A bioinformatics approach to quantify the effects of the underlying regulatory mechanisms on natural variation in gene expression by allele-specific expression analysis in *Arabidopsis thaliana* accessions using RNA-Seq Data

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Ganga Jeena

aus Gahana, Nainital, India

Köln, 2021

Die vorliegende Arbeit wurde am Max-Planck-Institut für Pflanzenzüchtungsforschung in Köln in der Abteilung für Pflanzenzüchtung und Genetik (Direktor: Prof. Dr. ir. Maarten Koornneef) in der Arbeitsgruppe von Dr. José M Jiménez-Gómez angefertigt.

The work described in this thesis was conducted at the Max Planck Institute for Plant Breeding Research in the Department of Plant Breeding and Genetics (Director: Prof. Dr. ir. Maarten Koornneef) under the supervision of Dr. José M Jiménez-Gómez.

Max Planck Institute for
Plant Breeding Research

IMPRS Cologne

Berichterstatter:	Prof. Dr. Korbinian Schneeberger Prof. Dr. Achim Tresch
Prüfungsvorsitzender:	Prof. Dr. Alga Zuccaro
Beisitzer:	Dr. Gregor Langen
Tag der Disputation:	August 3, 2020

ABSTRACT

Natural variation in gene expression plays a crucial role in evolution, natural selection and improved response to environmental cues. Expression divergence may be caused by localised polymorphisms within the gene or regulatory regions (cis) or induced by differences in regulatory factors (trans). The relative contribution of cis- and trans-regulatory variants can be quantified by genome-wide analysis of Allele-specific expression (ASE) with few F1 samples, unlike the conventional eQTL mapping methods. Differential allelic expression patterns between reciprocal crosses reflect the effect of parent-of-origin and genomic imprinting, while deviation in gene expression of hybrids from parentals reflects heterosis events.

This project aims to define an optimal methodology for ASE analysis using RNA-seq data to quantify the effects of cis- and trans-regulatory variants on gene expression differences. To develop this protocol, we used simulated and sequenced RNA-Seq data from leaf tissue in three biological replicates of *Arabidopsis thaliana* accessions Cape Verde Island (Cvi), Landsberg erecta (L*er*), and in their reciprocal hybrids (Cvi x L*er*, L*er* x Cvi). Factors like the genome of reference, gene region, frequency and location of SNPs, are crucial in expression quantification from mapped reads and were considered to determine a benchmarking criteria for the pipeline.

The ASE pipeline was used to perform analysis of RNA-Seq from six accessions of *Arabidopsis thaliana* namely An-1, Bor-4, Bur-0, Knox-10, Sha and L*er* from varying geographical locations and their reciprocal crosses with L*er*. The allelic ratio within F1 hybrids was compared to the parental ratio to quantify the regulatory effects of cis, trans, cis x trans and compensatory. Approximately 46% of genes with atleast one diagnostic SNP could be analysed for expression profiles. Nearly 1200 genes showed significant differential expression in parentals. Only genes with similar expression measure over exons and SNPs were analysed for allelic divergence in hybrids. ASE analysis could determine nearly 575 genes with different patterns of regulated expression variation. Of these $\sim$55% were regulated by cis-, $\sim$39% by trans-effects, $\sim$5% by compensatory and over less than 1% show cis x trans effect. Relative levels of regulatory effects were similar across all hybrids.

Interestingly, cis-regulated patterns were found to be prominent and conserved across multiple crosses. Most trans-effects were unique to crosses and thus explained more diversity. The cis-effects exhibit large expression divergence compared to trans-effects.

Parent-of-origin effects was assessed for genes with statistically significant allelic bias (cis-effects) by comparing the allelic ratio of reciprocal hybrids. Majority of allele-biased genes ($\sim$79%) were common between pairs of reciprocal hybrids and showed similar magnitude and direction of allelic variation. Only $\sim$9% and $\sim$12% genes reflected parental-effect with preferential paternal and maternal expressed alleles respectively.

Furthermore, gene expression of hybrids and parentals were compared to find heterosis events like overdominance, dominance, and additive effects. Preliminary analysis of expression inheritance patterns revealed that Overdominance are extremely rare events ($\sim$1%), majority of which are regulated by trans-effects.

Primary Metabolite profiling using GC/MS identified and measured relative levels of 37 metabolites across the six accessions and ten reciprocal hybrids. Proline levels were seen to exhibit maximum diversity. Proline biosynthesis genes were tested for association with proline levels variation. AtP5C1 Pyrroline-5-carboxylate (P5C) reductase, which plays a crucial role in proline synthesis, exhibit a strong correlation of transcript and proline levels, as experimentally validated by Verbruggen et al. in 1993.

In conclusion, I provide evidence to discuss the effects of parameters on expression quantification from RNA-Seq data, and suggest a framework for ASE analysis. I further used the optimised strategy to quantify effects of regulatory variants, inheritance patterns, parent-of-origin effects and assess association with phenotypic traits.

ZUSAMMENFASSUNG

Die natürliche Variation der Genexpression spielt eine entscheidende Rolle für die Evolution, die natürliche Selektion und die verbesserte Reaktion auf Umweltsignale. Eine Expressionsveränderung kann durch einzelne Polymorphismen innerhalb eines Gens oder der regulatorischen Regionen (cis) oder durch Unterschiede in den regulatorischen Faktoren (trans) verursacht werden. Der relative Beitrag von cis- und transregulatorischen Varianten kann im Gegensatz zu den herkömmlichen eQTL-Kartierungsmethoden durch genomweite Analyse der allelspezifischen Expression (ASE) mit wenigen F1-Proben quantifiziert werden. Differenzielle, allelische Expressionsmuster zwischen reziproken Kreuzungen spiegeln den Effekt des Ursprungselternteils und des genomischen Abdrucks wider, während Abweichungen in der Genexpression von Hybriden von Elternteilen Heterose-Ereignisse widerspiegeln.

Diese Doktorarbeit zielte darauf ab, eine optimale Methodik für die ASE-Analyse unter Verwendung von RNA-seq-Daten zu definieren, um die Auswirkungen von cis- und transregulatorischen Varianten auf Unterschiede in der Genexpression zu quantifizieren. Um dieses Protokoll zu entwickeln, verwendeten wir simulierte und sequenzierte RNA-Seq-Daten aus Blattgewebe von drei biologischen Replikaten von Arabidopsis thaliana-Akzessionen und von ihren reziproken Hybriden. Faktoren wie das Referenzgenom, die Genregion, die Häufigkeit und die Position von SNPs sind für die Expressionsquantifizierung anhand von alignierten Reads von entscheidender Bedeutung und wurden zur Bestimmung eines Benchmarking-Kriteriums für die Pipeline herangezogen.

Die ASE-Pipeline wurde weiter verwendet, um die RNA-Seq von sechs Akzessionen von Arabidopsis thaliana, nämlich An-1, Bor-4, Bur-0, Knox-10, Sha und L*er*, von verschiedenen geografischen Orten und deren wechselseitigen Kreuzungen mit L*er* zu analysieren. Das Allelverhältnis innerhalb von F1-Hybriden wurde mit dem Elternverhältnis verglichen, um die regulatorischen Wirkungen von cis, trans, cis x trans und kompensatorisch zu quantifizieren. Ungefähr 46% der Gene mit mindestens einem diagnostischen SNP konnten auf Expressionsprofile analysiert werden. Nahezu 1200 Gene zeigten eine signifikante unterschiedliche Expression bei den Eltern. Nur Gene mit einem ähnlichen Expressionsmaß über Exons und SNPs wurden auf allelische Divergenz in Hybriden analysiert. Die ASE-Analyse konnte fast 575 Gene mit unterschiedlichen Mustern regulierter Expressionsvariation bestimmen. Davon entfielen 55% auf cis-, 39% auf trans-Effekte, 5% auf kompensatorische Effekte und mehr als 1% auf cis x trans. Das relative Ausmaß regulatorischer Effekte war bei allen Hybriden ähnlich.

Interessanterweise wurde festgestellt, dass cis-regulierte Muster über mehrere Kreuze hinweg auffällig und konserviert sind. Die Mehrzahl der Trans-Effekte war für Kreuze einzigartig und erklärte somit mehr Vielfalt. Die cis-Effekte zeigen im Vergleich zu Trans-Effekten eine große Expressionsdivergenz.

Der Parent-of-Origin-Effekt wurde für Gene mit statistisch signifikanter allelischer Verzerrung (cis-Effekte) durch Vergleich des Allelverhältnisses von reziproken Hybriden bewertet. Die Mehrheit der allelabhängigen Gene (79%) war zwischen Paaren reziproker Hybride gemeinsam und zeigte eine ähnliche Größe und Richtung der allelischen Variation. Nur $\sim$9% und $\sim$12% der Gene spiegelten den elterlichen Effekt mit bevorzugten väterlichen bzw. mütterlich exprimierten Allelen wider.

Darüber hinaus wurde die Gene expression von Hybriden und Eltern verglichen, um Heteroseereignisse zu finden, d. H. überdominanz, Dominanz und additive Effekte. Eine vorläufige Analyse der Expressionsvererbungsmuster ergab, dass überdominanz extrem seltene Ereignisse ($\sim$1%) sind, von denen die meisten durch Trans-Effekte reguliert werden.

Primäres Metaboliten-Profiling unter Verwendung von GC/MS identifizierte und maß relative 37 Metaboliten über die sechs Akzessionen und zehn reziproken Hybriden. Prolinspiegel hatten die größten Unterschiede in den Samples. Prolin-Biosynthesegene wurden auf ihre Assoziation mit der Variation des Prolinspiegels getestet. AtP5C1-Pyrrolin-5-carboxylat (P5C) Reduktase, die eine entscheidende Rolle bei der Prolinsynthese spielt, weist eine starke Korrelation zwischen Transkript- und Prolinspiegeln auf, wie von Verbruggen et al. 1993 experimentell bestätigt.

Zusammenfassend möchte ich die Auswirkungen von Parametern auf die Expressionsquantifizierung anhand von RNA-Seq-Daten diskutieren und einen Rahmen für die ASE-Analyse vorschlagen, der zur Quantifizierung der Auswirkungen von regulatorischen Varianten, Vererbungsmustern, Eltern-Ursprungs-Effekten verwendet und für die Assoziation mit phänotypischen Merkmalen benutzt wurde.

TABLE OF CONTENTS

1. INTRODUCTION **2**

1.1. Natural variation . . . 2
1.2. QTL studies . . . 2
1.3. Expression-QTL analysis . . . 2
1.4. Categories of eQTLs . . . 3
1.5. ASE analysis . . . 4
1.6. Using RNA-Seq to detect ASE . . . 6
1.7. Advantages of RNA-Seq . . . 7
1.8. Advantages of Pyrosequencing . . . 7
1.9. Caveats of gene and allelic expression analysis . . . 7
1.10. Plant system used . . . 11
1.11. Parent-of-origin effect on allelic expression . . . 12
1.12. Inheritance patterns of expression divergence . . . 12
1.13. Primary metabolites profiling . . . 14

2. SCIENTIFIC OBJECTIVES **17**

2.1. Development of ASE analysis method . . . 17
2.2. Determination of genome-wide ASE in *Arabidopsis thaliana* . . . 17
2.2.1. Determination of parental effect on allelic expression . . . 17
2.2.2. Determination of expression inheritance patterns . . . 17
2.3. Integrated transcriptomics and metabolomics analysis . . . 18

3. MATERIALS AND METHODS **19**

3.1. Method development and benchmarking criteria . . . 19
3.2. Reported study . . . 19
3.3. Simulation analysis . . . 20
3.3.1. Plant material and RNA-seq data . . . 20
3.3.2. Sequence alignment . . . 20
3.3.3. Polymorphism detection . . . 21
3.3.4. Pseudo-reference genome construction . . . 21
3.3.5. Evaluation of impact of reference genome . . . 22
3.3.6. Simulation of differentially expressed RNA-seq data from pseudo-reference genomes . . . 22
3.3.7. Comparison of reference and pseudo-reference genome . . . 22
3.4. Allele-specific expression analysis . . . 22
3.4.1. Plant samples . . . 22
3.4.2. RNA-seq library preparation . . . 23
3.4.3. Re-sequencing library preparation . . . 23

3.4.4. Polymorphism detection . . . 24
3.4.5. Transcript abundance quantification . . . 24
3.4.5.1. Mapping and selection of reads . . . 24
3.4.5.2. Library size normalisation . . . 24
3.4.5.3. Selection of exonic reads . . . 24
3.4.5.4. Counting exonic reads . . . 25
3.4.5.5. Selection of allelic SNPs . . . 26
3.4.5.6. Counting SNP reads . . . 26
3.4.6. Selection criteria of genes for differential gene expression analysis 26
3.4.7. Differential expression analysis . . . 27
3.4.8. Comparison of expression variation over exons and SNPs . . . 27
3.4.9. Elucidation of expression divergence pattern . . . 27
3.5. Software packages used . . . 27
3.5.0.1. Selection criteria of genes for ASE analysis . . . 27
3.5.1. Verification of allelic ratio . . . 28
3.5.2. Functional enrichment . . . 28
3.6. Determination of parent-of-origin effect . . . 29
3.7. Inheritance of expression patterns . . . 29
3.8. Metabolite analysis . . . 29
3.8.1. Primary metabolite profiling by GC/TOF-MS . . . 29
3.8.2. Metabolite data analysis . . . 29

4. RESULTS AND DISCUSSION **31**
4.1. Allele-Specific expression analysis . . . 31
4.2. Method development . . . 32
4.2.1. Mapping of reads to reference genome . . . 33
4.2.2. Detection of polymorphic sites . . . 34
4.2.3. Distribution profile of SNPs . . . 35
4.2.4. Accessibility of gene for allelic expression analysis . . . 35
4.2.5. Effect of parental-specific pseudo-reference genome . . . 38
4.2.5.1. Improvement in read mapping . . . 38
4.2.5.2. Enhanced accuracy of expression estimation . . . 39
4.2.5.3. Influence of reference genome on differential expression analysis . . . 40
4.2.6. Pseudo-reference genome construction and assessment . . . 43
4.2.7. Expression estimation using SNPs . . . 44
4.2.8. Significance of gene region for DE analysis . . . 48
4.2.9. Benchmarking criteria for ASE analysis . . . 49

4.3. ASE analysis 53
4.3.1. Cis-regulatory mechanisms contribute to large scale expression divergence 56
4.3.2. Trans-effects are more unique to accessions 58
4.3.3. Comparison of RNA-Seq and Pyrosequencing in determination of allelic bias 59
4.3.4. Inheritance patterns of expression 61
4.3.5. Parent-of-origin effect 63
4.4. Metabolite analysis 64
4.4.1. Primary metabolite profiling by GC/TOF-MS 64
4.4.2. Metabolite data analysis 64
4.4.3. Pathway analysis 65
4.4.4. Correlation analysis of metabolite and transcriptomic profiles . . 66
4.4.5. Hierarchical clustering analysis to identify similarity based patterns 66
4.4.6. Multivariate analysis of metabolite levels to determine variation . 68
4.4.7. Proline concentration 68
4.4.8. Strong correlation of expression of P5C1 proline pathway gene and proline accumulation/concentration levels 69

5. CONCLUSIONS **71**

6. KEY LIMITATIONS **73**

7. FUTURE PROSPECTS **74**

REFERENCES **75**

LIST OF TABLES **82**

LIST OF FIGURES **83**

LIST OF ABBREVIATIONS **85**

SUPPLEMENTARY TABLES **87**

ARTICLES PUBLISHED DURING Ph.D. **90**

ACKNOWLEDGEMENTS **92**

DEDICATION **96**

ERKLÄRUNG **98**

CURRICULUM VITAE **100**

1. INTRODUCTION

1.1. Natural variation

Differential Gene expression is an important source of phenotypic variation of a given trait in individuals of a population (Kliebenstein *et al.*, 2006; West *et al.*, 2007). These variations are the basis of plant science research to understand plant-fitness, response to environmental cues, evolution, natural selection during domestication, and hybridization. Plant breeders and molecular biologists use natural variation as a tool to identify novel functional genes, understand the genetic factors influencing the heredity and intensity of traits. Different research studies aim to decipher the molecular mechanisms underlying natural phenotypic variation and to understand the interactions between different genetic factors. Farmers and plant scientists have been utilizing such information to select for desirable traits and develop new resistant varieties of crops with increased yield, improved fruit quality, and flavor with enhanced shelf life. Several reported methodologies have simplified detecting the genetic basis of phenotypic variation. However, it remains a bottleneck to uncover the underlying molecular mechanisms.

1.2. QTL studies

QTL analysis is carried out to determine the genetic regions controlling variation in a phenotype in a segregating population. Quantitative traits are mapped to chromosomal regions responsible for the variation of the given trait, which is called Quantitative loci or QTL. To detect or map QTLs requires considerable efforts, and a large number of individuals from segregating populations such as F2s, Recombinant Inbred Lines (RILs), Near-Isogenic Lines (NILs) or Isogenic Lines (ILs).

1.3. Expression-QTL analysis

Gene transcript levels show considerable tissue- and stage-specific variation under various environmental conditions. Variable expression of different genes enables enhanced plant response to various stimuli, and high survival rate under different stress conditions. Several studies observed quantitative genetic variation for gene expression levels. Jansen and Nap utilized the fact and established the concept of genetical genomics (Jansen and Nap, 2001). The expression levels are considered a quantitative trait, and the QTL analysis is performed to determine the genomic region controlling the expression (expression QTL or eQTL).

Analysis of gene expression enables understanding of the molecular mechanisms underlying the variation in expression. eQTL mapping in association with previously carried out QTL mapping help reduces the number of candidate genes for the QTL. A combination

of these two strategies also enables one to elucidate whether differences in expression are responsible for the observed phenotypic variation.

Numerous candidate genes involved in seed development in wheat and cell-wall digestion in Corn were identified using eQTLs in combination with QTL mapping (Shi *et al.*, 2007). In *Arabidopsis*, QTL mapping have been also used to study genetic interactions, and construct regulatory networks (Keurentjes *et al.*, 2007; West *et al.*, 2007).

1.4. Categories of eQTLs

eQTLs are further categorized as cis- or trans- depending on the location of the regulatory polymorphisms. Cis-eQTLs are caused by sequence polymorphisms located within or near the gene, for instance in the promoter region. Trans-eQTLs result from polymorphisms in other gene (s) that regulate the expression of the given gene via transcription factors.

Cis-eQTL investigation provides the most probable list of candidate genes which could be additionally used for QTL analysis. Polymorphisms in a receptor protein kinase, ERECTA, have been reported to have a pleiotropic effect on several traits. This is an example of trans-regulatory variants.

abundance of a gene. However, several genes may regulate the expression of a target gene. Hence, a trans-regulated variation in one or a few of several regulatory factors will have a small effect on the total gene expression.

On the other hand, one gene may control the expression of several genes, and exhibits a pleiotropic effect on various traits. A polymorphism in such a regulatory gene will influence the expression of several coregulated genes. Thus, trans-regulatory variants may have a large number of small-effects.

Global expression profiles of *Arabidopsis thaliana* in two independent studies for RILs of different accessions reported significantly different number of eQTLs, and relative cis-trans- contribution. For L*er* x Cvi 4000 eQTLs (Keurentjes *et al.*, 2007), and Bay-0 x Sha about 36000 eQTLs (West *et al.*, 2007) were identified. In L*er* x Cvi, cis-, and trans-regulatory variants were found to be equally abundant. However, in Bay-0 x Sha, 86% eQTLs were due to trans-effects, and only 14% changes were attributed to cis.

This may have been the result of the difference in the number of lines, and replicates considered for each study (160 RILs with one replicate each in L*er* x Cvi, and 211 RILS

with two independent replicates per RIL in Bay-0 x Sha).

A similar study in barley used one replicate per 139 lines reported 28-39 % cis eQTLs (Potokina *et al.*, 2008), while in maize 57-70% of cis eQTLs were reported (Stupar *et al.*, 2007; Guo *et al.*, 2004, 2008). In *Populus* 57% of 30 genes considered showed cis-regulated variation (Street, 2006).

1.5. ASE analysis

The relative contribution of alleles to the total transcript level in an F1 hybrid is known as Allele-Specific Expression (ASE). An alternative approach to conventional cis-eQTL mapping, which requires a large number of recombinant lines, is to perform an ASE analysis for each gene on a genome-wide scale in F1 hybrid.

Cis-regulated expression variation in two different individuals of a species or two different species results in differential expression of alleles in their hybrid (Figure 1). ASE analysis carried out in F1 hybrids of two different accessions varying in a certain trait can reveal the contribution of cis- and trans-regulatory variations (Cowles *et al.*, 2002; Yan *et al.*, 2002). ASE in hybrids is believed to promote hybrid vigor and may be an important source of phenotypic variation (Guo *et al.*, 2006a). ASE analysis is performed by measuring differential allelic expression in a F1 heterozygous individual By comparing the allelic ratios in hybrids to the expression differences between parental lines, cis- and trans- regulation events can be (distinctly identified / distinguished). Allelic ratios of transcripts similar to the observed expression difference between parentals indicate presence of cis-regulatory variants (Figure 2). However, if there is no significant difference in allelic transcripts in hybrid but the genes shows differential expression in parents, it might be due to regulatory variants acting in trans (Wittkopp *et al.*, 2008). ASE is a useful methodology to determine allelic imbalance specific to tissue, environmental condition or developmental stage, and imprinting effect i.e. allelic imbalance due to parent-of-origin effects.

Allele specific expression seems to be a common phenomenon across organisms such as humans (Lo *et al.*, 2003), mice (Doss *et al.*, 2005), *Drosophila* (Wittkopp *et al.*, 2004), maize (Guo *et al.*, 2008), Barley (von Korff *et al.*, 2009), and *Arabidopsis* (Zhang and Borevitz, 2009).

ASE analysis has been used to uncover the underlying cis-, trans-, and cis-by-trans effects on the variable gene expression in *Drosophila* (Wittkopp *et al.*, 2004). Investigation of ASE pattern in embryo tissue of maize revealed that, of the genes found to exhibits ASE in F1 hybrids, 20% showed parent-of-origin effect (Springer and Stupar, 2007b). Maternal

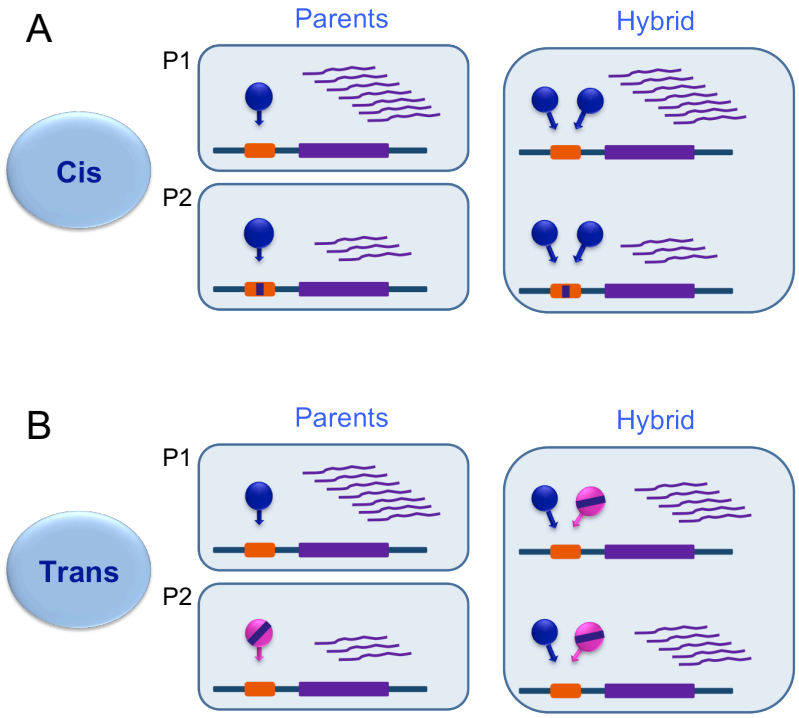


Figure 1. Schematic representation of cis- and trans-regulated expression differences. The image shows basic difference in (A) cis- and (B) trans-effects, by comparing differential parental expressions levels in parentals P1 and P2 (skyblue rectangle) to corresponding allelic transcripts in hybrid. The horizontal line represents genomic region with coding sequence (violet box) and promoter region (orange box). The relative transcript abundance is depicted as purple strands. Polymorphisms in the promoter (cis) cause allele-specific expression differences in heterozygous individual. The trans-regulatory variants do not show allelic imbalance.

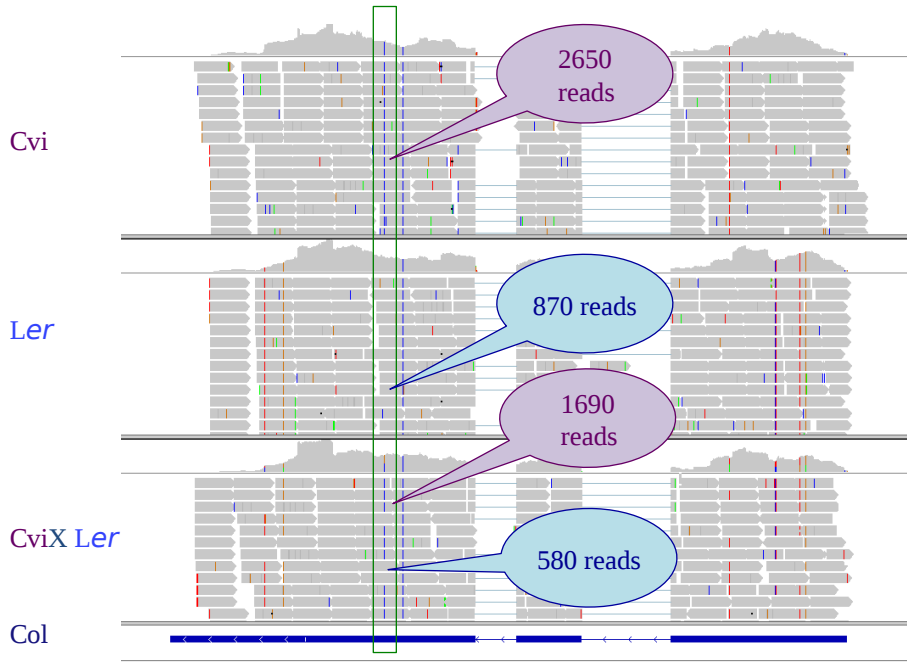


Figure 2. Allele-specific expression analysis for Cvi, L*er* and the F1 hybrid - Cvi x L*er*. Reads (grey boxes) map to reference (blue box at the bottom) Col-0, with conserved SNPs at particular location (blue vertical lines within green rectangle). In hybrid Cvi x L*er* some reads are with Cvi-specific SNPs while others correspond to L*er* SNPs. Read counts for each depicted with same colour scheme. Hybrid show 3: 1 ratio of Cvi and L*er* allele which is equivalent to expression difference between parents (3: 1 for Cvi: L*er*).

and paternal biased differential expression in maize endosperm has been documented. 24-31% of genes showed tissue-specific expression (Springer and Stupar, 2007b). Tissue-specific ASE has been reported in cotton (Adams and Wendel, 2005; Adams, 2007) and barley (von Korff *et al.*, 2009). Differential allelic imbalance in response to various stress conditions have also been examined in barley (von Korff *et al.*, 2009) and maize (Guo *et al.*, 2004; Springer and Stupar, 2007b; Stupar *et al.*, 2007).

1.6. Using RNA-Seq to detect ASE

High throughput sequencing technologies have enabled genome-wide scale studies. RNA-Seq is one of the most complex NGS applications. RNA-Seq allows relatively accurate determination of gene expression levels, alternative splicing, allele specific expression of transcripts, which are not easily accessible from previously utilised hybridisation based approaches. Furthermore, it allows the assessment of gene expression more precisely than with microarrays if sufficient coverage is achieved.

Background and cross-hybridization issues, and the inability to measure non-relative abundance of RNA transcripts does not allow detection of subtle changes in gene expression levels. In contrast to hybridisation-based methods, tag-based sequencing methods measure absolute transcript abundance without previous knowledge of gene sequences.

But the latter procedures are quite laborious, and have been of limited use.

It has now become possible with the use of RNA-Seq to characterize all transcripts expressed within a specific cell or tissue at a particular stage with great potential to determine correct splicing patterns, structure of genes, and expression profiles in physiological and pathological adaptation.

1.7. Advantages of RNA-Seq

1. Enable identification and quantification of transcripts of novel unannotated genes and spliced isoforms.
2. Provide evidence against incorrect annotated genes and exon/intron boundaries.
3. Produce low background signal.
4. Larger dynamic range of expression levels allows relatively accurate determination of difference in expression.
5. Have high levels of reproducibility for technical and biological replicates.

1.8. Advantages of Pyrosequencing

Pyrosequencing is the DNA sequencing methodology which depends on detection of inorganic pyrophosphates (PPi) released during DNA polymerisation. The released amount of pyrophosphates are emitted as light signal at each enzymatic processes.

Pyrosequencing is rapid and can generate massive high-throughput sequences upto 200bp for low costs. It offers a significant advantage in determining allelic imbalance (Korir and Seoighe, 2014) and imprinting effects (Tabano *et al.*, 2015). In addition, it provides a robust control to compare the accuracy of expression quantification method using RNA-Seq data. Therefore, to test and improve power of our ASE method, we have used the pyrosequencing results to validate allelic ratio of randomly selected genes.

