R_{Pa}} = P(C)\times P(CtG)\times P(C)\times P(CtT)\times P(C)\times P(CtA)\times P(G)\times P(GtT)\times P(G)\times P(GtA)\times P(T)\times P(TtA)$$

$$= \frac{1.25}{13}\times\frac{1}{16\times1.25}\times\frac{1.25}{13}\times\frac{1}{16\times1.25}\times\frac{1.25}{13}\times\frac{1}{16\times1.25}\times\frac{1.25}{13}\times\frac{1}{16\times1.25}\times\frac{1.25}{13}\times\frac{49}{16\times1.25}\times\frac{0.25}{13}\times\frac{1}{16\times0.25}$$

$$= 6.05084E-13$$

$$Odds_R = \frac{P_{R_{Pa}}}{P_{R_{Pb}}} = \frac{0.1367}{6.05084E-13} = 2.2593E+11$$

**Likelihood estimates of left vs. right haplotype**

$$R = \frac{Odds_L}{Odds_R} = \frac{2.2452E-13}{2.2593E+11} = 9.9371084E-25$$

Therefore, there is ample evidence that the left haplotype belongs to population A and the right haplotype belongs to population B.

**Table S3.B15 Output Data from phase-Stitcher**

| f1/hybrid | popA\|popB |
|---|---|
| A\|C | A\|C |
| T\|G | T\|G |
| G\|T | G\|T |
| C\|A | C\|A |

**Supplementary Materials S3.C: ShortVariantPhaser in Detail**

**Example showing a detailed computation of ShortVariantPhaser using the haphedge data structure.**

**Overview and objective**

The titular Hap-Hedge algorithm aims to determine the most likely diploid (pair of haplotypes) using the method described in the article "Reference-based phasing using the haplotype Reference Consortium panel" (Loh, Danecek, et al., 2016) . This includes encoding input alleles given in the reference panel, running Positional Burrows-Wheeler Transform, and then determining the most likely diploid.

Here we describe the algorithm analytically using simple example data. We intend to give definitive clarity on the algorithm using this example data and provide explicit visualization of each of its steps. Here is the tabular data we will be using,

**Table S3.C1 Example Data**

| Site | Genotype 1 | Genotype 2 | Genotype 3 | Target Genotype |
|---|---|---|---|---|
| S_1 | A \| G | A \| G | A \| C | A/G |
| S_2 | T \| C | T \| C | C \| T | T/C |

*Note* the" Target Genotype" column in the table in Figure B1; haplotypes are separated using a forward slash,"/", instead of a bar,"|", to convey the target haplotypes are unphased. We may also include unphased haplotypes in the reference panel itself. But for now, the above table depicting three fully phased genotype sequences over two variant sites serves as our easy-to-use example.

**Encoding Alleles**

We encode alleles in a binary form (we'll use 0 s and $1s$ ) as per-site operation. Meaning we assign each of the alleles in the target genotype a binary which need not be consistent over variant sites, and this encoding extends to the rest of the panel. Using our example data in Figure 1 , the first and second sites may be encoded as,

$$S_1: A \to 0, G \to 1$$
$$S_2: T \to 0, C \to 1$$

Applying this transformation to the entire panel,

**Table S3.C2 Encoded Data**

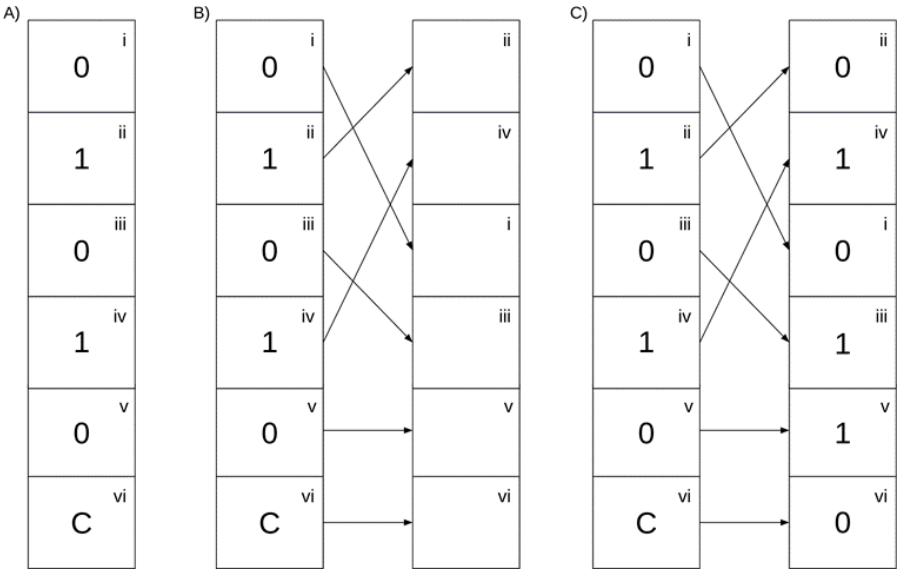
| Site | Genotype 1 | Genotype 2 | Genotype 3 | Target Genotype |
|---|---|---|---|---|
| S_1 | 0 \| 1 | 0 \| 1 | 0 \| $C$ | $0/1$ |
| S_2 | 0 \| 1 | 0 \| 1 | 1 \| 0 | $0/1$ |

We may note that the allele labeled initially as" C" has not changed in site S_1. Naturally, this is because it does not fit into our binary encoding; and to the Positional Burrows-Wheeler Transform, this can be interpreted as an" empty" or" none type" object less than both 0 and 1.

**Running Positional Burrows-Wheeler Transform**

The Positional Burrows-Wheeler Transform (PBWT) is a proper numerical technique for deriving information from strings. In our case, we have binary sequences comprising haplotypes. Moreover, the" positional" part of the PBWT implies that we could make good use of visualization:

**Figure S3.C1 Positional Burrows-Wheeler Transform on Encoded Data**




In the above figure, we take three steps. In step (A), we draw a column of boxes made distinct by the upper numeral, and we fill these boxes with the alleles from the first site. In step (B), we reorder the boxes by their filled object with the schema 1 > 0 > (anything else). Additionally, similar binaries are distinct. Meaning we can impose the ordering still preserves the relative ordering of those binaries; that is, if one 0 came before another 0, then that order is also preserved. Finally, step (C) is filling the second column of boxes (now rearranged) with the alleles from the second site. We can imagine repeating these steps over multiple sites; simply loop the instructions of filling the boxes with alleles and then reordering them.

Ultimately, the goal of this transformation is to produce transition matrices between these binaries. We will have one transition matrix because we only have the transitions from the two variant sites. However, for N sites, we would have $N - 1$ transition matrices. For our example data, we derive our transition matrix from part $(C)$,

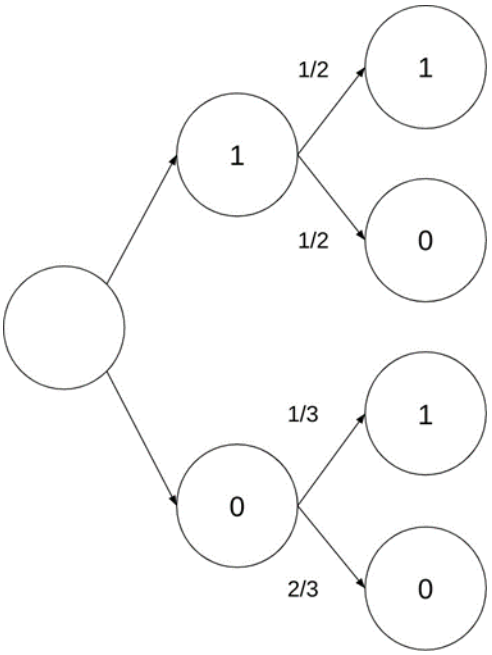
$$T = \begin{bmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{bmatrix}$$

In more detail, from the first site to the second site, we transition from 0 to 0 exactly 1/3 of the 3 transitions. Additionally, we transition from 1 to 0 once of the two transitions. Lastly, we simply disregard the" C." This ends the construction of the transition matrix from the PBWT.

**Calculating most-likely diploid**

The ultimate step in determining the most likely diploid, is constructing the HapHedge data structure. This is a rather simply step because we only have the one transition:

**Figure S3.C2 HapHedge Data Structure**




We should be keen to note why there are no transition probabilities between the very firstnode and the two incipient nodes: we know the two haplotype paths must be distinct, so one path must start at the node labeled "1" at the first site, and one path must start at the node labelled "0" at the first site. We effectively have two possible diploids,

**Table S3.C3 Possible diploid haplotype configurations with probabilities**

| Site | Diploid 1 | Diploid 2 |
|---|---|---|
| S_1 | 0 \| 1 | 0 \| 1 |
| S_2 | 0 \| 1 | 1 \| 0 |
| Probability | $(2/3)(1/2) = 1/3$ | $(1/3)(1/2) = 1/6$ |

And we arrive at the result that "Diploid 1" described in the table is our most likely diploid. Undoing our binary encoding, our phased target haplotypes, because of this algorithm, are (A, T) and (G, C) as one might readily interpret from the panel by inspection.

# CHAPTER IV: ALLELE-SPECIFIC EXPRESSION OF CANDIDATE GENES FROM LG2 QTL IN *ARABIDOPSIS LYRATA* USING F1 HYBRIDS


## Abstract

Allele-specific expression (ASE), also called allelic imbalance in gene expression, is a biological process in which a heterozygous locus transcribes unequal levels of transcripts from the two alleles. Analysis of ASE helps identify allele and gene expression basis of variation in a phenotype caused by cis-regulatory variation.

This chapter tests variation in ASE (allele-specific expression) of alleles between the Mayodan and Spiterstulen genome in F1 hybrids, specifically in the LG2 QTL, using RNA sequence data. Specifically, we tested ASE in a few candidate genes (*PIN1*, *PIN3*, *PILS2*, *BRC2*), which lie in LG2 (Chromosome 2) of the *Arabidopsis lyrata* reference genome. We picked these specific candidates based on a previous study conducted in *A. lyrata* populations (Mayodan and Spiterstulen) (Leinonen et al., 2013; Remington et al., 2013), which indicated that the genetic basis of adaptive variation in the life history of these two populations is in LG2 QTL.

The auxin efflux transport carrier *PIN3* showed statistically significant ASE (*P*-value=1.053203e-07) and high expression of My allele followed by *PILS2* (*P*-value=0.00062). We also did a global ASE assessment and found other interesting genes in the QTL regions; *AL2G30710* shows higher expression of the Mayodan allele, and *AL2G27860* shows higher expression of the Spiterstulen allele and are involved in phytohormone response. Functionally, *AL2G30710* is an ethylene-responsive transcription factor, and *AL2G27860* is a positive regulator of cytokinin-mediated development, which could affect plant developmental variation or work together with *PIN3* to mediate developmental variation.

## Introduction

Allele-specific expression (ASE), also called allelic imbalance in gene expression, is a phenomenon where a single gene or locus in a heterozygous individual transcribes unequal levels of transcripts from the two alleles. Due to cis-regulatory variation, maternal and paternal alleles are unequally regulated and expressed in an allele-specific manner. ASE can be identified by quantifying the difference between RNA expression of the haplotypes, particularly at heterozygous loci.

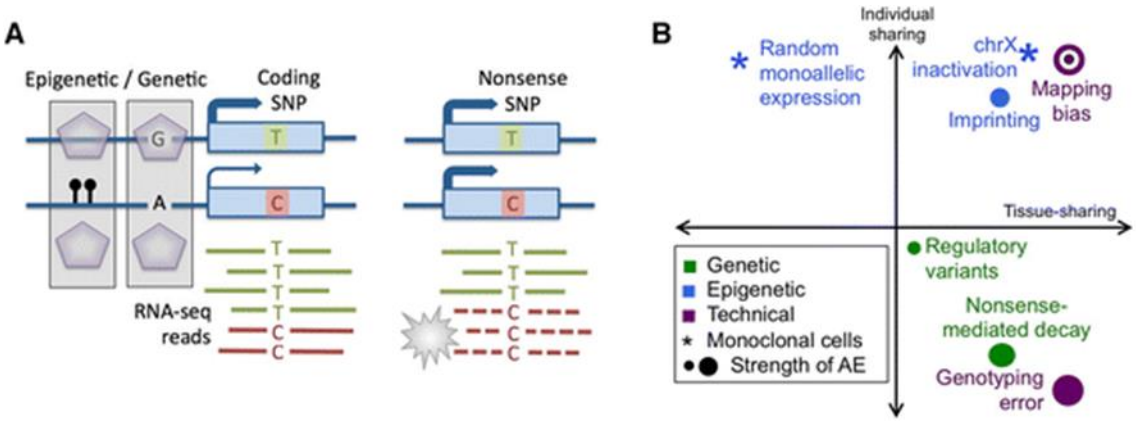
ASE could happen due to changes in the mRNA stability, differential transcription rate, and other processes that interfere with transcript abundance, like mutations that affect the binding of regulatory elements. The magnitude of allelic differences varies from purely monoallelic expression to subtle quantitative effects. The factors that regulate ASE are associated with genes and act in *cis* to change the transcript quantity. These cis-factors could be epigenetic marks or genetic variants that identify and interact unequally with paternal and maternal alleles. In contrast, the trans-acting factors would influence the expression of both parental alleles equally in the absence of cis-acting variation. Therefore, the ASE assessment plays a significant role in identifying a disease or phenotypic state caused by the expression levels of specific alleles (Raghupathy et al., 2018; Wittkopp et al., 2004).

### Importance of ASE Analyses

The integration of transcriptome and genome data has been widely used to understand genome function. For example, ASE, which integrates the allelic variation information with the level of expression of the alleles, helps assess and identify biological processes like nonsense-mediated decay (see Figure 4.1) (Castel et al., 2015). In addition, ASE helps identify maternal vs. paternal sources of genetic variation and uses that information on genetic variants (alleles) to

quantify variations in the level of gene expression, which could be a potential basis for variation in a particular phenotype.

**Figure 4.1 Allelic Expression With Its Sources**




**A**. Illustration of ASE. **B.** Biological sources of AE; the x-axis denotes the approximate sharing of AE across tissues of an individual. The y-axis shows the estimated sharing of AE signal in one tissue across different individuals. Source: (Castel et al., 2015).

**Allelic Variation for the Genetic Basis of Phenotype**

One way to identify gene-phenotype associations is GWAS, which has become popular with human genomics and a few other organisms where a plethora of genomics data is available. However, the GWAS study has some limitations or shortcomings. It requires large amounts of data from many samples, and the analyses must be genome-wide. The genotype-phenotype associations established through these methods are mostly not causal and need further testing and verification. Also, it does not provide a strong case for establishing the gene expression to phenotype basis of adaptation. ASE is a robust approach for identifying the gene-expression basis of phenotypic differences, as two alleles are exposed to the same trans-factors but different cis-factors. It also helps dissect the allelic and expression level basis of the association between genotype and phenotype.

## ASE Identification and Quantification

Genomic techniques like gene expression profiling or eQTL (expression QTL) analysis can help assess allelic differences. These approaches also help identify regulatory variations in cis- and trans-acting factors (Keurentjes et al., 2007; Kirst et al., 2005; Schadt et al., 2003) but cannot directly distinguish between cis- and trans-regulatory causes of expression variation. ASE analysis, by contrast, analyzes gene expression differences between two alleles caused by epigenetic and cis-acting regulatory variations.

ASE can be measured using sequencing methods that can help identify two copies of the transcripts or the alleles. For example, one of the methods for ASE analyses uses RNA sequencing technologies (RNAseq), which can help identify the allelic origin of some or most of the reads using SNP or INDEL markers (Castel et al., 2015; Lalonde et al., 2011). In addition, the RNAseq method helps further with the quantitative estimate of expression (i.e., the counts of the sequence reads).

ASE using RNAseq, therefore, provides a unique opportunity to understand the molecular mechanisms of metabolic disorders and complex health conditions (Lister et al., 2008; Nagalakshmi et al., 2008). However, there are some limitations to this approach. First, the data from RNA-seq alone does not provide enough information required to identify the origin of the transcript, except pointing out that they are different; i.e., it needs to be clear if a particular allele came from the paternal or maternal chromosome. Plus, quantifying ASE variation is another challenge because there is a chance of systemic biases in alignment due to the allelic differences (Degner et al., 2009). The alignment bias originates from the fact that we are trying to quantify variation in the counts of the two alleles arising from two homologs but by aligning reads to a single/haploid reference genome. In such a situation, a particular allele or haplotype (either

maternal or paternal) could be more similar to the reference, in which case the more similar allele will align at a higher frequency than the other allele giving a false estimate of ASE (Degner et al., 2009; Munger et al., 2014).

There are a few approaches to addressing the biased ASE estimate:

- Align DNA sequence from the hybrids to set a baseline for alignment bias and use that as an expected ratio of alignments against the observed alignment of the RNA sequence data.
- The ASE regulatory variation can also be inferred by measuring the ratio of the transcripts obtained from two different alleles generated by comparing the ratios calculated in parental mix RNAs and F1 hybrid (Wittkopp et al., 2004). The alleles present in the F1 hybrid sample possess similar regulatory factors as they are located in the same nucleus. Hence, the hybrid genes with cis-acting regulatory variation will show a biased ASE ratio if there is a bias in expression between two parental alleles. However, the genes possessing trans-acting regulatory variation will exhibit different ASE ratios in parental mixes than the F1 hybrid.
- Prepare a diploid genome by integrating known phased variants onto the reference genome representative of variants that the F1 contains and create a personalized diploid genome (Munger et al., 2014; Raghupathy et al., 2018; Rozowsky et al., 2011). Then competitively align RNAseq reads from the F1s against the diploid genome.

**Aims and Rationale**

This study's motivation is to identify the genetic basis of resource allocation tradeoffs using divergent populations (Mayodan and Spiterstulen) from outcrossing and perennial *Arabidopsis lyrata*. I will refer to Mayodan as My and Spiterstulen as Sp in the rest of the

chapter. *A. lyrata* is a close relative of *Arabidopsis thaliana,* a model organism used for research in plant genetics. As an outcrossing model, *A. lyrata* provides a better opportunity to identify novel alleles, QTLs, and genotypes and understand the genetic basis of adaptation in different environments. In addition, its relationship to *A. thaliana* offers a rich resource of genetic methods, materials, and databases for functional analyses.

Studies by Leinonen (2011) showed that two study populations of *A. lyrata* exhibit local adaptation and strong life-history differences. Further analyses of variation in local adaption by Leinonen (2013) showed that the most significant effect of QTL for resource allocation tradeoffs and adaptation in a cross between these phenotypically divergent populations was on chromosome 2 (LG2). In the same study (Leinonen et al., 2013) and Remington et al. (2013), using SEM (structural equation modeling) and an analysis of lateral shoots architecture by Remington et al. (2015) suggested that the early developmental differences between the two populations are causal to life-history tradeoffs during and after the reproductive season.

Following those findings, one of the main goals has been to test a few candidate genes that underlie the LG2 QTL and are central to the plants' developmental process. Previously, in Chapter 2, we explicitly tested those differences' chemical/hormonal basis in our two study populations. As a result, we obtained some evidence supporting a functional role for auxin transport, with Mayodan individuals showing a greater rate of auxin transport. This chapter aims to take this further by testing if those differences correlate with variation in the expression of genes involved in auxin homeostasis and transport. We are also interested in other genes in the LG2 QTL region with functions potentially relevant to shoot architecture development and potential contributors to life history differences between the two populations. A few candidate genes of interest located in LG2 QTL (region spanning 2:12,875,693 to 2:16,344,528, and

containing a total of 521 genes) are *PIN1*, *PIN3*, *PILS2*, and *BRC2*. *PIN1* and *PIN3* are essentially involved in auxin homeostasis and transport, and *PILS2* is mainly involved in auxin homeostasis and accumulation. At the same time, *BRC2* encodes a TCP transcription factor, a *tb1* ortholog found in maize (Doebley et al., 1995, 1997) that arrests axillary bud development and prevents axillary bud outgrowth.

Unlike in the auxin transport assay, where we tested for differences in transport responses in two populations, we use F1 hybrids to identify differentially expressed alleles. With ASE, the differences in genetic basis are more readily dissectable because we can distinguish the alleles between parents and quantify them, making it a robust method to identify population-level effects on gene expression and, hence, the variation in life-history traits. However, if the genetic basis for life-history variations is instead differences in protein sequences caused by non-synonymous substitutions or InDels and not the variation in the quantity of the gene expression, ASE would not be able to pinpoint the genetic basis of adaptive divergence in life-history traits.

Since individuals from Mayodan show greater apical dominance and a higher rate of auxin transport, we predict that the candidate genes *PIN1* and *PIN3* will show higher expression of Mayodan alleles because these genes are associated with promoting auxin transport. We also predict that the candidate gene *PILS2* will show higher expression of Mayodan alleles because higher auxin transport will need higher auxin homeostasis and accumulation. Furthermore, since *BRC2* expression arrests axillary bud outgrowth and Mayodan individuals exhibit lower lateral shoot development, we predicted *BRC2* expression to be higher for Mayodan alleles.

In addition, we also looked at the global ASE variation in several other genes contained in LG2 QTL. Again, we expected to see the higher expression of alleles that functionally promote the development of lateral vegetative shoots.

## Methods

### Experimental Design

The seeds for parents (Mayodan and Spiterstulen) were obtained as described in Chapter 2 and grown in the same manner in Chapter 2, Experiment #2. We established crosses between 3 unrelated Mayodan and Spiterstulen parents and obtained F1 seeds. The F1 hybrid seeds were sown as described in Chapter 2, Experiment #2, and allowed to grow until lateral vegetative shoots started developing. It was essential to select these F1s at this stage because this stage represents the early developmental time point that determines the fate of the meristems (whether the meristem should transition into a reproductive shoot or remain vegetative), providing the very basis of life-history tradeoffs from a developmental standpoint. We sampled twelve F1 hybrids for RNA extraction.

For RNA extraction, we removed the leafy tissue and the roots by clipping and retaining the whole shoot and lateral meristems for RNA extraction. Next, we extracted mRNA from the entire vegetative shoot tissue using MasterPure™ Plant RNA Purification Kit (Illumina - catalog # MPR09010) and tested for quality using a nanodrop spectrometer. Next, we selected the four samples with the highest-quality RNA representing two full sibs (2ms02g, 2ms03g) from the same family and two unrelated F1 (2ms01e, 2ms04h) for library preparation using the poly-A selection method to enrich coding transcripts and eliminate the rRNAs. The mRNA library was then sequenced at David H. Murdock Research Institute (DHMRI) at North Carolina Research Campus in Kannapolis sequencing center. The mRNAs from each F1 individual were individually barcoded and ran in a single lane on the Illumina Hiseq to generate100 bp paired-end read.

**Genomic Sequence Reads**

We obtained additional complete sets of genomic sequence data for both My and Sp populations from Outi Savolainen's lab based on previous research done by Mattila (Mattila et al., 2017), which were also 100 bp paired-end reads. Both genome and RNA sequence reads were quality tested using FastQC, and adapters, including low-quality reads at the end, were trimmed using Trimmomatic; see **Supplementary Materials S4.B AND S4.C** for more details.

**Variant Calling**

We aligned the RNAseq reads to the *A. lyrata* reference genome (T. T. Hu et al., 2011) using rnaSTAR (STAR Manual 2.7.10a, 2022), and called variants using GATK (Genome Analysis ToolKit) (Van der Auwera et al., 2013). In addition, genome sequence reads were aligned to the same reference genome using BWA (Burrows-Wheeler Aligner) (Li et al., 2008; Li & Durbin, 2009), and variants were called using GATK (Van der Auwera et al., 2013), and multisample VCFs was generated. Specific details and parameters from alignment to variant calling are available in **Supplementary Materials S4.B** for genome sequence data and in **Supplementary Materials S4.C** for RNA sequence data.

**Haplotype Phasing**

The variants called from both genome sequence and RNA sequence data were read-backed-phased using the application called Phaser (Castel et al., 2015, 2016). Phaser-produced RBP VCF files were parsed using VCF-Simplify (unpublished, https://github.com/everestial/VCF-Simplify) to prepare analyzable haplotype blocks. The haplotype files were further phased into extended haplotypes (for variants called from population genome data) using Phase-Extender (discussed in Chapter 3, unpublished). For haplotype phase extension, we used all the samples in the cohort plus the common allele of 24 A. lyrata genomes

(generated at the Max Planck Institute for Developmental Biology by Sang-Tae Kim and D. Weigel) (Arnold et al., 2015). Finally, RBP haplotypes from F1s were phased genome-wide using Phase-Stitcher (discussed in Chapter 3, unpublished); in this case, the haplotypes phased by Phase-Extender were used as population reference panels. Specific parameters and steps used in phasing are provided in step 6 of **Supplementary Materials S4.C**.

**Preparation of Custom Diploid Genome and GFF**

The initial step in quantifying ASE using RNA-seq data is to align the sequence reads to a transcriptome or genome and ensure unbiased alignment. The main challenge for accurate ASE estimation is to overcome the alignment bias. This can be solved by finding a haploid genome that equally represents both the parents or by establishing an individualized diploid genome sequence representing each parent; this diploid genome can be used for competitive alignment of the sequence reads.

In our case, we created a diploid individualized genome for each F1s. The variants prepared from F1 RNAseq reads and phased using Phase-Stitcher were used to prepare the variant template for the custom diploid genome. Not using the diploid genome would have increased ASE bias because Mayodan reads are closer to the *A. lyrata* reference genome (T. T. Hu et al., 2011), which was prepared from a single North American genotype (strain MN47 from Michigan, USA). Another reason was that we did not have any genome sequence from F1s or genome or RNA sequence from each parent, complicating the elimination of biases. We then patched the template variants (phased VCFs) prepared for each F1 with homozygous variants prepared from the population genome sequence data (Mattila et al., 2017) using python scripts built in-house and created a final and personalized diploid phased variants template. Finally, this personalized phased variant was patched onto the *A. lyrata* reference genome (T. T. Hu et al.,

2011) using g2gtools (Munger et al., 2014) to create a final personalized and diploid version of the genome and GFF (Gene Feature Format) file for each F1 sample. These steps and parameters are documented in step 6 of **Supplementary Materials S4.B** and made available on GitHub as ASE-CADG (unpublished, Allele-specific expression – using Competitive Alignment on Diploid Genome, https://github.com/everestial/ASE-CADG).

**Alignment of RNAseq to Custom Diploid Genome**

We first aligned the RNAseq reads to the haploid *A. lyrata* reference genome (T. T. Hu et al., 2011) on all the scaffolds using rnaSTAR 2.5.4b (STAR Manual 2.7.10a, 2022). Then we selected reads that aligned to haploid LG2 (or Chromosome 2) for alignment to the personalized diploid genome. Finally, we aligned the selected RNAseq reads to the custom diploid genome prepared for each F1 sample using rnaSTAR 2.5.4b (STAR Manual 2.7.10a, 2022) using a 2-pass alignment protocol. The alignment was limited to only LG2. The percentage of unique alignments for each F1s was > 40% **Supplementary Materials S4.**, Table S4.C1). So, the scope of the result presented in this chapter is limited to what was observed in the LG2 only. Subsequent mention of the global data and observation refers to globally at the level of LG2.

We set alignment parameters to only select and score the best alignment to an individual haplotype; however, it would be possible to have the same score across both haplotypes at a single location if scores were the same. Finally, we parsed the SAM files containing the reads aligned to personalized diploid genomes using custom python scripts developed in-house to count and prepare the following metrics for each gene for each sample:

1. unq_C_My – number of unique reads that aligned to Mayodan haplotype.
2. unq_C_Sp – number of unique reads that aligned to Spiterstulen haplotype.
3. bi_C_My – number of reads that aligned equally to Mayodan and Spiterstulen haplotype.

