Growth-related systems genetics analyses and hybrid performance prediction in canola

Dissertation

zur Erlangung des

Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät I

– Biowissenschaften –

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt

von Herrn **Dominic Knoch**

geboren am 31.10.1989 in Halle (Saale)

Gutachter:

1. Prof. Dr. Thomas Altmann (Martin-Luther-Universität Halle-Wittenberg)
2. Prof. Dr. Bernd Weisshaar (Universität Bielefeld)
3. Prof. Dr. Stefan Scholten (Georg-August-Universität Göttingen)

Arbeit eingereicht am: 25.10.2019
Datum der öffentlichen Verteidigung: 10.03.2020

Preface

The work presented here is part of the DFG-funded project '*PREDICT*: Omics-based models for prediction of hybrid performance in oilseed rape'. Parts of this work have been published as a research article in Plant Biotechnology Journal, 'Strong temporal dynamics of QTL action on plant growth progression revealed through high-throughput phenotyping in canola', May 24th 2019. I am the first author of this manuscript and the share of my own work comprises the following parts: contributions to the design of the phenotyping experiments, performing the experiments, data analyses, and writing and editing of the manuscript.

Dedication

This dissertation is dedicated to my beloved parents Inge-Lore and Roland Knoch, my family and my friends. Thank you for all your support along the way.

Table of Content

page

page

page

List of Figures

page

List of Tables

List of Supplementary Figures

page

List of Supplementary Data

Supplementary Data S1 - S15 and an electronic version (PDF file) of this dissertation have been deposited on the attached Compact Disc (CD).

List of Abbreviations

ACR	**a**bsolute **c**hange **r**ate
ANOVA	**an**alysis **o**f **va**riance
ATP	**a**denosine **t**riphos**p**hate
BC	**b**ack**c**ross
BH	**B**enjamini-**H**ochberg procedure
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
BLASTn	**n**ucleotide **BLAST**
BLUE	**b**est **l**inear **u**nbiased **e**stimator
BLUP	**b**est **l**inear **u**nbiased **p**rediction
bp	**b**ase **p**air
BPH	**b**est-**p**arent **h**eterosis
BRAD	**Bra**ssica **D**atabase
bzh	dwarf **B**REI**ZH** gene
CDS	**co**d**i**ng **s**equence
Chr.	**chr**omosome
CNV	**c**opy **n**umber **v**ariation
cv	**c**ross-**v**alidation
cv.	**c**ulti**v**ar
CV	**c**oefficient of **v**ariation
CVPPP	**c**omputer **v**ision **p**roblems in **p**lant **p**henotyping
DAF	**d**ays **a**fter first **f**lowering
DAS	**d**ays **a**fter **s**owing
DEG	**d**ifferentially **e**xpressed **g**ene
DH	**d**oubled-**h**aploid
DNA	**d**esoxyribo**n**ucleic **a**cid
DSV	**D**eutsche **S**aat**v**eredelung AG
DTF	**d**ays **t**o onset of **f**lowering
DW	**d**ry **w**eight
EDTA	**E**thylene**d**iamine**t**etraacetic **a**cid
eGBLUP	**e**xtended **gBLUP**
Eq.	**eq**uation
eQTL	**e**xpression **QTL**
ERCC	**e**xternal **RNA c**ontrols **c**onsortium
F_1	**1**st **f**ilial generation
FAF	**FA**NTASTIC **F**OUR-like protein
FDR	**f**alse **d**iscovery **r**ate
FLUO	static **fluo**rescence
FW	**f**resh **w**eight
G	**g**enomic data
gBLUP	**g**enomic **b**est **l**inear **u**nbiased **p**rediction
GBS	**g**enotyping **b**y **s**equencing
GCA	**g**eneral **c**ombining **a**bility
GC-MS	**g**as **c**hromatography–**m**ass **s**pectrometry
GEBV	**g**enomic **e**stimated **b**reeding **v**alue
GLM	**g**eneralised **l**inear **m**odel
GMD	**G**olm **M**etabolome **D**atabase
GO	**g**ene **o**ntology
GS	**g**enomic **s**election
GSL	**g**lucosinolate
GWA	**g**enome-**w**ide **a**ssociation
GWAS	**g**enome-**w**ide **a**ssociation **s**tudy
H^2	broad sense **h**eritability
HCA	**h**ierarchical **c**luster **a**nalysis
HEAR	**h**igh **e**rucic **a**cid **r**apeseed
HOLLi	**h**igh **o**leic / **l**ow **li**nolenic acid
HSB	**h**ue, **s**aturation, **b**rightness (colour model)
HSV	**h**ue, **s**aturation, **v**alue (colour model)
HTP	**h**igh-**t**hroughput **p**henotyping
IAP	**I**ntegrated **A**nalysis **P**latform (software)
IPK	Leibniz-**I**nstitut für **P**flanzengenetik und **K**ulturpflanzenforschung
JLU	**J**ustus-**L**iebig **U**niversity Gießen
kb	**k**ilo **b**ase pairs (10^3 bp)
Lab	**l**ightness from black to white, **a**: green to red, **b**: blue to yellow (colour model)
LASSO	**l**east **a**bsolute **s**hrinkage and **s**election **o**perator

LC-MS	**l**iquid **c**hromatography-**m**ass **s**pectrometry
LD	**l**inkage **d**isequilibrium
LiDAR	**l**ight **d**etection **a**nd **r**anging
M	**m**etabolic data
MAF	**m**inor **a**llele **f**requency
MAS	**m**arker-**a**ssisted **s**election
Mb	**m**ega **b**ase pairs (10^6 bp)
MCMC	**M**arkov **c**hain **M**onte **C**arlo
MLM	**m**ixed **l**inear **m**odel
MPH	**m**id-**p**arent **h**eterosis
mQTL	**m**etabolite **QTL**
mRNA	**m**essenger **RNA**
MS1 & 2	**m**ale **s**terile 1 & 2
MSL	**m**ale **s**terile **L**embke system
MST	**m**ass **s**pectral **t**ag
MTA	**m**arker-**t**rait **a**ssociation
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
NIR	**n**ear-**i**nfra**r**ed
NMR	**n**uclear **m**agnetic **r**esonance
NPZ	**N**orddeutsche **P**flanzen**z**ucht Hans-Georg Lembke KG
NPZi	**NPZ I**nnovation GmbH
P	**p**henotypic data
PAR	**p**hotosynthetically **a**ctive **r**adiation
PAV	**p**resence / **a**bsence **v**ariation
PC	**p**rincipal **c**omponent
PCA	**p**rincipal **c**omponent **a**nalysis
PGPR	**p**lant **g**rowth-**p**romoting **r**hizobacteria
PLA	**p**rojected **l**eaf **a**rea
Pol	**pol**linator
PVE%	**p**henotypic **v**ariance **e**xplained in percent (**%**)
px	**pix**el
qRT-PCR	**q**uantitative **r**eal-**t**ime **p**olymerase **c**hain **r**eaction
QTL	**q**uantitative **t**rait **l**ocus / **l**oci
QTN	**q**uantitative **t**rait **n**ucleotide
REML	**re**stricted **m**aximum **l**ikelihood
RF	**r**andom **f**orest
RGR	**r**elative **g**rowth **r**ate
RIN	**R**NA **i**ntegrity **n**umber
RKHS	**r**eproducing **k**ernel **H**ilbert **s**pace
RNA	**r**ibo**n**ucleic **a**cid
RNA-Seq	**RNA**-**seq**uencing
rpm	**r**otations **p**er **m**inute
RR-BLUP	**r**idge **r**egression **BLUP**
rRNA	**r**ibosomal **RNA**
SCA	**s**pecific **c**ombining **a**bility
SD	**s**tandard **d**eviation
SE	**s**ingle **e**nd
SEA	**s**ingular **e**nrichment **a**nalysis
SNP	**s**ingle **n**ucleotide **p**olymorphism
SS	**s**um of **s**quares
svd	**s**ingular **v**alue **d**ecomposition
T	**t**ranscriptomic data
TAE	**T**ris-**A**cetat-**E**DTA (buffer)
TAIR	**T**he **A**rabidopsis **I**nformation **R**esource
TF	**t**ranscription **f**actor
tpm	**t**ranscripts **p**er **m**illion
Tris	**tris**(hydroxymethyl)-aminomethane
UV	**u**ltra**v**iolet
VIF	**v**ariance **i**nflation **f**actor
VIGS	**v**irus-**i**nduced **g**ene **s**ilencing
VIS	**vis**ible light

1. Introduction

1.1. Origin, uses and breeding of canola / rapeseed

Canola / rapeseed is the leading oilseed crop in Canada, Australia, China, and Europe, and second in global production. Rapeseed has diverse uses, as edible vegetable oil for human consumption, as animal feed due to the protein rich meal, as industrial feedstock, and as renewable energy source, for example for the production of biodiesel (Lu *et al.*, 2011). *Brassica napus* belongs to the large eudicotyledon family of the *Brassicaceae* consisting of about 3,000 plant species, including cabbage, cauliflower, kale (all *B. oleracea*), turnip (*B. rapa*), black (*B. nigra*) and oriental mustard (*B. juncea*), and the model plant *Arabidopsis thaliana*. Rapeseed is a natural allopolyploid hybrid (*Brassica napus* L., AACC, 2n = 4x = 38) of *B. oleracea* (contributed the C subgenome) and *B. rapa* (contributed the A subgenome) that emerged approx. 7,500 years ago (Chalhoub *et al.*, 2014). Allopolyploidy is not considered as a natural steady state, but as an evolutionary snapshot and intermediate condition after hybridization or genome duplication events (Doyle and Sherman-Broyles, 2017). In plant breeding polyploidy has often been induced in diploids to obtain desirable characteristics like seedless fruits or a higher seed yield (Sattler *et al.*, 2016). The *Brassica* genus has a propensity for genome duplications and genome merging as illustrated by the 'triangle of U' (Figure 1; Nagaharu, 1935) which illustrates that combinations of the genomes of three ancestral diploid *Brassica* species *B. oleracea*, *B. rapa* and *B. nigra* result in the three common tetraploid vegetable and oilseed crop species *B. carinata*, *B. juncea* and *B. napus*.

Historical remnants suggest that rapeseed was already cultivated 2000 B.C. in India. It has undergone a short, but intense breeding history since the 19th century (Hannoufa *et al.*, 2014; Mason and Snowdon, 2016) resulting in diverse high quality oilseed rape varieties, including 'double-low' varieties free of erucic acid and with a low content of seed oil glucosinolates, high oleic / low linolenic acid (HOLLi) varieties, and high erucic acid rapeseed (HEAR) whose oil (colza oil) is used for production of plastics, lubricants, lacquers and detergents, while their meal can be used as a livestock feed, as seed glucosinolate levels have been reduced.

Modern rapeseed breeding is focused primarily on three main targets: seed yield, seed quality and disease resistance. The yield potential of rapeseed depends to a large extent on flowering time, and flowering time adaptation (Wang *et al.*, 2011*a*). Based on these characteristics, three main

subgroups / primary gene pools can be distinguished: European winter-type oilseed rape, Asian semi-winter-type rapeseed and North American spring-type canola (Wang *et al.*, 2011*a*). Through gene transfer and mutagenesis, breeders have developed canola that is tolerant to herbicides, such as triazine-tolerant or glyphosate-tolerant canola, allowing for effective control of weeds in the field (Stanton *et al.*, 2010). In addition, breeding efforts have been focused to develop varieties resistant against clubroot (Chen *et al.*, 2016; Mei *et al.*, 2019) and other pathogens (Mitrousia *et al.*, 2018).

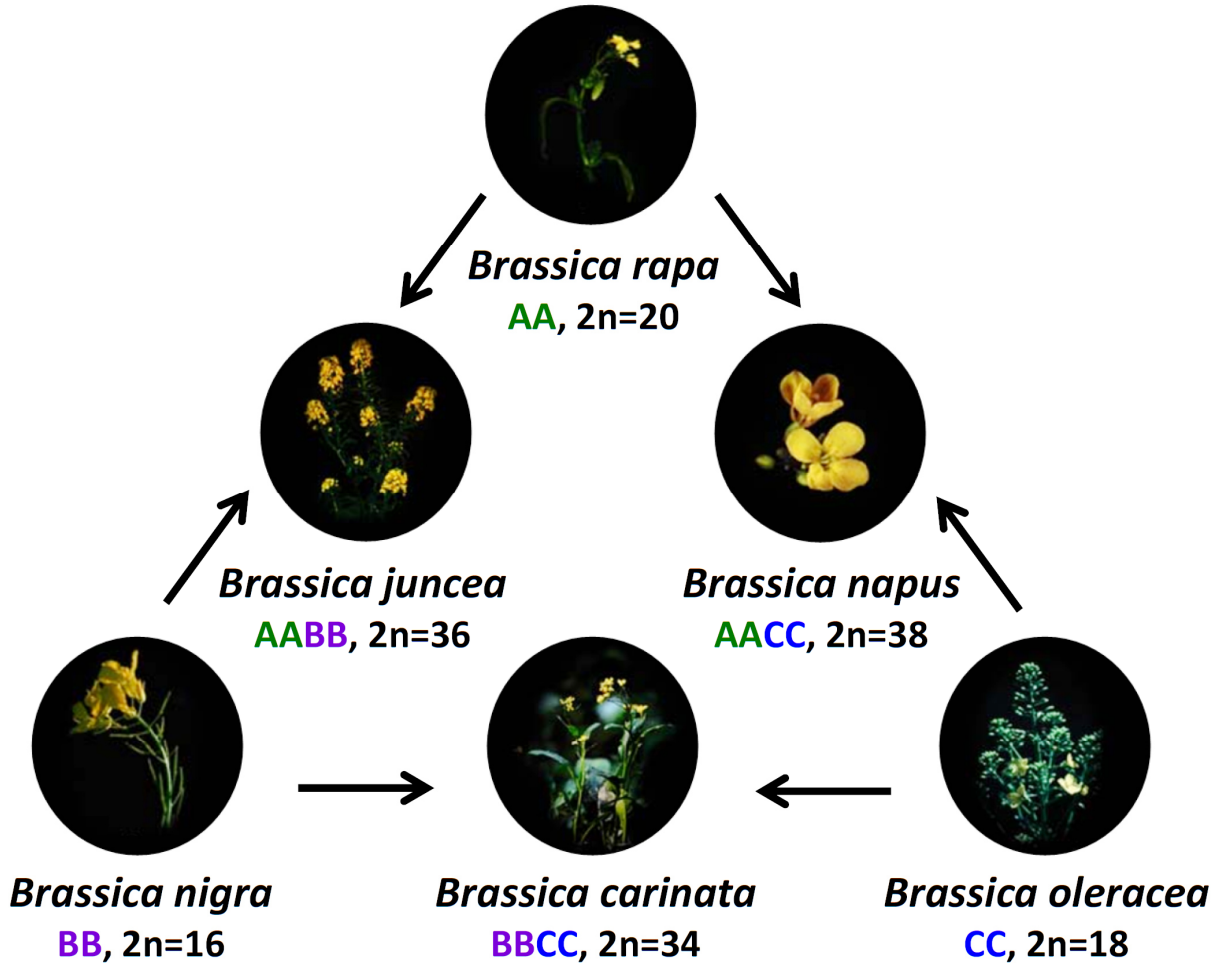


Figure 1. Relationship between members of the *Brassica* genus

Shown is the *Brassica* triangle of species as described by Nagaharu U (1935) with the A, B and C genomes and their respective amphidiploids (tetraploids). These species originated from spontaneous chromosome doubling via meiotic nondisjunction after interspecific hybridization events in regions where the respective diploid progenitors had an overlapping geographical distribution. This figure is a modified version of a figure published by Snowdon (2007).

Although rapeseed breeding is still mostly done using classical techniques, new biotechnological methods have been incorporated, such as the use of cytoplasmic male-sterility (Thompson, 1972; Wei *et al.*, 2019) and anther and microspore cultures (Keller *et al.*, 1975; Custers, 2003) to generate double haploids (Prem *et al.*, 2012), which have shortened the breeding process by years. The use of glasshouses and laboratories also boosted breeding, as now more than one generation per year is possible. In particular 'speed breeding' may hold the potential to shorten breeding cycles and to develop new varieties faster (Ghosh *et al.*, 2018). Future breeding will also further focus on N-efficiency, frost tolerance, improved nutrient use efficiency and resistance / tolerance to drought stress. Hybrid varieties are forecasted to be 'the future' of rapeseed breeding due to their outstanding heterotic features. Commercial varieties of oilseed rape are predominantly hybrids all over the world (Liu *et al.*, 2018*a*), for instance more than 80 % of cultivated rapeseed in Germany are hybrid varieties at the present day. These varieties perform better than open-pollinated varieties, in particular under stressful environmental conditions. Hybrid oilseed rape plants can be sown later, show higher disease resistance, and have enormous vitality and compensation ability, securing high, stable and consistent yield (Qian *et al.*, 2007; Zhang *et al.*, 2017*a*; Liu *et al.*, 2018*b*).

1.2. Recent advances in high-throughput phenotyping (HTP)

The phenotype of a plant is defined as the set of observable characteristics resulting from the complex interactions of its genotype with the environment, and the amount by which the expression of an individual genotype can be modified is termed its plasticity (Bradshaw, 1965). Phenomics, which studies the variety of phenotypic plant traits, is a key to understand genetic functions and the impact of environmental effects on plants. Crop plant performance is affected by the plant's genetic constitution and the characteristics of the surrounding environment. Many environmental factors, such as temperature, light, the weather and soil characteristics, along with biotic factors, for example weeds, diseases, insects and agricultural management, affect plant growth, quality and productivity. Several plant phenotyping platforms have been developed to study quantitative polygenic (complex) traits including biomass, plant colouration, growth-related traits, seed yield, the responses / resilience to biotic or abiotic stresses, as well as architectural traits. However, with the rapid development of genomics, the ability to acquire phenotypic data

has become the bottleneck in many plant genomic studies and crop breeding, as the phenotyping technologies were slow, labour-intensive, inaccurate in measurements of traits and often destructive (Furbank and Tester, 2011).

In recent years, technological advances have resulted in the availability of high-throughput phenotyping (HTP) offering non-invasive, image-based methods to analyse complex plant traits (Barabaschi *et al.*, 2016). With the increased throughput such systems allow to analyse effectively many plants (genotypes and replicates), necessary to provide power for genetic studies, in a reasonable amount of time. Consequently, many aspects of plant growth and morphological traits have been studied in depth for diverse model and crop plants, including Arabidopsis (Granier *et al.*, 2006; Hartmann *et al.*, 2011; Tisné *et al.*, 2013), maize (Junker *et al.*, 2015; Cabrera-Bosquet *et al.*, 2016; Muraya *et al.*, 2017; Zhang *et al.*, 2017*b*), rice (Yang *et al.*, 2014; Hairmansis *et al.*, 2014; Schilling *et al.*, 2015), barley (Honsdorf *et al.*, 2014; Neumann *et al.*, 2015) and rapeseed (Fanourakis *et al.*, 2014; Hatzig *et al.*, 2015; Kjaer and Ottosen, 2015) using mapping populations and natural accessions. The development of new platforms and techniques (Yang *et al.*, 2013; Li *et al.*, 2014*b*; Rahaman *et al.*, 2015; Roitsch *et al.*, 2019) for HTP allowed the efficient generation of measurements and to assess multiple traits at the same time, including plant growth and developmental traits, in a high-throughput manner at multiple time points. Examples include applications of visible light imaging for shoot biomass estimation (Walter *et al.*, 2007; Vanhaeren *et al.*, 2015), fluorescence sensing to quantitatively analyse photosynthesis (Tschiersch *et al.*, 2017), multispectral / hyperspectral imaging (Matsuda *et al.*, 2012; Sun *et al.*, 2019), near infrared spectroscopy for identifying physiological changes (Kuroki *et al.*, 2019) and thermal imaging to detect water stress (Munns *et al.*, 2010; Prashar and Jones, 2016). In addition, 3D imaging technologies, such as 3D laser scanning or light detection and ranging (Jimenez-Berni *et al.*, 2018; Rebetzke *et al.*, 2019) are emerging. However, they have not found wide applications in plant breeding yet due to the huge amount of complex data (‘big data’ issues, Zhao *et al.*, 2019) and high acquisition costs.

Although field phenotyping, manually, or by aerial or ground-based vehicles, constitutes an important category of phenotyping (Li *et al.*, 2014*b*; Chawade *et al.*, 2019), phenotyping under controlled, adjustable and reproducible conditions is a key requirement to effectively dissect genetic and environmental variances and to identify causal genetic determinants. Various types of

stationary (Walter *et al.*, 2007), vehicle-based (Tanger *et al.*, 2017; Chen *et al.*, 2019), self-propelled (Kicherer *et al.*, 2015; Salas Fernandez *et al.*, 2017), or portable phenotyping platforms (Ecarnot *et al.*, 2013) have been developed. Some of these platforms are fully automated facilities (Granier *et al.*, 2006; Golzarian *et al.*, 2011; Junker *et al.*, 2015) that allow precise control of the environment, and also include sensing equipment to log environmental changes such as light intensity, air humidity or soil water content.

Crop traits can be classified into three categories: morphological, physiological, and pathological. Traits such as shape, colour (spectral reflectance), texture, pattern and size are classified as morphological traits as they are direct measures of the forms and structures of the plant or their organs. Photosynthesis, respiration, nutrition, hormone responses, stress resistance and plant water relations are classified as physiological traits. Pathological traits as measures of plant diseases caused by pathogens or environmental conditions are often difficult to quantify directly and therefore predominantly assessed indirectly by morphological measurements (Fang and Ramasamy, 2015; Zhang and Zhang, 2018). Besides the above ground shoot part, plant roots are critical for plant growth and development. Plant roots are often complex, three-dimensional systems. Many traits such as root length, root density, number of roots, total surface, lateral root number, solidity or root convex area can be derived from imaging data obtained by scanning of washed roots (‘shovelomics’), rhizotrons or nuclear magnetic resonance (NMR) imaging systems (Pflugfelder *et al.*, 2017; Shi *et al.*, 2018; Atkinson *et al.*, 2019; Arifuzzaman *et al.*, 2019). However, root analysis is even more challenging than the analysis of the aerial parts of plants.

At the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, several high-throughput phenotyping facilities (Junker *et al.*, 2015) have been established and are continuously improved and enhanced. These systems have been previously used for the analysis of Arabidopsis (Junker *et al.*, 2015; Arend *et al.*, 2016; Tschiersch *et al.*, 2017), rice (Schilling *et al.*, 2015), barley (Chen *et al.*, 2014; Neumann *et al.*, 2015), maize (Muraya *et al.*, 2017), and rapeseed (Pommerrenig *et al.*, 2018). In addition to the generation of raw phenotyping data, the subsequent image processing and analysis require versatile tools and pipelines. HTPheno (Hartmann *et al.*, 2011), IAP (Klukas *et al.*, 2014), PlantCV (Fahlgren *et al.*, 2015; Gehan *et al.*, 2017) or Deep Plant Phenomics (Ubbens and Stavness, 2017) are just a few examples of such recently developed tools and software packages.

1.3. Rapeseed / canola genomics and genome-wide association studies

The availability of the *Brassica napus* reference genome sequence (Chalhoub *et al.*, 2014) and development of the high-density 60K (Clarke *et al.*, 2016; Mason *et al.*, 2017) and 15K (TraitGenetics internal development, currently unpublished) single nucleotide polymorphism (SNP) genotyping arrays have enabled genomic studies that greatly improved our understanding of the genetic basis underlying key agronomic traits in *Brassica napus*. With advances in sequencing and genotyping technologies, including genotyping by sequencing (GBS, Bayer *et al.*, 2015; Lees *et al.*, 2016) and array technologies, it has been feasible to generate genotype information for many lines in a high-throughput and cost-efficient manner as shown in recent studies (Wu *et al.*, 2016*b*; Stein *et al.*, 2017; Li *et al.*, 2018*a*; Zhang *et al.*, 2019*a*; Zhao *et al.*, 2019*a*). However, as SNP identification in polyploid genomes is complex and challenging, SNP discovery and array development has been seriously slowed in rapeseed (You *et al.*, 2018). A milestone boosting genomic analyses in rapeseed was the sequencing, annotation and publication of the *Brassica napus* genome of European winter oilseed cultivar Darmor-*bzh* (2n = 38, AACC), which has a size of approximately 1,130 Mb (Chalhoub *et al.*, 2014). In total, 101,040 gene models were estimated, which is roughly four-times the number of genes of its close relative *A. thaliana*. In addition to SNPs, copy number variation (CNV) and presence-absence variation (PAV) can provide complementary information (Stein *et al.*, 2017) for genetic analyses. In oilseed rape, in particular segmental deletions caused by widespread homoeologous exchanges (Samans *et al.*, 2017; Hurgobin *et al.*, 2018) were shown to be associated with trait variation (Qian *et al.*, 2016; Schiessl *et al.*, 2017; Stein *et al.*, 2017; Hatzig *et al.*, 2018). However, in contrast to its close relative *A. thaliana*, genetic analyses in rapeseed are complicated by the redundancy of genes due to the evolutionary recent 'collision' of the A and C subgenomes, their high homology and an extensive linkage disequilibrium, especially in the C subgenome (Wu *et al.*, 2016*a*).

With advances in high-throughput sequencing technologies, transcriptome sequencing (RNA-Seq) became available as frequently used omics-technology (Guo *et al.*, 2017; Shah *et al.*, 2018; Shahid *et al.*, 2019). However, transcriptomic approaches in rapeseed are challenged by the high similarity between the two subgenomes. Still, transcriptomic and associative-transcriptomic approaches (Harper *et al.*, 2012) have been used to study several aspects and traits of *B. napus* including the genetic architecture of erucic acid content and variation of tocopherol isoforms

(Havlickova *et al.*, 2018), lipid biosynthesis (Chen *et al.*, 2015) or freezing stress (Pu *et al.*, 2019). In addition, studies focus on traits relevant for breeding, including lodging resistance (Miller *et al.*, 2018), seed glucosinolates (Lu *et al.*, 2014), several yield-determining traits (Lu *et al.*, 2017), oil accumulation (Wan *et al.*, 2017) or thousand-seed weight (Geng *et al.*, 2018). The epigenome as additional omics layer has also been addressed in rapeseed, including genome-wide DNA methylation analyses using genome bisulfite sequencing and studies of small RNA expression (Takahashi *et al.*, 2018).

The metabolome constitutes another important layer of the omics-cascade. Almost all metabolic processes in living cells need enzyme catalysis in order to facilitate fast reaction rates to sustain life processes. Metabolic pathways depend on enzymes to catalyse individual reaction steps. The metabolome refers, similar as the term transcriptome, to the entire set of low molecular-weight compounds (metabolites) at a certain time point in a biological sample, which can either be a single cell, an organ or an entire organism (Jordan *et al.*, 2009). Profiling of metabolites provides a snapshot of the physiological status of a plant. Metabolites are defined as the intermediates or end products of metabolism and represent a dynamic system that can change rapidly. The metabolome comprises a huge range of very heterogeneous classes of chemical substances such as alcohols, amino acids, carbohydrates, lipids, nucleotides or organic acids. In general, the plant metabolism can be subdivided into primary metabolism, directly involved in the normal growth, development, and reproduction, and specialised metabolism including pigments, antibiotics or other compounds with functions that help plants to survive in specialised ecological niches (Mithen *et al.*, 1995) and under changing environmental conditions (Del Carmen Martínez-Ballesta *et al.*, 2013; Yang *et al.*, 2018). Metabolite levels can be regarded as the penultimate response of biological systems to genetic or environmental changes (Fiehn, 2002) and are most closely linked to the (morphological) phenotype of a plant. Moreover, metabolites in turn can influence gene expression and protein functions (Saito and Matsuda, 2010), or the metabolome can act as a buffer to changing environmental conditions (Gibon *et al.*, 2006). In particular mass spectrometric approaches, like gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS), but also nuclear magnetic resonance (NMR) approaches, allow in a targeted or untargeted way to simultaneously detect and quantify a wide range of small molecules in a high-throughput manner (Zampieri *et al.*, 2017). Owing to the

complexity of the metabolome and the diverse properties of metabolites, no single analytical platform is capable to detect all metabolites (Zhang *et al.*, 2012).

Beside phenotype information lines need to be characterised genetically to relate phenotypes to genes and their variation. For the successful identification of causal loci and genetic variants, besides a sufficient marker density, a sufficient degree of phenotypic variation in the traits of interest and a sufficient number of genotypes to provide the statistical power for the analysis are essential. Quantitative traits like seed yield, vegetative plant growth, early plant height and biomass production in rapeseed are under complex genetic control and are strongly influenced by the environment (Shi *et al.*, 2009; Zhao *et al.*, 2016). Dissecting the genetic basis of such traits is of high relevance to fundamental research and to crop improvement strategies alike. Previous studies applied quantitative trait locus (QTL) mapping and genome-wide association analyses using natural variation in populations to identify QTL / alleles for growth (Yong *et al.*, 2015), seed yield (Radoev *et al.*, 2008; Luo *et al.*, 2017*b*) and yield-related traits (Chen *et al.*, 2007; Yang *et al.*, 2012; Cai *et al.*, 2016; Dong *et al.*, 2018) in rapeseed. In some cases, genes underlying these QTL were also identified (Zeng *et al.*, 2011; Liu *et al.*, 2015*a*; Li *et al.*, 2018*c*). These GWAS analyses applied mixed linear models (MLM) or generalised linear models (GLM) to link phenotypic and genetic variation. However, these models have been shown to suffer from a trade-off between detection power and type I errors (false-positive results). Recently, a new method for genome-wide association studies, FarmCPU (Fixed and random model Circulating Probability Unification) has been proposed by Liu *et al.* (2016). The method iteratively performs marker tests with pseudo quantitative trait nucleotides (QTNs) as covariates in a fixed effects model and optimization on pseudo QTNs in a random effects model, which controls false positives and effectively reduces false negatives. The method was successfully applied in several studies and various species (Li *et al.*, 2016*c*; Hu *et al.*, 2017; Ravelombola *et al.*, 2017; Martinez *et al.*, 2018; Wang *et al.*, 2018*a*; Ward *et al.*, 2019).

The ultimate goal for breeders is to identify favourable lines in breeding populations according to their genotypes, and ideally, to stack multiple beneficial alleles for different genes in one genotype. Most of these studies however, have focused so far only on a limited number of phenotypic traits, and metabolic (mQTL) and expression (eQTL) studies in canola are rare (Qu *et al.*, 2016; Li *et al.*, 2018*b*; Yu *et al.*, 2018*a*). More importantly, in most studies only single time

points were hitherto analysed, although gene expression patterns are known to change during developmental progression.

In Arabidopsis, previous studies on projected leaf area at 12 different time points, on parameters derived from growth models, and on end-point biomass data revealed time-specific and general QTL affecting growth dynamics (Bac-Molenaar *et al.*, 2015). Similar observations were made regarding temporal patterns of biomass accumulation in barley (Neumann *et al.*, 2017), plant development and height in triticale (Busemeyer *et al.*, 2013; Würschum *et al.*, 2014*b*,*a*) and temporal expression of tiller number in wheat (Ren *et al.*, 2018). Dynamic QTL for plant height and for stress-responsive and several root traits at different developmental stages were also reported in upland cotton (Liang *et al.*, 2014; Pauli *et al.*, 2016; Shang *et al.*, 2016). In triticale, genetic dynamics underlying biomass yield were assessed in three developmental stages (Liu *et al.*, 2014). Interestingly, besides detecting QTL active in all stages, some QTL contributed to biomass development only in one or two of the stages. A recent study in maize assessed the genetics of growth dynamics at 11 different developmental time points and reported main-effect QTL and epistatic interactions with different patterns of expression and reversing allele effects (Muraya *et al.*, 2017). In *B. napus*, dynamic QTL for plant height were described that showed opposite genetic effects in different periods / stages and experiments (Wang *et al.*, 2015). However, multiple time point analyses to uncover the genetic basis for biomass and growth as dynamic traits have so far not been addressed in canola. The studies mentioned above highlight the need to investigate QTL / allele effects by time-series data to efficiently elucidate growth processes and to detect stage-specific loci that would likely be missed by analysing single or end-point data only.

1.4. Molecular genetics of vegetative plant growth and biomass accumulation

Early plant growth and biomass formation are crucial traits for productivity and seed yield, and early plant biomass (and heterosis) has been shown to be correlated with canola seed yield at the mature stage (Basunanda *et al.*, 2010; Zhao *et al.*, 2016). Moreover, early stage growth is of special importance for young seedlings to provide efficient ground coverage and to avoid competition with weeds in the fields. Growth-related heterosis manifests at a very early stage of seedling development in Arabidopsis (Meyer *et al.*, 2004, 2012) and in canola (Basunanda *et al.*, 2010), and plays a key role in field establishment. A recent study also linked early changes of gene

activity in developing seedlings of hybrids relative to parents to hybrid vigour in Arabidopsis (Zhu *et al.*, 2016). Hence, a better understanding of early plant growth is of great importance for breeding.Biomass accumulation in plants has been shown to be a complex trait, regulated and affected by various intrinsic developmental programmes / networks and environmental cues (Lima *et al.*, 2017). Plant growth can be defined as an irreversible increase in size of the plant that involves both, cell proliferation and cell expansion, whereby the timing of their transition has been shown to be of crucial importance (Gonzalez *et al.*, 2009, 2012). Increased biomass production can be achieved by two strategies, first by improvements of agricultural practices and second by genetic modifications that would increase plant growth and produce augmented plant dry matter. For instance plant cell wall features have been altered to increase the efficiency of biofuel production (Furtado *et al.*, 2014; Allwright and Taylor, 2016). Moreover, the nutrient supply, environmental stresses and also plant growth-promoting rhizobacteria (PGPR), a group of microorganisms that colonise the rhizosphere and roots of many plant species, have been shown to affect plant growth (Saghafi *et al.*, 2019).

Plant growth is influenced greatly by external environmental factors, still it appears that the size of plant organs is intrinsically determined by internal developmental factors. Intrinsic plant organ size, e.g. the petal size *of A. thaliana*, has been shown to be remarkably constant within individuals of a species (Mizukami and Fischer, 2000). A very large number of genes is involved in the control of plant growth and productivity in agriculture and many different pathways have been reported to be involved in biomass production. In Arabidopsis more than 70 growth promoting genes have hitherto been identified as reviewed by Lima *et al.* (2017).

One major determinant of plant growth is the cell cycle. It is, together with cell division, directly responsible for the number of cells which determines growth rate and overall organ size. Many of its molecular components have been shown to be conserved, like the cyclin-dependent kinases (CDKs), the cyclins (CYCs) and the multi-subunit E3 ubi-quitin ligase anaphase-promoting complex / cyclosome (APC/C, Inzé and De Veylder, 2006; Lima *et al.*, 2010; Inagaki and Umeda, 2011). In plants one major regulator of CDK activity are the inhibitor of CDK/KIP-related proteins (ICK/KRP, Verkest *et al.*, 2005), and down-regulation of ICK genes increased CDK activity and stimulated cell proliferation ultimately resulting in larger organs in *A. thaliana* (Cheng *et al.*, 2013). Moreover it has been shown that the S-phase regulating transcription factor E2F acts as positive

regulator of cell division (Vandepoele *et al.*, 2005). Overexpression of the APC/C, an E3 ubiquitin ligase that controls cell cycle transitions, accelerated plant growth and increased biomass production (Rojas *et al.*, 2009; de Freitas Lima *et al.*, 2013). SAMBA mutants, affected in a negative regulator of the Arabidopsis APC10, produced larger seeds and leaves (Eloy *et al.*, 2012). Also mutants of other cell cycle regulators like DA1-1 and EOD1-2 / BIG BROTHER have been reported to be significantly altered in plant organ size and biomass accumulation (Vanhaeren *et al.*, 2014, 2016, 2017).

Plant development and growth integrates many endogenous and environmental signals, whereby several growth regulating hormones including abscisic acid (ABA), auxins, brassinosteroids (BRs), cytokinins, ethylene and gibberellins (GAs) have been shown to be essential (Vanstraelen and Benková, 2012). In particular, GAs and BRs were linked to an increase of shoot biomass. GAs play an important role in the regulation of seed germination, flower and fruit development, stem elongation and leaf expansion (Hedden and Sponsel, 2015). The overexpression of the enzyme gibberellin 20-oxidase (GA 20-OX), which is rate-limiting in GA biosynthesis, resulted in increased biomass in poplar and maize (Voorend *et al.*, 2016; Jeon *et al.*, 2016), however, with negative side effects (significantly decreased leaf area or more slender stems). BRs are plant steroid hormones that regulate stem and root growth, floral initiation and fruit development (Zhu *et al.*, 2013). Two types of proteins, brossinosteroid-insensitive 1 (BRI1) receptor kinase, and brassinosteroid-associated kinase 1 (BAK1), a co-repressor, have been shown to be involved in BR perception (Li *et al.*, 2002). In rice, a *bri1* mutant has been reported to produce 35 % higher biomass compared to control plants at high planting density (Morinaka *et al.*, 2006). *DWARF4* is another important player catalysing a rate-limiting step in BR biosynthesis. Plants showed upon overexpression of *DWARF4* larger leaves and a higher number of branches (Choe *et al.*, 2001; Sahni *et al.*, 2016).

Transcription initiation is the initial step in which genes are selected for expression and modulation of expression levels (Joshi *et al.*, 2016) and several transcription factors have been shown to enhance growth when ectopically overexpressed, as reviewed by Gonzalez *et al.* (2009). Well studied examples from Arabidopsis include growth-regulating factor 5 (GRF5), its interacting protein angustifolia 3 (AN3/GIF1, Horiguchi *et al.*, 2005), the NAM/CUC transcription activator NAC1 (Xie *et al.*, 2000), aintegumenta (ANT, Mizukami and Fischer, 2000) and the At-hook transcription factor HRC1 (Century *et al.*, 2008). On the other side, several transcription factors,

such as auxin response factor 2 (ARF2, Okushima *et al.*, 2005), the HDZip transcription factor ATHB16 (Wang *et al.*, 2003) or the WD-40 transcriptional repressor (RON2, Cnops *et al.*, 2004) appear to repress plant growth. The overexpression of the heterosis-associated apetala2/ethylene-responsive element binding protein (AP2/EREBP) class transcription factor AP2L1 in Arabidopsis led to enlarged organs, increased biomass, and improved seed production (Li *et al.*, 2013*b*). Also manipulation of WRKY TFs like WRKY76 (Raineri *et al.*, 2015), NAC TFs (Grover *et al.*, 2014) that control growth of secondary cell walls, and basic helix-loop-helix TFs (bHLH, Noh *et al.*, 2015) resulted in increased biomass production. In addition, the homodomain-leucine zipper (HD-Zip) and GRAS TF families are important in cellular processes, causing morphophysiological alterations in plants (Hirsch and Oldroyd, 2009; Brandt *et al.*, 2014).

Beside these regulators, also photosynthesis and energy metabolism and involved genes may be targets to improve plant growth and biomass accumulation. For example, in rice, CO_2 assimilation was strongly correlated with the expression of the transcription factor gene *Higher Yield Rice* (*HYR*) and a high expression enabled enhancing the photosynthesis capacity and biomass accumulation of different organs (Ambavaram *et al.*, 2014). However, structural components of light reaction centres are highly conserved among plants (Rosado-Souza *et al.*, 2015) and several strategies to improve CO_2 assimilation and plant biomass failed due to the limited capacity of some species to utilise the photosynthesis products (Kirschbaum, 2011). Moreover, it has been suggested that the increase of biomass in *A. thaliana* hybrids, compared to parental lines, is not related to the photosynthetic rates, but rather attributable to an increase in the number of chloroplasts per cell and a higher chlorophyll content (Fujimoto *et al.*, 2012).

To provide energy for vital processes and to synthesise new organic material, all plants maintain sophisticated and complex metabolic networks. The two most essential nutrients thereby are carbon (C) and nitrogen (N), and keeping of the C/N ratio balance is essential for plant growth (Zheng, 2009). In Arabidopsis, overexpression of the nin-like protein 7 (NLP7) which modulates nitrate sensing and metabolism, enhances N assimilation and growth (Yu *et al.*, 2016). Another key enzyme of nitrogen metabolism, the glutamine synthetase (GS) has been overexpressed in pea resulting in high increases in plant fresh and dry weight (Oliveira *et al.*, 2002). Starch and soluble sugars play a central role in carbon metabolism and regulation of cellular physiology. Notably, the starch content at the end of the light period has been shown to be negatively correlated with

biomass and to be under circadian control (Sulpice *et al.*, 2009; Graf *et al.*, 2010). Substitution of mutants of the endogenous Arabidopsis isoform of starch branching enzyme (SBE) by a maize endosperm-enzyme significantly increased shoot dry weight of transgenic plants (Liu *et al.*, 2016*c*). Plant productivity also depends on source-sink relationships. Hence, it is not surprising that also the manipulation of genes involved in sugar transport and metabolism, like the sucrose synthase (SUS), the sucrose phosphate synthase (SPS) or the sucrose-phosphatase phosphatase (SPP) directly affect plant growth (Flügge *et al.*, 2011; Maloney *et al.*, 2015).

As additional factors, hybridity and ploidy (Fort *et al.*, 2016), as well as epigenetic regulation by miRNAs such as miR156 (Schwab *et al.*, 2005) have been reported to be involved in biomass accumulation. Also the size of the shoot apical meristem (SAM), in whose central zone cells divide to maintain a pluripotent stem cell population and at whose peripheral zone cells competent to differentiate are generated, has been considered to influence final leaf area and biomass (Gonzalez *et al.*, 2012). For instance, the *clavata1* mutant has been shown to have a larger SAM and an increased leaf initiation rate (Kwon *et al.*, 2005). Furthermore, the number of cells recruited to the leaf primordium has been shown to play a role in final organ size. In the *struwwelpeter* (swp) mutant, which shows an altered expression of an RNA polymerase II transcription mediator, smaller leaves containing fewer cells were produced (Autran *et al.*, 2002).

Considering all the examples of transgenic modifications causing severe effects mentioned above, it is likely that also natural variation in these genes and pathways affects growth and biomass accumulation. However, the analysis of subtle changes of complex biomass-related phenotypic traits like the projected leaf area (PLA), plant height or the estimated volume of a plant, have been challenging, as they require precise, repeated and often time-consuming measurements. This challenge has been addressed by the emergence of new high-throughput phenotyping technologies that allow a reliable, non-invasive and versatile acquisition of diverse phenotypic traits.

1.5. Heterosis – the genetic basis of hybrid vigour

Heterosis, or hybrid vigour, is an important, intensely studied but still poorly understood genetic phenomenon. The term heterosis was introduced by Shull in 1914 (Shull, 1948). It refers to the advantage of heterozygous offspring over their homozygous parents with regard to fitness-related

traits, such as seed yield. However, despite the high agronomic value, the genetic basis of heterosis, its mechanistic understanding and its prediction for improved efficiency of crop breeding still remains an elusive goal (Chen, 2013). As the heterotic effects peak in the F_1 and are lost during inbreeding, it is generally thought that heterosis is based on the contributions of numerous genetic factors, each only with a small effect and that, at least to some extent, the combined action of heterozygous alleles is involved. Moreover, a positive and highly significant correlation between the genetic distance of cultivars and mid-parent heterosis has been described in canola (Lefort-Buson *et al.*, 1987; Ali *et al.*, 1995). Three main classic hypotheses, dominance, overdominance and pseudo-overdominance, which try to explain the phenomenon at the genetic level have been proposed, and experimental evidence has been obtained for all of them (Lippman and Zamir, 2007). In addition, partial dominance, epistatic interactions (Shi *et al.*, 2011), co-regulated gene networks (Basunanda *et al.*, 2010) and epigenetic factors were also found to contribute to heterosis (Kawanabe *et al.*, 2016; Shen *et al.*, 2017; Fujimoto *et al.*, 2018). A model of physiological dominance has been proposed by Sewall Wright suggesting that heterosis of plant performance is an intrinsic property of nonlinear relationships between traits (Wright, 1934; Fiévet *et al.*, 2018; Vasseur *et al.*, 2019). According to a metabolic heterosis model, the gene expression in hybrids at the mid-parent level generates hybrid vigour by counterbalancing opposing detrimental expression levels in the parental lines on a genome-wide scale (Kacser and Burns, 1981; Springer and Stupar, 2007). An optimal distribution of enzyme quantities may approach maximal metabolic fluxes in hybrids, while in inbred lines several enzymes may be expressed at non-optimal amounts (de Vienne *et al.*, 2001; Fiévet *et al.*, 2010).