1.9. Caveats of gene and allelic expression analysis

(i) **Reference genome**

Mapping of reads against an available genome or transcriptome sequence reference is the first step to measure the expression from RNA-seq data. The limitation of the method arises due to unavailability of well-annotated genomes for most of the species and accessions and therefore, the annotated genome of nearest species is used as a reference. However, the species under consideration may differ markedly from or be very similar to the reference selected, as exemplified by the number and positions of polymorphisms. When two species are compared against each other using non-specific reference, the reads from a similar species are more likely to map

than the reads from the distant species. This bias in mapping reduces the accuracy of expression estimation. Hence, it becomes important to evaluate the impact of reference genome in intra-specific and intra-specific comparisons (Degner *et al.*, 2009).

(ii) **Exonic and intronic reads** (Ameur *et al.*, 2011)
Since reads are obtained from cDNA of processed spliced mRNA, the expression measurement should account for reads mapped in exon-exon junctions. However, several reads are observed to map across intergenic regions and introns creating noise in the expression (Figure 3A). The intronic reads may indicate partially-to-unprocessed mRNAs, background noise, genomic contamination, functional non-coding RNAs, or/and alternatively splicing isoforms (Figure 3B). It is argued that the exonic junctions are the most reliable region for estimation of expression. To test the argument and to select a useful metric, a comparative evaluation against control was performed using exons, introns and UTRs reads, reads mapping to exons and UTRs, and reads over only exon-exon junctions.

(iii) **SNP abundance** (Griffith *et al.*, 2010)
High number of SNPs improve the power of method for allelic transcript measure. The inter-specific comparisons have the power of higher number of differential bi-allelic sites. However, the polymorphism rate is relatively lower for the intra-specific accessions. In addition, not all genes can be evaluated for expression difference owing to lack of informative SNPs. Also a major number of genes have single SNPs and hence it affects the reliable measure across allelic transcripts. A comparison have been performed to evaluate the impact of analysis using read measure across single SNP and multiple SNPs.

(iv) **SNP location and read coverage** (Griffith *et al.*, 2010)
Depending on the SNP location within or outside gene coding regions, within exons or UTRs, there might be differences in the coverage over the SNPs (Figure 4). Especially more reads tend to map to exonic locations and the read mappability decrease over intronic regions, $5'$ and $3'$ UTRs.

Additionally, for the allelic estimation within hybrids, the distinguishable SNPs must be sufficiently mapped over by the allelic reads. However, the SNPs with low allelic coverage within hybrids and accessions, are rendered unaccountable for the expression quantifiaction. SNP distribution profile has been sketched across the genome to examine the accessibility abundance and coverage.

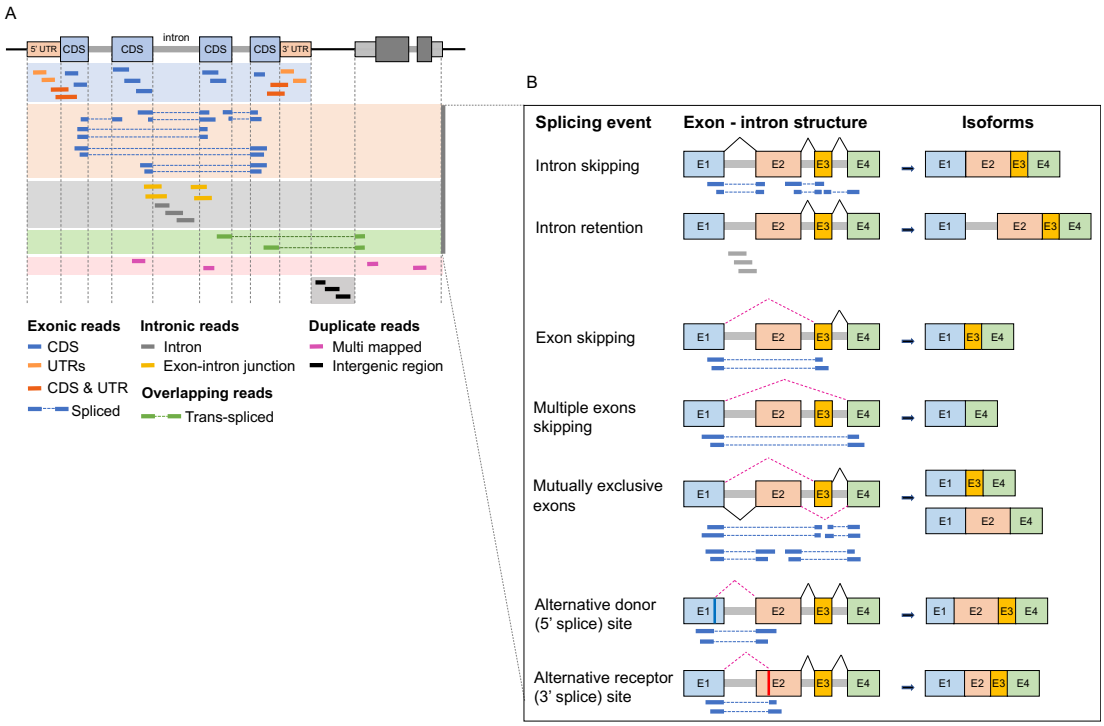


Figure 3. Categorisation of aligned reads. (A) The figure is an schematic representation of mapped transcriptomic reads onto the genome shown as black line connecting the large rectangles depicting segments namely, CDS, UTRs, introns. The genomic coordinates exons and introns represent the exon-intron boundries highlighted by dashed margin lines. Reads that completely map within the margins can be easily assigned to a location, and classified as exonic, intronic, or duplicate reads. Others may splice over and map onto exons of same gene, or overlap exons of other genes. The categories of mapped reads are shown in the legend. (B) Schematic representation of splicing events listed along the first column with spliced reads indicating the spliced location. The third column show the resulting alternate isoforms.

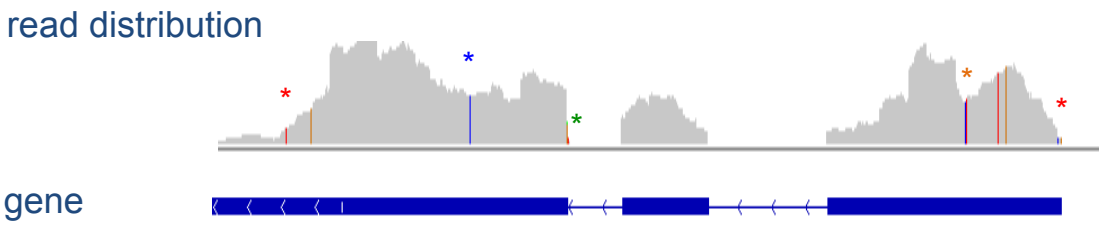


Figure 4. Position-dependent read distribution pattern over SNPs. A snapshot of gene expression profile with multiple SNPs. The read intensity depends on the location of SNP within the gene. The pattern of location effect is illustrated by the coverage plot marked by color coded stars. (red) depicts the SNPs at terminal gene region, towards which the read coverage gradually declines and thus, the very few reads cover the SNPs. (blue) represents SNPs within the inner exonic regions, which are usually read-dense, due to high and uniform coverage towards the center. (green) corresponds to the SNPs at exon-intron junctions, supported by fewer reads spanning the junctions, hence a lower coverage. (orange) points to SNP intense regions, wherein, the coverage may be highly restrained by the mismatches allowed per read while mapping.

(v) **Gene-length** (Oshlack *et al.*, 2009)
The expression level is quantified by counting the number of reads mapped over gene. However, this measure is directly proportional to the reads mapped and expected expression of gene (Rapaport *et al.*, 2013). The longer an expressed gene is, the higher the probability of it being sequenced is (due to the nonlinear amplification of fragments during RNASeq) and therefore higher the number of sequenced reads. Given the same amount of expression for two genes, one significantly shorter than the other in length, this bias reduces the power to detect differential expression of the shorter gene due to low coverage. Hence, the expression level has to be normalised to account for gene length bias as discussed in Mortazavi *et al.* (2008).

(vi) **Magnitude of expression change** (Oshlack *et al.*, 2009)
The levels by which expression changes for genes across species, defines the Efficiency of expression analysis method is largely decided by the detectable magnitude of expression deviation. The genes showing extremely high magnitude of change are most likely to be detected. However, the statistical power to detect the small scale changes might in turn be driven by several factors including the biological variation and the experimental setup. Hence, it becomes important to evaluate the impact on magnitude of variation for measuring expression and further use it for quantification of differential allelic transcripts within hybrids.

(vii) **Differential coverage over exons and SNPs**
The expression is measured by the reads mapped over the exons and used for determination of differential expression. However, to quantify the allelic transcripts within hybrids, the SNPs are used to distinguish between the allelic reads. Coverage over SNPs is driven by multiple factors including the number of SNPs, the allowed number of mismatches for a read to map, the length of the gene and the overall gene expression. SNP sites within highly expressed genes are more likely to be mapped by reads.

In addition, the SNPs in the intronic regions, $5'$- and $3'$-UTRs have fewer reads mapping to them unlike the SNPs within exonic regions. Over-expressed genes with single SNP should have extremely high coverage over the SNPs. Ideally, within the genes with multiple SNPs the reads should be distributed more or less uniformly over most SNPs.

However, it is very likely that the SNPs are unevenly mapped, with some overrepresented, while others underrepresented. Hypothetically, the deleterious unstable

mutations should not be very well represented by the reads and thus, should have relatively lower coverage. Since, the level of gene expression, the abundance and location of SNP defines the read coverage over the SNPs, it may be equivalent to or differ substantially from the representation at the exons. The statistical power of the method to determine allelic expression relies strongly on the proportion of similarity in the measure of transcripts from the SNPs and exons. Hence, a comparative assessment for high similarity in the exonic and SNP estimate of gene expression is absolutely inevitable.

1.10. Plant system used

Arabidopsis thaliana also known as thale cress is a member of Brassicaceae family. It is a small annual flowering plant with short life cycle. It was the first plant genome to be sequenced, has a small genome of about 125 Mb distributed in 5 chromosomes and contains approximately 27,000 genes. It shows a wide geographical distribution, found growing in Asia, Europe, and northwestern Africa (Figure 5). 1135 natural accessions of *Arabidopsis thaliana* have been characterized and genomes are made available to community as part of 1001 Genome Project (https://1001genomes.org/).

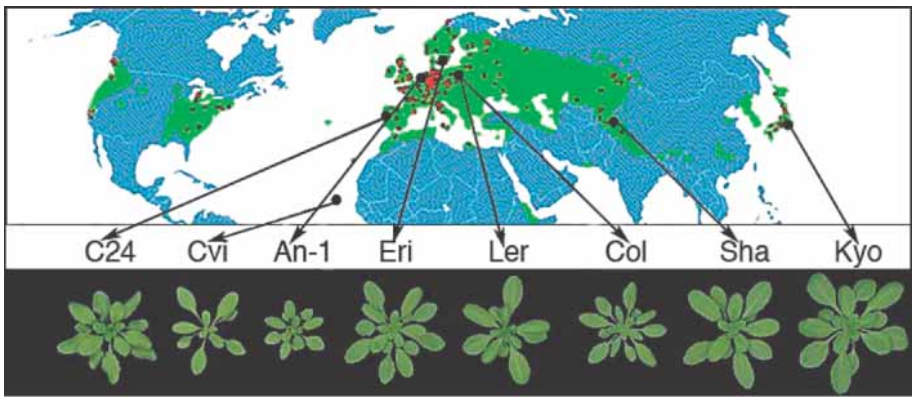


Figure 5. The geographical distribution of *Arabidopsis thaliana*. The dots marked on the map depict location of origin for the various accessions, which show phenotypic variation in shape, size and number of leaves (Van Norman and Benfey, 2009).

Reproduction in *Arabidopsis* occurs primarily by self-pollination. Therefore, geographically isolated *Arabidopsis* population exhibits genetic and phenotypic differences which are often adaptive to specific environment. As a result, immense diversity in morphology such as number, shape and size of leaves (Figure 5), physiology e.g. flowering time, disease resistance, and development is observed across various accessions. Wild type ecotypes such as Columbia (Col-0) and Cape Verde Islands (Cvi), and mutant lines such as Landsberg *erecta* (L*er)* are widely used for genetic and genomic studies.

1.11. Parent-of-origin effect on allelic expression

Unlike the allelic expression (ASE) caused by cis- or trans- variants, genomic imprinting is mainly an epigenetic phenomenon resulting in biased allelic expression depending on parent-of-origin. Imprinting have been exhaustively studied in humans (Tycko, 2010), mice (Shen, 2014), castor (Xu *et al.*, 2014), rice (Chen and Begcy, 2020), maize (Guo *et al.*, 2006b), and *Arabidopsis thaliana* (Pignatta *et al.*, 2014).

Depending on the maternally or paternally derived allele, the expression may be completely silenced resulting in monoallelic expression, or activated, enhanced, or supressed with differentially expressed alleles. Hence the allelic imbalance can result from parental-induced bias or genetic variants. To distinguish between the underlying cause of allelic bias in expression, comparison of allelic expression in reciprocal crosses is performed. For this purpose ASE analysis in RNA-Seq data from reciprocal hybrids proves to be useful in determination of parental bias of allelic expression. Further, the observed patterns are be differentiated into the following categories (Figure 6) :

i. Genomic imprinting, where expression is biased towards maternal or paternal allele (plausible imprinting effect), further categorised as:
    a. MEG, maternally expressed (paternally imprinted) genes.
    b. PEG paternally expressed (maternally imprinted) genes.
ii. Genetic ASE, where allelic bias is unrestricted by parent-of-origin allele.

Imprinting can be caused due to differential methylation in tissue-dependent manner. Thus, different tissues may exhibits varying degrees of imprinting effect. In this study, we have used only the leaf tissues of *Arabidopsis thaliana* accessions. Hence, it might limit the capacity of our method to determine any significant biologically relevant imprinting effects. Previous reports in *Arabidopsis thaliana* (Gehring *et al.*, 2006) have highlighted the importance of maternal and paternal allele regulation in the initial phases of seed development and hence the tid due to specific imprinting effects.

1.12. Inheritance patterns of expression divergence

The inheritance patterns of expression divergence within hybrids with respect to corresponding parentals may help in determination of crucial regulatory effects which result in beneficial phenotypical changes measured in terms of plant growth, weight, survival and adaptation to environment and metabolite content, grain yield etc. Thus, comparative study of inheritance patterns can be used to detect events of heterosis (Figure 7).

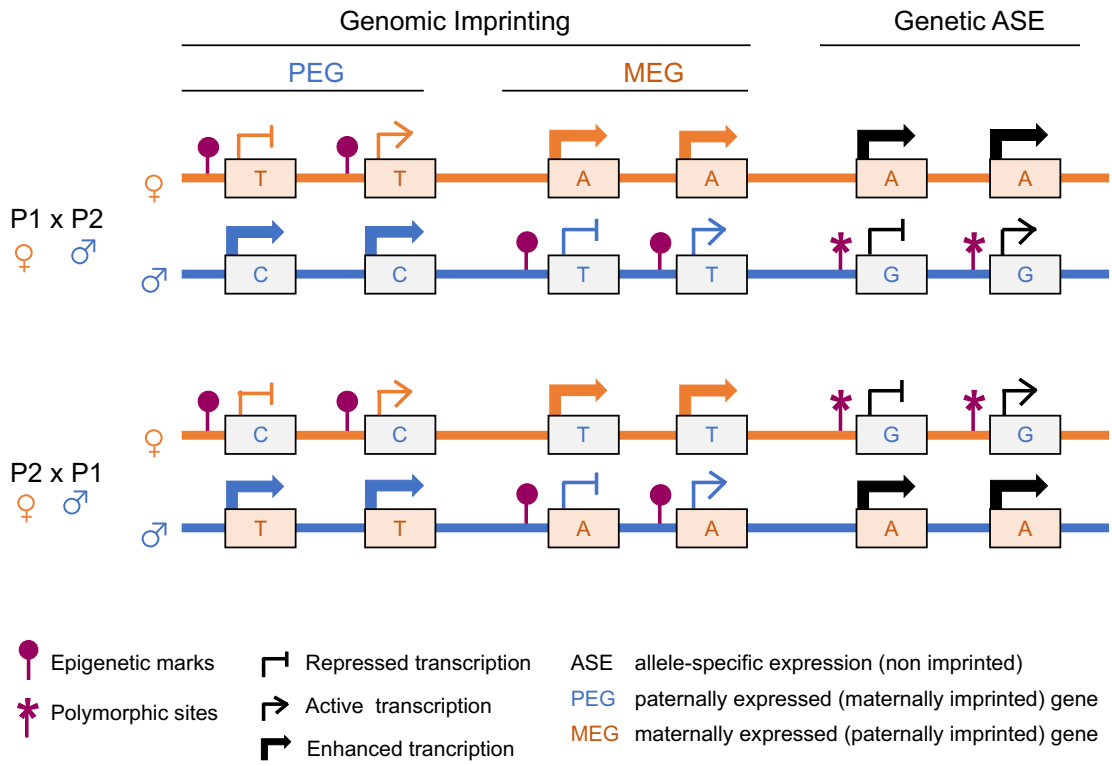


Figure 6. Schematic representation of parent-of-origin effect on allelic expression. The figure depicts the allelic bias due to genomic imprinting, and genetic allele specific expression effect. P1 (TTAAAA) and P2 (CCTTGG) represent two parental genotypes. P1 x P2 and P2 x P1 represent reciprocal crosses. Activation and enhanced transcription are represented by arrows, and repression by blocked arrows. In each cross, maternal allele is depicted by orange and paternal allele by blue colour. Non-imprinted or alleles without parental bias (allele-specific effect) are highlighted by black colour. Transcription of MEG, maternally expressed (paternally imprinted) are shown by orange colour, and paternally expressed (maternally imprinted) in blue colour.

Depending on degree of similarity between parentals (P1, P2) expression, the genes can be divided into two categories:

I. P1≈P2, with no difference in parental expression, and
II. P1≠P2, with significantly different parental expression.

Depending on the gene expression in F1 hybrid, the genes with similar parental expression (P1≈P2) can be differentiated into the following:

A. additive, F1≈MPV (mid parent value), where expression in F1 hybrid is equivalent to average of the two parental expression, and
B. non-additive, F1≠MPV. Non-additive events can then be categorised as:
    i. activation, (F1>MPV), where the F1 value is exceeds MPV.
    ii. repression, (F1<MPV), where the F1 values is less than MPV.

For the genes with significantly different expression in parentals (P1≠P2) the inheritance patterns can be also classified as following heterosis events:

A. additive, F1≈MPV.
B. non-additive, F1≠MPV, which can be further subcategoried as:
    i. dominance, where the hybrid expression is equivalent to either the
        a. over expressed high-parent value (F1≈HP), or
        b. underexpressed low-parent value (F1≈LP).
    ii. overdominant, (F1>HP), where the hybrid expression is higher than high-parent value,
    iii. underdominant, (F1<LP), where the hybrid expresson is lower than the lower-parent value.

Both overdominance and underdominance are categorised as transgressive expression, where the hybrid expression value lying outside the range of parental expression.

1.13. Primary metabolites profiling

Metabolite profiling determines, characterises and quantifies large number of metabolites in a robust manner by established protocols for extraction, separation and analysis of cellular chemicals. Metabolomics refers to the identification and measurement of all metabolites within a system. Several methods and techniques for comprehensive, selective, and

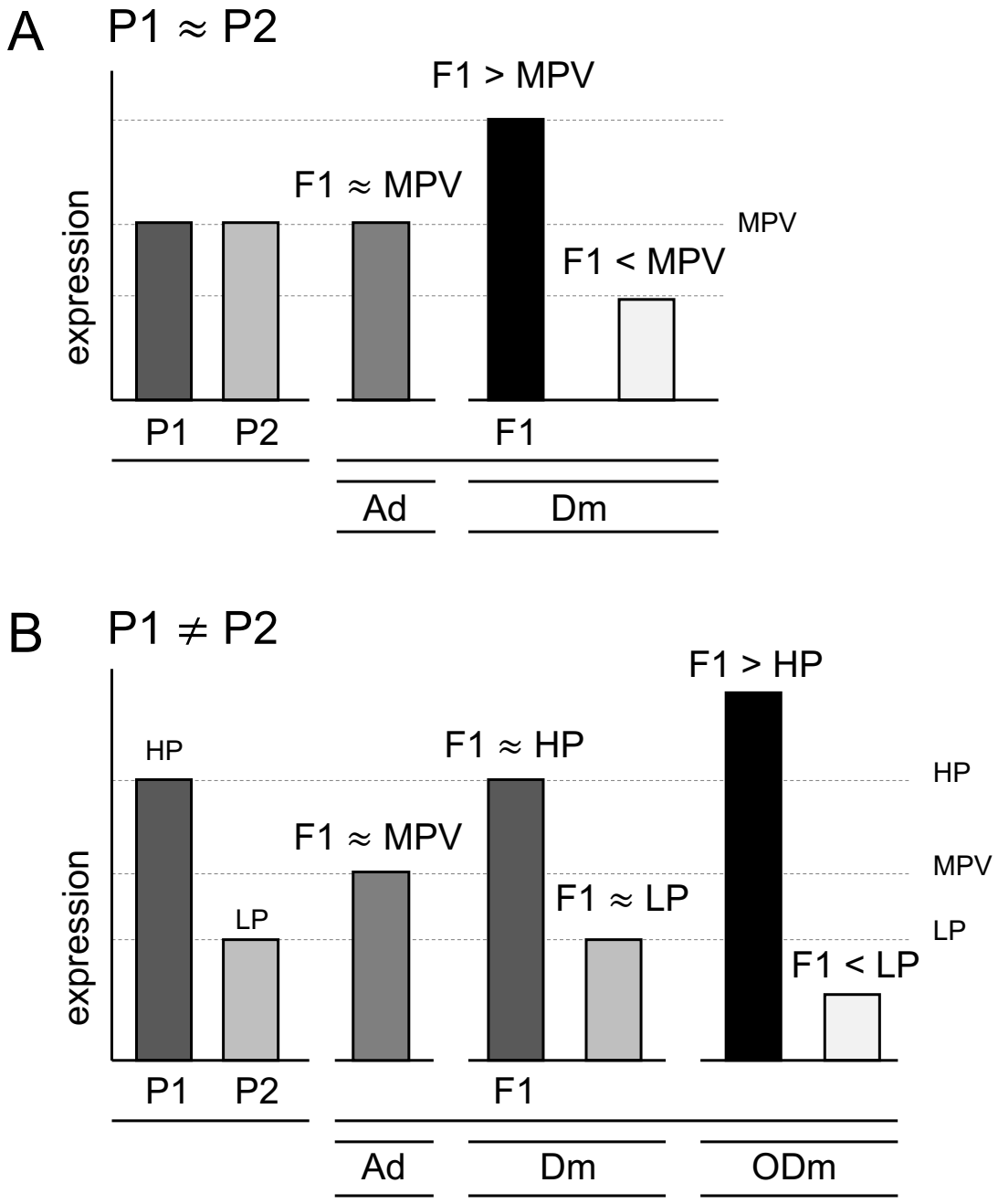


Figure 7. Schematic representation of gene expression inheritance patterns categories. The figure depicts the heterosis events depending on the expression pattern of hybrids compared to the two parentals. (A) P1≈P2, depicts genes for which there is no significant difference in the expression values of two parentals P1 and P2. (B) P1≠P2, depicts genes for which there is significant difference in the expression values of two parentals P1 and P2. Further, F1, represents hypothetical expression values in F1 hybrid. MPV, the mean parental value, average of two parental expression values. HP, higher parent, parental with higher value for gene expression. LP, lower parent, parental with lower value of gene expression.

sensitive measurement of most possible metabolites have been discussed by Schauer *et al.* (2005) in addition to overview of technologies for metabolite profiling, limitations, and data analysis methods in plant biology. Analysis detecting correlations can help determine unidentified relationships while principal component analysis can be help elucidate hidden patterns from variation in measured metabolite levels (Tohge, 2009).

Integrated data analysis of high throughput transcriptomics and metabolomics data can help in determination of interdependent patterns and coregulatory networks of genes and metabolic pathways. Recent surge in plant studies using integrated approaches have highlighted the crucial role of efficient statistical and informatics tools to analysis plant-metabolite data in a biological context. Major databases of metabolic pathways like AraCyc (*http://www.arabidopsis.org/tools/aracyc*; (Mueller *et al.*, 2003)), PMN (*http://www.plantcyc.org/*) and KEGG (*https://www.genome.jp/kegg/*; (Kanehisa and Goto, 2000)), provide an exhaustive and inevitable resource for crucial analysis and understanding of complex pathways.

2. SCIENTIFIC OBJECTIVES

The major scientific objective of this project was to decipher the evolutionary patterns underlying cis- and trans- regulatory variants and their reponses to the environment by performing ASE analysis in *Arabidopsis thaliana* hybrids.

2.1. Development of ASE analysis method

Our foremost objective was to determine and define an optimal methodology to measure ASE in RNA-Seq data. We compared real and simulated RNA-Seq datasets from leaf tissue of three biological replicates each of Cvi, L*er* and Cvi x L*er* for this project. For developing a robust ASE analysis workflow addressing the major caveats addressed in the section 1.9. on the page 7, the following tasks were carried out:

(i) Establish a minimal number of polymorphisms required to accurately detect ASE.
(ii) Establish a protocol to generate a near perfect parental-specific psuedo reference genome.
(iii) Compare the effect of reference and pseudo-reference on the expression quantification and establish the reference to be used for downstream ASE analysis.
(iv) Compare expression calculated from reads mapped over SNPs and exons independently to establish standard protocol for read counting.
(v) Use statistical tests to determine differential expression using Cvi and L*er* data.

2.2. Determination of genome-wide ASE in *Arabidopsis thaliana*

The subsequent objective was to use the successfully benchmarked and established ASE method for comparative evaluation of the RNA-Seq data from leaf tissues of the plants samples mentioned in section 3.4.1.. Thereafter, the ASE results were to be used for quantitative comparison and characterisation of functional classes of genes exhibiting the expression differences.

2.2.1. Determination of parental effect on allelic expression

Further objective was to distinguish genetic allelic changes from imprinting effects reflecting the parent-of-origin effect on allelic expression. A simplistic approach to achieve this goal would be used and ASE ratios in reciprocal hybrids will be compared.

2.2.2. Determination of expression inheritance patterns

With sufficient expression data from the parentals and corresponding hybrids, it is achievable to determine and quantify inheritance patterns of expression, and further ascertain heterosis effects. Thereafter, an integrated analysis of ASE and heterosis effects will reveal the relative contribution of regulatory variations.

2.3. Integrated transcriptomics and metabolomics analysis

Changes in metabolite concentration levels are direct responses of plants to developmental and environmental cues. Thus, there exists an enormous amount of diversity in the metabolite content at different stages, and variable tissues. Comparison of expression and metabolite profiles across same samples (described in section 3.4.1.) could unravel the complex associations. Therefore, our final objective was to carry out an exhaustive analysis of primary metabolites profiled across the same plants samples. Further, to assess correlations between divergence in expression and metabolite levels. With this exercise, our end goal was to identify candidate genes with strong correlation between genetically induced expression divergence and phenotypic (metabolite) variation.

The further sections describe at length, the experimental and analytical approaches to address each objective in a methodological approach. The results have been duly reported and discussed in next segments.

3. MATERIALS AND METHODS

3.1. Method development and benchmarking criteria

The major aim of the present study was to perform genome-wide analysis for the determination of Allele-Specific Expression (ASE) patterns using RNA-seq data. Therefore, the primary objective was to develop an efficient method with high sensitivity and specificity, specifically customised for our data-set and species under the plan. The major caveats and challenges in expression and allelic ratio estimation from NGS transcriptomics data (enlisted in subsection 1.9. on page 7) were addressed during method development. To examine the implication on expression variation quantification, and improve the efficiency and accuracy of the method, simulated values of differential expression were used as a measure of control. The deviation from the control was analyzed for the parameters selected to minimize the false positive rate.

The results were used to standardise the parameters including the threshold of gene expression at parental level, the minimum read count over SNPs, and minimum fold change at the exonic levels. The pattern of exonic and SNP congruity has been assessed for the selected criteria to determine the threshold values for detectable and robust similarity.

Additionally, the workflow allowed exclusion of genes with inconsistent and incomparable expression estimates at SNP and exonic level. The fine selection of genes with high degree of concordance for transcript abundance in exons and SNP increases the predictability of variation at allelic level. Thus, to develop an efficient ASE prediction workflow with enhanced accuracy for true positive estimation, major limiting factors were addressed in great details.

3.2. Reported study

The study reported in this thesis is intra-specific and the accessions under consideration lack well-annotated genome except for a recently annotated genome for L*er*. Hence, the available genome of closest relative accession *Arabidopsis thaliana* Columbia (Col-0) from TAIR-10 version is used. Selecting L*er* genome as reference would strongly bias the results as mapping of reads from other accessions would differ markedly from the mapping of L*er* reads to the L*er* genome.

This would influence the analysis of differential expression. Hence, selection of Col-0 genome, which is specific to neither of the parental accessions per hybrid, would rule out uni-directional bias to some extent. Although evidently, the level of similarity to relative accessions might pose a significant challenge. Hence a step-by-step approach was used for

comprehensive evaluation of method using Col-0 genome and accession-specific pseudo-reference genomes for expression analysis.