4. bi_C_Sp – number of reads that aligned equally to Mayodan and Spiterstulen haplotype. The value for this variable is the same as bi_C_My.
5. mul_C_My – number of reads aligned to Mayodan haplotype and aligned to more than two other regions in the genome.
6. mul_C_Sp – number of reads aligned to Spiterstulen haplotype and aligned to more than two different regions in the genome.

The integer value in the "NH" field of the SAM file (the one aligned to the genome) produced by rnaSTAR (STAR Manual 2.7.10a, 2022) was used to count the reads aligned to each haplotype.

Additionally, other counts were derived for our statistical analyses:

7. unq_C_sum = unq_C_My + unq_C_Sp
8. mul_C_sum = mul_C_My + mul_C_Sp
9. total_C_My = unq_C_My + bi_C_My + mul_C_My
10. total_C_Sp = unq_C_Sp + bi_C_Sp + mul_C_Sp
11. total_C_sum = total_C_My + total_C_Sp

Finally, we obtained sum aggregated measures for all samples. In the formulas below *j* refers to the specific sample and *J* refers to the overall number of samples.

12. unq_C_total = $\sum_{j=1}^{J}(unq\_\mathrm{C\_My}_j + unq\_\mathrm{C\_Sp}_j) = \sum_j^J unq\_\mathrm{C\_Sum}_j$
13. mul_C_total = $\sum_{j=1}^{J}(mul\_\mathrm{C\_My}_j + mul\_\mathrm{C\_Sp}_j) = \sum_j^J mul\_\mathrm{C\_Sum}_j$
14. total_C_total = $\sum_{j=1}^{J}(total\_\mathrm{C\_My}_j + total\_\mathrm{C\_Sp}_j) = \sum_j^J total\_\mathrm{C\_Sum}_j$

For data analyses, we selected genes that passed the following filters:

- Genes that have the expression of more than ten counts (either Mayodan or Spiterstulen) in at least one of the samples.

- Genes for which either mul_C_total does not exceed 20% of totalC_total or unqC_total exceeds 20% of totalC_total, i.e., either, unq_C_total > 0.2 * total_C_total , or, mul_C_total < 0.2 * total_C_total.

Filters are used to avoid a situation where one or both My and Sp expressions are zero (or very low) and leave expressed genes ratio very high or at infinity.

**Data Analyses**

We did statistical tests in R (version 4.0.5). We tried two approaches to check the difference in allele expression between Mayodan and Spiterstulen alleles. In the first approach, we made a basic comparison of ASE using exact binomial tests. In this approach, tests were done individually for each sample and each gene between the two alleles at 0.05, 0.01, and 0.001 levels of significance.

We used package DESeq2 (version 1.30.1) (Love et al., 2014) in the second approach and applied the Wald test for the test of significance. Although the DESeq2 package is not developed for the ASE analysis, it helps test the data generated for ASE. For ASE analyses, we treated unique reads count aligned to My haplotype as the first group and unique reads count aligned to Sp haplotype as the second group. Thus, we can fit the negative binomial model for each gene and perform ASE as the differential expression testing. The specification of the model is described in the following formula:

$$K_{ij} = NB(s_{ij}q_{ij}, \alpha_i)$$

Where $K_{ij}$ refers to the raw counts of gene $i$ for sample $j$ (in our case, samples are 2ms01e, 2ms02g, 2ms03g, 2ms04h), $s_{ij}$ is the mean counts, $q_{ij}$ is the normalization factor, and $\alpha_i$ is the dispersion for the gene.

After the model is fit, we estimated coefficients for each sample group and their standard error. The coefficients are the estimates for the $Log_2$FoldChanges for each sample group. Once the model is fitted with maximum likelihood, we apply the Wald test-based parameters that have been estimated by maximum likelihood. Finally, we test for the primary hypothesis:

$$H_0: Log2FoldChange = 0$$

**Preparation of Heatmap**

We assessed the similarity between samples by first applying regularized-logarithm (*rlog*) transformation to the data. We then estimate "Euclidean distances" between the samples and visualize sample distances with the heatmap.

"Euclidean distance" between samples $x$ and $y$ is estimated as,

$$dist_{x,y} = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

where $x_j$ and $y_j$ refer to the jth gene expression from samples $x$ and $y$, and $J$ indicates the total number of genes in the samples. We also estimate the PCA using the same *rlog* transformed data. PCA reduces the dimensionality of the data while retaining most of the variation in the data set, by identifying directions, called principal components, along which the variation in the data is maximal.

## Results

**Data After Filter**

Our data consisted of 4 sample datasets ("2ms01e", "2ms02g", "2ms03g" and "2ms04h"). Initially, each dataset contained 3017 unique genes (a total of 12068 observations). However, we processed the data via several filters we discussed earlier. After applying those filters, the sample

dataset was reduced to 1761 observations per dataset (a total of 7044 observations) for the whole LG2 (whole chromosome 2).

**General Expression Statistics**

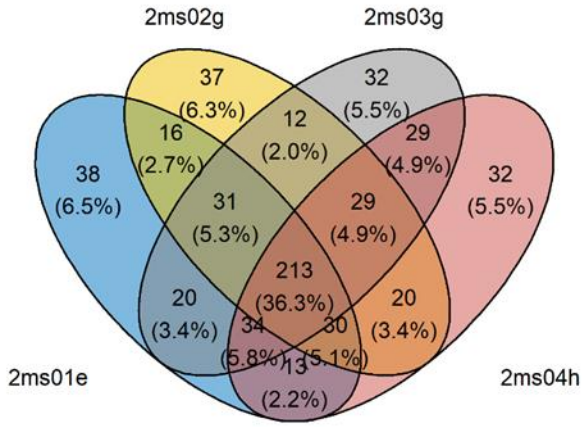
**Table 4.1 Summary Statistics of the Number of Genes (Second Column, N) and the Average Number of Reads (Third Column, Mean), IQR (Interquartile Range), Across All the Samples' Genes**

| Sample | N | Mean | Std. dev. | IQR | Min | %25 Q | %50 Q | %75 Q | Max |
|---|---|---|---|---|---|---|---|---|---|
| **My** | | | | | | | | | |
| 2ms01e | 1580 | 530.5861 | 1391.753 | 466.0 | 0 | 66 | 218 | 532.0 | 34392 |
| 2ms02g | 1552 | 576.0077 | 1324.961 | 530.5 | 0 | 74 | 230 | 604.5 | 28150 |
| 2ms03g | 1598 | 496.2178 | 1071.471 | 465.5 | 0 | 64 | 206 | 529.5 | 20864 |
| 2ms04h | 1595 | 571.8934 | 2302.589 | 482.0 | 0 | 66 | 214 | 548.0 | 76562 |
| **Sp** | | | | | | | | | |
| 2ms01e | 1580 | 541.3785 | 1862.259 | 450 | 0 | 64 | 202 | 514 | 57822 |
| 2ms02g | 1552 | 516.1456 | 1196.342 | 498 | 0 | 58 | 211 | 556 | 27098 |
| 2ms03g | 1598 | 473.7059 | 1100.764 | 435 | 0 | 60 | 181 | 495 | 24710 |
| 2ms04h | 1595 | 579.4683 | 2477.427 | 466 | 0 | 58 | 204 | 524 | 85362 |
| **unqC_Sum** | | | | | | | | | |
| 2ms01e | 1580 | 1071.9646 | 3078.293 | 925.0 | 16 | 146.0 | 430 | 1071.0 | 79846 |
| 2ms02g | 1552 | 1092.1534 | 2436.139 | 978.5 | 16 | 151.5 | 485 | 1130.0 | 55248 |
| 2ms03g | 1598 | 969.9237 | 2071.002 | 905.5 | 16 | 142.0 | 404 | 1047.5 | 45574 |
| 2ms04h | 1595 | 1151.3618 | 4702.487 | 926.0 | 16 | 134.0 | 436 | 1060.0 | 161924 |

*Note:* %25Q, %50Q and %75Q refer to the three quartile of number of reads.Some of the genes have high gene expression in all or several samples. For example, Figure 4.2 shows the number of genes in each sample with total gene expression for My and Sp (*unq_C_My* +

*unq_C_Sp*) higher than 75% quantile. Among these genes, 213 have a total expression in the 4th quartile in all samples.

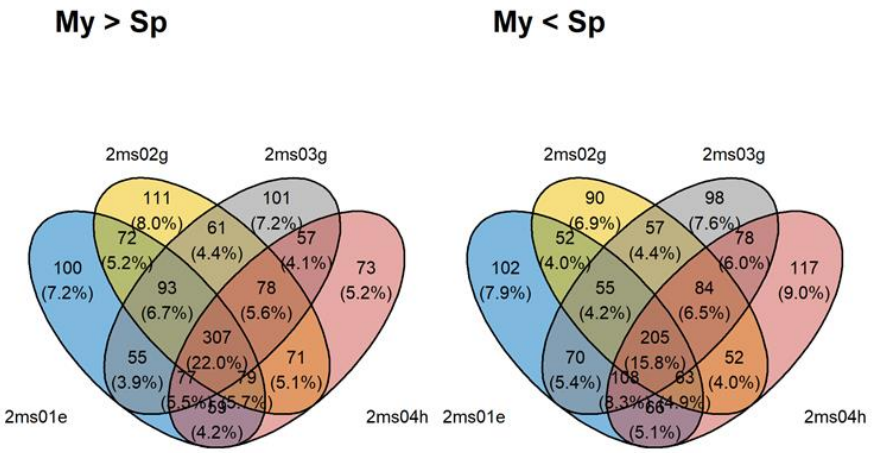
**Figure 4.2 Venn Plot of Number of Genes With Total Expression Greater than the 3rd quartile (> 3Q) of the Sample**




*Note:* The overlapping sets show the number of genes exhibiting the same (> 3Q) expression in the intersected samples.

The Venn diagrams (Figure 4.3) show that among all the genes, there are 512 (My > Sp: 307, Sp > My: 205) genes that have the same direction difference across all the samples. The total share of such genes is 37.8% among the total amount of genes. The following sections discuss genes that have statistically significant differences in gene expression between alleles (using Wald test) and are in the same direction.

**Figure 4.3 Venn Diagram Showing Number of Genes With the Same Directional Difference in Expression Across All Four Samples**




**Binomial Tests**

More than 50% of the genes showed significant ASE (*P*-value < 0.001) across all the samples in binomial tests, see Table 4.2, which seems unrealistic. For the rest of the analyses, we used the result from the Wald test provided by the DESeq2 package (Love et al., 2014).

**Table 4.2 Summary Statistics for the Number of Genes Showing Significant ASE (P-value < 0.001) Under Binomial Tests**

| Sample | $P-value < 0.001$ | $0.001 < P-value < 0.01$ | $0.01 < P-value < 0.05$ | $P-value > 0.05$ (not significant) |
|---|---|---|---|---|
| 2ms01e | 761 | 135 | 139 | 545 |
| 2ms02g | 880 | 111 | 111 | 450 |
| 2ms03g | 818 | 113 | 155 | 512 |
| 2ms04h | 855 | 122 | 129 | 489 |

*Note:* In this test, we applied one extra filter; we only analyzed genes with *unq_C_sum* > 16, which reduced the total number of observations to 6325.

**Heatmap and PCA**

The heatmap (Figure 4.4) and PCA (Figure 4.5) show the distances between samples and haplotypes. The heatmap is based on the rlog-transformed data and represents sample-to-sample

differences in gene expression. Distances between My groups are relatively lower than for Sp groups. For all My samples, the distance lies from 47.11 to 51.70, while the distance for Sp samples lies between 50.94 and 59.63. Thus, expression in Sp haplotypes seems more variable than expression in My haplotypes across samples.

Results from the PCA analysis support the conclusion obtained from the heatmap. We can see that the values for My vs. Sp are clustered together (i.e., values representing My expression are more closely grouped than those representing Sp expression across samples). This difference especially shows ups for the pairs of My-2ms03g – My-2ms04h and Sp-2ms03g – Sp-2ms04h.

**Figure 4.4 Heatmap of Sample Distances for All the Sample and Haplotype Pairs**

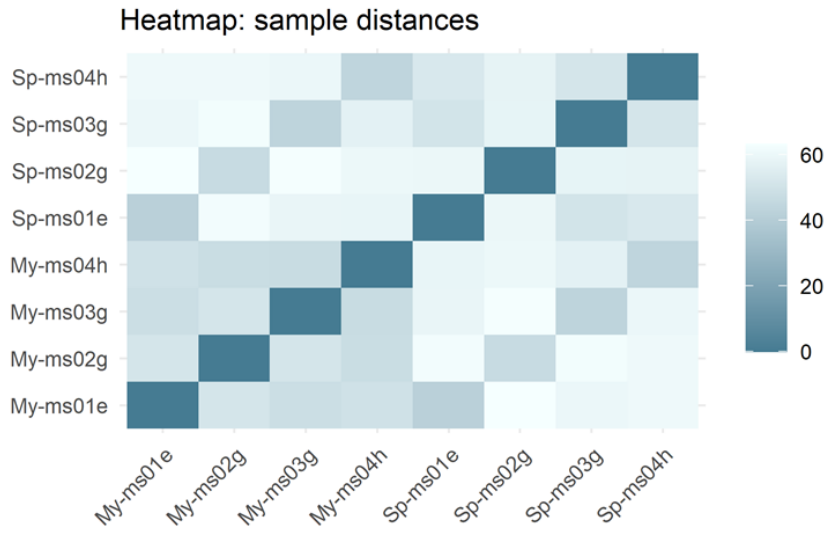
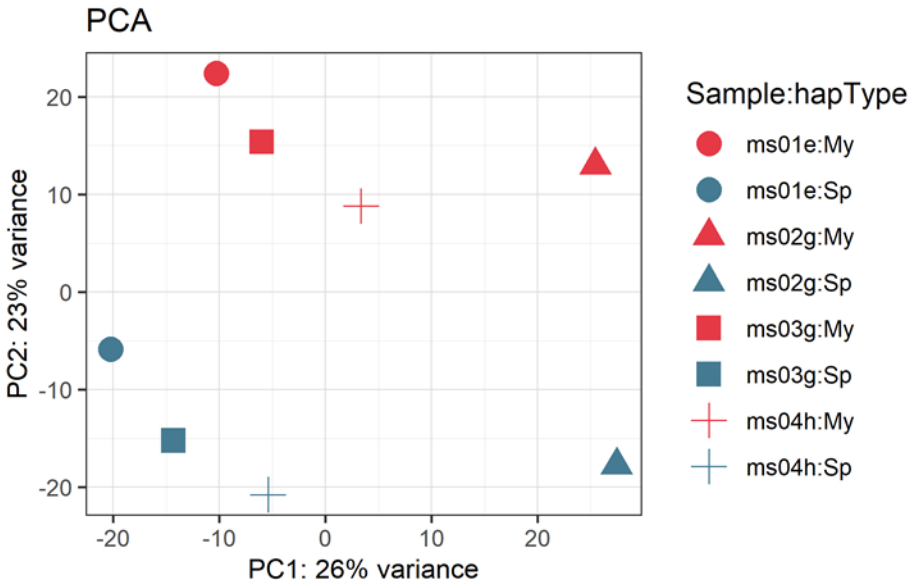


**Figure 4.5 PCA Plot for Samples Distance for All the Samples and Haplotype Pairs**




**Results from Wald Test**

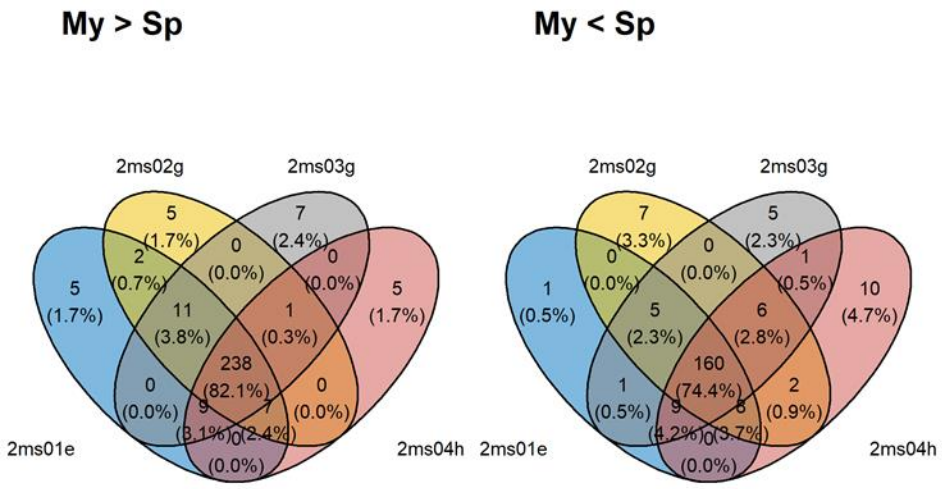
The Wald test was used to test the significance of differences between two groups (in our case, My and Sp). This test works for the whole dataset and gives an overall picture of the ASE and its significance across all samples. The results from the Wald test indicate that a relatively higher number of genes show ASE in favor of My alleles, compared to the Sp alleles, but still statistically not that significant. Overall, the number of genes that do not have significant differences (*P*-value > 0.05) is almost three times higher than those with significant differences, see Table 4.3. This estimate of ASE appears to be more realistic compared to the one provided by the Binomial test.

**Table 4.3 Number of Genes Showing ASE at Different Significance Levels Under Wald Test**

| *P*-value interval | $P - value < 0.001$ | $0.001 < P - value < 0.01$ | $0.01 < P - value < 0.05$ | $P - value > 0.05$ (not significant) |
|---|---|---|---|---|
| Number of genes | 196 | 90 | 173 | 1302 |

Of all the genes, 459 genes show a significant difference in ASE between My and Sp on the 5% confidence level. Among those, 268 genes show higher expression for the My allele than the Sp allele (238 genes with the same difference direction across all the samples, see Figure 4.6), and 191 genes show higher expression of the Sp allele (160 genes with the same difference direction across all samples, see Figure 4.6). This points to one of the common observations in ASE analyses, that the allele from a female parent usually shows higher expression than a male parent (Shao et al., 2019; Springer & Stupar, 2007), also known as parent-of-origin-effects, and all the F1s in our experiments have My cytoplasm (i.e., My is the mother for all the F1s). However, another possibility that the My alleles are showing higher ASE for most of the genes is that we first aligned the original RNAseq data to the haploid reference and then selected the reads aligned to LG2 to align it to the diploid genome. This could have created some bias because reads originating from the Mayodan haplotype could have higher alignment to the haploid reference because it is prepared based on a single North American strain. However, we still see a higher expression of Sp alleles for several genes, providing support that this bias must have been minimal. Moreover, there are many genes with both high Log2FoldChange and low *P*-value (see Figure 4.7) in both directions. A large number of genes with the insignificant difference is shown in Figure 4.8.

**Figure 4.6 Venn Diagram Showing the Number of Genes With Significant ASE (Based on Wald Test) in Either Direction (My > Sp, My < Sp) Across All the Samples**

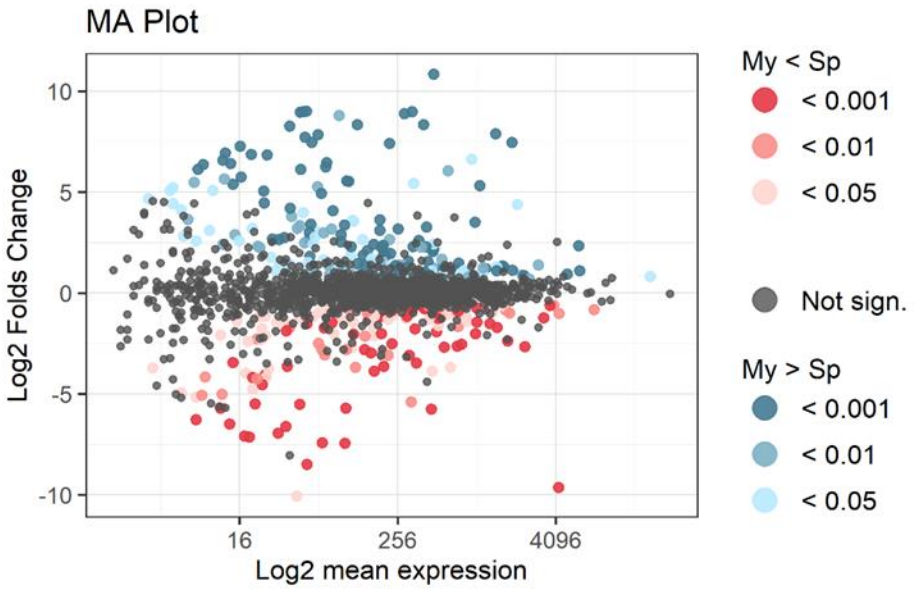



*Note:* 238 genes show significant (*P*-value < 0.05) ASE and have the expression of My > Sp in all the samples; 160 genes show significant (*P*-value < 0.05) ASE and have expression of My < Sp in all the samples.

**Figure 4.7 Volcano Plot Showing the Distribution of Log2FoldChange (X-axis) Against the -log10P-Value From the Wald Test**


*Note:* Blue dots show ASE in favor of Sp, and red dots show ASE in favor of My alleles with varying levels of significance justified by the intensity of the color.

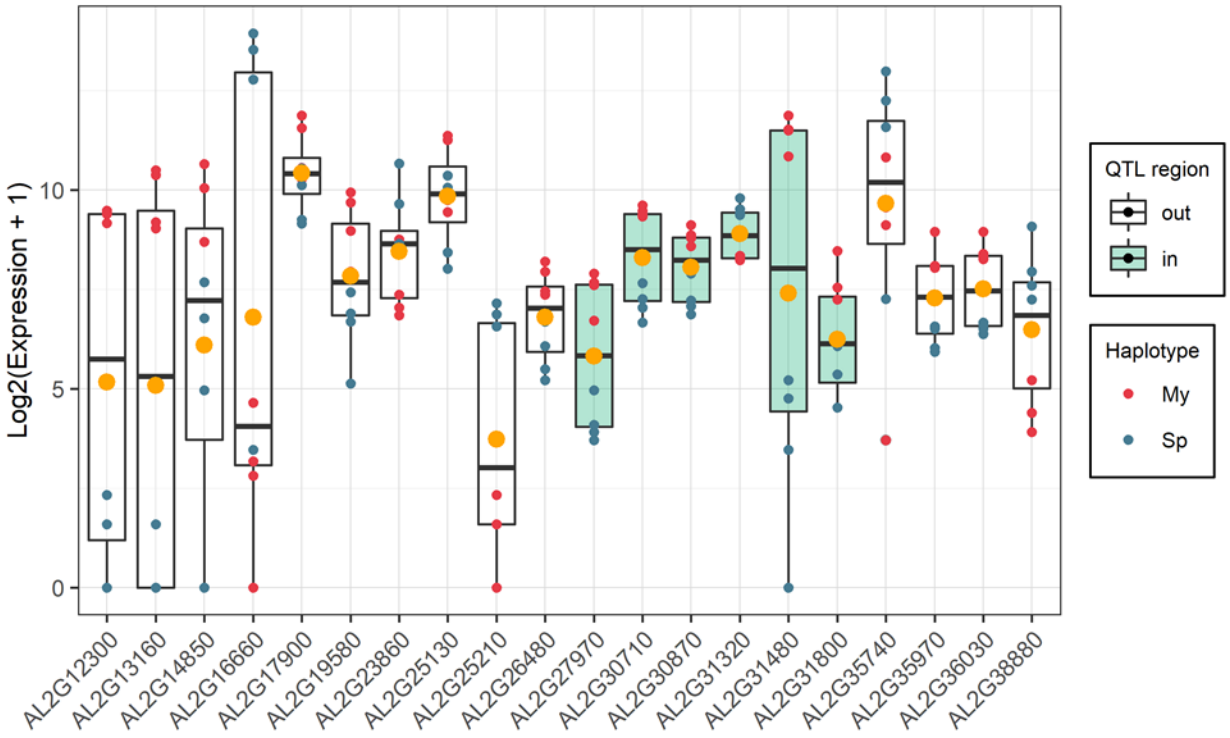
**Figure 4.8 MA Plot Showing Log2FoldChange (Y-axis) Against the Log2 of the Mean Expression Across Samples; P-values Are From the Wald Test**




*Note:* Blue dots show ASE in favor of Sp alleles, and red dots show ASE in favor of My alleles with varying levels of significance justified by the intensity of the color.

The data (Table 4.4) and boxplot (Figure 4.9) represent the twenty genes with the lowest adjusted *P*-values (False Discovery Rate adjusted). Some of the genes have a high expression difference due to the lower base of the opponent allele. For example, the gene with ID *AL2G12300* and *AL2G13160* have a mean of Sp allele expression less than 2, while gene *AL2G25210* has My allele expression equal to two (see Table 4.4). However, most other genes have both alleles expressed at higher levels. There are six genes from the QTL region among the top 20 reported by the Wald test *P*-value (see boxplots shaded in light green in Figure 4.9). The coordinates for resource allocation QTL in LG2 start at 2:12,875,693 (with gene FKF1, also called ADO3, at the left end of the QTL region) and ends at 2:16,344,528.

**Figure 4.9 According to the Wald Test, the Top 20 Genes With Significant ASE**

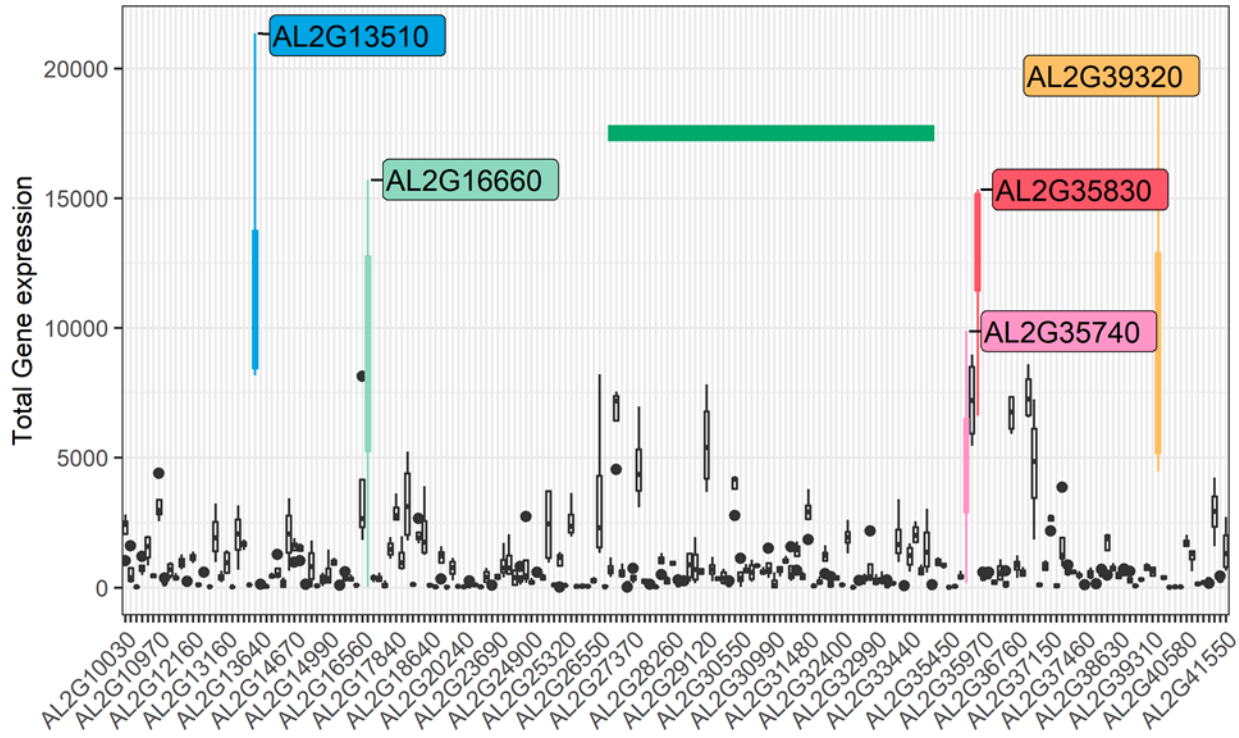



*Note:* Raw counts of total gene expression of My and Sp haplotypes are used for all of these plots. For better visualization, expression levels are transformed with log2(x +1) where x is the number of raw counts. Genes on the x-axis are ordered by genomic position (beginning to end). The bars shaded in green are loci within the QTL region.