In the model crucifer *A. thaliana*, biomass heterosis is established early during vegetative development (three to six days after sowing). However, depending on which accession is used as crossing partner, is not already present in the seed (Meyer *et al.*, 2004, 2012). In winter oilseed rape, significant biomass heterosis was also observed during early seedling development, and these hybrids also showed significant yield heterosis at later developmental stages (Basunanda *et al.*, 2010). Given these similarities, and due to the small genome, the fast generation cycle and the evolutionary close distance to rapeseed, *A. thaliana* constitutes an ideal system for heterosis research. Knowledge gained in the model plant Arabidopsis might be transferred in a straightforward manner into the crop plant canola.

The most important element in implementing hybrid breeding is the recognition of a heterotic pattern that supports high-yielding lines (Zhao *et al.*, 2015). In oilseed rape, hybrids today already make up the major share of the international seed market (Stahl *et al.*, 2017). However, in comparison to other important hybrid crops like maize, canola displays relatively low levels of F_1 heterosis (for example MPH of approx. 30 % for grain yield) as reported by Radoev *et al.* (2008). This can be attributed to the fact that in contrast to outcrossing crops like maize, hybrid breeding in self-pollinating crops like rapeseed began only a few decades ago after suitable hybrid seed production systems, for example cytoplasmic male-sterility, were developed, and therefore, large and well defined heterotic pools as in maize have not been established yet (Melchinger and Gumber, 1998; Kole, 2007). However, several attempts were made to broaden the genetic diversity and to develop heterotic gene pools for rapeseed hybrid breeding (Qian *et al.*, 2007; Girke *et al.*, 2012; Jesske *et al.*, 2013; Li *et al.*, 2014*c*). Thus, the breeding industry has considerable interest in utilisation of heterosis to improve plant performance, under optimal and stressful conditions. The identification of new superior hybrids among the millions of possible crosses requires extensive breeding programmes, involving the generation of various testcrosses between breeding material, extensive multi-location / multi-year field trials to generate phenotypic data and to test hybrid performance. As such programmes are both work and cost intensive (Desta and Ortiz, 2014) the prediction of hybrid performance is highly desirable. A straight-forward preselection of a few hundred, most favourable hybrids with high success rate, could substantially reduce the volume of the labour-intensive and time-consuming field trials (Xu *et al.*, 2016; Kadam *et al.*, 2016) and greatly impact the efficiency of hybrid breeding (Longin *et al.*, 2015).

1.6. Genomic selection and prediction of hybrid performance

In the past, individual plants with desirable phenotypic characteristics were selected and used as progenitors of the next generation. In contrast, modern plant breeding is based on the discoveries of Darwin and Mendel about evolution and inheritance (Borlaug, 1983), shifting the breeding from a parent-oriented to an offspring-oriented way. With the *Green Revolution* in the 1960's, this development reached a peak resulting in many new varieties and increased food production.

The overall aim of plant breeding is to improve crops by fixing and improving phenotypic traits. Quantitative (complex) phenotypic traits are determined by many genes of small effect and their

interactions with the environment (Lander and Schork, 1994; Heino, 2014). Knowledge about the underlying genetic architecture of traits is of crucial importance to support breeders because parental lines with the best phenotypes, in particular regarding seed yield and other heterotic traits, are not necessarily the best progenitors for a breeding programme (Melchinger and Gumber, 1998). Hence, breeders aim to detect quantitative trait loci (QTL), genomic regions that are associated with such traits, by linkage mapping and/or association mapping. With the help of markers in proximity and genetically linked to traits of interest, such traits can be selected indirectly by marker-assisted selection (MAS). In this way, beneficial genes can be selectively introduced in populations or genotypes, and ideally also stacked, or detrimental genes excluded (Collard and Mackill, 2008). Marker-assisted breeding and its variants such as marker-assisted recurrent selection (MARS, Lande and Thompson, 1990), a two-step approach involving the selection of significant markers and combination with phenotypic information in a selection index, and marker-assisted backcrossing (MABC) found implementation in plant breeding for a series of crop plants (Collard and Mackill, 2008; Jiang, 2015). However, MAS relies on available information from QTL mapping studies that often identify only few QTL with overestimated effects (Xu, 2003; Schön *et al.*, 2004). The relative efficiency of MAS decreases with an increasing number of QTL and decreasing trait heritability (Moreau *et al.*, 1998), which makes it less effective for complex polygenic traits such as seed yield.

A more recent method in plant breeding is genomic selection (GS) in which marker data are combined with pedigree information and phenotypic data to build a model that accurately predicts the performance of individuals with the aim to select progenitors in a breeding programme (Crossa *et al.*, 2010). Bernardo (1994) proposed to use a genomic relationship matrix estimated from DNA. The matrix defines the covariance between individuals based on observed similarity at the genomic level, rather than on the expected similarity based on pedigree information. Furthermore, Meuwissen *et al.* (2001) suggested to estimate all marker effects simultaneously. This approach is referred to as genomic best linear unbiased prediction (gBLUP).

For prediction of traits with genetic markers, phenotypic and genotypic data need to be collected. Marker effects are estimated in a training population using a statistical model describing the relationship between marker and phenotype data with genotypic data as predictors for the phenotype. Next, new individuals of a validation population are genotyped and the phenotypes

are predicted (genomic estimated breeding values, GEBVs) with the previously established model, and promising genotypes are then selected for a breeding programme. Advantages hereby are that in contrast to many phenotypic data, the genotype of plants can be evaluated in a very early developmental stage and the genotyping costs are constantly decreasing (Zhao *et al.*, 2015). For the prediction of hybrid performance, not the hybrids themselves are genotyped, but rather the parental (inbred) lines. With the help of testcrosses, the average performance of an inbred line in hybrid combinations (general combining ability, GCA), the relative performance of certain combinations compared to the average performance of the lines (specific combining ability, SCA) of the parental lines or the hybrid performance itself can be estimated (Sprague and Tatum, 1942). In order to evaluate the quality of the predictions, phenotypic data of the predicted individuals from the validation population are collected and correlated with the predicted values (prediction accuracy). The accuracy of predictions has been shown to be dependent on the statistical model applied, the type and density of genetic markers, the size, composition and ratio between training and validation population (Akdemir and Isidro-Sánchez, 2019). In addition, the overall genetic relatedness of the population (Albrecht *et al.*, 2014), the genetic architecture and the heritability of the trait which is to be predicted play major roles (Morgante *et al.*, 2018; Wang *et al.*, 2018*b*). Gene effects and the distribution of linkage disequilibrium (LD) between genetic markers, genotype × environment interactions and quantitative trait loci (QTL) have also been shown to be of importance (Windhausen *et al.*, 2012; Desta and Ortiz, 2014; Dan *et al.*, 2016). Hence, a major disadvantage and limitation of this approach is that good predictions are only achieved for highly similar populations and application to other gene pools / varieties might result in lower accuracies of the model.

Genome-wide regression / prediction is a powerful tool to analyse and predict quantitative polygenetic traits (Meuwissen et al. 2001) and various genome-wide prediction approaches for hybrid prediction have been explored and applied for plant populations over the last decades (Crossa *et al.*, 2014; Heslot *et al.*, 2015; Mangin *et al.*, 2017; Hickey *et al.*, 2017). Models include gBLUP and its equivalent ridge regression BLUP (RR-BLUP, de Vlaming and Groenen, 2015), but they are restricted to the incorporation of additive effects only. To overcome this limitation, advanced models have been proposed, including extended gBLUP (eGBLUP) which includes epistatic marker effects into the model (Jiang and Reif, 2015). Taking epistasis into account can

increase prediction accuracies as previously shown (Wang *et al.*, 2012; Muñoz *et al.*, 2014; He *et al.*, 2016). The assumption made by RR-BLUP that genetic effects are evenly spread across the genome (homoscedastic marker variances) was not satisfactory and Meuwissen *et al.* (2001) tried to relax it using Bayesian models. This led to the development of various constantly improved Bayesian implementations such as BayesA, BayesB, BayesC*π,* BayesD*π*, Bayesian LASSO (Meuwissen *et al.*, 2001; Park and Casella, 2008; Habier *et al.*, 2011). However, these models have a high computational demand (Lorenz *et al.*, 2011) and pose the difficulty to choose an appropriate prior distribution of the marker effects (Piepho, 2009). Haploblock-based genome-wide prediction models have been developed as well and reported to outperform marker-based models in some scenarios (Calus *et al.*, 2008; Jiang *et al.*, 2018; Jan *et al.*, 2019).

In addition, kernel-based methods like reproducing kernel Hilbert space regression (RKHS) have been exploited for predictions (Gianola and van Kaam, 2008). They contain a great deal of flexibility and no assumptions of linearity, which may render them superior in their ability to capture nonadditive genetic effects. However, studies have shown that there is no universally best prediction model (Momen *et al.*, 2018). Hence, in addition to superior predictive models and algorithms, also alternative sets of predictors have been addressed. There is evidence that genomic prediction may not be capable to capture all complex gene interactions and downstream regulatory processes, even with complete sequence information available (Zhu *et al.*, 2012; Ritchie *et al.*, 2015). Hence, the utilisation of endophenotypes such as metabolite abundances and gene expression was proposed to improve the prediction of complex trait, as they are expected to reflect more closely the variability across genotypes than genomic data *per se* (Mackay *et al.*, 2009; Patti *et al.*, 2012; Civelek and Lusis, 2014).

Previous studies in maize (Riedelsheimer *et al.*, 2012*a*; Feher *et al.*, 2014), rice (Dan *et al.*, 2016, 2018; Xu *et al.*, 2016; Wang *et al.*, 2019) and Arabidopsis (Meyer *et al.*, 2007; Gärtner *et al.*, 2009; Steinfath *et al.*, 2010) have shown that metabolite levels also may have high predictive power and can in some, but not all, cases (Zhao *et al.*, 2015) improve prediction accuracies. Recent studies illustrate the predictive value of transcriptome data (Swanson-Wagner *et al.*, 2006; Fu *et al.*, 2012; Zenke-Philippi *et al.*, 2016), including small RNAs (Seifert *et al.*, 2018). Compared to genomic data, transcripts have the advantage that they are independent of marker LD and are therefore better suited for prediction across heterotic pools (Frisch *et al.*, 2010). Consequently, downstream -omics

data, including expression data and metabolite profiles, are expected to integrate interactions within and between biological layers, thus they may capture physiological epistasis (Westhues *et al.*, 2017). With the examples mentioned above, it has been shown that endophenotypes provide reasonable predictive abilities relative to those of genetic markers and their integration with genetic markers can significantly improve predictive abilities (Guo *et al.*, 2016; Westhues *et al.*, 2017; Schrag *et al.*, 2018; Wang *et al.*, 2019).

1.7. Aims of this work

Based on previous work on biomass and heterosis prediction in Arabidopsis (Meyer *et al.*, 2007; Steinfath *et al.*, 2010) and maize (Riedelsheimer *et al.*, 2012), this work is built on the hypothesis that specific allelic combinations of regulatory genes, their downstream gene expression, as well as elicited metabolite profiles, are associated with improved vegetative growth, hybrid performance and seed yield in hybrids. The project featured three main goals: first, to evaluate omics-based models for prediction of hybrid performance in spring-type oilseed rape that can be effectively implemented in a commercial breeding programme; second, to elucidate links between vegetative growth, transcript and metabolite levels; and third, to identify candidate genetic loci / genes causing trait variation.

To address these goals, prior to this work, a hybrid population of 950 genotypes had been generated and evaluated in the field at multiple locations across Europe, analysing various agronomic traits. Complementarily, at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), detailed phenotyping data should be generated by growing 475 diverse pollinators from a commercial canola breeding programme and two elite male-sterile tester lines, comprising the parental lines of the hybrids, in the IPK automated high-throughput phenotyping platform (Junker *et al.*, 2015). Image data obtained at an early vegetative state should be complemented by endophenotypes, polar metabolite and global transcriptome (RNA-Seq) profiles, from the same plants. These extensive data sets should be utilised for correlation analyses, and in combination with array-derived SNP and CNV data for genome-wide-association studies (GWAS) to identify genetic loci associated with trait variation and candidate genes. As phenotypic data covering the early vegetative growth would be generated as a time series, this study should allow temporal analyses of genetic determinants contributing to growth-related traits.

2. Materials and methods

2.1. Genetic material and generation of an F_1 hybrid population

The experimental materials consisted of a total of 479 genotypes with double-low seed quality (low erucic acid, low glucosinolate content) from a spring-type *B. napus* (canola) breeding programme (Data S1) that showed contrasting patterns of general combining ability (GCA). The materials were carrying introgressions from the diploid progenitors of *B*. *napus*. The largest proportion, 475 lines, comprised genetically diverse pollinator lines that could be attributed to three breeding pools (denoted as breeding pools 1, 2, and 3). Some of the lines exhibited a high degree of heterozygosity. Two elite male-sterile testers (MS1 and MS2) from a pool of testers carrying the *Male Sterility Lembke* (MSL) sterility system (NPZ Lembke, Hohenlieth, Germany), and two commercial elite genotypes, 'Achat' and 'Campino' (*CR 3430*) were included. An F_1 hybrid population with 950 genotypes was generated by the commercial partners in this project by crossing all 475 pollinators to the two male-sterile testers.

2.2. Field experiments by commercial partners and agronomic traits

Field trials were performed in the year 2012 in a nested design by commercial partners NPZi and DSV whereby the 950 hybrids were evaluated at eight different locations across Denmark (Abildgard, Dyngby), Germany (Roßleben, Hohenlieth), Poland (Słupia, Krzyżewo), Latvia (Jelgava) and Estonia (Viljandi) at commercial plant breeding testing sites. A total of nine trials were performed with 35 trial ~ location combinations. Each trial evaluated plants at three to four individual locations and each tested hybrid had four replicates across all trials / locations. Four commercial lines ('Achat', 'Osorno', 'Mirakel' and 'DLE 1108') were included as standards in all trials at each location and tested either unreplicated or as duplicates in each trial. Various traits of commercial importance were assessed including the content of total seed glucosinolate (GSL; µmol/g seed), the days to onset of flowering (DTF; measured as number of days from sowing until 50 % flowering plants per plot), the seed oil yield (dt/ha), the trait seedling emergence (visual observation ranging from a minimum value of 0 to maximum 10), the seed oil content

(% volume per seed dry weight), the seed protein content (% volume per seed dry weight) and seed yield (dt/ha).

2.3. Plant cultivation under controlled conditions and experimental design

Prior to the main experiments three smaller pilot experiments were performed to optimise the cultivation and phenotyping of rapeseed in the glasshouse containing the phenotyping platform. The first pilot experiment addressed the questions of the substrate to be used, possible unequal germination rates of different lines, the adequate number of plants per pot for effective phenotyping, and suitable environmental conditions. In the second the growth speed of plants under the selected conditions was monitored, and the third experiment was performed to sample plant material for an initial metabolic analysis to address the question whether early (14 days after sowing, DAS) or later (28 DAS) plant material should be sampled. In these experiments, a set of commercial breeding lines and hybrids were used (Data S1). The results of these experiments led to the conditions and final experimental design as described below.

Plants were cultivated under controlled environmental conditions in an incomplete randomised block design (Data S2) in four glasshouse experiments (1413RCM, 1419RCM, 1442RCM and 1447RCM) in spring and winter 2014. An additional experiment (1504RCM) with a selection of 120 hybrids was performed in spring 2015 and data were included in the calculation of the BLUEs, but not in the GWAS, as no array data for the hybrids are available.

Experiments were carried out in the IPK phenotyping facility for large plants (Junker *et al.*, 2015) comprising a cultivation, transportation and imaging system with 396 mobile carriers. Each genotype was replicated three times. A container with nine plants comprised one replicate. Four reference lines were included as 'Checks' in higher replication ('*Achat*' n=12, '*Campino*' (*CR 3430*) n=12, and the two male sterile testers 'M1' and 'M2' each n=6 per experiment, respectively) in all cultivations. Plants were grown in large 25 litre square containers (Bamaplast S.r.l., Massa e Cozzile, Italy) in *red substrate 2* (Klasmann-Deilmann GmbH, Geeste, Germany) to provide enough space for the plants to grow and to avoid pot size effects (Poorter *et al.*, 2012), and covered with a blue rubber mat to facilitate image analysis. Before sowing, seeds were stratified for three days at 4 °C on moist filter paper in Petri dishes to trigger uniform germination. To ensure homogenous plant density, two seeds per position were sown, but were thinned to one

seedling per position at 5 DAS. A controlled climate regime was applied, on the one hand in consideration of natural field-like conditions and on the other hand to ensure consistency of conditions among the experiments. Temperatures were kept constant with 10 °C (dark phase) and 15 °C (light phase) during the entire growth period and the natural radiation was supplemented by additional illumination of 205–245 µmol $m^{-2}s^{-1}$ PAR using SonT Agro high pressure sodium lamps. The light period was set to 16 h light from 06:00 h to 22:00 h. These conditions correspond to a typical spring scenario in central Europe. Relative air humidity was set to a target value of 65 %. Watering was performed with an automated balance / watering station by target weight of the containers to maintain 80 % field capacity, pH 5.5. Containers were shuffled each day by one row and every second day by one block (eleven neighbouring plants in one row) in the system to minimise position effects. Initially, nine plants per container were cultivated (Figure S1 and S2).
At 14 DAS four plants per container were sampled to provide enough material for subsequent molecular / biochemical analyses resulting in approximately 7,920 plant samples. The remaining five plants were grown until 28 DAS.

2.4. Extraction and analysis of image-derived phenotypic data

Over a duration of three weeks (between 6 DAS and 28 DAS), plants were subjected to a daily imaging routine involving automated capturing of top and side-view images. Three types of illumination and camera systems in the IPK automated non-invasive plant phenotyping system for large plants were used as described in Junker *et al.* (2015). Visible light (VIS), static fluorescence (FLUO) and near-infrared (NIR) image data were acquired. Each carrier was imaged with two cameras within each system with one top view and four / three side-views taken at (0°, 45°, 180°, 225°) from 6 to 13 DAS and (0°, 45°, 135°) from 15 to 27 DAS, resulting in more than 84,000 individual images per experiment.

2.4.1. Automated high-throughput plant phenotyping and image analysis

Automated image analysis was performed using the IPK Integrated Analysis Platform, IAP Version 2.07 (Klukas *et al.*, 2014) implementing a customised pipeline combining top and side view images, and including image pre-processing, segmentation and feature extraction resulting in a total of 1,194 image-derived phenotypic traits. The pipeline comprised the following steps: 01 'Load images', 02 'Filter images by angle', 03 'Rotate images', 04 'Align camera images', 05 'Colour balancing VIS', 06 'Background correction FLUO', 07 'Circular colour balancing', 08 'Copy image set to mask', 09 'Crop sides', 10 'Auto-tune FLUO-segmentation', 11 'Median filter FLUO', 12 'Colour segmentation Lab', 13 'Colour segmentation HSV', 14 'Colour Segmentation HSB', 15 'Auto-tuning VIS-segmentation (k-means)', 16 'Morphological operations', 17 'Auto-tuning small noise removal', 18 'Adaptive NIR-segmentation', 19 'Auto-tuning FLUO-derived masking of other images', 20 'Morphological operations', 21 'Separate objects', 22 'Skeletonize', 23 'Skeletonize NIR', 24 'Calculate colour- and intensity-histograms', 25 'Calculate colour intensities', 26 'Calculate convex hull', 27 'Calculate areas', 28 'Calculate width and height (side)', 29 'Calculate texture features', 30 'Detect leaf centre points', 31 'Calculate volume estimations', 32 'Run post-processors', 33 'Move mask set to image set', 34 'Highlight null images', 35 'save result images'. Setting for all steps were empirically determined and optimised for the experimental setup, whereby different settings for the early and late phases and for top and side view images were combined in one pipeline. Small adjustments were performed for each individual experiment. The extracted traits comprised 128 (10.7 %) geometric traits giving insights into general plant morphology, 930 (77.9 %) plant colouration-related traits corresponding to pigmentation, 104 (8.7 %) fluorescence-related traits, mainly associated with chlorophyll fluorescence and 32 (2.7 %) near-infrared-related traits linked to water content and water dynamics. As these traits were partially redundant and inter-correlated, a two-step procedure was applied to filter them to a meaningful core set for further analysis. First, traits were filtered for broad-sense heritability higher than 0.7 for at least one day, reducing the set of phenotypic traits to 571. Second, stepwise variable selection using variance inflation factors (until VIF ≤ 10) was applied to minimize the multi-collinearity (Chen *et al.*, 2014), further reducing the number of traits. In total, 123 traits, including a subset of 32 manually selected traits of particular interest, were kept for subsequent analysis. In particular, four traits were focused on: the estimated biovolume, ´combined geometry

vis volume iap (voxel)´ which was estimated by combining information from top- and side-view (Junker *et al.*, 2015), projected leaf area ´top geometry vis area (px^2)´ which was derived from VIS top-view, early plant height ´side geometry fluo height (px)´ derived from FLUO side view, and plant colour uniformity ´side intensity vis lab a stddev´ extracted from VIS side-view images, respectively. Colour uniformity is given as the standard deviation of the a-values in the L*a*b* colour space of the plant pixels. The lower this value, the more uniform is the plant colour. Leaf colour differs between young and old leaves and therefore this trait may act as a proxy for the range of maturation stages of leaves within a given plant and thus of its rate of development. Traits represent information obtained from the analysis of whole containers, including nine plants at early and five plants at later stages, respectively. Shoot fresh weight (g) was determined on the basis of all five plants by cutting the shoots directly above ground level and by weighing using a medium-scale balance at 28 DAS. Dry weight was measured after drying the plant material for 3 days at 80 °C. Approximately 9,900 samples were collected and analysed during the five glasshouse experiments.

2.4.2. Post-processing and statistical analyses of phenotypic data

All statistical analyses were performed in the R software environment for statistical computing version 3.4.2 (R Core Team, 2019) and graphics and RStudio Version 1.1.419. Image derived traits were obtained from 6 DAS to 13 DAS and from 15 DAS to 27 DAS. An outlier-correction was performed in a combined approach of manual exclusion (carriers with insufficiently germinated plants) and a threshold-based filtering procedure (median ± 3 standard deviations) for each experiment, day and trait separately. A single-step analysis of the phenotypic data was performed. Best linear unbiased estimators (BLUEs, Data S3) were estimated using the 'lme4' package in R (Bates *et al.*, 2015) based on a linear mixed model for each image-derived phenotypic trait and each day separately (Eq. 1) or in case of end-point biomass data (Eq. 2). In the models, Y denotes the phenotypic value of a trait for each genotype, G represents the fixed effect of the Genotype, E the random effect of the Experiment, GxE the Genotype-Experiment-Interaction, C the random effect of the included 'Checks', CxE the Check-Experiment-Interaction, P the Position in the pot, PxE the Position-Experiment-Interaction and e the residual error (errors were assumed to be normally, independently, and identically distributed). Broad-sense heritabilities (H^2) for each trait

were estimated by Eq. 3, where σ_G^2 and σ_e^2 denote the variance components of the genotype and the residual variance, respectively, which were estimated using and n_0 the number of experiments or in case of the end-point biomass data the number of plant replicates per genotype (Nakagawa and Schielzeth, 2010; He *et al.*, 2016). Variance components σ_G^2 and σ_e^2 were estimated by restricted maximum likelihood (REML) and extracted from the mixed linear models (Eq. 1 or Eq. 2) in R 'lme4' assuming that all effects were random effects.

Eq. 1*:* $$Y = G + E + GxE + C + CxE + e$$

Eq. 2: $$Y = G + E + P + GxE + C + CxE + PxE + e$$

Eq. 3: $$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{n_0}\sigma_e^2}$$

2.4.3. Calculation of relative growth and absolute change rates

Absolute change rates (ACRs) and relative growth rates (RGRs) were calculated using previously published procedures (Hunt, 1990). Relative growth rates were determined for the estimated biovolume, projected leaf area, and early plant height (Eq. 4). To compensate for a potential growth bias due to the applied plant rotation / shift in image acquisition, growth rates were calculated with minute precision, as image acquisition date and time were documented. In addition, absolute change rates were calculated (Eq. 5) for plant colour uniformity. BLUEs for ACR and RGRs were subsequently estimated as described above (Eq. 1).

Eq. 4: $$RGR = \frac{(log_e W_2 - log_e W_1)}{(t_2 - t_1)}$$

Eq. 5: $$ACR = \frac{(W_2 - W_1)}{(t_2 - t_1)}$$

2.5. Sampling of early vegetative shoot material and post-processing

Shoot material (mainly leaves) of the four inner plants of each container was sampled at 14 DAS, starting seven hours after onset of illumination (Figure S3). Plant material was immediately shock-frozen in liquid nitrogen to quench the sample. Sampling of all 396 carriers was accomplished within three hours to minimise effects of diurnal rhythm. Deep frozen plant material was homogenised for 2 min using two 8 mm steel balls and a cryogenic plant grinder and dispensing system (Labman Automation Ltd., Stokesley, United Kingdom) at -60 °C. The four replicates per carrier were combined and the material was mixed by shaking the vial for one minute. Equal amounts of plant material of all replicates from the different experiments were pooled and mixed by shaking the vial for one minute. Samples from the first experiment were omitted, as due to the breakdown of the cooling system in the glasshouse deviant temperatures and an altered developmental speed might have biased the data. Subsequently, three aliquots of 15 mg (± max. 1.5 mg) fresh weight were generated by the dispensing system and filled in 1.4 ml Micronic storage tubes for extraction of polar metabolites. From the very same pooled material two 50 mg (± max. 1.5 mg) fresh weight aliquots were generated manually for subsequent total RNA extraction. Material was stored at -80 °C until use. Total RNA and polar metabolites were isolated from the same material.

2.6. Metabolite profiling in early vegetative tissue

The metabolite profiling analyses were performed in collaboration with Dr. David Riewe, a former colleague from the IPK who performed the annotation of polar primary metabolites and subsequent initial normalization procedures. The analytical work was performed jointly.

2.6.1. Extraction of polar leaf metabolites

Metabolite extraction was conducted in six batches with 96 samples each, via a previously described liquid–liquid extraction protocol (Lisec *et al.*, 2006; Riewe *et al.*, 2012, 2016) that was adjusted to 96 tubes / rack format and smaller volumes. The protocol was implemented on a liquid handling system (Biomek® FXP, Beckman Coulter GmbH, Krefeld, Germany; see Figure S10).

Polar metabolites were extracted from 15 mg deep frozen homogenised plant material in 0.625 ml chilled extraction buffer (2.5:1:1 v/v MeOH/$CHCl_3$/H_2O; plus internal standards L-Valine-d_8, 98 atom %, Campro Scientific GmbH, https://campro-webshop.eu and L-Alanine-2,3,3,3-d_4, 98 atom %, Isotec Inc., https://www.sigmaaldrich.com, each 1:1,000). Sample racks were inverted, briefly shaken using a vortex mixer, and incubated under shaking conditions for 15 min using a mixer mill (RETSCH, 20 Hz, 4 °C), followed by ultrasonication for 15 min at 4 °C. After the incubation, samples were briefly centrifuged and 250 µl of water were added. Samples were inverted, again briefly shaken, centrifuged at 2,000 g for 5 min, and sealed with aluminium foil and 'Parafilm M' to avoid evaporation during the automated aliquotation. Three aliquots of 50 µl of the supernatant of the upper polar phase were transferred into glass vials (CZT Klaus Trott, Kriftel, Germany). The polar phase was dried in a vacuum concentrator, vials were filled with argon, capped, and stored in sealed plastic bags containing silica desiccant at -80 °C.

2.6.2. Gas chromatography – mass spectrometry analyses (GC-MS)

Aliquots of the polar phases were in-line derivatised directly prior to injection according to Erban *et al.* (2007) in a Gerstel MPS2-XL autosampler (Gerstel, Muehlheim/Ruhr, Germany) and analysed in split mode (1:4) using a LECO Pegasus HT time-of-flight mass spectrometer (LECO, St. Joseph, MI, USA) hyphenated with an Agilent 7890 gas chromatograph (Agilent, Santa Clara, CA, USA) as previously described by Riewe *et al*. (2012, 2016). Samples were analysed in four larger batches and blocks of 20 samples. A total of 27 quality control pools (pooled material from all samples), 36 carrier replicates (pools of plants from one individual carrier) and 8 negative controls (extraction procedure with empty vials; 'blanks') were included in the analysis for quality control. Analyte mass spectra were deconvoluted using the LECO ChromaTOF software including the Statistical Compare package, and peaks were annotated by querying the electron impact spectra library provided by the Golm Metabolome Database (GMD, http://gmd.mpimpgolm.mpg.de). Quantitative peak information was extracted using the R and the 'TargetSearch' package (Cuadros-Inostroza *et al.*, 2009). After filtering for contaminations (sample to blank ratio > 2) and redundant analytes, 154 analytes of tentative biological origin, 64 of known and 90 of unknown chemical structure, were quantified.

2.6.3. Normalization of metabolite data

Metabolite intensities were normalised regarding sample weight, measurement day and median of the respective metabolite per analysed batch. Metabolite intensities were not normalised using the internal standards as this increased the standard deviation in the pooled samples. Furthermore, outliers were removed (median ±4× SD), and metabolite data were power transformed to ensure an approximate normal distribution (Box and Cox, 1964). The full list of annotated metabolite peaks (raw data and processed data) is provided as Data S4. A principal component analysis (PCA) was performed on the centred and scaled metabolite data using the pca function in R using the 'pcaMethods' package (Stacklies *et al.*, 2007).

2.7. Transcriptome analyses

The following analyses were performed in collaboration with Dr. Axel Himmelbach, who is in charge of the IPK sequencing facility, and with Prof. Dr. Andrea Bräutigam, the former group leader of the 'Network analysis and modelling' research group who performed the initial mapping of transcript sequences from the pilot experiment with Kallisto, the analysis of differentially expressed genes (DEGs), the pathway analysis using MapMan, and advised on data analyses.

2.7.1. RNA-extraction and quality assessment

To optimise the extraction procedure of total RNA from rapeseed shoot tissue, extraction tests with various commercially available kits (GeneJET RNA Purification Kit, RNeasy Plant Mini Kit, The InviTrap Spin Plant RNA Mini Kit, NucleoSpin RNA Plant Kit and the SurePrep Plant/Fungi Total RNA Purification Kit) and two extraction protocols (TRIzol method / Hot Borat method) were performed and the quantity and quality of extracted RNA was evaluated. Four replicates per method were extracted with 50-100 mg of pooled plant material (Achat & Campino) as input. Quality and quantity of extracted total RNA was assessed using a NanoDrop Spectrophotometer, an RNA-agarose gel (1.5 % Agarose; Figure S4), Qubit 2.0 and Bioanalyzer measurements.

For the main experiments, total RNA was isolated from each sample (aliquots of the same pooled shoot material as used for the metabolite extraction) using the GeneJET Plant RNA Purification

Mini Kit (Thermo Fischer Scientific Inc., Waltham, USA) according to the manufacturer's protocol and eluted in 50 µl nuclease-free water. The ERCC RNA Spike-In Mix (Thermo Fischer Scientific, Waltham, Massachusetts, USA) was added as internal standard with 2 µl in a 1:100 dilution into the extraction buffer for subsequent normalization. One µl of total RNA was used to initially assess quantity and quality after extraction using both, a NanoDrop ND-1000 and a NanoDrop One Microvolume UV-Vis Spectrophotometer. The ratio of absorbance at 260 nm and 280 nm was used to assess the purity of RNA. A ratio appreciably lower than 2.0 might indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. The ratio of absorbance at 260 nm and 230 nm was used as a secondary measure to assess RNA purity. Extraction of samples with a 260/280 ratio or 260/230 ratio < 2.0 and/or a low concentration (< 500 ng/µl) was repeated. Subsequently, 10 µl aliquots of each RNA sample were taken and extracted total RNA was stored at -80 °C. These aliquots were split and diluted for further quality checks and precise fluorometric quantification using the Quant-iT™ RNA Assay Kit, broad range (Thermo Fischer Scientific, Waltham, Massachusetts, USA) adapted to plate format, according to manufacturer's instructions and a TECAN plate photometer. RNA integrity number (RIN) was checked for each fourth RNA sample. To this end, RNA was diluted (volume of 5 µl; concentration of 300 ng/µl) and 1 µl was analysed using a Bioanalyzer 2100 following the manufacturer's instructions. All tested RNAs had a RIN > 8.0. After final quantification and quality assessment frozen RNA stocks were thawed and diluted (volume of 30 µl; concentration of 100 ng/µl). Randomised samples in 96 well plates were forwarded to the IPK sequencing facility.

2.7.2. RNA-sequencing and data analysis

RNA-sequencing was performed by the IPK's internal sequencing facility with two pilot experiments and a partitioned main experiment. For the first pilot experiment, six TruSeq RNA libraries (Illumina, San Diego, USA) were generated. For the second pilot and the main experiment, cDNA libraries were constructed using the Lexogen SENSE mRNA-Seq Library Prep Kit V2 (Lexogen GmbH, Vienna, Austria). All were sequenced using 100 bp single end (SE) reads on a HiSeq 2500 platform (Illumina, San Diego, USA). For the different runs, either 20 or 96 samples were multiplexed per flowcell lane aiming for an output of approximately 7 million reads per sample. For samples with a low number of reads, additional sequencing runs were performed or new

libraries were generated. In total, 27 flow cell lanes on 5 (high-output and rapid-run) flow cells were sequenced, resulting in a total of approximately 520 Gbases / 4.8 billion single-end reads. Adapter-trimmed raw reads were obtained from the sequencing facility and were further quality trimmed using the Trimmomatic software v0.36 (Bolger *et al.*, 2014) with the following options: SE, HEADCROP:6, LEADING:20, TRAILING:20, SLIDINGWINDOW:4:15, and MINLEN:50. Sequence quality (Phred score) was exemplarily assessed for raw and trimmed data by the fastQC software (Figure S5). The HEADCROP option of Trimmomatic was only applied for reads obtained from the Lexogen libraries due to reduced base quality (Phred score) at the 5'-end of the reads. Read normalization using the internal standard (ERCC Spike-in Mix) was not performed, as it decreased the quality of the data set. For the first pilot experiment, adapter sequences were removed by cutadapt v1.8.1 (Martin, 2011) and reads trimmed and mapped with clc-assembly-cell v4.3.0 (CLC bio, Aarhus, Denmark) to the ERCC Spike-in Mix sequences. For the second pilot experiment, reads were mapped with Kallisto (Bray *et al.*, 2016) to the coding sequence of *B. napus* (CDS, Darmor-*bzh* v4.1) using default settings. Differentially expressed genes (DEGs) were identified using R and the 'edgeR' package (Robinson *et al.*, 2010) with Bonferroni correction ($q < 0.01$). Reads of the main experiment were exemplarily mapped to ribosomal DNA sequences (5S, 5.8S, 18S, and 25S rRNA gene sequences of *Arabidopsis thaliana* obtained from the TAIR10 database) and organellar genomes (mitochondrial [NC_008285.1] and chloroplast genomes [NC_016734.1] obtained from the NCBI database).

Trimmed high-quality sequences were concatenated and aligned to the Darmor-bzw NRGene reference assembly (.fasta) file using Hisat2 v2.0.4 (Kim *et al.*, 2015) using default settings. The assembly and annotation data were provided by the project partners (Lab of Prof. Dr. Rod Snowdon) at the Justus-Liebig University (JLU) Gießen. On average, 70-85 % of the reads could be mapped to the reference genome. The mean overall alignment rate was 82.2 %, whereby 67.3 % of the reads could be uniquely aligned and 14.9 % multiple times. Counting of features was performed using HTSeq software v0.6.1p1 (Anders *et al.*, 2015) and the NRGene annotation (.gff3) file with information about 126,667 annotated genes and the following settings: -t exon, -i Parent, and -s no. In the sampled shoot tissue approx. 78 % of the transcripts were detectable (> 0 counts in at least one sample) with the applied sequencing depth at the examined time point 14 DAS.

Subsequently, raw counts were normalised for sequencing depths and transcript length using the 'tpm' procedure in R statistical software. 15.4 % and 9.2 % of the transcripts could be detected at a level of more than 5 and 10 tpm (median across all samples), respectively. Filtered tpm data (cutoff: median tpm ≥ 5) was transcript-wise filtered for outliers (median ±4× SD) was used for follow up analyses. Filtered tpm data (Data S13) were subjected to GWAS analyses and predictions. A principal component analysis of transcript data was performed on the centred and scaled tpm data in R using the pca function of the 'pcaMethods' package (Stacklies *et al.*, 2007).

2.7.3. Gene network inference with ensemble of trees (GENIE3)

A gene regulatory network (in the form of a weighted adjacency matrix) was inferred from expression data, using ensembles of regression trees. For this purpose, the GENIE3 function from the eponymous R package (Huynh-Thu *et al.*, 2010) was utilised. The function was parallelised using 60 cores on a Linux server. All transcripts (median tpm ≥ 5) in the form of an expression matrix (gene x samples) were treated as candidate regulators and as targets were restricted to biomass (FW) for which potential regulators were calculated. Random Forests (RF) was selected as tree-method and the number of trees in an ensemble the target was set to 10,000. The number of candidate regulators randomly selected at each tree node (for the determination of the best split) was the square root of the total number of candidate regulators (default setting). As output a weighted adjacency matrix of the inferred network is obtained. In this particular case, the weighted adjacency matrix was a vector with gives the importance of the link from regulatory gene to the target trait (FW). This vector with the weights of the regulatory links was further filtered for the 100 top-ranked links.

2.7.4. GO term enrichment analyses

Gene ontology (GO) term enrichment was performed using the 'agriGO' webtool (http://bioinfo.cau.edu.cn/agriGO/) with the option singular enrichment analysis (SEA) and default settings. The top loadings with GO terms and an absolute value ≥ |0.02| from the 4th principal component (PC) of the principal component analysis (PCA), which separates the genotypes by biomass (n= 152) were extracted. The filtered subsets of negative (n= 74) and positive

(n= 78) loadings, and a customised annotated *Brassica napus* reference used as input files (Data S14). Gene ontology (GO) information was obtained from Genoscope (http://www.genoscope.cns.fr/brassicanapus/). Results were visualised with the same tool, using the display category 'Biological Process', a minimal number of mapping entries (n= 5), the statistical test method 'hypergeometric', the multi-testing adjustment method 'Yekutieli (FDR under dependency)' and a significance level of 0.01. The top 100 candidates for FW and DW, respectively, identified by the GENIE3 network analysis, were extracted and duplicated entries were filtered. The resulting 69 unique genes with GO terms (Data S14) were subjected to a GO term enrichment analysis using the 'agriGO' v2.0 (http://systemsbiology.cau.edu.cn/agriGOv2/) with the same settings as described for the loadings of the 4th PC.

2.8. Genotype data

Genotype calls, copy-number variations and the reference genome with annotations were kindly provided by the project partners from the Justus-Liebig University (JLU) Gießen, Dr. Fabian Grandke, Dr. Birgit Samans and Prof. Dr. Rod Snowdon.

2.8.1. Reference genome and gene annotations

To ensure the unique positioning of as many markers as possible, an enhanced version of the *Brassica napus* cv. Darmor v4.1 reference genome assembly (Chalhoub *et al.*, 2014) was used, generated by incorporating long read information (NRGene, DeNovoMAGIC™; unpublished data from Prof. Dr. David Edwards, University of Western Australia) into the pseudomolecules. Genes were predicted *de novo* using a MAKER pipeline with AUGUSTUS and SNAP. Subsequently the transcriptome was annotated by mapping the transcript sequences on: 1) the *B. napus* Darmor-*bzh* v.4.1, 2) a concatenated *Brassica* AC-genome assembly comprising the *B. rapa* v1.5 (Wang *et al.*, 2011*b*) and the *B. oleracea* TO1000 (Parkin *et al.*, 2014), and 3) the *A. thaliana* TAIR10 transcriptomes, respectively, using the basic local alignment search tool (BLASTn). Transcripts were counted as hit if they reached a minimum similarity of 80 % over 40 % of the target transcript. If annotations in multiple genomes were obtained, they were prioritised in the order *B. napus*, *B. oleracea / rapa*, *A. thaliana*. The *B. napus* Darmor-*bzh* v.4.1, *B. rapa*

Chiifu-401–42 and *B. oleracea* TO1000 transcriptomes were functionally annotated using Blast2GO (Conesa *et al.*, 2005) version 3.0.8 with default settings. The Arabidopsis TAIR transcriptome annotations were downloaded from the TAIR homepage (https://www.arabidopsis.org/).

2.8.2. Array data analysis and calling of genotypes

The 477 genotypes (pollinators and male-sterile lines) were genotyped using the *Brassica* Infinium 60k genotyping array that was developed based on *Brassica oleracea* and *rapa* genome sequences (Illumina Inc., San Diego, CA; USA), as described previously by Jan *et al.* (2016). Raw data were initially filtered to exclude SNPs without positional information in the *Brassica rapa* and *Brassica oleracea* genomes. SNP genotypes were called using R and the 'gsrc' package (Grandke *et al.*, 2017). Subsequently, probe oligonucleotide sequences were mapped to the enhanced *Brassica napus* cv. Darmor reference genome assembly using the basic local alignment search tool (BLASTn) with 95 % similarity over a length of 50 bp. Markers showing multiple BLASTn hits in the genome were removed. For genome-wide association studies (GWAS) SNPs were coded in numerical format (0=AA, 1=AB, 2=BB) using the 'GAPIT' R package (Lipka *et al.*, 2012; Tang *et al.*, 2016). Furthermore, markers with minor allele frequencies (MAF) smaller than 0.01 and markers with more than 10 % missing values or more than 25 % heterozygous calls were removed.

To identify copy-number variations (CNVs), the SNP positions together with the signal intensity values were used to define blocks of similar intensity. If the blocks' values exceeded the applied thresholds, they were classified as deletions or duplications. This set of copy-number variations, also generated with the R package 'gsrc', complemented the obtained SNP data. CNVs were included as 0=normal, 2=deletion or duplication. Deletions and duplications were separately tested against the normal state and the other events (reciprocally, either the duplications or the deletions) were treated as missing values. For CNVs, positions were shifted by ±1 bp to avoid identical marker positions. A total of 16,311 markers comprising 13,201 unique, single-copy SNPs, 3,106 deletions and 4 duplications remained after filtering and were used for subsequent analyses (Data S5).

2.8.3. Analysis of population structure

Population structure was analysed using the programme STRUCTURE, version 2.3.4 (Pritchard *et al.*, 2000), marker data of all 477 lines and the 'admixture' model. Population clustering for K= 1 to 10 was performed with a burn-in period of 10.000, 10.000 MCMC replications and three iterations per K. The lambda parameter was inferred and adjusted to λ= 0.304. The mean Ln probability [L(K)] and population clustering for *K*= 2 to 5 is shown in Figure S6.