Remarkable advances in the high-throughput transcript quantification methods in recent years have given rise to well-established tools which address several caveats in expression analysis. However, the tools to determine patterns associated with expression divergence and underlying mechanisms, especially working on RNA-seq, are limited in number and restricted to specificity of sample under consideration, or still being improved upon. In the present project ASE analysis approach using hybrid background (Cowles *et al.*, 2002; Yan *et al.*, 2002) has been carried out to determine the underlying regulatory patterns. The patterns have been further investigated to distinguish the expression variation regulated by genetic variants in cis- from trans- as proposed and established by Wittkopp *et al.* (2004).

3.3. Simulation analysis

To decide the requisite metrics for the analysis method differentially expressed RNA-seq data was simulated using Flux-Simulator v1.1 (Griebel *et al.*, 2012). The gene expression levels and magnitude of expression deviation measured from the simulated reads were used as control. It was thereafter used to compare the sensitivity of the expression quantification method in downstream analysis.

3.3.1. Plant material and RNA-seq data

RNA extracted from leaf tissue of *Arabidopsis thaliana* accessions Cape Verde Island (Cvi) and Landsberg *erecta* (L*er*) was sequenced for nearly 10-15 million single end Illumina 96-mer reads for four replicates of Cvi and five of L*er* replicate. Nearly 6-12 million of 60-90 mer reads were retained after trimming adapter and low quality bases at the ends.

3.3.2. Sequence alignment

To quantify the transcript abundance and differential gene expression the reads were aligned using TopHat (Trapnell, Pachter and Salzberg, 2009; Trapnell, Roberts, Goff, Pertea, Kim, Kelley, Pimentel, Salzberg, Rinn and Pachter, 2012) release 2.0.6 to to TAIR10 version of *Arabidopsis thaliana* reference genome of Columbia (Col-0) with the following parameters:

–splice-mismatches 2 –max-multihits 1 –max-insertion-length 12 –max-deletion-length 12 –read-gap-length 12 –read-edit-dist 12 –segment-mismatches 3 –library-type fr-unstranded

3.3.3. Polymorphism detection

To detect the polymorphisms of Cvi and L*er* against Col-0, customised Linux Shell and Perl scripts, command line tools from Picard and GATK were used. Firstly, reads mapping to multiple locations in the genome were discarded. Further, duplicated reads were selectively removed using MarkDuplicates from Picard (*http://picard.sourceforge.net*). Following this, the indels were realigned using the IndelRealigner tool in GATK (DePristo *et al.*, 2011).

Then, GATK variant discovery tool, UnifiedGenotyper, was used to determine sequence variants between each accession and Col-0. Using a custom Perl script, reads were divided in five 20 mer segments from left to right. Variants covered by the same segment in all reads, or those supported by first or the last segments of a single read were discarded. The reads located in regions flanking large indels or rearrangements were filtered using a custom perl script. Variants covered by the same segment in all reads, or that are supported by segments 1 or 5 in a single read are discarded. Bi-allelic variants with a phred-scaled variant quality higher than 30 were selected for further analysis. Finally, the homozygous polymorphisms of selected variants, located in exons, were chosen to calculate allele-specific expression and construction of pseudo reference genome.

3.3.4. Pseudo-reference genome construction

The parental-specific pseudo-reference genome for Cvi and L*er* were created independently by incorporating the polymorphisms detected in Cvi and L*er* against Col-0 including Single Nucleotide Polymorphisms (SNPs), Insertions and Deletions (InDels) in the Col-0 genome.The pseudo-reference is the genome with higher sequence similarity to the respective accession in contrast to a divergent reference genome.

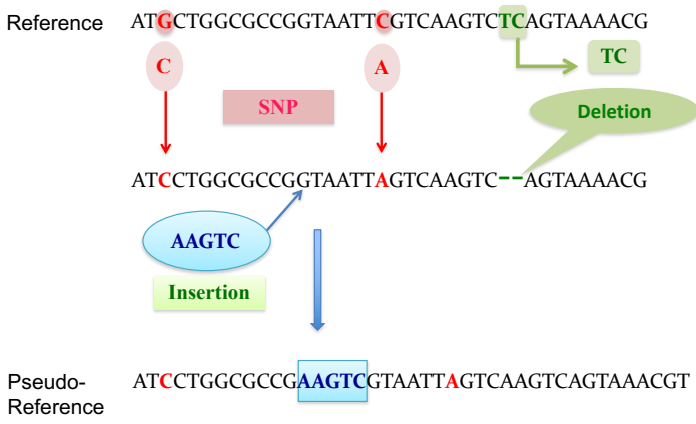


Figure 8. Schema for construction of parent-specific pseudo-reference genome. The reference genome was altered to create accession-specific genome sequence by using the polymorphic sites against the particular accession. The bases are modified at the SNP loci and the insertions are incorporated at the specific locations. The sites determined as deletions are removed from the genome.

3.3.5. Evaluation of impact of reference genome

To determine the effect of genome of reference over the estimation of expression, a comparative evaluation of Col-0 and accession-specific pseudo-reference genomes was carried out. The genomes were independently assessed in terms of sensitivity for the quantification of gene expression and estimation of differentially expressed genes.

3.3.6. Simulation of differentially expressed RNA-seq data from pseudo-reference genomes

RNA-seq data for Cvi and L*er* pseudo-reference genomes was simulated using Flux-Simulator v1.1 (Griebel *et al.*, 2012). Expression profile was tweaked to simulate true differential expression between the generated transcripts. Nearly 50 million RNA-seq reads were generated for each of the three replicates of Cvi and L*er*.

3.3.7. Comparison of reference and pseudo-reference genome

To determine the effect of reference genome the simulated RNA-seq reads from Cvi and L*er* were mapped against Col-0 and accession-specific pseudo-references respectively. Reads that mapped to unique position in the genome were considered for further analysis. The differential gene expression between Cvi and L*er* was quantified independently from Col-0 and pseudo-reference genomes. The observed results were compared to the expected simulated expression profile to evaluate the accuracy and sensitivity of each genome. Perl package Bio::DB::Sam was used for interacting with SAM and BAM alginment files. Customised Perl scripts were used for read count and pseudo-reference construction.

3.4. Allele-specific expression analysis

3.4.1. Plant samples

To perform Allele-specific expression (ASE) analysis, natural accessions of *Arabidopsis thaliana* Antwerpen (An-1), Borky (Bor-4), Burren (Bur-0), Knox (Knox-10), Shakdara (Sha),and Landsberg *erecta* (L*er*-0) have been used. Reciprocal hybrids were generated using L*er* as paternal and maternal parent against each of the other 5 accessions. The F1 hybrid plants An-1 x L*er* and L*er* x An-1, Bor-4 x L*er* and L*er* x Bor-4, Bur-0 x L*er* and L*er* x Bur-0, Knox-10 x L*er* and L*er* x Knox-10, Sha x L*er* and L*er* x Sha, were examined independently for ASE analysis and then compared to determine any plausible parentally-biased allelic expression.

Five replicates from each of the aforementioned accessions and their reciprocal hybrids were grown together in an environmental chamber set to 12 hours light 12 hours dark

photoperiods. leaf samples were collected at bolting time and frozen in liquid nitrogen. Samples were then pulverized in a mortar and divided in three aliquots. One aliquot from three biological replicates for each genotype were used for RNA-seq. One aliquot from one replicate of the parental lines was used for re-sequencing. Aliquots of all 5 biological replicates for all genotypes were used for metabolite analysis.

Single-end reads of length 96 bases were generated with an average quality of 35 per read. Sequence depth varied for the accessions and replicates. The number of reads sequenced varied from 12.2 million to 66.3 million with and an average of 19.1 million for parentals and 45.4 million for hybrids.

3.4.2. RNA-seq library preparation

For RNAseq experiments, total RNA was isolated using Qiagen RNeasy Plant Mini Kit (Qiagen Sciences, Maryland, USA) according to the manufacturer's instructions. RNA-seq libraries were obtained using the Illumina Truseq protocol according to manufacturer's instructions. The sequenced RNA-Seq reads corresponding to the plant samples mentioned in section 3.4.1. have been summarised in the Figure 9.

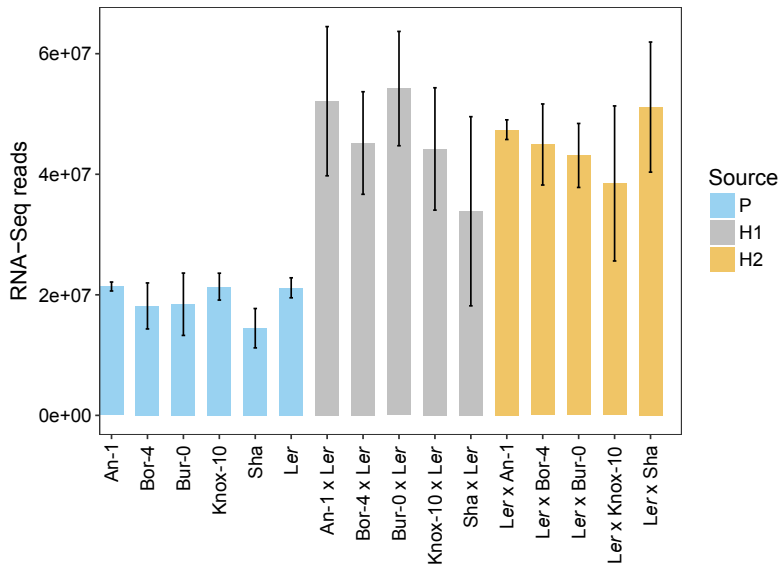


Figure 9. Overview of RNA-Seq sequenced from *Arabidopsis thaliana* accessions and reciprocal hybrids. The barplot summarises the transcriptomic reads obtained from plant samples; P (parental accessions), reciprocal hybrids H1 (P(m) x L*er*(p)), and H2 (L*er*(m) x P(p)).

3.4.3. Re-sequencing library preparation

For re-sequencing of the parental accessions, genomic DNA was isolated using Qiagen DNeasy Plant Mini Kit (Qiagen Sciences, Maryland, USA) according to the manufacturer's instructions. Sequencing libraries were obtained using the Illumina Truseq protocol according to manufacturer's instructions.

3.4.4. Polymorphism detection

Re-sequenced datasets were subjected to polymorphism detection using DNA reads aligned to the TAIR10 reference for *Arabidopsis thaliana*. For this DNA reads were aligned using Bowtie2 with default parameters (Langmead and Salzberg, 2012). In these alignments, reads mapping to more than one location were removed using a custom Linux command, duplicated reads were removed using the MarkDuplicates command from Picard (*http://picard.sourceforge.net*), and indels were realigned using the IndelRealigner module of GATK (DePristo *et al.*, 2011).

Sequence variants between each accession and Col-0 were calculated simultaneously for all accessions using the UnifiedGenotyper module in GATK with parameters *stand_call_conf* and *stand_emit_conf* set to 30 and 10. Variants located in regions flanking large indels or rearrangements were filtered using a custom Perl script. In this script, reads are divided in 20bp segments that are numbered from 1 to 5 from left to right. Variants covered by the same segment in all reads, or supported by segments 1 or 5 in a single read are discarded. Besides, only bi-allelic variants with a Phred-scaled variant quality higher than 30 were retained. Only homozygous polymorphisms located in exons were used to calculate allele specific expression.

3.4.5. Transcript abundance quantification

3.4.5.1. Mapping and selection of reads The Reads for each sample were mapped to TAIR10 release of *Arabidopsis thaliana* Col-0 reference genome using TopHat (Trapnell *et al.*, 2009, 2012) release 2.0.6. Low quality bases at terminal ends with less than 30 phred score were trimmed. The reads mapping onto to unique locations on Col-0 were considered for further analysis.

3.4.5.2. Library size normalisation Large genes have higher probability of more reads getting sequenced. Therefore, mRNA abundance was normalised as reads per kilobase per million (rpkm), which is a countermeasure against biased sequencing depth due to gene length.

3.4.5.3. Selection of exonic reads The bam files were assessed for coordinates of mapped reads and compared them to the exon coordinates (location) on the genome. In the Gene Feature format (GFF) file of Col-0 TAIR-10 version, start and end coordinate of the exons, CDS and UTRs corresponding to each gene are listed. Customised Perl scripts were used to compare the mapping location of reads to the exon coordinates in GFF file. This excluded the reads overlapping intergenic regions and those that mapped only onto introns due to intron retention events. However, the reads mapping over exon-intron junctions were selected as it includes the expressed exonic part of alternatively spliced

transcript. Constitutive reads that map over single exons, and spliced reads that mapped over to more than one exon were considered as exonic reads summarised as:

$$Constitutivereads: E\_s < R\_s < R\_e < Ee \quad (1)$$

$$Exon-intronjunctionreads: E\_s < R\_s < E\_e < R\_e \quad (2)$$

$$Spliced-read: E1\_s < R\_s < E1\_e < E2\_s < R\_e < E2\_s \quad (3)$$

where,

- R_s, read_start_pos
- R_e, read_end_pos
- E_s, exon_start_pos
- E_e, exon_end_pos

3.4.5.4. Counting exonic reads Using customised Perl scripts exonic reads were assigned to corresponding exons per gene.

Counting unique mapped reads is similar to a set-problem, where each gene can be considered as set of mutually exclusive independent exons, and each exon is a subset of corresponding reads.
Thus, the expression can be estimated as collectively exhaustive reads from all exon sets per gene. The quantification method is summarised in the equation 4:

$$E_{rc} = \sum_{k=0}^{Ne-2} (-1)^k \left(\sum_{i=1}^{Ne} nr\left(E_i \cap ....E_{((i+k)\%Ne)}\right) \right) + nr\left(E_1 \cap E_2 \cap ....E_{Ne}\right) \quad (4)$$

where,

- E_{rc} is exonic read count
- N_e is the number of exons in the gene,
- nr is the number of reads over set, and
- E_i represents i^{th} exon

3.4.5.5. Selection of allelic SNPs The SNPs heterozygous between the each pair of parentals corresponding to the hybrids were selected for ASE analysis. These SNPs were considered diagnostic for a pair of parentals P and L*er* as P≠L*er* i.e, SNP loci with distinct nucleotide base at P and L*er* irrespective of the corresponding base in Col-0.

3.4.5.6. Counting SNP reads To quantify allelic transcript variation, for every gene reads mapped over SNPs in the exons were counted. Counting number of reads at every SNP gives information about the coverage over SNPs. However, this method is likely to introduce bias in quantification due to redundant information for the reads aligning and overlapping multiple SNPs. Hence, each read was examined for the number of SNPs covered in its entire length to ensure that reads over multiple SNPs were counted exactly once. The read count measure have been summarized above and the equation for SNP reads count is as follows:

$$S_{rc} = \sum_{k=0}^{Ns-2} (-1)^k \left(\sum_{i=1}^{Ns} nr\left(S_i \cap ....S_{((i+k)\%Ns)}\right) \right) + nr\left(S_1 \cap S_2 \cap ....S_{Ns}\right) \tag{5}$$

where,

- S_{rc} is snp read count,
- N_s is the number of snps in the gene,
- nr is the number of reads over set, and
- S_i represents i^{th} snp

3.4.6. Selection criteria of genes for differential gene expression analysis

The following parameters was used to select genes for differential expression analysis between each pair of parentals:

- genes with ≥ 1 SNP
- genes with rpkm $\geq$1 in triplicates of at least one parental accession
- SNPs mapped by ≥ 10 reads

This enabled selection of genes with reliable expression levels for comparison between the pair of parentals. For the selected genes reads mapped over the SNPs were counted.

And a reliable coverage of 10 reads per SNP was selected as selection threshold (Figure 22 on page 54)

3.4.7. Differential expression analysis

Differential expression (DE) analysis was performed using edgeR exact test (Robinson *et al.*, 2010) independently from exonic and SNP reads for each pair of parentals. Benjamini and Hochbergś algorithm was used to control the false discovery rate (FDR). Genes with FDR < 0.05 were selected as significantly differentially expressed.

3.4.8. Comparison of expression variation over exons and SNPs

The genes with high degree of similarity in magnitude and direction of expression difference between parentals measured at both SNPs and exon levels were selected to be further examined for allelic variation. This included genes with varying but a conserved pattern of expression change reflected proportionally at the exons and SNPs.

3.4.9. Elucidation of expression divergence pattern

Expression divergence was measured as fold change on log scale (logFC) between parentals and their corresponding alleles in hybrids. The parental ratios were compared against the allelic ratios in hybrid to determine allele specific expression using edgeR in R. The significance were tested using the multiple correction test of Benjamini Hochberg. Change in expression was considered significant at 5% FDR.

3.5. Software packages used

Perl package Bio::DB::Sam was used for interacting with Sam and Bam files. Cutomised perl scripts were used for Read count, SNP annotation, and pseudo-reference construction. Differential expression analysis was carried out using the exact test of the edgeR (Robinson *et al.*, 2010) Bioconductor package ("classic edgeR") in R.

3.5.0.1. Selection criteria of genes for ASE analysis The differential expression for parentals using read count over exons was compared with estimation over SNPs. The expression was only considered significantly different (sig) at FDR below 5% (sig). The genes were further selected to test and classify into specific regulatory mechanisms. The criteria used is summarised in the following table:
The following parameters were then used to select genes for ASE analysis:

(i) exon logFC $\simeq$ SNP logFC
(ii) parental logFC ≥ 1.5

exon	snp	sign (exon_logFC * SNP_logFC)	selection	class
sig	sig	+	1	cis trans cis x trans
sig	non-sig	NA	0	NA
non-sig	sig	NA	0	NA
non-sig	non-sig	+	1	compensatory conserved

Table 1. Comparison of expression divergence corresponding exons and SNPs to select genes for ASE analysis. Expression divergence ratio of parentals corresponding to exons and SNPs were compared to assess significance (sig; non-sig). The direction of log fold change in expression is shown by mathematical signs of plus "+" (same) and minus "-" (opposite). Genes with differential magnitude and direction of expression change in exons and SNPs were not selected (0). Genes with similar magnitude and same direction expression change ("+") were selected (1) for ASE analysis. The possible regulatory divergence categories for genes selected are shown in the column "class".

Finally, the allelic expression profile in hybrids were compared with parental expression changes and the genes were classified as per the plausible regulatory mechanisms summarised in the Table 2.

DE Parents	DE Hybrid	sign (P_logFC * H_logFC)	class
sig	sig	+	cis
sig	non-sig	NA	trans
non-sig	sig	NA	compensatory
sig	sig	-	cis x trans
non-sig	non-sig	+	conserved

Table 2. Classification of cis- and trans- regulatory divergence patterns. Expression divergence ratio of parentals and hybrid are compared to assess for significance (sig; non-sig). Direction of log fold change in expression is depicted by "+" (same) "-" (opposite). The category of regulatory divergence patterns is decided based on the two parameters as shown in the "class" column.

3.5.1. Verification of allelic ratio

Allelic ratios from randomly selected cases for cis-, trans-, compensatory- and cisxtrans-categories were further subjected to comparison with pyrosequencing for verification of the results.

3.5.2. Functional enrichment

The genes detected in to be regulated in cis- and trans- category were further analysed for functional categories using MapMan version 3.6.0 (Thimm *et al.*, 2004).

3.6. Determination of parent-of-origin effect

The cis- regulatory divergence patterns were compared between each pair of reciprocal hybrids to determine possible parentally biased allelic expression or parent-of-origin effect on expression variation.

3.7. Inheritance of expression patterns

The gene expression of parentals and hybrids were compared to determine expression inheritance patterns. The genes in cis-, and trans- categories were further checked for their impact on inheritance patterns.

3.8. Metabolite analysis

3.8.1. Primary metabolite profiling by GC/TOF-MS

To determine the impact of natural variation in gene expression on on metabolic phenotype, the plant samples and the reciprocal hybrids were subjected to metabolite extraction for primary metabolites at Max Planck Institute for Molecular Plant Physiology, Golm. Metabolite analysis by GC/MS was performed from aliquots of 50 mg tissue fresh weight in 2 ml Eppendorf tubes. Extraction was performed as described by Lisec *et al.* (2006). Chromatograms and mass spectra were evaluated by using TagFinder 4.0 software (Luedemann *et al.*, 2008) and using Xcalibur 2.1 software (Thermo Fisher Scientific, Waltham, USA). Metabolites were identified in comparison to database entries of authentic standards (Kopka *et al.*, 2005; Schauer *et al.*, 2005). Peak areas of the ed to the internal standard (ribitol) and fresh weight of the samples. Identification and annotation of detected peaks were shown in Supplemental Table SX with information of "Recommendations for Reporting Metabolite Data" described in Fernie *et al.* (2011).

3.8.2. Metabolite data analysis

The metabolites were classified according to the biological role and pathways involved and analysed for variation in relative concentration across the accessions and hybrids. To determine the intra-specific variation of the metabolites from the samples under consideration, the data were normalised for each pair of parental samples and respective hybrids against the mean of metabolite value for the group. Genes involved in pathways of each metabolite were examined for expression profiles.

To determine the correlation of gene expression and metabolite concentration, metabolites with maximum abundance and variability were selected. The expression values of the pathway-specific genes for the most variable metabolites were tested for correlation with metabolite profiles. The case found with maximum correlation signaling a high degree of

the association has been reported here. Pathway information was collected from metabolic pathway databases: AraCyc (*http://www.arabidopsis.org/tools/aracyc*; (Mueller *et al.*, 2003)), PMN (*http://www.plantcyc.org/*) and KEGG(*https://www.genome.jp/kegg/*; (Kanehisa and Goto, 2000)).

4. RESULTS AND DISCUSSION

4.1. Allele-Specific expression analysis

Plant response to varying biotic and abiotic factors have been of considerable importance from an agricultural and ecological perspective. Change of environment generally triggers the change in the gene expression, consequently affecting the function of a gene or the plant phenotype. Therefore, to estimate and predict the phenotypic changes, it is crucial to dissect the changes in expression, determine and understand the causative regulatory mechanisms.

Since gene transcription is a multi-factor regulated process, with several genes regulated by multiple transcription factors from regulatory genes, miRNAs, epigenetic marks, the uni-cellular environment of the alleles. Therefore, within a specific system, trans-variants have small-effects on expression as compared to cis-variants (Springer and Stupar, 2007b). However, cis-variants are more prominent and the effects tend to be conserved, while the trans-factors have a variable impact on expression divergence.

Therefore, it is of considerable importance to attain a robust and reliable method for differential allelic expression estimation. Also, it is crucial to thoroughly examine the relative effect of major caveats (described in section 1.9. on the estimation, analysis results, and interpretation. Therefore the primary requirement of ASE study is a thorough and methodological development of a structured pipeline with due consideration to the following aspects :

(i) **Polymorphism profile**, to determine
   (a) relative difference to genome of reference
   (b) distribution of SNPs
   (c) diagnostibility of SNPs
   (d) coverage and accessibility of SNPs

(ii) **Gene profile**, to account for
   (a) genes with reliable read coverage for polymorphism detection and expression analysis
   (b) genes with diagnostic SNPs
   (c) genes with accountable allelic coverage of informative SNPs
   (d) genes with quantifiable allelic expression variation

(iii) **Genome of reference**, to assess
   (a) the impact on mappability of reads

(b) the impact on estimation of expression variation

(iv) **Differential Coverage over Exons and SNPs**, to determine

(a) the exonic and SNP measure of expression variation

(b) the impact of exonic and SNP on estimation of expression variation

4.2. Method development

Allele specific analysis method using RNA-Seq data was developed using RNA-Seq reads from leaf tissue of *Arabidopsis thaliana* accessions, Cvi and L*er*, and hybrid Cvi x L*er*. The RNA was extracted from three biological replicates for each parental accession and the hybrids. Col-0 genome was selected and used as genome of reference for the initial mapping and expression estimation. Additionally, pseduo-reference genomes for Cvi and L*er* have been developed to compare and examine the relative impact of the reference genomes. With an objective to develop robust estimation protocol for Allele-Specific expression variation, the above stated factors have been examined and assimilated during the method development protocol along the following schema: Further the results from

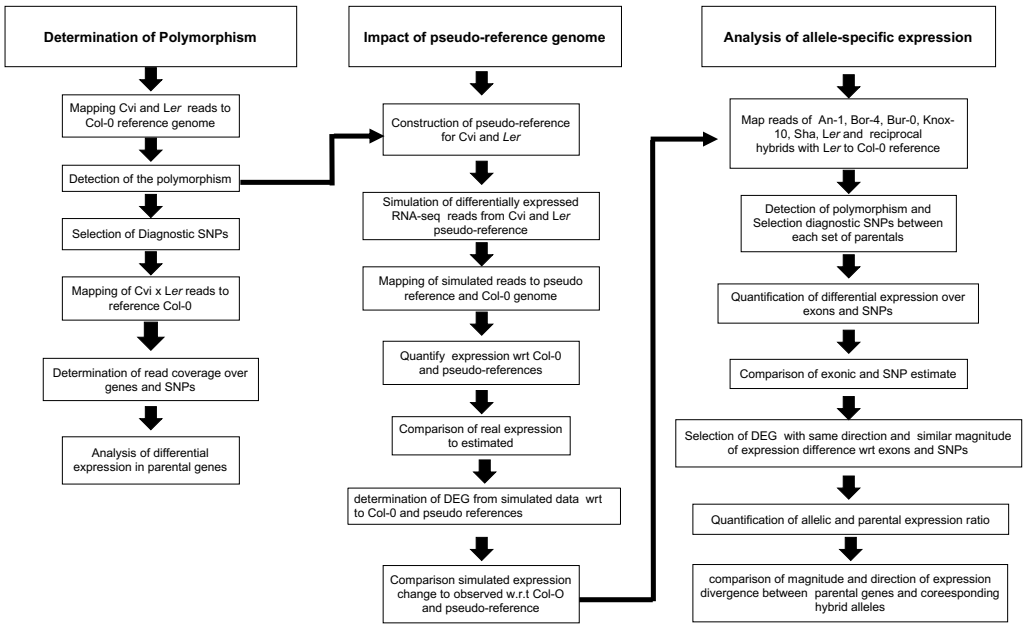


Figure 10. Sumarised workflow of method development schema for ASE analysis.

ASE were categorised into regulatory divergent patterns based the following criteria:

(a) **Cis-regulated** genes, with differentially expressed hybrid alleles and parental transcripts.

(b) **Trans-regulated** genes, with differentially expressed parental transcripts and similarly expressed hybrid alleles.

(c) **Compensatory** genes, with similarly expressed parental transcripts and differentially expressed hybrid alleles.

(d) **Conserved** genes, with parental transcripts and hybrid alleles similarly expressed.
(e) **Cis x Trans** genes, with differentially expressed parental transcripts and alternately expressed hybrid alleles.

4.2.1. Mapping of reads to reference genome

Gene expression is the measure of transcript abundance, which can be estimated using RNA-Seq data. The prerequisite for the expression analysis is the measure of the reads sequences for each transcript, which is reflected by the number of reads mapping to the genes. The RNA-Seq reads, therefore, are mapped against a genome of reference. The reference genome is selected based on availability and similarity to the species under consideration. Since the Col-0 genome is from the *Arabidopsis thaliana*, hence this study is more of an intra-specific and inter-accession expression variation. RNA-Seq reads from Cvi and L*er* were mapped to the Col-0 reference genome. Reads that mapped uniquely with not more than 12 mismatches were used for determination of polymorphic sites against Col-0.

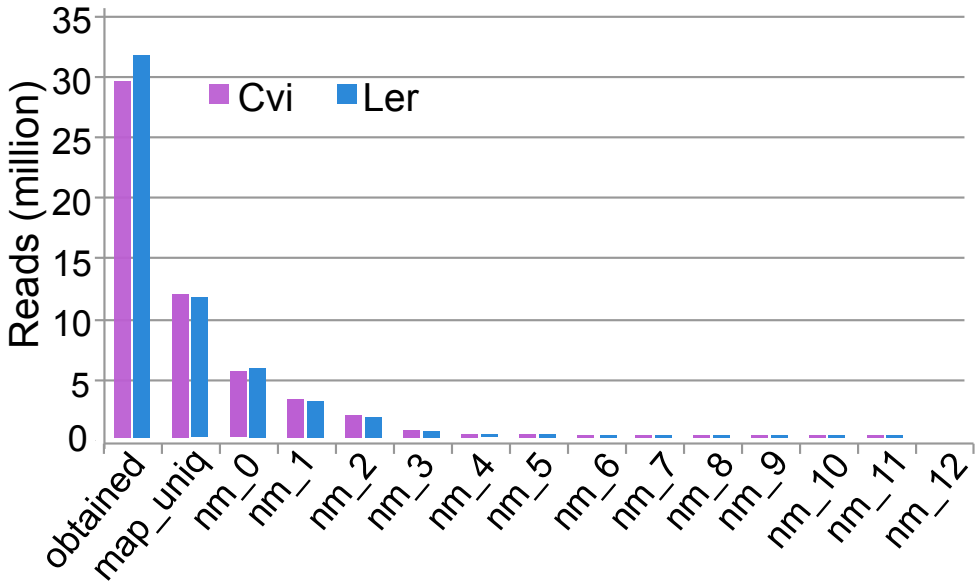


Figure 11. Mapping profile of Cvi and L*er* reads. The bar plot shows the number of RNA-Seq reads (Y-axis) sequenced from Cvi (magenta) and L*er* (blue) accessions and relative number of reads mapped uniquely to Col-0 reference genome, with mismatches (nm) - 0,1,2 etc; mapped using Tophat.v.2.0

Nearly 40% of Cvi reads and 36% of L*er* reads mapped uniquely to a single region on the reference genome. Approximately 88% of these uniquely mapping reads mapped with only 0 to 2 mismatches. Reads mapping uniquely with up to 12 mismatches were used for the determination of polymorphic sites against Col-0 reference genome.