The boxplot (Figure 4.10) represents genes with ASE at a *P*-value < 0.001 level of significance. The genes are arranged by their genomic position (beginning to end). Six genes are highlighted, with colored gene IDs showing comparatively high expression. We also provide the same plots for genes with different significance criteria (see, **FigureS4.A1** and **FigureS4.A2**). These criteria depend on both the magnitude of log2FoldChange and the *P*-values. These criteria are:

1. Log2FoldChange > 2 & *P*-value < 0.05 (104 genes) ; see **FigureS4.A1**.
2. Log2FoldChange > 4 & *P*-value < 0.01 (62 genes); see **FigureS4.A2**.

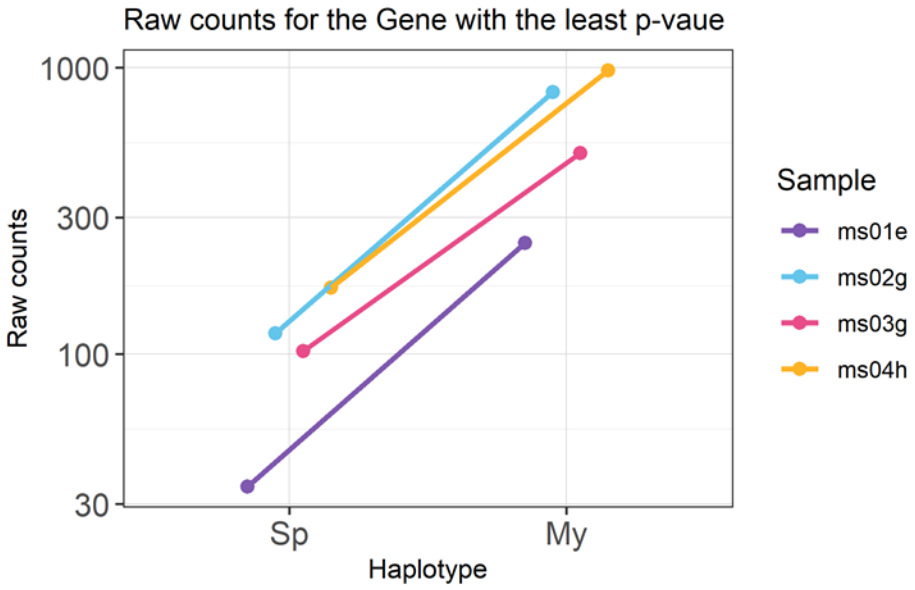
**Figure 4.10 Genes Showing ASE With Wald Test P-value < 0.001**




*Note:* The green shaded horizontal bar spans the QTL region. There is no space for all gene IDs' name, so only some are shown on the X-axis. The box plot for each gene is based on raw counts of My and Sp alleles from each sample. Unlike in Figure 4.9, a logarithm with a pseudo count of 1 is not added for generating this plot. Figure 4.9 showed 20 genes with the lowest *P*-values. Some of the genes in this plot can have high expression and low *P*-value but are still not as low as the ones mentioned in Figure 4.9, e.g., AL2G13510 has a *P*-value = 9.10E-09, which is still not enough to get into top 20. The highest *P*-value among the top 20 was 3.19E-14.

*AL2G19580* gene encodes an RNAse H-like protein and showed the most significant differences in ASE in our study, with My showing strong expression in all the samples; *P*-value = 2.726e-47, see Figure 4.11. As expected, the difference is rather severe in each sample.

**Figure 4.11 Raw Counts of My and Sp Alleles for the Gene (AL2G19580) With the Least P-value (P-value = 2.726e-47)**




We also identified the top 20 genes with the most significant ASE in the QTL region (Table 4.5). Among those, 4/20 (*AL2G30710*, *AL2G30870*, *AL2G27860*, *AL2G34280*) are transcription factor regulators, 3/20 (*AL2G27970*, *AL2G33570*, *AL2G32660*) are involved in protein binding, 3/20 (*AL2G31800*, *AL2G27470*, *AL2G30460*) support transporter activity, and 2/20 (*AL2G34280*, *AL2G30290*) are involved in managing biotic, abiotic stress. The locus *AL2G34280* encodes a MYB transcription factor for pathogen defense. The gene *AL2G30710* encodes an ethylene-responsive transcription factor, and *AL2G27860* is a positive regulator of cytokinin levels and cytokinin-mediated development. One gene, *AL2G27470*, encodes a membrane transporter for Gibberellic Acid. The top 20 genes ranked by fold difference in allelic expression are shown in Table 4.6.

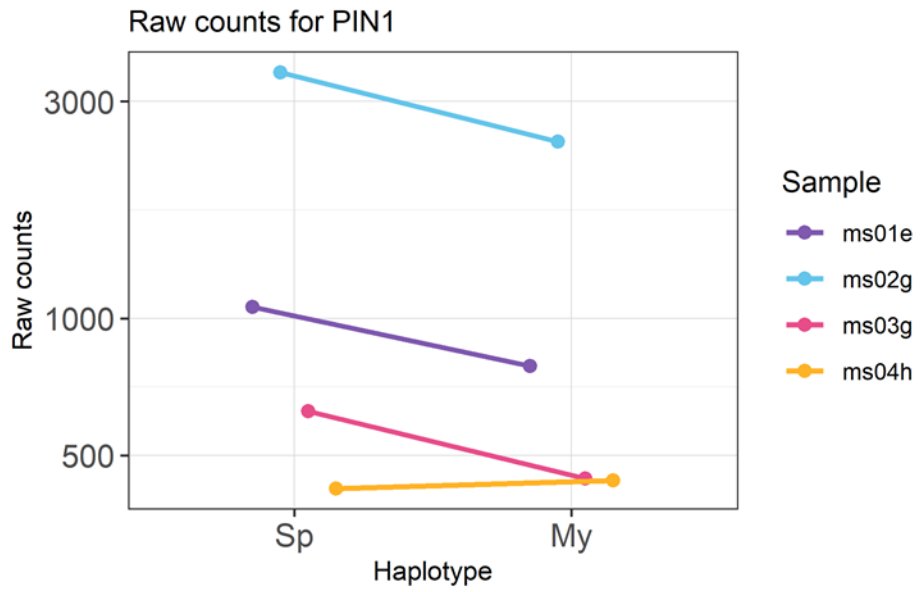
**Results for Candidate Genes**

We also looked into the candidate genes vital in plant morphological development, which we hypothesized earlier, *PIN1*, *PIN3*, *BRC2*, and *PILS2*, along with two other genes in the *TCP* family (*TCP15* and *TCP22*) and a key regulator of transition to flowering (*AP1*).

***PIN1***

*PIN1* showed higher expression of the Sp allele in 3 out of 4 samples (Figure 4.12, Figure A5). However, the difference is modest, resulting in a *P*-value equal to 0.08.

**Figure 4.12 Raw Counts (Y-axis) for PIN1 Observed for My and Sp Alleles (X-axis) Across Samples**

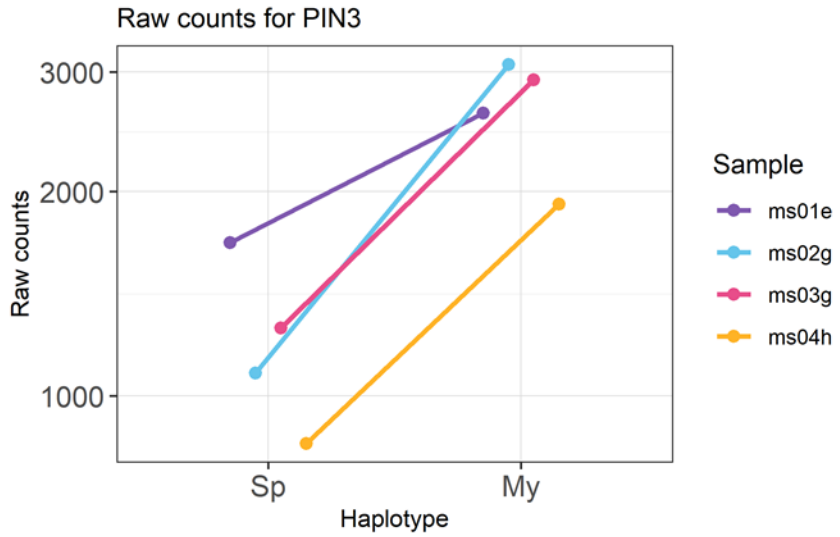



***PIN3***

*PIN3* showed higher expression of the My allele in all the samples (Figure 4.13, Figure A6) and is one of the top 20 gene in the QTL region (15$^{th}$ position, see Table 4.5) and is at 60$^{th}$/1761 position in the overall genes tested. The quantitative difference between two alleles seems to be exceeding even that for the gene with the lowest *P*-value, but the fold change is still

high for *AL2G19580*. However, in the case of the *PIN3* gene, the magnitude still varies strongly across samples. The Wald-test *P*-value for this gene is 1.053203e-07.

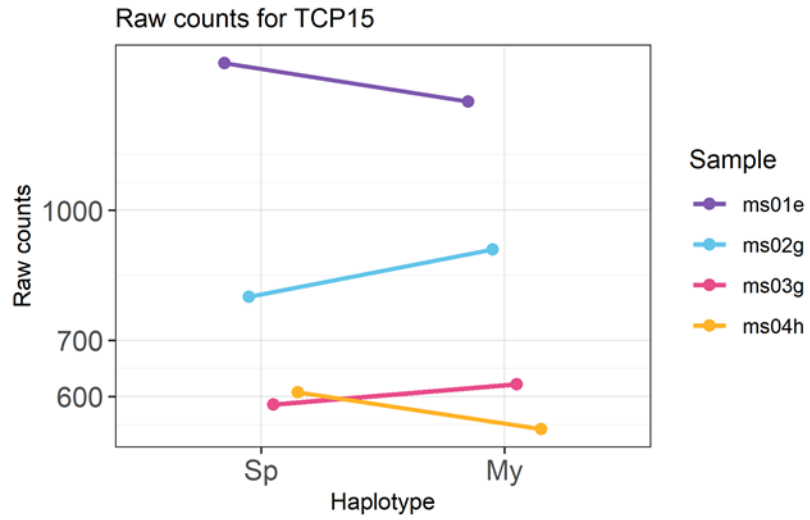
**Figure 4.13 Raw Counts (Y-axis) for PIN3 Observed for My and Sp Alleles (X-axis) Across Samples**




***TCP15***

*TCP15* gene does not show any ASE trend across samples (Figure 4.14, Figure A7). There are two samples with Sp alleles exceeding My and two vice versa. The Wald test *P*-value for this gene is also relatively high, at 0.981, which means that the difference in expression is insignificant; see Figure 4.14.

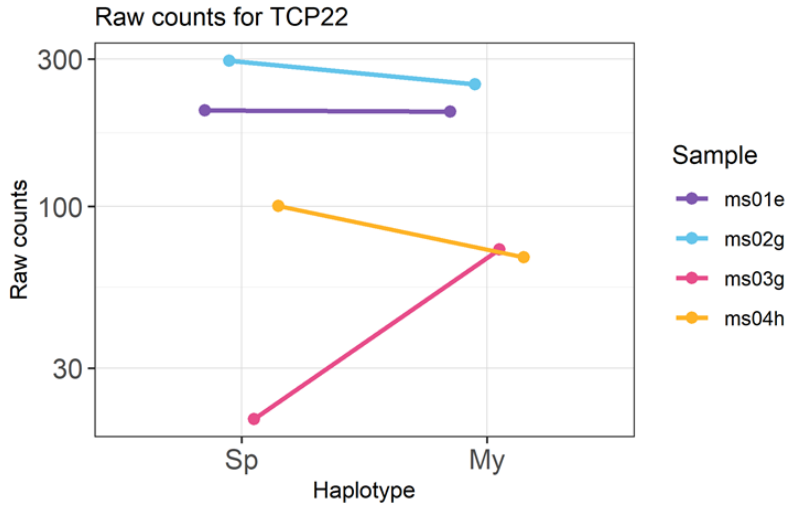
**Figure 4.14 Raw Counts (Y-axis) for TCP15 Observed for My and Sp Alleles (X-axis) Across Samples**




***TCP22***

Figure 4.15 and Figure S4.A8 represent raw counts for the TCP22 gene. Sample 2ms04h is the only sample with strong gene expression differences. The *P*-value for TCP is as well relatively high at 0.816.

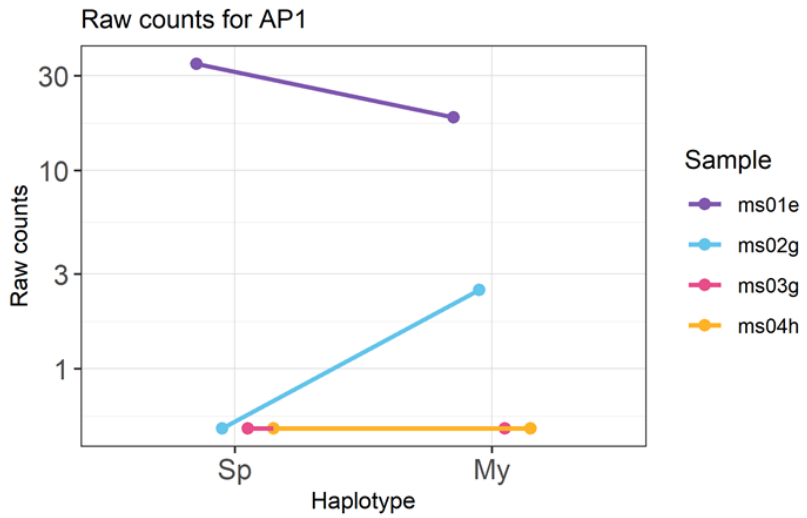
**Figure 4.15 Raw Counts (Y-axis) for TCP22 Observed for My and Sp Alleles (X-axis) Across Samples**



***AP1***

Gene *AP1* has a deficient gene expression in samples 2ms03g and 2ms04h. However, there is a clear difference in raw counts in the other two samples. The direction of the difference is different, however (Figure 4.16, Figure A9) . Therefore, the results of the Wald test indicate that there's no significant difference in gene expression (*P*-value = 0.91).

**Figure 4.16 Raw Counts (Y-axis) for AP1 Observed for My and Sp Alleles (X-axis) Across Samples**



***PILS2***

ASE differences for *PILS2* (*AL2G30670*) are statistically significant (*P*-value=0.00062), but My shows less than 2-fold greater expression levels than Sp, see Figure 4.17.

**Figure 4.17 Raw Counts (Y-axis) for PILS2 Observed for My and Sp Alleles (X-axis) Across Samples**

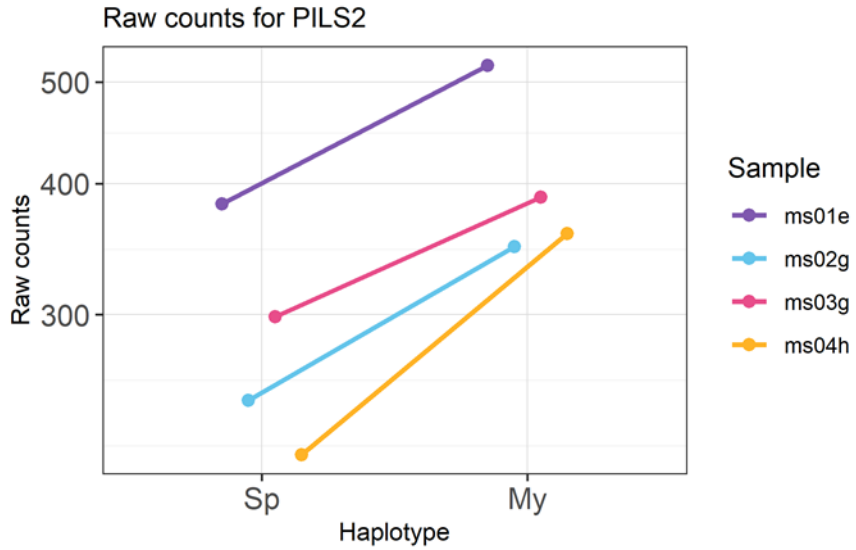


**Table 4.4 Top 20 Genes on the Entire Chromosome 2 With the Most Significant ASE (by P-value) Reported by the Wald Test**

| Gene ID | Mean (unqC_My) | SD (unqC_My) | Mean (unqC_Sp) | SD (unqC_Sp) | *P*-value (FDR) | in QTL region |
|---|---|---|---|---|---|---|
| AL2G12300 | 658.5 | 58.84159527 | 1.5 | 1.914854216 | 6.14E-24 | |
| AL2G13160 | 964 | 481.0100484 | 0.5 | 1 | 5.93E-19 | |
| AL2G14850 | 767 | 704.6408068 | 85.5 | 91.17565465 | 4.94E-18 | |
| AL2G16660 | 9.5 | 10.24695077 | 8618 | 6750.671473 | 1.98E-44 | |
| AL2G17900 | 2373.5 | 1201.733609 | 943.5 | 440.8246061 | 3.78E-35 | |
| AL2G19580 | 636 | 327.8414251 | 106 | 56.09515725 | 2.73E-47 | |
| AL2G23860 | 209.5 | 148.4710971 | 804 | 574.6488783 | 7.92E-27 | |
| AL2G25130 | 1655 | 1019.456718 | 745 | 524.2454896 | 8.48E-15 | |
| AL2G25210 | 2 | 1.632993162 | 90.5 | 55.60275773 | 3.19E-14 | |
| AL2G26480 | 218.5 | 60.56126375 | 62 | 29.52964612 | 3.65E-15 | |
| AL2G27970 | 184 | 56.35601121 | 18 | 8.164965809 | 3.12E-14 | + |
| AL2G30710 | 696 | 60.37659591 | 145.5 | 42.12283625 | 5.47E-20 | + |
| AL2G30870 | 460 | 71.49825173 | 158.5 | 53.30103188 | 5.93E-19 | + |
| AL2G31320 | 308.5 | 10.63014581 | 738 | 104.9317238 | 8.73E-21 | + |
| AL2G31480 | 2848.5 | 784.4460891 | 18 | 16.08311744 | 1.60E-30 | + |
| AL2G31800 | 190 | 118.0169479 | 37.5 | 20.80865205 | 9.21E-15 | + |
| AL2G35740 | 780.5 | 750.7904723 | 4025.5 | 3315.012871 | 6.61E-14 | |
| AL2G35970 | 324 | 112.0833024 | 77 | 17.4737899 | 1.83E-16 | |
| AL2G36030 | 361.5 | 86.53900855 | 92 | 7.831560083 | 1.63E-25 | |
| AL2G38880 | 39 | 32.68026928 | 281.5 | 175.4565473 | 7.36E-23 | |

*Note:* The genes with "+" in the column "in QTL region" belong to LG2 QTL, which showed strong life-history differences between our study populations.

**Table 4.5 Top 20 Genes in the LG2 QTL Regions, With the Most Significant ASE (by P-value) Reported by the Wald Test**

| Gene ID | Gene Name | unqC_My | unqC_Sp | direction | Padj (FDR) | Annotations |
|---|---|---|---|---|---|---|
| AL2G31480 | - | 11394 | 72 | My > Sp | 1.60E-30 | - |
| AL2G31320 | HGPT | 1234 | 2952 | Sp > My | 8.73E-21 | phosphoribosyltransferase |
| AL2G30710 | - | 2784 | 582 | My > Sp | 5.47E-20 | Ethylene responsive transcription factor |
| AL2G30870 | ATWHY2 | 1840 | 634 | My > Sp | 5.93E-19 | Transcription regulator |
| AL2G31800 | - | 760 | 150 | My > Sp | 9.21E-15 | Transmembrane transport |
| AL2G27970 | - | 736 | 72 | My > Sp | 3.12E-14 | Protein binding |
| AL2G34280 | MYB95 | 1088 | 5208 | Sp > My | 7.60E-14 | MYB transcription factor |
| AL2G28140 | - | 1010 | 6 | My > Sp | 2.25E-12 | Amino transferase |
| AL2G27130 | ENDO 2 | 7826 | 1865 | Sp > My | 1.68E-11 | Response to stress |
| AL2G33570 | - | 2184 | 6074 | Sp > My | 2.04E-11 | Zinc finger, protein binding |
| AL2G33070 | scpl6 | 958 | 256 | My > Sp | 7.78E-10 | Proteolysis |
| AL2G32500 | - | 4998 | 2774 | My > Sp | 3.91E-09 | - |
| AL2G27860 | ATSOFL2 | 204 | 744 | Sp > My | 6.00E-09 | Positive regulator of cytokinin levels |
| AL2G27470 | - | 11994 | 6782 | My > Sp | 3.09E-08 | Gibberelic acid membrane transport |
| AL2G30460 | PIN3 | 10522 | 4870 | My > Sp | 1.05E-07 | Auxin efflux transmembrane transport |
| AL2G32660 | - | 144 | 1222 | Sp > My | 1.10E-07 | Protein binding |
| AL2G30290 | - | 506 | 4 | My > Sp | 1.21E-07 | Defense response |
| AL2G31290 | - | 968 | 1680 | Sp > My | 2.34E-07 | Nucleic acid binding domain |
| AL2G31440 | PTP1 | 1120 | 594 | My > Sp | 6.87E-07 | Protein dephosphorylation |
| AL2G29070 | - | 950 | 1566 | Sp > My | 7.17E-07 | |

**Table 4.6 Top 20 Genes in the LG2 QTL Regions, With the Greatest Log2 Fold Difference In ASE expression (Reported by the Wald Test)**

| gene_ID | gene_Name | start | end | unqC_My | unqC_Sp | direction | padj | Log2Fold (My/Sp) | ABS(Log2Fold Change) |
|---|---|---|---|---|---|---|---|---|---|
| AL2G29760 | - | 14401428 | 14403086 | 1216 | 6 | My > Sp | 0.239741 | 7.441759 | 7.441759 |
| AL2G31480 | - | 15107502 | 15110686 | 11394 | 72 | My > Sp | 1.60E-30 | 7.286289 | 7.286289 |
| AL2G28140 | - | 13562675 | 13564212 | 1010 | 6 | My > Sp | 2.25E-12 | 7.174212 | 7.174212 |
| AL2G33420 | AtRABA6a | 15925702 | 15926926 | 130 | 0 | My > Sp | 7.32E-07 | 7.033423 | 7.033423 |
| AL2G30290 | - | 14635146 | 14636342 | 506 | 4 | My > Sp | 1.21E-07 | 6.663914 | 6.663914 |
| AL2G28230 | - | 13593457 | 13594282 | 0 | 90 | Sp > My | 0.065418 | -6.50779 | 6.507795 |
| AL2G32650 | - | 15620819 | 15622127 | 0 | 78 | Sp > My | 0.865459 | -6.30378 | 6.303781 |
| AL2G32540 | - | 15586449 | 15591281 | 0 | 60 | Sp > My | 0.000314 | -5.93074 | 5.930737 |
| AL2G31780 | - | 15225835 | 15226752 | 4 | 248 | Sp > My | 2.92E-06 | -5.63807 | 5.638074 |
| AL2G32490 | - | 15564449 | 15566342 | 0 | 46 | Sp > My | 0.026045 | -5.55459 | 5.554589 |
| AL2G30630 | - | 14770030 | 14771631 | 38 | 0 | My > Sp | 0.021017 | 5.285402 | 5.285402 |
| AL2G27220 | - | 13039018 | 13042554 | 0 | 30 | Sp > My | 0.240909 | -4.9542 | 4.954196 |
| AL2G28010 | - | 13501247 | 13502612 | 2 | 90 | Sp > My | 7.22E-05 | -4.92283 | 4.922832 |
| AL2G27280 | - | 13077719 | 13084998 | 518 | 20 | My > Sp | 0.000952 | 4.627273 | 4.627273 |
| AL2G27900 | AtbZIP | 13440351 | 13441078 | 2 | 64 | Sp > My | 0.001356 | -4.43741 | 4.437405 |
| AL2G28690 | - | 13828335 | 13831656 | 4 | 66 | Sp > My | 0.00291 | -3.74416 | 3.744161 |
| AL2G32800 | - | 15681217 | 15683572 | 2 | 38 | Sp > My | 0.058945 | -3.70044 | 3.70044 |
| AL2G27970 | - | 13470624 | 13474668 | 736 | 72 | My > Sp | 3.12E-14 | 3.335696 | 3.335696 |
| AL2G32660 | - | 15623379 | 15624766 | 144 | 1222 | Sp > My | 1.10E-07 | -3.0763 | 3.0763 |
| AL2G32530 | - | 15579500 | 15583054 | 88 | 10 | My > Sp | 0.004609 | 3.016302 | 3.016302 |

## Discussion

This research identifies that *PIN3* and *PILS2* show significant ASE in the predicted direction (My > Sp) among the few hypothesized candidates. One of the anticipated candidates, *BRC2*, almost did not show any expression. *PIN1*, on the contrary, showed slightly higher expression of Sp alleles. Our research gives some interesting insights; we see the strong expression of Mayodan alleles of *PIN3* and *PILS2* from the resource allocation QTL region, while *PIN1* shows almost equal bi-allelic expression, with the Sp allele showing only a little higher expression. The expression of the *BRC2* gene might be limited to certain tissues, especially meristems, because their specific role is to arrest the growth of the meristems, and extraction of mRNA from the whole shoot could have diluted their expression levels.

The genomic position of *PIN3* is almost precisely in the middle of the resource allocation QTL region. *PIN3* is an essential auxin transport protein associated with the control and localization of auxin toward the cell membrane's lateral side in response to gravitropic (Müller et al., 1998; Ottenschläger et al., 2003) and phototropic stimulation (Ding et al., 2011; Friml et al., 2002; T. Hu et al., 2021; Savaldi-Goldstein et al., 2007). *PIN3* expression is coupled with elevated auxin biosynthesis for SAS (Shade Avoidance Syndrome) (Keuskamp et al., 2010; Tao et al., 2008). Low R:FR (red/far-red light) ratio stimulates *PIN3* abundance, induces a lateral cellular reorientation of *PIN3,* and elevates auxin levels. The low R:FR environment itself regulates *PIN3* gene expression, further promoting PIN3 protein abundance and localization, directing its own (auxin) transport (Keuskamp et al., 2010). This environment-based adaptive significance of *PIN3* is observed in competitive experiments that comprise high plant densities. Low R:FR induces elongation and development of the hypocotyls, thereby reducing the fitness of *PIN3* mutant in low R:FR conditions compared to wild-type variants (Friml et al., 2002). In

previous Chapter 2, we found somewhat inconclusive support for the role of differences in auxin transport underlying life-history differences between the two study populations. The observed apparent cis-regulatory differences in *PIN3* allelic expression in this study could easily be the cause.