2.8.4. Analysis of linkage disequilibrium (LD) and decay

Pairwise linkage disequilibrium (LD) was analysed for each chromosome in R using the 'LDheatmap' package (Shin *et al.*, 2006) for the SNP marker data across all 477 canola lines. LD-decay was calculated in R for both subgenomes separately (Hill and Weir, 1988; Remington *et al.*, 2001; Marroni *et al.*, 2011), as considerable differences between the A and C subgenome have previously been reported (Wu *et al.*, 2016*a*).

2.9. Genome-wide association studies (GWAS)

Recently, a new method for genome-wide association studies, FarmCPU (Fixed and random model Circulating Probability Unification) has been proposed by Liu *et al.* (2016), which controls false positives and effectively reduces false negatives. The method iteratively performs marker tests with pseudo quantitative trait nucleotides (QTNs) as covariates in a fixed effects model and optimization on pseudo QTNs in a random effects model. To some extent, this process is capable to remove the confounding between testing markers and kinship, to prevent overfitting of the model, and to control false positives simultaneously. Genome-wide association analyses (GWAS) were conducted in R version 3.4.3 'FarmCPU' on BLUEs of the traits of the 477 canola lines using the filtered set of 16,111 numerically coded SNP (n= 13,201) and CNV (n= 3,110) markers. Analyses were performed in Rstudio on a CentOS 7.2 Linux server (HP ProLiant DL580 Gen9 with HP D3600 Array, 4x Intel Xeon E7-8880v3@2.3 GHz processors, 144 cores, 1TB RAM, 2x480 GB SSD, 2x 600GB SAS, 12x 8TB SAS). As the programme does not allow for missing marker information in the numeric genotype input file, missing data were replaced by heterozygous

states. Kinship was calculated using the FARM-CPU algorithm. Principal component analysis was performed on the centred genotype data using the pca function of the 'pcaMethods' package in R (Stacklies *et al.*, 2007). The first ten principal components (PCs) were calculated and the first four PCs were included into the GWAS model to correct for hidden population structure. The maxLoop parameter was increased to 100 and the optimal threshold for *p*-value selection of the model in the first iteration was estimated by the FarmCPU.P.Threshold function and set to 0.00001 for all traits. Subsequently, *p*-values of marker-trait associations were adjusted for multiple comparisons using FDR (Benjamini and Hochberg, 1995). Only associations with adjusted *p*-values smaller 0.1 were considered as statistically significant and used for further analyses. The phenotypic variance explained (PVE%) by a significant marker was estimated in R (Eq. 6). The sum of squares (SS) and residuals (e) were extracted from the ANOVA fitted with a linear model incorporating the phenotypic values and all significant markers in decreasing order of their *p*-value.

Eq. 6: $$PVE\%_{\,sig.marker} = (SS_{\,sig.marker} \,/\, SS_{\,all\,sig.markers} + e) * 100$$

2.10. Co-localisation of associations and permutation analyses

QTL from each omics-layer were binned in overlapping intervals of 1 Mb to identify potential clustering and QTL hotspots. In addition, co-localisation of associations for metabolites, transcripts and combined image-derived traits and end-point biomass were investigated. In a stringent approach, associations were treated as co-localised when exactly the same marker was reported to be significant for at least two traits. As the number of markers was limited and the number of detected associations (metabolites, n= 206, transcripts, n= 26,391, image-derived traits, n= 4,613; at $p\text{-value}_{(FDR)} \leq 0.05$; with PVE ≥ 2 %) was higher than the number of markers (n= 16,311), a co-localisation of traits by random chance is expected. To estimate the degree of random co-localisation, permutation analyses were performed, distributing the detected associations randomly to all markers. This procedure was repeated 10,000 times and the maximum number of co-localised traits extracted from each iteration. From the obtained distribution, the 95 % quantile was calculated and compared with the actual number of co-localisations detected for each of the three significance levels tested ($p\text{-value}_{(FDR)} \leq 0.05$, ≤ 0.01 and ≤ 0.001).

2.11. Correlation analyses between data sets

Pearson correlations between data sets (phenotypic traits of the parental lines and the hybrids, parental metabolite levels, parental transcript expression levels, agronomic traits of the hybrids, and calculated mid- and best-parent heterosis) were performed using the cor.test function from the 'stats' package, followed by a multiple testing correction of *p*-values using FDR procedure (p.adjust function from 'stats' package) in R. Pearson correlation analyses were parallelised using 60 cores on a Linux server.

2.12. Regions of interest and identification of candidate genes

Candidate gene regions were defined as LD blocks harbouring a significant trait-associated marker in which flanking markers had strong LD ($r^2 > 0.6$), and were extended to the left and right unrelated marker, respectively. All genes within the respective LD-block were considered for candidate gene identification. For significant markers outside of LD blocks, the 100 kb flanking regions on either side were searched for candidate genes as suggested by Zhou and Han *et al.* (2017). Candidate genes were prioritised according to their annotation and gene ontology (GO). A comprehensive list of all genes within the intervals and selected candidate genes for all evaluated traits can be found in Data S6.

2.13. Genomic and omics-based prediction models and model evaluation

Two types of prediction methods were employed: first, (genomic) best linear unbiased prediction (gBLUP; Eq. *7*; Whittaker *et al.*, 2000; Meuwissen *et al.*, 2001), which only considers additive marker effects using a marker-based relationship matrix, and second, reproducing kernel Hilbert space regression (RKHS; Eq. 7; Gianola and van Kaam, 2008), a non-linear regression model which captures both the additive and non-additive effects using a marker-based distance matrix. The gBLUP as well as the RKHS models were implemented using R (R Core Team, 2019) and the mmer2 function from 'sommer' package (Covarrubias-Pazaran, 2016) to solve the mixed model equations. For RKHS, two more packages were employed: the 'AlphaMME' package (https://bitbucket.org/hickeyjohnteam/alphamme) to transform Euclidean distance-matrices into

Gauss-matrices, and the 'rrBLUP' package (Endelman, 2011) to estimate the corresponding tuning (bandwidth) parameters.

The general gBLUP model was defined by (Eq. 7) in which y is an n × 1 vector of phenotypic values (BLUEs), n the number of hybrids, μ a vector of fixed effects that represent the overall mean. In the model g_α, g_β and g_γ are n × 1 vectors of random effects, and Z_α, Z_β and Z_γ are the design matrices assigning genetic values to hybrids using markers (g_α), transcripts (g_β), and metabolites (g_γ), respectively. Marker-based genetic values, transcriptomic values and metabolic values of the hybrids were modelled as random effects with $g_\alpha \sim N(0, G_\alpha \sigma_\alpha^2)$, $g_\beta \sim N(0, G_\beta \sigma_\beta^2)$ and $g_\gamma \sim N(0, G_\gamma \sigma_\gamma^2)$, respectively. The term σ_α^2 denotes the genomic variance estimated using SNP markers, σ_β^2 the transcriptomic variance and σ_γ^2 the metabolic variance and G_α, G_β and G_γ were the realised additive relationship matrices calculated based on VanRaden (2008). The residuals *e* follow a normal distribution $e \sim N(0, I\sigma^2)$, where I is the identity matrix.

Analogously to the previously described approach, the general statistical model for Reproducing Kernel Hilbert Space Regression (RKHS) was defined by (Eq. 7) where $g_\alpha \sim N(0, K_\alpha \sigma_\alpha^2)$, $g_\beta \sim N(0, K_\beta \sigma_\beta^2)$ and $g_\gamma \sim N(0, K_\gamma \sigma_\gamma^2)$ are random effects meassured by the genetic markers, transcriptome and metabolome data, respectively. K_α, K_β and K_γ denote the Gaussian Kernels based on SNP, transcriptomic- and metabolic markers, respectively. For the RKHS regression, marker matrices were first transformed into Euclidean distance matrices. Gaussian Kernels were subsequently calculated using Euclidean distance matrices between individuals based on the respective marker types and a bandwidth parameter *h*. The bandwidth parameters were estimated from the respective log-likelihood profile generated using the kin.blup function of the 'rrBLUP' R package.

Eq. 7: $$y = 1\mu + Z_\alpha g_\alpha + Z_\beta g_\beta + Z_\gamma g_\gamma + e$$

Predictions were performed for the seven agronomic traits (BLUEs) calculated for the set of 950 hybrids. Three sets of predictors (G = genomic, T = transcriptomic and M = metabolic data) were generated for the parental lines (475 pollinators and 2 male-sterile testers). Parental lines were 'crossed' *in silico* by combining the two respective parental matrices to extrapolate hybrid profiles (Werner *et al.*, 2017). The resulting matrices had the dimension 'number of hybrids' (n= 950) times

'number of features' (n_G = 13,201, n_T = 19,479, n_M = 154). Columns of the predictor matrices were centred and standardised to result in values between 0 and 2 (G) and 0 and 1 (T and M), respectively. Predictions were performed for each set of predictors (G, T, M) separately, and in combinations of two or three (G+T, G+M, T+M; G+T+M). A cross-validation (cv) scheme with 100 cv-cycles was applied, separating the data set in a training set (75 %) and a validation set (25 %), where the phenotypes (BLUEs) of the validation set were masked and predicted. Prediction accuracies were obtained as average Pearson product-moment correlation coefficients between predicted ($\hat{y}$) and observed phenotypes (y).

2.14. Hybrid performance and heterosis

Hybrid performance and heterosis, mid-parent heterosis (MPH) and best-parent heterosis (BPH), respectively, were examined and compared with regard to group assignment of the respective male-sterile parental tester (MS1 and MS2). MPH was calculated as difference between hybrid performance (F_1) and the mean value of the two parents [MP = (P1 + P2) / 2] for end-point biomass, projected leaf area and estimated biovolume at all time points with available data (Eq. 8). BPH was calculated as difference between hybrid performance (F_1) and the better performing parent (Eq. 9). To investigate potential heterotic patterns between parental components from the two different subsets, within-group and between-group crosses were tested for significant differences using a Student's *t*-test.

Eq. 8: $$MPH = \frac{(F1-MP)}{MP} \times 100$$

Eq. 9: $$MBH = \frac{(F1-BP)}{BP} \times 100$$

3. Results

The work presented here pursued three main objectives. On the one hand, the predictive power of genetic, molecular (transcriptome) and biochemical (metabolome) markers / factors and their combinations in hybrid performance prediction models should be evaluated. Data for a population of 950 hybrids that had been evaluated in the field for agronomically important traits like seed yield, seed oil content and flowering time was available for this work. The 477 parental lines of the hybrids should be grown under controlled conditions in a climatized glasshouse and data from different -omics layers should be gathered and used individually and in combination in the predictive models. On the other hand, genome-wide association studies (GWAS) should be carried out using the extensive -omics data sets to identify candidate genetic loci and genes for quantitative traits. In addition, co-localisation and correlation studies should be performed to elucidate links between vegetative growth, gene expression and metabolite levels.

3.1. Generation of -omics data sets

Complementarily to the field data for the set of 950 hybrids, extensive data sets were generated for the parental lines, 475 diverse pollinator lines and two male-sterile testers, which had been used to create the hybrids. Array-derived genotype data (SNPs and CNVs) for the parental lines were provided by the project partners from the Justus-Liebig University (JLU) Gießen. The genotype information was complemented by extensive -omics data sets including global transcriptome (RNA-Seq) profiles, polar primary metabolite (GC-MS) profiles, as well as detailed image-derived phenotyping data. The following chapter describes the generation of the individual data sets and gives an overview about the scale and complexity of the phenotypic and the molecular data.

3.1.1. Field experiments and statistical evaluation of agronomic traits

Field trials were performed during the 2012 growing season by the commercial partners NPZi and DSV at plant breeding testing sites. An F_1 hybrid population with 950 genotypes was generated by the commercial partners by crossing two male-sterile testers (MS1 and MS2) with the 475

genetically diverse pollinators that could be attributed to three breeding pools (denoted as breeding pools 1, 2, and 3). The resulting 950 hybrids were evaluated in nine trials at eight different locations across Denmark, Germany, Poland, Latvia, and Estonia, whereby each line was replicated four times in the nested design with 35 trial ~ location combinations. Seven traits of agronomic / commercial importance were assessed: the content of total seed glucosinolates (GSL; µmol/g seed), the days to onset of flowering (DTF; measured as number of days from sowing until 50 % of the plants per plot flowered), seed oil yield (dt/ha), seedling emergence (visual observation ranging from a minimum value of 0 to a maximum of 10), seed oil content (% volume per seed dry weight), seed protein content (% volume per seed dry weight) and seed yield (dt/ha). However, none of the traits with the exception of seed yield was scored at all locations. Calculation of adjusted values (BLUEs) was necessary, as in particular raw data for the traits seed oil yield, DTF and seedling emergence showed a bimodal distribution due to location effects. The BLUEs of all seven traits followed an approximate normal distribution (Figure 2), but due to missing data BLUEs could be calculated only for 929 of the 950 hybrids. The coefficients of variation ranged from 0.84 % for DTF to 20.82 % for total seed GSL content, which was the phenotypic trait (BLUEs) with the highest variability. Broad sense heritability values (H^2) for all traits were estimated across the different trials / field locations. H^2 values ranged from 0.34 for the trait seedling emergence to 0.92 for total seed GSL content (Table 1).

Table 1. Summary statistics for agronomic traits evaluated in field trials at eight locations

Trait [a]	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	CV (%)	H^2 (%)
Days to onset of flowering (DTF)	167.34	170.11	171.05	171.18	172.05	176.20	0.84	85
Seedling emergence (good = 9 to 10)	4.10	6.10	6.41	6.39	6.68	8.96	7.33	34
Seed GSL (µmol/g)	4.43	7.77	9.13	9.15	10.42	17.23	20.82	92
Seed oil yield (dt/ha)	9.85	13.72	14.44	14.34	15.07	17.74	7.47	82
Seed oil content (%)	43.81	47.25	48.20	48.22	49.13	53.03	2.92	90
Seed protein (%)	18.17	20.86	21.51	21.48	22.18	24.30	4.72	82
Seed yield (dt/ha)	23.22	29.92	31.13	31.04	32.29	37.64	5.89	62

[a] Best linear unbiased estimators (BLUEs) were calculated across the field trials conducted at eight different locations across Europe in 2012

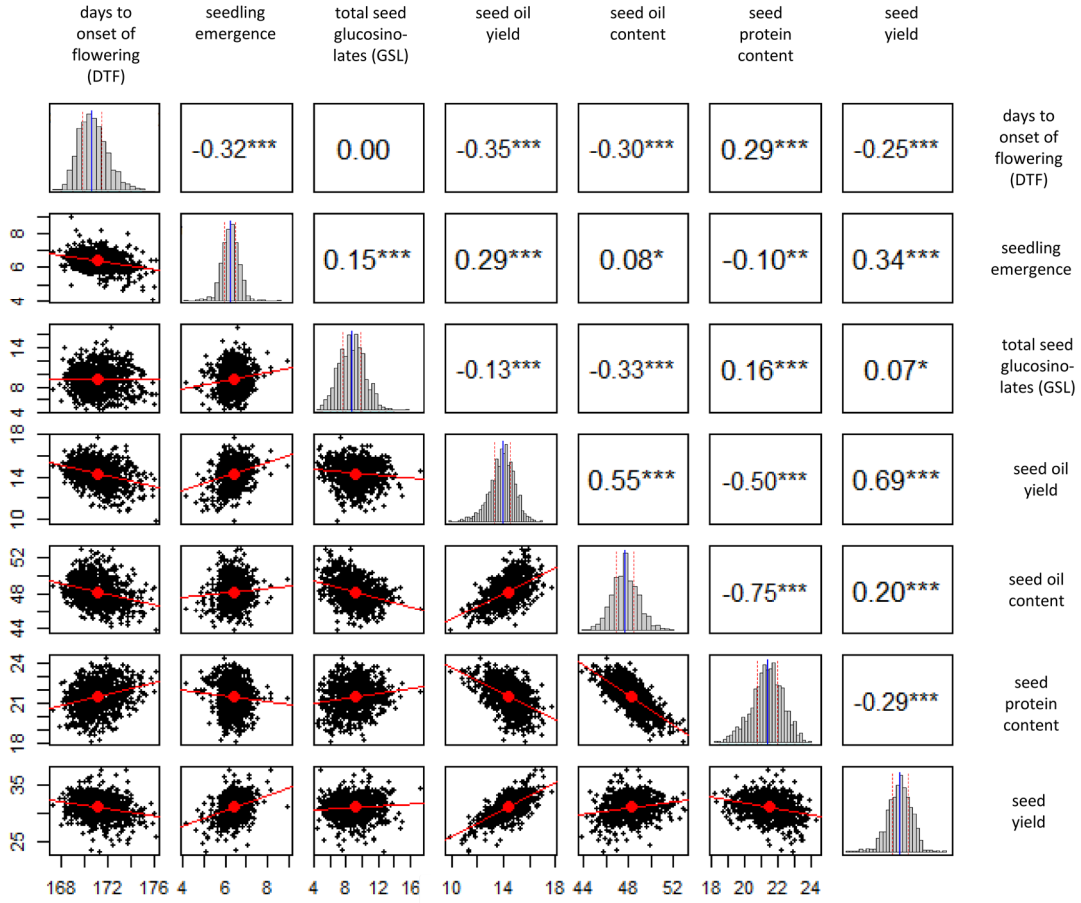


Figure 2. Overview of agronomic trait correlations

The matrix panel plot shows an overview of the agronomic traits analysed in the field trials. Shown are BLUEs of the seven agronomic traits. The upper triangle displays the Pearson correlation coefficients and significance of the correlations (alpha: * = 0.05; ** = 0.01 and *** = 0.001). The lower triangle displays the corresponding bivariate scatter plots of the relations. The red dot and the red line correspond to the ellipse centre point and the linear regression fit. The diagonal displays the histograms of the trait distribution. The blue solid line and the dashed red lines correspond to the median and the 1st and 3rd quantile of the data distribution, respectively.

3.1.2. Genomic data, copy number variations and population structure

Genotyping of the 477 parental lines was performed on the *Brassica* 60k SNP Infinium consortium array (Illumina Inc., San Diego, CA; USA) as described previously by Jan *et al.* (2016). Beyond single nucleotide polymorphisms (SNPs), copy number variation (CNV) and presence-absence variation (PAV) can provide complementary genetic information (Stein *et al.*, 2017). SNPs and CNVs were called in a combined approach from the array data as previously described (Grandke *et al.*, 2017). A total of 16,311 markers comprising 13,201 unique, single-copy SNPs, 3,106 deletions and

4 duplications (Data S5) were jointly used in the subsequent genome-wide association study. Some of the genotypes exhibited a high degree of heterozygosity (median= 11.2 %; max= 61.8 %). Pairwise marker LD-matrices (R^2) were calculated individually for each chromosome. LD-decay was derived for both subgenomes (A & C) separately, based only on the SNP marker data (Figure S7). The CNV markers were excluded from the calculation as they were not necessarily linked with SNP marker-derived LD blocks. A faster LD-decay was detected in the A subgenome compared to the C subgenome, with half decay values of approximately 400 kb and 3.9 Mb determined for the A and C subgenome, respectively. Multiple larger genomic regions of high LD ($R^2 > 0.6$) were detected, especially on the C-subgenome chromosomes (Data S7).

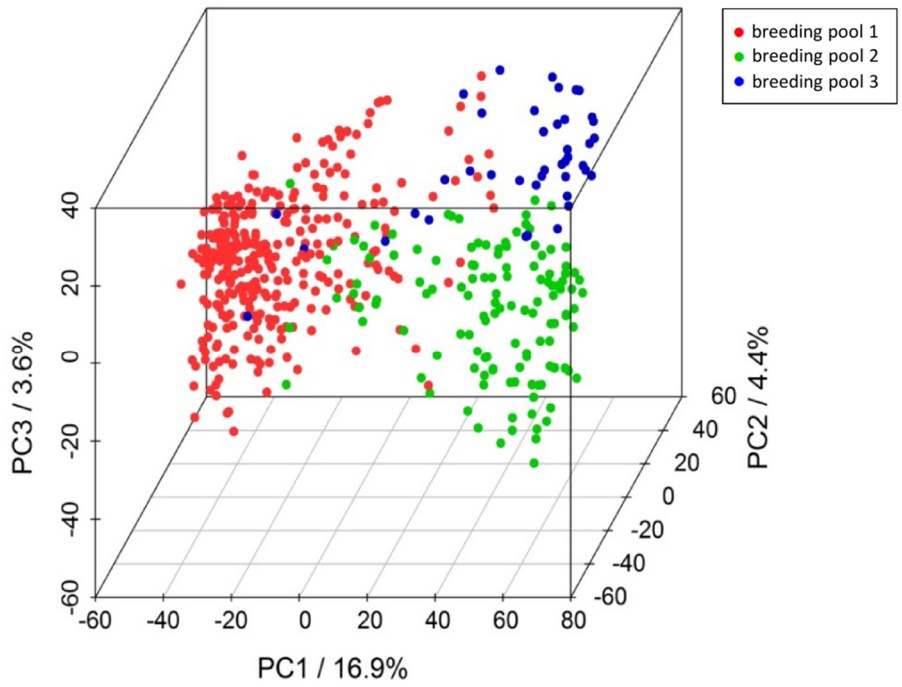


Figure 3. Visualisation of breeding pools by principal component analysis (PCA)

PCA was performed on 477 canola lines using a panel of 13,201 SNP and 3,110 CNV markers. Data were centred and the calculation was done by singular value decomposition (svd) of the data matrix. Proportions of explained variance of principal components (PCs) 1, 2 and 3 are indicated on the axes. Different colours indicated in the key correspond to canola breeding pools from which the investigated lines were selected.

A principal component analysis (PCA) of the population was performed using the combined SNP and CNV data sets (Figure 3). The first ten principal components explain a cumulative variance of approx. 40% (PC1: 16.9 %, PC2: 4.4 %, PC3: 3.6 %, PC4: 3.3 %, PC5: 2.8 %, PC6: 2.0 %, PC7: 1.8 %,

PC8: 1.8 %, PC9: 1.6 % and PC10: 1.5 %). The genotypes cluster into three larger groups corresponding to the three breeding pools. In addition, the STRUCTURE programme (Pritchard *et al.*, 2000) was used to assess the population structure. This analysis indicated the presence of two larger population groups and several potential subpopulations (Figure S6). The first three clusters coincided to a substantial degree with the breeding pools, but many genotypes showed pronounced admixture. As a consequence, the first four principal components, each accounting for more than 3 % of the total variance, were included as covariates into the GWAS analyses, as recommended by the developers of the 'FarmCPU' R package (Liu *et al.*, 2016*a*).

3.1.3. High-throughput phenotyping and image-derived traits

For a period of 21 days (between 6 DAS and 27 DAS) image data were obtained daily from three different camera systems and different views (Figure 4). These camera systems operate using visible light (VIS), static fluorescence (FLUO) and near-infrared (NIR). In total, five phenotyping experiments were performed and approximately 420,000 individual raw images were obtained. In the first four experiments (1413RCM, 1419RCM, 1442RCM and 1447RCM) the 477 parental lines were analysed in an incomplete randomised block design with three replicates per genotype, whereby each replicate comprised one pot with nine plants each. In a fifth experiment (1504RCM) a selection of 120 hybrids, 60 high and 60 low performers according to their seed yield in the field trials, was phenotyped with three replicates (pots with nine plants each) per genotype. As this large amount of data cannot be handled manually, automated image analysis was performed using the IPK Integrated Analysis Platform, IAP Version 2.0.7 (Klukas *et al.*, 2014) implementing a customised pipeline combining top and side view images. A total of 1,194 image-derived phenotypic traits could be derived from this analysis, including 128 (10.7 %) geometric traits giving insights into general plant morphology, 930 (77.9 %) traits related to plant colouration, 104 (8.7 %) traits related to static chlorophyll fluorescence and 32 (2.7 %) traits related to water content and water dynamics (Figure 4). Using a linear mixed model approach, best linear unbiased estimators (BLUEs) for all genotypes across the different phenotyping experiments, as well as estimations of broad-sense heritability (H^2) were calculated. As many of the traits were partially redundant and inter-correlated, a two-step procedure was applied to limit them to a core set for further analyses. Firstly, traits were filtered for broad-sense heritability higher 0.7 for at least one day, reducing the

phenotypic traits to 571. Secondly, stepwise variable selection (VIF) was applied as described by Chen *et al.* (2014) reducing the number of traits to 123 (Figure S8).

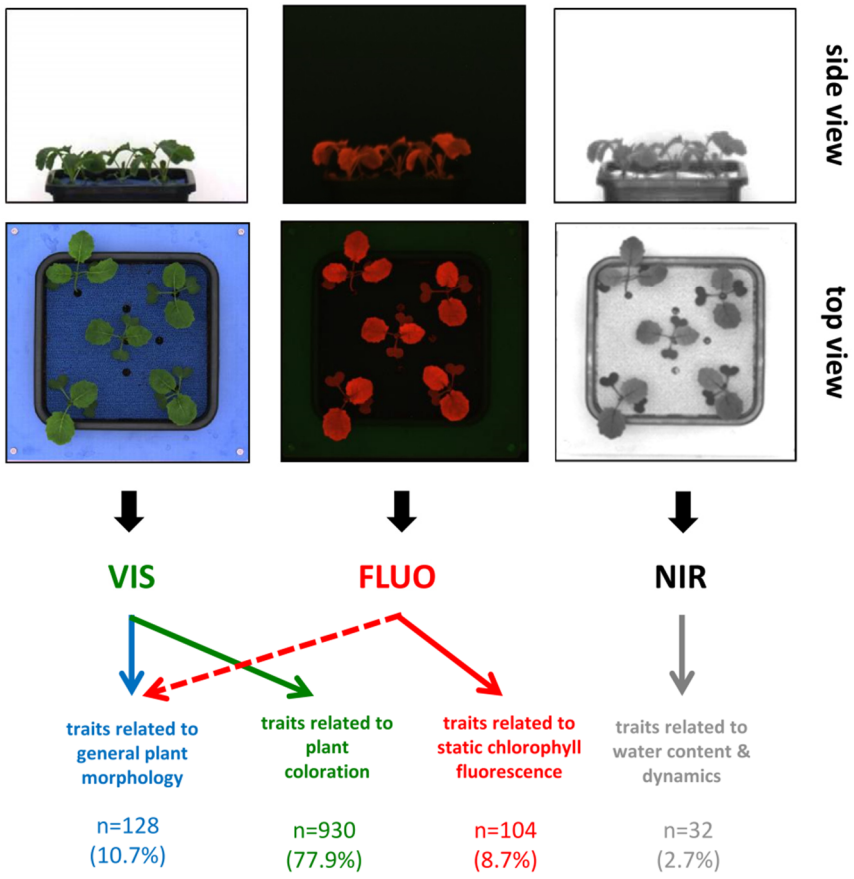


Figure 4. High-throughput phenotyping and image analysis

High-throughput plant phenotyping was performed in the IPK phenotyping platform for large plants during early vegetative growth. For a period of 21 days (6 DAS to 27 DAS) plants were imaged daily. Raw images were obtained from three different camera systems (VIS, FLUO and NIR) from top and side views with different angles. Automated image analysis was performed using IAP Version 2.0.7 (Klukas *et al.*, 2014) and a customised pipeline. 'n' refers to the number of traits analysed.

A subset of 32 traits related to plant growth and biomass including plant height, area, volume and compactness estimates was removed by the stepwise VIF function. As these traits were of particular interest for this study, they were manually retained for further analyses. Four phenotypic traits with high heritability (projected leaf area, estimated biovolume, early plant

height and colour uniformity) were selected for a detailed analysis as they reflect plant growth and biomass accumulation particularly well. In addition, for these four traits relative growth (RGRs) or absolute change rates (ACRs) were calculated over fifteen three-day intervals, for example 8-11 DAS, to assess effects over time, as described in the materials and methods section.

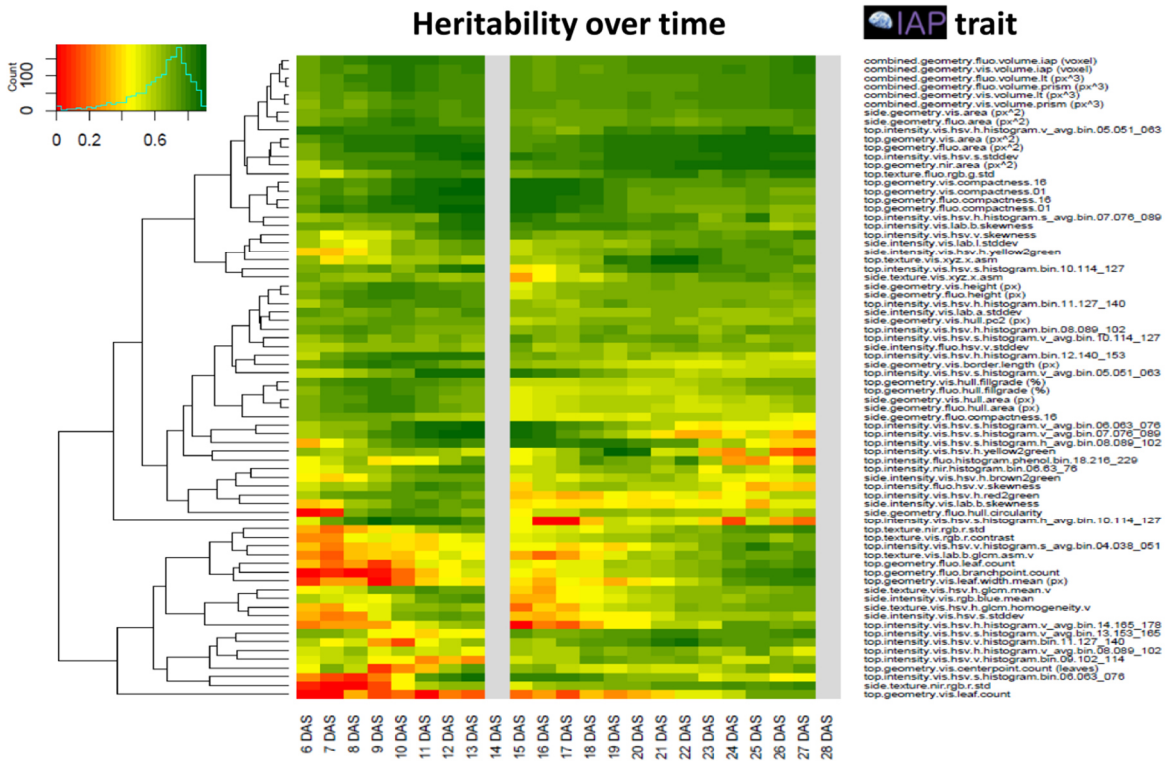


Figure 5. Trait heritabilities at different time points

The heritabilities of individual image-derived traits (H^2) were calculated separately for each time point during the experiment. The colour gradient (red to green) corresponds to H^2 values between 0 and 1. Traits were listed according to a hierarchical clustering with the dendrogram displayed on the left side. Only the subset of 74 traits without missing data is displayed. At 14 DAS and 28 DAS (grey colour) no imaging of the plants was performed, therefore heritabilities were not calculated.

Analysing the change of heritability of individual phenotypic traits over time, in general three types of behaviour can be discriminated (Figure 5). Some traits display an overall high heritability throughout the whole time period (for example projected leaf area, estimated biovolume and compactness), while other traits display a high heritability at early stages and reduced heritability at later stages (for example hull fill grade and the brown to green ratio) and vice versa

(for example branch point count and leaf width). At two time points (14 DAS and 28 DAS) no imaging of the plants was performed. At 14 DAS, the time point when most plants had observable epicotyls, the four inner plants around the central plant (Figure S1 b) were removed and sampled for molecular / biochemical analyses. At 28 DAS the remaining five plants in each pot were harvested to assess shoot fresh and dry weight.

3.1.4. Untargeted metabolome analyses via GC-MS

Global metabolite profiles were recorded by GC-MS analysis from pooled shoot material of three different phenotyping experiments sampled at 14 DAS. The time point of 14 DAS had been determined in a pilot experiment, as material sampled at 14 DAS showed an overall lower variation between genotype replicates and hence potentially more discriminative power than the material sampled at 28 DAS (Table 2). A complete separation of genotypes by their profiles was not achieved as indicated in the PCA plot (Figure S9), but a multifactorial ANOVA indicated a higher number of significantly altered metabolites discriminating genotypes in the earlier sampled material (69 vs. 52 metabolites with *p*-value $_{Bonferroni} \leq 0.05$, at 14 DAS and 28 DAS, respectively). Plant material from three phenotyping experiments was pooled and polar metabolites were extracted using a liquid-liquid extraction protocol (Lisec *et al.*, 2006; Riewe *et al.*, 2012, 2016) that was implemented during this work on a robotic liquid handling system (Figure S10).

In total 154 metabolites, 64 of known and 90 of unknown chemical structure, passed all quality filters (see materials and methods) and were quantified relatively. As expected, in the negative controls without plant material nearly no metabolites were detectable. After removal of potential outliers and Box-Cox power transformation, metabolite data were subjected to subsequent analyses. To assess data quality, a PCA analysis was performed. The first four PC groups contribute 23 %, 8.4 %, 5.8 %, and 4.9 % to the metabolic variance, respectively. Although no separation of groups (breeding pools 1, 2, and 3; quality control pools; negative controls; reference lines) in the first two PC groups was observed (Figure S11 b), the third PC partially separates lines of ‘breeding pool 2’ and ‘breeding pool 3’ from the other samples (Figure S11 c). Comparing the overall metabolite intensities (total ion count) between groups, no substantial differences between the breeding pools could be detected. Quality control pools, which had been included to assess the stability of measurements across the long GC-MS analysis with many samples, display the lowest

metabolic variance among all groups (Figure S11 d) and cluster in the centre of the PCA plots (Figure S11 b-c). The four reference lines ('Achat', 'Campino', 'MS1' and 'MS') were analysed in higher replication as pools of plants from individual pots / experiments. Reference lines displayed no difference in the overall metabolite intensities (total ion count) in comparison to the other samples. However, they indicated substantial metabolic variation between the replicates (data not shown).

Table 2. Standard deviations for the metabolomics pilot experiment

Line	H7	L13	L4	L7	Total
SD at 14 DAS	24.50 %	17.45 %	19.19 %	17.21 %	27.95 %
SD at 28 DAS	27.50 %	27.26 %	27.06 %	21.06 %	32.33 %

Median standard deviations (SD) across all relatively quantified metabolites in percent. Four canola lines (H7, L13, L4 and L7) provided by the commercial project partners (NPZi & DSV) were analysed at two time points, 14 and 28 days after sowing (DAS). Standard deviations are given for the four lines individually and across all genotypes (total). The data shown represent eight replicates per genotype / time point combination.

3.1.5. Transcriptome analyses by RNA-sequencing

In an RNA-Seq pilot experiment, a comparative analysis of four genotypes contrasting in biomass was performed. The pilot experiment was also used to answer the question, whether pooling of samples from the different phenotyping experiments might be feasible. To this end, four genotypes, 'Pol 229' (low biomass), 'Pol 396' and 'Pol 467' (medium biomass), and 'Pol 419 (high biomass) were analysed with three 'replicates' (pools of four plants from a single pot / a single phenotyping experiment) and as one 'pool' (twelve plants pooled from different pots / different phenotyping experiments) for each genotype (Figure S12 a). A principal component analysis indicated one sample as outlier in the first principal component (Figure S12 b). In the second principal component, it was possible to at least partially separate the genotypes. In addition, a hierarchical clustering analysis showed a clustering of genotypes and indicated that the 'pools' and 'replicates' of each genotype cluster together (Figure S12 c). Based on the results of the pilot experiment, it was concluded that it was appropriate to analyse pooled material from different phenotyping experiments.

In the main RNA-sequencing experiment, total RNA was extracted from aliquots of the same pooled shoot material sampled 14 DAS that was used for the analysis of polar metabolites.

After quality and quantity assessment, total RNA was submitted to the IPK sequencing facility for library generation and sequencing using an Illumina HiSeq 2500 System. Combined for all sequenced lines, a total of approx. 520 Gb (approx. 4.8 billion reads) were generated, covering each genotype with on average approx. 9.5 million reads. More than 96 % of all 477 samples could be covered with at least 7 million reads (Figure S5 a). Generated raw data showed an overall good quality as indicated in the plot of the Phred quality scores (a property which is logarithmically related to the base-calling error probabilities; Ewing and Green, 1998) per position in a representative example (Figure S5 b). A slight decrease of read quality (reduced base-quality) could be observed to some degree at the 5'-end and in particular at the 3'-end of reads of the reads compared to TruSeq RNA libraries used in the first pilot experiment. After quality-trimming, data were concatenated for each line and mapped using Hisat2 to the NRGene Darmor-*bzh* reference genome assembly that had been used to anchor the array-derived SNP markers. Overall, 82.2 % of the reads could be aligned to the NRGene Darmor-*bzh* reference genome, 67.3 % of them uniquely, which is approx. 5 % higher than the alignment rate using the published Darmor-*bzh* v4.1 reference genome version, as exemplarily determined using the sequencing read data of Pollinator 211. Reads were exemplarily mapped to ribosomal DNA sequences (5S, 5.8S, 18S, and 25S rRNA gene sequences of *Arabidopsis thaliana* obtained from the TAIR10 database) and organellar genomes (mitochondrial [NC_008285.1] and chloroplast genomes [NC_016734.1] obtained from the NCBI database). On average, 9 % of reads mapped to ribosomal sequences and 8 % to organellar sequences. Raw counts per transcript were obtained with HTSeq and normalised by sequencing depth and transcript length (tpm, Wagner *et al.*, 2012). In total, 54,521 genes (43 % of all 126,667 *de novo* annotated genes) were expressed (detected with median tpm > 0, Figure S5 c) in the sampled shoot material. The 19,479 transcripts (15.38 %) that were quantified at a median level ≥ 5 tpm across all samples were used for subsequent analyses.

3.2. Omics-based prediction of hybrid performance in canola

The following chapter covers the applied objective of the 'Predict' project. The aim was to evaluate the potential of omics-based data sets gathered from the parental lines to be employed in effective prediction of hybrid performance in field and in glasshouse cultivations, and to

investigate whether prediction accuracies achieved with genomic (marker) data can be improved through combinations with -omics data. Aspects of this work were performed in collaboration with Dr. Christian R. Werner (The Roslin Institute, The University of Edinburgh, UK).

3.2.1. Prediction of hybrid performance using individual and combined data sets

Using the best linear unbiased estimators (BLUEs) of all seven agronomic traits, (genomic) best linear unbiased predictions (gBLUP) were performed with different -omics data sets, comprising molecular markers (n= 13,201; unique single-copy SNPs), transcripts (n= 19,479; ≥ 5 tpm) and metabolites (n= 154). These data sets were used individually and in all possible combinations for prediction analyses. Prediction accuracies, the correlation between predicted and observed values, for the tested traits across 100 cycles of cross-validations (3 : 1 = training : validation population) are illustrated in Figure 6. Across all models and traits, prediction accuracies ranged from 0.245 for the trait seedling emergence using all available data sets to 0.72 for total seed GSL content using transcript data only. In all cases prediction accuracies strongly depended on the heritability of the traits with the trait seedling emergence (H^2= 0.34) showing the lowest prediction accuracies followed by seed yield (H^2= 0.62), seed oil yield and seed protein content (H^2= 0.82), DTF (H^2= 0.85), seed oil content (H^2= 0.90) and GSL (H^2= 0.92) with the highest prediction accuracies (Table 1). No significant differences between the prediction accuracies of any of these data sets / combinations were observed for the trait seedling emergence using an Analysis of Variance (ANOVA), followed by a post-hoc Tukey test. For all other traits, the prediction models solely based on metabolite data showed significantly lower prediction accuracies. Other than that, average values of the prediction models with different data sets and combinations thereof were similar for a given trait. Only for the trait total seed GSL content additional significant differences were detected. The model using the genotype data only displayed lower mean prediction accuracy than the other models (except metabolites only). Models including the transcriptome data yielded the highest accuracies for GSL. Thus, in case of the trait total seed GSL content, a significant increase in prediction accuracy could be achieved by adding transcriptome data to the predictive models. For all other traits, using this particular population and data sets, no significant increase in prediction accuracy could be achieved by multi-omics-based predictions.

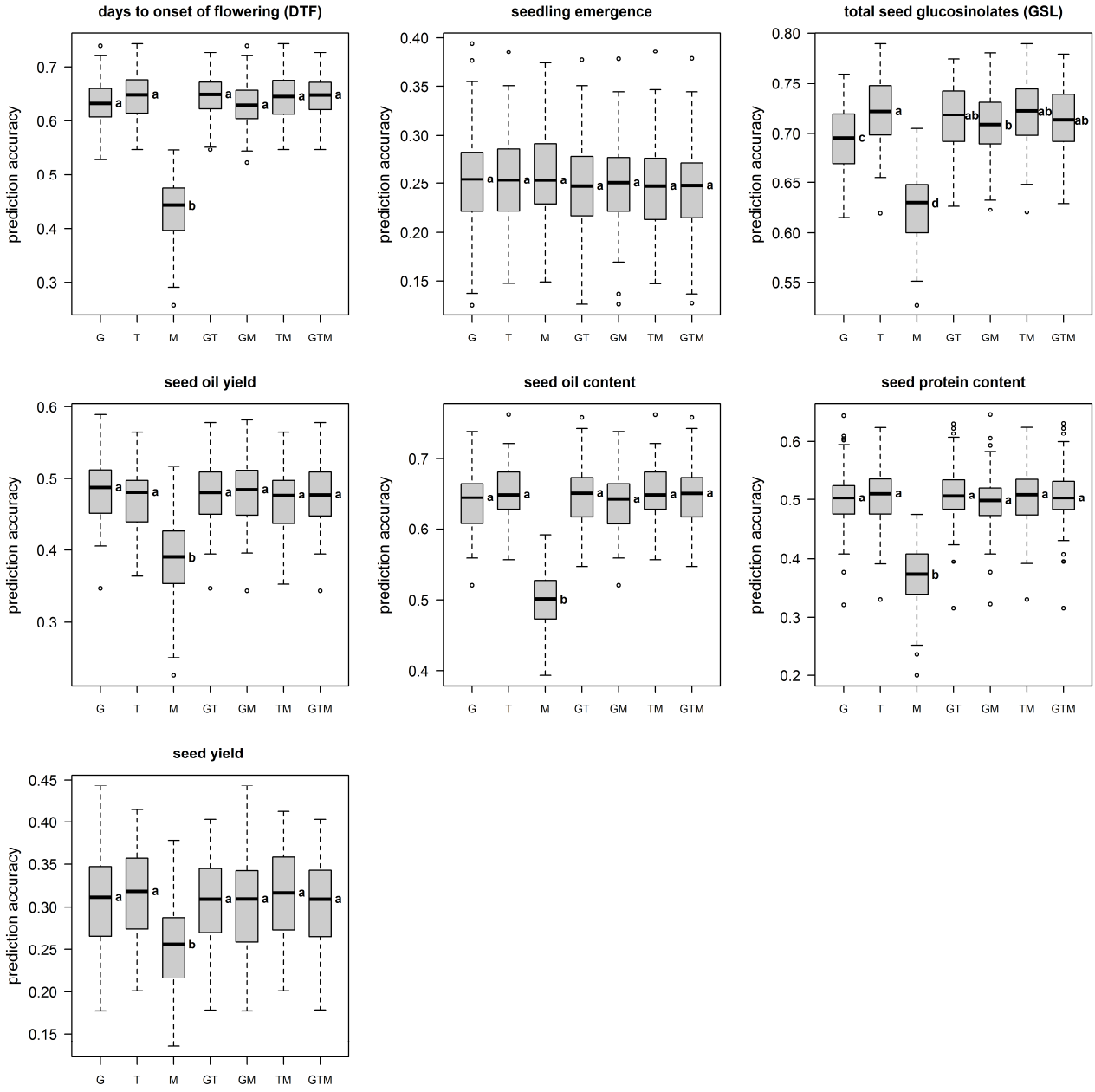


Figure 6. Prediction of hybrid performance by gBLUP models using -omics data sets

A summary of (genomic) best linear unbiased predictions (gBLUP) of hybrid performance is given as boxplots. Seven agronomic traits, assessed in multi-location field trials, were analysed. The prediction accuracies of the models were defined as the correlation between the true and the predicted phenotypic values. A cross-validation scheme with 100 cycles was applied, separating the data set in a training set (75 %) and a validation set (25 %). The different -omics data sets (predictors) were obtained from the parental lines and are denoted as: G, genomic data; T, transcriptomic data; M, metabolite data and their respective combinations G+T, G+M, M+T and G+T+M. The paternal -omics data sets (T & M) were obtained from plants cultivated in the glasshouse. Letters beside the boxes indicate significant differences between predictor sets determined by a one-way ANOVA followed by a post-hoc Tukey's multiple comparison test.