4.2.2. Detection of polymorphic sites

The uniquely mapped reads from Cvi and L*er* were further processed using GATK Variant Caller to determine polymorphisms in uniquely mapped Cvi and L*er* reads against Col-0 reference genome. The differences in the polymorphic rates of the species under consideration against the genome of reference, can result in biased mapping of reads from the species similar to the reference. Therefore, the abundance of variants against the reference must be compared to account for any bias in mapping, and hence expression estimate. The Figure 12 depicts the polymorphic sites in Cvi and L*er* against Col-0. 99.89% of the detected SNPs were found to be di-allelic with homozygous base for at

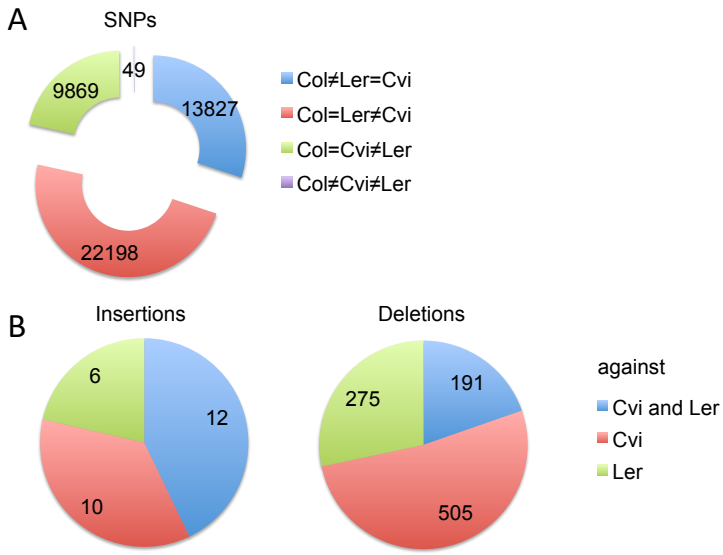


Figure 12. Comparison of polymorphism profile in Cvi and L*er* against Col-0. The figure shows intra-specific polymorphism rates between the three accessions. (A) SNPs. The pie chart depicts the number of loci on Col-0, at which either or both Cvi or L*er* show a SNP. The legend summarises the allelic composition of SNP loci as following equations:

(i) Col≠ L*er*=Cvi, shows loci with heterozygous SNP between L*er* and Col-0, with same base for L*er* and Cvi.
(ii) Col=L*er*≠Cvi, are the SNPs between Cvi and L*er* where L*er* and Col-0 have the same base.
(iii) Col=Cvi≠L*er*, are the SNPs between Cvi and L*er* for which Cvi and Col-0 have the same base.
(iv) Col≠L*er*≠Cvi, are the SNPs between Cvi and L*er* where both have a different base than Col-0.

(B) Insertions and deletions. The pie charts show the number of Col-0 genomic sites at which insertion or deletion was detected from the mapped reads of Cvi and L*er*:

(i) Cvi and L*er*, refer to number of Col-0 loci at which both Cvi and L*er* show insertion or deletion.
(ii) Cvi, refer to unique insertion and deletion sites in Cvi.
(iii) L*er*, refer to unique insertion and deletion sites in L*er*.

least 2 accessions (Col≠L*er*=Cvi, Col=L*er*≠Cvi, Col=Cvi≠L*er*). The di-allelic loci indicate single base change during evolution in either or both the parents with respect to Col-0. SNPs at Tri-allelic loci (Col≠L*er*≠Cvi) made up only 0.11% (49) suggesting that multiple base changes are rare events.

The SNPs heterozygous can distinguish hybrid alleles and are referred to as diagnostic SNPs hereafter. The diagnostic di-allelic and tri-allelic SNPs heterozygous at least for Cvi and L*er* (L*er*≠Cvi), accounted for more than 69.9% of total identified SNPs. While, the remaining 30.10% made up the non-diagnostic SNPs with homozygous Cvi and L*er* allele.

Number of variants against Col-0 were found to be less in L*er* as compared to Cvi indicating a relatively high degree of genetic similarity between L*er* to Col-0. This may cause to biased mapping in favour of L*er*.

4.2.3. Distribution profile of SNPs

Allele-specific expression analysis in hybrids needs prominently covered allelic SNPs in gene regions. Hence, it is crucial to assess and annotate the pattern of SNPs. Therefore, the SNPs determined using unique reads from the parentals, Cvi and L*er*, mapping to Col-0 were filtered for diagnostic SNPs, which could distinguish Cvi from L*er* allele i.e, Cvi(X):L*er*(Y), where X and Y denotes a nucleotide base ATGC. The selected SNPs were assigned to respective genomic locations using the annotated gene feature format file (GFF) for Col-0 (TAIR10 release). Thereafter, reads from hybrids pooled from all three biological replicates of the Cvi x L*er* and L*er* x Cvi, were mapped to Col-0 reference, to get an overall estimation. Eventually, the SNPs were examined for relative coverage by reads within the hybrids at coding sequence regions (CDS), five-prime untranslated regions (5′-UTRs) and three-prime (3′) UTRs collectively. This allowed the identification of SNPs sufficiently covered by allelic reads essential for estimation of allelic transcript abundance.

A total of 45943 loci were identified as SNP against Col-0 genome. When the location of these SNPs in Col-0 reference was annotated, 45087 (98.1%) of SNPs could be assigned to genes and the remaining 856 (1.8%) to intergenic regions. 44080 (∼95.9%) SNPs mapped to exons with nearly 76.7% (35273), 4.6% (4115) and 16.8% (7725) SNPs were detected within CDS, 5′ and 3′ UTRs respectively. 1007 SNPs (∼2.2%) mapped to intronic regions. Since the reads were generated from mRNAs (transcripts), the intronic SNPs indicate intron-retention events. Notably, only 20576 (∼44.8%) SNPs within coding portions were covered by at least 1 hybrid read. Hence, more than 50% of SNPs could not be used used for allelic differential expression analysis (Figure 13 on page 36.

4.2.4. Accessibility of gene for allelic expression analysis

Genome-wide expression analysis aims to determine the divergence patterns for most genes. In additional, divergence patterns in parentals provide a road map to determine

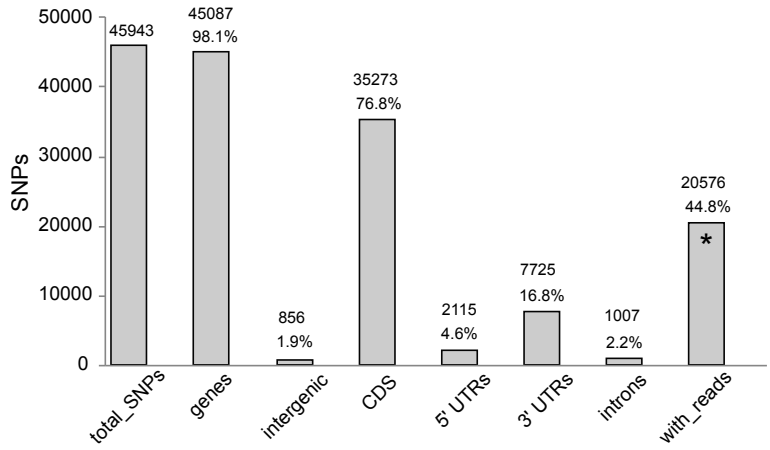


Figure 13. Abundance and distribution profile of SNPs across the genome of Col-0 against Cvi and *Ler*. The figure reveals the number of SNPs located within gene and intergenic regions, CDS, 5′ and 3′ UTRs, and introns. It also shows the SNPs covered by allelic reads from hybrids.

allelic expression patterns in hybrids and identify the underlying regulatory mechanisms. Transcription is a dynamic process and subject to change in biological and experimental conditions and technical constraints. Read information for every gene and allele might not be available due to technical limitations (poorly sequenced genes) and/or experimental conditions (unexpressed genes) or lack of SNPs. Hence, it is inevitable to ascertain the accessibility of the genes with respect to SNP abundance, location and read coverage.

For the present study, the comparative abundance of genes in genomic regions and their usability for expression analysis have been evaluated and represented in the Figure 14 on page 37. Allele-specific expression studies rely strongly on SNPs for allelic distinction. Therefore genes without SNPs cannot be considered for ASE analysis. Also, genes with SNPs should be sufficiently represented by the hybrid allelic reads. Genes without sufficient allelic reads can not be selected for determining allelic expression patterns. It is equally important that the allelic SNPs have adequate coverage to improve the specificity of the method to quantify differential expression. Hence, only genes for which both alleles have a certain level of reads can be considered for further analysis.

The magnitude of expression divergence can vary from subtle to large scales. While more substantial differences are easier to detect, the detection of more trivial changes are heavily dependent on the statistical power of the method to identify biologically relevant variations. This can be achieved with low sensitivity but noise may be introduced. Hence, a minimum threshold for variation level (log fold change $\geq$ 1.5) was chosen as per conventional reports, to filter out technical noise.The successive filtering criteria and corresponding gene profile is depicted in the Figure 14 on the page 37.Gene annotation file in

the GFF format (general feature format) for *Arabidopsis thaliana* available from TAIR10 release contains information about 28775 genes including coordinates of exons, CDS, 5′ UTRs, 3′ UTRs, isoforms etc. 11120 genes (38.6%) genes were found to contain SNPs between Cvi and L*er*, while no SNPs were detected for 61.3% genes. SNP containing genes were selected for downstream expression analysis. Nearly 37.9% (10906) genes have SNPs located within in CDS, and five-prime (5′) and three-prime (3′) UTRs.

Further, the genes were examined for read coverage over the SNPs within hybrid Cvi x L*er*. The hybrid reads were allocated to parental alleles based on the parent-specific SNPs. To count the reads for each allele reads overlapping the SNPs were used (Exonrc). The read count over SNPs (SNPrc) per allele was determined by counting each read overlapping multiple SNPs only once. From the genes with SNPs in the coding portion, only 13.7% (3963) genes with 45% SNP positions were mapped by at least one hybrid read (Figure 13).

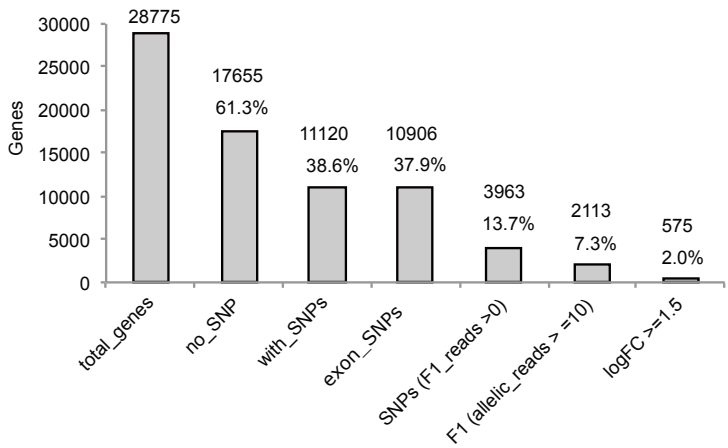


Figure 14. Availability of genes for allelic expression analysis. Bar graph plots the number of Col-0 genes (Y-axis) with diagnostic SNPs between Cvi and L*er*, located within the exons (CDS and UTRs), genes with reads mapping over SNPs in F1 hybrid Cvi x L*er*, genes with minimum 10 reads per allele in hybrid and those with minimum log fold change of 1.5 between the parentals.

Alleles with low or no coverage may reduce the accuracy of differential expression analysis. Therefore, a minimum coverage of 10 reads per allele was used as a threshold to select genes for ASE analysis. Only 2113 genes (∼7.3%) passed the set threshold with at least 10 reads for each allele per gene.

These genes were further assessed for allelic imbalance and differential expression analysis was performed using edgeR (Robinson *et al.*, 2010). As per most widely used in differential expression studies, a standard threshold of a miminum of 1.5 log fold change was used to finally select 575 (only ∼2.0%) genes with allelic divergence within the Cvi x L*er*, with significance level at FDR 0.05. Of these 328 genes showed L*er*>Cvi as stronger expressed alleles (L*er*>Cvi) and 247 with Cvi as higher expressed allele (L*er*<Cvi.

4.2.5. Effect of parental-specific pseudo-reference genome

4.2.5.1. Improvement in read mapping Read mapping bias may occur while mapping to common reference owing to alleles with high degree of similarity to the genome (Degner *et al.*, 2009; Satya *et al.*, 2012). The use of a pseudo-reference genome might reduce the mapping bias (Satya *et al.*, 2012). Therefore, to assess the influence on expression and read mapping, pseudo-reference genomes were created separately for Cvi and L*er* by incorporating the accession-specific polymorphisms including SNPs, insertions, and deletions in the Col-0 genome. The pseudo-reference with the polymorphic sites incorporated should allow more reads to map to it simply as a matter of enhanced sequence similarity (Satya *et al.*, 2012). Reads that remain unmapped to the distant genome due to a large number of mismatches, SNPs, or very large Indels, may map correctly to the pseudo-reference owing to the high degree of similarity and reduction in mismatches.

A comparison of mapping against Col-0 with pseudo-reference genome would enable evaluation of the degree of improvement for the present study. Hence, to assess the mappability, simulated reads from Cvi and L*er* were mapped against Col-0 and respective pseudo-references. When mapped against pseudo-reference, almost ∼99% genes had more reads mapping to them (Figure 15A). The genome-wide comparison confirmed the relative increase in mapped reads across pseudo-reference. Although, the increment observed was a minimal fraction of the reads mapping to Col-0, yet it underscores the importance of parental-specific reference genome for expression analysis. The technical challenges including the difference in the extent and quality of genome sequencing methods, assembly, and annotation for the species under examination, may restrain the use of this proposed method. However, the improvement is expected to be considerably higher during the comparison across divergent species with high degree of variation. The computational pipelines AlleleSeq (Rozowsky *et al.*, 2011), MMSEQ (Rozowsky *et al.*, 2011)

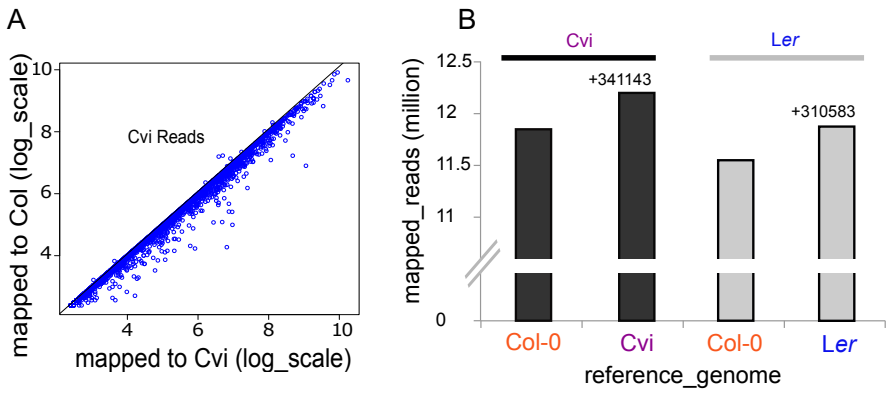


Figure 15. Comparison of Col-0 and pseudo-references for read mapping. (A) Reads mapped per gene. The scatter plot shows the expression (log normalised) for each gene (blue dot) mapped to Col-0 (Y-Axis) against Cvi (X-axis). Dots lying on the diagonal line indicates genes with little difference in expression estimates in both Col-0 and Cvi. Degree of deviation below the diagonal line indicate indicate genes with more reads mapped onto Cvi. Dots above the diagonal line indicate genes with more reads mapped onto Col-0. B. Reads mapped per genome. The bar plot compares the number of reads mapped (Y-axis) from Cvi (black) and L*er* (gray) to Col-0 and accession-specific pseudo-reference (X-axis). The reads and references are indicated accession-specific colour code for Col-0 (orange), Cvi (magenta) and L*er* (blue).

and Allim (Pandey *et al.*, 2013) demonstrate the beneficial use of pseudo-references for determination of allele-specific expression.

4.2.5.2. Enhanced accuracy of expression estimation Improvement in the read mappability should increase the accuracy of gene expression measure. The reads with higher similarity to parental reference should map more accurately thus reducing the erroneous estimation. Additionally, the reads mapped to the pseudo-reference should provide high similarity with the real gene expression. The simulated expression profile provided a control estimate of the reads per gene. The simulated reads were mapped to Col-0 and Pseudo-references separately and the expression level for each gene was calculated from the uniquely mapped reads. To assess the influence of pseudo-reference on the power of estimation, a comparison of control versus estimated expression was performed. The deviation from the expected (expected/observed) ratio of 1 (log FC =0) reflects change in measured expression which is caused by under or overrepresentation of genes by the reads. Thus, a log fold change of 0 represents that the gene expression is estimated with absolute accuracy as it is similar to the simulated value. A positive deviation for a gene indicate very low number of mapped reads, and hence a lower than known estimate of transcript measure. Similarly, a negative deviation for any gene implies higher than expected reads mapped, which could be the consequence of reads from multiple gene copies. Depending on the number of unmapped reads the magnitude of expression may deviate negligibly or strongly. Most of the deviation is expected to be positive, with low proportion of negative deviation expected for overrepresented genes. With high mapping quality the deviation for most genes should be non-significant and equivalent to the expected in-

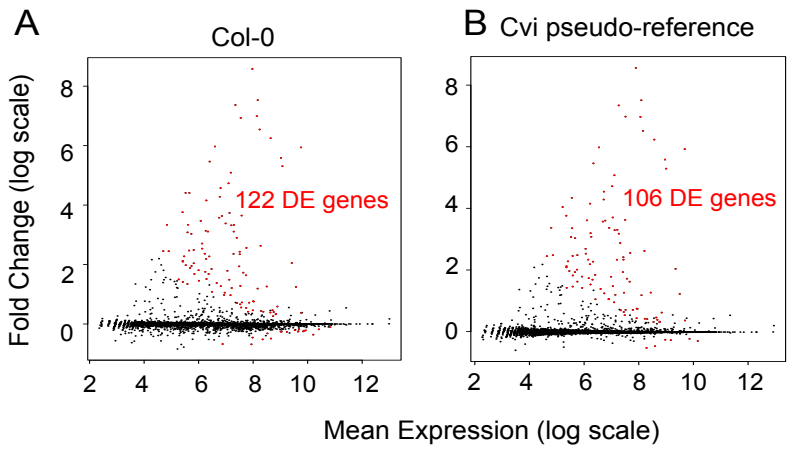


Figure 16. Degree of deviation from expected expression profile. The scatter plot shows comparison of (A) Col-0 and (B) Cvi pseudo-reference to evaluate deviation from expected [log(expected)-log(observed)] gene expression levels measured from simulated RNA-Seq reads (control). Each dot represents a gene, for which mean expression levels is plotted along X-axis, and fold-change (log scaled) between expected and observed expression is plotted along Y-axis. Expression quantified from mapped simulated reads is expected to be equivalent to the control (fold change $\sim$0) for all expressed genes. However, any significant deviation (FDR $<$ 0.05) from the expected expression represents over (fold change $<$ 0) or underrepresented (fold change $>$ 0) genes (highlighted in red colour). The number of genes with differential expression have been stated within each plot.

tensity. Hence, the genes with significant differences in the observed and expected ratio, provide benchmark to characterise sensitivity of expression computing method. Using this strategy, we compared the Col-0 and pseudo-reference genomes for assessing the measurement accuracy.

4.2.5.3. Influence of reference genome on differential expression analysis Use of parent-specific reference improves expression quantification per gene (Figure 16). Higher accuracy in expression measure are likely to improve detection of differential expression. Therefore, to test this hypothesis the reference genomes were evaluated for any added advantage in estimating expression change. Expression profile for simulated Cvi reads was tweaked to generate L*er* reads with set of differential expression profiles for random set of genes. The expression profiles provided for simulations were compared to generate set of genes with significant differential expression (control) using edgeR at FDR$<$0.05.

Next, the analysis was independently carried out with Col-0, accession-specific pseudo-reference to generate corresponding sets of statistically significant DEGs. Comparison of the three datasets was used to evaluate the sensitivity of Col-0 and pseudo-reference to determine expression divergence. The list of genes identified as significantly differ-

entially expressed using Col-0 and pseudo-reference were compared against the control list. Based on the observations, the genes were further characterized as true positives (TP), false positives (FP), and false negatives (FN) as per following criteria: Correctly identified

dE	Sim	Col-0	Class$_{ref}$	Pref	Class$_{pref}$
sig	1	1	TP$_{ref}$	1	TP$_{pref}$
sig	1	0	FN$_{ref}$	0	FN$_{pref}$
sig	0	1	FP$_{ref}$	1	FP$_{pref}$

Table 3. Comparison of Col-0 and Psuedo-reference genome for determination of expression divergence. Table compares the sensitivity and specificity of Col-0 and Pref in determination of significant (sig) differential expression (dE) against the known differentially expressed genes in simulated data (sim). Genes were further classified into true-positives (TP), false-negatives (FN) and false-positives (FP) depending on accuracy of detection by ref and pref.

differentially expressed genes were categorised as true-positives for the specific genome, for eg. TP$_{ref}$ for Col-0 and TP$_{pref}$ for pseudo-reference. On the contrary, the truly differentially expressed genes which could not by identified from Col-0 or pseudo-reference were classified as false-negatives (FN$_{ref}$ and FN$_{pref}$) for the respective references.

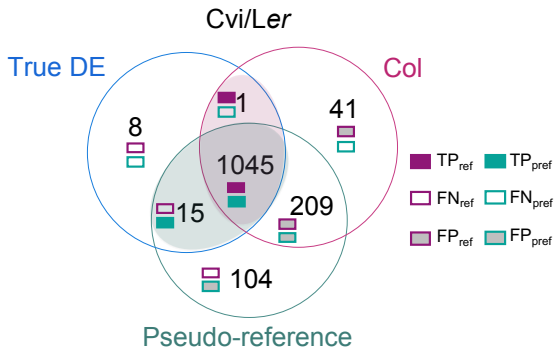


Figure 17. Evaluation of sensitivity of genome of reference to detect expression divergence. The venn diagram depicts the number of differentially expressed genes (True DEG) identified using Col-0 (magenta) and Parent-specific pseudo-reference (Pref; green) genomes. The venn sets are colour coded to represent Col-0 (magenta), Pref (green) and simulated data (blue). The components of venn-diagram are further highlighted using boxes to indicate the relative number of TP (filled box), FP (empty box), FN (grey box). The box lines are colour coded as per the reference ref (magenta) and Pref (green).

Both methods detect a large proportion of true positive DEGs with pseudo-references allowing for detection of more true-positives at the cost of relatively more false positives. Interestingly, more than the expected number of differentially expressed genes were determined with both reference (False positives). Although, such genes have nearly similar expression levels across the accessions in simulated data, nevertheless, show phenomenal variation when mapped against reference genomes. The variations in expression might be similarly represented, irrespective of the genome, as evident from the common false positives, The relative number of true positives, false negatives and false positives were used to calculate the efficiency of each genome. The quantification of accuracy measures provide an overview of the competency of the method as well as the reliability of the results.

As a code of conduct, the performance of statistical tests, computational methods, or bioinformatics pipelines are measured as sensitivity, specificity, positive and negative predictive values and accuracy. The following equations depicts the factors used for calculation:

$$Sensitivity = TP/(TP+FN) \tag{6}$$

$$Specificity = TN/(TN+FP) \tag{7}$$

$$Precision = TP/(TP+FP) \tag{8}$$

$$Accuracy = (TP+TN)/(TP+FN+TN+FP) \tag{9}$$

While, sensitivity is the proportion of correctly identified positives (true positive rate), specificity refers to the correctly identified negatives (true negative rate). Additionally, Positive Predictive Value (PPV) or Negative Predictive Value (NPV) are used to describe and compare the accuracy of the statistical test for the experimental condition. The PPV is the percentage of correctly identified positives from the positive results. Sensitivity and specificity are generally inversely proportional and if used alone might be only a partial measure of tool efficiency. Rather, calculation of Precision (Positive predictive value) and accuracy from combination of sensitivity and specificity allow a more comprehensive and unbiased evaluation of the method. Hence, we employ the later strategy to select for relatively high precision and more accurate method for further estimation and analysis of expression profiles.

For our biological system under consideration, use of genome-specific pseudo-references show a slight improvement in read mappability. Additionally, the expression measure improve with the use of pseudo-reference. However, when compared against Col-0 as

a genome of reference, the determination of differential expression is rather worse, as evident from a higher false positive rate. Thus, the concept and usage of pseudo-reference may not necessarily provide an added significant advantage over the Col-0 genome for computing differential expression.

However, this observation could be particular for our intra-specific comparison. Large degree of inter-accession variation and the polymorphism rates to be relatively are lower as compared to the variation at inter-specific level. Alternatively, the detection of higher than expected number of differentially expressed genes using pseudo-references could imply higher sensitivity of the pseudo-reference. The use of accession-specific genomes enables detection of substantially insignificantly low levels of expression differences, especially the genes which have expression variation so low that the differences are not evaluated as significant in the simulated data-set. These difference when assessed using pseudo-reference as significant can be termed as false-positives. This is further strongly exemplified by the relatively higher number of true-positives detected using pseudo-reference are higher than by Col-0 genome. Therefore, we propose evaluating more stringent parameters while using pseudo-references to increase power of detection.

Finally, based on the observed improvement in read-mapping with higher accuracy of expression estimation and more true-positives detection by pseudo-reference, we propose that the use of pseudo-reference should strongly benefit allelic-imbalance studies in inter-species with larger scale of variation. However, the ratio of true-positives to false-positives is higher using Col-0. Therefore, we use Col-0 as a genome of reference for estimation of differential expression and allele-specific expression analysis with experimental data. A comparative overview of ASE results of the hybrids of natural accessions An-1 x L*er*, Bor-4 x L*er*, Bur-0 x L*er*, Knox10 x L*er* and Sha x L*er* have been discussed In the section 4.3. on page 53.

Nevertheless, the advantage of pseudo-reference genomes cannot be neglected entirely for genome-wide scale studies. Therefore, we provide a framework for generating accession-specific references (Figure 8). In addition, we used the framework to construct accession-specific-references for the parentals in an iterative manner. A comparative assessment of polymorphic intensity, and subsequent improvement in read mapping post each iteration have been discussed in the section 4.2.6.

4.2.6. Pseudo-reference genome construction and assessment

Results from analysis of simulated supported the use of parent-specific pseudo-references genome for expression analysis. However, in addition to gene expression divergence, the

primary requirement for ASE analysis is allelic ratio quantification within hybrids. Hence, to decide upon a reference genome, parental-reference genomes were constructed for all six *Arabidopsis thaliana* accessions. To obtain the pseudo-reference genomes reads of each accession were independently mapped Col-0 reference to determine polymorphic sites, which were then subsequently incorporated onto Col-0 genome to generate corresponding accession-specific-reference (iteration 1). The process was iteratively repeated for 25 iterations, using the modified reference post each iteration until no more polymorphic sites could be determined.

A comparison was made to assess the saturation of polymorphisms, ad improvement in mapped reads. It was observed that after fifth iteration, the rate of polymorphisms reduced drastically for all accessions (figure 18A). A log scale comparison shown for An-1 in the Figure 18B, showed that 25 iterations are needed for saturation of any existing polymorphisms, and corresponding improvement in mapped reads (figure 18C).

Differences in polymorphism density highlights the phylogenetic distances between the accessions. A comparative assessment of polymorphic rates amongst the six selected accessions determined Sha as the most divergent and L*er* the least divergent from Col-0 (Figure 18A). As expected, modified references show drastic reduction in the number of polymorphisms. The change after iteration 5 are too low to be clearly identifiable. Therefore, the log-transformed differences after each iterations are shown in Figure 18B for An-1. Reads mapped to modified references show saturation and little improvement after iteration 10. Similar pattern was observed in the remaining parental accessions. Parental-specific pseudo-references in the publication **??** were created following the same approach.

4.2.7. Expression estimation using SNPs

The comparative analysis and observations from simulated RNA-Seq analysis highlighted the challenges of expression studies and the impact of the genome of reference on read-mapping, quantification of expression, and assessment of variation in transcript levels.