The other protein, PILS2, constitutes a protein from a distinct family that evolved independently but is structurally similar to PIN proteins (Feraru et al., 2012). While *PINs* mediate long-distance auxin transport, *PILS2,* a putative auxin carrier, localizes to the endoplasmic reticulum (ER) and regulates intracellular auxin accumulation and auxin homeostasis (Barbez et al., 2012; Feraru et al., 2012; Mohanta et al., 2015). The *PILS* protein family is conserved throughout the plant lineage, including unicellular algae (*Ostreococcus tauri* and *Chlamydomonas reinhardtii*). However, the PIN proteins are absent in those algae, indicating that PILS proteins evolved before PINs. This suggests that intracellular auxin transport and auxin compartmentalization are evolutionarily older than directional, cell-to-cell PIN-dependent auxin transport (Barbez et al., 2012). In our study, we found higher expression of My alleles of *PILS2*. This correlated higher expression of *PIN3* and *PILS2* suggests that higher transport of auxin-mediated by *PIN3* could be supported by higher auxin compartmentalization and supply of auxin-mediated by *PILS2,* contributing to higher apical dominance in Mayodan individuals and likely resource allocation tradeoffs.

The locus *AL2G34280* regulates the MYB transcription factor for pathogen defense and is not a likely candidate for life-history variation. Among other genes with the lowest ASE *P*-values (Table 4.5) in this QTL region, *AL2G30710, AL2G27860,* and *AL2G27470* could have a role in developmental variation since both are involved in phytohormone response. Functionally, *AL2G30710* is an ethylene-responsive transcription factor, and *AL2G27860* is a positive

regulator of cytokinin levels and cytokinin-mediated development, which could affect developmental variation or work together with *PIN3* to mediate developmental variation. The gene, *AL2G27470,* encodes for membrane transporter for Gibberellic Acid and would be interesting, but annotations in other studies describe it as being expressed in the root endodermis. However, in our research, this gene is clearly expressed in the main shoot of the plants. Overall, these three genes are involved in phytohormone response, and they could have a role in developmental variation. However, they don't seem nearly as good a possibility as *PIN3* (*AL2G30460*).

Based on these data, *PIN3* emerges as a strong candidate gene for this QTL region that explains the life-history variation between the two populations of *Arabidopsis lyrata* we studied. The phototropic-based adaptive responses could be the driving factor for the evolution of diverged resource allocation tradeoffs in two study populations. Since the Mayodan population has a relatively more extended seasonal growth period, it provides an opportunity for elevated auxin synthesis and lateral transport mediated by higher expression of *PIN3*. However, the mechanistic process by how *PIN3* would provide the adaptive basis still needs thorough analyses.

A more detailed list of genes showing significant ASE variation is provided in **Supplementary Materials S4.E** (sorted by *P*-value) and **Supplementary Materials S4.E** (sorted by genomic position).

## Future Studies

We could derive interesting future studies based on these results; a potential follow-up study would be to do a transgenic exchange of My alleles (for gene *PIN3*, *PILS2*) onto the Sp genotypes and vice-versa and study their effects on life-history traits. Another way of testing for

the effects of these genes/alleles would be to generate CRISPR knock-outs and check for their effects on life-history traits. Additional genomic and transcriptomics studies can also test if NPA affects gene expression variation in the parental populations, mainly Mayodan. This analysis can also include other genes involved in developmental variation *AL2G30710, AL2G27860,* and *AL2G27470*. Finally, other experiments involving verification of ASE expression could be done in the parental population using qPCR, which was one of the original goals of this chapter but was missed due to technical difficulties. One of the technical difficulties was that we were not successful in growing two parental populations and having them flower simultaneously. Also, *A. lyrata* is a perennial; they take about 6 months to reach the developmental stage necessary for the experiment, unlike *A. thaliana*, which is ready in about 1 month. It would also be beneficial to check if the *PIN1* gene shows coding polymorphisms likely to affect protein function. Mutations in the coding region could influence protein function and drive variation in apical dominance not due to variation in gene expression but through variation in protein function (i.e., coding polymorphisms could induce differences in auxin transport). This is potentially important because *PIN1* was a stronger *a priori* candidate than *PIN3*. Another vital investigation would be to evaluate flanking region polymorphisms in *PIN3* and *PILS2* for gain/loss of transcription factor binding sites. This analysis could provide insights into the cis-regulatory mechanisms of gene expression differences.

## Supplementary Materials: Chapter IV

### Abbreviations

My: Mayodan

Sp: Spiterstulen

SE: Standard error estimate

DnaSP: DNA Sequence Polymorphism

SAM: Sequence Alignment Map

MA plot: M (log ratio) and A (mean average) scales plot

### Supplementary Materials S4.A: Extra Diagrams

**Figure S4.A1 Box Plot Showing the Distribution of Total Gene Expression for Genes With Log2FoldChange > 2 and Wald Test P-value < 0.05**

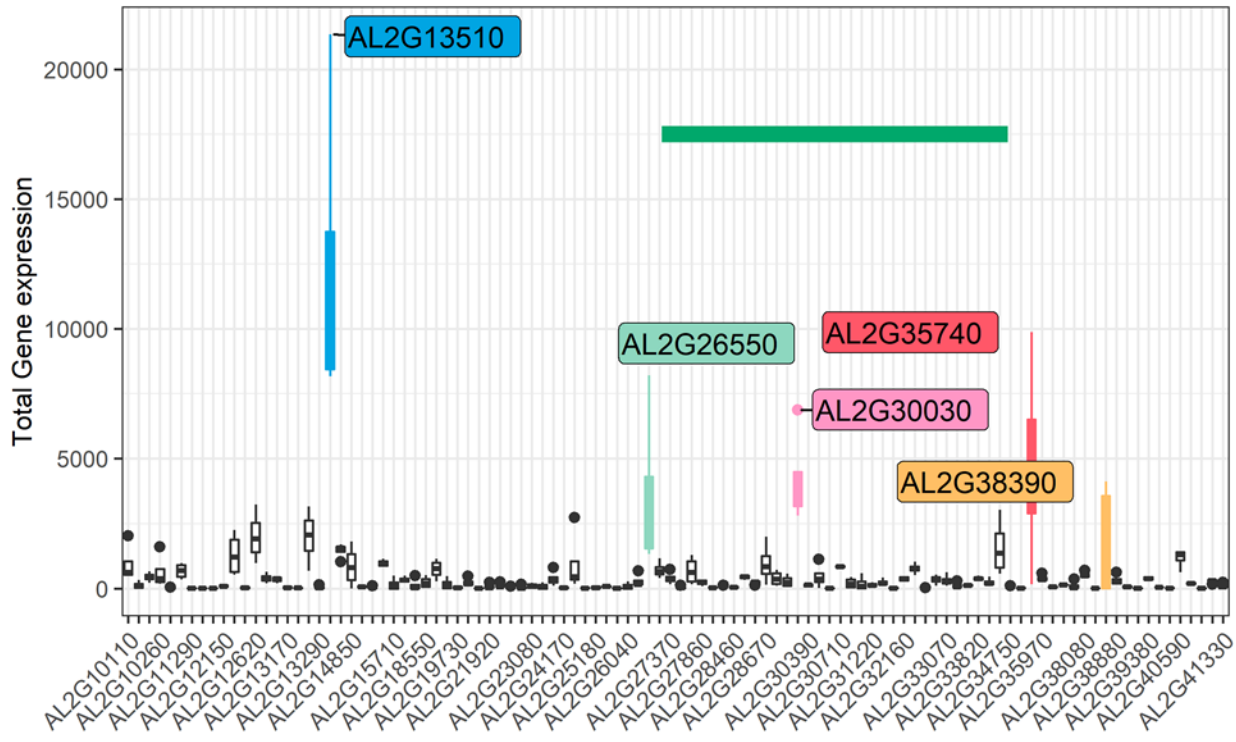



**Figure S4.A2 Box Plot Showing the Distribution of Total Gene Expression for Genes With Log2FoldChange > 4 and Wald Test P-value < 0.01**

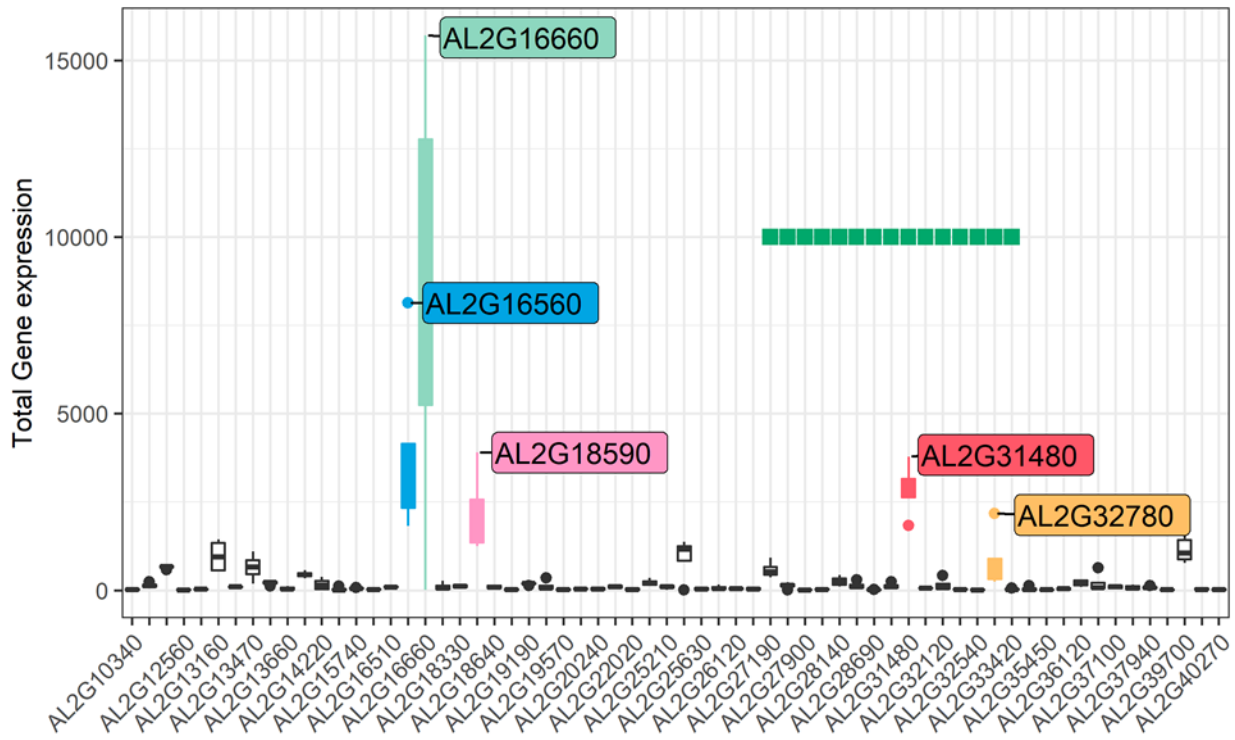
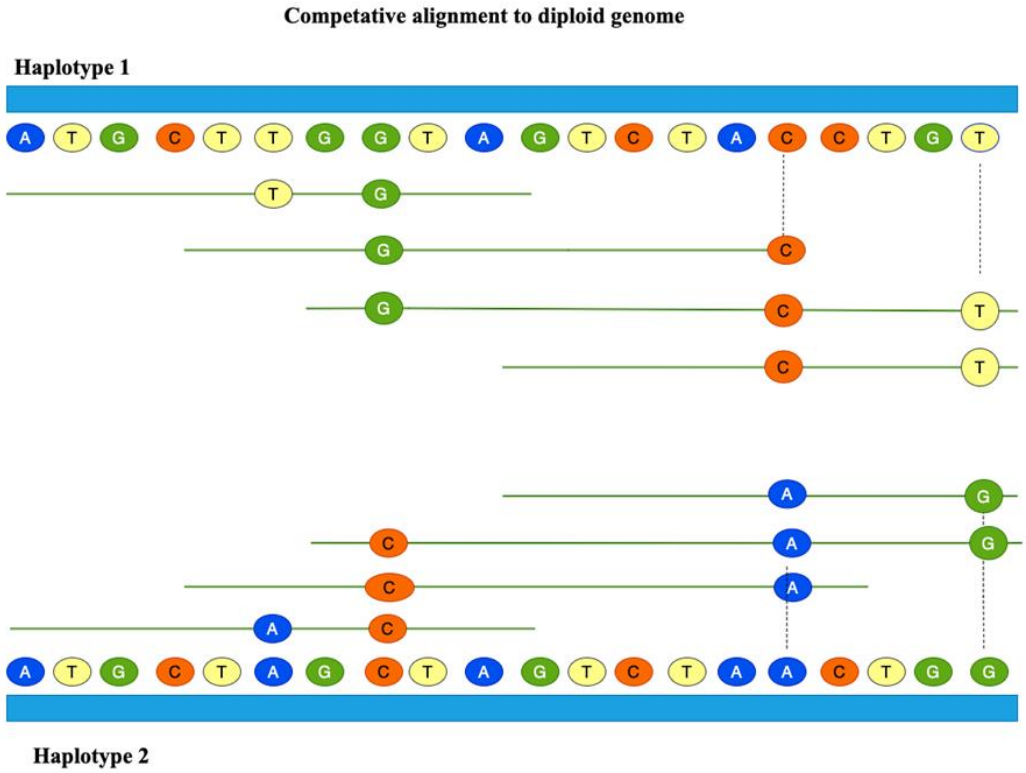
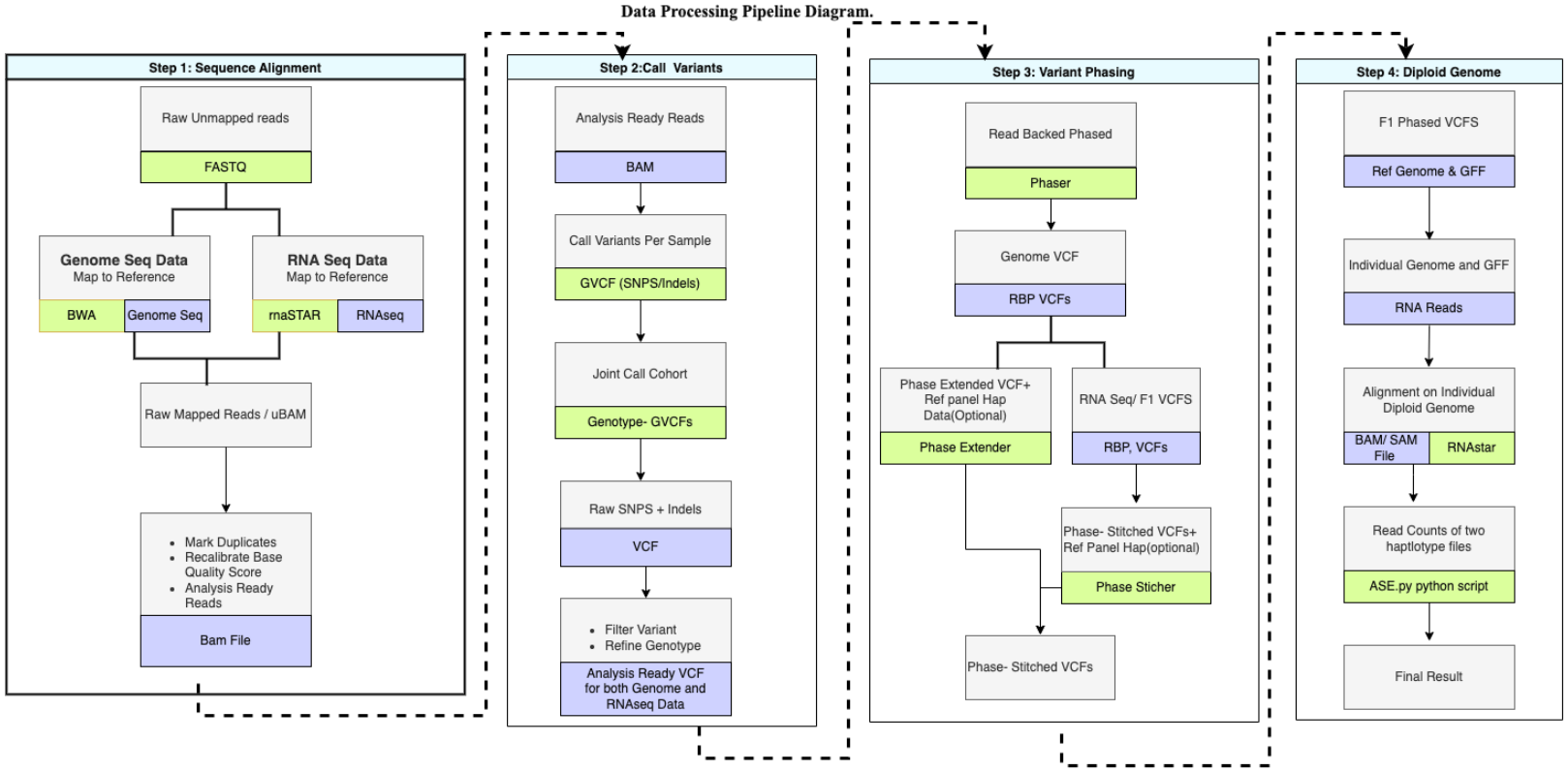

Total Gene expression
15000
10000
5000
0
AL2G16660
AL2G16560
AL2G18590
AL2G31480
AL2G32780
AL2G10340
AL2G12560
AL2G13160
AL2G13470
AL2G13660
AL2G14220
AL2G15740
AL2G16510
AL2G16660
AL2G18330
AL2G18640
AL2G19190
AL2G19570
AL2G20240
AL2G22020
AL2G25210
AL2G25630
AL2G26120
AL2G27190
AL2G27900
AL2G28140
AL2G28690
AL2G31480
AL2G32120
AL2G32540
AL2G33420
AL2G35450
AL2G36120
AL2G37100
AL2G37940
AL2G39700
AL2G40270

**Figure S4.A3 A schematic of competitive alignment on the diploid genome**




*Note:* Haplotype 1 contains haplotype "T-G-C-T", whereas haplotype 2 contains haplotype variant "A-C-A-G". ASE can be estimated by counting the allele expressed at each variant site; e.g., from "Haplotype 1" 3 alleles of "G" are expressed at the 8$^{th}$ position, with "Haplotype 2" showing the same number of "C". Since both alleles are equally expressed there is no significant ASE difference. Another way is to count the total number of reads aligned, which in this case, "Haplotype 1" has 4 reads expressed and "Haplotype 2" also has 4 reads expressed, still showing no significant ASE difference. In our method, we use the latter to count the total number of reads aligned to the haplotype (or gene/transcriptomic) region.

**Figure S4.A4 Bioinformatics data processing pipeline diagram for haplotype phasing and ASE analysis**




*Note:* This pipeline contains 4 stages. The first stage deals with sequence alignment, the second stage deals with calling variants, the third stage deals with phasing variants, and the last/fourth stage deals with the alignment of the reads competitively on a diploid genome, then counts the read aligned to two haplotypes for ASE analysis.

**Figure S4.A5 Bar plot showing raw counts for PIN1 Observed for My and Sp Alleles Across Samples**

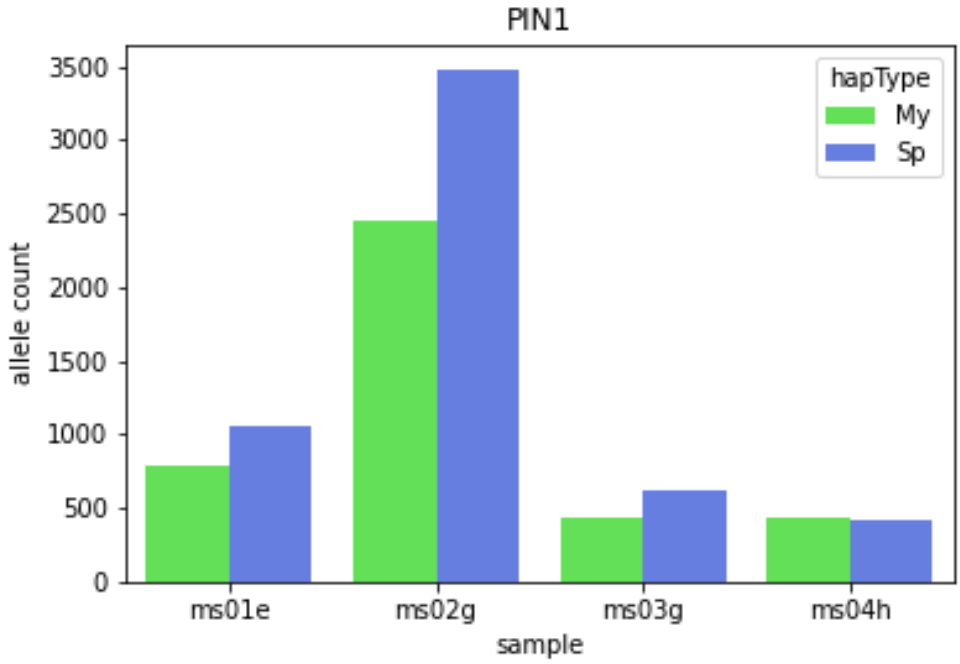



**Figure S4.A6 Bar plot showing raw counts for PIN3 Observed for My and Sp Alleles Across Samples**

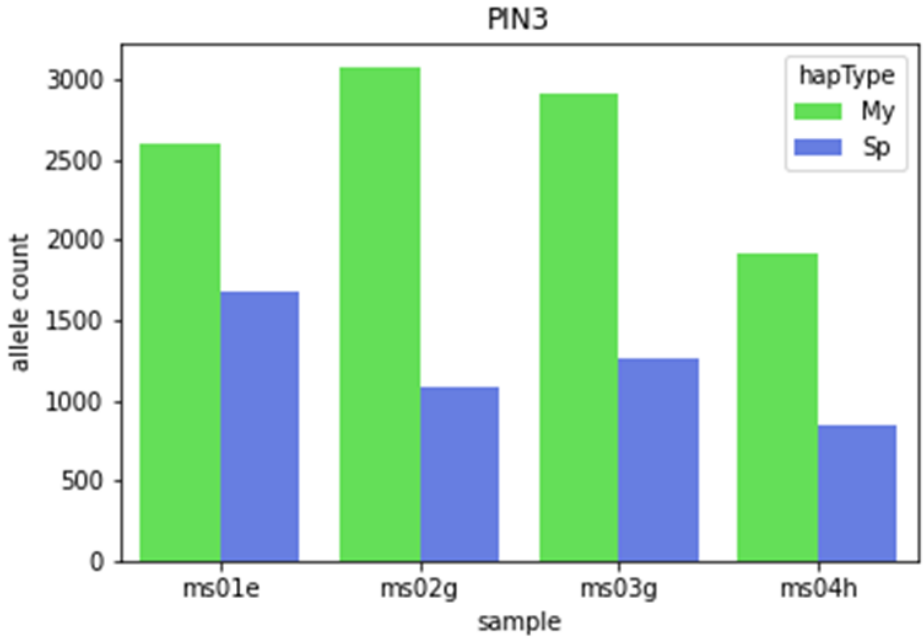


**Figure S4.A7 Bar plot showing raw counts for TCP15 Observed for My and Sp Alleles Across Samples**

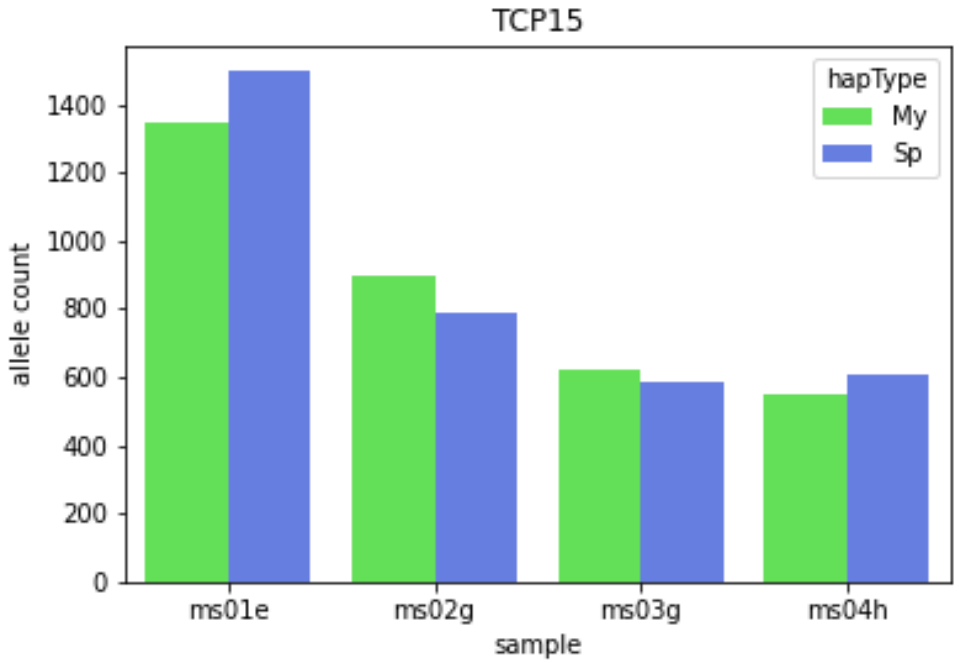



**Figure S4.A8 Bar plot showing raw counts for TCP22 Observed for My and Sp Alleles Across Samples**

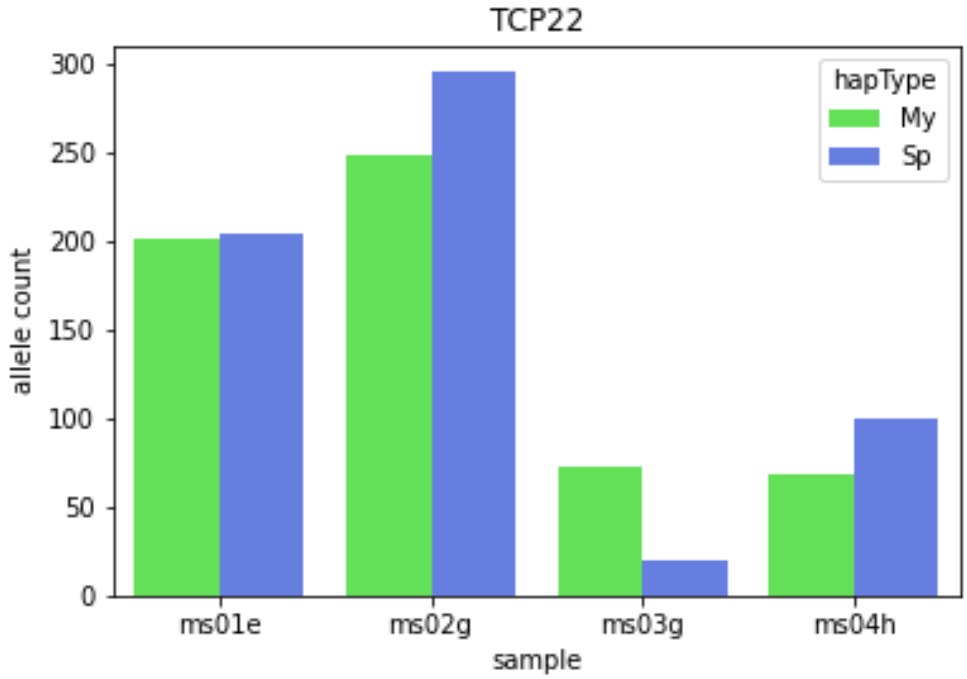


**Figure S4.A9 Bar plot showing raw counts for AP1 Observed for My and Sp Alleles Across Samples**

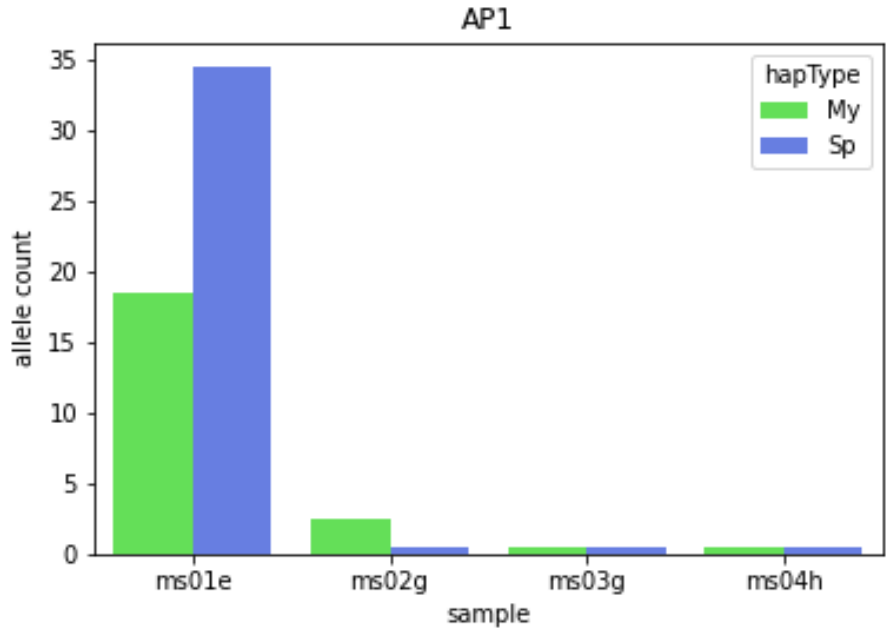
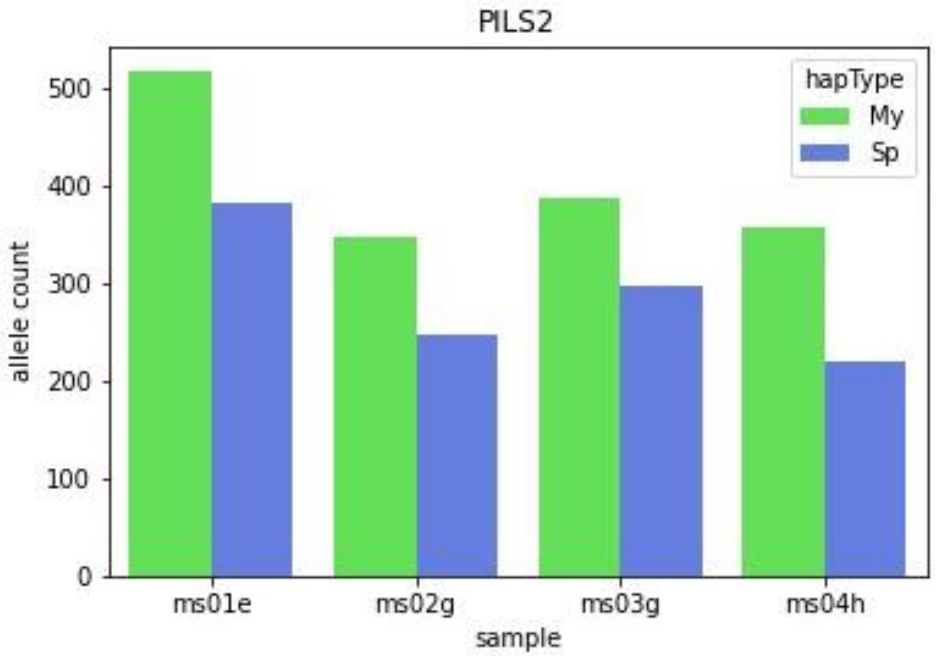