3.2.2. Comparison of the predictive abilities of gBLUP and RKHS models

In addition to gBLUP prediction (Habier *et al.*, 2007, 2013; Goddard, 2009), which has routinely been used as a base-line model in numerous other studies, a second method class, reproducing kernel Hilbert space regression (RKHS; Gianola and van Kaam, 2008) was employed for prediction of hybrid performance (Figure S13).

In contrast to gBLUP, RKHS exploits both the additive and to some extent additive × additive epistatic effects among markers. In general, prediction accuracies for the RKHS models followed a similar pattern than the gBLUP models. The lowest prediction accuracies were obtained for the trait seedling emergence. Prediction accuracies clearly correlated with trait heritability.

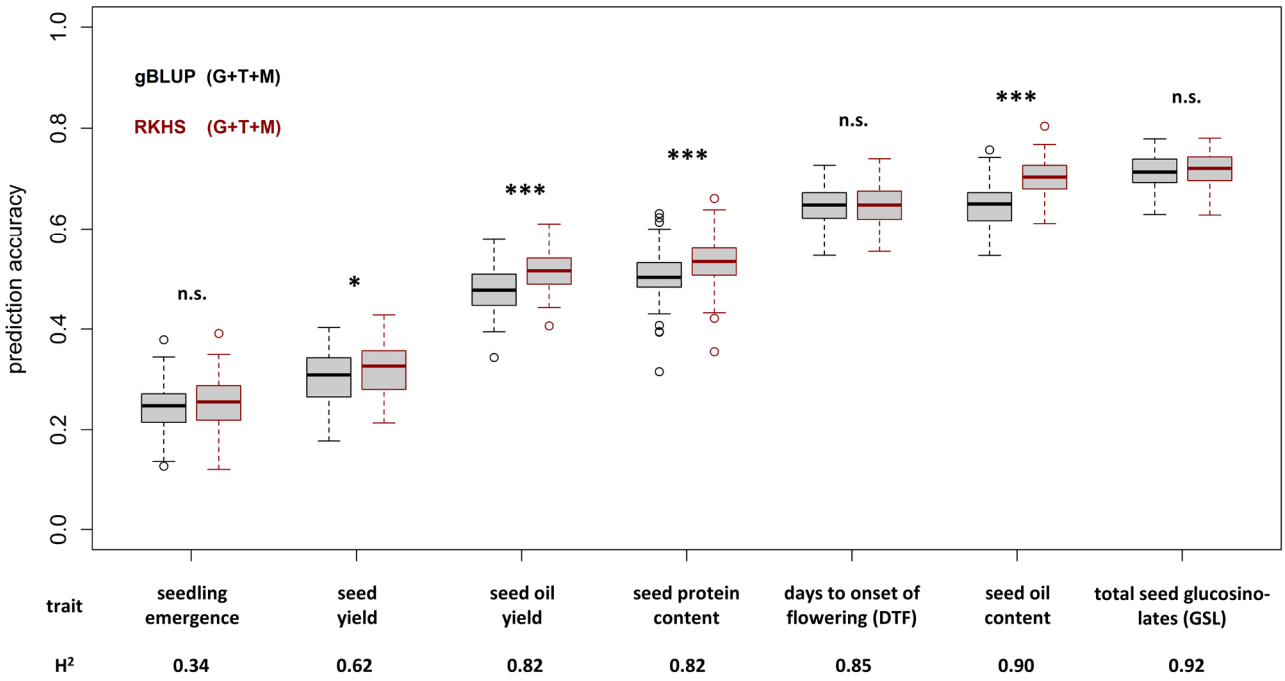


Figure 7. Comparison of Reproducing Kernel Hilbert Space (RKHS) and gBLUP models

A comparative analysis of gBLUP and RKHS models for hybrid prediction is shown as black and red boxplots, respectively. The prediction accuracies were defined as the correlation between the true and the predicted phenotypic values. A cross-validation scheme with 100 cycles was applied, separating the data set in a training set (75 %) and a validation set (25 %). Exemplarily, only the combination of all three -omics data sets as predictors (genomic, transcriptomic and metabolite data; G+T+M) is shown. Asterisks above the plots indicate significant differences between gBLUP and RKHS models determined by Welch's two sample *t*-test (alpha: * = 0.05; ** = 0.01 and *** = 0.001). The broad-sense heritability (H^2) for each of the seven analysed traits is indicated at the bottom of the figure.

In the RKHS models the metabolite data individually also yielded the lowest prediction accuracies. Two traits, total seed GSL content and seed oil content showed significantly higher prediction accuracies for RKHS models (T, G+T, T+M, G+T+M) including transcriptome data compared to models not including transcripts (G, M, G+M; data not shown). In direct comparison, RKHS models were able to outperform gBLUP for the traits seed yield, seed oil yield, seed protein content and seed oil content (Figure 7). Substantial increases in prediction accuracy of up to 3.6 % for seed oil yield, 3.0 %, for protein content, and 5.5% for seed oil content could be achieved by RKHS. No significant improvement of prediction accuracies was obtained for the trait seedling emergence. Although prediction accuracy could not be improved for all traits, RKHS models were in no case inferior to the gBLUP models.

3.2.3. Hybrids display strong mid- and best-parent heterosis

A fifth phenotyping experiment was performed in the glasshouse with a selection of 120 of the 950 hybrids analysed before. These hybrids were selected by the breeders with respect to their seed yield in the field trials whereby the 60 lines with highest overall seed yield and the 60 lines with lowest seed yield were selected. The collection of phenotypic data for these hybrids in combination with the data of the 477 parental lines from the previous four experiments provided the basis to calculate best-parent heterosis (BPH) and mid-parent heterosis (MPH) values. For this purpose, end-point biomass (fresh weight and dry weight), projected leaf area and estimated biovolume were chosen because of their high heritability values (Figure S14). Overall, far more positive than negative mid-parent heterosis was detected, and even for BPH a trend towards positive heterosis was observed (Figure 8). For projected leaf area and estimated biovolume, calculations have been performed for all 21 days individually on the basis of BLUEs across all five phenotyping experiments. In the investigated population, strong positive, as well as negative MPH could be detected ranging from -38.8 % to 138.5 % for projected leaf area (Figure 8 a-b) and from -37.7 % to 135.4 % for estimated biovolume (Figure 8 e-f).

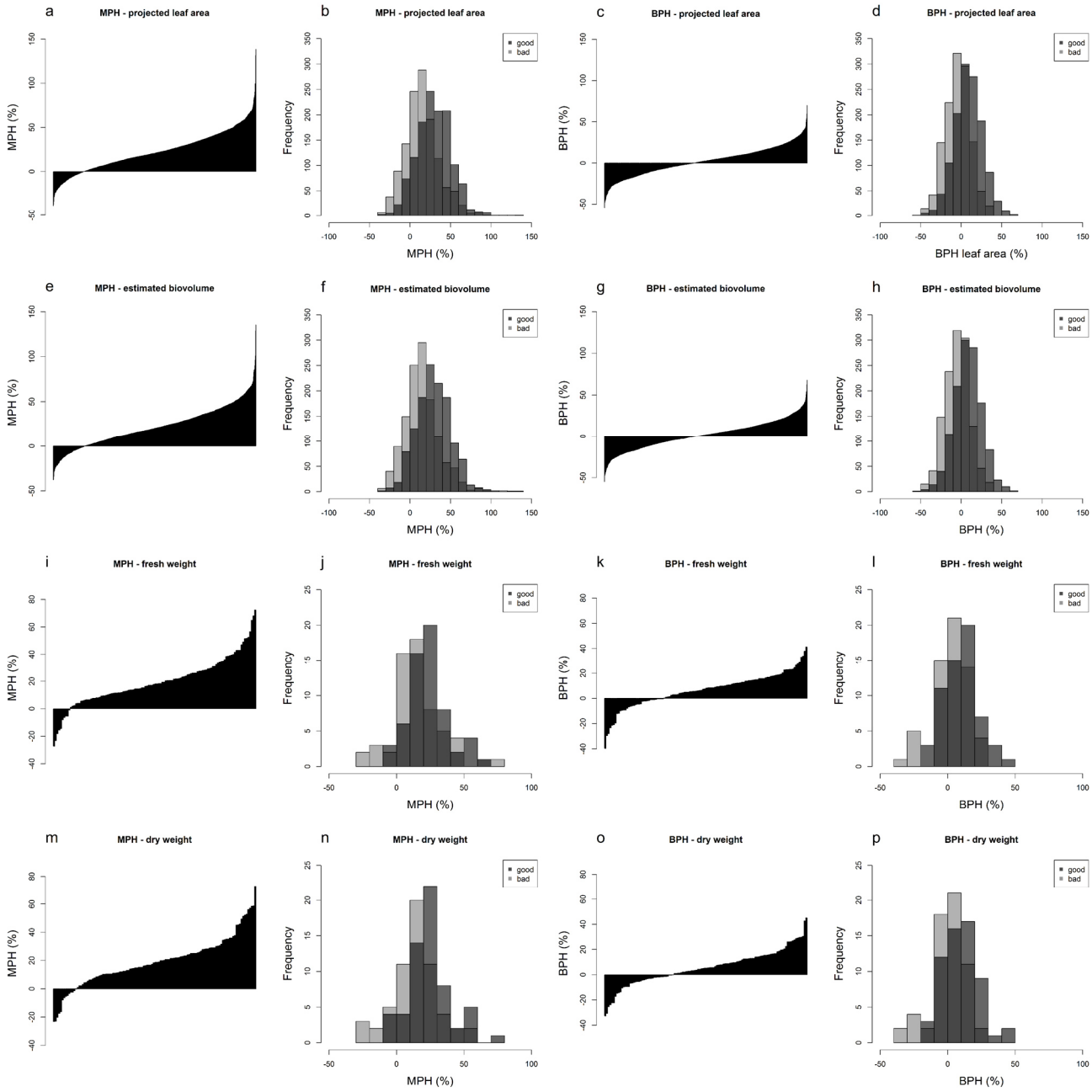


Figure 8. Hybrids display strong heterosis for biomass and growth-related traits

Overview figure of mid-parent (MPH) and best-parent (BPH) heterosis for four selected traits: projected leaf area (PLA), estimated biovolume (VOL), fresh weight (FW) and dry weight (DW). The panels **a**, **e**, **i** and **m** display barplots of MPH and the panels **b**, **f**, **j** and **n** show MPH values as histograms with hybrids distinguished by 'good' and 'bad' seed yield in the field trials, indicated by partially transparent dark and light grey, respectively. The panels **c, g, k** and **o** diplay the BPH values calculated for the same hybrids and traits. The panels **d**, **h**, **l** and **p** show the histograms for BPH traits. FW and DW were determined at 28 DAS. The MPH and BPH values for projected leaf area and estimated biovolume are shown combined for all 21 days of phenotyping (6 to 27 DAS).

Determined BPH values ranged from -54.8 % to 70.0 % for projected leaf area (Figure 8 c-d) and from -54.8 % to 68.2 % for estimated biovolume (Figure 8 g-h). MPH values ranged from -27.21 % to 72.27 % for FW (Figure 8 i-j) and from -23.21 % to 73.01 % for DW, respectively (Figure 8 m-n). BPH values ranged from -39.56 to 41.25 % for FW (Figure 8 k) and from -32.66 to 45.06 % for DW (Figure 8 o-p), respectively.

Comparing the ratio of positive to negative heterosis (MPH and BPH) at the individual time points, a shift over time towards more positive values was observed (data not shown). Notably, grouping the 120 lines into 'good' and 'bad' hybrids with respect to seed yield data from the field trials, distinct differences in their distributions were detected. The set of 'good' hybrids displayed significantly higher MPH (Figure 8 b, f, j and n), as well as BPH values (Figure 8 d, h, l and p) for all four traits (Welch Two Sample *t*-test, two-sided, *p*-value < $2.2e\text{-}^{16}$) compared to the set of 'bad' hybrids. Significant differences between the two groups of hybrids originating from the crosses either with MS1 or MS2 as mother line were detected. Crosses with MS1 displayed substantially higher MPH values for leaf area and MPH for biovolume than crosses with MS2. In addition, BPH for leaf area and biovolume were substantially higher for crosses with MS1 at the earlier time points up to 13 DAS (data not shown). However, no significant differences for end-point biomass heterosis (FW and DW; MPH and BPH) were detected at 28 DAS between hybrids originating from crosses with MS1 or MS2, respectively.

3.2.4. Prediction of early vegetative growth of hybrids in the glasshouse

Although the set of 120 hybrids grown in the glasshouse represents only 13 % of the whole set of 950 hybrids, and the power might be reduced in comparison to the whole data set, end-point biomass data (FW & DW) of the hybrids was used to perform prediction analyses. As parental lines and selected hybrids were grown in the same facilities under the same controlled environmental regime, it was hypothesised that end-point biomass might reflect more closely the parental data sets than the hybrid yield data from the field. Prediction analyses were performed with the same models and predictor sets (G, T, and M), and their combinations as described above. For the combined predictor sets (G+T+M), prediction accuracies (gBLUP) of 0.62 and 0.66 were achieved for FW and DW, respectively (Figure S15). Notably, in contrast to the predictions of the seven agronomic traits evaluated in the field trails, the metabolites as predictors achieved

substantially higher prediction accuracies for plant biomass assessed in the glasshouse. The prediction accuracies using the metabolite data were comparable to those using the transcript data as predictors (Figure S15). In addition, the time series data for projected leaf area and estimated biovolume (6 to 27 DAS) were analysed using gBLUP models and the combined data sets (G+T+M). Prediction accuracies are low for early time points, but increase over time and reach saturation at a value of approx. 0.6 for both traits (Data S8). Prediction of MPH and BPH values for both traits yielded overall low prediction accuracies (≤ 0.4). In contrast to the *per se* value predictions, the highest prediction accuracies, for BPH (median: leaf area = 0.39; biovolume = 0.4), were determined at the first measuring day (6 DAS), with a second peak observed around 14 DAS, the time point when material of the parental lines was sampled for transcriptome and metabolome analyses (Data S8).

3.3. Comprehensive analyses of the -omics data sets

The extensive multi-level -omics data sets that were generated for the parental lines were utilised to perform correlation analyses to identify potential links between the omics-layers. In combination with the array-derived genotype data, genome-wide association (GWA) and co-localisation studies were performed to identify candidate genetic loci and genes potentially underlying these loci. The focus was placed on the functional interrelationship between biomass, vegetative growth, gene expression and metabolite profiles.

3.3.1. Correlation analyses between -omics data sets

Pairwise Pearson correlation analyses were performed between the different -omics data sets, with 154 polar metabolites, 19,479 transcripts and 2,691 phenotypic traits (including individual traits at 21 days, the growth rates described earlier, and end-point biomass). In total, 532 significant pairwise correlations between transcripts and phenotypic traits, 331 between metabolites and transcripts and only 22 between metabolites and phenotypic traits, with moderate correlation coefficients $|r| \geq 0.4$ and $p\text{-values}_{FDR} \leq 0.05$ were recorded (Data S9; Figure S16). Three questions were addressed, (i) is there a linear relationship between molecular traits and end-point biomass, (ii) can molecular traits be linked to projected leaf area and

estimated biovolume at the days shortly before or after the sampling time point, and (iii) is it possible to derive significant features related to biomass and growth by analysing the intersections of pair-wise relations between all three data sets?

Only weak correlations ($|r| \sim 0.3$) between biomass (FW or DW) and polar metabolites (for example '2-hydroxy-Glutaric acid', 'Indole-3-acetonitrile' and 'Unknown MST 68'), and only seven correlated transcripts with $|r| \sim 0.4$ could be detected: 'BnaC02g08760D', an elongation factor 1b α-subunit protein; 'maker-scaffold124-snap-gene-8.174' showing homology to ATPAP18, a purple acid phosphatase; 'maker-scaffold296-snap-gene-35.40', which shows homology to the Arabidopsis *EGY3*; 'BnaC02g16010D' encoding a nucleic acid- ob-fold-like protein; 'BnaC06g28700D' annotated as signal recognition particle signal 72 kDa; 'BnaC02g04720D', a peptidyl-prolyl cis-trans isomerase; and 'BnaA07g19110D', a putative protein kinase family protein.

As levels of only few transcripts displayed linear relationships with FW or DW, a complementary approach was applied to detect potential regulators of biomass. Expression data (median ≥ 5 tpm) were subjected to a gene network inference analysis using a Random Forests (RF) based method (GENIE3 algorithm, Huynh-Thu *et al.*, 2010) while restricting the targets to biomass (FW and DW) only. The ten highest ranked potential regulatory transcripts of fresh and dry weight are described in Table 3 and are shown in Figure S17. Among them were 'BnaC06g28700D' (positively correlated), 'BnaC02g04720D' (negatively correlated), and 'maker-scaffold124-snap-gene-8.174' (negatively correlated), which were also mentioned above with correlations ($|r| \geq 0.4$) to end-point biomass (FW or DW). Five of the candidates were shared between the ten top-ranked transcripts of fresh and dry weigh (Table 3, displayed in bold font). The other five candidates in each of the two data sets were found among the top 50 candidates of the reciprocal data set (data not shown). The output of the analysis was restricted to the 100 top-ranked regulatory links between transcripts and end-point biomass (fresh weight, Data S10) and the candidates were subjected to a Singular gene ontology (GO) term Enrichment Analysis (SEA) using the 'agriGO' v2.0 web tool (http://systemsbiology.cau.edu.cn/agriGOv2/). The terms 'translation' (GO:0006412), 'peptide biosynthetic process' (GO:0043043), 'ribosome' (GO:0005840), and 'structural constituent of ribosome' (GO:0003735) were found among the most significantly enriched GO terms.

Table 3. The 10 top-ranked transcripts associated with biomass identified by GENIE3

Target trait [a]	potential regulatory transcript	weight [b]	best match [c]	gene description [d]	r [e]
fresh weight	maker-scaffold59-augustus-gene-9.44	0.003471627	BnaA07g21340D	auxin efflux carrier family protein	0.3915
fresh weight	snap_masked-scaffold154-processed-gene-33.17	0.003466832	BnaA02g20630D	lipid-binding protein at4g00165-like	-0.3661
fresh weight	**maker-scaffold402-augustus-gene-1.35**	**0.003355078**	**BnaC06g28700D**	**signal recognition particle 72 kDa**	**0.4485**
fresh weight	**maker-scaffold763-augustus-gene-27.30**	**0.003159457**	**BnaC06g05870D**	**peroxisomal membrane mpv17 pmp22 family protein**	**-0.3971**
fresh weight	maker-scaffold28-augustus-gene-24.36	0.002922718	BnaA07g25010D	60s ribosomal protein l17	0.3972
fresh weight	**maker-scaffold171-augustus-gene-46.49**	**0.002815322**	**BnaC06g32010D**	**mlp-like protein 28**	**-0.3691**
fresh weight	maker-scaffold364-snap-gene-12.49	0.002614644	BnaCnng17010D	serine hydroxymethyltransferase 2	0.3975
fresh weight	maker-scaffold174-augustus-gene-100.20	0.002561317	(AT4G03120)	(C2H2 and C2HC zinc fingers superfamily protein)	-0.3594
fresh weight	**maker-scaffold28-snap-gene-38.140**	**0.002483514**	**BnaA07g23560D**	**serine hydroxymethyltransferase 2**	**-0.3740**
fresh weight	**maker-scaffold574-augustus-gene-31.44**	**0.002445149**	**BnaC02g04720D**	**peptidyl-prolyl cis-trans isomerase chloroplastic-like**	**-0.4135**
dry weight	maker-scaffold124-snap-gene-8.174	0.004173532	(NA)	(NA)	-0.4038
dry weight	**maker-scaffold763-augustus-gene-27.30**	**0.003860580**	**BnaC06g05870D**	**peroxisomal membrane mpv17 pmp22 family protein**	**-0.3583**
dry weight	**maker-scaffold402-augustus-gene-1.35**	**0.002996914**	**BnaC06g28700D**	**signal recognition particle 72 kDa**	**0.4042**
dry weight	maker-scaffold226-augustus-gene-38.6	0.002993855	BnaC02g13540D	glucose-6-phosphate phosphate translocator	0.2893
dry weight	maker-scaffold226-augustus-gene-39.20	0.002668281	BnaA02g27940D	transcription factor ilr3	0.3539
dry weight	**maker-scaffold28-snap-gene-38.140**	**0.002607634**	**BnaA07g23560D**	**serine hydroxymethyltransferase 2**	**-0.3378**
dry weight	maker-scaffold154-snap-gene-32.32	0.002469829	BnaA02g20750D	dormancy auxin associated protein	-0.3624
dry weight	maker-scaffold206-augustus-gene-43.65	0.002287003	BnaA04g29230D	40s ribosomal protein s5	-0.1153
dry weight	**maker-scaffold574-augustus-gene-31.44**	**0.002182786**	**BnaC02g04720D**	**peptidyl-prolyl cis-trans isomerase chloroplastic-like**	**-0.3732**
dry weight	**maker-scaffold171-augustus-gene-46.49**	**0.002175093**	**BnaC06g32010D**	**mlp-like protein 28**	**-0.3691**

[a] best linear unbiased estimators (BLUEs) for fresh and dry weight were calculated across the glasshouse experiments
[b] weight from the adjacency matrix of the inferred network by GENIE3 that were used to sort the regulatory links
[c] best match (gene) using the basic local alignment search tool (BLASTn) as described in Materials and methods; 'NA' indicates no unique hit matching the filtering criteria
[d] gene description based on homology to the best Arabidopsis gene hit obtained from the TAIR data base (https://www.arabidopsis.org/)
[e] Pearson correlation coefficients between transcripts and fresh and dry weight, respectively

potential regulatory transcripts associated with both, fresh and dry weight, are indicated in bold font

Eleven candidates (Data S15), ten encoding ribosomal proteins and one, 'BnaC06g28700D' encoding the 'signal recognition particle 72 kDa', were annotated with the GO term 'intracellular ribonucleoprotein complex' (GO:0030529), which showed the most significant enrichment (*p*-value = 1.2e-5; FDR = 0.00038).

Regarding the time points close to the sampling day (14 DAS +/- 3 days), only few significant correlations between transcripts and growth-related traits ($|r| \sim 0.4$) were detected: 'BnaC02g08760D' negatively correlated with projected leaf area at 17 DAS, 'BnaC02g01040D' also negatively correlated with projected leaf area at 16 and 17 DAS, and 'Bra000292' and 'BnaC07g05770D' positively correlated with early plant height at 11 DAS. Broadening the time frame and filtering for correlations at multiple days, for early plant height (between 6 and 11 DAS) two transcripts, one with homology to 'Bra000292', a gene putatively involved in cell wall organization, and 'BnaC07g05770D', a peroxidase (PER64) involved in stem lignification and mainly expressed in the shoot in Arabidopsis (Yi Chou *et al.*, 2018), were significantly correlated. One transcript ('BnaC06g28700D', signal recognition particle 72 kDa) showed positive and three other transcripts ('BnaC02g08760D', translation elongation factor EF1B/ribosomal protein S6 family protein; 'BnaC02g01040D', a putative mitochondrial ATP synthase β-subunit, and 'maker-scaffold124-snap-gene-8.174'), negative correlations with projected leaf area in a later phase (Data S9).

Due to the low correlation between metabolic and phenotypic traits, no significant triangular relations between all three data sets (metabolites, transcripts, and growth-related traits) could be detected. However, 141 correlations between the metabolite and transcript layers ($|r| \geq 0.4$, involving metabolites of known structure) were detected (Data S9). Among the highest ranking correlations with metabolites of known chemical structure were 'BnaA06g01540D' (β-glucosidase 18) ~ indole-3-acetonitrile (r= -0.582); 'Bra022161' (JASMONATE RESPONSIVE 1) ~ indole-3-acetonitrile (r= 0.544), 'BnaC01g00550D' (a putative aminotransferase) ~ β-alanine (r= -0.574); 'BnaC03g39190D' (GO: response to sucrose stimulus) ~ sucrose (r= -0.572); and 'BnaC07g45200D' (an arginine decarboxylase) ~ putrescine (r= 0.536). Moreover, correlations of transcripts with multiple metabolites were detected, including 'maker-scaffold19-snap-gene-11.66' (11 correlated metabolites, mostly with amino acids or derivatives; displays homology to the Arabidopsis AILP1; annotated as N-terminal nucleophile aminohydrolase), 'BnaC01g05300D' (11 correlated

metabolites; annotated as α-vacuolar processing enzyme; GO terms related to proteolysis and lytic vacuole) or 'BnaC04g41510D' (seven correlated metabolites; annotated as glutathione S-transferase). In addition, correlations between paternal transcript levels in the pollinators and agronomic traits in the hybrids were performed. Three correlations ($|r| \geq 0.4$) were identified: 'BnaA07g19110D', a protein kinase family protein, was positively ($r= 0.414$), and 'maker-scaffold296-snap-gene-35.40' ($r= -0.4067$) and 'BnaA07g21340D' ($r= -0.4091$), encoding an auxin efflux carrier family protein, were negatively correlated with seed oil content of the hybrids. Lower correlations ($0.3 \leq |r| \leq 0.4$) were also detected for the traits seed oil content, seed protein content, seed oil yield, seed glucosinolate content and days to onset of flowering (Data S9).

3.3.2. The expression of the *Brassica* subgenomes and biomass accumulation

An explorative principal component analysis of the transcriptome data indicated a clustering of genotypes in the 4th principal component corresponding to the breeding pools underlying the population investigated (Figure 9 a). In addition, the same PC group discriminated lines by end-point biomass. Two subsets, one with overall higher and one with overall lower biomass were partially separated (Figure 9 b). To further explore this observation, the scaled, centred and Log_2 transformed loadings (transcripts with median ≥ 5 tpm) of the 4th principal component were extracted and subjected to a deeper analysis. First, the top ranking positive ($n= 78$) and negative ($n= 74$) loadings with an absolute loading value > $|0.02|$ were separately subjected to a gene ontology (GO) term enrichment analysis using the 'agriGO' web tool (http://bioinfo.cau.edu.cn/agriGO/). The analysis revealed significant enrichments, among others for the GO terms 'biosynthetic process' (GO: 0009058), 'gene expression' (GO: 0010467) and 'translation' (GO: 0006412) in the negative loadings (Figure 10 a). In the positive loadings, 'cellular biosynthetic process' (GO: 0044249), 'cellular metabolic process' (GO: 0044237), 'generation of precursor metabolites and energy' (GO: 0006091), 'proton transport' (GO: 0015992) and again 'translation' (GO: 0006412) were found among the enriched terms (Figure 10 b). Another interesting observation was that the subgenome contribution differed between the top negative and positive loadings. For the negative loadings (direction of lower biomass), 28 transcripts were contributed from genes of the A subgenome and 46 transcripts from genes of the C subgenome, respectively.

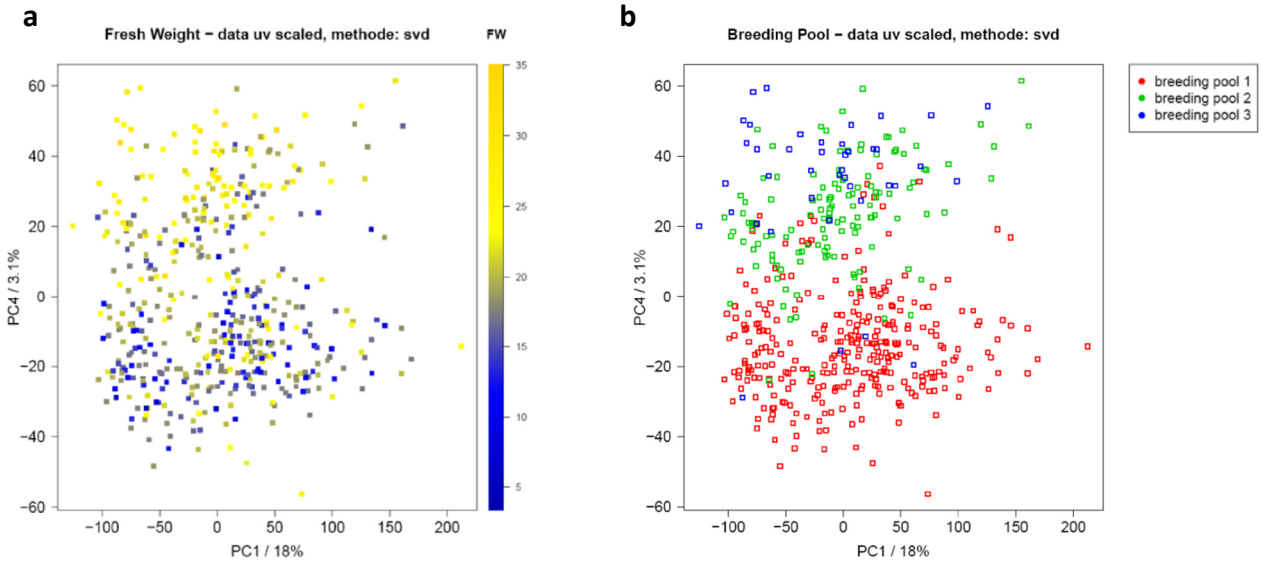


Figure 9. Transcript profiles separate genotypes according to biomass and breeding pools

Principal component analysis was performed on filtered transcript data (median tpm ≥ 5) for all 477 genotypes. Transcript data were centred and scaled (z-scores). The PCA calculation was done by singular value decomposition (svd) of the data matrix. The first four PCs explained 18 %, 9.7 %, 3.8 % and 3.1 % of variance, respectively. **a** Scatter plot of PC1 and PC4 with samples coloured according to their biomass (fresh weight BLUEs) using a gradual scale (colour gradient blue, low biomass to yellow, high biomass). **b** The same PCA plot with genotypes coloured according to their affiliation to one of the breeding pools.

In contrast, the top positive loadings (direction of higher biomass) contained 51 transcripts from genes of the A subgenome and 27 transcripts from genes of the C subgenome, respectively. The hypothesis that a higher expression of particular classes of genes from the A subgenome is associated with higher biomass production is also supported by the RNA-Seq pilot experiment. Here a set of genotypes were analysed including four lines selected based on their overall biomass production (FW 28 DAS) during the phenotyping experiments. The four lines 'Pol 229' (low biomass), 'Pol 396' and 'Pol 467' (medium biomass), and 'Pol 419' (high biomass) were sequenced as individual replicates (pool of four individuals from one experiment) and as pools of material from all experiments. Two of these lines ('Pol 229' and 'Pol 419') were picked for a detailed biological analysis. First, differentially expressed genes (DEGs) between the two lines were identified. In total, 1,153 coding sequences (CDS, 1.1 %) were higher expressed in the high biomass line 'Pol 419', 658 derived from the A subgenome (1.4 %) and 489 from the C subgenome (0.9 %), respectively.

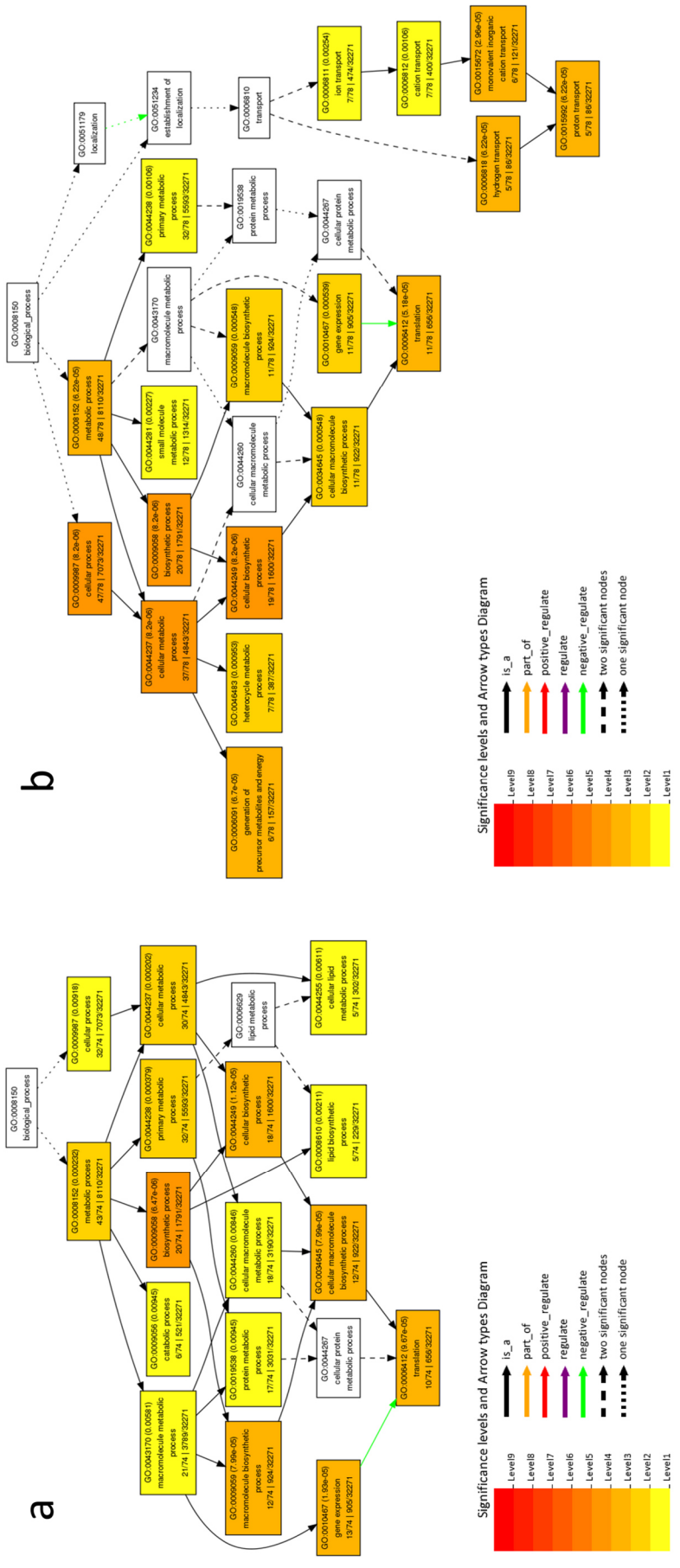


Figure 10. Gene ontology terms associated with biomass accumulation (loadings of PC4)

Gene ontology (GO) term enrichment (singular enrichment analysis, SEA) was performed using the 'agriGO' webtool. The top loadings of PC4 (n= 152), which separates genotypes by biomass (see Figure 9), were filtered for absolute values ≥ |0.02|. The analysis was performed on the split subsets of **a** negative (n= 74) and **b** positive (n= 78) loadings. As background a customised annotated *Brassica napus* reference was used (see materials and methods). Gene ontology (GO) information was obtained from Genoscope (http://www.genoscope.cns.fr/brassicanapus/). Results were visualised using the display category 'Biological Process'. The minimal number of mapping entries was set to n= 5, the statistical test method 'hypergeometric' was selected and a significance level of 0.01 was applied.

In the low biomass line ‘Pol 229’, 1,503 CDS were higher expressed with 612 attributed to the A subgenome (1.4 %) and 890 to the C subgenome (1.6 %). Thus, the line ‘Pol 229’ displayed an even distribution between the subgenomes, as expected from the ratio of transcripts encoded by the A and C subgenomes, while ‘Pol 419’, the high biomass line, had relatively more genes from the A than the C subgenome significantly higher expressed. This observation is consistent with the results of the top positive PCA loadings reported in the previous paragraph. Moreover, an analysis of gene ontology terms was performed. In both lines GO terms related to translation were found in the top enriched CDS which is consistent with the observations from the analysis of PC4 in the main experiment. Both lines also display enrichment for separate terms, for instance terms related to α-amino acid metabolism and nucleosome assembly for ‘Pol 229’ and terms related to purine biosynthesis for ‘Pol 419’. In addition to the GO term enrichment analysis, transcripts were mapped to pathways using Mapman (Figure S18). Transcripts of ‘Pol 229’ were significantly enriched in protein biosynthesis and glucosinolate synthesis, while those of ‘Pol419’ were enriched in photosynthesis and light reaction.

3.3.3. Candidate genes putatively affecting biomass heterosis in canola

Early biomass heterosis is an agronomically important trait, but difficult to predict by the parental *per se* performance (Gärtner *et al.*, 2009; Steinfath *et al.*, 2010). Thus, to identify molecular factors associated with superior hybrid performance (heterosis), BPH and MPH values for projected leaf area and estimated biovolume calculated for different days (6 to 27 DAS), as well as MPH and BPH values for end-point biomass (28 DAS) were correlated with the expression data of the pollinators (male parental lines). Overall, only low to moderate correlation coefficients ($|r| \leq 0.5$) were observed. Moreover, due to the high number of statistical tests and the relatively low number of samples (n= 120), *p*-values would not pass a multiple-testing correction. Still these correlations might reflect biological links between individual transcripts and biomass heterosis. For end-point biomass the highest correlation ($|r| = 0.42$) has been observed between dry weight MPH and transcript ‘maker-scaffold378-snap-gene-2.74’ which was not annotated, but shows homology to the Arabidopsis AGG3 (*AT5G20635*) encoding an atypical heterotrimeric G-protein gamma-subunit. In contrast, substantially higher correlations were observed between the parental transcript data and the *per se* hybrid biomass values (Data S9). The highest correlation ($r = 0.55$)

was found for the parental expression of ‘BnaA01g31350D’, annotated as ‘superoxide dismutase’ and hybrid fresh weight.

Analysing the correlations for all time points, leaf area MPH and biovolume MPH were highest correlated with the transcripts ‘maker-scaffold378-snap-gene-2.74’, ‘BnaC04g13490D’ annotated as ‘cdt1-like protein chloroplastic-like’ and ‘BnaCnng47650D’ encoding a FANTASTIC FOUR-like protein. However, no significant correlations ($|r| \geq 0.3$) of these three transcripts with growth-related traits (estimated biovolume, projected leaf area or end-point biomass) in the parent were determined.

Notably, when samples were grouped by their origin from crosses with MS1 or MS2, ‘BnaCnng47650D’ shows substantially higher correlation for the MS1 subset than for the MS2 subset. This is exemplarily shown for ‘BnaCnng47650D’ at the time point with the highest correlation coefficient at 9 DAS (Figure 11 a). The differences were most distinctive at earlier time points (Figure 11 b). If the two subsets are further divided into ‘good’ and ‘bad’ hybrids regarding their seed yield in the field, significant differences in leaf area MPH can be determined (Figure 11 c). Although no significant differences could be detected regarding the expression of ‘BnaCnng47650D’ in the grouped pollinators, a tendency towards higher expression of ‘BnaCnng47650D’ in the ‘good / MS1’ subset could be observed (Figure 11 d).

Moreover, MS1 plants themselves showed a higher expression of the gene compared to MS2 plants. Similar differences between the MS1 and the MS2 subsets were also observed for the other two transcripts ‘maker-scaffold378-snap-gene-2.74’ and ‘BnaC04g13490D’ (data not shown).

For leaf area and biovolume BPH values at different time points, the highest correlated transcript ($|r| \geq 0.4$) was ‘BnaAnng16580D’ described as ‘tata-associated factor ii 58’, which was positively correlated with leaf area BPH 19 DAS to 24 DAS and volume BPH 20 DAS to 24 DAS.

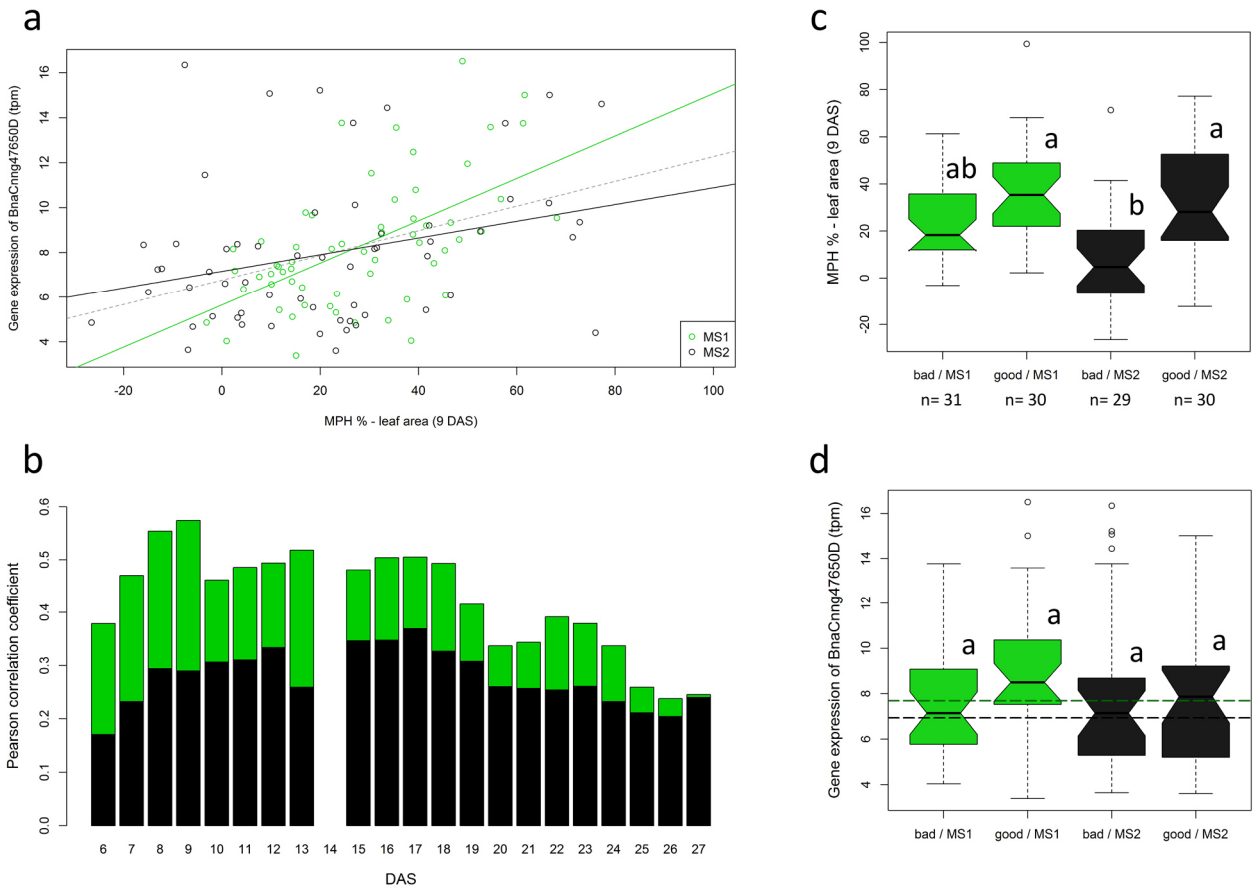


Figure 11. *BnaCnng47650D* as a candidate gene for biomass heterosis

In panel **a**, the correlation between MPH values for projected leaf area at 9 DAS and the abundance of the 'BnaCnng47650D' transcript in the parental pollinators is shown (grey dashed line; r= 0.396, *p*-value= 8.388e-6, n= 120). Higher and lower correlations were observed when analysing the MS1 (green line; r= 0.574, *p*-value= 1.586e-6, n= 61) and MS2 (black line; r= 0.290, *p*-value= 0.026, n=59) subsets individually. The gene 'BnaCnng47650D' encodes a FANTASTIC FOUR-like protein. **b** Pearson correlation coefficients of MPH values for projected leaf area ~ 'BnaCnng47650D' for all 21 days from 6 DAS to 27 DAS. Green (background) and black colours (foreground) indicate the MS1 and MS2 subset, respectively. **c** Boxplots for MPH values for projected leaf area for the subsets further grouped into 'good' (high seed yield) and 'bad' (low seed yield) hybrids. Different letters indicate significant differences between grouped samples ('bad / MS1' 'good / MS1', 'bad / MS2', 'good / MS2') determined by a one-way ANOVA followed by a post-hoc Tukey's multiple comparison test. **d** Boxplots of 'BnaCnng47650D' expression in the parental pollinator lines in tpm. Same letters indicate no significant differences between all four sets. The dashed green and black horizontal lines correspond to the expression levels of 'BnaCnng47650D' in the male-sterile lines MS1 and MS2, respectively.