The degree to which the SNPs read count will reflect total transcripts from gene, depends upon multiple factors including:

(i) SNP density within a gene,
(ii) SNPs loci within gene region,
(iii) gene expression, and
(iv) maximum mismatches allowed while mapping

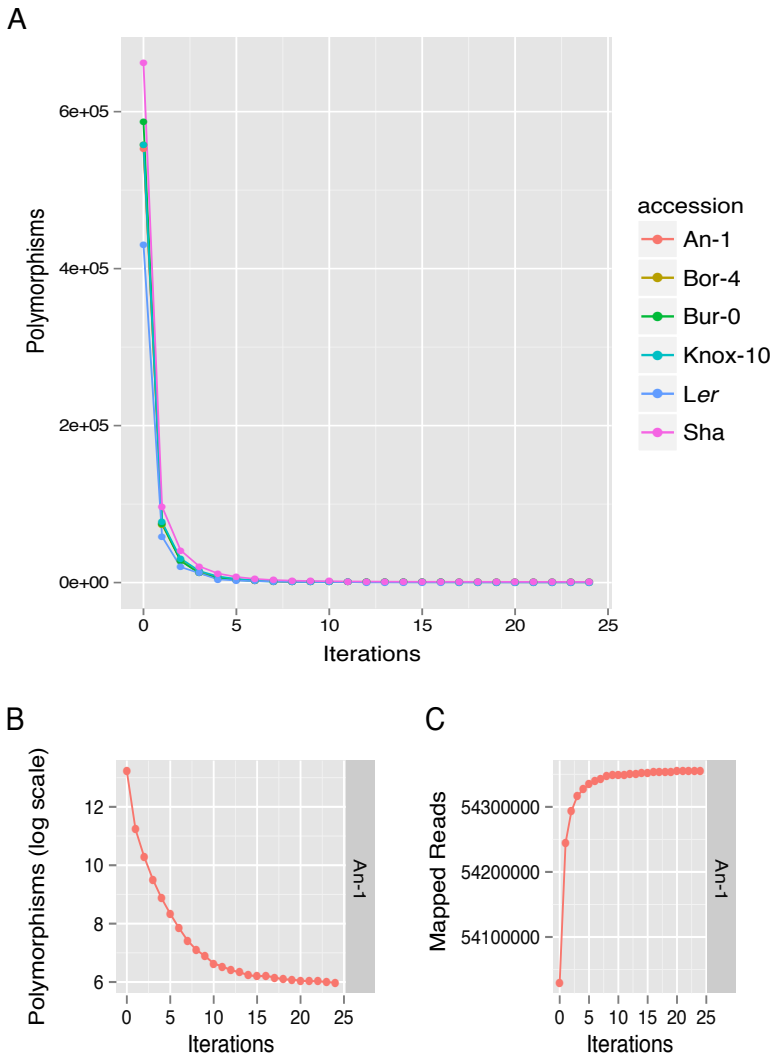


Figure 18. Iterative process of generating and improving parent-specific pseudo-reference (pref). The figure compares the Col-0 (iteration 0) against pseudo-references generated by incorporating detected polymorphisms, and remapping reads to modified reference after each iteration (iteration 1-25). (A) Polymorphism rate per iteration. The graph shows relative number of polymorphisms in the six parental accessions (legend) against Col-0 (iteration 0) and that of modified parental pseudo-references generated after each iteration. Accession-specific polymorphisms are incorporated into the Col-0 genome to generate a parental pseudo-reference at iteration 1. To improve the accession-genomes further, reads are mapped onto new reference. Subsequent decrease in polymorphic rate is shown after every iteration (X-axis). The process is repeated with the reference generated after every cycle of mapping and tweaking the and process is repeated in 25 iterations (X-axis) and decrease in the polymorphic sites after each. (B) Polymorphisms in An-1. Rate of polymorphic sites in the modified An-1 references produced after each iteration are compared to the polymorphic sites in original An-1 detected against Col-0 reference. (C) Effect on reads mapped on An-1. The plot compares number of An-1 reads mapped (Y-axis) to the Col-0 and subsequent An-1 references obtained after each iteration. It shows an Improvement in the mapped reads against Col-0 and modified An-1 genome after each iteration.

The above stated factors determine the abundance and uniformity of distribution of reads over the SNPs. Read coverage over genes is relatively higher around the inner region of the exons and gradually decrease towards exon-intron junctions, $3'$ and $5'$ ends. The reads over introns due to intron-retention events are relatively fewer. Certain exons may have high read distribution due to multiple reads from alternatively spliced isoforms. As discernible from mapping profile of a gene in the Figure 4. The read density over SNPs in the middle of the exons tend to be relatively higher than over those at the exonic ends and exon-intron junctions, introns, $3'$ and $5'$ ends and UTRs.

The reads tend to distribute more uniformly over multiple SNPs. However, single SNPs tend to be either overrepresented or underrepresented. Thus, with uniformly distributed reads over SNPs, the read measure correspond highly to gene transcripts. However, overrepresented or underrepresented SNPs tend to diverge strongly.

The measure of abundance of allelic-transcripts is highly dependent on the coverage of allelic SNPs.

Hence, we highly recommend evaluation expression divergence estimation over SNPs. The gene expression profile of simulated data was assessed as described in the section 4.2.5.2. on 39. The read coverage over SNPs was used to calculate transcript abundance over SNPs. The reads mapping over genes (observed expression), and read count over SNPs were independently compared against the number of reads simulated per gene (expected expression). The two cases were compared to estimate how accurately can the read coverage over SNPs, and over full length genes represent the gene expression.

The degree of deviation from the actual transcript abundance estimated over the whole-length of gene or exons was measured as log fold change of observed versus expected (control) read count. The log fold change value of 0 indicates that the expression level of gene is estimated precisely and hence the ratio of $exp : obs$ is 1. Positive deviation on a log scale implies that the observed number of transcript reads are lower than the expected, which could be explained by the number of unmapped reads. Duplicate genes, paralogs, and/or polygenic copies may be highly overrepresented by multiply mapping reads. Therefore, a certain percentage of negative deviation from the expression is also expected (Figure 16A).

To visualize the scale of divergence at SNP level, the simulated gene expression was compared to the read distribution over the SNPs. The divergence can arise from the difference in the read numbers. Hence, to correct for this bias, the read count from SNPs were pro-

vided to edgeR (Robinson *et al.*, 2010) along with gene reads. The assumption of the cases as separate libraries enabled normalisation against the difference in overall read numbers as well as the average number of reads per gene across the cases.

Ultimately multiple testing was implemented and the deviations were selected to be statistically significant for adjusted p-value less than 0.05. Figure 19 represent the log fold change of normalised read count over SNPs and genes. As evident from the MA plots, the estimation over SNPs is relatively less similar to the real expression.

Hence, most significantly deviating genes (red dots) with a positive deviation, depict the SNPs with lower than expected coverage. Additionally, some genes also show the negative deviation indicate SNPs with more reads than expected.

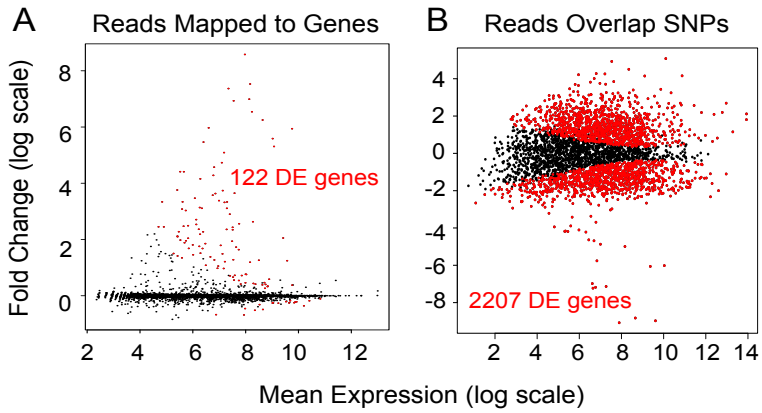


Figure 19. Comparison of expression estimation from read overlapping gene and only SNPs. The scatter plots represents mean expression per gene (shown as dots). The Y-axis shows the log fold change reflecting deviation from expected expression. (A) Cvi reads mapped to gene regions (Grc) with or without SNPs on Col-0 reference genome. (B) only overlapping SNPs (SNPrc) on the genes. Any significant deviation is highlighted in red colour and number of genes deviating are stated in the plot. Expression measured over reads mapped to genes is relatively more similar to expected profile than the SNPrc. It indicates that SNPs have high possibility of over and under-representation as discuss in above sections.

As evident from the Figure 19A, most genes show high similarity to simulated expression profile and only a few under- and overrepresented genes show significant divergence. In contrast, a large number of genes in Figure 19B show significant deviation for expression estimated over SNPs indicating unequal coverage over SNPs.

The large number of DEG underline the limitation of SNPs for transcript measure. Importantly, our observations highlight the emerging necessity to assess the influence of SNPs over the allelic measure. To increase the accuracy of ASE methods, we recommend selection of SNPs with equivalent expression profiles in the parentals. We have established a benchmarking protocol to filter for qualifying SNPs and it is described in further seg-

ments.

4.2.8. Significance of gene region for DE analysis

Depending on gene expression, the reads may be distributed differentially over exons, UTRs and introns. Besides, several studies reported natural variation in alternatively-spliced isoforms with intron-retention events indicating the possibility of differential read mapping over gene regions. The genes with retained introns are likely to be mapped by additional reads. Consequently, the read intensity over SNPs in these regions will tend to be highly distant from the actual gene expression profiles. The read bias of location-specific SNPs pose the next challenge to be addressed for improvement of ASE analysis method. To evaluate the gene region, for which the SNP read intensity is least diverging from the expected expression, a comparative evaluation for sensitivity was performed.

Considering the observations in the previous sections, Col-0 was selected as a standard reference for current analysis. The simulated expression profile of each Cvi gene was used as a control. The reads were then mapped to Col-0 genome. Firstly, the reads mapping to SNPs located in CDS, UTRs, and introns (full-length gene) were quantified as gene read count (Grc) data-set. Next, the reads mapping to SNPs in the introns were excluded from Grc to create a data-set specific to exons and UTRs. This is referred in further segments as CDS_UTR read count. Lastly, the reads mapping onto SNPs in UTRs were excluded to generate exon-specific CDS read count data-set.

Further, the three datasets were compared independently against the simulated expression (control) and divergence was measured as the ratio of expected versus observed. The divergence of the estimated measure was selected as significant for an FDR$<$0.05.

In each of the three comparisons, the genes which showed significant deviation from real expression were identified as false-positives (FP and those with least divergent expression estimate were selected as true-positives (TP). The ratio of true-positives to and false positives for each gene region was used to measure sensitivity.

The SNPs in UTRs and introns are likely to show more differential read distribution than the SNPs in CDS. Importantly, the restraining the estimate of SNPs only from exons and UTRs improve the expression estimate. The results indicate that using SNPs within exons and UTRs is most sensitive method to estimate transcript abundance. Therefore, we use the read counts from the SNPs within CDS and UTR region for the downstream analysis.

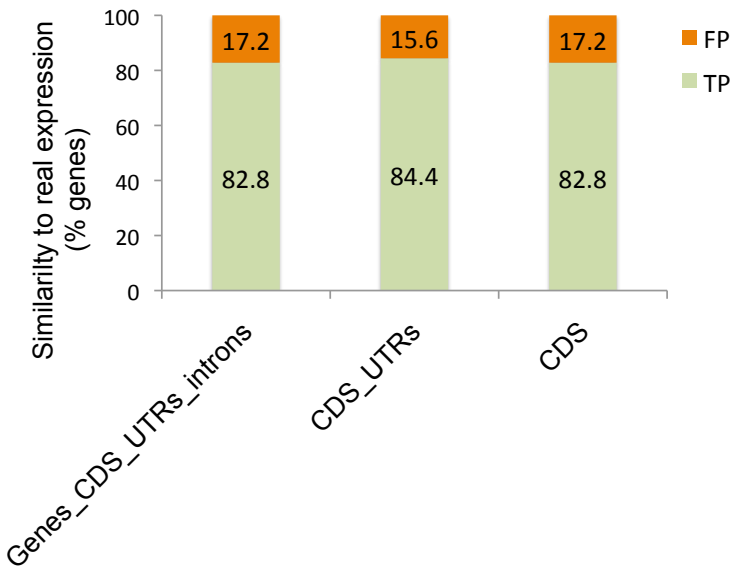


Figure 20. Expression estimation over SNPs in different gene regions. Barplot shows the TP (orange) and FP (light green) ratios (Y-axis) from estimated expressions on SNP levels within specified gene regions. TP represents proportion of genes with reads overlapping SNPs within specified regions highly similar to simulated (expected) expression. FP represents significant deviation from expected abundance levels.

4.2.9. Benchmarking criteria for ASE analysis

Accuracy of ASE analysis relies strongly on the SNPs to distinguish the allelic transcripts within hybrids and the reads mapped over SNPs to quantify transcript abundance per allele. Figure 4 depicts the variation in read distribution over SNPs determined by the gene expression levels, SNP location and density and mapping parameters. SNPs within highly expressed genes are more likely to be enriched with reads than the SNPs in the minimally expressed genes. Genes with single SNP might be differentially represented over the SNPs in contrast to the read coverage over multiple SNPs. Ideally, within the gene with high SNP density, the reads should be distributed rather uniformly over most SNPs. Consequently, for an over-expressed gene with a high number of SNPs, the estimation from SNP reads might be higher than the corresponding gene expression. On the other hand, poorly expressed gene with one or few SNPs may provide a lower estimate for the expression profile.

The gene region is an additional crucial determining factor for relevance of expression elucidation from SNPs. SNPs in the introns, 5′- and 3′-UTRs have fewer reads mapping to them unlike the preferentially covered exons.

Therefore, to ensure selection of best possible SNPs for optimal allelic measure we eval-

uated the gene regions for most closely represented SNPs. The results are discussed in the section 4.2.8. on page 48. The observations underlined the region-dependent SNP coverage and the corresponding error rate in measured expression profile. The magnitude of gene expression is reflected with high sensitivity by the read distribution of SNPs within exons and UTRs. Based on the results, we recommend use of exonic and UTR reads for expression quantification and designating related SNP reads for comparing allelic profiles.

The differential estimation of expression profiles of SNPs from exons raise a few concerns. Indeed, these observations outline the next crucial challenge for ASE quantification to alleviate bias in the method. The number of genes with non-uniformly represented SNPs is quite high, even within region with the highest similarity with the overall expression. Discarding such genes might reduce crucial candidates. The degree of similarity to gene expression is a decisive factor controlling the efficiency and accuracy of the ASE results. Comparative assessment for similarity in the exonic and SNP estimate of gene expression, is inevitable to select candidates with congruency in SNP and exonic reads. ASE in SNPs from these genes is likely to fine-select for high proportion of true-positives.

However, it is noteworthy to consider the fact that the available methods for differential expression have been primarily developed to address the large scale gene-to-gene comparisons. Hence, gene-to-SNP comparisons pose a challenge given the sheer factor of difference in the numbers itself. To address this caveat beforehand, the gene and SNP data were provided as different samples. The library normalisation in edgeR takes into account the differences within and across the samples.

The normalisation enables reduction of noise, as evident, from the large number of genes determined with similar expression profile across SNP.

Notably, the variation of transcript profiles at the SNPs from the gene emphasizes the need to shortlist the candidate genes with similar expression profiles.

However, it is extremely important to assay the SNPs for differential allelic expression in hybrids. Pronounced differences in the estimated expression levels using SNPs and exons strongly bias computation of variation. The allelic variation determined from SNPs in the hybrid may be technical bias and result in misinterpretations. Therefore, to minimise the noise in the variations, computed simply as a factor of over- and under-estimated allelic levels in SNPs, the scale of error rate must be determined for DE estimation in exons versus SNPs.

Based on the DE analysis the genes can be broadly distinguished in four classes:

(i) genes with significant expression change in same direction for exons and SNPs
(ii) genes with differences significant only over exons
(iii) genes with significant changes only over SNPs

The robustness of method can be established depending on the relative proportion of genes with concordant variation at exon and SNP to the uniquely identified changes. Increase in the number of genes with exonic and SNP read data reflecting similar scale of changes, will reduce the probability of erroneous variation, and improve the predictability. Hence, an exhaustive rating of expression divergence in exons and SNPs was performed and used to fine-select the threshold limits.

RNA-Seq data from leaf-tissue of An-1 and L*er* was used for this experiment. The reads were mapped to Col-0 and the SNPs were determined against Col-0 and subsequently classified as diagnostic SNPs if the base at the loci was distinct between An-1 and L*er*. Next, the reads spanning exons and UTRs were accounted for as exonic reads for An-1 and L*er* and processed in edgeR for DE analysis.

The Figure 21 depicts the distribution of genes with varying degree of change in expression. The scale on the X-axis represent the log fold change of An-1 versus L*er* and shows the most deviating genes.

Majority of genes do not show expression changes between the parental accessions An-1 and L*er*. Less than 3% genes show significant (FDR < 0.05) changes in expression, with only 2% showing strong differences. Nearly all large-scale differences are detected as statistically significant. The large scale changes are most likely due to cis- and small-scale changes are due to combination of cis-, trans- and other patterns. To elucidate the underlying regulatory mechanisms the parental ratio is compared to allelic ratio and summarised in Figure 23.

The expression divergence was considered significant at FDR $< 0.05\%$. Similar analysis was performed for read count over SNPs within exonic and UTR regions within the hybrid An-1 x L*er*. The differentially expressed genes determined using reads mapped over exons were compared to corresponding allelic profile from reads mapped to SNPs, and find the pairs for which the magnitude and direction of variation is similar. Significant expression changes determined by exons and SNPs have been compared against each other and plotted on the scatter plots on the left panel in the Figure 22. with the bar plots

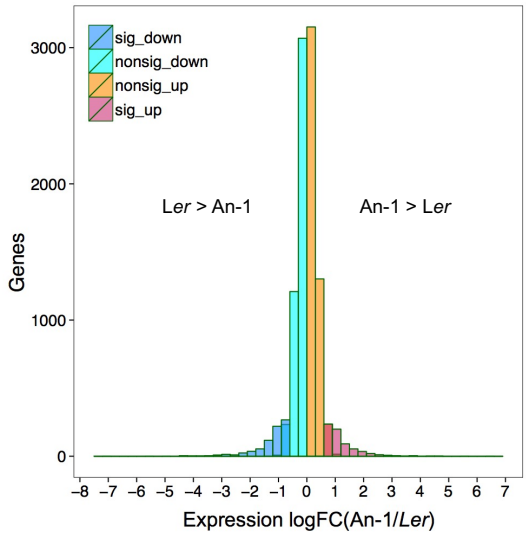


Figure 21. Magnitude of expression divergence between An-1 vs L*er*. The barplot compares the magnitude of expression divergence between parentals An-1 and L*er*. The log fold changes in parental genes have been divided into multiple bins depending on the degree of variation ranging from 0 (no variation) to 8 (strongest) with corresponding genes (X-axis) shown by bars. Most genes show negligible and non-significant change (log fold change $<$ |1.5|) for up-regulated for An-1 $>$ L*er* (orange) and down-regulated An-1 $<$ L*er* (skyblue). Less than 3% genes exhibits strong and significant expression changes (log fold change$>$ |1.5|) for up-regulated An1 $>$ L*er* (pink) and down-regulated An-1 $<$ L*er* (blue).

on the right panel with number of DE genes determined by each category. The exon and SNP profile was examined for expression variation. The stringency in the thresholds was increased and a comparison was made at each level.

A comparative assessment of the results ES consistency in results with change of specific parameters. As a raw control, read data for every gene was used without a threshold filtered and thereafter subsequently increasing constraints. An interesting pattern in reduction of noise from SNP data surfaced.

On the scatterplot for null threshold, the genes detected to be show similar magnitude and direction of expression change over both exons and SNPs are highlighted in grey. These genes arrange around the diagonal from lower left corner to upper right corner. If the differences are significant only over SNPs, the genes are shown by yellow coloured dots and corresponding bar in the adjacent bargraph. These cases reflect over- and under-mapped SNPs. However, if the differences are significant over only the exons, the genes are depicted as cyan dots and cyan bar states the corresponding percentage. One of the reasons could be a low SNP density, or low coverage over SNP location.

Although, the proportion of DE genes detected by both exons and SNPs was rather low at 45%. However, the percentage of genes that are significantly variable and detectable over

only SNPs is alarming 44%. Interestingly only for 11% of DE genes, the corresponding information is not valid at SNPs. Such genes can not be selected for allele-specific information, because the SNPs lack the power and accuracy for allelic measure. Thus, this comparison gives us a first hint of the genes to exclude from ASE analysis.

Interestingly, DE analysis for gene with normalised read measure of rpkm$\geq$1 reduced the difference between exon and SNP specific DE estimation. Subsequently, the genes with normalised read count of mimimum 1 rpkm, were retained for DE analysis within exons and in SNPs. A drastic reduction in the proportion of DE genes from SNPs indicate improvement in the elimination of noise. In the DE genes almost 15% more genes showed congruent expression variation in the exons and SNPs.

Further, the genes with less than 5 reads over the SNPs were discarded and the rest was assessed for DE analysis in exons and SNPs. The overall pattern changed slightly and almost an equal and low proportion of DE genes were determined uniquely whereas 66% of DE genes were identified in both conditions. A further increment of SNP read count limit to 10 reads did not show drastic changes to improve the results.

Genes with higher than log fold change of 1 were processed and a significant increment in the concordant results from exon and SNPs was observed to be almost 86%, with a substantial reduction in SNP-biased variation to only 2%. Increasing the log fold change limit for exons to higher than 1.5 eliminated almost 40% of genes, with a significant improvement in the consistency between SNPs and exons. Additional constraint by increasing the exon log fold change to greater than 2, proved to be most stringent and the almost 92 percent of gene detected to be differentially expressed (DE) agreed with exonic and SNP read count. The SNP and exon-specific DE genes were significantly lower.

We used these results to standardize the parameters including the threshold of gene expression at parental level, the minimum read count over SNPs, and minimum fold change at the exonic levels. The pattern of exonic and SNP congruity has been assessed to determine the threshold values for detectable and robust similarity. Additionally, the evaluation enables the exclusion of genes with inconsistent expression estimate over SNPs and exons. The fine selection of genes with a high degree of concordance of read abundance in exons and SNP increases the predictability of variation at the allelic level.

4.3. ASE analysis

The workflow for ASE analysis was applied to differentiate and quantify relative contribution of underlying regulatory mechanisms of intra-specific expression variation be-

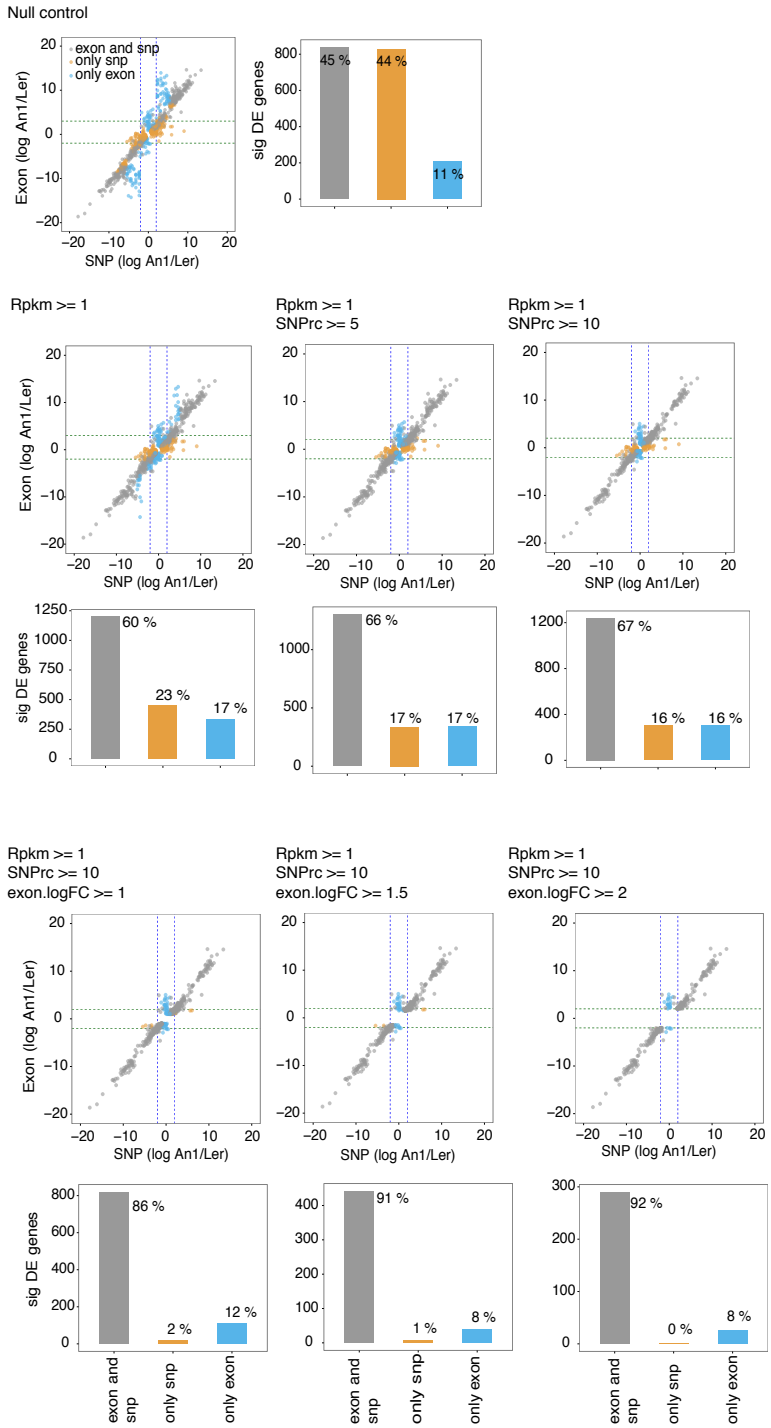


Figure 22. Exonic versus SNP estimate of expression changes. Comparison of effect of various parameters (left panel) in expression divergence estimation over exons and SNPs. It is crucial to the selection of thresholds for improvement in the method to distinguish ASE patterns.

tween *Arabidopsis thaliana* accessions. Relative allelic contribution in hybrids were compared against allelic ratio in corresponding set of parental accessions. For the analysis, the RNA-Seq data from leaves of each of the three replicates of *Arabidopsis thaliana* accessions An-1, Bor-4, Bur-0, Knox-10, Sha, and L*er* were mapped onto Col-0 using TopHat version 2.0.6 (Trapnell *et al.*, 2009; Kim *et al.*, 2013). SNPs for each accession against Col-0 were determined from the mapped reads using GATK (McKenna *et al.*, 2010). The RNA-Seq data from leaves of triplicates of the reciprocal hybrids An-1 x L*er*, Bor-4 x L*er*, Bur-0 x L*er*, Knox-10 x L*er*, Sha x L*er*, L*er* x An-1, L*er* x Bor-4, L*er* x Bur-0, L*er* x Knox-10, and L*er* x Sha were mapped independently against Col-0 as the reference genome. Diagnostic SNPs between set of parentals were selected to distinguish allelic reads in corresponding hybrids.

The expression was quantified for each set of parentals from reads mapped onto exons (*exonrc*) and SNPs (*SNPrc*). The expression was further normalised to account for differences in library size and gene length using read count per million (RPKM). The genes were selected for downstream ASE analysis if they had expression with at least 1 RPKM in all replicates of at least 1 of the parental accesions per set. The gene list was reduced to the genes with at least 1 SNP, with minimum of 10 reads overlapping exonic SNPs. Further, expression ratio between each set of parentals was quantified independently from *exonrc* and *SNPrc* to determine statistically significant difference at FDR$<$0.05 using edgeR (Robinson *et al.*, 2010).

The genes with same direction and similar magnitude of expression divergence estimated at SNPs and exons were selected for ASE analysis. The allelic ratio for each hybrid was quantified from SNPrc. Direction and magnitude of expression divergence of genes and alleles for each set of parentals and corresponding hybrids (Figure 23A).

Genes with log fold change $>$ 1 and statistically significant differences (FDR$<$0.05) in parentals with similar magnitude and same direction of change in alleles were classified as cis-effects (magenta dots along the diagonal in Figure 23A). The ones with non-significant deviation in alleles, in contrast to significant changes in parents were classified as trans (green dots in Figure 23A).

The Figure 23A reflects the An-1/L*er* ratio. Similar analysis was performed for each pair of parentals and reciprocal hybrids with L*er* in both directions.

The relative proportion of cis, trans, cisxtrans, and compensatory regulatory effects were comparable among all sets as summarised in the Figure 23B. Approximately 575 genes

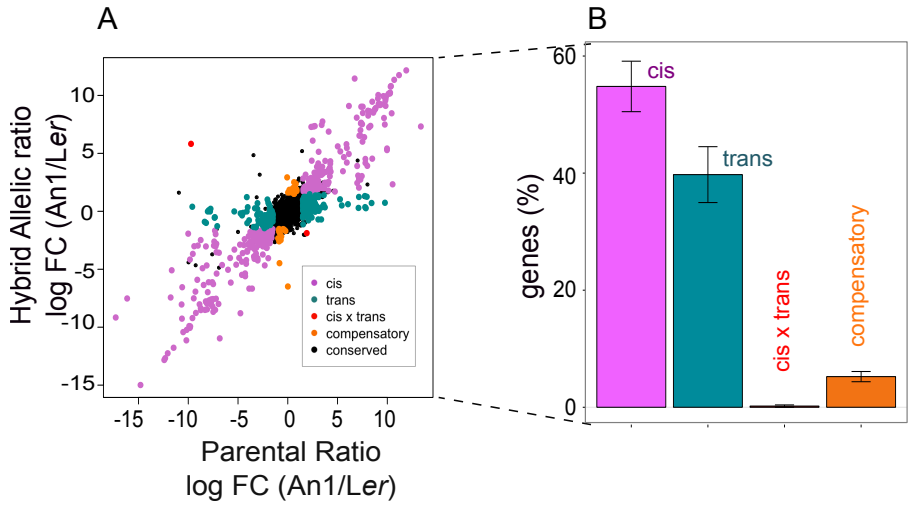


Figure 23. Determination and quantification of cis- and trans-regulatory divergence in *Arabidopsis thaliana* accessions (A) Scatter plot showing the relative allelic expression ratio (log fold change An-1/L*er*) in the hybrid An-1 x L*er* (Y-axis) and corresponding gene expression ratio (log fold change An-1/L*er*) of parentals (X-axis). Each coloured dot represent a gene showing statistically significant difference in expression (FDR<0.05). The colour codes correspond to the underlying regulatory mechanisms- cis (magenta), trans (green), cis x trans (red), compensatory (orange). The genes with no significant change with conserved expression pattern both in hybrids and parentals are shown as black dots. (B) Barplot summarises the relative cis, trans, cis x trans, compensatory effects (X-axis) quantified from each pair of parentals corresponding to the hybrids and reciprocal hybrids. The relative contribution of the regulatory effects are measured as the percentage of genes (Y-axis) under each category.

under different regulatory patterns were determined across the hybrids. Of these nearly 55% genes were regulated by cis-effects, ∼39% by trans-effects, ∼5% by compensatory and less than 1% controlled by cis x trans effect (Figure 23B).