**Figure S4.A10 Bar plot showing raw counts for PILS2 Observed for My and Sp Alleles Across Samples**



**Supplementary Materials S4.B: GENOME SEQ PIPELINE**

The original bash scripts and pipeline for this procedure is available at,

- https://github.com/everestial/phase-extender

- https://github.com/everestial/ASE-CADG

**Mayodan sample names:** MA605, MA611, MA622, MA625, MA629, Ncm8

**Spiterstulen sample names:** Sp154, Sp164, Sp21, Sp3, Sp76, SpNor33

**F1 sample name:** 2ms01e, 2ms02g, 2ms03g, 2ms04h

**Step 01: Trim the adapter from My and Sp fastq files**

*trimSoftware= ~ /Trimmomatic-0.35/trimmomatic-0.35.jar*

*item= "MA605"*

*java -jar trimSoftware PE -phred33 ${item}_R1.fastq ${item}_R2.fastq Trm_PE-${item}_R1.fastq Trm_SE-${item}_R1.fastq Trm_PE-${item}_R2.fastq Trm_SE-${item}_R2.fastq ILLUMINACLIP:all_all_adapters_primers-RNAseq_updated01.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36*

**Step 02: Index the reference genome prior to alignment (if the reference genome and appropriate indexes are already prepared it won't be required). See the scripts below to prepare index for the tools: bwa, samtools and picard.jar**

*GATK3p8= /GenomeAnalysisTK-3.8/GenomeAnalysisTK.jar*

*Picard= /Picard2.16/picard.jar*

*RefGenome= /RefGenomeN_index/lyrata_genome.fa*

*TempDir= /temp_files/*

*samtools=/usr/local/apps/samtools/0.1.19-gcc412/samtools*

**2.1 prepare the bwa index of the reference genome**

This step creates 4-8 different index files with extensions *.sa *.amb, etc.

*bwa index -a bwtsw ${RefGenome}*

**2.2 generate index for samtools // this creates a *.fai index file**

*${samtools} faidx ${RefGenome}*

**2.3 generate sequence dictionary index for picard.jar // this creates a *.dict index file**

*java -jar ${Picard} CreateSequenceDictionary REFERENCE=${RefGenome}*

*OUTPUT=lyrata_genome.dict*

**Step 03: Generate an unmapped BAM from FASTQ**

*GATK3p8= /apps/GenomeAnalysisTK-3.8/GenomeAnalysisTK.jar*

*Picard= /apps/Picard2.16/picard.jar*

*RefGenome= /apps/RefGenomeN_index/lyrata_genome.fa*

*TempDir= /apps/temp_files/*

java -Xmx4G -jar ${Picard} FastqToSam FASTQ=Trm_PE-MA605_R1.fastq FASTQ2=Trm_PE-MA605_R2.fastq OUTPUT=uBAM_MA605.bam READ_GROUP_NAME=MA605_C080WACXX_7 SAMPLE_NAME=MA605 LIBRARY_NAME=C080WACXX_4 PLATFORM_UNIT=GTGAAA.7 PLATFORM=illumina SEQUENCING_CENTER=EU TMP_DIR=${TempDir}

**Step 04: Map and Mark Duplicates With Mate Cigar values**

Function/Purpose: Jointly realign the aligned reads around the indel regions for all the samples.

java -jar ${GATK3p8} -T PrintReads -R ${RefGenome} -I ${item}_duplicates_merged.bam -o ${item}.deDuplicated.bam -rf DuplicateRead

# index all the files using samtools

samtools index *.*bam

**Step 05: First create Realignment interval file.**

Run jointly on all samples (from population)

java -Xmx4g -jar ${GATK3p8} -T RealignerTargetCreator -R ${RefGenome} -I MA605.deDuplicated.bam -I MA611.deDuplicated.bam -I MA622.deDuplicated.bam -I MA625.deDuplicated.bam -I MA629.deDuplicated.bam -I Ncm8.deDuplicated.bam -o all_MA-Indels.intervals

Then run IndelRealigner for realignment of the reads around the target indels to correct any SNP artifactual. Local realignment helps to identify the most parsimonious alignment.

java -jar ${GATK3p8} -T IndelRealigner -R ${RefGenome} -I MA605.deDuplicated.bam -I MA611.deDuplicated.bam -I MA622.deDuplicated.bam -I MA625.deDuplicated.bam -I MA629.deDuplicated.bam -I Ncm8.deDuplicated.bam -targetIntervals all_MA-Indels.intervals -nWayOut '.indel_realigned.bam' --consensusDeterminationModel USE_SW

This step creates a file called *deDuplicated.realigned.bam containing all the original reads but with better local alignments in the targeted regions using the interval file"all_MA-Indels.interval". Finally, we use --consensusDeterminationModel USE_SW to generate alternate consensus using Smith-Waterman.

**Step 06: This step filters the paralogous and ambiguous alignment by mapping quality and coverage.**

**6.1: Filter BAM files**

samtools sort mapQ20.${item}.bam mapQ20.sorted.${item}

samtools index mapQ20.sorted.${item}.bam

rm mapQ20.${item}.bam  # remove the unrequired file

**6.2: prepare a genome coverage file (text file) in bedgraph format**

bedtools genomecov -ibam mapQ20.sorted.${item}.bam -bg > genomecov.mapQ20_${item}.bg

**6.3: First, prepare a bed file with high coverage (for each sample)**

For My Samples

awk '($1 < 9 && $3-$2 > 5 && $4 > 21); ($1 == "scaffold_9" && $3-$2 > 5 && $4 > 47); ($1 == "scaffold_10" && $3-$2 > 5 && $4 > 127)' genomecov.mapQ20_MA605.bg > highCoverage_MA605.bg

Note: For this filtering we need new version of samtools (1.3)

SamtoolsV1_3= /Samtools1.3/Samtools1.3

${SamtoolsV1_3} view -b -h -L highCoverage_${item}.bg -U good_Coverage.mapQ20.${item}.bam mapQ20.sorted.${item}.bam -o highCoverage.${item}.bam

samtools sort good_Coverage.mapQ20.${item}.bam good_Coverage.mapQ20.sorted.${item}

# only provide the prefix

samtools index good_Coverage.mapQ20.sorted.${item}.bam

rm highCoverage.${item}.bam

rm good_Coverage.mapQ20.${item}.bam

rm mapQ20.sorted.${item}.bam

**Step 07: Joint Variant Call on all Samples.**

**7.1 Run single-sample gvcf calls**

java -jar -Xmx32g -Djava.io.tmpdir=${TempDir} ${GATK3p8} -T HaplotypeCaller -R ${RefGenome} -I good_Coverage.${item}.bam --emitRefConfidence GVCF -o ${item}.raw.snps.indels.g.vcf

java -jar -Xmx32g -Djava.io.tmpdir=${TempDir} ${GATK3p8} -T HaplotypeCaller -R ${RefGenome} -I realigned_${item}.bam --emitRefConfidence GVCF -o ${item}.raw.snps.indels.g.vcf

**7.2 If there are lots of samples run combineGVCF on a batch of 200s**

# Our samples size is small, so skipping this step

**7.3 : Joint Genotyping by running GenotypeGVCFs on all the samples together**

#

https://software.broadinstitute.org/gatk/documentation/tooldocs/org_broadinstitute_gatk_tools_walkers_variantutils_GenotypeGVCFs.php

*java -jar -Xmx32g -Djava.io.tmpdir=${TempDir} ${GATK3p8} -T GenotypeGVCFs -R ${RefGenome} --variant MA605.raw.snps.indels.g.vcf --variant MA611.raw.snps.indels.g.vcf --variant MA622.raw.snps.indels.g.vcf --variant MA625.raw.snps.indels.g.vcf --variant MA629.raw.snps.indels.g.vcf --variant Ncm8.raw.snps.indels.g.vcf --variant Sp3.raw.snps.indels.g.vcf --variant Sp21.raw.snps.indels.g.vcf --variant Sp76.raw.snps.indels.g.vcf --variant Sp154.raw.snps.indels.g.vcf --variant Sp164.raw.snps.indels.g.vcf --variant SpNor33.raw.snps.indels.g.vcf --variant ms01e.raw.snps.indels.g.vcf --variant ms02g.raw.snps.indels.g.vcf --variant ms03g.raw.snps.indels.g.vcf --variant ms04h.raw.snps.indels.g.vcf -o raw_variants.AllSamples.MySpF1.vcf*

**Step 08: We take the Joint Genotyped VCF data and run Variant Filtration (using either VQSR or Hard Filtering parameters)**

**8.1: Separate the truth Set variants Vs. Other Set of Variants**

## We are taking the variants data provided by Detlef's lab as a truth set.

## So, any variants called in our set that matches or intersects with their variants are highly true.

**8.1.1 : variants common between my call and Detlef's lab VCF - i.e concordant variants**

*java -jar ${GATK3p8} -T SelectVariants -R ${RefGenome} -V raw_variants.AllSamples.MySpF1.vcf --concordance Detlef.Variants.asPASSED.vcf -o HighConf_Variants.AllSamples.vcf*

**8.1.2: Identify the discordant set of Variants**

# i.e the variants in my call but missing in Detlef's lab VCF - i.e Disconcordant variants

*java -jar ${GATK3p8} -T SelectVariants -R ${RefGenome} -V raw_variants.AllSamples.MySpF1.vcf --discordance Detlef.Variants.asPASSED.vcf -o Remaining.Variants.AllSamples.vcf*

**Step 8.2 (for SNPs) : Now, start filtering the variants from Discordant set.**

# Since Discordant set contains variants that may or may not be true, we do further filtering based on several "Quality parameters".

# In this Discordant set we take variants only from GenomicVariants (i.e Excluding the variants from RNAseq Samples) for quality control.

# Some GATK flags that are useful during this filtering: --removeUnusedAlternates , --excludeNonVariants , --excludeFiltered

**8.2.1: Separate GenomeReSeq Variants from RNAseq variants**

# removing RNAseq samples - use flag "-xl_sn" ; & remove Un-used Alternate alleles - use flag "--removeUnusedAlternates"

# remove sites that are uncalled (GT = ./.) after sample selection

# - use flag "'vc.getCalledChrCount() == 0' will select the sites that was noCall (./.) in all the samples

# - -invertSelect" will invert this selection and also include the variants where at least variant in one sample is called (including GT = 0/0).

*java -jar ${GATK3p8} -T SelectVariants -R ${RefGenome} -V Remaining.Variants.AllSamples.vcf -selectType SNP --removeUnusedAlternates -o DNA_Samples.raw_SNPs.vcf -xl_sn ms01e -xl_sn ms02g -xl_sn ms03g -xl_sn ms04h -select 'vc.getCalledChrCount() == 0' –invertSelect*

**8.2.2: Identify the discordant set of Variants**

# i.e the variants in my call but missing in Detlef's lab VCF - i.e Disconcordant variants

*java -jar ${GATK3p8} -T SelectVariants -R ${RefGenome} -V raw_variants.AllSamples.MySpF1.vcf --discordance Detlef.Variants.asPASSED.vcf -o Remaining.Variants.AllSamples.vcf*

**Step 8.3 (for SNPs) : Now, start filtering the variants from Discordant set.**

# Since Discordant set contains variants that may or may not be true, we do further filtering based on several "Quality parameters".

# In this Discordant set we take variants only from GenomicVariants (i.e Excluding the variants from RNAseq Samples) for quality control.

# Some GATK flags that are useful during this filtering: --removeUnusedAlternates , --excludeNonVariants , --excludeFiltered

**8.4.1: Separate GenomeReSeq Variants from RNAseq variants**

# removing RNAseq samples - use flag "-xl_sn" ; & remove Un-used Alternate alleles - use flag "--removeUnusedAlternates"

# remove sites that are uncalled (GT = ./.) after sample selection

# - use flag "'vc.getCalledChrCount() == 0' will select the sites that was noCall (./.) in all the samples

# - -invertSelect" will invert this selection and also include the variants where at least variant in one sample is called (including GT = 0/0).

*java -jar ${GATK3p8} -T SelectVariants -R ${RefGenome} -V Remaining.Variants.AllSamples.vcf -selectType SNP --removeUnusedAlternates -o DNA_Samples.raw_SNPs.vcf -xl_sn ms01e -xl_sn ms02g -xl_sn ms03g -xl_sn ms04h -select 'vc.getCalledChrCount() == 0' –invertSelect*

**8.4.2: Set the hard filter parameters (for SNPs)**

# Mark the Filtered SNPs and filterName from genomic VCF data

*java -jar -Xmx8g ${GATK3p7} -T VariantFiltration -R lyrata_genome.fa -V DNA_Samples.raw_SNPs.vcf --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filterName "my_snp_filter" -o DNA_Samples.Filtered_SNPs.vcf*

# Select passed variants (SNPs) - use either of below codes

*java -jar -Xmx8g ${GATK3p7} -T SelectVariants -R lyrata_genome.fa -V DNA_Samples.Filtered_SNPs.vcf -o DNA_Samples.Passed_SNPs.vcf -select 'vc.isNotFiltered()'*

**8.4.3 (for InDels) : Varian Filtration from Discordant set.**

Select Variants (InDels) and exclude samples from RNAseq

*java -jar -Xmx8g ${GATK3p8} -T SelectVariants -R lyrata_genome.fa -V Remaining.Variants.AllSamples.vcf -selectType INDEL --maxIndelSize 20 --removeUnusedAlternates -o DNA_Samples.raw_InDels.vcf -xl_sn ms01e -xl_sn ms02g -xl_sn ms03g -xl_sn ms04h -select 'vc.getCalledChrCount() == 0' -invertSelect*

# also removes UnUsed Alternates and sites that are uncalled for all selected samples

**8.4.4: Set the hard filter parameters (for InDels)**

# Mark the Filtered InDels and filterName from genomic VCF data

*java -jar -Xmx8g ${GATK3p7} -T VariantFiltration -R lyrata_genome.fa -V DNA_Samples.raw_InDels.vcf --filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" --filterName "my_indel_filter" -o DNA_Samples.Filtered_InDels.vcf*

*java -jar -Xmx8g ${GATK3p7} -T SelectVariants -R lyrata_genome.fa -V DNA_Samples.Filtered_InDels.vcf -o DNA_Samples.Passed_InDels.vcf -select 'vc.isNotFiltered()'*

**8.5 : Remove the sites that are heterozygote InDels in all called samples at that site.**

# i.e the sites that have atleast one Homozygous allele is retained

*java -jar ${GATK3p7} -T SelectVariants -R lyrata_genome.fa -V DNA_Samples.Passed_InDels.vcf -o DNA_Samples.Passed.InDels_allHetsRemoved.vcf -select 'vc.getHetCount() == vc.getCalledChrCount()/2' –invertSelect*

**8.6.1: Purpose/Function: Filter the variants from RNAseq Data**

# In the previous step 08-A-C we created high quality Final variants for Genome Data. In this workflow we prepare high quality Final Variants for RNAseq data.

# RNAseq data cannot/shouldnot be filtered based on the depth of the coverage. Because highly expressed genes will be expressed more and will have high coverage. Also, RNAseq reads mostly cover Genic regions, so paralog alignment from duplicated regions is not as pervasive as in GenomeReq data. Also, the paralog alignment for duplicated genes are only possible in local cluster within the GENE, rather than through out the whole genome.

# So, the fix for this problem is to filter the RNAseq aligned BAM files. We take the highly confident variants from Genomic VCFs and any covert it to a bedfile. After this we filter the RNAseq BAMs based on this bedfile. Any reads not touching the bed regions are filtered away.

# This works under the assumption that any reads that contains and/or overlapps a true variant (from BED regions) should be an ortholog reads.

**8.6.2 : Make a Bed file from all the passed Genomic VCF:**

## Use the self created python parser: VcfToBed

# https://github.com/everestial/VcfToBed

# use python > version 3.6

*python VcfToBed.py DNA_Samples.Passed_Variants.Final.vcf DNA_Samples.Passed_Variants.Final.bed*

**8.6.3 : Now filter the BAMs from RNAseq Data**

## Only retain the reads that 1) are mapQ > 40 2) touch the true set of variants from above BED file.

## **Note: While parallelizing with 'samtools' make sure to include the whole task/command within " " because parallel sees '>' as redirection rather than output. The '>' may also be enclosed within " " as ">" to work the problem out.

# filter the reads

*parallel --tmpdir ${TempDir} --jobs 4 "${samtools} view -b -q 40 realigned_{}.bam -L DNA_Samples.Passed_Variants.Final.bed > realigned_{}Filtered.bam" ::: ms01e ms02g ms03g ms04h*

*parallel --tmpdir ${TempDir} --jobs 4 "${samtools} index realigned_{}Filtered.bam realigned_{}Filtered.bai" ::: ms01e ms02g ms03g ms04h*

# Create flagstat report for non-filtered and filtered BAMs

*parallel --tmpdir ${TempDir} --jobs 4 "${samtools} flagstat realigned_{}.bam > realigned_{}.BAM.report.txt" ::: ms01e ms02g ms03g ms04h*

*parallel --tmpdir ${TempDir} --jobs 4 "${samtools} flagstat realigned_{}Filtered.bam > realigned_{}Filtered.BAM.report.txt" ::: ms01e ms02g ms03g ms04h*

# To Do- Now, compare the flagstat report to check how much reads got filtered (both number and percentage).

**8.7: Now, again call the variants from filtered RNAseq BAM files in GVCF mode**

Now, call Variants from RNAseq data in GVCF mode (using Parallel).

*parallel --jobs 4 java -jar -Xmx16g -Djava.io.tmpdir=${TempDir} ${GATK3p8} -T HaplotypeCaller -R ${RefGenome} -I realigned_${}_Filtered.bam --emitRefConfidence GVCF -o realigned_Filtered_{}.raw.snps.indels.g.vcf ::: ms01e ms02g ms03g ms04h*

**8.8 : Joint Genotyping by running GenotypeGVCFs on all the samples together**

https://software.broadinstitute.org/gatk/documentation/tooldocs/org_broadinstitute_gatk_tools_walkers_variantutils_GenotypeGVCFs.php

*java -jar -Xmx16g -Djava.io.tmpdir=${TempDir} ${GATK3p8} -T GenotypeGVCFs -R ${RefGenome} --variant MA605.raw.snps.indels.g.vcf --variant MA611.raw.snps.indels.g.vcf --*

*variant MA622.raw.snps.indels.g.vcf --variant MA625.raw.snps.indels.g.vcf --variant MA629.raw.snps.indels.g.vcf --variant Ncm8.raw.snps.indels.g.vcf --variant Sp3.raw.snps.indels.g.vcf --variant Sp21.raw.snps.indels.g.vcf --variant Sp76.raw.snps.indels.g.vcf --variant Sp154.raw.snps.indels.g.vcf --variant Sp164.raw.snps.indels.g.vcf --variant SpNor33.raw.snps.indels.g.vcf --variant realigned_Filtered_ms01e.raw.snps.indels.g.vcf --variant realigned_Filtered_ms02g.raw.snps.indels.g.vcf --variant realigned_Filtered_ms03g.raw.snps.indels.g.vcf --variant realigned_Filtered_ms04h.raw.snps.indels.g.vcf -o raw_variants.Set02.AllSamples.MySpF1.vcf*

**8.9 : Now, separate RNAseq Data Variants from Genome Variants.**

# useful flags: "-sn" for selecting required Samples

# "--removeUnusedAlternates": remove Un-used Alternate alleles

# remove sites that are uncalled (GT = ./.) after sample selection

# - use flag "'vc.getCalledChrCount() == 0' will select the sites that was noCall (./.) in all the samples

# - -invertSelect" will then invert this selection and also include the variants where at least variant in one sample is called (including GT = 0/0).

# SelectVariants from RNAseq Samples

*java -jar -Djava.io.tmpdir=${TempDir} ${GATK3p7} -T SelectVariants -R lyrata_genome.fa -V raw_variants.Set02.AllSamples.MySpF1.vcf --removeUnusedAlternates -o RNAseq_Samples.raw_Variants.vcf -sn ms01e -sn ms02g -sn ms03g -sn ms04h -select 'vc.getCalledChrCount() == 0' -invertSelect*

# Now, convert the these variants as PASS

*head -n -0 RNAseq_Samples.raw_Variants.vcf |tee >(grep '^#' > header.txt) >(grep '^#' -v | awk '$7 = "PASS" {print $0}' OFS="\t" > passed.txt)*

**Supplementary Materials S4.C: RNASEQ PIPELINE**

NOTE: The original bash scripts and pipeline for this procedure is available at

- https://github.com/everestial/phase-stitcher

- https://github.com/everestial/ASE-CADG

Data for this pipeline were sequenced at David Murdoch Center. It has four samples. Each sample has 100 bp paired-end reads.

***Initial Step*: Data check and adapter trimming.**

**Quality check** – using *fastqc* application (as in **Supplementary Materials S4.A** that was done for genome sequence reads)

**Adapters trimming** - using *trimmomatic* application.

*java -jar / apps/Trimmomatic-0.35/trimmomatic-0.35.jar PE -phred33 2ms01e_R1.fastq 2ms01e_R2.fastq Trm_PE-2ms01e_R1.fastq Trm_SE-2ms01e_R1.fastq Trm_PE-2ms01e_R2.fastq Trm_SE-2ms01e_R2.fastq ILLUMINACLIP:all_all_adapters_primers-RNAseq_updated02.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36*

**Step 01**: This step is used to align RNAseq reads to the haploid reference genome, transcriptome using rnaSTAR software; then extract the SNPs, InDels which will be further utilized in phasing.

We map RNAseq reads to the reference genome using the 2-pass approach.

**1.1 : Create index using reference genome and GTF/GFF align (1st pass)**

*mkdir 1sTpassGenomeDir*

*STAR --runThreadN 8 --runMode genomeGenerate --genomeDir 1sTpassGenomeDir --genomeFastaFiles lyrata_genome.fa --sjdbGTFfile lyrata.ensemble.gtf --sjdbOverhang 100*

Outcome: 1) makes index files for given reference genome.

Ensure that the input *gtf* file is compatible with the reference genome.

**1.2: First pass alignment directory for each RNAseq sample**

*mkdir 1passAlignment*

# sample 2ms01e

*mkdir 1passAlignment/2ms01e*

*STAR --runThreadN 16 --runMode alignReads --genomeDir 1sTpassGenomeDir --readFilesIn Trm_PE-2ms01e_R1.fastq Trm_PE-2ms01e_R2.fastq --outFileNamePrefix 1passAlignment/2ms01e/2ms01e --outFilterMultimapNmax 10 --outSAMmapqUnique 60 --outSAMtype BAM SortedByCoordinate --outReadsUnmapped Fastx --outSAMattributes All --alignIntronMin 10 --quantMode TranscriptomeSAM GeneCounts*

# Comment: --outSAMmapqUnique Integer0to255 makes the SAM file compatible with GATK; set it at 60.

# With --quantMode GeneCounts option STAR will count number of reads per gene while mapping. A read is counted if it overlaps (1nt or more) one and only one gene. This option requires annotations (GTF or GFF with –sjdbGTFfile option) at the genome generation step, or at the mapping step. STAR outputs read counts per gene into ReadsPerGene.out.tab file with 4 columns which correspond to different strandedness options. - see STAR manual for details.

# With --quantMode TranscriptomeSAM will also output the reads aligned to transcriptome reads (check for  Aligned.toTranscriptome.out.bam file - this bam file can be used for identification of diagnostic alleles), this flag combined with --quantTranscriptomeBan

IndelSoftclipSingleend (default) prohibit indels, soft clipping and single-end alignments - to make it compatible with RSEM (and probably EMASE).