3.4. Multi-omics genome-wide association studies

The generation of extensive omics data sets (time-resolved phenotypes, transcripts and metabolite profiles) for a large and diverse set of rapeseed lines provided the opportunity to study the genetic basis of trait variation at different omics-levels. To further this goal, all traits were subjected to genome-wide association analyses. Detected marker-trait associations were analysed for enrichment in particular regions of the genome for each of the data sets and compared for co-localisations between the different omics-layers. The results of these analyses are summarised in the following section.

3.4.1. Identification of phenotypic, expression and metabolite QTL

Extensive -omics data sets were generated for 2,691 phenotypic traits (including individual phenotypic traits at 21 days, growth rates for a selection of four traits calculated over three-day intervals and end-point biomass), 154 polar metabolites and expression data for 19,479 transcripts (expressed at ≥ 5 tpm). For all phenotypic traits combined, a total of 15,789 marker-trait associations / quantitative trait loci (QTL) were detected at a $p\text{-value}_{FDR} \leq 0.05$ (Data S11). For 2,150 (80 %) of the phenotypic traits at least one associated locus could be identified. On average, detected associations explain only a minor proportion of phenotypic variance (mean of 1.8 %), but large effects were identified too, for example a deletion at marker 'Bn-scaff_15877_1-p293578_del' was associated with the leaf width mean at 16 DAS and explained 41,9 % of phenotypic variance. In summary, 77 QTL (0.5 %) explain more than 10 %, 8,570 QTL (54.3 %) between 10 % and 1 %, and 5,236 QTL (33.2 %) less than 1 % of phenotypic variance (PVE). If too many QTL were fitted into the ANOVA model, no sum of squares could be extracted for the least significant QTL. Thus, for 1,906 QTL (12 %) no PVE values were obtained due to errors in the model fit. Notably, the most significant associations were detected with a series of deletion markers covering nearly the complete chromosome C03 and the trait yellow to green ratio. A total of 2,255 associations, 1,925 of them deletions, were detected for this particular trait at various days.

Likewise, data for all 19,479 transcripts (expressed at ≥ 5 tpm) were individually subjected to GWAS analyses to identify expression-QTL (eQTL). For 14,683 transcripts (75.4 %) at least one eQTL was identified. In total, 86,013 associations were detected with a *p*-value $_{FDR}$ ≤ 0.05. PVE of these eQTL ranged between < 0.1 % and > 70 %. The highest percentage of 71.7 % PVE was found for a cis-eQTL on chromosome A07 for 'BnaC07g36930D' annotated as 'peptide methionine sulfoxide reductase b2'. In summary, 6,012 eQTL (7.4 %) explain more than 10 %, 43,789 eQTL (50.1 %) between 10 % and 1 %, 26,323 eQTL (32.5 %) between 1 % and 0.1 % and 4,844 eQTL (6 %) less than 0.1 % of phenotypic variance. Notably, 26.3 % of associations were detected with a deletion as genetic marker, with a deletion on chromosome C02 associated with 'BnaC02g01750D' coding for a 'putative proline-rich family protein' showing the highest explained variance (49.6 % PVE). The eQTL were classified either as cis-eQTL or trans-eQTL. As there is no universal definition, cis-eQTL were defined as associated markers within an interval of ± 500 kb (or alternatively ± 1 Mb) of the transcription start of the gene itself and trans-eQTL as associations outside of this interval or on another chromosome. Overall, a higher number of trans-eQTL than cis-eQTL was detected. In total, 8,140 (10,007) cis-eQTL and 77,873 (76,006) trans-eQTL were identified for both intervals, respectively. Cis-eQTL explained substantially more phenotypic variance than trans-eQTL (Figure 12 a). Cis-eQTL defined by either a ± 500 kb or a ± 1 Mb interval did not show any substantial differences in their distribution of explained phenotypic variance. Approximately 40 % of the cis-eQTL exhibited more than 10 % PVE, and approx. 50 % between 1 % and 10 % PVE. On the other side, only around 3 % of the trans-eQTL showed more than 10 % PVE and around 40 % less than 1 % PVE. For many transcripts associated with multiple eQTL, the most significant eQTL represents a cis-eQTL likely corresponding to the gene locus itself as indicated exemplarily for 'BnaA04g27920D' encoding a potential ribulose bisphosphate carboxylase small chain 2b precursor on chromosome A04 (Figure 12 b).

For all polar primary metabolites, a total of 544 marker-trait associations, 257 for metabolites of known and 287 for metabolites of unknown chemical structure, were detected at a *p*-value $_{FDR}$ ≤ 0.05. In summary, for 102 (66 %) of the 154 metabolites, at least one QTL could be detected. The average mQTL explained 2.2 % of phenotypic variance. Eleven mQTL (2 %) explain more than 10 %, 352 mQTL (64.7 %) between 10 % and 1 % and 181 mQTL (33.3 %) less than 1 % of phenotypic variance. The highest PVE, 31.9 %, was detected for the marker-trait association

(MTA) between ‘Bn-scaff_16361_1-p506447’ and a metabolite of unknown chemical structure. The highest PVE for a metabolite of known structure was observed for marker ‘Bn-scaff_16361_1-p439309’ on chromosome C08 and the amino acid proline, explaining 14.7 % of the metabolite’s variance.

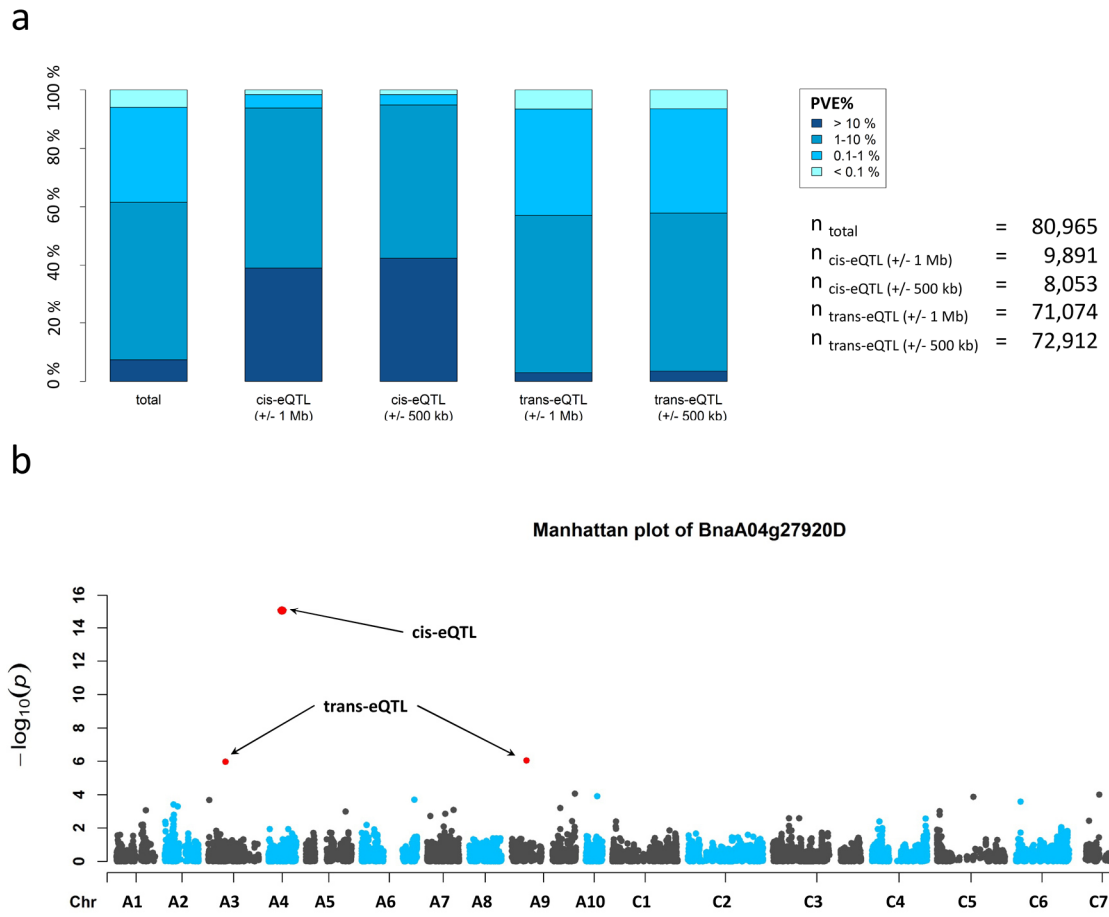


Figure 12. Detection of cis- and trans-eQTL

Panel **a** shows an overview of detected expression QTL (eQTL) and explained phenotypic variance (PVE%). The leftmost bar shows the distribution of PVE% for all 80,965 eQTL with a p-value $_{FDR} \leq 0.05$. The blue colour code refers to the PVE% of the eQTL grouped by: > 10 %, 1-10 %, 0.1-1 % and < 0.1 % PVE. The four other bars group the eQTL in cis- and trans-eQTL, defined by either ± 500 kb or ± 1 Mb intervals around the transcription start of the respective gene. Panel **b** shows the Manhattan plot of the transcript ‘maker-scaffold184-augustus-gene-48.19’ annotated as ‘BnaA04g27920D’. Three significant marker-trait associations (MTAs) on chromosomes A4, A3 and A9 are indicated by red dots (p-value $_{FDR} \leq 0.05$). One MTA was classified as cis-eQTL and two MTAs as trans-eQTL.

3.4.2. QTL for phenotypic, expression and metabolite traits cluster in hotspots

For a subsequent analysis, associations were filtered for eQTL that explain more than 2 % of phenotypic variance, reducing the number of associations to 26,391 (14.2 %). These eQTL (8,610 cis-eQTL and 17,781 trans-eQTL) were binned in overlapping intervals of ± 1 Mb and tested for their distribution. The eQTL were not equally distributed across the 19 chromosomes. Some chromosomal regions were depleted of QTL while others show substantial overrepresentation (hotspots) compared to the average of the chromosome (Figure S19). The highest number of eQTL was detected on chromosome C03 (n= 2,760) and the lowest number on chromosome C07 (n= 415). Moreover, pronounced hotspots with more than 150 co-localised eQTL could be detected on chromosomes A02, A03, C02, C03, C05, C06, C08, and C09 (Figure S19 a). The largest number of co-localised eQTL was detected in an interval (2 to 3 Mb) on chromosome C05. In this region a high overrepresentation of trans-eQTL was found (16 cis-eQTL vs. 453 trans-eQTL). In general, for the A subgenome eQTL seem to be more evenly distributed, while for the C subgenome eQTL tend to cluster more often in hotspot regions. Similar patterns were observed for metabolite QTL (mQTL, (Figure S19 b), as well as for phenotypic QTL (Figure S19 c), but the hotspot regions were not identical. For phenotypic traits, pronounced co-localisation of QTL was observed on chromosomes A02, A03, C02, C03, C05, C06 and C09 with a notable hotspot of more than 150 QTL in a small region of chromosome C02. Regions with a particular high density of mQTL were observed on chromosomes A03, A05, C03, C06 and C08. However, in total, much fewer mQTL than phenotypic or expression QTL were detected due to the overall much lower number of traits.

3.4.3. QTL co-localisation analyses across the three omics-layers

Genome-wide association analyses were performed for all generated data sets. A total of 102,346 marker-trait associations (MTAs) were detected with a $p\text{-value}_{FDR} \leq 0.05$ (Table 4). Co-localisation of associations was investigated between the three omics-layers of transcriptome, metabolome and phenome (including end-point biomass and growth rates; Figure 13 a). As the medians of PVE for these mQTL, eQTL and QTL were 1.41 %, 1.31 % and 1.37 %, respectively, and the median significant $p\text{-values}_{FDR}$ were in all cases around 0.01, in a first step associations were filtered for

p-values $_{FDR}$ ≤ 0.01 and PVE ≥ 2 %, and a co-localisation analysis was performed with the selected MTAs. Traits were regarded as co-localised when significant associations with the very same marker were detected. This threshold combination resulted in a total of 16 markers with co-localised mQTL, eQTL and QTL. Furthermore, different *p*-values $_{FDR}$ and PVE combinations were tested. The combination of PVE ≥ 5 % and *p*-values $_{FDR}$ ≤ 0.05 resulted in no co-localisation at all, while a filter with the thresholds *p*-values $_{FDR}$ ≤ 0.001 and PVE ≥ 1 % yielded a total of 9 markers with co-localising traits of the different omics-layers. Taking these results into account, the *p*-values $_{FDR}$ threshold was loosened to ≤ 0.05 and associations of all three sets were pre-filtered for PVE ≥ 2 %, resulting in 31,264 associations (n_{QTL} = 4,667, n_{eQTL} = 26,391 and n_{mQTL} = 206). With these thresholds, at 32 markers co-localisation between all three omics layers were observed. As the number of quantitative traits analysed, in particular the number of transcripts, is higher than the number of markers used for the genome wide-association analyses, it is likely that co-localisation of traits might be found by random chance. To quantify the degree of random co-localisation, permutation analyses were performed. To this end, detected associations were randomly shuffled over all marker positions and the 95 % quantile was calculated and compared to the actual number of detected co-localisations. Under the assumption of a random and equal distribution of QTL over the whole genome, substantially more co-localised QTL were detected (n= 32), than expected by random chance (n= 13, Figure 13 b).

Table 4. List of marker-trait associations for each data set at different significance levels

Data set	Number of traits	MTAs at *p*-value $_{FDR}$ ≤ 0.05	MTAs filtered for *p*-value $_{FDR}$ ≤ 0.05 and ≥ 2 PVE% [d]	MTAs filtered for *p*-value $_{FDR}$ ≤ 0.01 and ≥ 2 PVE% [d]	MTAs filtered for *p*-value $_{FDR}$ ≤ 0.001 and ≥ 1 PVE% [d]	MTAs filtered for *p*-value $_{FDR}$ ≤ 0.05 and ≥ 5 PVE% [d]
Metabolites (M)	154	n= 544	**n= 206**	n= 132	n= 106	n= 38
Transcripts (T) [a]	19,479	n= 86,013	**n= 26,391**	n= 20,966	n= 19,659	n= 11,346
Phenotypic traits (P) [b]	2,689	n= 15,770	**n= 4,659**	n= 2,786	n= 1,927	n= 778
End-point biomass (P)	2	n= 19	**n= 8**	n= 6	n= 4	n= 4
Co-localizations (T, M, P)			**n= 32**	n= 16	n= 9	n= 0
Permutation threshold [c]			**n= 13**			

[a] filtered for median over all samples ≥ 5 tpm

[b] including image-derived traits for 21 time points (6-27 DAS) and calculated relative growth / absolute change rates

[c] estimated number of random co-localisations obtained by 10,000 permutations of the data set

[d] estimated percentage of phenotypic variance explained by the genetic marker

Notably, the deletion marker ‘Bn-scaff_28509_1-p33495_del’, which is likely part of a larger deletion, co-localised with a hotspot of 153 eQTL (filtered by *p*-values $_{FDR} \leq 0.05$ and PVE ≥ 2 %) on chromosome C03 (position 26.6 Mb). Furthermore, this hotspot co-localised with an mQTL for glucose and six QTL for colour-related traits based on visible-light top view images. The deletion (‘Bn-scaff_28509_1-p33495_del’) was only detected in ten closely related lines. These lines display severely reduced transcript levels for many genes within the approx. 32 Mb large region (nearly 40 % of chromosome C03), confirming the large (potentially heterozygous) deletion or a series of deletions. Moreover, these lines display a significantly higher glucose level (Welch Two Sample *t*-test; *p*-value= 4.668e-07, Figure S20). In an attempt to prioritise the transcripts within the eQTL hotspot, correlations between transcripts and metabolites were used. For the transcripts within the eQTL hotspot no significant correlation to glucose could be detected. However, some genes were annotated with GO terms related to glucose metabolism (‘BnaA03g27500D’, a putative v-type proton ATPase subunit and ‘BnaC09g39650D’, a probable galactose-1-phosphate uridyltransferase) or response to glucose stimulus (‘BnaC03g37820D’, a putative phosphoglycerate kinase). Another co-localisation was detected for SNP marker ‘Bn-scaff_19244_1-p313887’ on chromosome C01 (position 492 kb). This hotspot includes an mQTL for β-Alanine, 11 eQTL and another QTL for a colour-related trait. The highest correlated (r= -0.574) among the transcripts was ‘BnaC01g00550D’, encoding an alanine-glyoxylate aminotransferase. Another notable example is the deletion marker ‘Bn-scaff_19026_1-p369147_del’ on chromosome C03 (position 36.3 Mb) including mQTL for fructose-6-phosphate, glucose-6-phosphate, sucrose and three metabolites of unknown structure, 44 eQTL forming a smaller hotspot and two QTL, one for a colour-related trait and one for the 'skewness' of the brightness values of the plant pixels (Figure 13 a). The highest correlated transcript within this hotspot (r= -0.5716) was ‘BnaC03g39190D’, which is correlated with sucrose and annotated with the GO terms ‘response to sucrose stimulus’ and ‘response to fructose stimulus’. The transcript shows homology to the Arabidopsis gene *AT3G15630* encoding a protein of unknown function.

Focusing only on mQTL and eQTL, a total of 127 markers with co-localisations, 68 involving metabolites of known chemical structure, were identified. Among them are examples with clear links between metabolism and transcript function like ‘Bn-A04-p16744035_del’ linking mQTL for leucine and isoleucine to an eQTL of BnaA04g29360D, encoding a putative amino acid transporter

family protein, or ‘Bn-A06-p9119114’ linking an mQTL for valine to an eQTL for ‘BnaA06g15650D’, encoding a potential branched-chain-amino-acid aminotransferase, but also many other co-localisations with no obvious link between transcripts and metabolism. When broadening the co-localisation analysis to all QTL for biomass (FW) regardless of the explained phenotypic variance and filtered eQTL, eight markers with co-localised eQTL and biomass QTL were detected. Notably, among them were two transcripts (‘BnaA07g23560D’ and ‘BnaCnng17010D‘) with eQTL on chromosome C02 (deletion marker ‘Bn-scaff_16116_1-p487063_del’) that were also detected among the ten top-related transcripts in the network analysis with GENIE3.

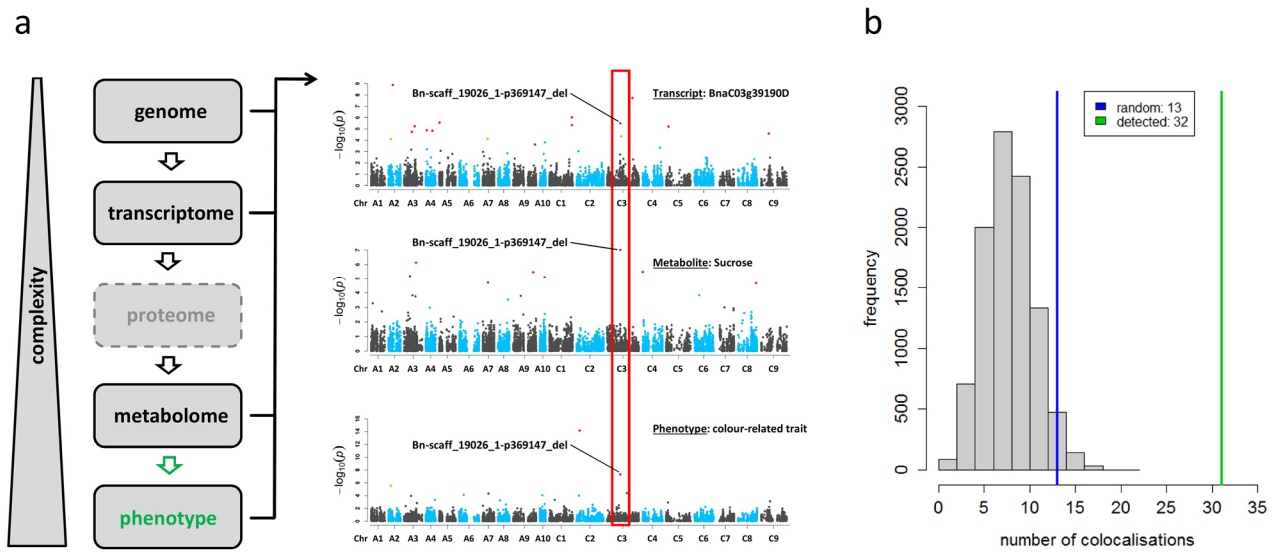


Figure 13. Co-localisation of marker-trait associations across the omics-cascade

Panel **a** displays an example for QTL co-localisation across different omics-layers, including the transcriptome, metabolome and phenome. Associations were filtered for *p*-values $_{FDR} \leq 0.05$ and PVE ≥ 2 % prior to analysis. The genetic marker ‘Bn-scaff_19026_1-p369147_del’, a deletion on chromosome C03, was found to be associated with 6 metabolites, 44 transcripts and two phenotypic traits. Exemplarily shown are the Manhattan plots for the transcript of ‘BnaC03g39190D’ (top), the metabolite sucrose (middle) and the colour-related trait ‘top.intensity.vis.hsv.h.histogram.v_avg.bin.14.165_178’ at 23 DAS (bottom). HSV represents a colour model defined by *Hue*, *Saturation* and *Value*. The trait describes the average colour intensity derived by dividing the HSV *hue* by the HSV *value* of the 14th colour bin. The histogram in panel **b** shows the distribution of the maximum number of co-localisations, obtained from a permutation analysis with 10,000 iterations. The blue and the green vertical lines correspond to the determined number of random (0.95 quantile of the distribution; n= 13) and observed (n= 32) co-localisations, respectively.

3.5. Strong temporal dynamics of QTL action on plant growth in canola

A major challenge of plant biology is to unravel the genetic basis of quantitative (complex) traits. Taking advantage of recent technical advances in high-throughput phenotyping in conjunction with genome-wide association studies, genotype-phenotype relationships were elucidated at high temporal resolution. The following chapter is a revised version of the results and discussion part of the research article 'Strong temporal dynamics of QTL action on plant growth progression revealed through high-throughput phenotyping in canola' published in Plant Biotechnology Journal, May 24th 2019.

3.5.1. Capturing of dynamic growth by high-throughput phenotyping

The diverse spring-type canola breeding population consisting of 477 genotypes with 'double-low' seed quality (low erucic acid, low glucosinolate content), was investigated at an early vegetative growth phase. Automated high-throughput phenotyping was applied daily using the previously described IPK phenotyping platform for large plants (Junker *et al.*, 2015), and image analysis was performed with the in-house image analysis pipeline (IAP) to derive estimations of growth related traits at multiple time points (Klukas *et al.*, 2014). Examples of acquired raw plant images are provided in Figure S2. The following section focuses on the detailed analysis of a sub-selection of four growth-related traits, over the time course of the whole experiment. After quality checks, estimates of biovolume, projected leaf area, early plant height as well as colour uniformity were obtained for 21 consecutive time points from 6 to 27 DAS, covering approximately the first growth phase of rapeseed development from completely unfolded cotyledons to four or more unfolded leaves. All four traits showed broad phenotypic variation resulting in medium to high coefficients of variation (Data S3), with highest values for biovolume and lowest values for colour uniformity. Biovolume and projected leaf area displayed exponential increases over time, while early plant height increased in a linear manner. Colour uniformity increased during the first days, but remained at an almost constant level during the later phase (Figure S21 a-d). Image derived phenotypes were complemented by manually determined end-point fresh weight (FW) and dry weight (DW) values at 28 DAS (Figure S22 a). Both, fresh and dry weight were strongly correlated (r= 0.969, Figure S22 b) and highly correlated with the image-derived biovolume estimates at the

latest time point (27 DAS), with r= 0.929 and r= 0.926 for FW and DW, respectively (Figure S22 c-d and Data S9). These high correlations indicate that biovolume estimates can serve as a suitable proxy for the actual plant biomass. To assess the repeatability and quality of the phenotypic data, broad-sense heritabilities (H^2) were estimated (Figure S14, Data S3). Over the whole experiment, H^2 for image-derived phenotypes ranged between 0.528 (early plant height at 15 DAS) and 0.874 (projected leaf area at 26 DAS). High H^2 values of 0.895 and 0.878 were also obtained for fresh and dry weight, respectively, facilitating the temporal analysis of trait relationships and forming a solid basis for genetic analyses.

3.5.2. Predominantly minor and medium effect QTL contribute to growth

BLUEs of image-derived phenotype data for projected leaf area, estimated biovolume, early plant height and colour uniformity at 21 time points, as well as manually determined biomass (FW and DW) at 28 DAS were used for genome-wide association studies using Fixed and Random Model Circulating Probability Unification, implemented in the 'FarmCPU' R package (Liu *et al.*, 2016). For manually determined biomass, 22 significant marker-trait associations (MTAs) were detected at a $p\text{-value}_{(FDR)} \leq 0.1$ (Figure 14 and Table 5), with thirteen and nine MTAs for fresh weight (Figure 14 a) and dry weight (Figure 14 b), respectively. This moderate *p*-value threshold was applied for comparability between QTL for end-point biomass and time-resolved traits. Despite the high phenotypic correlation (r=0.969), only three shared MTAs for FW and DW, one on chromosome A10 and two on C02, were identified.

Genome-wide association analyses performed for data measured at all 21 time points with the moderate threshold ($p\text{-value}_{(FDR)} \leq 0.1$) revealed a total of 787 MTAs, including 191 associations for estimated biovolume, 200 MTAs for projected leaf area, 182 MTAs for early plant height, and 192 MTAs for colour uniformity. A moderate *p*-value threshold was chosen in the first step, as subsequently another filter was applied to enrich for QTL of interest. There were no substantial differences in the number of associations between the A and the C subgenomes. The majority of detected associations could be attributed to unique, single-copy SNP markers (84 % of all associations). A substantial number of CNVs (deletions and duplications) also showed trait associations independently of the two SNP alleles (Grandke *et al.*, 2017; Mason *et al.*, 2017). To reduce the list to robust candidate regions, detected MTAs were further filtered to retain only

loci showing significant associations for at least three consecutive time points (Data S11). Most of the detected MTAs explained only a small percentage of phenotypic variance (< 5 PVE%, Figure S23) and were equally distributed over the subgenomes. Only 40 (3.8 %) marker-trait associations with larger effects (> 5 PVE%) were detected, for example Bn-A04-p4409752 explaining up to 8.64 % phenotypic variance of biomass (fresh weight).

Table 5. Information about markers associated with end-point biomass

Trait	Marker ID	Chromosome	Position (bp)	MAF	*p*-value	*p*-value (FDR)	Effect	PVE% [a]
fresh weight	Bn-A04-p4409752	A04	5,462,587	0.4937	1.89E-08	0.0002	1.0762	8.64
fresh weight	Bn-A01-p7850092	A05	1,595,585	0.0430	8.62E-07	0.0028	-2.6757	4.01
fresh weight	Bn-A07-p9632473	A07	15,644,870	0.2715	1.05E-05	0.0191	-0.9127	5.07
fresh weight	Bn-A07-p5114831	A08	5,664,005	0.3637	5.54E-08	0.0003	0.9170	1.75
fresh weight	Bn-A08-p16771030	A08	26,455,071	0.2966	4.72E-09	0.0001	-1.2357	1.37
fresh weight	Bn-A10-p10672359	A10	10,601,845	0.3019	9.28E-06	0.0191	0.8144	0.02
fresh weight	Bn-A10-p13343454 [b]	A10	12,120,357	0.2117	1.48E-07	0.0006	-1.1502	2.16
fresh weight	Bn-scaff_16116_1-p487063_del	C02	25,078,453	0.0629	3.01E-05	0.0377	-0.8749	0.33
fresh weight	Bn-scaff_18702_1-p589589	C02	27,593,710	0.0639	8.45E-06	0.0191	-1.4439	1.21
fresh weight	Bn-scaff_16545_1-p862530 [b]	C02	50,263,120	0.4874	1.05E-05	0.0191	-0.8823	1.51
fresh weight	Bn-scaff_21705_1-p175010 [b]	C02	54,034,064	0.3344	1.60E-05	0.0261	0.8604	1.33
fresh weight	Bn-scaff_16200_1-p503123	C07	17,183,655	0.3176	2.52E-05	0.0342	1.5186	0.57
fresh weight	Bn-scaff_16197_1-p3022518	C08	41,613,071	0.2809	1.92E-05	0.0285	0.7434	0.44
dry weight	Bn-A04-p9426523	A04	13,935,829	0.1908	1.71E-05	0.0466	-0.0705	2.11
dry weight	Bn-A10-p11817272	A07	2,411,921	0.2002	1.63E-06	0.0066	-0.0755	4.85
dry weight	Bn-A10-p13343454 [b, c]	A10	12,120,357	0.2117	3.05E-05	0.0709	-0.0623	3.56
dry weight	Bn-scaff_16804_1-p178142	C02	9,108,149	0.1122	1.07E-07	0.0009	-0.1186	5.08
dry weight	Bn-scaff_16545_1-p862530 [b, c]	C02	50,263,120	0.4874	4.81E-05	0.0872	-0.0411	0.00
dry weight	Bn-scaff_21705_1-p175010 [b]	C02	54,034,064	0.3344	4.03E-07	0.0022	0.0736	1.48
dry weight	Bn-scaff_21312_1-p895326	C03	11,220,963	0.0398	8.64E-09	0.0001	0.2735	5.91
dry weight	Bn-scaff_15766_1-p117110	C07	14,697,010	0.2904	6.13E-06	0.0200	0.1230	1.47
dry weight	Bn-A10-p10083397_del [c]	C09	59,994,601	0.0273	3.70E-05	0.0755	0.1283	0.89

[a] estimated percentage of phenotypic variance explained by the genetic marker

[b] common marker-trait associations (MTAs) shared between fresh & dry weight

[c] *p*-value (FDR) ≤ 0.1

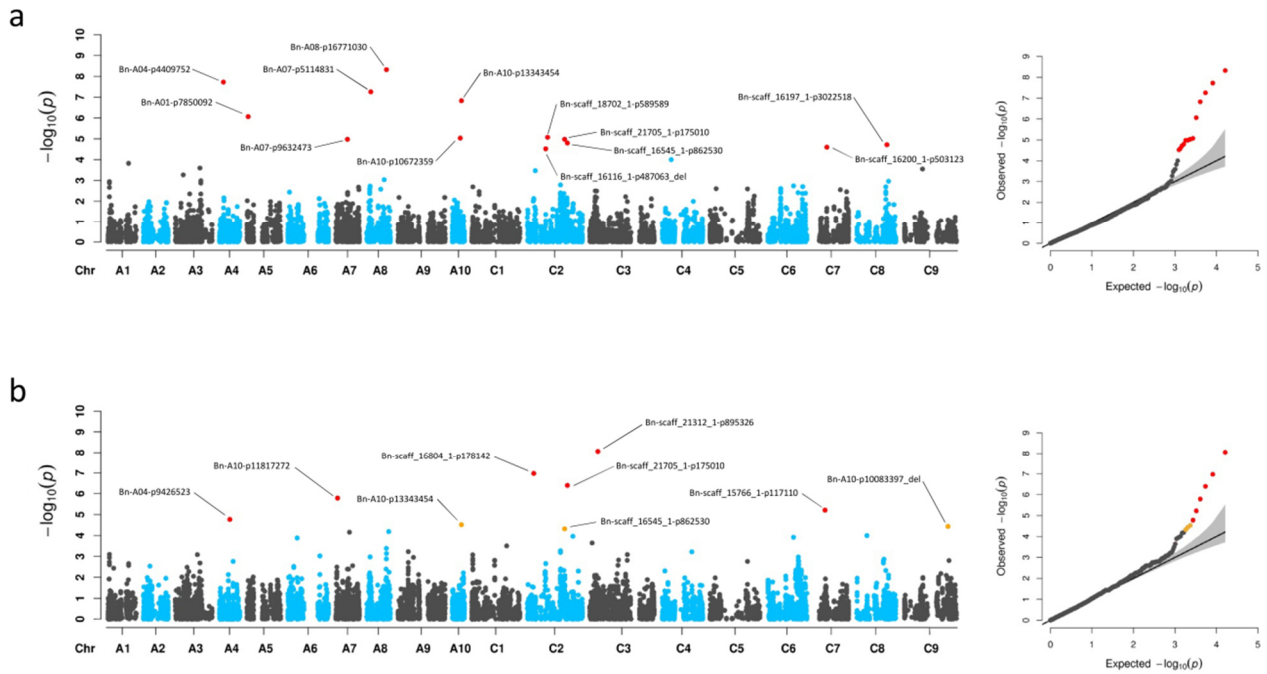


Figure 14. Genome-wide marker-trait associations for end-point biomass

a Manhattan plot (left) and quantile-quantile plot (right) for fresh weight (FW) at 28 days after sowing (DAS). **b** Manhattan plot (left) and quantile-quantile plot (right) for dry weight (DW) at 28DAS. GWAS was performed in R 'FarmCPU' on BLUEs estimated using three replicates (carriers) with five plants each. Significant marker-trait-associations (MTAs) are shown with marker-IDs. MTAs with *p*-values $_{FDR} \leq 0.05$ or 0.1 are indicated by red and orange dots, respectively.

3.5.3. Identification of dynamic growth QTL in canola

The time-resolved design of the phenotyping experiments enabled the tracking of the effects of individual markers over the course of 21 days of early growth between six and 27 DAS. In summary, 14, nine, four and thee MTAs for projected leaf area, estimated biovolume, early plant height and colour uniformity were detected to be significant at three consecutive days, respectively (Figure 15, Data S11). To further address the dynamic nature of these traits, relative growth rates (RGRs) for projected leaf area, estimated biovolume and early plant height, as well as absolute change rates (ACRs) for colour uniformity were calculated over 15 intervals of three days to integrate the effects over longer periods (Figure S21 e-f).

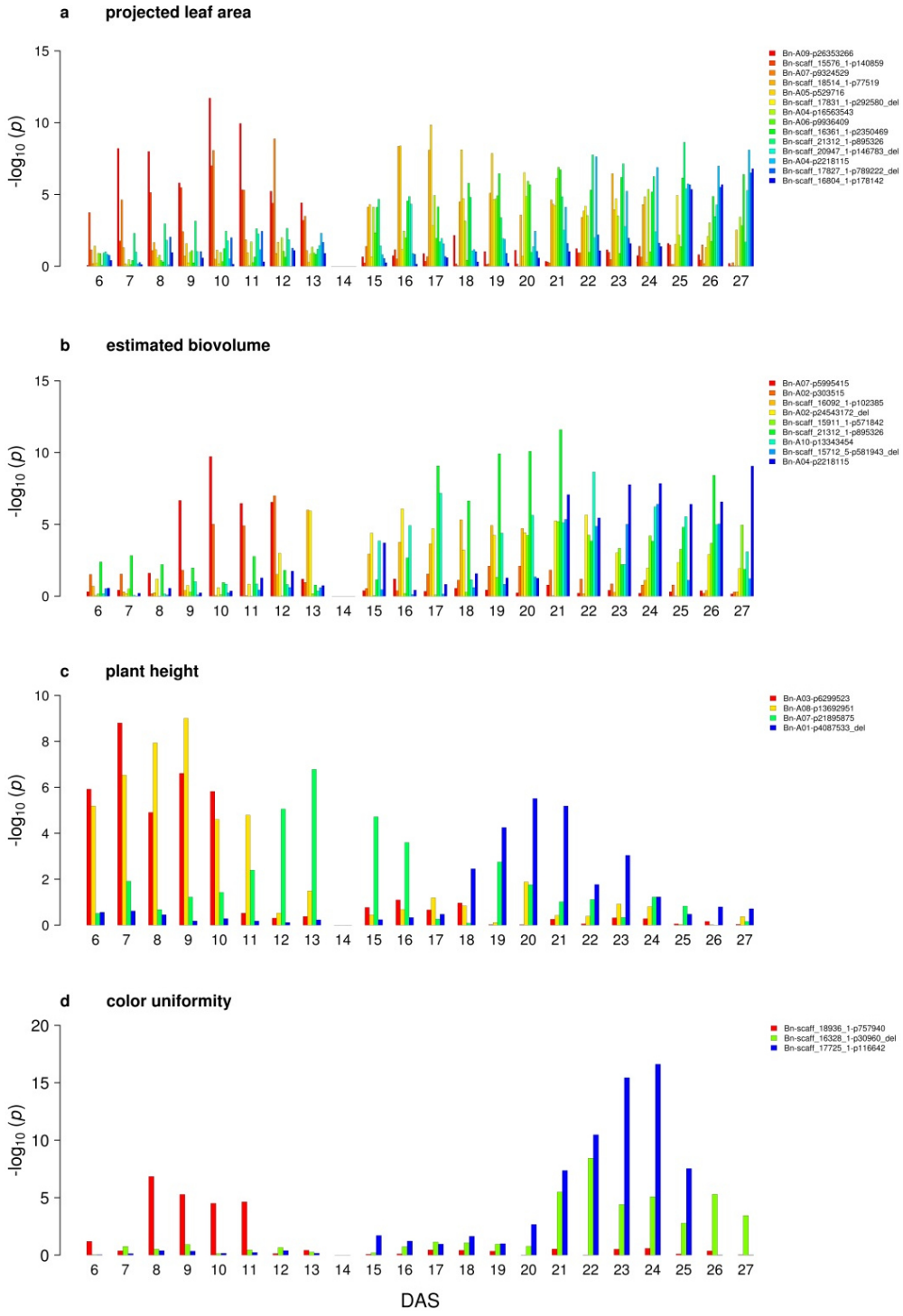


Figure 15. Dynamic associations detected during cultivation from 6 to 27 days after sowing

GWAS was performed on BLUEs of **a** projected leaf area, **b** estimated biovolume, **c** early plant height and **d** plant colour uniformity in R / package 'FarmCPU'. Different colours indicate markers with *p*-value $_{(FDR)} \leq 0.1$ at three consecutive days, with the colour gradient corresponding to the temporal pattern. DAS denotes days after sowing. BLUEs were estimated using three replicates (carriers) with nine and five plants for 6 to 13 DAS and 15 to 27 DAS, respectively. No data were recorded at 14 DAS due to sampling of shoot material.

Highest relative growth rates, especially for plant height, were detected at the beginning of the cultivation with a decreasing trend over time attributable either to an actual decrease in growth or to a bias due to overlapping leaves. Absolute change rates for colour uniformity were more stable than the relative growth rates during development. Growth rates were subsequently mapped with the same approach as the single time point data. Moreover, GWAS was successfully applied to RGR traits of multiple successive time intervals resulting in the detection of a total of 268 significant marker-trait associations (MTAs). In summary, 100 MTAs for biovolume RGRs, 76 MTAs for leaf area RGRs, 73 MTAs for plant height RGRs, and 19 MTAs for the colour uniformity changes were detected for the individual intervals.

To focus on particularly robust MTAs, the growth rate associations were further filtered for at least two consecutive significant intervals, as done previously for the *per se* traits at individual time points. For colour uniformity ACRs, no consecutive significant associations were found. Two MTAs for leaf area RGRs at intermediate growth intervals, four MTAs for biovolume RGRs distributed evenly over the entire examined growth period, and two MTAs for plant height at a very early phase were detected (Figure 16). Allelic effects of loci did not only increase and decrease with time, tending to diminish after a certain interval, but for a substantial fraction of MTAs (16/30 for *per se* trait MTAs and even 8/8 for RGR MTAs), allele effects reversed over time (Figure S24 and S25).

3.5.4. Shared associations and novel candidate genes for growth dynamics

Among the 36 identified markers displaying temporal dynamic patterns, nine were shared between different traits. In particular ‘Bn-A10-p13343454’ showed association with projected leaf area, estimated biovolume and early plant height, as well as with fresh weight and dry weight. The marker ‘Bn-scaff_21312_1-p895326’ was associated with projected leaf area, biovolume and dry weight, while ‘Bn-scaff_16804_1-p178142’ was shared between projected leaf area and dry weight. The other six SNP and CNV markers: ‘Bn-A02-p24543172_del’, ‘Bn-A04-p2218115’, ‘Bn-scaff_15911_1-p571842’, ‘Bn-scaff_16361_1-p2350469’, ‘Bn-scaff_17831_1-p292580_del’ and ‘Bn-scaff_20947_1-p146783_del’ were associated with both, projected leaf area and biovolume.