Thus, cis- effects were found to be more prominent, causing in expression divergence of ˜55% differentially regulated genes. The ASE patterns in *Arabidopsis thaliana* accessions are in lines with the previous reports on maize (50-70% by (Guo *et al.*, 2008)). Several different proportion of cis- and trans- effects have also been reported in other studies, viz. 51% cis, 66% trans in *Drosophila* (McManus *et al.*, 2010), 27% cis, 29% trans in *Arabidopsis* hybrids (Zhang and Borevitz, 2009).

4.3.1. Cis-regulatory mechanisms contribute to large scale expression divergence

Cis-regulated expression variation is directly imparted by localised genetic changes, which likely have a drastic impact on gene expression and associated plant traits.

The degree to which cis- and trans- genetic regulatory variants affect the magnitude of expression could have serious fitness and evolutionary consequences for the plants. Depending upon the fitness consequences of multiple mutations, the deleterious ones are likely to

be selected against, resulting in natural selection, accumulation and fixation of beneficial ones over long evolutionary periods. Therefore, it become crucial to assess the relative impact of cis- and trans- variants on magnitude of expression divergence. The expression divergence between each pair of parental accessions imparted due to cis- or trans- regulatory effects was plotted in Figure 24. This analysis shows that genes under cis-effects

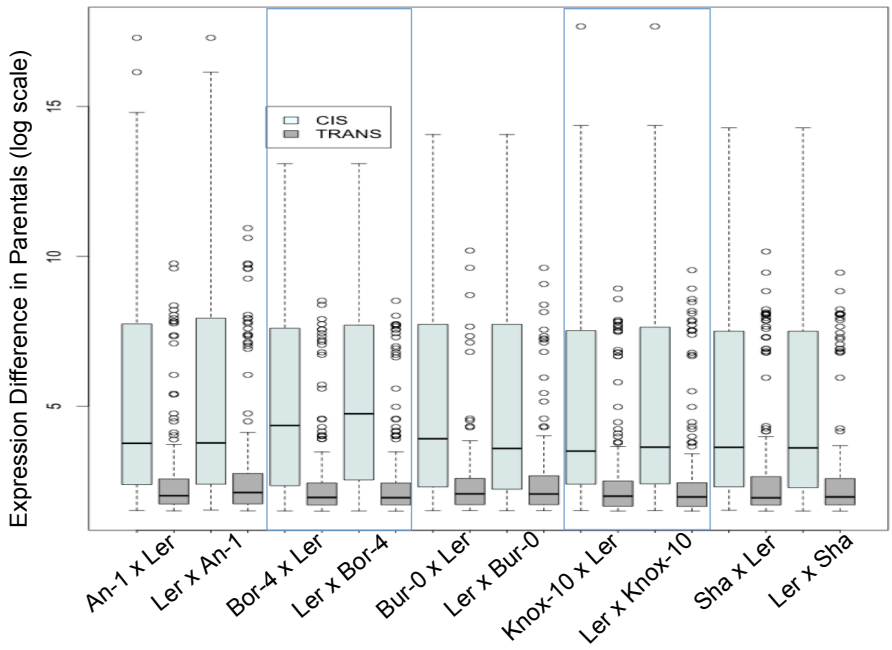


Figure 24. Magnitude of expression divergence (log fold change) due to effect of cis- and trans- regulatory variants. Boxplot summarises the relative effects of cis and trans variants on the intra-specific expression divergence in parentals. Y-axis represents log fold changes in expression divergence between parental accessions for genes under cis- and trans- regulatory effects in each set of reciprocal hybrids shown in X-axis (e.g. log(An-1/L*er*) for An-1 x L*er* and L*er* x An-1).

exhibits a large diversity in expression variation and the genes with large fold changes in expression are regulated in cis. Unlike the lower fold changes in trans-regulated genes. Similar observations have been reported previously in Drosophila (McManus *et al.*, 2010) and maize (Lemmon *et al.*, 2014).

Cis-variants unlike trans-variants induce larger degree of variation in the expression, which are likely to have a substantial role during evolutionary process as explained for maize (Lemmon *et al.*, 2014). The small scale divergence in expression due to trans-regulated variants could be explained by the fact that trans-regulatory mechanisms have a higher degree of pleiotropy and influence multiple genes in the similar manner unlike the cis-mechanisms which have specific targeted effect on expression.

4.3.2. Trans-effects are more unique to accessions

To determine if the genetic regulatory variants imparting the expression divergence were common in the multiple *Arabidopsis thaliana* accessions under consideration, ASE results from the 5 hybrids with L*er* as paternal parent were selected for the analysis.

Reciprocal crosses could add to effect of parent-of-origin or plausible imprinting effects and hence were not considered further for this analysis. The genes determined to be regulated in cis- and trans- from each hybrid were combined to generate the pattern specific gene sets. The cis- and trans- genes from each hybrid were considered as subsets and compared against the summarised gene sets of corresponding pattern. The cis and trans-genes common to multiple sets (hybrids) and those unique to each set (hybrid) were compared (Figure 25).

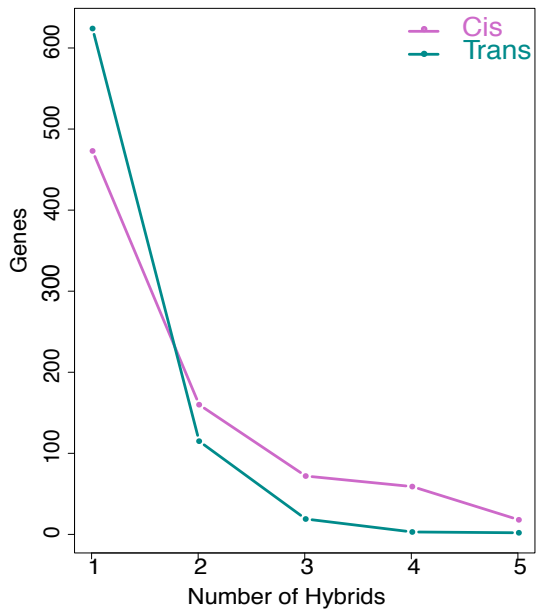


Figure 25. Conserved cis- and trans-effects across the *Arabidopsis thaliana* hybrids. The line graph shows the number of genes under cis- (magenta) and trans-effects (green) on Y-axis and the number of hybrids with L*er* as paternal parent on X-axis. Each dot correspond a value which represent the number of hybrids in which the cis- and trans- genes show common expression divergence pattern.

More genes under cis-effects were found to be common in multiple hybrid sets. In contrast a large proportion of trans- genes were found unique to single hybrid, which indicates that unlike the more common cis-effects, trans-regulated variation is more hybrid-specific.

It could be attributed to two major factors. Firstly, the effect of common parental in the

hybrids. Since, L*er* is the common paternal parent in the hybrids considered for analysis, the cis-effects detected in all 5 hybrids may reflect the L*er* specific cis-variants. Secondly, the effect of the different maternal parent of the hybrids. Despite the different contributing maternal parent, the cis-effects common in more than 1 hybrid can be hypothetically explained by the effect of common polymorphisms against Col-0 reflecting cis-variants likely to have been integrated into the corresponding *Arabidopsis thaliana* accessions during adaptation over evolutionary period.

Also, it is very well understood that the trans-effects arise as a result of distant acting variations, normalised under same environment within a hybrid. Trans-variants might be differentially regulated in the hybrids depending on the contributing pair of parentals. Since the regulatory variants of L*er* maternal parent are common in all hybrids, the paternal parent different in each hybrid might be the determinant factor, imparting the unique trans-regulated expression divergence.

4.3.3. Comparison of RNA-Seq and Pyrosequencing in determination of allelic bias

ASE analysis was carried out in all hybrids and corresponding accession. Fours different methods were used. Firstly, to assess effect of relative SNP abundance on estimation of expression differences, allelic expression variation was quantified separately from reads mapped over single SNP per gene (Figure 26A), and reads mapped only over all SNPs (Figure 26B). Secondly, to determine any bias of reference genome, expression divergence was quantified from Col-0 (Figure 26C) and parent-specific pseudo-references (Pref) (Figure 26D).
The allelic ratio of randomly selected genes under the the regulatory category cis-, trans-, cis x trans and compensatory, measured from RNA-Seq, using each of the above stated criteria was compared against the corresponding allelic frequency obtained from Pyrosequencing (Figure 26). As confirmed by Pyrosequencing, estimating expression differences over all SNPs correlates better with RNA-Seq based allelic ratio. Use of Col-0 shows a higher correlation between RNA-Seq and Pyrosequencing ratios suggesting that the use of Col-0 is slightly better for intra-specific expression variation measure. The analysis supports using all reads overlapping over all possible diagnostic SNPs for allelic quantification. Also the evidence shows that the use of pseudo-references do not have any significant improvement in allelic quantification within the intra-specific comparison of *Arabidopsis thaliana* accessions under consideration. Hence, it was decided to use every SNP in the genes, and Col-0 was selected as the genome of reference for downstream expression analysis.

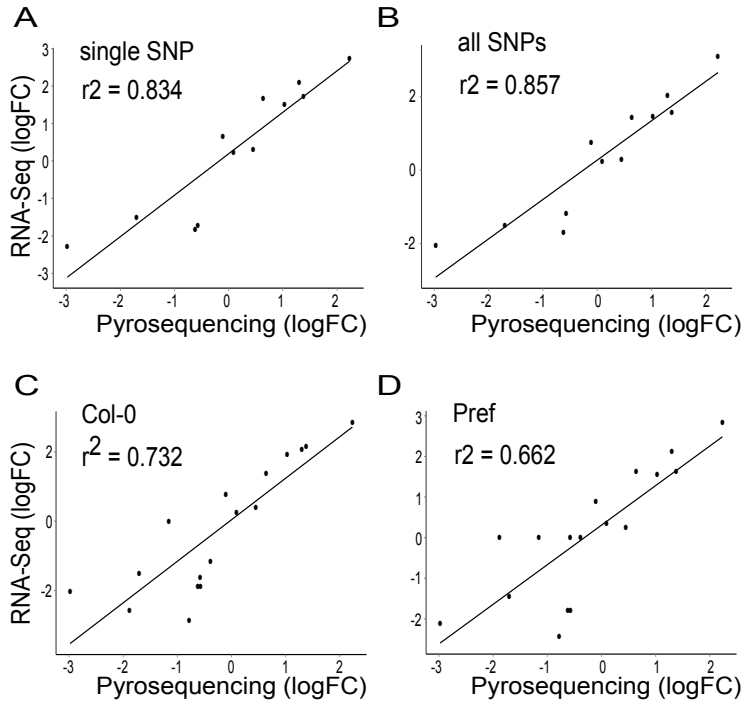


Figure 26. Comparison of Allelic Ratio quantified from RNA-Seq and Pyrosequencing Method. Scatter plot showing correlation between RNA-Seq and Pyrosequencing based quantification of allelic ratio using the following criteria: I.(A) allelic ratio quantified from reads overlapping only single SNP, and (B) all SNPs per gene; II.(C) allelic ratio measure using Col-0 reference genome, and (D) parental-specific pseudo-reference genomes (Pref); A higher value of coefficient of correlation indicates relatively better allelic ratio estimation for a method.

4.3.4. Inheritance patterns of expression

Plant expression data (section 3.4.1.) was sequenced with an objective to quantify Allelic effects (Figure 23). The same data can be investigated from a different perspective with an aim to identify inheritance patterns of gene expression in hybrids. Inheritance expression patterns can further uncover the plausible heterosis events (Figure 7 on page 15). To distinguish and quantify effect of regulatory variation on the inheritance patterns of gene expression in hybrids, comparison of parental and hybrid gene expression was performed. Depending on the pattern exhibited in hybrids with respect to corresponding parentals, each gene was further classified into the categories of heterosis events (Figure 7A) namely overdominance, high-parent dominance, low-parent dominance, underdominance and additive. To quantify the relative cis- and trans effects on the expression patterns, genes of each category were compared against the results of allele-specific expression (discussed in section 4.3. on page 53).

Inheritance patterns were determined from preliminary analysis of gene expression in parentals and hybrids considering the criteria (Figure 7) Majority of the expressed genes ($\sim$69.3%) were detected with conserved mode of expression i.e, the expression profiles of these genes in hybrids and parentals do not differ significantly (FDR$<$0.05). The remaining $\sim$30% genes were detected exhibiting heterosis patterns in hybrids. Majority of genes were detected to exhibits dominance ($\sim$14.69%) and additive effects ($\sim$10.98%). Similar observations were reported in cotton (Bao *et al.*, 2019). Amongst the non-additive expression categories, Overdominance ($\sim$2.51%) and underdominance ($\sim$2.51%) were the rarest heterotic events. Dominance was found to be most prominent in hybrids with $\sim$7.72% genes depicting high-parent dominance, and $\sim$6.97% genes showing low-parent dominance.

Genes with additive and non-additive expression patterns were further examined to determine and quantify the underlying regulatory mechanisms i.e, cis- and trans- effects. $\sim$76% genes were found without exclusive cis- and trans-effects (depicted as "others" in the figure 7B). $\sim$14.3% genes were detected to be under cis- and $\sim$9.6% genes under trans- effects.

Hence, cis-effects were found to cause more heterotic events in comparison to trans-effects. This can be explained by the scenario where in a F1 hybrid both parental alleles within same environment are under the influence common trans-regulatory factors. Hence, the difference in parental accessions due to trans-effetcs can be overcome in hybrids. However, expression deviation caused by allelic effects by cis-variants will likely be reflected within the hybrids.

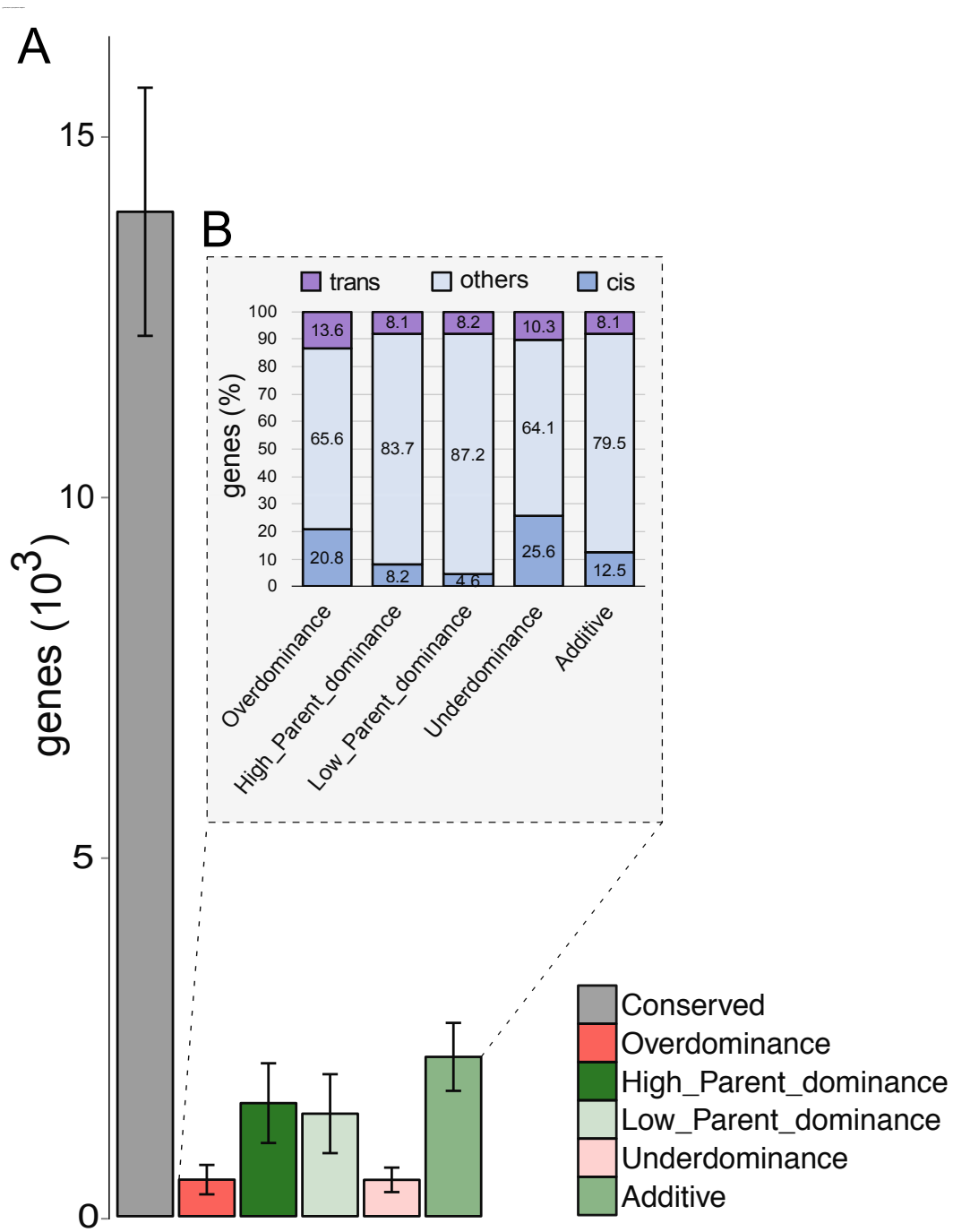


Figure 27. cis- and trans effects on expression inheritance patterns in hybrids. (A) Bar plot shows the number of genes corresponding to the categories of expression patterns in F1 hybrids with respect to parental accessions. (B) Stacked Bar plot showing the relative cis- and trans effects on genes corresponding to each inheritance category.

The cis-variants may strongly affect the hybrid expression as evident from relative proportions of ∼20.8% cis- and ∼13.6% trans-effects causing overdominance. Similar pattern is observed for underdominance where ∼25.6% cis- and only ∼10.3% trans-effects contribute to heterosis. The difference in also observed for genes with additive effect in hybrids regulated by ∼12.5% cis- and ∼8.1% trans-effects. However, cis- and trans-variants have comparable influence on genes under dominance, with high-parent dominance regulated by ∼8.2% cis- and ∼8.1% trans, and low-parent-dominance influenced by ∼4.6% cis- and ∼8.2% trans-variants.

The fact that more than ∼76% were not exclusively regulated by cis- and trans- effects highlight the multi-factor regulation of gene expression within hybrids which may also include compensatory, combined (cis and trans), and antagonistic (cis x trans) effects. Most importantly, cis-effects are responsible for more drastic expression divergence underlying the rarest over- and under-dominance events.

4.3.5. Parent-of-origin effect

Preliminary assessment was performed to determine plausible parent-of-origin effect on the observed bias in allelic expression within the hybrids. This simplistic model comprises of comparison between reciprocal hybrids and corresponding parental genes (shown in the Figure 6 on page 13).

Therefore, to distinguish genetically induced allelic bias from those influenced by parent-of-origin, for each pair of reciprocal hybrids, the gene expression values were compared against the corresponding set of parentals (Figure 28). The genes with significant allelic bias which were tested for ASE, were selected for imprinting analysis. Depending upon the up and downregulated parental allele in each pair of reciprocal hybrids, the genes were further categorised into plausible maternally expressed genes (MEG), paternally expressed genes (PEG), and non-imprinted genetic allele-specific expressed genes (ASE) as per the schema in the Figure 28. Majority of the differentially expressed hybrid alleles were independent of parental induced bias (∼78.65% ASE). Importantly large proportion of genes detected under plausible imprinting were expressed maternally (∼12.36% MEG), and only ∼8.99% genes were expressed paternally. These findings are in concordance with the previous patterns in previous reports on *Arabidopsis thaliana* (Fort *et al.*, 2017; Tuteja *et al.*, 2019; Gehring *et al.*, 2006), rice (Shao *et al.*, 2019), and mice (Chen and Begcy, 2020). Using the ASE results from reciprocal hybrids the parentally-biased allelic expression could be determined and distinguished from the non-imprinted allelic effect. Hence this method provides a simplified and efficient identification of plausible imprinting effects. Notably, the imprinting effects are tissue-specific and mostly regulated

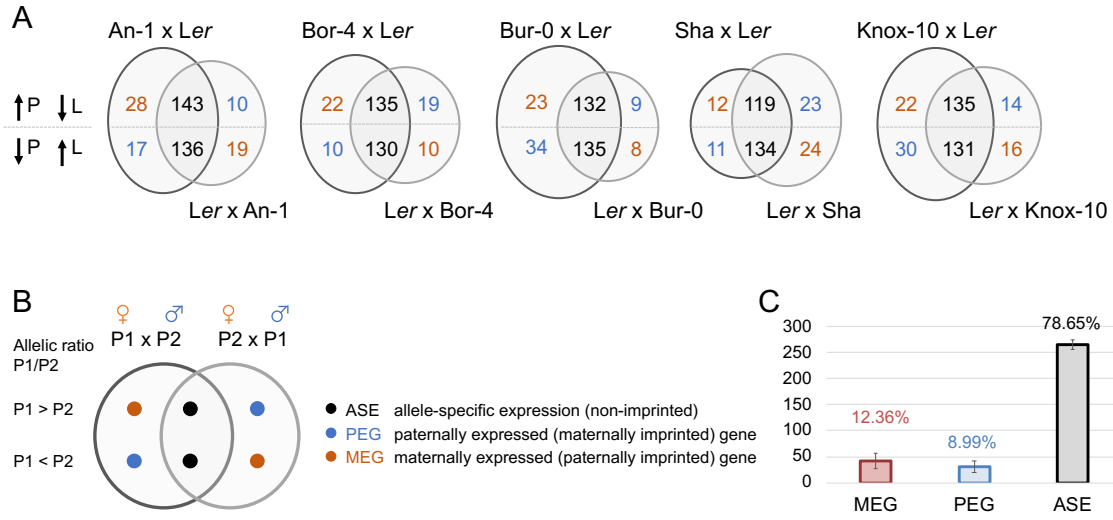


Figure 28. Comparison of cis-effects between reciprocal hybrids to determine parental-effect on allelic bias. (A) Venn diagram comparing each pair of reciprocal hybrids. The upper hemisphere represents upregulated maternal (orange) and down regulated paternal (blue). The lower hemisphere represents the downregulated maternal (orange) and the upregulated paternal (blue) allele. The common ares of venn diagram represents the differentially expressed alleles independent of maternal and paternal bias. P represents parental accession other than L*er*, L depicts L*er*. (B) Reference schema to decipher alleles affected by parent-of-origin by venn diagrams of pair of reciprocal hybrids, each category represented by coloured dots, i.e, ASE (black), PEG (blue), MEG (orange). (C) Barplot showing relative proportion of genes determined to be under parental influence on allelic expression.

by maternal allele in embryo and endosperm during initial seed developmental stages (Gehring *et al.*, 2011; Springer and Stupar, 2007a). Hence, data source plays a crucial role in detection of imprinting effects. With a goal to develop efficient protocol for ASE detection and quantification RNA-Seq was obtained from leaf tissue. With use of distinct RNA-Seq data from growing seed tissues, and implementation of above proposed methodology, we expect to determine many more genes expressed maternally and paternally.

4.4. Metabolite analysis

4.4.1. Primary metabolite profiling by GC/TOF-MS

Metabolite profiling was carried out to detect and quantify primary metabolites in the plant samples and the reciprocal hybrids used for ASE analysis. Metabolite concentration was measured as peak areas of the mass (m/z) fragments and were normalised to the internal standard (Ribitol) and fresh weight of the samples.

4.4.2. Metabolite data analysis

The quantified primary metabolites were further classified according to the biological role and pathways involved as per Gibon *et al.* (2006). To determine the intra-specific variation of the metabolites from the samples under consideration, the data

was normalised for each pair of parental samples and respective hybrids against the mean of metabolite value for the group (mean normalised value). Further, the relative normalised metabolite levels were compared across the plant accessions and hybrids. Genes involved in pathways of each metabolite were examined for expression profiles. To determine the association of the natural variation in gene expression to the relative metabolite concentration, the concentration of most abundant and variable metabolites were compared to the expression profiles of pathway-specific genes. Pathway information were downloaded from the three major metabolic pathway databases namely, AraCyc (*http://www.arabidopsis.org/tools/aracyc*; (Mueller *et al.*, 2003)), PMN (*http://www.plantcyc.org/*) and KEGG (*https://www.genome.jp/kegg*; (Kanehisa and Goto, 2000)). The expression values of the pathway specific genes for the most variable metabolites were tested for correlation the metabolite profiles. The case found with maximum correlation signalling high degree of association have been reported in this thesis.

4.4.3. Pathway analysis

AraPath is a comprehensive database for pathway analysis in plant genomics with data from *Arabidopsis thaliana*. It comprises information for 4332 gene set categories from different sources including literature and annotation related databases KEGG (109), AraCyc (224), Gene Ontology (941) and Plant Ontology (230), transcription factors (33) and microRNA (309) target genes, co-expressed genes (48) from various gene expression studies.

AraPath metadata file (*http://bioinformatics.sdstate.edu/arapath/AraPath_all_v1.gmt*) with information of database, corresponding metabolite pathway, annotation and related list of genes was downloaded.

AraCyc and KEGG categories were extracted and further processed using customised Perl scripts to select *Arabidopsis thaliana* specific gene lists. Gene sets of AraCyc and KEGG consists of 594 and 887 *Arabidopsis thaliana* genes respectively.

Metabolite genes which could be tested for ASE were selected. It reduced the data to 11 AraCyc pathway set and 9 KEGG.

Gene set categories from AraCyc were selected for profiled metabolites. For the selected metabolite, the gene were separated according to the biosynthesis, degradation, metabolism or synthesis.

4.4.4. Correlation analysis of metabolite and transcriptomic profiles

38 identified primary metabolites were classified were categorised in different classes based on biological function or metabolic pathway as per MapMan Classification. Correlation analysis was performed to determine any existing correlation between metabolite concentration and gene expression profiles. Genes showing high correlation coefficient (R^2 >=0.8) for each metabolites were quantified (Figure 29).

Majority of the identified primary metabolites comprised of amino acids (42.11%), followed by eight Organic acids, four polyamines, and three sugars (Figure 29). In addition amino acids comprised of majority of genes showing strong correlation between metabolite and transcript profiles.

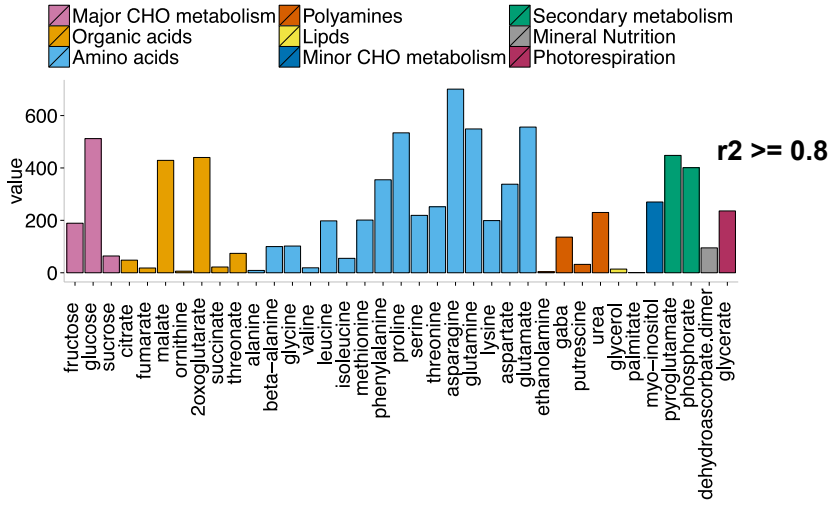


Figure 29. Categories of primary metabolite in *Arabidopsis thaliana* leaves and corresponding genes with high degree of correlation (R^2 >=0.8). The barplot depicts the number of genes with strong correlation between metabolite and transcript profiles across the accessions and hybrids. The categories of metabolites have been shown in distinct colours.

4.4.5. Hierarchical clustering analysis to identify similarity based patterns

Hierarchical clustering analysis helps uncover the association and impact of the metabolite phenotype of respective genotypes. To determine degree of similarity in metabolite levels across *Arabidopsis thaliana* accessions and reciprocal hybrids, hierarchical clustering of mean normalised value of metabolite levels was carried out.

Heatmap in the Figure 30 depicts the normalised metabolite concentrations across the accessions and hybrids. The plot represents specific groups with relatively high and low intensity for specific accessions.