**Step 02. Add read groups, sort, mark duplicates, and create index**

*java -jar / apps/picard-tools-2.5.0/picard.jar AddOrReplaceReadGroups I=2ms01eAligned.sortedByCoord.out.bam O=2ms01e_sorted.bam SO=coordinate RGID=TTAGGC RGLB=F1_hybrid RGPL=Illumina RGPU=C4AM3ACXX RGSM=2ms01e CREATE_INDEX=true*

*java -jar / apps/picard-tools-2.5.0/picard.jar MarkDuplicates I=2ms01e_sorted.bam O=2ms01e_dedupped.bam CREATE_INDEX=true VALIDATION_STRINGENCY=SILENT M=2ms01e_output.metrics*

**Step 03 - note: Using GATK app requires creating a sequence dictionary file for** reference genome.

*java -jar -Xmx16g -Djava.io.tmpdir= /temp_files/ /apps/picard-tools-2.5.0/picard.jar CreateSequenceDictionary R=lyrata_genome.fa O=lyrata_genome.dict*

**# Step 03.1: Split'N'Trim and reassign mapping qualities**

*java -jar -Xmx16g -Djava.io.tmpdir= /temp_files/ apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T SplitNCigarReads -R lyrata_genome.fa -I 2ms01e_dedupped.bam -o 2ms01e_split.bam -U ALLOW_N_CIGAR_READS*

**Step 03.2: Variant Calling**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T HaplotypeCaller -R lyrata_genome.fa -I 2ms01e_split.bam -dontUseSoftClippedBases -stand_call_conf 20.0 -stand_emit_conf 20.0 -o 2ms01e_raw.vcf*

# Variant Filtration

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T VariantFiltration -R lyrata_genome.fa -V 2ms01e_raw.vcf -window 35 -cluster 3 -filterName FS -filter "FS > 30.0" -filterName QD -filter "QD < 2.0" -filterName MappingQ -filter "MQ < 25.0" -o 2ms01e_filtered.vcf*

**Step 04: IndelRealigment**

## Realignment around the indels: Evidence of hidden indels > 1) presence of mismatches 2) softclips. InDels in reads (especially near the ends) can trick the mappers into mis-aligning with mismatches. Note: Know indels sites may be supplemented for the preparing the realignment target site (but not required).

**4.1 Run the RealignmentTargetCreator to find the positions on the chromosome that may require realiment**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T RealignerTargetCreator -R lyrata_genome.fa -I 2ms01e_split.bam -o 2ms01e_realigner.intervals*

#Note: this run a) requires index file for both input *.bam as well as the reference *.fasta files, so check for the index files. b) creates a intervals.list for futher downstream realignment.

**4.2 And then run IndelRealigner for realignment of the reads around the target indels to correct for any snp artifactuals. Local realignment helps to identify most parsimonious alignment.**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T IndelRealigner -R lyrata_genome.fa -I 2ms01e_split.bam -*

*targetIntervals 2ms01e_realigner.intervals -o realigned_2ms01e.bam --consensusDeterminationModel USE_SW*

#//This creates a file called *realigned_readsMA605.bam* containing all the original reads, but with better local alignments in the regions that were targeted using interval file 2ms01e_realigner.intervals. We are using --*consensusDeterminationModel USE_SW* to generate alternate consensus using 'Smith-Waterman'; this model requires lots of computational power but is mostly accurate.

**Step 05: Use phaser for creating RBP blocks.**

```
vcf_file=passed_variants.All_samples.vcf.gz
# Note: use gziped files for phASER
# index the bam files
samtools index "realigned_${item}.bam"
# make a sub-directory on the fly
rm -rf ${OPATH}/"${item}_phased"; mkdir ${OPATH}/"${item}_phased"
```

**5.1 now, run phASER for each sample on a for-loop**

```
python phaser.py --threads 2 --vcf ${vcf_file} --bam "realigned_${item}.bam" --paired_end 1 --mapq 20 --baseq 10 --sample "$item" --o ${OPATH}/"${item}_phased"/"${item}_Only_Chr2" --id_separator - --haplo_count_bam 1 --chr 2 --write_vcf 1 --as_q_cutoff 0.025 --include_indels 1 --unique_ids 1 --output_network "${item}_network" --show_warning 1 --debug 1 &> ${OPATH}/"${item}_phased"/"${item}.debug.log"
```

**5.2 Create path, variables and fileName for corresponding files.**

*vcf_file=passed_variants.All_samples.vcf.gz  # Note: use gziped files for phASER*

*echo "Read the vcf file for all samples (:"*

# main output Dir

*rm -rf PHASER_OUTPUT_GenomeData; mkdir PHASER_OUTPUT_GenomeData  # creates an output dir if it doesn't exist*

*OPATH=PHASER_OUTPUT_GenomeData*

*echo "Create main output dir named: '${OPATH}' (:"*

# Run phASER for each genome samples/individuals

*samtools index "${item}.deDuplicated.indel_realigned.bam"*

# make sub-directory on the fly, if there is already a directory with that name (will be removed) -

*rm -rf ${OPATH}/"${item}_phased"; mkdir ${OPATH}/"${item}_phased"*

# now, run phASER for each sample on a for-loop

*python phaser.py --threads 2 --vcf ${vcf_file} --bam "${item}.deDuplicated.indel_realigned.bam" --paired_end 1 --mapq 20 --baseq 10 --sample "$item" --o ${OPATH}/"${item}_phased"/"${item}_Only_Chr2" --id_separator - --haplo_count_bam 1 --chr 2 --write_vcf 1 --as_q_cutoff 0.025 --include_indels 1 --unique_ids 1 --output_network "${item}_network" --show_warning 1 --debug 1 &> ${OPATH}/"${item}_phased"/"${item}.debug.log"*

**5.3 (optional )**

### """ Description:  This is just a alternative method

## this is run to

# - use reference Genome to impute the phased variants

# - impute homVar (all samples) found in two distince populations

# - create chain files, and update GTF/GFF file

## prepared to align the RNAseq reads only from Chr2 to diploid Chr2

## This can be fully extended to include data analyses for all chromosomes and full genome updated GTF/GFF

**5.4: Phase Stitching**

*python pHASE-Stitcher-Markov/markov_final_test/Stitcher_using_1stOrderMarkov_InteractiveMode.py --vcf1 My_seq.test.vcf --vcf2 Sp_seq.test.vcf --pop1 My --pop2 Sp --output 2ms04h_test --het_vcf RNA_seq.test.vcf --f1_sample 2ms04h*

*python stitcher/pHASE-Stitcher-Markov/markov_final_test/Stitcher_using_1stOrderMarkov_InteractiveMode.py --vcf1 MY.phased_variants.Final.vcf --vcf2 SP.phased_variants.Final.vcf --pop1 My --pop2 Sp --output ${item}_Chr2 --het_vcf RNAseq.phased_variants.Final.vcf --f1_sample ${item}*

**5.5: Find fix HomVar(GT) in both population**

# This variants can be use for imputing the variants in phased genome

## Part A: - select vcf site/lines that are fixed homozygous variants (GT = 1/1 or 2/2 ...)

## Note: involves two runs for each vcf

## first to select allHom Ref or Var and second to select sameAllHomVar in all samples

## for Mayodan

*java -jar /apps/GenomeAnalysisTK-3.7/GenomeAnalysisTK.jar -T SelectVariants -R lyrata_genome.fa -V MY.phased_variants.Final.vcf -o My.AllHomVar.vcf -select 'vc.getHomVarCount() == 6'*

# change sample size (6) if multiple vcf with different size is used

# now select the lines that are same HomVar in all samples

*java -jar / /apps/jvarkit/dist/vcffilterjs.jar -e 'function accept(vc) {for(var i=1;i<vc.getNSamples();i++) if(!vc.getGenotype(0).sameGenotype(vc.getGenotype(i))) return false; return true;}accept(variant); ' My.AllHomVar.vcf -o My.AllHomVarSameGT.vcf*

## same selection for SpiterStulen

*java -jar / apps/GenomeAnalysisTK-3.7/GenomeAnalysisTK.jar -T SelectVariants -R lyrata_genome.fa -V SP.phased_variants.Final.vcf -o Sp.AllHomVar.vcf -select 'vc.getHomVarCount() == 6'*

# now select the lines that are same HomGT in all samples

*java -jar / apps/jvarkit/dist/vcffilterjs.jar -e 'function accept(vc) {for(var i=1;i<vc.getNSamples();i++) if(!vc.getGenotype(0).sameGenotype(vc.getGenotype(i))) return false; return true;}accept(variant); ' Sp.AllHomVar.vcf -o Sp.AllHomVarSameGT.vcf*

**5.6: imputation status complete for Chromosome #2**

## Now, take the Haplotype Portrait file and run with

# 1) impute_F1_genotypes.py

# code for mining HomVar(GT_bases) - One run works for all(mostly) samples

# check codes

# 2) impute_F1_genotypes_part02_withPandas.py - still not an interactive program

# run for each sample separately - careful with samples names !! ??

**Step 06: Create a diploid genome**

This is done to:

## prepared to align the RNAseq reads only from Chr2 to diploid Chr2

## This can be fully extended to include data analyses for all chromosomes and full genome updated GTF/GFF

**6.1 Preparation of Diploid genome - Using g2gTools :**

## But, this method has a slight modification - in SNP haplotype file

## Modification in prepraration of left vs. right SNP_InDelsPatched Genome (see explanation)

**6.1.1: gzip and index the vcf file**

# Only take Chr2 - Remove this step later

# Split using pyfaidx

*faidx -x lyrata_Chr2.My.fa 2*

# outcome: This will output only chr2 in the fasta file

# SNPs

*bgzip -c F1_2ms04h.imputed.SNP.haplotype.vcf > F1_2ms04h.imputed.SNP.haplotype.vcf.gz*

*tabix -p vcf F1_2ms04h.imputed.SNP.haplotype.vcf.gz*

# InDels

*bgzip -c F1_2ms04h.imputed.InDel.haplotype.vcf > F1_2ms04h.imputed.InDel.haplotype.vcf.gz*

*tabix -p vcf F1_2ms04h.imputed.InDel.haplotype.vcf.gz*

**6.2 : Create a chain-file using InDels**

# using lyrata_genome(1-10).fa

# using --diploid option active

# A: Create chain file using InDels

*g2gtools vcf2chain -f ${REF} -i ${VCF_INDELS} -s ${STRAIN} -o ${STRAIN}/REF-to-${STRAIN}.chain*

*g2gtools vcf2chain -f lyrata_Chr2.only.fa -i F1_2ms04h.imputed.InDel.haplotype.vcf.gz -s 2ms04h -o REF_to_2ms04h.chain –diploid*

# Outcome: This will create left - REF_to_2ms04h.left.chain and right - REF_to_2ms04h.right.chain chainfiles.

# If --diploid mode is not active, HetVar will be tossed while creating chain files

# B: Patch the SNPs

g2gtools patch -i lyrata_Chr2.only.fa -v F1_2ms04h.imputed.SNP.haplotype.vcf.gz -s 2ms04h -o lyrata_2ms04h_SNPsPatched.fa --diploid

# Outcome: This will create left - lyrata_2ms04h_SNPsPatched.l.fa and right - lyrata_2ms04h_SNPsPatched.r.fa genome

# If --diploid mode is not active HetVar will be tossed while creating genome

# C: Incorporate Indels

*g2gtools transform -i ${STRAIN}/${STRAIN}.patched.fa -c ${STRAIN}/REF-to-${STRAIN}.chain -o ${STRAIN}/${STRAIN}.fa*

# Note: So, no we plan to add InDels to SNPs patched genome

# So, use left-SnpPatchedGenome and left-chainfile to add the left-indels to create left-SnpInDelPatchedGenome

# prep Left-Genome:

*g2gtools transform -i lyrata_2ms04h_SNPsPatched.l.fa -c REF_to_2ms04h.left.chain -o lyrata_2ms04h_SNPs_InDelsPatched.left.fa -d*

# prep Right-Genome:

*g2gtools transform -i lyrata_2ms04h_SNPsPatched.r.fa -c REF_to_2ms04h.right.chain -o lyrata_2ms04h_SNPs_InDelsPatched.right.fa -d*

**6.3: Create custom gene-annotation (i.e update GTF/GFF file using chain-file)**

# Now, we create custom gene annotation with respect to the new custom genome.

# We can also create custom annotation database (so we can extract from custom genome) in the following steps:

*g2gtools convert -c ${STRAIN}/REF-to-${STRAIN}.chain -i ${GTF} -f gtf -o ${STRAIN}/${STRAIN}.gtf*

# GTF for Left-Genome:

*g2gtools convert -c REF_to_2ms04h.left.chain -i lyrata.Chr2_only.gtf -f gtf -o GTF_for2ms04h.Left.gtf -d*

# GTF for Left-Genome:

*g2gtools convert -c REF_to_2ms04h.right.chain -i lyrata.Chr2_only.gtf -f gtf -o GTF_for2ms04h.Right.gtf -d*

# We can also use gff file - Using this for our ASE data analyses

# GFF for Left-Genome:

*g2gtools convert -c REF_to_2ms04h.left.chain -i lyrata.Chr2.fromRawat_02.gff -f gtf -o GFF_for2ms04h.Left.gff*

# GFF for Right-Genome:

*g2gtools convert -c REF_to_2ms04h.right.chain -i lyrata.Chr2.fromRawat_02.gff -f gtf -o GFF_for2ms04h.Right.gff*

**6.4 Add pre/suf-fix to the Diploid Genome/GTF,GFF files**

# First rename the chormosomes - Do manually

*lyrata_2ms04h_SNPs_InDelsPatched.left.fa > lyrata_2ms04h_Chr2.My.fa # left chromosome*

*lyrata_2ms04h_SNPs_InDelsPatched.right.fa > lyrata_2ms04h_Chr2.Sp.fa # right chromosome*

# Rename the GTF/GFF files too

*GFF_for2ms04h.Left.gff > lyrata_2ms04h_Chr2.My.gff*

*GFF_for2ms04h.Right.gff > lyrata_2ms04h_Chr2.Sp.gff*

**6.5 Add appropriate pre/suffixes in the fasta and GTF files**

# suffix added manually for fasta files

# add suffix in GTF/GFF files

""" prefix was added to the gtf file, column 1, where the chromsome label if present

From:

2 version-2 gene 1 1011 0.42 - .
ID=AL2G10010;Name=AL2G10010;Note=Protein_Coding_gene

2 version-2 transcript 1 1011 0.42 - .
ID=AL2G10010.t1;Parent=AL2G10010

To:

2_My version-2 gene 1 1011 0.42 - .
ID=AL2G10010;Name=AL2G10010;Note=Protein_Coding_gene

2_My version-2 transcript 1 1011 0.42 - .

ID=AL2G10010.t1;Parent=AL2G10010

"""

**6.6 Add suffix in column #1 (chr names) in gtf files**

# using Sed tool:

*#sed 's/^\S*/&_My/' lyrata_2ms04h_Chr2.My.gff > lyrata_2ms04h_Chr2_My.gff*

*#sed 's/^\S*/&_Sp/' lyrata_2ms04h_Chr2.Sp.gff > lyrata_2ms04h_Chr2_Sp.gff*

# using Awk

*awk -v PRE='_My' '{$1=$1PRE; print}' OFS="\t" lyrata_2ms04h_Chr2.My.gff > lyrata_2ms04h_Chr2_My.gff*

*awk -v PRE='_Sp' '{$1=$1PRE; print}' OFS="\t" lyrata_2ms04h_Chr2.Sp.gff > lyrata_2ms04h_Chr2_Sp.gff*

## also add suffix to all the variable_names in gene features (column 9)

*awk -v PRE='_My' '{gsub(/;/,PRE";",$9); sub(/$/,PRE,$9); print}' OFS='\t' lyrata_2ms04h_Chr2_My.gff > lyrata_2ms04h_Chr2_test_My.gff*

*awk -v PRE='_Sp' '{gsub(/;/,PRE";",$9); sub(/$/,PRE,$9); print}' OFS='\t' lyrata_2ms04h_Chr2_Sp.gff > lyrata_2ms04h_Chr2_test_Sp.gff*

## Now, merge the genome and GTF/GFF files

*cat lyrata_2ms04h_Chr2.My.fa lyrata_2ms04h_Chr2.Sp.fa > lyrata_2ms04h_Chr2.MySp.fa*

*cat lyrata_2ms04h_Chr2_test_My.gff lyrata_2ms04h_Chr2_test_Sp.gff > lyrata_2ms04h_Chr2_MySp.gff*

**Step 07: Alignment to Diploid Genome, GTF/GFF using STAR in TrancriptomeSAM, Gene Counts Mode**

## Set variable names and corresponding files to it.

*samtoolsV3= /Samtools1.3/Samtools1.3*

*ref_genome=lyrata_genome.fa*

# output Directory

*rm -rf BAM_To_FastQ; mkdir BAM_To_FastQ OPATH=BAM_To_FastQ*

*samtools index realigned_${item}.bam*

# select alignment reads only from Chr2

*samtools view -h -b realigned_${item}.bam 2 > "${OPATH}/chr2.${item}.bam"*

# sort the bam file

*samtools sort -n "${OPATH}/chr2.${item}.bam" "${OPATH}/chr2.sorted.${item}"*

# now, BAM to FastQ conversion

*/Samtools1.3/Samtools1.3 fastq -1 ${OPATH}/chr2_R1_${item}.fastq -2 ${OPATH}/chr2_R2_${item}.fastq -O -s ${OPATH}/chr2.single_${item}.fastq -t ${OPATH}/chr2.sorted.${item}.bam --reference ${ref_genome}*

**Step 08: Variant Calling**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T HaplotypeCaller -R lyrata_genome.fa -I realigned_2ms01e.bam -I realigned_2ms02g.bam -I realigned_2ms03g.bam -I realigned_2ms04h.bam -dontUseSoftClippedBases -stand_call_conf 20.0 -stand_emit_conf 10.0 -o F1_hybrids_raw.vcf -nt 1 -nct 8*

**8.2 Variant Filtration**

**# Select Variants (SNPs)**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T SelectVariants -R lyrata_genome.fa -V F1_hybrids_raw.vcf -selectType SNP -o F1_hybrids_SNPs_raw.vcf*

**# Filter SNPs**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T VariantFiltration -R lyrata_genome.fa -V F1_hybrids_SNPs_raw.vcf --filterExpression "AS_FS > 60.0" --filterName "AS_FS_fail" --filterExpression "AS_QD < 2.0" --filterName "AS_QD_fail" --filterExpression "AS_MQRankSum < -12.5" --filterName "AS_MQRankSum_fail" --filterExpression "AD < 3" --filterName "AlleleDepth_fail" -o F1_hybrids_SNPs_filtered.vcf*

## The filtration of variants (cluster of 3 SNPs) in a window of 35 bp was not applied, because in the reads from Spiterstulen haplotype we were expecting there would be more SNPs cluster within the window of 35 bp.

# Instead filtration was done by adding one more filtering parmeter AS_MQRankSum

# See the link for AS_MQRankSum https://software.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_annotator_AS_MappingQualityRankSumTest.php

**# Select passed variants (SNPs)**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T SelectVariants -R lyrata_genome.fa -V F1_hybrids_SNPs_filtered.vcf -o F1_hybrids_SNPs_passed.vcf -select 'vc.isNotFiltered()'*

**8.3 (InDels): Variant Filtration**

# Select Variants (InDels)

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T SelectVariants -R lyrata_genome.fa -V F1_hybrids_raw.vcf -selectType INDEL -o F1_hybrids_InDels_raw.vcf*

# Filter InDels

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T VariantFiltration -R lyrata_genome.fa -V F1_hybrids_InDels_raw.vcf --filterExpression "AS_FS > 200.0" --filterName "AS_FS_fail" --filterExpression "AS_QD < 2.0" --filterName "AS_QD_fail" --filterExpression "AS_MQRankSum < -20" --filterName "AS_MQRankSum_fail" --filterExpression "AD < 3" --filterName "AlleleDepth_fail" -o F1_hybrids_InDels_filtered.vcf*

# Select passed variants (InDels)

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T SelectVariants -R lyrata_genome.fa -V F1_hybrids_InDels_filtered.vcf -o F1_hybrids_InDels_passed.vcf -select 'vc.isNotFiltered()'*

**8.4 (SNPs and InDels): Merge Variants**

*java -jar -Xmx16g -Djava.io.tmpdir=/temp_files /apps/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T CombineVariants -R lyrata_genome.fa --variant:SNPs F1_hybrids_SNPs_passed.vcf --variant:InDels F1_hybrids_InDels_passed.vcf -o F1_passed_variants.vcf*

**Table S4.C1 RNAseq data alignment metrics.**

| *Sample* | # of Reads | % Unique | %Multi | %Too many | %Unmapped |
|---|---|---|---|---|---|
| 2ms01e | 2549672 | 42.53 | 56.59 | 0.12 | 0.76 |
| 2ms02g | 2757392 | 38.76 | 60.47 | 0.08 | 0.68 |
| 2ms03g | 2352049 | 42.01 | 57.29 | 0.08 | 0.62 |
| 2ms04h | 2734175 | 42.17 | 56.98 | 0.09 | 0.76 |

*Note:* This alignment metrics are for reads aligned on a diploid genome (using RnaSTAR) for each sample. The "% Unique columns" shows the percentage of uniquely aligned reads and other columns shows percentage of reads aligned in that category.

**Supplementary Materials S4.D: Codes Used for ASE Analysis With DESeq2**

(1) [R script] Upload the DESeq2 package

*library(DESeq2)*

(2) Loading data

*counts_2ms01e <- read.table( "final_counts_2ms01e.txt", sep="\t", header = TRUE)*

*counts_2ms02g <- read.table("final_counts_2ms02g.txt", sep="\t", header = TRUE)*

*counts_2ms03g <- read.table("final_counts_2ms03g.txt", sep="\t", header = TRUE)*

*counts_2ms04h <- read.table("final_counts_2ms04h.txt", sep="\t", header = TRUE)*

(3) Merge/join Dataframes

counts.merged.ms1e2g3g4h <- merge(

counts_2ms01e, counts_2ms02g,

```
  by=c("contig", "start", "transcript_ID")) %>%
  merge(counts_2ms03g,
      by=c("contig", "start", "transcript_ID")) %>%
  merge(counts_2ms04h,
      by=c("contig", "start", "transcript_ID"))
```

(4) Apply filters

```
counts.ms1e2g3g4h.Abv10Counts <- filter(
  counts.merged.ms1e2g3g4h,
  unqC_My.ms01e>10 |unqC_Sp.ms01e>10 |unqC_My.ms02g>10
|unqC_Sp.ms02g>10  |unqC_My.ms03g>10 |unqC_Sp.ms03g>10
|unqC_My.ms04h>10 |unqC_Sp.ms04h>10)
```

(5) Sequence alignment of Spiterstulen population of PIN3 gene

```
counts.ms1e2g3g4h.Summed <- cbind(
  counts.ms1e2g3g4h.Abv10Counts,
  unqC_total = counts.ms1e2g3g4h.Abv10Counts %>%
select(matches("unqC"))%>%
    rowSums(),
  mulC_total =
    counts.ms1e2g3g4h.Abv10Counts %>%
    select(matches("mulC")) %>%
    rowSums(),
  totalC_total = counts.ms1e2g3g4h.Abv10Counts %>%
    select(matches("totalC")) %>%
```

```
    rowSums())
```

(6) Apply filters

```
counts.ms1e2g3g4h.filt01 <- filter(
  counts.ms1e2g3g4h.Summed,
  mulC_total < .20*totalC_total |
    unqC_total > .20*totalC_total) %>%
  select(-unqC_total, -mulC_total, -totalC_total) %>%
  arrange(start, contig)
```

(7) Design the expression data

```
expressionData <- data.frame(
  counts.ms1e2g3g4h.filt01, row.names = "transcript_ID") %>%
  select(matches("unqC", ignore.case = TRUE))
```

(8) Design covariate factors for the test

```
# sample level variates
sampleID <- factor(rep(c("ms01e", "ms02g", "ms03g", "ms04h"), each = 2))
      # haplotype level variates
hapType <- factor(rep(c("My", "Sp"), 4))
     # maternal cytoplasm level variates
cytoplasm = factor(rep("Ma", 8))
# family level variates
familyGroup <- factor(rep(c("e", "g", "g", "h"), each=2))
## create a dataframe detailing the above co-variate:sample relationship
exp_factors = data.frame(sampleID = sampleID,
```

```
                    hapType = hapType,
                    familyGroup = familyGroup,
                    maternalEff = cytoplasm)
    # experimental design set up for ASE
    expDesign <- model.matrix(~0 + sampleID + hapType)
```

(9) Provide count data as matrix, colData and design

```
    ASE_Matrix <- DESeqDataSetFromMatrix(countData = expressionData,
                                colData = exp_factors,
                                design = ~ 0 + sampleID + hapType)
```

(10) Set normalization factors for all columns to 1

```
    sizeFactors(ASE_Matrix) <- rep(1, 2*4)
```

(11) Run the DESeq using the design - fit the model using "local"

```
    dds.ASE_Data <- DESeq(ASE_Matrix, fitType = "local", betaPrior = FALSE)
```

(12) Build the results table

```
    # Without threshold
    result.ASE_Data <- results(
      dds.ASE_Data, contrast = c("hapType", "My", "Sp"),
      alpha = 0.05)
    # With threshold
    result.ASE_Data.LFC1 <- results(dds.ASE_Data, lfcThreshold = 1,
                            contrast = c("hapType", "My", "Sp"))
```

**Supplementary Materials S4.E: Files repo**

1. File showing significant ASE differences (sorted by *P*-values) is on Github repo https://github.com/everestial/ASE-CADG , file name "wald_all_genes_*P*Value_byValues.csv"
2. File showing significant ASE differences (sorted by genomic position) is on Github repo https://github.com/everestial/ASE-CADG , file name "wald_all_genes_*P*Value_byPosition.csv"

# CHAPTER V: CONCLUSION

## Dissertation Goal 1; Chapter II

My first goal (Chapter 2) was to determine whether differences in apical dominance and shoot architecture observed in the Mayodan and Spiterstulen populations can be explained by variation in the rate of auxin transport.

I tested variation in auxin transport rates in our two study populations using radiolabelled 3H-IAA (a synthetic auxin) in the inflorescence shoots. I found weak evidence of variation in the rate of auxin transport between populations; in the predicted direction, with Mayodan individuals showing higher auxin transport. The results might have been confounded by the variation in the diameter of inflorescences as Spiterstulen individuals mostly have a thicker diameter which is visually apparent but which we did not measure. We would expect the amount of transport to increase with diameter, so the actual differences in the transport rate might have been more significant if those were taken into account.

In Experiment II (auxin inhibition assay), I found weak evidence that auxin transport inhibition affects life-history traits, i.e., NPA treatment reduces the diameter and increases lateral shoot rating. These are the expected directions if auxin transport differences cause life-history differences between our study populations. While the effects of NPA on inflorescence numbers were not significant, the control group had a higher number of inflorescences during the reproductive period, as expected.

Overall, the evidence is not strong, but it does point to the direction as predicted that auxin transport is a likely candidate in shaping life-history differences between our study populations. The plants treated with NPA also had high mortality, thus reducing the test's statistical power. This mortality could be due to altered auxin dynamics causing direct toxic effects of NPA or the ecological consequences of the tradeoff or NPA affecting some

developmental pathways. In the NPA treatment group, we observed a delay in the apparent effects of transport inhibition on lateral shoot rating and the number of inflorescences, with effects showing up three months after discontinued NPA treatment. The findings are consistent with the idea that variation in traits such as apical dominance in early development can cascade through later developmental stages, changing the entire trajectory of life history. If the NPA treatment could have been continued without any adverse effects based on the dose of NPA, the measured differences in life-history traits between the two groups might have become greater. This provides evidence, although tentative, that genetic variation affecting auxin transport could underlie adaptive variation in life history in *A. lyrata*.