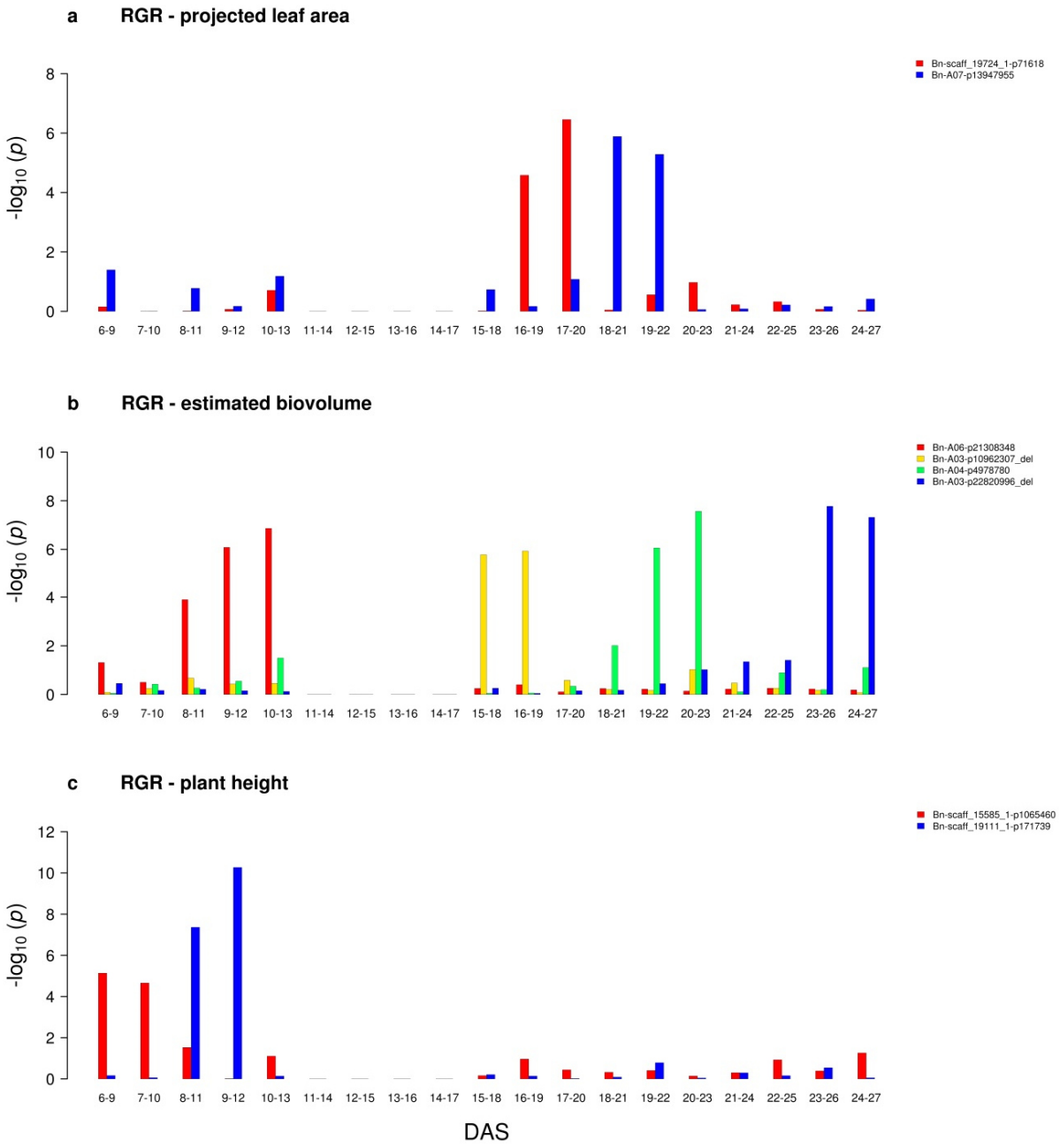


Figure 16. Dynamic associations detected for relative growth rates

GWAS was performed on BLUEs of **a** relative growth rates for projected leaf area, **b** relative growth rates for estimated biovolume and **c** relative growth rates for early plant height in R / package 'FarmCPU'. Different colours indicate markers with *p*-value $_{(FDR)} \leq 0.1$ at two consecutive intervals. DAS denotes days after sowing. BLUEs were estimated using three replicates (carriers) with nine and five plants for 6 to 13 DAS and 15 to 27 DAS, respectively. No data were recorded at 14 DAS due to sampling of shoot material.

From these nine markers, a subset of five was selected based on the number of associations and traits for detailed analysis. Candidate genes were identified in the corresponding regions on chromosomes A04, A10, C02, C03 and C08 (Table 6, Figure S26) by an LD-based confidence interval approach. Genes were selected within LD blocks ($r^2 \geq 0.6$) as shown for candidate region 5 on chromosome C08 (Figure 17), where the significantly associated marker 'Bn-scaff_16361_1-p2350469' forms an LD block with four of its neighbouring SNPs. The block spans a region of 368 kb and contains 72 genes, of which seven were selected as putative candidates based on their annotation: the citrate synthase CSY2; the MADS-box transcription factor SHP1; PAR2 involved in the brassinosteroid mediated signalling pathway; the pectinesterase PME35 implicated in cell wall modification; the bHLH transcription factor PIF5; the tetrapyrrole-binding protein GUN4 which regulates chlorophyll synthesis; and the flowering time control protein FPA also annotated to be involved in cell differentiation. In case of the absence of detectable LD, genes were selected in the 100 kb flanking regions on either side of the significant marker as suggested by Zhou and Han *et al.* (2017). A comprehensive list of all thus identified candidate genes for all evaluated traits can be found in Data S6.

Among the 361 genes in the intervals 1 to 5, 30 genes were selected as particularly interesting candidates based on their annotation and gene ontology (GO). Nine of these genes have annotations related to meristem development and cell growth, including S*epallata1* (BnaA10g18480D), *Longifolia1* (BnaA10g18650D), *Squamosa promoter binding 3* (BnaC03g18800D) and S*hatterproof1* (BnaC08g29530D). Several other genes are putatively involved in flowering time or cell wall biogenesis and modification, or were annotated as transcription factors. Further examination of the associated markers of the five candidate regions revealed that the allele distribution differs between the three breeding pools: for example, for 'Bn-A04-p2218115' the minor allele is underrepresented in breeding pool 2 and completely absent in breeding pool 3. In contrast, the minor allele of 'Bn-scaff_21312_1-p895326' is nearly absent in breeding pools 1 and 2, but although only present in the heterozygous state, is highly overrepresented in breeding pool 3.

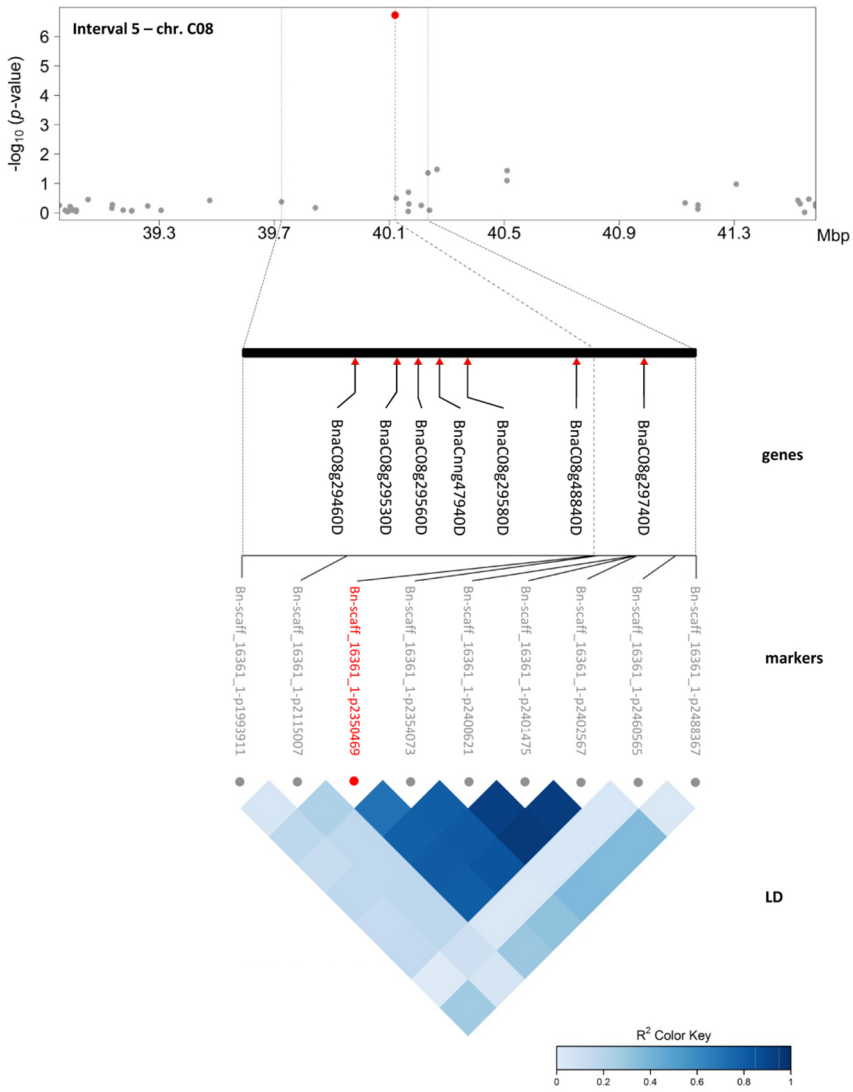


Figure 17. Manhattan plot for a representative marker-trait association in the candidate region 5 on Chr. C08 with selected candidate genes and correlations between markers

The Manhattan plot describes the marker-trait associations for the candidate region 5 on chromosome C08. The trait 'projected leaf area at 21 DAS' is shown as a representative trait for the 14 traits associated with the marker 'Bn-scaff_16361_1-p2350469' (Data S11). The significant associated SNP is indicated by a red dot. Grey dots represent surrounding non-significant markers in the region. Please note that the FarmCPU GWAS method, which iteratively uses fixed and random effect models and pseudo QTN as covariates, results in a different appearance of the Manhattan plots. Significant associations are illustrated by 'helicopters' rather than 'skyscrapers', see materials and methods. For reasons of clarity and comprehensibility, the zoom-in of the candidate region was extended to the next flanking SNP markers ('Bn-A04-p1895018' and 'Bn-A04-p2094818'). Red triangles indicate the positions of selected candidate genes (Table 6). The LD heatmap in the bottom section shows the correlations (r^2) between surrounding SNP markers. The markers 'Bn-scaff_16361_1-p2350469', 'Bn-scaff_16361_1-p2354073', 'Bn-scaff_16361_1-p2400621', 'Bn-scaff_16361_ 1-p2401475' and 'Bn-scaff_16361_1-p2402567' form an LD block ($r^2 \geq 0.6$).

Table 6. List of candidate regions and selected candidate genes

Interval	Marker	Chr.	Pos. (bp)	LD-block [a]	Interval start (bp)	Interval stop (bp)	Interval size (bp)	Number of genes	Number of MTAs	Traits	Selected candidates [b]	Arabidopsis homologe / putative function [c]
1	Bn-A04-p2218115	A04	2,103,821	no	2,003,821	2,203,821	200,000	44	14	leaf area, biovolume,	Bra014695	ARR17 / two-component response regulator
											BnaA04g02550D	WRKY55 / transcription factor, WRKY
											BnaA04g02600D	ANAC064 / transcription factor, NAC domain
											BnaC05g07680D	ANAC064 / transcription factor, NAC domain
2	Bn-A10-p13343454	A10	12,120,357	no	12,020,357	12,220,357	200,000	54	16	FW, DW, leaf area, biovolume, plant height	BnaA10g18330D	NIK1 / protein phosphorylation
											BnaA10g18440D	BZIP3 / transcription factor, basic-leucine zipper
											BnaA10g18480D	SEPALLATA1 / transcription factor, MADS-box
											BnaA10g18530D	RRT1 / O-fucosyltransferase, pectin biosynthetic process
											BnaA10g18590D	RGP2 / UDP-arabinose mutase, cell wall biogenesis
											BnaA10g18600D	COBRA-like protein / cell wall biogenesis
											BnaA10g18650D	LONGIFOLIA1 / regulation of cell growth
3	Bn-scaff_16804_1-p178142	C02	9,108,149	yes	8,497,706	10,116,820	1,619,114	148	9	DW, leaf area, biovolume, plant height, plant colour uniformity	BnaC02g11320D	SMAX1 / hydrolase, seedling development
											BnaC02g11400D	COL5 / transcription factor, zinc finger (B-box type)
											BnaC02g11890D	AT1G50890 / cell growth
											BnaC02g44440D	ZEP2 / transcription factor, zinc finger
											BnaC02g44470D	IAA33 / auxin-activated signaling pathway
											BnaC02g11520D	AIL5 / transcription factor, post-embryonic development
											BnaC02g11970D	AT5G56960 / transcription factor, bHLH
											BnaC02g12210D	GULLO4 / oxidoreductase activity, cell wall biogenesis
											BnaC02g12340D	EXP14 / α-expansin, cell growth
											BnaC03g72190D	CRK / CDPK-related kinase, ABA-activated signaling pathway
4	Bn-scaff_21312_1-p895326	C03	11,220,963	no	11,120,963	11,320,963	200,000	43	16	DW, leaf area, biovolume	BnaC03g18580D	IAA13 / auxin-activated signaling pathway
											BnaC03g18800D	SPL3 / transcription factor, SBP-box
5	Bn-scaff_16361_1-p2350469	C08	40,120,568	yes	39,843,456	40,211,817	368,361	72	14	leaf area, biovolume	BnaC08g29460D	CSY2 / citrate synthase
											BnaC08g29530D	SHP1 / transcription factor, MADS-box
											BnaC08g29560D	PAR2 / brassinosteroid mediated signaling pathway
											BnaCnng47940D	PME35 / pectinesterase, cell wall modification
											BnaC08g29580D	PIF5 / transcription factor, bHLH
											BnaC08g48840D	GUN4 / tetrapyrrole binding, chlorophyll biosynthetic process
											BnaC08g29740D	FPA / flowering time control, cell differentiation

[a] in case of the [a] in case of the absence of an LD-block flanking 100 kb regions on either side of the associated marker were screened for candidate genes;

[b] best match using BLASTn of transcript sequences to the *B. napus* Darmor-*bzh* v.4.1, the concatenated *Brassica* AC, and the *A. thaliana* TAIR10 transcriptomes; a full list of de-novo annotated genes within the five intervals, BLASTn results, descriptions and functional annotations (BLAST2GO) is available in Data S6

[c] closest homolog in *Arabidopsis thaliana*; putative function (selection) obtained from the Brassica (BRAD) & the TAIR databases

4. Discussion

Based on previous work on biomass and heterosis prediction in maize (Riedelsheimer *et al.*, 2012*a*) and the model crucifer and close *Brassica* relative *Arabidopsis thaliana* (Meyer *et al.*, 2007; Steinfath *et al.*, 2010), the study was built on the hypothesis that specific allelic combinations of regulatory genes, the gene expression of their targets, as well as elicited metabolite profiles, are associated with variation in vegetative growth and seed yield in hybrids. The study pursued three goals: first, to evaluate models for hybrid prediction in spring oilseed rape by using and combining information of multiple omics-layers; second, to relate vegetative growth to gene expression and metabolite content levels by correlation analyses; and third to perform genome-wide association (GWAS) and co-localisation studies to identify candidate genetic loci / genes associated with trait variation, in particular vegetative growth and biomass accumulation.

For this purpose, a hybrid population with 950 hybrid lines had been established and evaluated in field trials. Complementarily, detailed phenotyping data were generated by growing the parental lines of the hybrids (475 diverse pollinators from a commercial canola breeding programme and two elite male-sterile tester lines), as well as selected hybrids in the IPK automated high-throughput non-invasive phenotyping platform (Junker *et al.*, 2015). Image-derived trait data obtained at an early vegetative stage were complemented by polar metabolite and global transcriptome (RNA-Seq) profile data from the same individual plants. The extensive -omics data sets gathered from the parental lines were used individually and in combination for prediction of hybrid performance to evaluate if additional omics-information can be utilised to further improve genomic predictions. Two types of prediction methods were employed and compared, (genomic) best linear unbiased predictions (gBLUP) and reproducing kernel Hilbert space regression (RKHS). Furthermore, prediction of hybrid performance was broadened to data gathered from the glasshouse for a subset of hybrids and the extent of hybrid vigour (heterosis) was analysed. The comprehensive data sets generated were further utilised for correlation analyses, and in combination with array-derived SNP and CNV data for genome-wide-association studies (GWAS) to identify genetic loci associated with trait variation and candidate genes.

4.1. Generation of extensive -omics data sets

Prior to this work, an F_1 hybrid population with 950 hybrid lines had been generated and evaluated in extensive field trials for multiple agronomically important traits by the commercial project partners in the growing season of 2012. These experiments were designed as unreplicated nested field trials and were performed at eight different locations across Europe to test the hybrids under diverse environmental conditions. Each genotype was tested at three to four locations across Europe. This type of experimental design is common for commercial field studies of rapeseed, as it would be too expensive for the breeding companies to test all hybrid combinations in all environments. However, the same set of four reference lines (Achat', 'Osorno', 'Mirakel' and 'DLE 1108') was included in all field trials at all locations to compare and relate the different trials, to dissect the genotypic and the environmental components of trait variation, and to calculate BLUEs across the different field experiments.

Multiple traits of agronomic importance were scored in the different environments. However, not all traits were consistently scored in all trials and in all environments, resulting in missing values within the entire data set. Traits of particular interest to the breeders were scored at multiple locations including seed yield, the content of total seed glucosinolate, the days to onset of flowering, the seed oil yield, seedling emergence, the seed oil content and the seed protein content. Seed yield was the only trait scored at all locations. Additional less replicated traits including Alternaria (*Alternaria brassicae*) and Sclerotinia (*Sclerotinia sclerotiorum* and *S. minor*) resistance, moisture content, standability, straw length, linolenic or oleic acid content were scored in a single or only few of the environments and therefore excluded from the subsequent analyses.
The calculation of adjusted values (BLUEs) was necessary, as raw data showed location effects. In particular the traits seed oil yield, days to flowering and seedling emergence displayed a bimodal data distribution. BLUEs of the seven traits followed an approximate normal distribution (Figure 13), displayed moderate to high coefficients of variation (CV: 0.84 % to 20.82 %) and heritability (H^2: 0.34 to 0.92) making the data suitable for the follow-up analyses. Substantial variation in flowering time has been previously described between different years (Wang *et al.*, 2011*a*). Hence, differences in flowering time and seedling emergence between the different test locations across Europe with varying climate conditions were expected.

Moreover, the content of seed GSL has been shown to be largely regulated by environmental factors in rapeseed, and to be correlated with flowering time (Fu *et al.*, 2015; He *et al.*, 2018).

Complementary to the agronomic traits recorded for the hybrids in the field, extensive -omics data sets for the 477 parental lines were generated. These data included array-derived genotype data (SNPs and CNVs), global transcriptome (RNA-Seq) profiles, polar primary metabolite (GC-MS) profiles, as well as detailed high-throughput image-derived phenotyping data. The population used in this study featured two characteristics, on the one side a reasonable size and diversity making it suitable for genome-wide association studies, and on the other side it constituted heterogeneous breeding material, including F_3 to F_6 lines, BC_1F_4 lines, open pollinated DH lines and elite lines, making results transferable to ongoing breeding programmes. However, in contrast to pure inbred lines, one disadvantage of the material is its complexity and high heterozygosity impeding genetic analyses. The STRUCTURE analyses indicated the presence of population structure. The three major clusters coincide to a substantial degree with the breeding pools, but many individuals show pronounced admixture. These results are consistent with the finding that breeding pools do not necessarily fully reflect the genetic structure of a population, especially if they are not yet firmly established and have only undergone very few cycles of diversifying selection. Given the ambiguous results of the STRUCTURE analyses, the explained phenotypic variance was considered in addition and the first four PCs (> 3 % explained variance) were incorporated in the GWAS analyses.

The parental lines were genotyped using the *Brassica* Infinium 60k genotyping array (Clarke *et al.*, 2016; Mason *et al.*, 2017) and genotype data for 19,674 SNP markers were obtained and further filtered to 13,201 unique, single-copy SNPs (MAF ≥ 0.01 and less than 10 % missing calls). Although SNP markers cover all chromosomes, marker hotspots, for example on chromosome C05, and larger monomorphic regions (without SNPs), e.g. on chromosome A06 (18.7 Mb), were detectable. The former might be due to the design of the array, e.g. by an unequal distribution of the oligo-nucleotide probes across the *B. napus* genome. The latter can be attributed to the fact that the new *B. napus* cv. Darmor-*bzh* reference gene assembly (NRGene, assembly size of approx. 1,047 Mb) used for the positioning of the SNPs most likely contains larger repetitive regions not included

in the previous Darmor-*bzh* v4.1 reference (Chalhoub *et al.*, 2014, assembly size of approx. 850 Mb). This notion is supported by the fact that the shorter reference sequence covered only 79 % of the 1,130 Mb *Brassica napus* genome and included 95.6 % of *Brassica* expressed sequence tags (ESTs). In addition to single nucleotide polymorphisms (SNPs), copy number variation (CNV) and presence-absence variation (PAV) can provide complementary information, as they potentially show associations with phenotypic changes (Stein *et al.*, 2017). In consequence, the genotyping data was also used to derive CNV markers using R and the 'gsrc' package (Grandke *et al.*, 2017). The data analysis yielded 3,106 deletions and 4 duplications (Figure S27), additionally used as genetic markers for the subsequent analyses. Notably, a higher number of deletions relative to only a few duplications were detected, which is consistent with previous studies in canola (Cao and Schmidt, 2013; Zou *et al.*, 2018), and a CNV hotspot was observed on chromosome C03 for a set of ten related genotypes from an F_6 generation. Although resequencing of genotypes or genotyping by sequencing (GBS) would have resulted in a substantially higher number of detected polymorphisms that could be used as markers to increase the resolution of the GWAS analyses, the array-based genotyping was the method of choice, in particular with respect to the cost-benefit-ratio, at the time the data were generated. The transcriptome data might have been used to call additional SNPs, but for this purpose the sequencing depth was not high enough.

Plants were grown and phenotyped for a period of 21 days (from 6 to 27 DAS). Each of the 477 genotypes was replicated as three pots with 9 individual plants to provide enough replicates to reach sufficient statistical power and sufficient plant material for subsequent analyses. Previous studies using the IPK phenotyping facilities analysed only individual or a few traits. Schilling *et al.* (2015) for instance analysed projected side-view convex hull area in rice. Muraya *et al.* (2017) focused on estimated biovolume and derived growth rates in maize. They reported high broad-sense heritability values (0.71 to 0.81) for this particular trait, which is comparable to the finding in this study. Neumann *et al.* (2015) analysed multiple traits including biovolume estimates, plant height and compactness, and colour-based traits such as the average hue or the yellow to green ratio in barley plants using a drought stress scenario. In this study, comparably high H^2 values were reported. In a recent study, Pommerrenig *et al.* (2018) analysed 12 biomass-, architectural and colour-related traits obtained from the IAP image analysis pipeline in rapeseed.

In the present study, a total of 1,194 image-derived, mostly colour-related, phenotypic traits related to general plant morphology, plant colouration, static chlorophyll fluorescence and water content and dynamics were obtained and filtered to a core set of 123 traits for further analyses. The core set included in particular growth and biomass related traits such as plant height, area, volume and compactness estimates.

To cope with environmental differences between experiments and GxE interactions, adjusted means (BLUEs) for each phenotypic trait across the different experiments were calculated. The phenotypic traits displayed varying patterns of heritability over the phenotyping time interval. While some traits had an overall high heritability at all days, some displayed low heritability values at earlier or later stages, respectively. These temporarily low heritability values might be attributed to technical and/or environmental bias at the particular time points, or difficulties to correctly estimate certain parameters, for instance the leaf number at a very early stage. Plant material was harvested at two time points: At 14 DAS four plants per pot were harvested for subsequent molecular analyses, and at 28 DAS the remaining five plants were harvested to assess end-point biomass (FW & DW). This early time point at 14 DAS has been chosen as previous studies had shown that data obtained at an early stage harbour important information that could be utilised for the prediction of hybrid performance (Riedelsheimer *et al.*, 2012*a*). Moreover, the pilot GC-MS experiment described in this study yielded evidence that the earlier sampling time point results in more discriminative metabolite profiles than a later time point (28 DAS). Polar metabolites and total RNA were extracted from aliquots of the very same pooled plant material, sampled at 14 DAS, to derive closely related profiles that could be used effectively in the subsequent analyses. Due to a breakdown of the cooling system in the glasshouse during phenotyping experiment 1413RCM, higher temperatures during the first days of growth resulted in a faster growth of the plants and an advanced developmental stage compared to the other experiments (Figure S28). Notably, this bias did not negatively affect the calculation of the phenotypic traits as BLUEs as including experiment 1413RCM resulted in higher heritability values and a higher detection power for QTL compared to the situation when omitting these data (data not shown). This observation can most likely be attributed to the higher number of replicates and an effective correction of the environmental effect component by the mixed linear model.

In contrast, pooling of plants of different developmental stages could lead to biases in the transcriptome and metabolome data. Consequently, sample material from phenotyping experiment 1413RCM was omitted and only material of the three other phenotyping experiments (1419RCM, 1442RCM and 1447RCM) was pooled. This, however, results in a different number of plants per genotype (n= 4 to 12) that are represented in the pooled material.

Polar primary metabolites were analysed by gas chromatography–mass spectrometry (GC-MS). In total 154 metabolites, 64 of known and 90 of unknown chemical structure were quantified. Although these few metabolites cover only a sub-fraction of the whole rapeseed metabolome, the detected metabolites include key substances of the primary metabolism. Among the detected metabolites were substrates of the TCA cycle, multiple amino acids, organic acids, sugars and sugar alcohols. Due to the high number of samples analysed, samples had to be measured in four GC-MS experiments over a period of several days. An explorative data analysis (PCA) of the raw data indicated measurement day effects (Figure S11 a). Consequently, batch-normalisation of raw data for the day of measurement was performed. In addition, normalisation according to sample weight adjusted for the exact amount of plant material used for metabolite extraction. An additional normalisation using the internal standards (L-Valine-d_8 and L-Alanine-2,3,3,3-d_4) was not performed as this increased the standard deviation in the pooled samples, which indicates a reduction in overall quality of the data. Nearly no metabolites in the negative controls and the clustering of the quality control pools in the centre of the PCA plot indicated that there were no substantial contaminations due to the metabolite extraction or issues with the analytical procedure. Notably, a partial separation of genotypes according to the breeding pools in the third PC indicated that also the polar metabolite profiles to some extent reflect the genetic differences between the pools. Four genotypes (‘Achat’, ‘Campino’, ‘MS1’ and ‘MS2’) were included in a higher replication, and their results indicated substantial, but metabolite specific, variation between the replicates (data not shown). Instead of analysing pooled material replicated samples for each genotype could have been analysed. However, this would have greatly increased the number of samples and costs of the experiment. A reduction in sample size by reducing the number of genotypes was also inadvisable as this would have drastically reduced the statistical power of follow-up analyses, in particular for the GWAS analyses.

Two RNA-Seq pilot experiments were performed. The first focused on the implementation of a data normalisation method using the ERCC RNA spike-in mix. Two genotypes ('MS1' and 'Campino') were analysed with different dilutions (1:100; 1:1,000; 1:10,000) of the spike-in mix in the extraction buffer. However, the normalisation procedure was not applied for the main experiment, as the different spike-in RNAs were not correlated across samples (data not shown). The results of the second pilot transcriptome analysis (PCA and HCA, Figure S12) indicated that it was feasible to analyse pooled material from different phenotyping experiments. Moreover, a better separation of genotypes in a PCA analysis using only centred data, in which large expression values contribute more, compared to a PCA analysis with centred and scaled data, where only expression patterns matter, suggests that the highly expressed genes make a large contribution to the observed differences (data not shown). Total RNA was used to generate Lexogen SENSE mRNA-Seq Libraries that were sequenced on an Illumina HiSeq 2500 System. Raw data showed an overall good quality as indicated by the Phred quality scores. However, a fraction of 9 % of the reads could be mapped to ribosomal sequences, indicating that even after poly-A selection / ribosomal RNA depletion during the library preparation a substantial proportion of rRNA remained in the samples. In addition, 8 % of the reads map to organellar (chloroplastic and mitochondrial) sequences. A certain proportion of these reads may map to nuclear sequences that have been transferred from the organellar genomes to the nuclear genome during evolution (Bock and Timmis, 2008; Bock, 2017). Using the *de novo* annotated NRGene reference genome version, transcripts of 54,521 genes (43 % of all 126,667 genes) could be detected as expressed in the shoot tissue. This is comparable to what has been observed in *A. thaliana* (Tian *et al.*, 2019). 19,479 transcripts (15.38 %) were quantified at a median expression level ≥ 5 tpm across all samples and used for subsequent analyses. This is less in percentage, but is approximately the same number of genes that are expressed in Arabidopsis shoot tissue. Transcript data (tpm), which had been concatenated from the different sequencing runs for each genotype, were subjected to an explorative principal component analysis to derive an overview about the data structure of the 477 samples. In the first two principal components, samples clustered in one main group and a smaller subgroup, which was not explainable by any population or design-based factors (Figure S29 a). For this reason, a separate mapping was performed, using all generated data files separately. The separated groups of samples could be traced back to two particular library batches

during library preparation (Figure S29 b). As it was not clear which transcripts and to which degree individual transcripts were affected by this bias, no further normalisation was performed, to avoid further bias in the data. Alternatively, subsequent analyses were performed with both, the data set including and excluding the affected library batches (see below).

4.2. Omics-based hybrid prediction and potential applications in breeding

Rapeseed is one of the leading oilseed-crops worldwide. Hence, breeding companies are interested in constant improvement of the agronomical performance of rapeseed cultivars. Since the late 1980s there has been an increasing proportion of hybrid rapeseed grown and nowadays hybrid plants dominate the rapeseed fields worldwide due to their superior performance (heterosis effect, Gehringer *et al.*, 2007; Liu *et al.*, 2018). For example, according to breeders' information, on more than 80 % of Germany's rapeseed acreage hybrid varieties are grown. Also, in Canada and China, two of the world's most important producers of rapeseed/canola, predominately hybrid varieties are cultivated. However, for the development of new superior hybrids extensive and expensive breeding programmes are necessary. Hundreds of parental inbred lines have to be crossed and the F_1 progeny needs to be evaluated in field studies over multiple locations (and years). This limitation was partially overcome by the introduction of genomic prediction, which allows to estimate the performance of hybrids based on genetic marker information of the parental lines.

One main goal of this work was to evaluate if omics-based data sets gathered from the parental lines can be employed to effectively predict hybrid performance in the field and in the glasshouse. The hypothesis based on previous work in maize (Riedelsheimer *et al.*, 2012*b*) and Arabidopsis (Meyer *et al.*, 2007; Steinfath *et al.*, 2010) was that integration of different sets of predictors (genotype data, transcriptome profiles and metabolite profiles) can improve prediction accuracies. To further this goal, extensive data sets for the 477 parental lines were generated and used individually and in combinations to fit best linear unbiased prediction (gBLUP) and Reproducing Kernel Hilbert Space (RKHS) models. The different -omics data sets comprised molecular markers (n= 13,201; single-copy SNPs), transcripts (n= 19,479; ≥ 5 tpm) and metabolites (n= 154). The polar metabolite profiles constituted by far the smallest set of predictors, but was comparable in size to

previous studies in Arabidopsis by Steinfath *et al.* (2010) with 181 metabolites, maize by Riedelsheimer *et al.* (2012*a*) with 130 leaf metabolites, Feher *et al.* (2014) with 112 root metabolites, or Westhues *et al.* (2017) with 92 leaf and 283 root metabolites, respectively. However, other studies in rice reported the use of 525 (Dan *et al.*, 2016) or even 1,000 leaf and root metabolites (Xu *et al.*, 2016) quantified by GC-MS or LC-MS approaches. Xu *et al.* (2016) also reported poor prediction accuracies when using only a subset of 100 metabolites compared to the full set of 1,000 metabolites. For the seven agronomic traits analysed in the present study, seedling emergence, seed yield, seed oil yield, seed protein content, days to onset of flowering (DTF), seed oil content and seed glucosinolate content (GSL), BLUEs were calculated to adjust for environmental effects and predictions were performed. Although it was possible to effectively predict hybrid performance in oilseed rape, prediction accuracies were strongly dependent on the trait heritability, the genetic complexity of the trait, and the quality of the input phenotype data, e.g. field data assessed at multiple locations and different environments in sufficiently high replication. Phenotypic traits of low heritability like seedling emergence (H^2 = 0.34) or seed yield (H^2 = 0.62) could only be predicted with low to moderate prediction accuracy, while traits with a high heritability like seed oil content (H^2 = 0.90) and total glucosinolate content (H^2 = 0.92) were predictable with high accuracy. The median prediction accuracies ranged between 0.25 (seedling emergence) and 0.72 (GSL). This observation likely reflects the genetic complexity of the traits. Seed yield is known to be a highly polygenic trait heavily influenced by GxE interactions (Marjanović-Jeromela *et al.*, 2011; Escobar *et al.*, 2011). GSL content on the other side, although noticeably influenced by environmental factors (He *et al.*, 2018), is assumed to be controlled by a core set of biosynthesis and degradation genes as well as regulators (Grubb and Abel, 2006; Halkier and Gershenzon, 2006; Ishida *et al.*, 2014). In particular the genome and the transcriptome data showed a high predictive ability. Another observation was that also the number of predictors is likely an important factor that affects prediction accuracy because the metabolite data sets with the smallest number of predictors yielded the lowest prediction accuracies. The genotypic data set and the transcriptome data, both comparable in the number of predictors, yielded mostly comparable results. The transcripts provided approximately 6,000 more predictors than the molecular markers, but did not result in significantly higher prediction accuracies for most traits. These results indicate that with a certain number of predictors the largest part of the genome and

most causative genes and their effects are covered by the data set. This was also observed when including the CNV data into the predictive models did not increase prediction accuracies (data not shown). Hence only SNP data were used.

The principal component analysis of the transcriptome data had indicated a bias due to library preparation. For this reason, predictions were performed with both, the full data set including data of the two biased libraries, and with a reduced data set excluding them from the analysis of transcript data. Full and partial exclusion of data for 89 genotypes and 15 genotypes, respectively, resulted in consistently lower prediction accuracies (Figure S30), most likely due to the resulting reduced statistical power. In case of best linear unbiased prediction models, transcript data as predictors yielded significantly higher prediction accuracies for GSL compared to genetic markers alone. Moreover, combinations of multiple omics-layers incorporating transcript profiles increased prediction accuracies for GSL to the same level. This indicates that biological information, potentially by posttranscriptional regulation like RNA-processing, in addition to regulation of transcription *per se*, is covered by the transcriptome that is not reflected by the genotype (SNP) data. Alternatively, genes affecting the GLS content might be covered insufficiently by the set of genetic markers used. Further stacking of -omics data sets as predictors had no additional positive effect. Hence, the initial hypothesis that the integration of additional transcriptome and metabolite profiles in the prediction models can increase prediction accuracies could only be supported partially. These observations are in accordance with a recent study in maize (Westhues *et al.*, 2017) in which a trait-dependent increase of prediction accuracy was observed and leaf metabolites also yielded overall poor prediction accuracies. Similarly, Riedelsheimer *et al.* (2012*a*) reported prediction accuracies ranging from 0.6 to 0.8 for metabolites, which was on average 6.7 % lower compared to SNP data. Also Zhao *et al.* (2015) reported in a study in wheat that integration of metabolomic data did not result in superior predictions for grain yield compared to genomic prediction. However, they integrated only a very small number of 35 metabolites in their predictive models.

Genomic predictions are already in practical use by breeders since decades and have considerably advanced trait prediction over traditional pedigree-based BLUP, hence representing a well-

established alternative to large numbers of test crosses. Although omics-based (transcriptomic) predictions have shown only an increase of prediction accuracies for the trait GSL in the present rapeseed population with the obtained data sets, they still might be beneficial for other species, traits or other populations. However, it can be concluded that genomic data, which represent the most cost-effective data sets used in this study, are in most cases and for many traits sufficient to effectively predict hybrid performance. A recent study in semi-winter rapeseed demonstrated that already low-density marker sets comprising a few hundred to thousand markers enable high prediction accuracies in breeding populations with long-range LD (Werner *et al.*, 2018).

Beside best linear unbiased prediction (gBLUP) models, reproducing kernel Hilbert space regression (RKHS; Gianola and van Kaam, 2008), which in contrast to gBLUP is at least partially able to exploit additive × additive epistatic effects among markers, was employed for prediction of hybrid performance. In direct comparison to gBLUP models, the usage of RKHS could substantially improve (up to 5.5 % in case of seed coil content) the prediction accuracies for multiple agronomic traits, including seed yield, oil yield, seed protein content and oil content. A major role of epistasis influencing rapeseed yield was revealed (Luo *et al.*, 2017*a*) and it was shown that epistasis, together with heterozygous loci, plays an important role in yield heterosis (Radoev *et al.*, 2008). Epistatic interactions of loci, especially additive x additive epistasis, accounting for a high proportion of variance were also described for a number of yield-related traits in rapeseed, including biomass yield, flowering time, plant height, branch number, harvest index, seed oil content and seed protein content (Zhao *et al.*, 2006; Shi *et al.*, 2011; Li *et al.*, 2012; Würschum *et al.*, 2013). Dominant effects, on the other side, were found to account only for a small proportion of variance (Shi *et al.*, 2011). Notably, clustering of both, QTL and epistatic interactions in several chromosomes was observed. However, both individual QTL and epistatic interactions explained on average < 10 % of PVE, and as only two epistatic interactions of seed yield in different environments were detected, Shi and colleagues suggested that epistatic interactions of yield-correlated traits are extremely sensitive to the environmental variation. For the other four phenotypic traits, seedling emergence, days to flowering and total seed GSL, no statistically significant improvement could be observed in the present study. Interestingly, RKHS was in no case inferior to gBLUP, indicating that for some of the agronomic traits epistatic interactions

contribute to trait manifestation. However, these observed differences were in most cases only subtle compared to differences observed between different traits indicating that the trait heritability, the genetic complexity of the traits, and the quality and size of input phenotype data are more important than the prediction model, as previously reported by Werner *et al.* (2017). Nevertheless, RKHS or other models incorporating non-additive / epistatic effects like EGBLUP (Jiang and Reif, 2015) or Bayesian models (Meuwissen *et al.*, 2001; Habier *et al.*, 2011; Yang and Tempelman, 2012; Werner *et al.*, 2017; Fikere *et al.*, 2018) constitute for some traits a valuable alternative to gBLUP to predict hybrid performance. In addition to the prediction of agronomic traits obtained from the field trials, it was also possible to predict hybrid biomass and growth-related traits for the subset of 120 hybrids (only 13 % of the whole set of hybrids) grown under controlled conditions in the glasshouse. Prediction accuracies of 0.62 and 0.66, which is in the range of the prediction accuracies obtained for the agronomic traits from the field, were achieved using a combination of all available predictors (G+T+M) for FW and DW, respectively. Unfortunately, no direct comparisons between the predictability of field and glasshouse derived data was possible as no common traits were available. Prediction accuracies for biomass accumulation were highest for models including molecular markers (SNPs) and lower for the models using transcriptome profiles, metabolite profiles or a combination of both (Figure S15). These results are in contrast to what has been described in maize (Westhues *et al.*, 2017; Schrag *et al.*, 2018) where a set of 1,323 array-derived gene expression profiles showed excellent performance in the prediction of dry matter yield. Notably, metabolites yielded substantially higher prediction accuracies as predictors for biomass, assessed in the glasshouse, than for the agronomic traits from the field trials (Figure 6; Figure S15). However, further analyses are needed to dissect whether the differently sized and structured data sets and/or different growing conditions affect the prediction accuracies. Furthermore, it should be considered that the hybrids analysed in the glasshouse comprise only 120 selected lines and that both, metabolites and plant biomass are affected strongly by the environment.

Similar to end-point biomass, projected leaf area and estimated biovolume could be predicted at several time points. Prediction accuracies were relatively low for early time points, but increased over time and reached saturation at a value of approx. 0.6 for both traits. This observation may

reflect maternal effects in the earlier phases of plant growth and/or environmental effects during seed establishment, while the effects diminish as soon as plants establish new leaves and the shift from drawing nutrients released from the storage tissue to own photosynthesis occurs, as observed in Arabidopsis seeds and young seedlings (Meyer *et al.*, 2012). Alternatively, a high dispersion of data or an inaccurate registration of the features in the early phase might explain the low prediction accuracy at early time points. A general collinearity of prediction accuracies and trait heritability values seems not to be present, as projected leaf area and estimated biovolume show high heritability at early time points and changes follow a different trend (Figure 5; Data S3).

The parental lines, and a selection of hybrids in a separate experiment, were grown under comparable conditions and phenotyped for the same traits in the same way, which allowed the estimation of heterosis. For a selection of phenotypic traits (fresh weight, dry weight, projected leaf area and estimated biovolume) mid- and best-parent heterosis values were calculated. High heterosis was observed for end-point biomass, ranging from -39.56 to 41.25 % for FW and from -32.66 to 45.06 % for DW, respectively. Similarly, high mid-parent heterosis values were observed for growth-related traits. As previously found in Arabidopsis (Meyer *et al.*, 2004) MPH values for all four traits were mostly positive (Figure 8) and already detectable at early stages of development. Hybrid FW was overall only moderately correlated with the FW of the parental pollinators ($r = 0.48$). Substantial differences in this correlation were observed when grouping the hybrids into ‘good’ ($r = 0.62$) and ‘bad’ ($r = 0.21$) with respect to their seed yield in the field trials, indicating that there is a link between the *per se* performance of parental lines and biomass for at least a subset of the hybrids. Furthermore, these findings and the positive correlation between hybrid seed yield and hybrid biomass ($r = 0.52$) indicate that there is a link between biomass production and seed yield in canola, which is in concordance with previous studies (Basunanda *et al.*, 2010; Zhao *et al.*, 2016). Moreover, this relationship could be confirmed by a significant effect of the male-sterile mother lines. Hybrids with MS1 as female parent, which displayed significantly higher biomass than MS2, produced overall larger plants in comparison to hybrids originating from crosses with MS2. In addition, candidate genes identified by the correlation of the transcript data with leaf area and biovolume MPH displayed substantially higher correlations for hybrids from the MS1 than for the MS2 subset. This might be attributed to beneficial genetic determinants in MS1

compared to MS2, a higher general combining ability (GCA) or cytoplasmic effects. There must be specific interactions (epistasis) between the identified genes and factors in MS1 and/or MS2 that are not or at least rarely present in the parents. Another potential explanation are negative epistatic effects, e.g. the functions of genes that negatively affect growth are suppressed in the hybrids by factors from MS1 or MS2. However, for the three afore-mentioned transcripts that were highest correlated with estimated biovolume MPH, projected leaf area MPH or end-point biomass MPH, no significant correlations ($|r| \geq 0.3$) with growth-related traits in the parent were determined. Causative polymorphisms within the genes or in regulatory elements need to be identified and substantial differences in expression levels between MS1 and MS2, as observed for the transcript of '*BnaCnng47650D*', need to be verified by additional experiments. The two male-sterile lines had originally been selected for the generation of male-sterile testers, as they represent two subgroups of the breeding pool 1. Such heterotic pools with genetically different, less related parental genotypes are an important prerequisite to exploit heterosis in hybrid breeding. However, in oilseed rape such heterotic groups (Melchinger and Gumber, 1998) are not yet well established and genetic diversity is not as broad (Qian *et al.*, 2007; Rincent *et al.*, 2014) as for instance in maize, with the *flint* and *dent* pools and the *stiff-stalk* and *non-stiff-stalk* heterotic groups within the *dent* pool (Younas *et al.*, 2012; Liu *et al.*, 2019). This can be attributed in particular to a less intensive and shorter breeding history of canola compared to maize (Chalhoub *et al.*, 2014; Hu *et al.*, 2019). However, (semi)-resynthesised *B. napus* can be used to increase genetic diversity and to develop new heterotic gene pools that may harbour the potential for the development of new hybrid cultivars (Girke *et al.*, 2012; Zou *et al.*, 2018; Szała *et al.*, 2019). Using the best- and mid-parent heterosis values as quantitative traits for prediction resulted only in low to moderate prediction accuracies, highlighting the complexity of heterosis. This outcome is to some degree expected, as heterosis is the difference between the better parent (BPH) or mid-parent value (MPH) and the hybrids. These values are smaller than the *per se* hybrid performance and much more affected by the dispersion of both (parental and hybrid) values and potential measurement errors. Candidate genes for biomass heterosis could be identified by correlative approaches, although only weak to moderate correlations between individual features were detected. One candidate gene 'BnaCnng47650D' on chromosome C09 was deemed to be particularly interesting. Its Arabidopsis homolog (*AT5G22090*) encodes a FANTASTIC FOUR-like

protein (FAF) which is putatively involved in regulation of the shoot meristem size by modulating the CLAVATA3- WUSCHEL feedback loop (Wahl *et al.*, 2010). Experiments with FAF overexpressing lines indicated that the FAF proteins can repress WUSCHEL, which ultimately leads to an arrest of meristem activity and a marked reduction of meristem size. This finding is consistent with the hypothesis that meristem size may influence final leaf area due to variation in the number of founder cells recruited to form the leaf initial (Gonzalez *et al.*, 2012). It is conceivable that meristem size and genes associated with meristem activity may also play a role in the regulation of vegetative growth and the establishment of biomass heterosis.