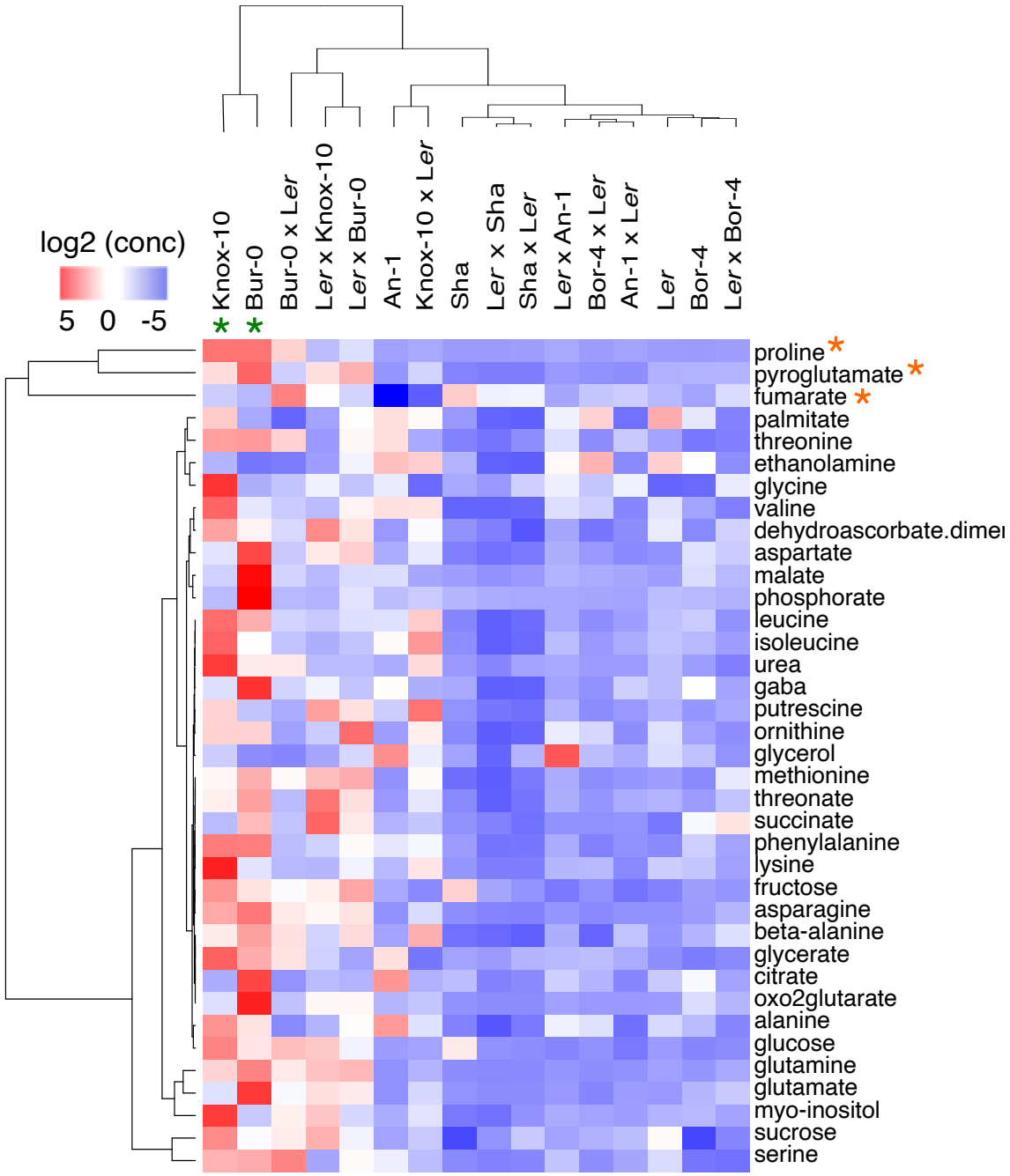


Figure 30. Hierarchical clustering analysis of primary metabolites from leaves of *Arabidopsis thaliana*. Heatmap showing the relative normalised intensity of metabolites (rows) in the accession and the hybrids (columns). The values along each column depict average metabolite levels across three replicates per sample. Metabolites marked with stars clustered with the variation profile different than the other metabolites. Parental accessions highlighted with stars cluster together with relatively high concentration for most metabolites. The colours spectrum depicts intensity varying from blue (low) to red (high).

4.4.6. Multivariate analysis of metabolite levels to determine variation

Comprehensive Principal component analysis (PCA) was carried out for metabolite levels of all plant samples to determine variable metabolites. This technique allows to uncover complex variation pattern of metabolite profiles. Interestingly ∼82.28% variance could be visualised along first component and ∼12% variance was determined along second component (Figure 31).

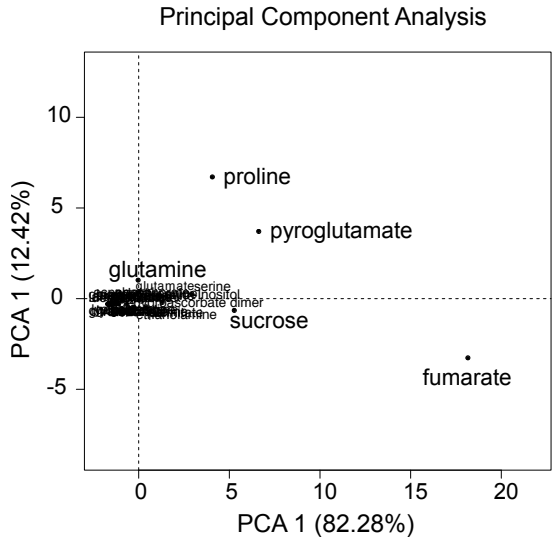


Figure 31. Principle component analysis (PCA) plot of metabolite levels in *Arabidopsis thaliana* accessions and hybrids.

The degree of separation between metabolites along X- and Y-axis allowed to decipher most variable metabolites. Majority of metabolites were found to group closely near the central axis depicting the least variation. Interestingly, proline, pyroglutamate, and fumarate were found to exhibits high degree of variation and separated along the first and second components (Figure 31).

4.4.7. Proline concentration

Proline levels were found to be extremely variable with quantifiable abundance. Therefore, relative levels of proline concentration was compared across all samples. proline was found to be most abundant in Bur-0 and Knox-10 with high degree of variation between respective replicates. L*er* and other parental accessions and hybrids show negligible variation and very low concentration of proline. Hybrids. Most hybrids show pattern similar to the respective parental accession except Knox-10 and Bur-0. However, proline content in Bur-0 x L*er*, L*er* x Knox-10, L*er* x Bur-0, Knox-10 x L*er* show a pattern similar to mean parent value. This data depicts similar pattern of concentration and variability of proline levels in hybrids compared to the parentals Bur-0 and Knox-10. Since, the value in hybrids is similar to mean value of the higher parent, it indicates an

additive heterosis effect (Figure 32).

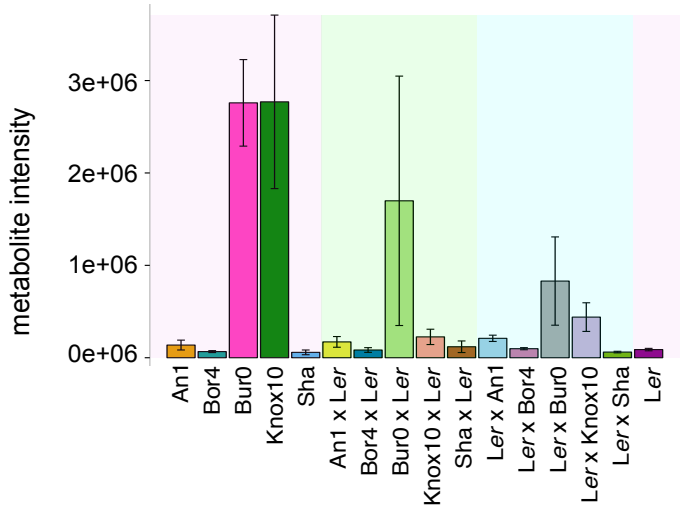


Figure 32. Comparison of proline levels across *Arabidopsis thaliana* accessions and reciprocal hybrids. The barplot depicts the relative proline levels across all plant samples from parental accessions and reciprocal hybrids.

4.4.8. Strong correlation of expression of P5C1 proline pathway gene and proline accumulation/concentration levels

Proline was found to be most abundant in Bur-0 and Knox-10 with high degree of variation between respective replicates. Ler and other parental accessions were found to have negligible variation and relatively lower concentrations of proline. Amongst the profiled metabolites proline levels showed highest degree of variation across the plant accessions and all reciprocal hybrid samples. Therefore, I selected the genes involved in proline biosynthesis pathway to determine any association of transcript and metabolite levels. Out of the three genes of proline biosynthesis pathway AT2G39800, AT3G55610, and AT5G14800 (P5C1), expression levels of P5C1 showed strong correlation with relative changes in proline levels across the profiled samples (Figure 33).

Proline levels can go as high as upto ∼20% of the amino acid pool post salt stress conditions in *Arabidopsis thaliana*. Proline accumulation in plants is carried out via biosynthesis from glutamate. The last step of the proline biosynthetic pathway is crucial and is catalyzed by pyrroline-5-carboxylate (P5C) reductase gene (AT-P5C1). As reported by Giberti *et al.* (2014) the P5C reductase mRNA level rise significantly to almost 5-fold post salt stress. We also see a very strong correlation of proline levels with At-P5C1 levels across the samples, which indicates feedback like mechanism of metabolite levels on transcript levels.

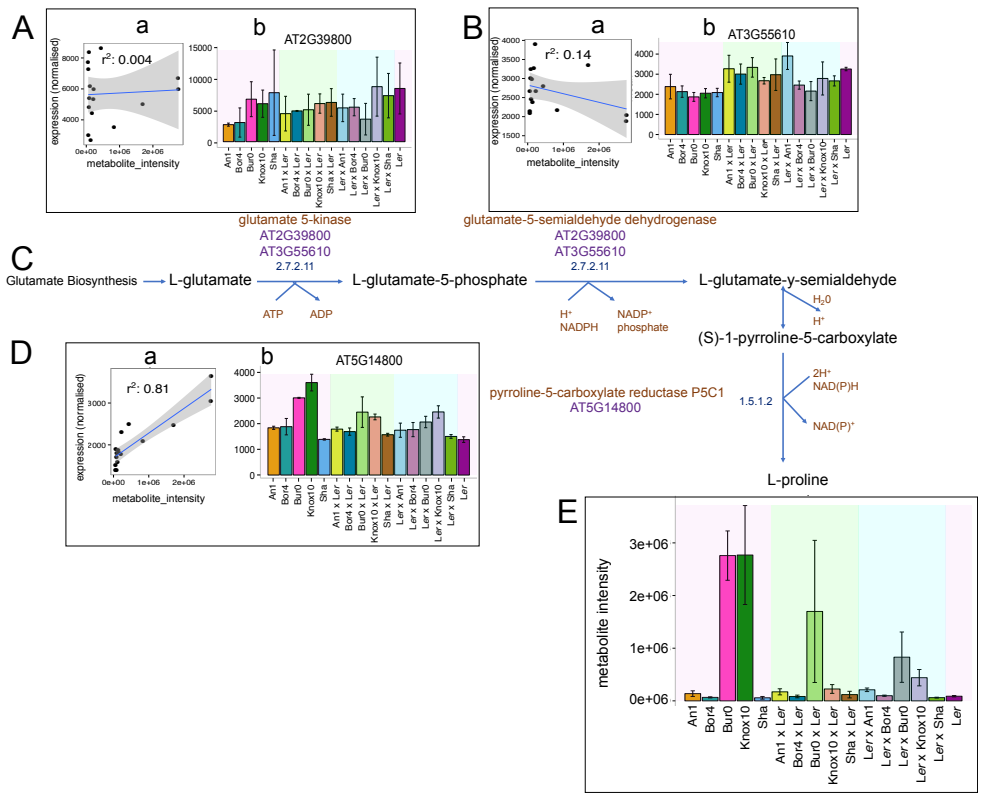


Figure 33. Schematic representation of proline biosynthesis pathway. (A) a. Correlation plot of AT2G39800 expression and proline concentration. b. Bar plot showing expression values of AT2G39800 gene across all accessions and hybrids. (B) a. Correlation plot of AT3G55610 expression and proline concentration. b. Bar plot depicting relative expression values of AT3G55610 gene in the plant samples. (C) Proline Biosynthesis Pathway (KEGG). (D) a. Correlation plot of AT514800 expression and proline concentration. b. Bar plot with expression values of AT514800 gene in the accessions and hybrids. (E) Relative concentration of proline in the samples from the accessions and hybrids.

5. CONCLUSIONS

This research work was undertaken with the major objective of dissecting the underlying mechanisms of natural variation in gene expression by genome-wide allele-specific expression (ASE) analysis. However, there are certain crucial challenges associated with the expression quantification which can introduce varying degrees of bias in estimation. Therefore, the foremost and crucial objective of this work was to define an optimal workflow with high accuracy.

Several key challenges associated with RNA-Seq data analysis have been addressed ingreat detail in several publications. However, this is the first reported study to present a comprehensive evidence based assessment of the challenges towards quantification of allelic expression. To achieve this, exhaustive comparison were made between real and simulated expression profiles to assess the accuracy and effect of the factors on estimation. Rigorous testing and benchmarking was carried out to set thresholds for the ASE protocol. The methodology was further used for ASE analysis of our plant samples. As an outcome we obtained transcript abundance for each gene in parental accessions, hybrids, and allelic measures within hybrids. Hence, we could move onto subsequent objectives, with a detailed strategy devised for each task.

Firstly, allelic and parental gene expression ratio was used to distinguish and quantify the relative contributions of cis-, trans-, cis-x-trans, compensatory effects. The comparison of cis- and trans-effects across multiple hybrids revealed the relative abundance, and accession-specificity regulatory variants. Next, to determine heterosis effects, quantify relative contribution of cis- and trans- effects towards the inheritance patterns of expression divergence, parental and hybrid gene expression levels were compared and examined for cis- and trans- effects. For the next downstream objective to distinguish genetic allelic bias from parent-of-origin effects, we used a simplistic approach to compare allelic ratio in reciprocal hybrid pairs. The allelic ratio for genes under genetic ASE did not change and showed similar magnitude and direction of bias. While the allelic ratio for imprinted genes showed bias in favour of maternal or paternal parent. The results were promising and remains to be experimentally validated and affirmed with the published dataset. The final and substantial objective of this project was to determine phenotypic effect of expression variation on primary metabolite levels by using an integrated analysis metabolites and transcriptomic profiles across the plant samples. Hence, the first task was to perform a comprehensive analysis of metabolite data in order to quantify relative concentration levels in each sample, and determine most variable and abundant metabolites. Thereafter, the most abundant metabolites were examined for correlation with the transcriptomic profiles of their pathway specific genes.

To summarise, we report an evidence based methodological approach to address the major caveats and challenges in the allele expression analysis from RNA-Seq in intra-specific accessions in this study. In addition, we report our key observations from in-depth analysis of expression and metabolite data. Firstly allelic analysis revealed that cis-effects were relatively more abundant, and prominently exhibit large expression differences and similar degree of expression change in multiple hybrid samples. Comparison of allelic ratio of reciprocal hybrids revealed that majority of cis-effects were genetically regulated and free from parental bias. While in most of the imprinted genes, maternal alleles were preferentially expressed. We also found overdominance and underdominance to be rare heterotic events regulated majorly by trans-effects. Proline levels were found to be most variable and highly abundant. Proline biosynthesis gene P5C1, Pyrroline-5-carboxylate (P5C) reductase, with instrumental role in proline synthesis exhibit significantly high correlation of transcript and proline levels in agreement to published reports.

6. KEY LIMITATIONS

Expression data from triplicates of 16 plant samples provided reliable depth and strong advantage for extensive testing and multi-directional comparisons to develop optimal ASE analysis methodology. Hybrid expression data set was an added advantage in study of heterosis.

However, subjecting the data to analysis of imprinting effects highlighted a key limitation of the project. It is crucial to list the observation here with an objective to consider it as a good lesson in project planning in future endeavours. It is to be noted that we did determine a few imprinted alleles. However, the project is limited in its capacity to distinguish the absolute imprinting effects with complete silencing of either allele. Also it is crucial to understand that the phenomena occurs in the very initial developmental stages and in a tissue dependent manner. Observed imprinting effects in other than embryo and endosperm tissues may be noise, non-functional, or effects with biological significance which need experimental validation.

The final and major setback was during the exhaustive analysis of variable metabolite and transcriptomic profiles to determine functional correlated networks. It was found that different classes of metabolites exhibiting varying number of strongly correlated genes. However, upon close examination of expression profiles of pathway specific genes for each metabolite, it was found that most showed very low or non-significant correlation. Of the three genes in proline biosynthesis pathway only AT- P5C1 showed strong significant correlation with metabolite levels of proline. Yet, it is important to list these finding to emphasise the understanding for future projects. Single significant outcome from the comprehensive analysis highlighted the core limitation of our approach for this objective. It is important to note here that both the transcriptomic and metabolite datasets have been obtained for a single time point and same condition. It heavily restricts our capacity to examine the differential data profiles across multiple conditions, which is the key aspect of most integrated network analysis projects.

Finally, to mention a major caveat of gene expression studies in general. It is well understood that the transcriptomics levels may not produce proteins in equivalent quantity. Rather the post-transcriptional and post-translational processes may result in the protein profiles different than expected from transcript abundance.

The downstream effects on phenotypes are even more complex. The multi-level complications pose an inevitable challenge in understanding and drawing conclusions from the observed expression changes and phenotypic effects.

7. FUTURE PROSPECTS

The thesis presents an in-depth discussion of the project aims, limitations, challenges addressed, implemented methodology, and observed results. However, due to limitations of data and strategy, specific preliminary observations remain unverified in-vitro. Notably, the association of expression variation by cis- and trans-effects with the causal genomic variants will be significant in molecular plant breeding strategies. It will also be instrumental in combining ChIP-Seq and DNase-seq data to identify the associated variants improving the understanding of natural variation. We used our method for intra-specific comparisons with a relatively lower degree of variation in the present study. It, however, remains to be tested in inter-specific analysis. It will be of vital interest to perform a network analysis with added datasets from varying conditions. The outcome could reveal significant and functional biological connections.

REFERENCES

References

Adams K.L. Evolution of Duplicate Gene Expression in Polyploid and Hybrid Plants. *Journal of Heredity*, 98(2):136–141, 2007. doi:10.1093/jhered/esl061.

Adams K.L. and Wendel J.F. Allele-specific, bidirectional silencing of an alcohol dehydrogenase gene in different organs of interspecific diploid cotton hybrids. *Genetics*, 171(4):2139–2142, 2005.

Ameur A., Zaghlool A., Halvardson J., Wetterbom A., Gyllensten U., Cavelier L. and Feuk L. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature structural & molecular biology*, 18(12):1435–1440, 2011.

Bao Y., Hu G., Grover C.E., Conover J., Yuan D. and Wendel J.F. Unraveling cis and trans regulatory evolution during cotton domestication. *Nature communications*, 10(1):5399, 2019. doi:10.1038/s41467-019-13386-w.

Chen C. and Begcy K. Genome-Wide Identification of Allele-Specific Gene Expression in a Parent-of-Origin Specific Manner. *Methods in molecular biology (Clifton, N.J.)*, 2072:129–139, 2020. doi:10.1007/978-1-4939-9865-4_11.

Cowles C.R., Hirschhorn J.N., Altshuler D. and Lander E.S. Detection of regulatory variation in mouse genes. *Nature genetics*, 32(3):432–437, 2002.

Degner J.F., Marioni J.C., Pai A.A., Pickrell J.K., Nkadori E., Gilad Y. and Pritchard J.K. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, 2009.

DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011.

Doss S., Schadt E.E., Drake T.A. and Lusis A.J. Cis-acting expression quantitative trait loci in mice. *Genome research*, 15(5):681–691, 2005.

Fernie A.R., Aharoni A., Willmitzer L., Stitt M., Tohge T., Kopka J., Carroll A.J., Saito K., Fraser P.D. and DeLuca V. Recommendations for reporting metabolite data. *The Plant Cell*, 23(7):2477–2482, 2011.

Fort A., Tuteja R., Braud M., McKeown P.C. and Spillane C. Parental-genome dosage

effects on the transcriptome of F1 hybrid triploid embryos of Arabidopsis thaliana. *The Plant journal : for cell and molecular biology*, 92(6):1044–1058, 2017. doi:10.1111/tpj.13740.

Gehring M., Huh J.H., Hsieh T.F., Penterman J., Choi Y., Harada J.J., Goldberg R.B. and Fischer R.L. DEMETER DNA Glycosylase Establishes MEDEA Polycomb Gene Self-Imprinting by Allele-Specific Demethylation. *cell*, 124(3):495–506, 2006.

Gehring M., Missirian V. and Henikoff S. Genomic analysis of parent-of-origin allelic expression in Arabidopsis thaliana seeds. *PLoS One*, 6(8):e23687, 2011.

Giberti S., Funck D. and Forlani G. 1-Pyrroline-5-carboxylate reductase from Arabidopsis thaliana: stimulation or inhibition by chloride ions and feedback regulation by proline depend on whether NADPH or NADH acts as co-substrate. *The New phytologist*, 202(3):911–9, 2014. doi:10.1111/nph.12701.

Gibon Y., Usadel B., Blaesing O.E., Kamlage B., Hoehne M., Trethewey R. and Stitt M. Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. *Genome biology*, 7(8):R76, 2006. doi:10.1186/gb-2006-7-8-R76.

Griebel T., Zacher B., Ribeca P., Raineri E., Lacroix V., Guigó R. and Sammeth M. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083, 2012.

Griffith M., Griffith O.L., Mwenifumbo J., Goya R., Morrissy A.S., Morin R.D., Corbett R., Tang M.J., Hou Y.C., Pugh T.J. *et al.* Alternative expression analysis by RNA sequencing. *Nature methods*, 7(10):843–847, 2010.

Guo M., Rupe M.A., Yang X., Crasta O., Zinselmeier C., Smith O.S. and Bowen B. Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theoretical and Applied Genetics*, 113(5):831–845, 2006a.

Guo M., Rupe M.A., Yang X., Crasta O., Zinselmeier C., Smith O.S. and Bowen B. Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 113(5):831–45, 2006b. doi:10.1007/s00122-006-0335-x.

Guo M., Rupe M.A., Zinselmeier C., Habben J., Bowen B.A. and Smith O.S. Allelic variation of gene expression in maize hybrids. *The Plant Cell Online*, 16(7):1707–1716, 2004.

Guo M., Yang S., Rupe M., Hu B., Bickel D.R., Arthur L. and Smith O. Genome-

wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSS) reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. *Plant Mol Biol*, 66(5):551–563, 2008. doi:10.1007/s11103-008-9290-z.

Jansen R.C. and Nap J.P. Genetical genomics: the added value from segregation. *TRENDS in Genetics*, 17(7):388–391, 2001.

Kanehisa M. and Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

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. and Jansen R.C. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, 104(5):1708–13, 2007. doi:10.1073/pnas.0610429104.

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R. and Salzberg S.L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36, 2013.

Kliebenstein D.J., West M.A.L., van Leeuwen H., Loudet O., Doerge R.W. and St Clair D.A. Identification of QTLs controlling gene expression networks defined a priori. *BMC Bioinformatics*, 7:308, 2006. doi:10.1186/1471-2105-7-308.

Kopka J., Schauer N., Krueger S., Birkemeyer C., Usadel B., Bergmüller E., Dörmann P., Weckwerth W., Gibon Y., Stitt M. *et al.* GMD@ CSB. DB: the Golm metabolome database. *Bioinformatics*, 21(8):1635–1638, 2005.

Korir P.K. and Seoighe C. Inference of allele-specific expression from RNA-seq data. *Methods Mol Biol*, 1112:49–69, 2014. doi:10.1007/978-1-62703-773-0_4.

Langmead B. and Salzberg S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

Lemmon Z.H., Bukowski R., Sun Q. and Doebley J.F. The role of cis regulatory evolution in maize domestication. *PLoS genetics*, 10(11):e1004745, 2014. doi: 10.1371/journal.pgen.1004745.

Lisec J., Schauer N., Kopka J., Willmitzer L. and Fernie A.R. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat Protoc*, 1(1):387–396, 2006. doi:10.1038/nprot.2006.59.

Lo H.S., Wang Z., Hu Y., Yang H.H., Gere S., Buetow K.H. and Lee M.P. Allelic

variation in gene expression is common in the human genome. *Genome research*, 13(8):1855–1862, 2003.

Luedemann A., Strassburg K., Erban A. and Kopka J. TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737, 2008.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. and DePristo M.A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, 2010. doi:10.1101/gr.107524.110.

McManus C.J., Coolon J.D., Duff M.O., Eipper-Mains J., Graveley B.R. and Wittkopp P.J. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome research*, 20(6):816–825, 2010.

Mortazavi A., Williams B.A., McCue K., Schaeffer L. and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

Mueller L.A., Zhang P. and Rhee S.Y. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology*, 132(2):453–460, 2003.

Oshlack A., Wakefield M.J. *et al.* Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, 4(1):14, 2009.

Pandey R.V., Franssen S.U., Futschik A. and Schlötterer C. Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Molecular ecology resources*, 13(4):740–745, 2013.

Pignatta D., Erdmann R.M., Scheer E., Picard C.L., Bell G.W. and Gehring M. Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *eLife*, 3:e03198, 2014. doi:10.7554/eLife.03198.

Potokina E., Druka A., Luo Z., Wise R., Waugh R. and Kearsey M. Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *The Plant Journal*, 53(1):90–101, 2008.

Rapaport F., Khanin R., Liang Y., Pirun M., Krek A., Zumbo P., Mason C.E., Socci N.D. and Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, 14(9):R95, 2013.

Robinson M.D., McCarthy D.J. and Smyth G.K. edgeR: a Bioconductor package

for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Rozowsky J., Abyzov A., Wang J., Alves P., Raha D., Harmanci A., Leng J., Bjornson R., Kong Y., Kitabayashi N. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1), 2011.

Satya R.V., Zavaljevski N. and Reifman J. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic acids research*, page gks425, 2012.

Schauer N., Steinhauser D., Strelkov S., Schomburg D., Allison G., Moritz T., Lundgren K., Roessner-Tunali U., Forbes M.G., Willmitzer L., Fernie A.R. and Kopka J. GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett*, 579(6):1332–1337, 2005. doi:10.1016/j.febslet.2005.01.029.

Shao L., Xing F., Xu C., Zhang Q., Che J., Wang X., Song J., Li X., Xiao J., Chen L.L., Ouyang Y. and Zhang Q. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *PNAS; Proceedings of the National Academy of Sciences*, 116(12):5653–5658, 2019.

Shen S.Q. Hybrid Mice Reveal Parent-of-Origin and Cis- and Trans-Regulatory Effects in the Retina. *PLoS One*, 2014.

Shi C., Uzarowska A., Ouzunova M., Landbeck M., Wenzel G. and Lübberstedt T. Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. *BMC Genomics*, 8:22, 2007. doi:10.1186/1471-2164-8-22.

Springer N.M. and Stupar R.M. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *The Plant cell*, 19(8):2391–402, 2007a. doi:10.1105/tpc.107.052258.

Springer N.M. and Stupar R.M. Allelic variation and heterosis in maize: how do two halves make more than a whole? *Genome research*, 17(3):264–275, 2007b.

Street N.R. The genetics and genomics of the drought response in Populus. *The Plant Journal*, 2006.

Stupar R.M., Hermanson P.J. and Springer N.M. Nonadditive expression and parent-of-origin effects identified by microarray and allele-specific expression profiling of maize endosperm. *Plant physiology*, 145(2):411–425, 2007.

Tabano S., Bonaparte E. and Miozzo M. Detection of Loss of Imprinting by Pyrose-

quencing(®). *Methods in molecular biology (Clifton, N.J.)*, 1315:241–58, 2015. doi: 10.1007/978-1-4939-2715-9_18.

Thimm O., Bläsing O., Gibon Y., Nagel A., Meyer S., Krüger P., Selbig J., Müller L.A., Rhee S.Y. and Stitt M. mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6):914–939, 2004.

Tohge T. Metabolite profiling in plant biology: platforms and destinations. *Genome Biology*, 2009. doi:10.1186/gb-2004-5-6-109.

Trapnell C., Pachter L. and Salzberg S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D.R., Pimentel H., Salzberg S.L., Rinn J.L. and Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–578, 2012.

Tuteja R., McKeown P.C., Ryan P., Morgan C.C., Donoghue M.T.A., Downing T., O'Connell M.J. and Spillane C. Paternally Expressed Imprinted Genes under Positive Darwinian Selection in Arabidopsis thaliana. *Molecular biology and evolution*, 36(6):1239–1253, 2019. doi:10.1093/molbev/msz063.

Tycko B. Allele-specific DNA methylation: beyond imprinting. *Human molecular genetics*, page ddq376, 2010.

Van Norman J.M. and Benfey P.N. Arabidopsis thaliana as a model organism in systems biology. *Wiley interdisciplinary reviews. Systems biology and medicine*, 1(3):372—379, 2009. ISSN 1939-5094. doi:10.1002/wsbm.25.

von Korff M., Radovic S., Choumane W., Stamati K., Udupa S.M., Grando S., Ceccarelli S., Mackay I., Powell W., Baum M. and Morgante M. Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *The Plant journal : for cell and molecular biology*, 59(1):14–26, 2009. doi: 10.1111/j.1365-313X.2009.03848.x.