Future studies measuring transport differences with larger biological and technical replicates and the measurements of the diameter of the inflorescences (as a covariate) can provide more clarity on this issue. In addition, further analyses of transport differences can be done at population levels using auxin pulse-chase assays. Other studies involving optimized doses of NPA treatment can provide more robust insights into the role of auxin transport on life-history traits. Additionally, genetic insertion of My alleles (for auxin transport-related genes) on the Spiterstulen genotype background and vice-versa can test whether auxin transport genes result in life-history variation.

**Dissertation Goal 2; Chapter III**

My second goal (Chapter 3) was to develop algorithms and tools for phasing and assigning haplotypes in outcrossing populations.

I developed three different algorithms and methods to help with phasing haplotypes for unphased genotype and read-backed-phased genetic variants data. The three tools/algorithms are Phase-Extender, Phase-Stitcher, and ShortVariantPhaser and are designed to handle three

different types of data structure generated in concurrent variant genotyping. Developing these resources was essential for phasing RNA-seq reads in our F1 samples since we didn't have the parents' genome or transcriptome sequence data.

The results comparing Phase-Extender and ShapeIT showed that Phase-Extender could phase variants on par with ShapeIT, even using fewer reference panels. Additionally, Phase-Extender can help samples phase each other and provide a more controlled approach to haplotype phasing. Therefore, I expect that it has the potential to be useful for other investigators who need to phase genomes that do not have a large number of reference haplotypes or in the situation when a small number of sample cohorts are only to be used for haplotype phasing.

**Dissertation Goal 3; Chapter IV**

My third goal (Chapter 4) was to identify candidate genes underlying a key life-history QTL region by evaluating quantitative variation in the expression of alleles from Mayodan and Spiterstulen genomes.

This study found that the genes *PIN3* and *PILS2* show significant ASE in the predicted direction (My > Sp) among the few hypothesized candidates. Our research gives some interesting insights; we see strong expression of Mayodan alleles of *PIN3* and *PILS2* from the resource allocation QTL regions, while *PIN1* shows almost equal bi-allelic expression, with the Sp allele showing only a little higher expression. On the other hand, the expression level of *BRC2* was very low. The expression of the *BRC2* gene might be limited to certain tissues, especially meristems, because their specific role is to arrest the growth of the meristems. Also, the extraction of mRNA from the whole shoot could have diluted their expression levels, so we can not rule out a role for *BRC2* for these reasons. A few other genes from the QTL region also showed significant ASE, but none are obvious candidates based on their annotations.

The role of PINI3 and PILIS2 on life history can be further tested for their effects on life-history traits by transgenically exchanging My alleles (for gene *PIN3, PILS2*) onto the Sp genotypes and vice-versa and studying their effects on life-history traits. Another way of testing for the effects of these genes/alleles would be to generate CRISPR knock-outs and study their phenotypic effects. Another graduate student in the lab has developed CRISPR constructs for *PIN3* and *BRC2*, which will help test the role of population-specific alleles on life-history traits. However, based on the results, I recommend that *PILS2* be added to the list. Furthermore, additional genomic and transcriptomics studies can also be designed to test whether NPA affects gene expression variation in the parental populations, mainly Mayodan. Finally, other experiments involving verification of ASE expression could be done in the parental population using qPCR, which was one of the original goals of this chapter but was missed due to technical difficulties. One of the technical difficulties was that we were not successful in growing two parental populations and having them flower simultaneously. And, since A. lyrata is a perennial, they take about six months to reach that stage of required biological condition, unlike A. thaliana, which are ready in about one month.

Overall, the results from auxin transport/inhibition and ASE analyses converge on support for genetic variation in auxin transport as a mechanism underlying adaptive life-history variation. Furthermore, the results indicate *PIN3* as a likely candidate driving life history differences in our study population, which are strongly adapted to contrasting climatic environments. Although none of this is conclusive, it all points in the same direction. The gene *PIN3* is interesting because previous studies have not implicated auxin transporters as adaptive QTLs underlying life-history variation. In most cases, *PIN1* is the gene that is tested and researched for its role in apical dominance, not *PIN3*.

**Resource allocation is an "integrated complex phenotype."**

I want to reiterate that resource allocation can be understood as an "integrated complex phenotype" (mentioned previously in Chapter-I) where several fitness-related components of an organism integrate. While these are quantitative in nature, they form an integrative life-history pattern characterized by resource allocation tradeoffs and play a central role in adaptation to different climates. We emphasize that meaningful insights about variation in iteroparous plants require understanding fitness from a developmental perspective, where limited time for resource acquisition during a particular season and optimal investment of this resource pool to different functions along the life cycle is crucial for organisms' fitness. Similar integrated phenotypes appear to be essential for adaptation in other perennial plants (Gove et al., 2012; Kim & Donohue, 2011, 2012; Leinonen et al., 2012; Remington et al., 2015; Wang et al., 2009).

CHAPTER VI: REFERENCES

## References Chapter I


Aguilar-Martínez, J. A., Poza-Carrión, C., Cubas, P., Aguilar-Martínez, J. A., Poza-Carrión, C., & Cubas, P. (2007). Arabidopsis BRANCHED1 acts as an integrator of branching signals within axillary buds. *The Plant Cell*, *19*(2), 458–472. https://doi.org/10.1105/tpc.106.048934

Anderson, J. T., Willis, J. H., & Mitchell-Olds, T. (2011). Evolutionary genetics of plant adaptation. *Trends in Genetics*, *27*(7), 258–266. https://doi.org/10.1016/j.tig.2011.04.001

Baker, A. M., Burd, M., & Climie, K. M. (2005). Flowering phenology and sexual allocation in single-mutation lineages of Arabidopsis thaliana. *Evolution*, *59*(5), 970–978.

Barbez, E., Kubeš, M., Rolčík, J., Béziat, C., Pěnčík, A., Wang, B., Rosquete, M. R., Zhu, J., Dobrev, P. I., Lee, Y., Zažímalovà, E., Petrášek, J., Geisler, M., Friml, J., Kleine-Vehn, J., Zašímalovà, E., Petrášek, J., Geisler, M., Friml, J., & Kleine-Vehn, J. (2012). A novel putative auxin carrier family regulates intracellular auxin homeostasis in plants. *Nature*, *485*(7396), 119–122. https://doi.org/10.1038/nature11001

Bell, G. (1980). The costs of reproduction and their consequences. *American Naturalist*, 45–76.

Brown, J. S., & Venable, D. L. (1986). Evolutionary ecology of seed-bank annuals in temporally varying environments. *American Naturalist*, 31–47.

Callahan, H. S., Dhanoolal, N., & Ungerer, M. C. (2005). Plasticity genes and plasticity costs: a new approach using an Arabidopsis recombinant inbred population. *New Phytologist*, *166*(1), 129–140.

Charlesworth, B. (1971). Selection in density-regulated populations. *Ecology*, 469–474.

Charnov, E. L., & Schaffer, W. M. (1973). Life-history consequences of natural selection: Cole's result revisited. *American Naturalist*, 791–793.

Charnov, E. L., Schaffer, W. M., The, S., Naturalist, A., & Dec, N. N. (1973). The University of Chicago Life-History Consequences of Natural Selection : Cole ' s Result Revisited. *The American Naturalist*, *107*(958), 791–793.

Choi, M. S., Koh, E. B., Woo, M. O., Piao, R., Oh, C. S., & Koh, H. J. (2012). Tiller formation in rice is altered by overexpression of OsIAGLU gene encoding an IAA-conjugating enzyme or exogenous treatment of free IAA. *Journal of Plant Biology*, *55*(6), 429–435. https://doi.org/10.1007/s12374-012-0238-0

Clauss, M. J., & Koch, M. a. (2006). Poorly known relatives of Arabidopsis thaliana. *Trends in Plant Science*, *11*(9), 449–459. https://doi.org/10.1016/j.tplants.2006.07.005

Cohen, D. (1966). Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, *12*(1), 119–129.

Cole, L. C. (1954). The population consequences of life history phenomena. *Quarterly Review of Biology*, 103–137.

Cubas, P., Lauter, N., Doebley, J., & Coen, E. (1999). The TCP domain: A motif found in proteins regulating plant growth and development. *Plant Journal*, *18*(2), 215–222. https://doi.org/10.1046/j.1365-313X.1999.00444.x

Doebley, J., Stec, A., & Gustus, C. (1995). Teosinte Branched1. *Gene Expression*, *81*.

Doebley, J., Stec, A., & Hubbard, L. (1997). *The evolution of apical dominance in maize*.

Fisher, R. A. (1958). The genetical theory of natural selection. Рипол Классик.

Flatt, T., & Heyland, A. (2011). Mechanisms of Life History Evolution: The Genetics and Physiology of Life History Traits and Trade-Offs. *Genetics*, 540. http://books.google.com/books?hl=en&lr=&id=2SwjI6tPHPcC&oi=fnd&pg=PP2&dq=Mechanisms+of+Life+History+Evolution+The+Genetics+and+Physiology+of+Life+History+Traits+and+Trade-Offs&ots=O0-Y_D4sr_&sig=7gZXKqbL9xSaKeATNugGFbKS948

Friml, J. (2003). Auxin transport—shaping the plant. *Current Opinion in Plant Biology*, 7–12. https://doi.org/10.1016/S1369-5266(02)00003-1

Friml, J. J., Wiśniewska, J., Benková, E., Mendgen, K., Palme, K., Wiśniewska, J., Benková, E., Mendgen, K., & Palme, K. (2002). Lateral relocation of auxin efflux regulator PIN3 mediates tropism in Arabidopsis. *Nature*, *415*(6873), 806–809. https://doi.org/10.1038/415806A

Gallavotti, A. (2013). The role of auxin in shaping shoot architecture. *Journal of Experimental Botany*, *64*(9), 2593–2608. https://doi.org/10.1093/jxb/ert141

Gälweiler, L., Guan, C., Müller, A., Wisman, E., Mendgen, K., Yephremov, A., & Palme, K. (1998). Regulation of polar auxin transport by AtPIN1 in Arabidopsis vascular tissue. *Science (New York, N.Y.)*, *282*(5397), 2226–2230. https://doi.org/10.1126/science.282.5397.2226

González-Grandío, E., Poza-Carrión, C., Sorzano, C. O. S., Cubas, P., González-Grand\'io, E., Poza-Carrión, C., Sorzano, C. O. S., & Cubas, P. (2013). BRANCHED1 promotes axillary bud dormancy in response to shade in Arabidopsis. *The Plant Cell*, *25*(3), 834–850. https://doi.org/10.1105/tpc.112.108480

Grbic, V., & Bleecker, a B. (2000). Axillary meristem development in Arabidopsis thaliana [In Process Citation]. *The Plant Journal : For Cell and Molecular Biology*, *21*(2), 215–223.

Heidel, A. J., Clarke, J. D., Antonovics, J., & Dong, X. (2004). Fitness costs of mutations affecting the systemic acquired resistance pathway in Arabidopsis thaliana. *Genetics*, *168*(4), 2197–2206. https://doi.org/10.1534/genetics.104.032193

Holeski, L. M., Chase-Alone, R., & Kelly, J. K. (2010). The genetics of phenotypic plasticity in plant defense: trichome production in Mimulus guttatus. *The American Naturalist*, *175*(4), 391–400. https://doi.org/10.1086/651300

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., others, Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., … Guo, Y. L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, *43*(5), 476–481. https://doi.org/10.1038/NG.807

Karkkäinen, K., Løe, G., & Ågren, J. (2004). Population structure in Arabidopsis lyrata: evidence for divergent selection on trichome production. *Evolution*, *58*(12), 2831–2836.

Kebrom, T. H., Burson, B. L., & Finlayson, S. a. (2006). Phytochrome B represses Teosinte Branched1 expression and induces sorghum axillary bud outgrowth in response to light signals. *Plant Physiology*, *140*(3), 1109–1117. https://doi.org/10.1104/pp.105.074856

Kim, E., & Donohue, K. (2013). Local adaptation and plasticity of Erysimum capitatum to altitude: Its implications for responses to climate change. *Journal of Ecology*, *101*(3), 796–805. https://doi.org/10.1111/1365-2745.12077

Kimura, M. (1984). *The neutral theory of molecular evolution*. Cambridge University Press.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. In *Idengaku zasshi* (Vol. 66, Issue 4, pp. 367–386). https://doi.org/10.1266/jjg.66.367

Koch, M. A., & Matschinger, M. (2007). Evolution and genetic differentiation among relatives of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, *104*(15), 6272–6277.

Law, R. (1979). Optimal life histories under age-specific predation. *American Naturalist*, 399–417.

Leinonen, P. H., Remington, D. L., Leppälä, J., & Savolainen, O. (2012). Genetic basis of local adaptation and flowering time variation in Arabidopsis lyrata. *Molecular Ecology*.

Leinonen, P. H., Remington, D. L., & Savolainen, O. (2011). Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of arabidopsis lyrata. *Evolution*, *65*(1), 90–107. https://doi.org/10.1111/j.1558-5646.2010.01119.x

Leinonen, P. H., Sandring, S., Quilot, B., Clauss, M. J., Thomas, M. O., Ågren, J., & Savolainen, O. (2009). Local adaptation in european populations of arabidopsis lyrata (brassicaceae). *American Journal of Botany*, *96*(6), 1129–1137. https://doi.org/10.3732/ajb.0800080

Lewis, D. R., & Muday, G. K. (2009). Measurement of auxin transport in Arabidopsis thaliana. *Nature Protocols*, *4*(4), 437–451. https://doi.org/10.1038/nprot.2009.1

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Martin, A., & Orgogozo, V. (2013). The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. *Evolution*, *67*(5), 1235–1250. https://doi.org/10.1111/evo.12081

Martı´n-Trillo, M., Grandı´o, E. G., Serra, F., Marcel, F., Rodrı´guez-Buey, M. L., Schmitz, G., Theres, K., Bendahmane, A., Dopazo, H., & Cubas, P. (2011). Role of tomato BRANCHED1-like genes in the control of shoot branching. *The Plant Journal*, *67*(4), 701–714.

Mattila, T. M., Tyrmi, J., Pyhäjärvi, T., & Savolainen, O. (2017). Genome-Wide Analysis of Colonization History and Concomitant Selection in Arabidopsis lyrata. *Molecular Biology and Evolution*, *34*(10), 2665–2677. https://doi.org/10.1093/molbev/msx193

McKay, J. K., Richards, J. H., & Mitchell-Olds, T. (2003). Genetics of drought adaptation in Arabidopsis thaliana: I. Pleiotropy contributes to genetic correlations among ecological traits. *Molecular Ecology*, *12*(5), 1137–1151.

Mitchell-Olds, T., & Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature*, *441*(7096), 947–952. https://doi.org/10.1038/nature04878

Muller, M.-H. H., Leppälä, J., & Savolainen, O. (2007). Genome-wide effects of postglacial colonization in Arabidopsis lyrata. *Heredity*, *100*(1), 47–58. https://doi.org/10.1038/sj.hdy.6801057

Munné-Bosch, S. (2008). Do perennials really senesce? *Trends in Plant Science*, *13*(5), 216–220.

Niwa, M., Daimon, Y., Kurotani, K., Higo, A., Pruneda-Paz, J. L., Breton, G., Mitsuda, N., Kay, S. a, Ohme-Takagi, M., Endo, M., & Araki, T. (2013). BRANCHED1 interacts with FLOWERING LOCUS T to repress the floral transition of the axillary meristems in Arabidopsis. *The Plant Cell*, *25*(4), 1228–1242. https://doi.org/10.1105/tpc.112.109090

Obeso, J. R. (2002). The costs of reproduction in plants. *New Phytologist*, *155*(3), 321–348.

Okada, K., Ueda, J., Komaki, M., Bell, C., & Shimura, Y. (1991). Requirement of the Auxin Polar Transport System in Early Stages of Arabidopsis Floral Bud Formation. *The Plant Cell*, *3*(7), 677–684. https://doi.org/10.1105/tpc.3.7.677

Orr, H. A. (2005). The genetic theory of adaptation: a brief history. *Nature Reviews. Genetics*, *6*(2), 119–127. https://doi.org/10.1038/nrg1523

Petrásek, J., Friml, J. J., Petrášek, J., & Friml, J. J. (2009). Auxin transport routes in plant development. *Development*, *136*(16), 2675–2688. https://doi.org/10.1242/dev.030353

Poza-Carrión, C., Aguilar-Mart\'inez, J. A., & Cubas, P. (2007). Role of TCP gene BRANCHED1 in the control of shoot branching in Arabidopsis. *Plant Signal Behav*, *2*(6), 551–552.

Prusinkiewicz, P., Crawford, S., Smith, R. S., Ljung, K., Bennett, T., Ongaro, V., & Leyser, O. (2009). Control of bud activation by an auxin transport switch. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(41), 17431–17436. https://doi.org/10.1073/pnas.0906696106

Pyhäjärvi, T., Aalto, E., & Savolainen, O. (2012). Time Scales of divergence and speciation among natural populations and subspecies of Arabidopsis lyrata (Brassicaceae). *American Journal of Botany*, *99*(8), 1314–1322. https://doi.org/10.3732/ajb.1100580

Remington, D. L., Figueroa, J., & Rane, M. (2015). Timing of shoot development transitions affects degree of perenniality in Arabidopsis lyrata (Brassicaceae). *BMC Plant Biology*, *15*(1), 226. https://doi.org/10.1186/s12870-015-0606-2

Remington, D. L., Leinonen, P. H., Leppälä, J., & Savolainen, O. (2013). Complex genetic effects on early vegetative development shape resource allocation differences between Arabidopsis lyrata populations. *Genetics*, *195*(3), 1087–1102. https://doi.org/10.1534/genetics.113.151803

Riihimäki, M., Podolsky, R., Kuittinen, H., Koelewijn, H., & Savolainen, O. (2005). Studying genetics of adaptive variation in model organisms: flowering time variation in Arabidopsis lyrata. In *Genetics of Adaptation* (pp. 63–74). Springer.

Riihimäki, M., & Savolainen, O. (2004). Environmental and genetic effects on flowering differences between northern and southern populations of Arabidopsis lyrata (Brassicaceae). *American Journal of Botany*, *91*(7), 1036–1045.

Roff, D. a., & Fairbairn, D. J. (2007). The evolution of trade-offs: Where are we? *Journal of Evolutionary Biology*, *20*(2), 433–447. https://doi.org/10.1111/j.1420-9101.2006.01255.x

Ross-Ibarra, J., Wright, S. I., Foxe, J. P., Kawabe, A., DeRose-Wilson, L., Gos, G., Charlesworth, D., & Gaut, B. S. (2008). Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata. *PLoS ONE*, *3*(6). https://doi.org/10.1371/journal.pone.0002411

Sandring, S., & Ågren, J. (2009). POLLINATOR-MEDIATED SELECTION ON FLORAL DISPLAY AND FLOWERING TIME IN THE PERENNIAL HERB ARABIDOPSIS LYRATA. *Evolution*, *63*(5), 1292–1300. https://doi.org/10.1111/j.1558-5646.2009.00624.x

Sandring, S., RIIHIMÄKI, M.-A. a., Savolainen, O., & Ågren, J. (2007). Selection on flowering time and floral display in an alpine and a lowland population of Arabidopsis lyrata.

*Journal of Evolutionary Biology*, *20*(2), 558–567. https://doi.org/10.1111/j.1420-9101.2006.01260.x

Scarcelli, N., Cheverud, J. M., Schaal, B. A., & Kover, P. X. (2007). Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. *Proceedings of the National Academy of Sciences*, *104*(43), 16986–16991.

Schmickl, R., Jørgensen, M. H., Brysting, A. K., & Koch, M. a. (2010). The evolutionary history of the Arabidopsis lyrata complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolutionary Biology*, *10*, 98. https://doi.org/10.1186/1471-2148-10-98

Stearns, S. C. (1992). *The evolution of life histories* (Vol. 248). Oxford University Press Oxford.

Stern, D. L., & Orgogozo, V. (2008). The loci of evolution: How predictable is genetic evolution? *Evolution*, *62*(9), 2155–2177. https://doi.org/10.1111/j.1558-5646.2008.00450.x

Thomas, H. (2004). Do green plants age, and if so, how? In *Model Systems in Aging* (pp. 145–171). Springer.

Thomas, H., Thomas, H. M., & Ougham, H. (2000). Annuality, perenniality and cell death. *Journal of Experimental Botany*, *51*(352), 1781–1788.

Tian, D., Traw, M., Chen, J., Kreitman, M., & Bergelson, J. (2003). Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. *Nature*, *423*(May), 74–77. https://doi.org/10.1038/nature01575.1.

Toivainen, T., Pyhäjärvi, T., Niittyvuopio, A., & Savolainen, O. (2014). A recent local sweep at the PHYA locus in the Northern European Spiterstulen population of Arabidopsis lyrata. *Molecular Ecology*, *23*(5), 1040–1052. https://doi.org/10.1111/MEC.12682

Turner, T. L., Bourne, E. C., Von Wettberg, E. J., Hu, T. T., & Nuzhdin, S. V. (2010). Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nature Genetics*, *42*(3), 260–263.

van Noordwijk, a. J., & de Jong, G. (1986). Acquisition and Allocation of Resources: Their Influence on Variation in Life History Tactics. In *The American Naturalist* (Vol. 128, Issue 1, p. 137). https://doi.org/10.1086/284547

Vergeer, P., & Kunin, W. E. (2011). Life history variation in Arabidopsis lyrata across its range: Effects of climate, population size and herbivory. *Oikos*, *120*(7), 979–990. https://doi.org/10.1111/j.1600-0706.2010.18944.x

Vieten, A., Vanneste, S., Wisniewska, J., Benková, E., Benjamins, R., Beeckman, T., Luschnig, C., & Friml, J. (2005). Functional redundancy of PIN proteins is accompanied by auxin-dependent cross-regulation of PIN expression. *Development (Cambridge, England)*, *132*(20), 4521–4531. https://doi.org/10.1242/dev.02027

Wang, R., Farrona, S., Vincent, C., Joecker, A., Schoof, H., Turck, F., Alonso-Blanco, C., Coupland, G., & Albani, M. C. (2009). PEP1 regulates perennial flowering in Arabis alpina. *Nature*, *459*(7245), 423–427. https://doi.org/10.1038/nature07988

Wilczek, A. M., Roe, J. L., Knapp, M. C., Cooper, M. D., Lopez-Gallego, C., Martin, L. J., Muir, C. D., Sim, S., Walker, A., Anderson, J., & others. (2009). Effects of genetic perturbation on seasonal life history plasticity. *Science*, *323*(5916), 930–934.

Williams, G. C. (1966). Natural selection, the costs of reproduction, and a refinement of Lack's principle. *American Naturalist*, 687–690.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution (Vol. 1). na.

Young, T. P. (2010). Semelparity and Iteroparity. *Nature Education Knowledge*, *3*(10), 2.

DO: Add chapter references here for Chapter 3 and then for 2.

## References Chapter II

Blakeslee, J. J., Bandyopadhyay, A., Lee, O. R., Mravec, J., Titapiwatanakun, B., Sauer, M., Makam, S. N., Cheng, Y., Bouchard, R., Adamec, J., & others. (2007). Interactions among PIN-FORMED and P-glycoprotein auxin transporters in Arabidopsis. *The Plant Cell*, *19*(1), 131–147.

Casimiro, I., Beeckman, T., Graham, N., Bhalerao, R., Zhang, H., Casero, P., Sandberg, G., & Bennett, M. J. (2003). Dissecting Arabidopsis lateral root development. *Trends in Plant Science*, *8*(4), 165–171. https://doi.org/10.1016/S1360-1385(03)00051-7

Casimiro, I., Marchant, A., Bhalerao, R. P., Beeckman, T., Dhooge, S., Swarup, R., Graham, N., Inzé, D., Sandberg, G., Casero, P. J., Bennett, M., & others. (2001). Auxin transport promotes Arabidopsis lateral root initiation. *The Plant Cell Online*, *13*(4), 843–852. https://doi.org/10.1105/TPC.13.4.843

Checker, V. G., Kushwaha, H. R., Kumari, P., & Yadav, S. (2018). Role of Phytohormones in Plant Defense: Signaling and Cross Talk. *Molecular Aspects of Plant-Pathogen Interaction*, 159–184. https://doi.org/10.1007/978-981-10-7371-7_7

Chng, M. W., & Moore, K. A. (2020). Differences in inflorescence numbers and endogenous gibberellic acid levels in 'afterglow' bougainvillea. *HortTechnology*, *30*(6), 650–653. https://doi.org/10.21273/HORTTECH04673-20

Davies, P. J. (2010). The Plant Hormones: Their Nature, Occurrence, and Functions. *Plant Hormones: Biosynthesis, Signal Transduction, Action!*, 1–15. https://doi.org/10.1007/978-1-4020-2686-7_1

Dhonukshe, P., Grigoriev, I., Fischer, R., Tominaga, M., Robinson, D. G., Hašek, J., Paciorek, T., Petrášek, J., Seifertová, D., Tejos, R., Meisel, L. A., Zažímalová, E., Gadella, T. W.