4.3. The *Brassica* subgenomes contribute differently to biomass accumulation

As a first step, more than 55 million pair-wise Pearson correlations, were performed between the traits from the different -omics data sets (polar metabolites, n = 154; transcripts, n = 19,479 and phenotypic traits, n = 2,691). These analyses yielded 532 significant pairwise correlations between transcripts and phenotypic traits, 331 correlations between metabolites and transcripts and 22 correlations between metabolites and phenotypic traits ($|r| \geq 0.4$ and a $p\text{-value}_{FDR} \leq 0.05$), respectively. Comparing correlations between the different data sets, the transcripts and metabolites data sets yielded above average and by far the highest correlations. This might be attributed to the close link between the transcriptome, in particular for enzyme encoding genes, and the metabolome that represents an interface for gene-environment interactions and which is commonly seen as the closest link to the observable phenotype differences (Pinu *et al.*, 2019). Many of the highest correlations for metabolites of known chemical structure include genes that are annotated to be involved in related pathways (e.g. enzymes), as shown for example for sucrose and ‘BnaC03g39190D’ annotated with the gene ontology term ‘response to sucrose stimulus’, or β-Alanine and ‘BnaC01g00550D’ encoding a putative aminotransferase. Although only few correlations between polar metabolites and phenotypic traits were detected and correlations between transcripts and phenotypes (restricted to biomass and growth-related traits) were mostly low, the correlation analyses and the network analysis using GENIE3 yielded some interesting candidate genes like ‘BnaA07g21340D’ encoding a putative auxin efflux carrier family protein, ‘Bra000292’ putatively involved in cell wall organization, or ‘BnaC07g05770D’, a peroxidase (PER64) involved in stem lignification. Moreover, a potential link between biomass production and

the structural constituents of ribosomes and differences in translation could be identified that are worth investigating in follow-up studies.

Another main finding was the observation that the biomass accumulation of the canola lines could be related to global differences in their transcriptomes. An explorative principal component indicated a clustering of genotypes in the fourth PC separating them into lines with overall higher and overall lower early biomass production (Figure 9). Notably, this clustering also coincided with the affiliation of genotypes to the breeding pools. Lines from breeding pool 1 displayed an overall lower biomass accumulation than lines originating from the breeding pools 2 and 3.

A gene ontology (GO) term enrichment analysis using the loadings (negative and positive separately) of PC4 was performed. The analysis indicated an overrepresentation of genes involved in biosynthetic processes and gene expression in the negative loadings, and cellular biosynthetic processes, cellular metabolic processes, generation of precursor metabolites and energy and proton transport in the positive loadings. The term 'translation' was found to be enriched in both, negative and positive loadings. This finding is in line with the results from the network analysis. Similar results were also obtained from the pilot experiment (GO term enrichment and MapMan analyses) where two genotypes ('Pol 229', breeding pool 1, low biomass; and 'Pol 419', breeding pool 3, high biomass) were analysed. Differentially expressed genes (DEGs) with a potential function in protein biosynthesis were enriched in 'translation' (GO: 0006412) between the two lines. Previous findings in Arabidopsis showed that growth is associated with ribosome number and polysome loading (Piques *et al.*, 2009; Pal *et al.*, 2013; Czedik-Eysenberg *et al.*, 2016), and that growth rates are negatively correlated with protein turnover (Ishihara *et al.*, 2017). Moreover, enrichment in photosynthesis and light reaction was observed for 'Pol 419', the genotype with higher biomass production. Another observation was that the A and C subgenomes differed in their influence on biomass accumulation. Already in the analysis of differentially expressed genes (DEGs) in the pilot study, a tendency towards a higher contribution of genes from the A subgenome was observed in the genotype 'P419' (higher biomass), while the genotype 'Pol 229' (low biomass) showed a rather balanced ratio of the two subgenomes. This tendency could be confirmed by the results of the PCA analysis, where the top positive loadings (direction of higher biomass) comprised significantly more transcripts

from the A subgenome. The reverse situation was observed for the negative loadings (direction of lower biomass) including significantly more transcripts from the C than from the A subgenome. This does not necessarily imply that in general a higher expression of the A subgenome is associated with higher biomass production, but rather particular classes of genes with a difference in expression. However, the observed differences might also be attributed to the unique nature of the breeding population. This study included, besides many spring-type rapeseed lines, also lines derived from crosses with 'exotic' material. Among others, germplasm of the diploid progenitor genomes was included in the breeding programme by marker-assisted introgression. Moreover, crosses between spring-type and winter-type rapeseed were performed and exotic, genetically distant gene pools were utilised to facilitate the development of new heterotic pools within adapted germplasm. Previous studies reported different contributions from the two subgenomes to the transcriptome of *B. napus* (Higgins *et al.*, 2012; Wu *et al.*, 2018). In a resynthesised *B. napus* approximately one-third of the expressed gene pairs displayed an expression bias with a slight preference towards the A subgenome (Wu *et al.*, 2018). Another study described that partitioning of homeolog gene expression is largely established in *B. napus* with patterns of both genome dominance and genome equivalence, but with the absence of significant bias toward either subgenome (Chalhoub *et al.*, 2014). In this study, the A and C homeologs contributed similarly to gene expression for 17,326 (58.3%) of the gene pairs. In case of the remaining gene pairs (41.7 %), either the A or the C homeologs showed higher expression. However, differences between the subgenome contributions may occur in a developmental or tissue specific manner, as shown by Chalhoub *et al.* (2014), and also for other species (Zhang *et al.*, 2015; Liu *et al.*, 2015*b*). The findings in the present study highlight the important role of translation for biomass accumulation. Moreover, the differences in the contributions of the *Brassica* subgenomes to biomass production might be utilised to selectively optimise line performance and to boost future breeding programmes. To further this goal, additional in-depth analyses of differentially expressed candidate genes and the transcripts comprising the top loadings of PC 4 are necessary.

4.4. Association analyses and regions with effect on different -omics layers

A major challenge of plant biology is to unravel the genetic basis of quantitative (complex) traits and the underlying molecular mechanism. Phenotypic data and molecular data were used in

combination with array-derived genotype data for genome-wide association studies (GWAS), as well as in correlation analyses to link phenotypic and molecular traits, focusing in particular on vegetative growth and biomass accumulation.

Using the different -omics data sets, thousands of individual associations were detected by the genome-wide association studies using Fixed and random model Circulating Probability Unification (Liu *et al.*, 2016) at a p-value $_{FDR} \leq 0.05$. Compared to generalised linear model (GLM) and mixed linear model (MLM) approaches, this method features a low rate of false positive associations and a fast runtime, which, in combination with parallelization, allows the analysis of multiple phenotypic traits in a reasonable period of time. Although relatively new, the method was already applied successfully in several different studies, for example to identify genetic loci for drought tolerance in maize (Li *et al.*, 2016*c*), plant height in maize (Hu *et al.*, 2017), salt tolerance in cowpea (Ravelombola *et al.*, 2017), seed traits in soybean (Wang *et al.*, 2018*a*), or tolerance to pre-harvest sprouting and low falling numbers in wheat (Martinez *et al.*, 2018).

In summary, 15,789 QTL for 2,691 phenotypic traits (individual traits at 21 days, growth rates and end-point biomass), 86,013 eQTL for 19,479 transcripts (≥ 5 tpm) and 544 mQTL for 154 polar metabolites were identified. Most QTL in relation to the number of traits were detected for the phenotypic data set, with at least one QTL for 80 % of all traits, followed by the transcript data (75.4 %) and the metabolite data set (66 %). However, the average explained variance for eQTL (3.1 %) and mQTL (2.2 %) was higher than for the phenotypic QTL (1.8 %). Notably, 7.4 % of eQTL explained more than 10 % of PVE compared to 2 % for mQTL and only 0.5 % for the phenotypic QTL, indicating that transcript variation is at least to some degree controlled by a combination of larger effect and small effect loci. Metabolite variation and in particular phenotypic trait variation, on the other hand seem in most cases to be controlled by many small effect loci. Most associations were observed for SNP markers, but a substantial proportion of deletions were found to be associated with phenotypic trait variation, for example the deletions on chromosome C03. This is consistent with previous studies that showed that in particular segmental deletions caused by widespread homoeologous exchanges (Samans *et al.*, 2017; Hurgobin *et al.*, 2018) were associated with trait variation (Qian *et al.*, 2016; Schiessl *et al.*, 2017;

Stein *et al.*, 2017; Hatzig *et al.*, 2018). However, the number of associations with deletion markers is surprisingly high. In some cases, traits display dozens to more than a hundred associated CNV markers at a particular chromosome like the trait 'yellow to green ratio' at 23 DAS with 410 associated deletion markers on chromosome C03. These may indicate one or multiple larger deletions (see Figure S27), as groups of CNV markers are highly linked and map to large regions. Hence, in such cases the focus should first be placed on the highest significantly associated deletion markers in the particular regions, second taking the entire potentially deleted regions into account for a candidate gene search. As groups of deletion markers associated with particular traits were also found to be interspersed with SNP markers not associated with this trait, it will be important to determine whether other chromosomal rearrangements rather than (potentially heterozygous) deletions may be causal for these observations. In this work CNV markers were partitioned into deletion and duplication markers. The impact of this partitioning on the GWAS results needs further in-depth analysis and verification in order to assess its effect. Deletion markers yielded on average slightly more PVE than SNP markers (3.24 % vs. 2.83 %), as expected as they often cause more severe changes in the DNA sequence than point mutations. The eQTL were classified in either cis- or trans-eQTL, whereby nearly eight times more trans-eQTL than cis-eQTL were identified, which is substantially higher that in a previous QTL mapping study by Li *et al.* (2018*a*), but comparable to what has been reported in soybean (Bolon *et al.*, 2014). The cis-eQTL contributed a substantially higher percentage of explained phenotypic variance. However, the numbers of cis- and trans-eQTL and hence the attributed variances might be slightly biased due to the applied definition by size (500 kb or 1 Mb), potential shifts or errors in the genome annotation, or mix up homologous and/or homeologous genes. Moreover, it has been shown that QTL mapping in small to moderate populations were underpowered to detect QTL with small effects, resulting in a substantial overestimation of the effects of large QTL (the so-called 'Beavis effect'; Xu, 2003). With increasing size and power of QTL mapping experiments, large-effect QTL were often shown to fractionate into many, closely linked QTL with smaller effects (Ingvarsson and Street, 2011). Often, alleles of adjacent small-effect QTL have opposite effects as shown in a study on growth rate in *A. thaliana* (Kroymann and Mitchell-Olds, 2005).

Previous studies described an unequal distribution of eQTL across the genome (Kliebenstein, 2009; Fu *et al.*, 2009; Li *et al.*, 2013*a*, 2018*b*; van Muijen *et al.*, 2016) and mQTL hotspots (Schauer *et al.*, 2006; Joosen *et al.*, 2013; Alseekh *et al.*, 2015; Wen *et al.*, 2015; Knoch *et al.*, 2017) in various species. As previously described, QTL were found to be unevenly distributed across chromosomes and chromosomal positions. Regions depleted in QTL as well as regions with a substantial overrepresentation of QTL (hotspots) were detected for all data sets. In the A subgenome, eQTL seem to be more evenly distributed than in the C subgenome where they more often tend to cluster in hotspots. Similar patterns were observed for the mQTL and phenotypic QTL, but hotspots did not or only partially overlap between the sets. Fewer and less pronounced mQTL hotspots were found compared to the phenotypic or expression QTL, which can most likely be attributed to the overall much lower number of analysed features. Although different hotspots for the different sets were observed, the general QTL density seems to follow the gene density, as illustrated by Chalhoub *et al.* (2014). It has been hypothesised that the observation of clustered QTL could be explained by master regulating factors, like transcription factors, influencing the expression of various downstream targets and hence ultimately affecting trait variation (Lisec *et al.*, 2008).

To identify potential links through the different omics layers a co-localisation analysis for traits of the different omics-layers was performed. As many more QTL were detected than markers were used for the GWAS analysis, in particular many eQTL, stochastic co-localisations between phenotypic QTL, eQTL and mQTL should be expected. For this reason, a permutation approach was applied to estimate the number of co-localisations that might be expected by random chance. The observed number of 32 markers with co-localisation between all three omics-layers is much higher than the estimated threshold of 13 markers obtained from the permutation analysis. However, this number might still be biased as an equal distribution of QTL over the genome was assumed. Several interesting candidate links between the omics-layers were observed. For example, the deletion marker 'Bn-scaff_28509_1-p33495_del' coincided within a hotspot of 153 eQTL on chromosome C03 (position 26.6 Mb). Notably, 'Bn-scaff_28509_1-p33495_del' is likely part of a larger deletion, which was only detected in a narrow, related subset of genotypes (see Figure S27). This potential large deletion spans a region of approx. 12.8 Mb (from

'Bn-scaff_27881_1-p15080_del' to 'Bn-scaff_18917_1-p472625_del') and likely affects more than 1,800 putative genes. The eQTL hotspot further co-localised with an mQTL for glucose and six QTL for colour-related traits. Co-localisations were used in combination with the correlations to prioritise candidate genes. In this way, another region on chromosome C01 (position 492 kb) could be linked to changes in the abundance of β-Alanine, a stress-related compound (Parthasarathy *et al.*, 2019), the expression of the transcript of 'BnaC01g00550D', encoding an alanine-glyoxylate aminotransferase and changes in plant colour intensity. Moreover, the deletion marker 'Bn-scaff_19026_1-p369147_del' on chromosome C03 (position 36.3 Mb), which is likely part of the same large deletion mentioned above, could be linked to sugar metabolism, a transcript ('BnaC03g39190D') annotated with response to sugar stimuli and changes in a colour-related trait. In cases of genes encoding enzymes or transporters, which can directly affect the abundance of metabolites, a correlation between the transcript and the metabolite, as well as co-localisation of mQTL and eQTL is likely (Brotman *et al.*, 2011). This notion is confirmed for example by the co-localisation of an mQTL for amino acids and a cis-eQTL coinciding with a gene encoding a putative amino acid transporter protein. Moreover, an mQTL for valine and an eQTL corresponding to a potential branched-chain-amino-acid aminotransferase were co-localised. However, for many other co-localisations it was not possible to find a direct link between metabolites and transcripts using their gene annotations. The interpretation of the results may be complicated by an incomplete genome annotation and homology-based functional characterisation of the NRGene Darmor-*bzh* reference genome assembly. In addition, the incomplete coverage of the transcriptome with only 15.38 % of all *B. napus* genes quantified at ≥ 5 tpm in the sampled tissue / at the analysed time point may hamper the interpretation of the results as relevant genes, for instance low abundant transcription factors, might have not been covered. A deeper RNA-sequencing analysis might have been beneficial, but would have increased the sequencing costs disproportionately.

In concordance with previous studies (Wu *et al.*, 2016*a*; Zhou *et al.*, 2017*b*), a faster LD-decay was detected in the A subgenome (half decay of approx. 400 kb) compared to the C subgenome (half decay of approx. 3.9 Mb). The large LD blocks and a long-range LD decay especially in the C subgenome increase the number of potential candidate genes. The multiple larger genomic

regions of high LD ($R^2 > 0.6$), especially in the C-subgenome, possibly reflect regions that were preferentially selected or recently introgressed by breeders, or indicate the presence of larger structural variations within the population compared to the reference genome. Another limiting factor is that the metabolome and transcriptome was analysed only at a single time point. Only a fraction of genes is expressed at certain time point or in a certain tissue (Chan *et al.*, 2016; Wan *et al.*, 2017). Hence, it might be difficult to elucidate the complex and time-dependent interactions underlying traits like plant growth (Ni *et al.*, 2009; Farré, 2012). In particular growth and biomass accumulation are highly complex polygenetic traits (Gonzalez *et al.*, 2012; González and Inzé, 2015) that are affected by various different pathways and environmental influences (Lima *et al.*, 2017; Elferjani and Soolanayakanahally, 2018). As various different gene functions might have an impact on growth, it is difficult to prioritise genes by their annotation. Nevertheless, the analyses including the network analysis using GENIE3 pointed to candidate genes potentially associated with biomass production like 'BnaA07g23560D' and 'BnaCnng17010D', both putatively encoding for serine hydroxymethyl-transferases which had been associated with photorespiration (Jamai *et al.*, 2009). Notably, there are associations of candidates genes related to biomass accumulation identified in the present study and candidate genes identified in the study on growth dynamics in maize by Muraya *et al.* (2017). In the present study, 'BnaC02g08760D', the *B. napus* homolog of the Arabidopsis *AT5G19510*, described as encoding a 'translation elongation factor EF1B/ribosomal protein S6 family protein' was identified as negatively correlated with biomass and leaf area. Muraya *et al.* (2017) identified a functionally closely related protein (both involved in translational elongation), '*GRMZM2G029559*' the maize homolog of *AT1G09640*, described as 'translation elongation factor EF1B gamma chain' as candidate affecting biomass accumulation. Moreover, both studies identified purple acid phosphatases among candidate genes, *GRMZM2G138698* the maize homolog of the Arabidopsis PAP27 (*AT5G50400*) by Muraya *et al.* (2017) and 'maker-scaffold124-snap-gene-8.174' with homology to the Arabidopsis PAP18. Despite the obstacles mentioned above, the data sets generated in this work constitute a broad and extensive basis for follow up analyses. The candidates inferred in this study provide a basis for further in-depth analysis to validate associations and to carefully validate observed links, e.g. by virus-induced gene silencing (VIGS) approaches (Álvarez-Venegas *et al.*, 2014; Yu *et al.*, 2018*b*), or by characterisation of

mutants from TILLING populations (Wang *et al.*, 2008; Gilchrist *et al.*, 2013), or by CRISPR cas9 mutagenesis (Zhang *et al.*, 2019*b*; Zhai *et al.*, 2019; Zheng *et al.*, 2019). Subsequent detailed studies should focus on a time-series analysis with sampling and analysis of plant material at multiple time points to address this issue. This might be in particular interesting with respect to the analysis of biomass heterosis as it has been hypothesised that different alleles of genes might be beneficial only at particular developmental stages and in combination may contribute over time to the superior performance of hybrids / heterosis (principle of heterochrony; Cong *et al.*, 2002; Guo *et al.*, 2003; Sang *et al.*, 2019). The unique combination of data sets generated in this study and especially the detected eQTL provide an excellent entry-point for a follow up network analysis. In a first step the focus should be placed on the relation between transcripts and metabolites. Later information of biomass production could be integrated as an additional layer to the network.

4.5. Temporal dynamics of QTL action on early growth

The following chapter includes revised parts of the research article 'Strong temporal dynamics of QTL action on plant growth progression revealed through high-throughput phenotyping in canola' published in Plant Biotechnology Journal, May 24th 2019.

A main part of the study was the daily high-throughput phenotyping of the 477 diverse spring-type canola genotypes, originating from a breeding programme, using the IPK phenotyping system for large plants. Phenotypic traits related to general plant morphology, plant colouration, static chlorophyll fluorescence and water content and dynamics were obtained for a 21-day period from 6 to 27 DAS, covering an early vegetative stage of rapeseed development. Four phenotypic traits that reflect plant growth (projected leaf area, estimated biovolume, early plant height, and colour uniformity), with high heritability, as well as end-point biomass accumulation (FW & DW), were selected for a detailed analysis. The temporally resolved data for 21 days were used to calculate BLUEs that were subjected to genome-wide association studies to address the following questions: (i) Which key genomic regions are associated with growth-related traits and relative growth rates in the early phase of vegetative plant development? (ii) To what extent do identified regions

contribute to trait variance? (iii) Can dynamic, stage-specific contributions of loci for early growth be resolved by a time course analysis? (iv) Is it possible to nominate candidate genes that might be causal for the observed marker-trait associations?

For all analysed time points, a total of 809 associations for end-point biomass and the four selected phenotypic traits were detected at a *p*-value $_{(FDR)} \leq 0.1$. Most of the detected MTAs explained only a small percentage of phenotypic variance (< 5 PVE%, Figure S23). In total, only 40 (3.8 %) marker-trait associations with larger effects (> 5 PVE%) were detected, like marker 'Bn-A04-p4409752' explaining up to 8.64 % PVE of biomass (fresh weight). These findings are consistent with the hypothesis that biomass accumulation and growth-related traits are mostly governed by small effect loci and their interactions (Muraya *et al.*, 2017). No substantial differences in the number of associations were detected between the A subgenome and the C subgenome. The 22 significant MTAs identified for fresh weight and dry weight were compared to a list of 771 previously described QTL obtained from 13 publications analysing 45 growth, seed yield and quality-related traits (Li *et al.*, 2011, 2014*a*; Luo *et al.*, 2015; Tang *et al.*, 2015; Körber *et al.*, 2015, 2016; Liu *et al.*, 2016*b*; Lu *et al.*, 2016; Sun *et al.*, 2016; Wang *et al.*, 2016; Li *et al.*, 2016*a*, 2017; Zheng *et al.*, 2017). The marker 'Bn-scaff_18702_1-p589589' has been shown to be associated with plant height (Tang *et al.*, 2015). Seven other MTAs were in proximity (± 500 kb, based on NRGene marker positions) to previously described QTL: 'Bn-A04-p4409752' close to a QTL for stem dry weight (Lu *et al.*, 2016); 'Bn-A10-p11817272' close to a QTL for plant height (Sun *et al.*, 2016) and QTL for branch angle (Li *et al.*, 2017); 'Bn-A07-p9632473' co-localised with a QTL for flowering time (Wang *et al.*, 2016); "Bn-A08-p16771030" close to QTL for biomass yield and stem dry weight (Lu *et al.*, 2016) and a QTL for branching angle (Li *et al.*, 2017); "Bn-A10-p10672359" in proximity to a QTL for biomass yield, a QTL for stem dry weight (Lu *et al.*, 2016) and a QTL for plant height (Sun *et al.*, 2016); 'Bn-A10-p13343454' close to another QTL for branching angle (Li *et al.*, 2017); 'Bn-scaff_21312_1-p895326' close to QTL for stem dry weight, a QTL for biomass yield (Lu *et al.*, 2016) and a QTL for plant height (Li *et al.*, 2016*a*). Notably, despite the high phenotypic correlation between biomass values (r=0.969), only three shared MTAs for FW (13 MTAs) and DW (9 MTAs) were identified (Table 5). This highlights the fact that fresh weight is not identical to dry weight, as a similar FW may be reached by varying contributions of its

individual components, such as the content of cellulose, hemicellulose or lignin, and the water content. Moreover, this finding also suggests that for the other phenotypic traits some associated loci might be missed by filtering highly correlated traits in the previous step of stepwise variable selection using variance inflation factors.

The time-resolved design of the phenotyping experiments enabled the analysis of the temporal dynamics of detected growth QTL. To further this goal, a two-filter procedure was applied with a first moderate threshold (p-value $_{(FDR)} \leq 0.1$) filter for associations. Subsequently, a second filter was applied to enrich for robust QTL. Only markers that displayed sequentially significant association with measured phenotypic traits for three consecutive time points were considered and evaluated in more detail. In summary, 14, nine, four and three such persistent associations were detected for projected leaf area, estimated biovolume, early plant height and colour uniformity, respectively. Previous studies in Arabidopsis (Moore *et al.*, 2013; Bac-Molenaar *et al.*, 2015), rice (Al-Tamimi *et al.*, 2016), *Setaria* (Feldman *et al.*, 2017) and maize (Muraya *et al.*, 2017) addressed the dynamics of growth and time-dependent QTL mapping, but they did not provide such a high temporal resolution at a daily basis. A recent study in barley applied daily plant imaging for the time span of 32 to 59 days after planting, but focused on derived growth rates and end-point traits including fresh and dry weight, tiller number and plant height (Pham *et al.*, 2019). In accordance with the findings of developmental geneticists that genes are expressed selectively / differentially at different developmental stages and in different tissues (Nakabayashi *et al.*, 2005; Schmid *et al.*, 2005; Kudapa *et al.*, 2018; Lee *et al.*, 2019), the data generated in this study indicated that plant growth is the cumulative result of the interaction of various different genes and that the contributing sets of factors change during plant development. In contrast to a previous genome-wide association study in Arabidopsis, which revealed time-specific and general / constitutive QTL affecting growth dynamics (Bac-Molenaar *et al.*, 2015), in the present study in canola no constitutive QTL, but only time-specific associations were detected. The longest interval of consecutive significant effects was found in the present study for marker 'Bn-scaff_16361_1-p2350469' on chromosome C08 associated with projected leaf area between 16 to 27 DAS. Notably, one marker, 'Bn-scaff_16804_1-p178142' on chromosome C02, was found to be associated with both, the late projected leaf area at 25 to 27 DAS, and with end-point dry weight assessed at 28 DAS. The time-specific patterns observed suggest that QTL are under the

control of dynamic genetic regulation. The beneficial effect of an allele of an early QTL might lose its benefit with progression of development and another allele of a later QTL might take up the beneficial effect. The dynamic nature of the identified QTL implies that many associations with effects at earlier time points would likely not have been identified if biomass-associated traits had only been evaluated as integrated effects at the end of the experiment. Consequently, underlying genes might not be uncovered or the genetic value of the loci might be underestimated.

To further address the dynamic QTL effects, relative growth rates (RGRs) for projected leaf area, estimated biovolume and early plant height, as well as absolute change rates (ACRs) for colour uniformity were calculated integrating the effects over longer periods, and also subjected to a GWAS analysis. Different interval sizes, day-to-day intervals and intervals spanning two and three days were tested. Larger intervals were assumed to accumulate the effects over a longer period of time and hence are expected to show less scatter / less variation than shorter intervals. This assumption was confirmed by the observation that overall fewer significant MTAs and also fewer MTAs with consecutively significant effects were detected for shorter intervals (data not shown). Hence, the focus was placed on the three-day intervals. In total, 268 significant associations for relative growth rates and absolute change rates calculated for the 15 overlapping three-day intervals were detected. To enrich particularly interesting MTAs, the growth rate associations were further filtered to be significant for at least two consecutive significant intervals. Although no consecutive significant associations for colour uniformity ACRs were observed, such MTAs were detected for leaf area, plant height and biovolume RGRs (Figure 16). The detection of such dynamic growth rate QTL can also be attributed to the statistical power achieved in the present study through the large dimension and the setup of the experiments assessing each of the 477 analysed genotypes replicated in three large scale glasshouse experiments performed under controlled environmental conditions with nine (6 to 13 DAS) or five (15 to 27 DAS) individuals per replicate, respectively. The substantially lower number of RGR QTL active at two consecutive intervals vs. the total number of RGR MTAs may indicate that the majority of effects are restricted to very narrow time windows. Since RGR MTAs address the acute action of the genetic loci at the assessed time point, while the MTAs of *per se* traits reflect the cumulative effects of the loci that happened during the entire growth period up to the time point

of measurement, it is not surprising that the number of detected RGR MTAs is generally lower than the number of MTAs of *per se* traits and that there is only minor overlap between the MTAs of the two types of traits. A recent study analysed the genetic architecture of biomass accumulation in spring barley (Neumann *et al.*, 2017) by image analysis, and described temporal patterns similar to the findings for *per se* trait and growth rate MTAs in the present study. Muraya *et al.* (2017) detected MTA effects on RGR for a subset of the strongest *per se* trait MTAs and described the reversal of allelic effects over time for markers associated with relative growth rates. Similar observations were made in the present study on canola. In addition, reversal of allele effects over time was observed for a substantial fraction of MTAs (Figure S24 and S25).

As most dynamic growth / biomass-associated QTL actions tended to persist for periods of only a few days during early growth, it might be hypothesised that these QTL are associated with the initiation or development of new leaves. A particularly remarkable pattern with four dynamic QTL for early, intermediate and late time points was observed for the RGR of estimated biovolume (Figure 16 b). A manual analysis of leaf number for a subset of 30 lines at the different time points indicated that new leaves emerge on average in three to four-day intervals, coinciding with the observed pattern of dynamic growth QTL. However, to verify this observation, more in-depth analyses will be necessary that will require robust high-throughput quantification of leaf number in the acquired images. While promising advances in image analyses have been achieved in this direction, for instance by 'CVPPP challenges' (Pape and Klukas, 2015; Scharr *et al.*, 2016), further developments will be necessary to use automated image analyses towards this goal.
If the hypothesis of different QTL being involved in initiation and development of successive leaves can be supported, it indicates the exciting possibility that formation of each leaf (or more generally every organ) may be controlled by a distinct genetic programme triggered through certain leaf-specific loci. The time resolved genome-wide association analyses revealed temporal dynamics of QTL for early growth–related traits and growth rates. These findings highlight the need for stage-specific investigations in future studies to identify genes operating at different developmental phases. Muraya *et al.* (2017) proposed that genes corresponding to dynamic QTL are either selectively expressed at different growth stages or their functions are required or growth-limiting only in certain developmental phases.

For a selection of five candidate regions sharing dynamic associations for multiple growth traits, candidate genes were evaluated by an LD-based approach and prioritised based on their annotation. For example, nine of these genes are related to meristem development, vegetative phase change and cell growth in Arabidopsis, including S*epallata1* (*AT5G15800*; Pelaz *et al.*, 2000; Li *et al.*, 2016*b*); *BnaA10g18480D*), *Longifolia1* (*AT5G15580*; Lee *et al.*, 2006; *BnaA10g18650D*), *Squamosa promoter binding 3* (*AT2G33810*; Wu and Poethig, 2006; *BnaC03g18800D*) and S*hatterproof1* (*AT3G58780*; Favaro *et al.*, 2003; Battaglia *et al.*, 2006; *BnaC08g29530D*). Several other genes are putatively involved in flowering time or cell wall biogenesis and modification, or were annotated as transcription factors. A comparable previous genome-wide mapping study by Bac-Molenaar *et al.* (2015) in Arabidopsis identified candidate genes which were annotated to be involved in the determination of cell number and size, seed germination, embryo development, developmental phase transition, or senescence. Due to the large LD-blocks it is of pivotal importance that candidate genes identified in this study will be further analysed and validated in follow-up studies. For example, qRT-PCR approaches involving temporally and spatially resolved assessment of gene activity might be performed. Despite this limitation, the described dynamic QTL represent a well exploitable resource to deepen the knowledge of early plant growth and biomass accumulation.

A previous study in a doubled-haploid (DH) population derived from a cross between 'KenC-8' (spring-type *B. napus*), and 'N53-2' (winter-type *B. napus*) described a high positive correlation ($r = 0.83$) between seed yield and biomass yield (Zhao *et al.*, 2016). In addition, another study identified a weak, but significant correlation between heterosis for shoot weight and heterosis for seed yield in two large DH mapping populations and two corresponding sets of backcrossed test hybrids (Basunanda *et al.*, 2010). Hence, the introgression of beneficial alleles underlying dynamic growth-QTL, absent or underrepresented in conventional breeding pools, on the one side, as well as selection and stacking of beneficial alleles on the other side might help to enhance genetic gain for complex traits towards further improvement of yield performance in canola breeding. Moreover, it broadens the selection basis by introducing the factor temporal dynamics, facilitating marker-assisted selection to breed high-vigour cultivars.

5. Summary

Hybrid plants are gaining more and more importance in plant breeding due to heterosis, the superior performance of progeny compared to their inbred parents. Since the development of new superior hybrids requires work and cost intensive breeding programmes, the prediction of hybrid performance is of utmost interest to breeders. One objective of this work was to test the effectiveness of prediction models for hybrid performance in spring-type oilseed rape (canola) by using different omics profiles, individually or in combination. The basis of this study was an F_1 hybrid population consisting of 950 genotypes that had been evaluated for seed yield and six other breeding-relevant traits at multiple locations across Europe in commercial field trials. A subset of the hybrids was also evaluated in the glasshouse regarding early biomass production. Predictors for hybrid performance were generated from the 477 parental lines, including 13,201 single nucleotide polymorphisms (SNPs), 154 polar primary metabolites and 19,479 transcripts assessed at 14 days after sowing. Both, SNP marker and transcriptome data were similarly effective in predicting hybrid performance using (genomic) best linear unbiased prediction (gBLUP) models for the agronomic traits analysed. Exploiting transcriptome data alone or in combination resulted in significantly higher prediction accuracies for only one out of seven traits, seed glucosinolate content, compared to SNP markers only. Reproducing Kernel Hilbert Space regression models significantly outperformed gBLUP for four out of the seven agronomic traits, seed yield, seed oil yield, seed protein content and seed oil content, probably by capturing epistatic genetic effects. Prediction accuracies strongly depended on the trait, its underlying genetic architecture and heritability, no universally best prediction model was identified.

A major challenge of plant biology is to unravel the genetic basis of complex traits. In the second part of this study, the generated omics-profiles for the parental lines of the hybrid population, two male-sterile lines and 475 pollinators, were analysed in conjunction with high-throughput non-invasive phenotyping data that had been established for plants cultivated under controlled conditions in the glasshouse. The plants were imaged on a daily basis during the first four weeks of early vegetative development, and 123 traits, together with end-point biomass, were selected for a detailed analysis. Notably, an unequal contribution of transcripts from the two *Brassica napus* subgenomes to biomass formation was uncovered and gene ontology term enrichment analysis provided hints that differences in ‘translation’ may play a prominent role in biomass production.

Array-derived genome-wide SNP and copy-number variation marker data provided the framework for multi-omics genome-wide association analyses. A total of 31,264 quantitative trait loci (QTL), each explaining more than 2 % of phenotypic variance, were detected at a $p\text{-value}_{(FDR)} \leq 0.05$, including 206 metabolite-QTL (mQTL), 26,391 expression-QTL (cis- and trans-eQTL), and 4,667 phenotypic QTL. An uneven distribution of these QTL across the genome with pronounced hotspots was observed. Moreover, co-localisations of QTL across the different data sets, more than expected by random chance, were detected. As these co-localisations may correspond to key loci with effects on multiple omics-layers, candidate genes in these regions were prioritised by correlating transcript and metabolite features. Capitalising on the daily temporal resolution of phenotyping for four growth-related traits and derived growth rates, an in-depth and time-resolved analysis was performed. In total, 96 robust main effect marker-trait associations, significant for at least two consecutive days, were detected. Based on a linkage disequilibrium-based approach, candidate genes were identified at five selected loci with dynamic behaviour and effect on multiple traits. The candidates were involved in meristem development, cell wall modification and transcriptional regulation. The results of the temporal analysis highlight that early plant growth is a highly complex trait governed by several medium and many small effect loci, most of which act only during short developmental phases.

6. Zusammenfassung

Hybridpflanzen gewinnen in der Pflanzenzüchtung aufgrund des Heterosis-Effekts, der überlegenen Leistung der Nachkommen im Vergleich zu ihren Inzucht-Eltern, immer mehr an Bedeutung. Da die Entwicklung neuer überlegener Hybride jedoch arbeits- und kostenintensive Zuchtprogramme erfordert, ist die Vorhersage der Hybridleistung für Züchter von höchstem Interesse. Ein Ziel dieser Arbeit war es, die Effektivität von Vorhersagemodellen für Hybridleistung in Sommerraps (Canola) zu testen, indem verschiedene Omics-Profile, einzeln oder in Kombination, zur Vorhersage genutzt wurden. Grundlage dieser Studie war eine F_1-Hybrid-population bestehend aus 950 Genotypen, die in kommerziellen Feldversuchen an mehreren Standorten in ganz Europa auf Saatgutertrag und sechs weitere züchterisch relevante Merkmale untersucht wurden. Eine Auswahl der Hybriden wurde zudem auch im Gewächshaus auf Biomasseertrag zu einem frühen Zeitpunkt untersucht. Prädiktoren für die Hybridleistung, darunter 13.201 genetische Marker (SNPs), 154 polare Primärmetabolite und 19.479 Transkripte, wurden 14 Tage nach der Aussaat von den 477 Elternlinien erhoben. Sowohl SNP-Marker als auch Transkriptomdaten waren ähnlich effektiv bei der Vorhersage der Hybridleistung für die analysierten agronomischen Merkmale unter Verwendung von (*genomic) best linear unbiased prediction* (gBLUP) Modellen. Die Nutzung von Transkriptomdaten, allein oder in Kombination mit anderen Datensätzen, führte für nur eines von sieben Merkmalen, den Gehalt an Samen-Glukosinolaten, zu einer signifikant höheren Vorhersagegenauigkeit im Vergleich zu reinen SNP-Markern. Durch RKHS-Modelle (*Reproducing Kernel Hilbert Space regression*) konnte die Vorhersagekraft der gBLUP-Modelle für vier der sieben agronomischen Merkmale, Samenertrag, Samenausbeute, Samenölertrag, Samenproteingehalt und Samenölgehalt übertroffen werden. Dies ist wahrscheinlich auf die Erfassung epistatischer genetischer Effekte zurückzuführen. Die Vorhersagegenauigkeiten hängen stark von dem jeweiligen Merkmal, der zugrunde liegenden genetischen Architektur und der Heritabilität ab. Es wurde kein universell bestes Vorhersage-modell identifiziert.

Eine große Herausforderung der Pflanzenbiologie besteht darin, die genetische Grundlage komplexer Merkmale zu entschlüsseln. Im zweiten Teil dieser Arbeit wurden die generierten Omics-Profile für die Elternlinien der Hybridpopulation, zwei männlich-sterilen Mutterlinien und 475 Bestäubern, in Verbindung mit nicht-invasiven Hochdurchsatz-Phänotypisierungsdaten

analysiert, die für Pflanzen erstellt wurden, welche unter kontrollierten Bedingungen im Gewächshaus angezogen wurden. Die Pflanzen wurden in den ersten vier Wochen der frühen vegetativen Entwicklung täglich aufgenommen und 123 Merkmale sowie die Endpunktbiomasse wurden für eine detaillierte Analyse ausgewählt. Insbesondere wurde ein ungleicher Beitrag der Transkripte aus den beiden *Brassica napus* Subgenomen zur Biomassebildung aufgedeckt. Eine *Genontologie* (GO) Anreicherungsanalyse lieferte Hinweise darauf, dass Unterschiede in der Translation eine wichtige Rolle bei der Biomasseproduktion spielen könnten. Array-basierte genomweite SNP- und Kopienzahlvariationsmarkerdaten bildeten den Rahmen für multi-omics genomweite Assoziationsanalysen. Insgesamt konnten 31.264 *Quantitative Trait Loci* (QTL), die jeweils mehr als 2 % der phänotypischen Varianz erklären, mit einem *p*-Wert $_{(FDR)} \leq 0{,}05$ nachgewiesen werden, darunter 206 Metaboliten-QTL (mQTL), 26,391 Expression-QTL (cis- und trans-eQTL) und 4.667 phänotypische QTL. Eine ungleichmäßige Verteilung dieser QTL über das Genom mit ausgeprägten Hotspots wurde beobachtet. Darüber hinaus wurden Kolokalisierungen von QTL (mehr als zufällig erwartet) über die verschiedenen Datensätze hinweg festgestellt. Da diese Kolokalisationen Schlüsselpunkten mit Auswirkungen auf mehrere Omics-Schichten entsprechen können, wurden die Kandidatengene in diesen Regionen durch Korrelation von Transkript- und Metabolitenmerkmalen priorisiert. Ausgehend von der tageweisen zeitlichen Auflösung der Phänotypisierung wurde für vier wachstumsrelevante Merkmale und abgeleitete Wachstumsraten eine detaillierte und zeitaufgelöste Analyse durchgeführt. Insgesamt wurden 96 robuste Haupteffekt-Marker-Merkmals-Assoziationen, die für mindestens zwei aufeinander folgende Tage signifikant waren, identifiziert. Durch einen Kopplungsungleichgewicht-basierenden Ansatz wurden Kandidatengene an fünf ausgewählten Genloci mit dynamischem Verhalten und Wirkung auf mehrere Merkmale identifiziert. Die Kandidatengene waren an der Meristementwicklung, der Zellwandmodifikation und der transkriptionellen Regulation beteiligt. Die Ergebnisse der zeitaufgelösten Analyse zeigen, dass das frühe Pflanzenwachstum ein hochkomplexes Merkmal ist, von mehreren Genloci mit mittleren und vielen kleinen genetischen Effekten gesteuert wird, von denen die meisten nur in kurzen Entwicklungsphasen wirken.

7. References

Akdemir D, Isidro-Sánchez J. 2019. Design of training populations for selective phenotyping in genomic prediction. Scientific Reports **9**, 1446.

Albrecht T, Auinger H-J, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho H-P, Schön C-C. 2014. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theoretical and Applied Genetics. **127**, 1375–1386.

Ali M, Copeland LO, Elias SG, Kelly JD. 1995. Relationship between genetic distance and heterosis for yield and morphological traits in winter canola (*Brassica napus* L.). Theoretical and Applied Genetics **91**, 118–121.

Allwright MR, Taylor G. 2016. Molecular Breeding for Improved Second Generation Bioenergy Crops. Trends in Plant Science **21**, 43–54.

Alseekh S, Tohge T, Wendenberg R, *et al.* 2015. Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. The Plant Cell **27**, 485–512.

Al-Tamimi N, Brien C, Oakey H, Berger B, Saade S, Ho YS, Schmöckel SM, Tester M, Negrão S. 2016. Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. Nature Communications **7**, 13342.

Álvarez-Venegas R, Zhang Y, Kraling K, Tulsieram L. 2014. Flowering Without Vernalization in Winter Canola (*Brassica napus*): use of Virus-Induced Gene Silencing (VIGS) to accelerate genetic gain. Nova Scientia **3**, 29–50.

Ambavaram MMR, Basu S, Krishnan A, Ramegowda V, Batlang U, Rahman L, Baisakh N, Pereira A. 2014. Coordinated regulation of photosynthesis in rice increases yield and tolerance to environmental stress. Nature Communications **5**, 5302.

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics **31**, 166–169.

Arend D, Lange M, Pape J-M, Weigelt-Fischer K, Arana-Ceballos F, Mücke I, Klukas C, Altmann T, Scholz U, Junker A. 2016. Quantitative monitoring of *Arabidopsis thaliana* growth and development using high-throughput plant phenotyping. Scientific Data **3**, 160055.

Arifuzzaman M, Oladzadabbasabadi A, McClean P, Rahman M. 2019. Shovelomics for phenotyping root architectural traits of rapeseed/canola (*Brassica napus* L.) and genome-wide association mapping. Molecular genetics and genomics **294**, 985–1000.

Atkinson JA, Pound MP, Bennett MJ, Wells DM. 2019. Uncovering the hidden half of plants using new advances in root phenotyping. Current Opinion in Biotechnology **55**, 1–8.

Autran D, Jonak C, Belcram K, Beemster GTS, Kronenberger J, Grandjean O, Inzé D, Traas J. 2002. Cell numbers and leaf development in Arabidopsis: a functional analysis of the STRUWWELPETER gene. The EMBO journal **21**, 6036–6049.

Bac-Molenaar JA, Vreugdenhil D, Granier C, Keurentjes JJB. 2015. Genome-wide association mapping of growth dynamics detects time-specific and general quantitative trait loci. Journal of Experimental Botany **66**, 5567–5580.

Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Valè G, Cattivelli L. 2016. Next generation breeding. Plant Science **242**, 3–13.

Basunanda P, Radoev M, Ecke W, Friedt W, Becker HC, Snowdon RJ. 2010. Comparative mapping of quantitative trait loci involved in heterosis for seedling and yield traits in oilseed rape (*Brassica napus* L.). Theoretical and Applied Genetics **120**, 271–281.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software **67**, 1–48.

Battaglia R, Brambilla V, Colombo L, Stuitje AR, Kater MM. 2006. Functional analysis of MADS-box genes controlling ovule development in Arabidopsis using the ethanol-inducible alc gene-expression system. Mechanisms of Development **123**, 267–276.

Bayer PE, Ruperao P, Mason AS, *et al.* 2015. High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. Theoretical and Applied Genetics. **128**, 1039–1047.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **57**, 289–300.

Berger S, Sinha AK, Roitsch T. 2007. Plant physiology meets phytopathology: plant primary metabolism and plant-pathogen interactions. Journal of Experimental Botany **58**, 4019–4026.