West M.A., Kim K., Kliebenstein D.J., van Leeuwen H., Michelmore R.W., Doerge R. and Clair D.A.S. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics*, 175(3):1441–1450, 2007.

Wittkopp P.J., Haerum B.K. and Clark A.G. Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88, 2004.

Wittkopp P.J., Haerum B.K. and Clark A.G. Regulatory changes underlying expression differences within and between Drosophila species. *Nature genetics*, 40(3):346–50, 2008. doi:10.1038/ng.77.

Xu W., Dai M., Li F. and Liu A. Genomic imprinting, methylation and parent-of-origin effects in reciprocal hybrid endosperm of castor bean. *Nucleic acids research*, page gku375, 2014.

Yan H., Yuan W., Velculescu V.E., Vogelstein B. and Kinzler K.W. Allelic variation in human gene expression. *Science*, 297(5584):1143–1143, 2002.

Zhang X. and Borevitz J.O. Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics*, 182(4):943–954, 2009.

LIST OF TABLES

1 Comparison of expression divergence corresponding exons and SNPs to select genes for ASE analysis 28
2 Classification of cis- and trans- regulatory divergence patterns 28
3 Comparison of Col-0 and Psuedo-reference genome for determination of expression divergence 41
S.1 MapMan annotations for cis- genes in An-1 x L*er* and L*er* x An-1. 87
S.2 MapMan annotations for cis- genes of Bor-4 x L*er* and L*er* x Bor-4. . . . 87
S.3 MapMan annotations for cis- genes in Bur-0 x L*er* and L*er* x Bur-0 87
S.4 MapMan annotations for cis- genes in Knox-10 x L*er* and L*er* x Knox-10. 88
S.5 MapMan annotations for cis- genes in Sha x L*er* and L*er* x Sha. 88
S.6 MapMan annotations for trans- genes in An-1 x L*er* and L*er* x An-1. . . . 88
S.7 MapMan annotations for trans- genes in Bor-4 x L*er* and L*er* x Bor-4. . . 88
S.8 MapMan annotations for trans- genes in Bur-0 x L*er* and L*er* x Bur-0. . . 88
S.9 MapMan annotations for trans- genes in Knox-10 x L*er* and L*er* x Knox-10 88
S.10 MapMan annotations for trans- genes in Sha x L*er* and L*er* x Sha. 89

LIST OF FIGURES

1 Schematic representation of cis- and trans-regulated expression differences 5
2 Allele-specific expression analysis for Cvi, L*er* and the F1 hybrid - Cvi x L*er* . . . 6
3 Categorisation of aligned reads . . . 9
4 Position-dependent read distribution pattern over SNPs . . . 9
5 The geographical distribution of *Arabidopsis thaliana* . . . 11
6 Schematic representation of parent-of-origin effect on allelic expression . 13
7 Schematic representation of gene expression inheritance patterns categories 15
8 Schema for construction of parent-specific pseudo-reference genome . . . 21
9 Overview of RNA-Seq sequenced from *Arabidopsis thaliana* accessions and reciprocal hybrids . . . 23
10 Sumarised workflow of method development schema for ASE analysis . . 32
11 Mapping profile of Cvi and L*er* reads . . . 33
12 Comparison of polymorphism profiles in Cvi and L*er* against Col-0 . . . 34
13 Abundance and distribution profile of SNPs across the genome of Col-0 against Cvi and L*er* . . . 36
14 Availability of genes for allelic expression analysis . . . 37
15 Parent-specific pseudo-reference genomes improves read mappability . . 39
16 Degree of deviation from expected gene expression profile . . . 40
17 Evaluation of sensitivity of genome of reference to detect expression divergence . . . 41
18 Iterative process of generating and improving parent-specific pseudo-reference (pref) . . . 45
19 Comparison of expression estimation from read overlapping gene and only SNPs . . . 47
20 Expression estimation over SNPs in different gene regions . . . 49
21 Magnitude of expression divergence between An-1 vs L*er* . . . 52
22 Exonic versus SNP estimate of expression changes . . . 54
23 Determination and quantification of cis- and trans-regulatory divergence in *Arabidopsis thaliana* accessions . . . 56
24 Magnitude of expression divergence (log fold change) due to regulatory cis- and trans- regulatory variants . . . 57
25 Conserved cis- and trans-effects across the *Arabidopsis thaliana* hybrids . 58
26 Comparison of Allelic Ratio quantified from RNA-Seq and Pyrosequencing Method . . . 60
27 cis- and trans effects on expression inheritance patterns in hybrids 62

28 Comparison of cis-effects between reciprocal hybrids to determine parental-effect on allelic bias 64
29 Categories of primary metabolites in *Arabidopsis thaliana* leaves 66
30 Hierarchical clustering analysis of primary metabolites from leaves of *Arabidopsis thaliana* 67
31 PCA analysis of primary metabolites 68
32 Comparison of proline levels across *Arabidopsis thaliana* accessions and reciprocal hybrids 69
33 Schematic representation of proline biosynthesis pathway 70

ABBREVIATIONS

ASE	Allele-Specific expression
DE	Differentially Expressed
MEG	Maternally expressed (paternally imprinted) gene
PEG	Paternally expressed (maternally imprinted) gene
GC-MS	Gas Chromatography coupled to Mass Spectrometry
GC/TOF-MS	Time-of-flight mass spectrometry
MS	Mass Spectrometry
HTS	High-Throughput Sequencing
RNA-Seq	RNA Sequencing
SE	Sensitivity
SP	Specificity
TP	True Positive
FP	False Positive
PCA	Principle Component Analysis
TAIR	The Arabidopsis Information Resource
GATK	Genome Analysis Toolkit
GFF	Gene Feature Format
CDS	Coding Sequence
UTR	Untranslated Region
SNP	Single Nucleotide Polymorphism
InDel	Insertion and Deletion
logFC	logarithm of Fold Change in gene expression
rpkm	read per kilobase per million
Grc	Gene read count
SNPrc	SNP read count
Exonrc	Exon read count
FDR	False Discovery Rate
edgeR	Empirical Analysis of Digital Gene Expression Data in R
ARACyc	A Biochemical Pathway Database for Arabidopsis
PMN	Plant Metabolic Network
KEGG	Kyoto Encyclopedia of Genes and Genomes
RIL	Recombinant Inbred Lines

SUPPLEMENTARY TABLES

Table S.1. MapMan annotations for cis- genes of An-1 x L*er* and L*er* x An-1.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
30.2.17	signalling.receptor kinases.DUF 26	11 30 268 17701	2.12E-11	3.82E-09
20.1	stress.biotic	24 317 255 17414	7.33E-10	1.32E-07
20.1.7	stress.biotic.PR-proteins	15 166 264 17565	1.51E-07	2.72E-05
20	stress	26 617 253 17114	8.41E-06	1.51E-03
30.2	signalling.receptor kinases	18 362 261 17369	2.86E-05	5.15E-03

Table S.2. MapMan annotations for cis- genes of Bor-4 x L*er* and L*er* x Bor-4.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
20.1	stress.biotic	21 307 244 16588	4.22E-08	7.60E-06
20	stress	25 598 240 16297	1.33E-05	2.39E-03
20.1.7	stress.biotic.PR-proteins	12 163 253 16732	1.81E-05	3.26E-03
30.2.17	signalling.receptor kinases.DUF 26	6 37 259 16858	4.84E-05	8.71E-03
26.1	misc.cytochrome P450	10 134 255 16761	8.14E-05	1.47E-02
16	secondary metabolism	13 251 252 16644	2.45E-04	4.41E-02
26	misc	29 945 236 15950	6.86E-04	1.23E-01
16.5.1.3	secondary metabolism.sulfur-containing.glucosinolates.degradation	3 10 262 16885	9.29E-04	1.67E-01
3.8.2	minor CHO metabolism.galactose.alpha-galactosidases	2 2 263 16893	1.40E-03	2.51E-01
16.5.1	secondary metabolism.sulfur-containing.glucosinolates	4 34 261 16861	2.72E-03	4.89E-01

Table S.3. MapMan annotations for cis- genes of Bur-0 x L*er* and L*er* x Bur-0.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
30.2.17	signalling.receptor kinases.DUF 26	7 37 260 18348	2.78E-06	4.88E-04
20.1	stress.biotic	16 321 251 18064	3.04E-05	5.35E-03
30.2	signalling.receptor kinases	15 367 252 18018	4.26E-04	7.50E-02
16.5.1	secondary metabolism.sulfur-containing.glucosinolates	5 42 262 18343	5.43E-04	9.56E-02
16.5	secondary metabolism.sulfur-containing	5 47 262 18338	8.69E-04	1.53E-01
20.1.7	stress.biotic.PR-proteins	9 165 258 18220	9.26E-04	1.63E-01
20	stress	20 627 247 17758	1.15E-03	2.03E-01
16.5.1.1.1	secondary metabolism.sulfur-containing.glucosinolates.synthesis.aliphatic	3 15 264 18370	2.02E-03	3.55E-01

Table S.4. MapMan annotations for cis- genes of Knox-10 x L*er* and L*er* x Knox-10.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
20.1	stress.biotic	22 323 242 17112	1.16E-08	2.05E-06
20	stress	25 623 239 16812	1.45E-05	2.56E-03
26.1	misc.cytochrome P450	11 138 253 17297	1.48E-05	2.63E-03
20.1.7	stress.biotic.PR-proteins	12 167 252 17268	1.62E-05	2.86E-03
20.1.7.6.1	stress.biotic.PR-proteins.proteinase inhibitors.trypsin inhibitor	3 4 261 17431	1.10E-04	1.94E-02
27	RNA	12 1986 252 15449	1.71E-04	3.03E-02
30.2	signalling.receptor kinases	16 360 248 17075	1.76E-04	3.12E-02
26	misc	30 952 234 16483	1.92E-04	3.40E-02
20.1.7.6	stress.biotic.PR-proteins.proteinase inhibitors	3 6 261 17429	2.58E-04	4.56E-02

Table S.5. MapMan annotations for cis- genes of Sha x L*er* and L*er* x Sha.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
20.1	stress.biotic	19 338 235 17331	8.68E-07	1.70E-04
20.1.7	stress.biotic.PR-proteins	13 178 241 17491	3.49E-06	6.83E-04

Table S.6. MapMan annotations for trans- genes of An-1 x L*er* and L*er* x An-1.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
1.1.1.1	PS.lightreaction.photosystem II.LHC-II	3 6 129 17872	3.13E-05	4.48E-03

Table S.7. MapMan annotations for trans- genes of Bor-4 x L*er* and L*er* x Bor-4.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
15.1	metal handling.acquisition	3 5 136 17016	2.83E-05	3.96E-03
13.2.2.2	amino acid metabolism.degradation.glutamate family.proline	2 1 137 17020	1.94E-04	2.72E-02

Table S.8. MapMan annotations for trans- genes of Bur-0 x L*er* and L*er* x Bur-0.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
3.2.3	minor CHO metabolism.trehalose.potential TPS/TPP	3 3 285 18361	7.04E-05	1.59E-02
34.3	transport.amino acids	6 44 282 18320	1.16E-04	2.61E-02
16.8	secondary metabolism.flavonoids	7 65 281 18299	1.22E-04	2.75E-02

Table S.9. MapMan annotations for trans- genes of Knox-10 x L*er* and L*er* x Knox-10.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
10.7	cell wall.modification	4 32 170 17493	4.16E-04	6.86E-02

Table S.10. MapMan annotations for trans- genes of Sha x L*er* and L*er* x Sha.

Bin	BinName	Contingency	Pvalue	Adj.Pvalue
27.3.3	RNA.regulation of transcription.AP2/EREBP, APETALA2/Ethylene-responsive element binding protein family	6 62 171 17684	5.62E-05	8.82E-03
16.5	secondary metabolism.sulfur-containing	5 39 172 17707	7.06E-05	1.11E-02
10.7	cell wall.modification	5 39 172 17707	7.06E-05	1.11E-02
16.5.1.1.1.4	secondary metabolism.sulfur-containing.glucosinolates.synthesis.aliphatic. methylthioalkylmalate isomerase small subunit (MAM-IS)	2 0 175 17746	9.70E-05	1.52E-02

ARTICLES PUBLISHED DURING Ph.D.

***Arabidopsis thaliana* natural variation reveals connections between UV radiation stress and plant pathogen-like defense responses**

Authors	Thomas Piofczyk, **Ganga Jeena** and Ales Pecinka
Journal	Plant Physiology and Biochemistry
Date of Publication	February 2015
Personal contribution	Ganga Jeena created accession-specific pseudo-references

ACKNOWLEDGEMENTS

I sincerely thank my supervisor Dr. José M Jiménez-Gómez for the opportunity to work on carrying out this research work under his guidance. His critical feedback and support were instrumental in this project's work to bring the ideas to shape. I am very grateful to our department director Prof. Dr. ir. Maarten Koornneef for his crucial support throughout the project tenure and later for thesis completion.
I am very grateful to Prof. Dr. Korbinian Schneeberger and Prof. Dr. Achim Tresch for agreeing to evaluate this thesis and being a part of my thesis committee. Thanks to Prof. Dr. Schneeberger for his support in bringing the thesis to closure and required German translations in the thesis.
I thank my former supervisor Prof. Dr. Alga Zuccaro agreeing to be the chairperson of my Ph.D. thesis examination committee. I am grateful for her support during the finishing of this dissertation and the opportunity to work and learn in her group. I am grateful to Dr. Gregor Langen for his support and agreeing to be on the thesis committee as Beisitzer.

I wholeheartedly acknowledge the IMPRS, Max Planck Institute for Plant Breeding Research for funding my Ph.D. work, participation in workshops, conferences, and Ph.D. retreats. I want to convey my sincere gratitude to the former IMPRS scientific coordinator Dr. Olof P. Persson, and Graduate School of Biological Sciences Coordinator Dr. Isabell Witt for their crucial role in organising scientific training courses and workshops. I thank the Ms. Birgit Thron, Ms. Priska Dormels, and Mr. Stefan Schuller for their constant help with the administrative process, and SUSAN team for their crucial support with high-performance cluster.

I thank my peers from the Jiménez, Pecinka , and Koornneef group for the stimulating learning environment. Special thanks to Thomas Piofczyk for testing my thesis protocol of pseudo-reference genome construction for his publication. I also thank the colleagues in the lab of Prof. Zuccaro and Prof. Bucher and CEPLAS young scientists for stimulating and engaging scientific discussions, sharing common goals, motivations, bottlenecks, making it a great teamwork experience.
Thanks to my former supervisors at NIPGR, Dr. Debasis Chattopadhyay and Dr. Gitanjali Yadav or providing me the opportunity to work on Chickpea Genomics, wherein I gained the first insights into the challenges of Next Generation Sequencing data analysis for genomics and transcriptomic projects.
I express my sincere gratitude to the Bioinformaticians of our group Arunkumar Srinivasan and Malgorzata Ryngajllo for their ideas and suggestions of the project challenges, presentation preparations, and great deal of support. I greatly appreciate the enthusiasm and patience of Arun during countless stimulating brainstorming discussions, sharing his insights on statistical analysis, assessment and interpretation of results, help with learning nuances of R, advanced Perl and statistical concepts. I am very thankful to Malgorzata for her help with Pyrosequencing validation of RNA-seq results, MapMan analysis of ASE results, assistance with LaTeXdocumentation, uplifting and motivational discussions, swimming lessons and crucial support in the lab. I am very grateful to

my brother Virender Singh Jeena for formulating the mathematical equations to summarise the read-counting method for ASE analysis, and help with compilation of documents. I wholeheartedly thank Tripta Zhang for taking me under her wings right upon the day of my arrival, helping me settle down in the new environment. I am grateful for the kindness and generosity of Selva Kumari Ramasubramanian, Romel Ahmed and his family, Kerstin H Richau, Marcel von Roth, Ahmed Abdelsamad, Mohamed Suliman, Björn Pietzenuk, Geo J Velikakkam and Sushimita, Shaista and Ilyas Mohhamad, Luis and Jamuna, Shuoan and his family, Tiago, and Usman for making me feel at home with their heartwarming hospitality. Thanks to the company of Inga Schmalenbach, Niels Müller and Ute Tartler for making the workplace so warm and welcoming.
Thanks to all my colleagues and friends at MPIPZ and the University of Cologne for their kindness and patience, and most importantly for sharing their precious experiences and critical insights. I could evolve as a person thanks to their invaluable lessons.

I am indebted to Udhaya Ponraj, Vivek Halder, Arunkumar Srinivasan, Satish Kumar Eeda, and Chun Hsin Liu (Phoebe) for their care and unending support throughout, most importantly during excruciating times. Thanks to the extraordinary humor of Udhaya, her care, and patience, I could survive many phases of extreme anxiety and distress. Thanks to Phoebe for building her trust in me over the time, sharing her earnest love for music and swimming, and especially for calling me from across the world to make sure of my well being. I thank Artem Pankin for his critical questions that helped me develop a more rational perspective. Thanks to his inspirational enthusiasm for challenging sports, I could expand the horizon of my limits, learning a great deal from him in the process. I thank Van Anh Dao for her sturdy support, and crucial understanding, inspiring discussions, and countless walks exploring nature in and around Cologne.
I am grateful to and will cherish the valuable friendships of Deepak Bhandari, Aishwarya Ghule, Mahwish and Tajjamul Hussain, Vimal Rawat, Meenakshi Barua, Vandana, Arunima Shilpi, Vinay Dahiya and Jay, Kumari Billakurthi, Jatish Ponnu and Shalima, Kamlesh Sahu, Swati Puranik and Pranav Sahu, Kashif Nawaj, Shadab Azam, Debika Sarkar, Hanna Rovenich, Alan Wanke, Jayshree Pandey, and Sobia and Sandal.

I want to express heartfelt sense of gratitude towards my insightful teachers. Their trust in me was my safe ground and instilled a deep sense of security and inner confidence, propelling me further to the best of my abilities. Special thanks to my science teachers for feeding my unending curiosity and igniting the love for science. I forever cherish my literature teachers for their recognition of my enthusiasm for creative words and their continuous efforts towards improving my capabilities.

Thanks to my grandparents for their warmth and affection and my family back in mountains for living healthy and happy. Thanks to my younger siblings for growing up into beautiful responsible people, taking charge of the team and keeping our parents entertained in my absence. Most special thanks to my parents for their unending support throughout. I sincerely value and appreciate their struggles and dedication.

DEDICATION

I sincerely dedicate this dissertation to my parents Ms. Deepa and Mr. Harish Singh Jeena for their relentless efforts and countless sacrifices.

For standing strong and almost always alone against the gender biased social norms, in order to continuously support my education, unfazed by their limited means.

For their unbreakable resilience in the face of extreme hardships, which made it possible for me to explore wider horizons in pursuit of knowledge.

For their farsightedness and determination, an invaluable lesson for me to work hard towards continuously evolving into a person better and beyond my limits.

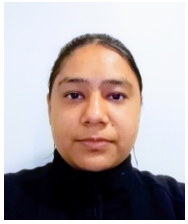
ERKLÄRUNG

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht wor – den ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Dr. José M Jiménez-Gómez betreut worden.

I declare that I have independently completed the dissertation I submitted, that the sources and tools used are completely cited and that parts of the dissertation - including tables, maps and figures -, taken from other sources (in the wording or the sense) in each individual case has been referred to as such. Further, I declare that this dissertation has not been submitted to any other Faculty or University and that - apart from the following partial publications - has not yet been published, and that I will not publish the dissertation before the end of the doctoral examination. I am aware of the requirements of the doctoral regulations. The doctoral project and Dissertation has been supervised by Dr. José M Jiménez-Gómez.

Köln, August 3, 2021 Ganga Jeena

CURRICULUM VITAE



Name	**Ganga Jeena**
Contact	+49-151-75622924
Email	ganga.jeena@gmail.com
Address	Weyertal 24, 50937, Cologne, Germany
Nationality	Indian
Date of Birth	10.12.1985
Place of Birth	Gahana, Nainital

Education Profile

2007 – 2009	M.Sc. Bioinformatics, Banasthali University, Rajasthan, India
2004 – 2007	B.Sc. Environmental Sciences, Maitreyi College, Delhi University, New Delhi, India

Research Profile

2018-2019	**Guest Scientist,** Zuccaro Lab
Institute	Biocenter, University of Cologne, Germany
2015-2017	**CEPLAS Bioinformatics Scientist,** Zuccaro Lab
Institute	Biocenter, University of Cologne, Germany
Supervisors	**Dr. Gregor Langen** and Prof. Dr. Alga Zuccaro
Projects	I. Genome assembly from PacBio sequences, with annotation, and analysis of fungal root endophyte *Rhynchosporium* strain 229 II. Transcriptomic analysis of bipartite interaction of *Arabis alpina* and*Rhynchosporium* strain 229
Collaborators	Dr. Juliana Almario, Prof. Dr. Marcel Bucher
Projects	III. Comparative transcriptomic analysis of tripartite interaction between *Hordeum vulgaris*, *Serendipita vermifera*, *Bipolaris sativa*
Collaborators	Dr. Debika Sarkar
Projects	IV. Genome assembly, annotation and comparative analysis of endophytic fungal species, *Serendipita indica*, *Serendipita herbamans*, *Chaetospermum artocarpi*, and *Chaetospermum camellieae* V. Transcriptomic assembly and analysis of fungal endophytes in root pelotons of *Neottia nidus*, *Chaetospermum artocarpi*, and *Chaetospermum camellieae*
Responsibilities	• Data management, quality check, processing, analysis, visualisation, communication, and coordinating submission to NCBI • Programming, method development and software testing and benchmarking • Scientific communication, discussions, presentations and co-authoring manuscripts

2011-2015	**IMPRS PhD Fellow**, Jiménez-Gómez lab
Institute	Max Planck Institute for Plant Breeding and Research, Cologne, Germany
Supervisors	**Dr José M Jiménez-Gómez,** Prof. Dr. Korbinian Schneeberger and Prof. Dr. Maarten Koornneef
Projects	I. Development of in-silico workflow for genome wide Allele-Specific expression analysis in *Arabidopsis thaliana* accessions II. Determination of heterotic and imprinting effects III. Integrated analysis of transcriptomic and metabolite profiles
Responsibilities	• Problem solving by computational approach, data management, processing, analysis, visualisation, communication of results • Scientific discussions, preparation of reports, oral and poster presentations
2010-2011	**Junior Research Fellow,** Debasis-Chattopadhyay lab
Institute	National Institute of Plant Genome Research, New Delhi, India
Supervisors	**Dr. Debasis Chattopadhyay**
Projects	I. Chickpea genome sequence assembly, annotation and analysis II. Genomic and transcriptomic SNP prediction of wild and cultivated Chickpea species
Collaborators	Dr. Sabhyata Bhatia, Dr. Mukesh Jain
Responsibilities	Data management, processing, analysis and communication of results
2009	**Master Dissertation II**
Institute	National Institute of Plant Genome Research, New Delhi, India
Supervisors	**Dr. Gitanjali Yadav**
Project	Analysis of essential oil composition of dicotyledonous plant of Apiaceae and Lamiaceae families.
2008	**Master Dissertation I**
Institute	Banasthali University, Tonk, Rajasthan, India
Supervisors	**Dr Pramod Katara**
Project	In-silico prediction of the regulatory element patterns of Human pathogen *Mycobacterium tuberculosis* H37Rv and *Pseudomonas syringae* tomato DC3000

Teaching Experience

2017	Linux and Perl programming workshop for PhD students (2 days)
2016, 2017	Bioinformatics programming and NGS data analysis practical course module for Master (M.Sc) students (3 days)
Responsibilities	Co- planning and preparation of course material, lecturer and Practical trainer

Fellowships

2011 International Max Planck Research Scholarship (**IMPRS**) for PhD,
Max Planck Institute for Plant Breeding Research, Cologne, Germany

2009 Department of Biotechnology (**DBT**) Project fellowship for Master thesis,
National Institute for Plant Genome Research, Delhi, India

Bioinformatics and Scientific Skills

- Genome and transcriptome assembly and annotation, comparative and functional analysis
- Next generation sequencing data from PacBio, Illumina, 454, processing, analysis and visualisation
- High performance cluster computing, Perl, R, Shell Scripting
- Linux (Ubuntu, CentOS, Fedora, Mint), MacOS, Windows system experience
- Scientific communication using written, oral and graphic medium

Scientific Training

2016 Hands-on session in R for RNA-Seq analysis,
Heinrich-Heine-Universität Düsseldorf, Germany

2015 Good Scientific Practices, Heinrich-Heine-Universität Düsseldorf, Germany

2014 Conflict Management Course; Scientific Presentation; Self and Management Skills
MPIPZ and University of Cologne, Germany

2013 Getting Funded Course; Data Visualisation Workshop; Project Management
MPIPZ and University of Cologne, Germany

2012 Plant Module Course; Scientific Writing Course; R programming course
MPIPZ and University of Cologne, Germany

2011 Advances and Challenges in Next Generation Sequencing & Bioinformatics of Genome Analysis, Delhi University, New Delhi, India

Conferences attended

2018 XXIVth Minisymposium on Plant Biology, University of Cologne

2018 1st Cologne Conference on Food for Future, Cologne, Germany

2017 2nd International BioSC Symposium -Towards an Integrated Bioeconomy, Germany

2017 2nd Cologne Excellent Women in Science Symposium,
University of Cologne, Germany

2015 Botanical Institute Research Colloquium, University of Cologne, Germany

2010 Indo-Swiss Bioinfomatics Symposium, Indian Institute of Technology (IIT-D),
New Delhi, India

2008 International Conference of Microbiology and Biotechnology,
Banasthali University, Rajasthan, India

Oral Presentations

2016	MPItM-ABRE Workshop, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany
2015, 16	CEPLAS Annual Symposium and Youth Research Retreats
2014	6th European Plant Science Retreat, Amsterdam, Netherlands
2012,13,14	IMPRS Annual PhD Retreats

Poster presentations

2017	12th conference of the VAAM special group Molecular Biology of Fungi, Jena, Germany
2016, 17	CEPLAS Annual Symposium, Germany
2014	EMBO Interdisciplinary Plant Development Conference, Sainsbury Laboratory, Cambridge, UK
2012, 13	Science Day, Max Planck Institute of Plant Breeding and Research, Cologne, Germany
2012	Natural Variation of Plants, PhD Summer School, Wageningen University, Netherlands

Event Planning and Coordination

2014	Career Day MPIPZ, Cologne, Germany (co-organiser)

Language Proficiency

Hindi, English, Kumauni	Fluent (Reading, Writing, Listening, Spoken)
German	Basic (Reading, Writing, Listening, Spoken)
Punjabi, Haryanvi, Urdu	Fluent (Listening) Basic (Spoken), good understanding

Creative Interests

- Reading
- Listening to soft, classical and folk music
- Hiking in nature and exploring wilderness, Yoga and endurance sports

Strengths

- Patience, thoughtfulness, and a high degree of empathy and compassion
- Attention to details and methodological approach to work

Cited Publications

2019 D. Sarkar*, H. Rovenich*, **G. Jeena***, S. Nizam, A. Tissier, G. U. Balcke, L. K. Mahdi, M. Bonkowski, G. Langen, and A. Zuccaro**.** The inconspicuous gatekeeper: endophytic serendipita vermifera acts as extended plant protection barrier in the rhizosphere. **New Phytologist,** 224(2):886–901.

2017 J. Almario, **G. Jeena**, J. Wunder, G. Langen, A. Zuccaro, G. Coupland, and M. Bucher. Root-associated fungal microbiota of nonmycorrhizal Arabis alpina and its contribution to plant phosphorus nutrition. **Proceedings of the National Academy of Sciences (PNAS),** 114(44):E9403– E9412.

2015 T. Piofczyk, **G. Jeena,** and A. Pecinka. Arabidopsis thaliana natural variation reveals connections between uv radiation stress and plant pathogen-like defense responses. **Plant Physiology and Biochemistry**, 93:34–43.

2015 R. Gaur, **G. Jeena**, N. Shah, S. Gupta, S. Pradhan, A. K. Tyagi, M. Jain, D. Chattopadhyay, and S. Bhatia. High density linkage mapping of genomic and transcriptomic snps for synteny analysis and anchoring the genome sequence of chickpea. **Scientific reports**, 5:13387.

2014 S. Kumari, S. Pundhir, P. Priya, **G. Jeena**, A. Punetha, K. Chawla, Z. Firdos Jafaree, S. Mondal, and G. Yadav. Essoildb: a database of essential oils reflecting terpene composition and variability in the plant kingdom. **Database.**

2013 M. Jain, G. Misra, R. K. Patel, P. Priya, S. Jhanwar, A. W. Khan, N. Shah, V. K. Singh, R. Garg, **G. Jeena**, et al. A draft genome sequence of the pulse crop chickpea (cicer arietinum l.). **The Plant Journal,** 74(5):715–729.

2012 R. Gaur, S. Azam, **G. Jeena**, A. W. Khan, S. Choudhary, M. Jain, G. Yadav, A. K. Tyagi, D. Chattopadhyay, and S. Bhatia**.** High-throughput snp discovery and genotyping for constructing a saturated linkage map of chickpea (cicer arietinum l.). **DNA research**, 19(5):357– 373.

2009 P. Katara, M. Agarwal, **G. Jeena**, S. Karkra, I. Sharma, and V. Sharma. In-silico prediction of the regulatory element patterns of human pathogen mycobacterium tuberculosis. **International Journal of Biotechnology & Biochemistry**, 5(1):7–14.

Cologne, 06.06.2020 **Ganga Jeena**