J., Stierhof, Y. D., Ueda, T., Oiwa, K., Akhmanova, A., Brock, R., Spang, A., & Friml, J. (2008). Auxin transport inhibitors impair vesicle motility and actin cytoskeleton dynamics in diverse eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(11), 4489–4494. https://doi.org/10.1073/PNAS.0711414105

Ding, Z., Wang, B., Moreno, I., Duplá Ková, N., Simon, S., Carraro, N., Reemmer, J., Pě Nčí K, A., Chen, X., Tejos, R., SkÅpa, P., Pollmann, S., Mravec, J., Petrá Š Ek, J., ZaÅimalova, E., Honys, D., Rolčí K, J., Murphy, A., Orellana, A., … Friml, J. (2012). ER-localized auxin transporter PIN8 regulates auxin homeostasis and male gametophyte development in Arabidopsis. *Nature Communications*, *3*. https://doi.org/10.1038/NCOMMS1941

Enders, T. A., & Strader, L. C. (2015). Auxin Activity: Past, present, and Future. *American Journal of Botany*, *102*(2), 180. https://doi.org/10.3732/AJB.1400285

Flasiński, M., & Hac-Wydro, K. (2014). Natural vs synthetic auxin: studies on the interactions between plant hormones and biological membrane lipids. *Environmental Research*, *133*, 123–134. https://doi.org/10.1016/J.ENVRES.2014.05.019

Friml, J. (2003). Auxin transport—shaping the plant. *Current Opinion in Plant Biology*, 7–12. https://doi.org/10.1016/S1369-5266(02)00003-1

Gälweiler, L., Guan, C., Müller, A., Wisman, E., Mendgen, K., Yephremov, A., & Palme, K. (1998). Regulation of polar auxin transport by AtPIN1 in Arabidopsis vascular tissue. *Science (New York, N.Y.)*, *282*(5397), 2226–2230. https://doi.org/10.1126/science.282.5397.2226

Ganguly, A., Lee, S. H., Cho, M., Lee, O. R., Yoo, H., & Cho, H.-T. (2010). Differential auxin-transporting activities of PIN-FORMED proteins in Arabidopsis root hair cells. *Plant Physiology*, *153*(3), 1046–1061. https://doi.org/10.1104/pp.110.156505

Jones, A. M. (1998). Auxin transport: down and out and up again. *Science (New York, N.Y.)*, *282*(5397), 2201–2202. https://doi.org/10.1126/SCIENCE.282.5397.2201

Kerk, N. M., & Feldman, L. J. (1995). A biochemical model for the initiation and maintenance of the quiescent center: implications for organization of root meristems. *Development*, *121*(9), 2825–2833. https://doi.org/10.1242/DEV.121.9.2825

Kitakura, S., Vanneste, S., Robert, S., Löfke, C., Teichmann, T., Tanaka, H., & Friml, J. (2011). Clathrin Mediates Endocytosis and Polar Distribution of PIN Auxin Transporters in Arabidopsis. *The Plant Cell*, *23*(5), 1920. https://doi.org/10.1105/TPC.111.083030

Kleine-Vehn, J., Wabnik, K., Martinière, A., Łangowski, Ł., Willig, K., Naramoto, S., Leitner, J., Tanaka, H., Jakobs, S., Robert, S., Luschnig, C., Govaerts, W., W Hell, S., Runions, J., & Friml, J. (2011). Recycling, clustering, and endocytosis jointly maintain PIN auxin carrier polarity at the plasma membrane. *Molecular Systems Biology*, *7*. https://doi.org/10.1038/MSB.2011.72

Kondhare, K. R., Patil, A. B., & Giri, A. P. (2021). Auxin: An emerging regulator of tuber and storage root development. *Plant Science*, *306*, 110854. https://doi.org/10.1016/J.PLANTSCI.2021.110854

Leinonen, P. H., Remington, D. L., Leppälä, J., & Savolainen, O. (2013). Genetic basis of local adaptation and flowering time variation in Arabidopsis lyrata. *Molecular Ecology*, *22*(3), 709–723. https://doi.org/10.1111/j.1365-294X.2012.05678.x

Lewis, D. R., & Muday, G. K. (2009). Measurement of auxin transport in Arabidopsis thaliana. *Nature Protocols*, *4*(4), 437–451. https://doi.org/10.1038/nprot.2009.1

Leyser, O. (2005). The fall and rise of apical dominance. *Current Opinion in Genetics and Development*, *15*(4), 468–471. https://doi.org/10.1016/j.gde.2005.06.010

Leyser, O. (2009). The control of shoot branching: An example of plant information processing. *Plant, Cell and Environment*, *32*(6), 694–703. https://doi.org/10.1111/j.1365-3040.2009.01930.x

Löfke, C., Luschnig, C., & Kleine-Vehn, J. (2013). Posttranslational modification and trafficking of PIN auxin efflux carriers. *Mechanisms of Development*, *130*(1), 82–94. https://doi.org/10.1016/J.MOD.2012.02.003

Meuwly, P., & Pilet, P.-E. (1991). Local treatment with indole-3-acetic acid induces differential growth responses in Zea mays L. roots on JSTOR. Planta. https://www.jstor.org/stable/23381357

Mravec, J., Skůpa, P., Bailly, A., Hoyerová, K., Křeček, P., Bielach, A., Petrášek, J., Zhang, J., Gaykova, V., Stierhof, Y. D., Dobrev, P. I., Schwarzerová, K., Rolčík, J., Seifertová, D., Luschnig, C., Benková, E., Zažímalová, E., Geisler, M., & Friml, J. (2009). Subcellular homeostasis of phytohormone auxin is mediated by the ER-localized PIN5 transporter. *Nature 2009 459:7250*, *459*(7250), 1136–1140. https://doi.org/10.1038/nature08066

Muday, G. K., & Haworth, P. (1994). Tomato root growth, gravitropism, and lateral development: correlation with auxin transport. *Plant Physiology and Biochemistry : PPB*, *32*(2), 193–203. https://europepmc.org/article/med/11540612

Okada, K., Ueda, J., Komaki, M., Bell, C., & Shimura, Y. (1991). Requirement of the Auxin Polar Transport System in Early Stages of Arabidopsis Floral Bud Formation. *The Plant Cell*, *3*(7), 677–684. https://doi.org/10.1105/tpc.3.7.677

Paciorek, T., Sauer, M., Balla, J., Wiśniewska, J., & Friml, J. (2006). Immunocytochemical technique for protein localization in sections of plant tissues. *Nature Protocols*, *1*(1), 104–107. https://doi.org/10.1038/nprot.2006.16

Petrásek, J., Friml, J. J., Petrášek, J., & Friml, J. J. (2009). Auxin transport routes in plant development. *Development*, *136*(16), 2675–2688. https://doi.org/10.1242/dev.030353

Prusinkiewicz, P., Crawford, S., Smith, R. S., Ljung, K., Bennett, T., Ongaro, V., & Leyser, O. (2009). Control of bud activation by an auxin transport switch. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(41), 17431–17436. https://doi.org/10.1073/pnas.0906696106

Rashotte, a M., DeLong, a, & Muday, G. K. (2001). Genetic and chemical reductions in protein phosphatase activity alter auxin transport, gravity response, and lateral root growth. *The Plant Cell*, *13*(7), 1683–1697. https://doi.org/10.1105/TPC.010158

Rashotte, A. M., Brady, S. R., Reed, R. C., Ante, S. J., & Muday, G. K. (2000). Basipetal Auxin Transport Is Required for Gravitropism in Roots of Arabidopsis. *Plant Physiology*, *122*(2), 481–490. https://doi.org/10.1104/PP.122.2.481

Reed, R. C., Brady, S. R., & Muday, G. K. (1998). Inhibition of Auxin Movement from the Shoot into the Root Inhibits Lateral Root Development in Arabidopsis. *Plant Physiology*, *118*(4), 1369–1378. https://doi.org/10.1104/PP.118.4.1369

Reinhardt, D., Mandel, T., & Kuhlemeier, C. (2000). Auxin regulates the initiation and radial position of plant lateral organs. *The Plant Cell*, *12*(4), 507–518. https://doi.org/10.1105/TPC.12.4.507

Reinhardt, D., Pesce, E. R., Stieger, P., Mandel, T., Baltensperger, K., Bennett, M., Traas, J., Friml, J., & Kuhlemeier, C. (2003). Regulation of phyllotaxis by polar auxin transport. *Nature*, *426*(6964), 255–260. https://doi.org/10.1038/NATURE02081

Remington, D. L., Figueroa, J., & Rane, M. (2015). Timing of shoot development transitions affects degree of perenniality in Arabidopsis lyrata (Brassicaceae). *BMC Plant Biology*, *15*(1), 226. https://doi.org/10.1186/s12870-015-0606-2

Remington, D. L., Leinonen, P. H., Leppälä, J., & Savolainen, O. (2013). Complex genetic effects on early vegetative development shape resource allocation differences between Arabidopsis lyrata populations. *Genetics*, *195*(3), 1087–1102. https://doi.org/10.1534/genetics.113.151803

Ruegger, M., Dewey, E., Hobbie, L., Brown, D., Bernasconi, P., Turner, J., Muday, G., & Estelle, M. (1997). Reduced naphthylphthalamic acid binding in the tir3 mutant of Arabidopsis is associated with a reduction in polar auxin transport and diverse morphological defects. *The Plant Cell*, *9*(5), 745–757. https://doi.org/10.1105/TPC.9.5.745

Shi, Q., Li, C., & Zhang, F. (2006). Nicotine synthesis in Nicotiana tabacum L. induced by mechanical wounding is regulated by auxin. *Journal of Experimental Botany*, *57*(11), 2899–2907. https://doi.org/10.1093/JXB/ERL051

Snyder, W. E. (1949). Some Responses of Plants to 2,3,5-Triiodobenzoic Acid. *Plant Physiology*, *24*(2), 195–206. https://doi.org/10.1104/PP.24.2.195

Steinmann, T., Geldner, N., Grebe, M., Mangold, S., Jackson, C. L., Paris, S., Gälweiler, L., Palme, K., & Jürgens, G. (1999). Coordinated polar localization of auxin efflux carrier PIN1 by GNOM ARF GEF. *Science (New York, N.Y.)*, *286*(5438), 316–318. https://doi.org/10.1126/science.286.5438.316

Thimann, K. V. (1988). history of the knowledge of auxin. Physiology and Biochemistry of Auxins in Plants : Proceedings of the Symposium, Held at Liblice, Czechoslovakia, September 28-October 2, 1987 / Edited by Milan Kutacek, Robert S. Bandurski and Jan Krekule. https://doi.org/10.3/JQUERY-UI.JS

Trewavas, A. J. (1992). What remains of the Cholodny-Went theory? A summing up. *Plant, Cell & Environment*, *15*(7), 793–794. http://www.ncbi.nlm.nih.gov/pubmed/11541817

Tsurumi, S., & Ohwaki, Y. (1978). Transport of 14C-lableled indoleacetic acid in Vicia root segments. *Plant and Cell Physiology*, *19*(7), 1195–1206. https://doi.org/10.1093/OXFORDJOURNALS.PCP.A075700

Vieten, A., Vanneste, S., Wisniewska, J., Benková, E., Benjamins, R., Beeckman, T., Luschnig, C., & Friml, J. (2005). Functional redundancy of PIN proteins is accompanied by auxin-dependent cross-regulation of PIN expression. *Development (Cambridge, England)*, *132*(20), 4521–4531. https://doi.org/10.1242/dev.02027

Waldie, T., & Leyser, O. (2018). Cytokinin Targets Auxin Transport to Promote Shoot Branching. *Plant Physiology*, *177*(2), 803–818. https://doi.org/10.1104/pp.17.01691

Wang, C., Liu, Y., Li, S. S., & Han, G. Z. (2015). Insights into the Origin and Evolution of the Plant Hormone Signaling Machinery. *Plant Physiology*, *167*(3), 872–886. https://doi.org/10.1104/PP.114.247403

Wisniewska, J., Xu, J., Seifartová, D., Brewer, P. B., Růžička, K., Blilou, L., Rouquié, D., Benková, E., Scheres, B., & Friml, J. (2006). Polar PIN localization directs auxin flow in plants. *Science (New York, N.Y.)*, *312*(5775), 883. https://doi.org/10.1126/SCIENCE.1121356

Wu, X., & McSteen, P. (2007). The role of auxin transport during inflorescence development in maize (Zea mays, Poaceae). *American Journal of Botany*, *94*(11), 1745–1755. https://doi.org/10.3732/AJB.94.11.1745

## References Chapter III

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., Qiu, R., Lee, C., & Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature 2013 500:7461*, *500*(7461), 207–211. https://doi.org/10.1038/nature12064

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R.,

… Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature 2012 491:7422*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Chory, J., Dean, C., Debieu, M., Meaux, J. De, Joseph, R., Faure, N., … Traw, M. B. (2011). *NIH Public Access*. *465*(7298), 627–631. https://doi.org/10.1038/nature08800.Genome-wide

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics 2011 12:10*, *12*(10), 703–714. https://doi.org/10.1038/nrg3054

Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y., & Lappalainen, T. (2016). Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nature Communications 2016 7:1*, *7*(1), 1–6. https://doi.org/10.1038/ncomms12817

Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, *7*(2), 111–122. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040591

Consortium, I. H., Others, & International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics 2001 29:2*, *29*(2), 229–232. https://doi.org/10.1038/ng1001-229

Delaneau, O., Coulonges, C., & Zagury, J.-F. F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, *9*(1), 540. https://doi.org/10.1186/1471-2105-9-540

Durbin, R. (2014). Efficient haplotype matching and storage using the Positional Burrows-Wheeler Transform (PBWT). *Bioinformatics (Oxford, England)*, *30*(9), 1266–1272. https://doi.org/10.1093/bioinformatics/btu014

Fan, H. C., Wang, J., Potanina, A., & Quake, S. R. (2011). Whole-genome molecular haplotyping of single cells. *Nature Biotechnology 2010 29:1*, *29*(1), 51–57. https://doi.org/10.1038/nbt.1739

Giakountis, A., Cremer, F., Sim, S., Reymond, M., Schmitt, J., & Coupland, G. (2009). Distinct Patterns of Genetic Variation Alter Flowering Responses of Arabidopsis Accessions to Different Daylengths. *PLANT PHYSIOLOGY*, *152*(1), 177–191. https://doi.org/10.1104/pp.109.140772

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/JOURNAL.PGEN.1000529

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., others, Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., … Guo, Y. L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, *43*(5), 476–481. https://doi.org/10.1038/NG.807

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., & Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics 2008 40:9*, *40*(9), 1068–1075. https://doi.org/10.1038/ng.216

Lo, C. (2014). Algorithms for Haplotype Phasing. *Undefined*.

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics 2016 48:11*, *48*(11), 1443–1448. https://doi.org/10.1038/ng.3679

Loh, P. R., Genovese, G., Handsaker, R. E., Finucane, H. K., Reshef, Y. A., Palamara, P. F., Birmann, B. M., Talkowski, M. E., Bakhoum, S. F., McCarroll, S. A., & Price, A. L. (2018). Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature 2018 559:7714*, *559*(7714), 350–355. https://doi.org/10.1038/s41586-018-0321-x

Loh, P. R., Palamara, P. F., & Price, A. L. (2016). *Fast and accurate long-range phasing in a UK Biobank cohort*. *48*(7). https://www.nature.com/articles/ng.3571

Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K., & Song, Q. (2010). Direct determination of molecular haplotypes by chromosome microdissection. *Nature Methods 2010 7:4*, *7*(4), 299–301. https://doi.org/10.1038/nmeth.1443

Naseri, A., Zhi, D., & Zhang, S. (2019). Multi-allelic positional Burrows-Wheeler transform. *BMC Bioinformatics*, *20*(11), 279. https://doi.org/10.1186/s12859-019-2821-6

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., … Campbell, P. J. (2012). The Life History of 21 Breast Cancers. *Cell*, *149*(5), 994–1007. https://doi.org/10.1016/J.CELL.2012.04.023

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H. Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C. S., … Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods 2015 12:8*, *12*(8), 780–786. https://doi.org/10.1038/nmeth.3454

Porubsky, D., Garg, S., Sanders, A. D., Korbel, J. O., Guryev, V., Lansdorp, P. M., & Marschall, T. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature Communications 2017 8:1*, *8*(1), 1–10. https://doi.org/10.1038/s41467-017-01389-4

Reich, D. E., Cargili, M., Boik, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., & Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature 2001 411:6834*, *411*(6834), 199–204. https://doi.org/10.1038/35075590

Rubinacci, S., Delaneau, O., & Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. *PLOS Genetics*, *16*(11), e1009049. https://doi.org/10.1371/journal.pgen.1009049

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V, Patterson, N. J., McDonald, G. J., & others. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., & others. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918.

Seo, J. S., Rhie, A., Kim, J. J. J. J., Lee, S., Sohn, M. H., Kim, C. U., Hastie, A., Cao, H., Yun, J. Y., Kim, J. J. J. J., Kuk, J., Park, G. H., Kim, J. J. J. J., Ryu, H., Kim, J. J. J. J., Roh, M., Baek, J., Hunkapiller, M. W., Korlach, J., … Kim, C. U. (2016). De novo assembly and phasing of a Korean human genome. *Nature 2016 538:7624*, *538*(7624), 243–247. https://doi.org/10.1038/nature20098

Shendure, J., & Aiden, E. L. (2012). The expanding scope of DNA sequencing. *Nature Biotechnology 2012 30:11*, *30*(11), 1084–1094. https://doi.org/10.1038/nbt.2421

Snyder, M. W., Adey, A., Kitzman, J. O., & Shendure, J. (2015). Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics 2015 16:6*, *16*(6), 344–358. https://doi.org/10.1038/nrg3903

Yang, H., Chen, X., & Wong, W. H. (2011). Completely phased genome sequencing through chromosome sorting. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(1), 12–17. https://doi.org/10.1073/PNAS.1016725108/-/DCSUPPLEMENTAL

Zhan, X. (2017). Lab: Phasing. In *Zhan Phasing Workshop*. https://portal.biohpc.swmed.edu/content/training/bioinformatics-nanocourses/gwas/zhan-phasing-workshop/

Zhang, C. Z., & Pellman, D. (2015). From Mutational Mechanisms in Single Cells to Mutational Patterns in Cancer Genomes. *Cold Spring Harbor Symposia on Quantitative Biology*, *80*, 117–137. https://doi.org/10.1101/SQB.2015.80.027623

Zhang, C. Z., Spektor, A., Cornils, H., Francis, J. M., Jackson, E. K., Liu, S., Meyerson, M., & Pellman, D. (2015). Chromothripsis from DNA damage in micronuclei. *Nature 2015 522:7555*, *522*(7555), 179–184. https://doi.org/10.1038/nature14493

## References Chapter IV

Arnold, B., Kim, S. T., & Bomblies, K. (2015). Single Geographic Origin of a Widespread Autotetraploid Arabidopsis arenosa Lineage Followed by Interploidy Admixture. *Molecular Biology and Evolution*, *32*(6), 1382–1395. https://doi.org/10.1093/MOLBEV/MSV089

Barbez, E., Kubeš, M., Rolčík, J., Béziat, C., Pěnčík, A., Wang, B., Rosquete, M. R., Zhu, J., Dobrev, P. I., Lee, Y., Zažímalovà, E., Petrášek, J., Geisler, M., Friml, J., Kleine-Vehn, J., Zašímalovà, E., Petrášek, J., Geisler, M., Friml, J., & Kleine-Vehn, J. (2012). A novel putative auxin carrier family regulates intracellular auxin homeostasis in plants. *Nature*, *485*(7396), 119–122. https://doi.org/10.1038/nature11001

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, *16*(1), 1–12. https://doi.org/10.1186/S13059-015-0762-6/FIGURES/7

Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y., & Lappalainen, T. (2016). Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nature Communications 2016 7:1*, *7*(1), 1–6. https://doi.org/10.1038/ncomms12817

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207–3212. https://doi.org/10.1093/bioinformatics/btp579

Ding, Z., GGalván-Ampudia, C. S., Demarsy, E., Łangowski, Ł., Kleine-Vehn, J., Fan, Y., Morita, M. T., Tasaka, M., Fankhauser, C., Offringa, R., & Friml, J. (2011). Light-mediated polarization of the PIN3 auxin transporter for the phototropic response in Arabidopsis. *Nature Cell Biology 2011 13:4*, *13*(4), 447–452. https://doi.org/10.1038/ncb2208

STAR manual 2.7.10a, no. 2.5.4a, 1 (2022). https://github.com/alexdobin/STAR

Doebley, J., Stec, A., & Gustus, C. (1995). Teosinte Branched1. *Gene Expression*, *81*.

Doebley, J., Stec, A., & Hubbard, L. (1997). *The evolution of apical dominance in maize*.

Feraru, E., Vosolsobě, S., Feraru, M. I., Petrášek, J., & Kleine-Vehn, J. (2012). Evolution and structural diversification of PILS putative auxin carriers in plants. *Frontiers in Plant Science*, *3*(OCT), 227. https://doi.org/10.3389/FPLS.2012.00227/ABSTRACT

Friml, J. J., Wiśniewska, J., Benková, E., Mendgen, K., Palme, K., Wiśniewska, J., Benková, E., Mendgen, K., & Palme, K. (2002). Lateral relocation of auxin efflux regulator PIN3

mediates tropism in Arabidopsis. *Nature*, *415*(6873), 806–809. https://doi.org/10.1038/415806A

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., others, Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., … Guo, Y. L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, *43*(5), 476–481. https://doi.org/10.1038/NG.807

Hu, T., Yin, S., Sun, J., Linghu, Y., Ma, J., Pan, J., & Wang, C. (2021). Clathrin light chains regulate hypocotyl elongation by affecting the polarization of the auxin transporter PIN3 in Arabidopsis. *Journal of Integrative Plant Biology*, *63*(11), 1922–1936. https://doi.org/10.1111/JIPB.13171/SUPPINFO

Keurentjes, J. J. B., Fu, J., Terpstra, I. R., Garcia, J. M., van den Ackerveken, G., Snoek, L. B., Peeters, A. J. M., Vreugdenhil, D., Koornneef, M., & Jansen, R. C. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences*, *104*(5), 1708–1713.

Keuskamp, D. H., Pollmann, S., Voesenek, L. A. C. J., Peeters, A. J. M., & Pierik, R. (2010). Auxin transport through PIN-FORMED 3 (PIN3) controls shade avoidance and fitness during competition. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(52), 22740–22744. https://doi.org/10.1073/pnas.1013457108

Kirst, M., Basten, C. J., Myburg, A. A., Zeng, Z. B., & Sederoff, R. R. (2005). Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. *Genetics*, *169*(4), 2295–2303. https://doi.org/10.1534/GENETICS.104.039198

Lalonde, E., Ha, K. C. H., Wang, Z., Bemmo, A., Kleinman, C. L., Kwan, T., Pastinen, T., & Majewski, J. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Research*, *21*(4), 545–554. https://doi.org/10.1101/GR.111211.110

Leinonen, P. H., Remington, D. L., Leppälä, J., & Savolainen, O. (2013). Genetic basis of local adaptation and flowering time variation in Arabidopsis lyrata. *Molecular Ecology*, *22*(3), 709–723. https://doi.org/10.1111/j.1365-294X.2012.05678.x

Leinonen, P. H., Remington, D. L., & Savolainen, O. (2011). Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of arabidopsis lyrata. *Evolution*, *65*(1), 90–107. https://doi.org/10.1111/j.1558-5646.2010.01119.x

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, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/GR.078212.108

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. https://doi.org/10.1016/J.CELL.2008.03.029/ATTACHMENT/D78DAC70-EEE1-49ED-A854-DD1252ACED80/MMC15.PDF

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Mattila, T. M., Tyrmi, J., Pyhäjärvi, T., & Savolainen, O. (2017). Genome-Wide Analysis of Colonization History and Concomitant Selection in Arabidopsis lyrata. *Molecular Biology and Evolution*, *34*(10), 2665–2677. https://doi.org/10.1093/molbev/msx193

Mohanta, T. K., Mohanta, N., & Bae, H. (2015). Identification and Expression Analysis of PIN-Like (PILS) Gene Family of Rice Treated with Auxin and Cytokinin. *Genes 2015, Vol. 6, Pages 622-640*, *6*(3), 622–640. https://doi.org/10.3390/GENES6030622

Müller, A., Guan, C., Gälweiler, L., Tänzler, P., Huijser, P., Marchant, A., Parry, G., Bennett, M., Wisman, E., & Palme, K. (1998). AtPIN2 defines a locus of Arabidopsis for root gravitropism control. *The EMBO Journal*, *17*(23), 6903–6911. https://doi.org/10.1093/EMBOJ/17.23.6903

Munger, S. C., Raghupathy, N., Choi, K., Simons, A. K., Gatti, D. M., Hinerfeld, D. A., Svenson, K. L., Keller, M. P., Attie, A. D., Hibbs, M. A., & others. (2014). RNA-Seq Alignment to Individualized Genomes Improves Transcript Abundance Estimates in Multiparent Populations. *Genetics*, *198*(1), 59–73.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, *320*(5881), 1344–1349. https://doi.org/10.1126/SCIENCE.1158441

Ottenschläger, I., Wolff, P., Wolverton, C., Bhalerao, R. P., Sandberg, G., Ishikawa, H., Evans, M., & Palme, K. (2003). Gravity-regulated differential auxin transport from columella to lateral root cap cells. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(5), 2987–2991. https://doi.org/10.1073/PNAS.0437936100/ASSET/7F4C60C9-403F-45F4-8FD4-212E78EFAEC6/ASSETS/GRAPHIC/PQ0437936006.JPEG

Raghupathy, N., Choi, K., Vincent, M. J., Beane, G. L., Sheppard, K. S., Munger, S. C., Korstanje, R., Pardo-Manual De Villena, F., & Churchill, G. A. (2018). Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics*, *34*(13), 2177–2184. https://doi.org/10.1093/BIOINFORMATICS/BTY078

Remington, D. L., Figueroa, J., & Rane, M. (2015). Timing of shoot development transitions affects degree of perenniality in Arabidopsis lyrata (Brassicaceae). *BMC Plant Biology*, *15*(1), 226. https://doi.org/10.1186/s12870-015-0606-2

Remington, D. L., Leinonen, P. H., Leppälä, J., & Savolainen, O. (2013). Complex genetic effects on early vegetative development shape resource allocation differences between Arabidopsis lyrata populations. *Genetics*, *195*(3), 1087–1102. https://doi.org/10.1534/genetics.113.151803

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., Bhardwaj, N., Rubin, M., Snyder, M., & Gerstein, M. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, *7*. https://doi.org/10.1038/MSB.2011.54

Savaldi-Goldstein, S., Peto, C., & Chory, J. (2007). The epidermis both drives and restricts plant shoot growth. *Nature 2006 446:7132*, *446*(7132), 199–202. https://doi.org/10.1038/nature05618

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., & others. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, *422*(6929), 297–302.

Shao, L., Xing, F., Xu, C., Zhang, Q., Che, J., Wang, X., Song, J., Li, X., Xiao, J., Chen, L. L., Ouyang, Y., & Zhang, Q. (2019). Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(12), 5653–5658. https://doi.org/10.1073/PNAS.1820513116/SUPPL_FILE/PNAS.1820513116.SD21.XLSX

Springer, N. M., & Stupar, R. M. (2007). Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *The Plant Cell*, *19*(8), 2391–2402. https://doi.org/10.1105/tpc.107.052258

Tao, Y., Ferrer, J. L., Ljung, K., Pojer, F., Hong, F., Long, J. A., Li, L., Moreno, J. E., Bowman, M. E., Ivans, L. J., Cheng, Y., Lim, J., Zhao, Y., Ballaré, C. L., Sandberg, G., Noel, J. P., & Chory, J. (2008). Rapid Synthesis of Auxin via a New Tryptophan-Dependent Pathway Is Required for Shade Avoidance in Plants. *Cell*, *133*(1), 164–176. https://doi.org/10.1016/J.CELL.2008.01.049

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the GenomeAnalysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *11*(1110), 11.10.1. https://doi.org/10.1002/0471250953.BI1110S43

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, *430*(6995), 85–88. https://doi.org/10.1038/NATURE02698

## References Chapter V


Gove, R. P., Chen, W., Zweber, N. B., Erwin, R., Rychtář, J., & Remington, D. L. (2012). Effects of causal networks on the structure and stability of resource allocation trait correlations. *Journal of Theoretical Biology*, *293*, 1–14. https://doi.org/10.1016/j.jtbi.2011.09.034

Kemi, U., Leinonen, P. H., Savolainen, O., & Kuittinen, H. (2019). Inflorescence shoot elongation, but not flower primordia formation, is photoperiodically regulated in Arabidopsis lyrata. *Annals of Botany*, *124*(1), 91–102. https://doi.org/10.1093/AOB/MCZ035

Kim, E., & Donohue, K. (2011). Population differentiation and plasticity in vegetative ontogeny: Effects on life-history expression in erysimum capitatum (brassicaceae). *American Journal of Botany*, *98*(11), 1752–1761. https://doi.org/10.3732/ajb.1100194

Kim, E., & Donohue, K. (2012). The effect of plant architecture on drought resistance: Implications for the evolution of semelparity in Erysimum capitatum. *Functional Ecology*, *26*(1), 294–303. https://doi.org/10.1111/j.1365-2435.2011.01936.x

Leinonen, P. H., Remington, D. L., Leppälä, J., & Savolainen, O. (2012). Genetic basis of local adaptation and flowering time variation in Arabidopsis lyrata. *Molecular Ecology*.

Remington, D. L., Figueroa, J., & Rane, M. (2015). Timing of shoot development transitions affects degree of perenniality in Arabidopsis lyrata (Brassicaceae). *BMC Plant Biology*, *15*(1), 226. https://doi.org/10.1186/s12870-015-0606-2

Remington, D. L., Leinonen, P. H., Leppälä, J., & Savolainen, O. (2013). Complex genetic effects on early vegetative development shape resource allocation differences between Arabidopsis lyrata populations. *Genetics*, *195*(3), 1087–1102. https://doi.org/10.1534/genetics.113.151803

Wang, R., Farrona, S., Vincent, C., Joecker, A., Schoof, H., Turck, F., Alonso-Blanco, C., Coupland, G., & Albani, M. C. (2009). PEP1 regulates perennial flowering in Arabis alpina. *Nature*, *459*(7245), 423–427. https://doi.org/10.1038/nature07988