Bernardo R. 1994. Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Science **34**, 20–25.

Bock R. 2017. Witnessing Genome Evolution: Experimental Reconstruction of Endosymbiotic and Horizontal Gene Transfer. Annual Review of Genetics **51**, 1–22.

Bock R, Timmis JN. 2008. Reconstructing evolution: gene transfer from plastids to the nucleus. BioEssays **30**, 556–566.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114–2120.

Bolon Y-T, Hyten DL, Orf JH, Vance CP, Muehlbauer GJ. 2014. eQTL Networks Reveal Complex Genetic Architecture in the Immature Soybean Seed. The Plant Genome **7**, 1–14.

Borlaug NE. 1983. Contributions of conventional plant breeding to food production. Science **219**, 689–693.

Box GEP, Cox DR. 1964. An Analysis of Transformations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **26**, 211–252.

Bradshaw AD. 1965. Evolutionary Significance of Phenotypic Plasticity in Plants. Advances in Genetics. Elsevier, 115–155.

Brandt R, Cabedo M, Xie Y, Wenkel S. 2014. Homeodomain leucine-zipper proteins and their role in synchronizing growth and development with the environment. Journal of Integrative Plant Biology **56**, 518–526.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology **34**, 525–527.

Brotman Y, Riewe D, Lisec J, Meyer RC, Willmitzer L, Altmann T. 2011. Identification of enzymatic and regulatory genes of plant metabolism through QTL analysis in Arabidopsis. Journal of Plant Physiology **168**, 1387–1394.

Busemeyer L, Ruckelshausen A, Möller K, Melchinger AE, Alheit KV, Maurer HP, Hahn V, Weissmann EA, Reif JC, Würschum T. 2013. Precision phenotyping of biomass accumulation in triticale reveals temporal genetic patterns of regulation. Scientific Reports **3**, 2442.

Cabrera-Bosquet L, Fournier C, Brichet N, Welcker C, Suard B, Tardieu F. 2016. High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. New Phytologist **212**, 269–281.

Cai G, Yang Q, Chen H, Yang Q, Zhang C, Fan C, Zhou Y. 2016. Genetic dissection of plant architecture and yield-related traits in *Brassica napus*. Scientific Reports **6**, 21625.

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics **178**, 553–561.

Cao HX, Schmidt R. 2013. Screening of a *Brassica napus* bacterial artificial chromosome library using highly parallel single nucleotide polymorphism assays. BMC Genomics **14**, 603.

Century K, Reuber TL, Ratcliffe OJ. 2008. Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. Plant Physiology **147**, 20–29.

Chalhoub B, Denoeud F, Liu S, *et al.* 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science **345**, 950–953.

Chan AC, Khan D, Girard IJ, Becker MG, Millar JL, Sytnik D, Belmonte MF. 2016. Tissue-specific laser microdissection of the *Brassica napus* funiculus improves gene discovery and spatial identification of biological processes. Journal of Experimental Botany **67**, 3561–3571.

Chawade A, van Ham J, Blomquist H, Bagge O, Alexandersson E, Ortiz R. 2019. High-Throughput Field-Phenotyping Tools for Plant Breeding and Precision Agriculture. Agronomy **9**, 258.

Chen ZJ. 2013. Genomic and epigenetic insights into the molecular bases of heterosis. Nature Reviews Genetics **14**, 471–482.

Chen SW, Liu T, Gao Y, Zhang C, Peng SD, Bai MB, Li SJ, Xu L, Zhou XY, Lin LB. 2016. Discovery of clubroot-resistant genes in *Brassica napus* by transcriptome sequencing. Genetics and Molecular Research **15**, doi.org/10.4238/gmr.15038243.

Chen D, Neumann K, Friedel S, Kilian B, Chen M, Altmann T, Klukas C. 2014. Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. The Plant Cell **26**, 4636–4655.

Chen Y, Sidhu HS, Kaviani M, McElroy MS, Pozniak CJ, Navabi A. 2019. Application of image-based phenotyping tools to identify QTL for in-field winter survival of winter wheat (*Triticum aestivum* L.). Theoretical and Applied Genetics **132**, 2591–2604.

Chen J, Tan R-K, Guo X-J, Fu Z-L, Wang Z, Zhang Z-Y, Tan X-L. 2015. Transcriptome Analysis Comparison of Lipid Biosynthesis in the Leaves and Developing Seeds of *Brassica napus*. PLOS ONE **10**, e0126250.

Chen W, Zhang Y, Liu X, Chen B, Tu J, Tingdong F. 2007. Detection of QTL for six yield-related traits in oilseed rape (*Brassica napus*) using DH and immortalized F2 populations. Theoretical and Applied Genetics **115**, 849–858.

Cheng Y, Cao L, Wang S, *et al.* 2013. Downregulation of multiple CDK inhibitor ICK/KRP genes upregulates the E2F pathway and increases cell proliferation, and organ and seed sizes in Arabidopsis. The Plant Journal **75**, 642–655.

Choe S, Fujioka S, Noguchi T, Takatsuto S, Yoshida S, Feldmann KA. 2001. Overexpression of DWARF4 in the brassinosteroid biosynthetic pathway results in increased vegetative growth and seed yield in Arabidopsis. The Plant Journal **26**, 573–582.

Civelek M, Lusis AJ. 2014. Systems genetics approaches to understand complex traits. Nature Reviews Genetics **15**, 34–48.

Clarke WE, Higgins EE, Plieske J, *et al.* 2016. A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. Theoretical and Applied Genetics. **129**, 1887–1899.

Cnops G, Jover-Gil S, Peters JL, Neyt P, De Block S, Robles P, Ponce MR, Gerats T, Micol JL, Van Lijsebettens M. 2004. The rotunda2 mutants identify a role for the LEUNIG gene in vegetative leaf morphogenesis. Journal of Experimental Botany **55**, 1529–1539.

Collard BCY, Mackill DJ. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences **363**, 557–572.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**, 3674–3676.

Cong B, Liu J, Tanksley SD. 2002. Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations. Proceedings of the National Academy of Sciences of the United States of America **99**, 13606–13611.

Covarrubias-Pazaran G. 2016. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLOS ONE **11**, e0156744.

Crossa J, Campos G de L, Pérez P, *et al.* 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics **186**, 713–724.

Crossa J, Pérez P, Hickey J, *et al.* 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity **112**, 48–60.

Custers JBM. 2003. Microspore culture in rapeseed (*Brassica napus* L.). In: Maluszynski M, Kasha KJ, Forster BP, Szarejko I (eds). Doubled Haploid Production in Crop Plants. Springer, 185–193.

Czedik-Eysenberg A, Arrivault S, Lohse MA, *et al.* 2016. The Interplay between Carbon Availability and Growth in Different Zones of the Growing Maize Leaf. Plant Physiology **172**, 943–967.

Dan Z, Chen Y, Xu Y, Huang J, Huang J, Hu J, Yao G, Zhu Y, Huang W. 2018. A metabolome-based core hybridization strategy for the prediction of rice grain weight across environments. Plant Biotechnology Journal **17**, 906–913.

Dan Z, Hu J, Zhou W, Yao G, Zhu R, Zhu Y, Huang W. 2016. Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). Scientific Reports **6**, 21732.

Del Carmen Martínez-Ballesta M, Moreno DA, Carvajal M. 2013. The physiological importance of glucosinolates on plant response to abiotic stress in Brassica. International Journal of Molecular Sciences **14**, 11607–11625.

Desta ZA, Ortiz R. 2014. Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science **19**, 592–601.

Dong H, Tan C, Li Y, *et al.* 2018. Genome-Wide Association Study Reveals Both Overlapping and Independent Genetic Loci to Control Seed Weight and Silique Length in *Brassica napus*. Frontiers in Plant Science **9**, 921.

Doyle JJ, Sherman-Broyles S. 2017. Double trouble: taxonomy and definitions of polyploidy. New Phytologist **213**, 487–493.

Ecarnot M, Bączyk P, Tessarotto L, Chervin C. 2013. Rapid phenotyping of the tomato fruit model, Micro-Tom, with a portable VIS-NIR spectrometer. Plant Physiology and Biochemistry **70**, 159–163.

Elferjani R, Soolanayakanahally R. 2018. Canola Responses to Drought, Heat, and Combined Stress: Shared and Specific Effects on Carbon Assimilation, Seed Yield, and Oil Composition. Frontiers in Plant Science **9**, 1224.

Eloy NB, Gonzalez N, Van Leene J, *et al.* 2012. SAMBA, a plant-specific anaphase-promoting complex/cyclosome regulator is involved in early development and A-type cyclin stabilization. Proceedings of the National Academy of Sciences of the United States of America **109**, 13853–13858.

Endelman JB. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome **4**, 250–255.

Escobar M, Berti M, Matus I, Tapia M, Johnson B. 2011. Genotype × Environment Interaction in Canola (*Brassica napus* L.) Seed Yield in Chile. Chilean Journal of Agricultural Research **71**, 175–186.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology **14**, 2611–2620.

Ewing B, Green P. 1998. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. Genome Research **8**, 186–194.

Fahlgren N, Feldman M, Gehan MA, *et al.* 2015. A Versatile Phenotyping System and Analytics Platform Reveals Diverse Temporal Responses to Water Availability in Setaria. Molecular Plant **8**, 1520–1535.

Fang Y, Ramasamy R. 2015. Current and Prospective Methods for Plant Disease Detection. Biosensors **5**, 537–561.

Fanourakis D, Briese C, Max JF, Kleinen S, Putz A, Fiorani F, Ulbrich A, Schurr U. 2014. Rapid determination of leaf area and plant height by using light curtain arrays in four species with contrasting shoot architecture. Plant Methods **10**, 9.

Farré EM. 2012. The regulation of plant growth by the circadian clock. Plant Biology **14**, 401–410.

Favaro R, Pinyopich A, Battaglia R, Kooiker M, Borghi L, Ditta G, Yanofsky MF, Kater MM, Colombo L. 2003. MADS-Box Protein Complexes Control Carpel and Ovule Development in Arabidopsis. The Plant Cell **15**, 2603–2611.

Feher K, Lisec J, Römisch-Margl L, Selbig J, Gierl A, Piepho H-P, Nikoloski Z, Willmitzer L. 2014. Deducing Hybrid Performance from Parental Metabolic Profiles of Young Primary Roots of Maize by Using a Multivariate Diallel Approach. PLOS ONE **9**, e85435.

Feldman MJ, Paul RE, Banan D, *et al.* 2017. Time dependent genetic analysis links field and controlled environment phenotypes in the model C4 grass Setaria. PLOS Genetics **13**, e1006841.

Fiehn O. 2002. Metabolomics--the link between genotypes and phenotypes. Plant Molecular Biology **48**, 155–171.

Fiévet JB, Dillmann C, de Vienne D. 2010. Systemic properties of metabolic networks lead to an epistasis-based model for heterosis. Theoretical and Applied Genetics. **120**, 463–473.

Fiévet JB, Nidelet T, Dillmann C, de Vienne D. 2018. Heterosis Is a Systemic Property Emerging From Non-linear Genotype-Phenotype Relationships: Evidence From in Vitro Genetics and Computer Simulations. Frontiers in Genetics **9**, 159.

Fikere M, Barbulescu DM, Malmberg MM, *et al.* 2018. Genomic Prediction Using Prior Quantitative Trait Loci Information Reveals a Large Reservoir of Underutilised Blackleg Resistance in Diverse Canola (*Brassica napus* L.) Lines. The Plant Genome **11**, 170100.

Flügge U-I, Häusler RE, Ludewig F, Gierth M. 2011. The role of transporters in supplying energy to plant plastids. Journal of Experimental Botany **62**, 2381–2392.

Fort A, Ryder P, McKeown PC, Wijnen C, Aarts MG, Sulpice R, Spillane C. 2016. Disaggregating polyploidy, parental genome dosage and hybridity contributions to heterosis in *Arabidopsis thaliana*. New Phytologist **209**, 590–599.

de Freitas Lima M, Eloy NB, Bottino MC, Hemerly AS, Ferreira PCG. 2013. Overexpression of the anaphase-promoting complex (APC) genes in Nicotiana tabacum promotes increasing biomass accumulation. Molecular Biology Reports **40**, 7093–7102.

Frisch M, Thiemann A, Fu J, Schrag TA, Scholten S, Melchinger AE. 2010. Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theoretical and Applied Genetics. **120**, 441–450.

Fu J, Falke KC, Thiemann A, Schrag TA, Melchinger AE, Scholten S, Frisch M. 2012. Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. Theoretical and Applied Genetics. **124**, 825–833.

Fu J, Keurentjes JJB, Bouwmeester H, *et al.* 2009. System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nature Genetics **41**, 166–167.

Fu Y, Lu K, Qian L, *et al.* 2015. Development of genic cleavage markers in association with seed glucosinolate content in canola. Theoretical and Applied Genetics. **128**, 1029–1037.

Fujimoto R, Taylor JM, Shirasawa S, Peacock WJ, Dennis ES. 2012. Heterosis of Arabidopsis hybrids between C24 and Col is associated with increased photosynthesis capacity. Proceedings of the National Academy of Sciences of the United States of America **109**, 7109–7114.

Fujimoto R, Uezono K, Ishikura S, Osabe K, Peacock WJ, Dennis ES. 2018. Recent research on the mechanism of heterosis is important for crop and vegetable breeding systems. Breeding Science **68**, 145–158.

Furbank RT, Tester M. 2011. Phenomics--technologies to relieve the phenotyping bottleneck. Trends in Plant Science **16**, 635–644.

Furtado A, Lupoi JS, Hoang NV, Healey A, Singh S, Simmons BA, Henry RJ. 2014. Modifying plants for biofuel and biomaterial production. Plant Biotechnology Journal **12**, 1246–1258.

Gärtner T, Steinfath M, Andorf S, Lisec J, Meyer RC, Altmann T, Willmitzer L, Selbig J. 2009. Improved heterosis prediction by combining information on DNA- and metabolic markers. PLOS ONE **4**, e5220.

Gehan MA, Fahlgren N, Abbasi A, *et al.* 2017. PlantCV v2: Image analysis software for high-throughput plant phenotyping. PeerJ **5**, e4088.

Gehringer A, Snowdon R, Spiller T, Basunanda P, Friedt W. 2007. New Oilseed Rape (*Brassica napus*) Hybrids with High Levels of Heterosis for Seed Yield under Nutrient-poor Conditions. Breeding Science **57**, 315–320.

Geng X, Dong N, Wang Y, Li G, Wang L, Guo X, Li J, Wen Z, Wei W. 2018. RNA-seq transcriptome analysis of the immature seeds of two *Brassica napus* lines with extremely different thousand-seed weight to identify the candidate genes related to seed weight. PLOS ONE **13**, e0191297.

Ghosh S, Watson A, Gonzalez-Navarro OE, *et al.* 2018. Speed breeding in growth chambers and glasshouses for crop breeding and model plant research. Nature Protocols **13**, 2944–2963.

Gianola D, van Kaam JBCHM. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics **178**, 2289–2303.

Gibon Y, Usadel B, Blaesing OE, Kamlage B, Hoehne M, Trethewey R, Stitt M. 2006. Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. Genome Biology **7**, R76.

Gilchrist EJ, Sidebottom CHD, Koh CS, Macinnes T, Sharpe AG, Haughn GW. 2013. A mutant *Brassica napus* (canola) population for the identification of new genetic diversity via TILLING and next generation sequencing. PLOS ONE **8**, e84303.

Girke A, Schierholt A, Becker HC. 2012. Extending the rapeseed gene pool with resynthesized *Brassica napus* II: Heterosis. Theoretical and Applied Genetics. **124**, 1017–1026.

Goddard M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136**, 245–257.

Golzarian MR, Frick RA, Rajendran K, Berger B, Roy S, Tester M, Lun DS. 2011. Accurate inference of shoot biomass from high-throughput images of cereal plants. Plant Methods **7**, 2.

Gonzalez N, Beemster GT, Inzé D. 2009. David and Goliath: what can the tiny weed Arabidopsis teach us to improve biomass production in crops? Current Opinion in Plant Biology **12**, 157–164.

González N, Inzé D. 2015. Molecular systems governing leaf growth: from genes to networks. Journal of Experimental Botany **66**, 1045–1054.

Gonzalez N, Vanhaeren H, Inzé D. 2012. Leaf size control: complex coordination of cell division and expansion. Trends in Plant Science **17**, 332–340.

Graf A, Schlereth A, Stitt M, Smith AM. 2010. Circadian control of carbohydrate availability for growth in Arabidopsis plants at night. Proceedings of the National Academy of Sciences of the United States of America **107**, 9458–9463.

Grandke F, Snowdon R, Samans B. 2017. gsrc: an R package for genome structure rearrangement calling. Bioinformatics **33**, 545–546.

Granier C, Aguirrezabal L, Chenu K, *et al.* 2006. PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in *Arabidopsis thaliana* permitted the identification of an accession with low sensitivity to soil water deficit. New Phytologist **169**, 623–635.

Grover A, Singh S, Pandey P, Patade VY, Gupta SM, Nasim M. 2014. Overexpression of NAC gene from *Lepidium latifolium* L. enhances biomass, shortens life cycle and induces cold stress tolerance in tobacco: potential for engineering fourth generation biofuel crops. Molecular Biology Reports **41**, 7479–7489.

Grubb CD, Abel S. 2006. Glucosinolate metabolism and its control. Trends in Plant Science **11**, 89–100.

Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D. 2016. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theoretical and Applied Genetics. **129**, 2413–2427.

Guo M, Rupe MA, Danilevskaya ON, Yang X, Hu Z. 2003. Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. The Plant Journal **36**, 30–44.

Guo YM, Samans B, Chen S, Kibret KB, Hatzig S, Turner NC, Nelson MN, Cowling WA, Snowdon RJ. 2017. Drought-Tolerant *Brassica rapa* Shows Rapid Expression of Gene Networks for General Stress Responses and Programmed Cell Death Under Simulated Drought Stress. Plant Molecular Biology Reporter **35**, 416–430.

Habier D, Fernando RL, Dekkers JCM. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**, 2389–2397.

Habier D, Fernando RL, Garrick DJ. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics **194**, 597–607.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics **12**, 186.

Hairmansis A, Berger B, Tester M, Roy SJ. 2014. Image-based phenotyping for non-destructive screening of different salinity tolerance traits in rice. Rice **7**, 16.

Halkier BA, Gershenzon J. 2006. Biology and Biochemistryof Glucosinolates. Annual Review of Plant Biology **57**, 303–333.

Hannoufa A, Pillai BVS, Chellamma S. 2014. Genetic enhancement of *Brassica napus* seed quality. Transgenic Research **23**, 39–52.

Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, Hattori C, Werner P, Bancroft I. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nature Biotechnology **30**, 798–802.

Hartmann A, Czauderna T, Hoffmann R, Stein N, Schreiber F. 2011. HTPheno: an image analysis pipeline for high-throughput plant phenotyping. BMC Bioinformatics **12**, 148.

Hatzig S, Breuer F, Nesi N, Ducournau S, Wagner M-H, Leckband G, Abbadi A, Snowdon RJ. 2018. Hidden Effects of Seed Quality Breeding on Germination in Oilseed Rape (*Brassica napus* L.). Frontiers in Plant Science **9**, 419.

Hatzig SV, Frisch M, Breuer F, Nesi N, Ducournau S, Wagner M-H, Leckband G, Abbadi A, Snowdon RJ. 2015. Genome-wide association mapping unravels the genetic control of seed germination and vigor in *Brassica napus*. Frontiers in Plant Science **6**, 221.

Havlickova L, He Z, Wang L, Langer S, Harper AL, Kaur H, Broadley MR, Gegas V, Bancroft I. 2018. Validation of an updated Associative Transcriptomics platform for the polyploid crop species *Brassica napus* by dissection of the genetic architecture of erucic acid and tocopherol isoform variation in seeds. The Plant Journal **93**, 181–192.

He Y, Fu Y, Hu D, Wei D, Qian W. 2018. QTL Mapping of Seed Glucosinolate Content Responsible for Environment in *Brassica napus*. Frontiers in Plant Science **9**, 891.

He S, Schulthess AW, Mirdita V, Zhao Y, Korzun V, Bothe R, Ebmeyer E, Reif JC, Jiang Y. 2016. Genomic selection in a commercial winter wheat population. Theoretical and Applied Genetics. **129**, 641–651.

Hedden P, Sponsel V. 2015. A Century of Gibberellin Research. Journal of Plant Growth Regulation **34**, 740–760.

Heino M. 2014. Quantitative Traits. In: Cadrin SX, Kerr LA, Mariani S (eds). Stock Identification Methods. Academic Press, 59–76.

Heslot N, Jannink J-L, Sorrells ME. 2015. Perspectives for Genomic Selection Applications and Research in Plants. Crop Science **55**, 1–12.

Hickey JM, Chiurugwi T, Mackay I, Powell W, Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants. 2017. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature Genetics **49**, 1297–1303.

Higgins J, Magusin A, Trick M, Fraser F, Bancroft I. 2012. Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species *Brassica napus*. BMC Genomics **13**, 247.

Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theoretical Population Biology **33**, 54–78.

Hirsch S, Oldroyd GED. 2009. GRAS-domain transcription factors that regulate plant development. Plant Signaling & Behavior **4**, 698–700.

Honsdorf N, March TJ, Berger B, Tester M, Pillen K. 2014. High-throughput phenotyping to detect drought tolerance QTL in wild barley introgression lines. PLOS ONE **9**, e97047.

Horiguchi G, Kim G-T, Tsukaya H. 2005. The transcription factor AtGRF5 and the transcription coactivator AN3 regulate cell proliferation in leaf primordia of *Arabidopsis thaliana*. The Plant Journal **43**, 68–78.

Hu S, Wang C, Sanchez DL, Lipka AE, Liu P, Yin Y, Blanco M, Lübberstedt T. 2017. Gibberellins Promote Brassinosteroids Action and Both Increase Heterosis for Plant Height in Maize (*Zea mays* L.). Frontiers in Plant Science **8**, 1039.

Hu D, Zhang W, Zhang Y, Chang S, Chen L, Chen Y, Shi Y, Shen J, Meng J, Zou J. 2019. Reconstituting the genome of a young allopolyploid crop, *Brassica napus,* with its related species. Plant Biotechnology Journal **17**, 1106–1118.

Hunt R. 1990. *Basic Growth Analysis*. Dordrecht: Springer Netherlands.

Hurgobin B, Golicz AA, Bayer PE, *et al.* 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. Plant Biotechnology Journal **16**, 1265–1274.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. PLOS ONE **5**, e12776.

Inagaki S, Umeda M. 2011. Cell-cycle control and plant development. International Review of Cell and Molecular Biology **291**, 227–261.

Ingvarsson PK, Street NR. 2011. Association genetics of complex traits in plants: Tansley review. New Phytologist **189**, 909–922.

Inzé D, De Veylder L. 2006. Cell cycle regulation in plant development. Annual Review of Genetics **40**, 77–105.

Ishida M, Hara M, Fukino N, Kakizaki T, Morimitsu Y. 2014. Glucosinolate metabolism, functionality and breeding for the improvement of Brassicaceae vegetables. Breeding Science **64**, 48–59.

Ishihara H, Moraes TA, Pyl E-T, Schulze WX, Obata T, Scheffel A, Fernie AR, Sulpice R, Stitt M. 2017. Growth rate correlates negatively with protein turnover in Arabidopsis accessions. The Plant Journal **91**, 416–429.

Jamai A, Salomé PA, Schilling SH, Weber APM, McClung CR. 2009. Arabidopsis photorespiratory serine hydroxymethyltransferase activity requires the mitochondrial accumulation of ferredoxin-dependent glutamate synthase. The Plant Cell **21**, 595–606.

Jan HU, Abbadi A, Lücke S, Nichols RA, Snowdon RJ. 2016. Genomic Prediction of Testcross Performance in Canola (*Brassica napus*). PLOS ONE **11**, e0147769.

Jan HU, Guan M, Yao M, *et al.* 2019. Genome-wide haplotype analysis improves trait predictions in *Brassica napus* hybrids. Plant Science **283**, 157–164.

Jeon H-W, Cho J-S, Park E-J, Han K-H, Choi Y-I, Ko J-H. 2016. Developing xylem-preferential expression of PdGA20ox1, a gibberellin 20-oxidase 1 from *Pinus densiflora*, improves woody biomass production in a hybrid poplar. Plant Biotechnology Journal **14**, 1161–1170.

Jesske T, Olberg B, Schierholt A, Becker HC. 2013. Resynthesized lines from domesticated and wild Brassica taxa and their hybrids with *B. napus* L.: genetic diversity and hybrid yield. Theoretical and Applied Genetics. **126**, 1053–1065.

Jiang G-L. 2015. Molecular Marker-Assisted Breeding: A Plant Breeder's Review. In: Al-Khayri JM, Jain SM, Johnson DV (eds.) Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools. Cham: Springer International Publishing, 431–472.

Jiang Y, Reif JC. 2015. Modeling Epistasis in Genomic Selection. Genetics **201**, 759–768.

Jiang Y, Schmidt RH, Reif JC. 2018. Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. G3: Genes, Genomes, Genetics **8**, 1687–1699.

Jimenez-Berni JA, Deery DM, Rozas-Larraondo P, Condon ATG, Rebetzke GJ, James RA, Bovill WD, Furbank RT, Sirault XRR. 2018. High Throughput Determination of Plant Height, Ground Cover, and Above-Ground Biomass in Wheat with LiDAR. Frontiers in Plant Science **9**, 237.

Joosen RVL, Arends D, Li Y, Willems LAJ, Keurentjes JJB, Ligterink W, Jansen RC, Hilhorst HWM. 2013. Identifying genotype-by-environment interactions in the metabolism of germinating arabidopsis seeds using generalized genetical genomics. Plant Physiology **162**, 553–566.

Jordan KW, Nordenstam J, Lauwers GY, Rothenberger DA, Alavi K, Garwood M, Cheng LL. 2009. Metabolomic characterization of human rectal adenocarcinoma with intact tissue magnetic resonance spectroscopy. Diseases of the Colon and Rectum **52**, 520–525.

Joshi R, Wani SH, Singh B, Bohra A, Dar ZA, Lone AA, Pareek A, Singla-Pareek SL. 2016. Transcription Factors and Plants Response to Drought Stress: Current Understanding and Future Directions. Frontiers in Plant Science **7**, 1029.

Junker A, Muraya MM, Weigelt-Fischer K, Arana-Ceballos F, Klukas C, Melchinger AE, Meyer RC, Riewe D, Altmann T. 2015. Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. Frontiers in Plant Science **5**, 770.

Kacser H, Burns JA. 1981. The molecular basis of dominance. Genetics **97**, 639–666.

Kadam DC, Potts SM, Bohn MO, Lipka AE, Lorenz AJ. 2016. Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. G3: Genes, Genomes, Genetics **6**, 3443–3453.

Kawanabe T, Ishikura S, Miyaji N, *et al.* 2016. Role of DNA methylation in hybrid vigor in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America **113**, E6704–E6711.

Keller WA, Rajhathy T, Lacapra J. 1975. *In vitro* production of plants from pollen in *Brassica campestris*. Canadian Journal of Genetics and Cytology **17**, 655–666.

Kicherer A, Herzog K, Pflanz M, Wieland M, Rüger P, Kecke S, Kuhlmann H, Töpfer R. 2015. An automated field phenotyping pipeline for application in grapevine research. Sensors **15**, 4823–4836.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods **12**, 357–360.

Kirschbaum MUF. 2011. Does enhanced photosynthesis enhance growth? Lessons learned from CO2 enrichment studies. Plant Physiology **155**, 117–124.

Kjaer KH, Ottosen C-O. 2015. 3D Laser Triangulation for Plant Phenotyping in Challenging Environments. Sensors **15**, 13533–13547.

Kliebenstein D. 2009. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. Annual Review of Plant Biology **60**, 93–114.

Klukas C, Chen D, Pape J-M. 2014. Integrated Analysis Platform: An Open-Source Information System for High-Throughput Plant Phenotyping. Plant Physiology **165**, 506–518.

Knoch D, Riewe D, Meyer RC, Boudichevskaia A, Schmidt R, Altmann T. 2017. Genetic dissection of metabolite variation in Arabidopsis seeds: evidence for mQTL hotspots and a master regulatory locus of seed metabolism. Journal of Experimental Botany **68**, 1655–1667.

Kole C (Ed.). 2007. *Oilseeds*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Körber N, Bus A, Li J, Higgins J, Bancroft I, Higgins EE, Parkin IAP, Salazar-Colqui B, Snowdon RJ, Stich B. 2015. Seedling development traits in *Brassica napus* examined by gene expression analysis and association mapping. BMC Plant Biology **15**, 136.

Körber N, Bus A, Li J, Parkin IAP, Wittkop B, Snowdon RJ, Stich B. 2016. Agronomic and Seed Quality Traits Dissected by Genome-Wide Association Mapping in *Brassica napus*. Frontiers in Plant Science **7**, 386.

Kroymann J, Mitchell-Olds T. 2005. Epistasis and balanced polymorphism influencing complex trait variation. Nature **435**, 95–98.

Kudapa H, Garg V, Chitikineni A, Varshney RK. 2018. The RNA-Seq-based high resolution gene expression atlas of chickpea (*Cicer arietinum* L.) reveals dynamic spatio-temporal changes associated with growth and development. Plant, Cell & Environment **41**, 2209–2225.

Kuroki S, Tsenkova R, Moyankova D, Muncan J, Morita H, Atanassova S, Djilianov D. 2019. Water molecular structure underpins extreme desiccation tolerance of the resurrection plant Haberlea rhodopensis. Scientific Reports **9**, 3049.

Kwon CS, Chen C, Wagner D. 2005. WUSCHEL is a primary target for transcriptional regulation by SPLAYED in dynamic control of stem cell fate in Arabidopsis. Genes & Development **19**, 992–1003.

Lande R, Thompson R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124**, 743–756.

Lander E, Schork N. 1994. Genetic dissection of complex traits. Science **265**, 2037–2048.

Lee YK, Kim G-T, Kim I-J, Park J, Kwak S-S, Choi G, Chung W-I. 2006. LONGIFOLIA1 and LONGIFOLIA2, two homologous genes, regulate longitudinal cell elongation in Arabidopsis. Development **133**, 4305–4314.

Lee SI, Muthusamy M, Nawaz MA, Hong JK, Lim M-H, Kim JA, Jeong M-J. 2019. Genome-wide analysis of spatiotemporal gene expression patterns during floral organ development in *Brassica rapa*. Molecular Genetics and Genomics, doi.org/10.1007/s00438-019-01585-5.

Lees CJ, Li G, Duncan RW. 2016. Characterization of *Brassica napus* L. genotypes utilizing sequence-related amplified polymorphism and genotyping by sequencing in association with cluster analysis. Molecular Breeding **36**, 155.

Lefort-Buson M, Guillot-Lemoine B, Dattee Y. 1987. Heterosis and genetic distance in rapeseed (*Brassica napus* L.): crosses between European and Asiatic selfed lines. Genome **29**, 413–418.

Li F, Chen B, Xu K, *et al.* 2014*a*. Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). DNA Research **21**, 355–367.

Li F, Chen B, Xu K, *et al.* 2016*a*. A genome-wide association study of plant height and primary branch number in rapeseed (*Brassica napus*). Plant Science **242**, 169–177.

Li H, Cheng X, Zhang L, *et al.* 2018*a*. An Integration of Genome-Wide Association Study and Gene Co-expression Network Analysis Identifies Candidate Genes of Stem Lodging-Related Traits in *Brassica napus*. Frontiers in Plant Science **9**, 796.

Li D, Fu X, Guo L, *et al.* 2016*b*. FAR-RED ELONGATED HYPOCOTYL3 activates SEPALLATA2 but inhibits CLAVATA3 to regulate meristem determinacy and maintenance in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **113**, 9375–9380.

Li R, Jeong K, Davis JT, Kim S, Lee S, Michelmore RW, Kim S, Maloof JN. 2018*b*. Integrated QTL and eQTL Mapping Provides Insights and Candidate Genes for Fatty Acid Composition, Flowering Time, and Growth Traits in a F2 Population of a Novel Synthetic Allopolyploid *Brassica napus*. Frontiers in Plant Science **9**, 1632.

Li F, Ma C, Wang X, *et al.* 2011. Characterization of Sucrose transporter alleles and their association with seed yield-related traits in *Brassica napus* L. BMC Plant Biology **11**, 168.

Li L, Petsch K, Shimizu R, *et al.* 2013*a*. Mendelian and non-Mendelian regulation of gene expression in maize. PLOS Genetics **9**, e1003202.

Li C, Sun B, Li Y, *et al.* 2016*c*. Numerous genetic loci identified for drought tolerance in the maize nested association mapping populations. BMC Genomics **17**, 894.

Li J, Wen J, Lease KA, Doke JT, Tax FE, Walker JC. 2002. BAK1, an Arabidopsis LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. Cell **110**, 213–222.

Li K, Yao Y, Xiao L, Zhao Z, Guo S, Fu Z, Du D. 2018*c*. Fine mapping of the *Brassica napus* Bnsdt1 gene associated with determinate growth habit. Theoretical and Applied Genetics. **131**, 193–208.

Li H, Zhang L, Hu J, *et al.* 2017. Genome-Wide Association Mapping Reveals the Genetic Control Underlying Branch Angle in Rapeseed (*Brassica napus* L.). Frontiers in Plant Science **8**, 1054.

Li L, Zhang Q, Huang D. 2014*b*. A review of imaging techniques for plant phenotyping. Sensors **14**, 20078–20111.

Li Y, Zhang X, Ma C, Shen J, Chen Q, Wang T, Fu T, Tu J. 2012. QTL and epistatic analyses of heterosis for seed yield and three yield component traits using molecular markers in rapeseed (*Brassica napus* L.). Genetika **48**, 1171–1178.

Li A, Zhou Y, Jin C, Song W, Chen C, Wang C. 2013*b*. LaAP2L1, a heterosis-associated AP2/EREBP transcription factor of Larix, increases organ size and final biomass by affecting cell proliferation in Arabidopsis. Plant & Cell Physiology **54**, 1822–1836.

Li Q, Zhou Q, Mei J, Zhang Y, Li J, Li Z, Ge X, Xiong Z, Huang Y, Qian W. 2014*c*. Improvement of *Brassica napus* via interspecific hybridization between *B. napus* and *B. oleracea*. Molecular Breeding **34**, 1955–1963.

Liang Q, Li P, Hu C, Hua H, Li Z, Rong Y, Wang K, Hua J. 2014. Dynamic QTL and epistasis analysis on seedling root traits in upland cotton. Journal of Genetics **93**, 63–78.

Lima M de F, Eloy NB, Pegoraro C, *et al.* 2010. Genomic evolution and complexity of the Anaphase-promoting Complex (APC) in land plants. BMC Plant Biology **10**, 254.

Lima M de F, Eloy NB, Siqueira JAB de, Inzé D, Hemerly AS, Ferreira PCG. 2017. Molecular mechanisms of biomass increase in plants. Biotechnology Research and Innovation **1**, 14–25.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics **28**, 2397–2399.

Lippman ZB, Zamir D. 2007. Heterosis: revisiting the magic. Trends in Genetics **23**, 60–66.

Lisec J, Meyer RC, Steinfath M, *et al.* 2008. Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations: Metabolic and biomass QTL in Arabidopsis. The Plant Journal **53**, 960–972.

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR. 2006. Gas chromatography mass spectrometry–based metabolite profiling in plants. Nature Protocols **1**, 387–396.

Liu W, Gowda M, Reif JC, Hahn V, Ruckelshausen A, Weissmann EA, Maurer HP, Würschum T. 2014. Genetic dynamics underlying phenotypic development of biomass yield in triticale. BMC Genomics **15**, 458.

Liu J, Hua W, Hu Z, Yang H, Zhang L, Li R, Deng L, Sun X, Wang X, Wang H. 2015*a*. Natural variation in ARF18 gene simultaneously affects seed weight and silique length in polyploid rapeseed. Proceedings of the National Academy of Sciences of the United States of America **112**, E5123-5132.

Liu X, Huang M, Fan B, Buckler ES, Zhang Z. 2016*a*. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genetics **12**, e1005767.

Liu J, Li M, Zhang Q, Wei X, Huang X. 2019. Exploring the molecular basis of heterosis for plant breeding. Journal of Integrative Plant Biology, doi.org/10.1111/jipb.12804.

Liu S, Snowdon R, Chalhoub B (Eds.). 2018*a*. *The Brassica napus Genome*. Cham: Springer International Publishing.

Liu J, Wang W, Mei D, Wang H, Fu L, Liu D, Li Y, Hu Q. 2016*b*. Characterizing Variation of Branch Angle and Genome-Wide Association Mapping in Rapeseed (*Brassica napus* L.). Frontiers in Plant Science **7**, 21.

Liu Y, Xu A, Liang F, *et al.* 2018*b*. Screening of clubroot-resistant varieties and transfer of clubroot resistance genes to *Brassica napus* using distant hybridization. Breeding Science **68**, 258–267.

Liu F, Zhao Q, Mano N, Ahmed Z, Nitschke F, Cai Y, Chapman KD, Steup M, Tetlow IJ, Emes MJ. 2016*c*. Modification of starch metabolism in transgenic *Arabidopsis thaliana* increases plant biomass and triples oilseed production. Plant Biotechnology Journal **14**, 976–985.

Liu X, Zhao B, Zheng H-J, *et al.* 2015*b*. Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Scientific Reports **5**, 14139.

Longin CFH, Mi X, Würschum T. 2015. Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theoretical and Applied Genetics. **128**, 1297–1306.

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L. 2011. Genomic Selection in Plant Breeding. Advances in Agronomy. Elsevier, 77–123.

Lu G, Harper AL, Trick M, Morgan C, Fraser F, O'Neill C, Bancroft I. 2014. Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in *Brassica napus*. DNA Research **21**, 613–625.

Lu C, Napier JA, Clemente TE, Cahoon EB. 2011. New frontiers in oilseed biotechnology: meeting the global demand for vegetable oils for food, feed, biofuel, and industrial applications. Current Opinion in Biotechnology **22**, 252–259.

Lu K, Peng L, Zhang C, *et al.* 2017. Genome-Wide Association and Transcriptome Analyses Reveal Candidate Genes Underlying Yield-determining Traits in *Brassica napus*. Frontiers in Plant Science **8**, 206.

Lu K, Xiao Z, Jian H, *et al.* 2016. A combination of genome-wide association and transcriptome analysis reveals candidate genes controlling harvest index-related traits in *Brassica napus*. Scientific Reports **6**, 36452.

Luo X, Ding Y, Zhang L, Yue Y, Snyder JH, Ma C, Zhu J. 2017*a*. Genomic Prediction of Genotypic Effects with Epistasis and Environment Interactions for Yield-Related Traits of Rapeseed (*Brassica napus* L.). Frontiers in Genetics **8**, 15.

Luo X, Ma C, Yue Y, *et al.* 2015. Unravelling the complex trait of harvest index in rapeseed (*Brassica napus* L.) with association mapping. BMC Genomics **16**, 379.

Luo Z, Wang M, Long Y, *et al.* 2017*b*. Incorporating pleiotropic quantitative trait loci in dissection of complex traits: seed yield in rapeseed as an example. Theoretical and Applied Genetics. **130**, 1569–1585.

Mackay TFC, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics **10**, 565–577.

Maloney VJ, Park J-Y, Unda F, Mansfield SD. 2015. Sucrose phosphate synthase and sucrose phosphate phosphatase interact in planta and promote plant growth and biomass accumulation. Journal of Experimental Botany **66**, 4383–4394.

Mangin B, Bonnafous F, Blanchet N, *et al.* 2017. Genomic Prediction of Sunflower Hybrids Oil Content. Frontiers in Plant Science **8**, 1633.

Marjanović-Jeromela A, Nagl N, Gvozdanović-Varga J, Hristov N, Kondić-Špika A, Marinković MVR. 2011. Genotype by environment interaction for seed yield per plant in rapeseed using AMMI model. Pesquisa Agropecuária Brasileira **46**, 174–181.

Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, Morgante M. 2011. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. Tree Genetics & Genomes **7**, 1011–1023.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**, 10–12.

Martinez SA, Godoy J, Huang M, Zhang Z, Carter AH, Garland Campbell KA, Steber CM. 2018. Genome-Wide Association Mapping for Tolerance to Preharvest Sprouting and Low Falling Numbers in Wheat. Frontiers in Plant Science **9**, 141.

Mason AS, Higgins EE, Snowdon RJ, Batley J, Stein A, Werner C, Parkin IAP. 2017. A user guide to the Brassica 60K Illumina Infinium™ SNP genotyping array. Theoretical and Applied Genetics. **130**, 621–633.

Mason AS, Snowdon RJ. 2016. Oilseed rape: learning about ancient and recent polyploid evolution from a recent crop species. Plant Biology **18**, 883–892.

Matsuda O, Tanaka A, Fujita T, Iba K. 2012. Hyperspectral imaging techniques for rapid identification of Arabidopsis mutants with altered leaf pigment status. Plant & Cell Physiology **53**, 1154–1170.

Mei J, Guo Z, Wang J, Feng Y, Ma G, Zhang C, Qian W, Chen G. 2019. Understanding the Resistance Mechanism in *Brassica napus* to Clubroot Caused by *Plasmodiophora brassicae*. Phytopathology **109**, 810–818.

Melchinger AE, Gumber RK. 1998. Overview of heterosis and heterotic groups in agronomic crops. In: Lamkey KR, Staub JE (eds) Concepts and Breeding of Heterosis in Crop Plants. (Crop Science Society of America, Madison, WI), pp. 29-44

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**, 1819–1829.

Meyer RC, Steinfath M, Lisec J, *et al.* 2007. The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America **104**, 4759–4764.

Meyer RC, Törjék O, Becher M, Altmann T. 2004. Heterosis of biomass production in Arabidopsis. Establishment during early development. Plant Physiology **134**, 1813–1823.

Meyer RC, Witucka-Wall H, Becher M, *et al.* 2012. Heterosis manifestation during early Arabidopsis seedling development is characterized by intermediate gene expression and enhanced metabolic activity in the hybrids. The Plant Journal **71**, 669–683.

Miller CN, Harper AL, Trick M, Wellner N, Werner P, Waldron KW, Bancroft I. 2018. Dissecting the complex regulation of lodging resistance in *Brassica napus*. Molecular Breeding **38**, 30.

Mithen R, Raybould AF, Giamoustaris A. 1995. Divergent selection for secondary metabolites between wild populations of *Brassica oleracea* and its implications for plant-herbivore interactions. Heredity **75**, 472–484.

Mitrousia GK, Huang YJ, Qi A, Sidique SNM, Fitt BDL. 2018. Effectiveness of Rlm7 resistance against *Leptosphaeria maculans* (phoma stem canker) in UK winter oilseed rape cultivars. Plant Pathology **67**, 1339–1353.

Mizukami Y, Fischer RL. 2000. Plant organ size control: AINTEGUMENTA regulates growth and cell numbers during organogenesis. Proceedings of the National Academy of Sciences of the United States of America **97**, 942–947.

Momen M, Mehrgardi AA, Sheikhi A, Kranis A, Tusell L, Morota G, Rosa GJM, Gianola D. 2018. Predictive ability of genome-assisted statistical models under various forms of gene action. Scientific Reports **8**, 12309.

Moore CR, Johnson LS, Kwak I-Y, Livny M, Broman KW, Spalding EP. 2013. High-throughput computer vision introduces the time axis to a quantitative trait map of a plant growth response. Genetics **195**, 1077–1086.

Moreau L, Charcosset A, Hospital F, Gallais A. 1998. Marker-assisted selection efficiency in populations of finite size. Genetics **148**, 1353–1365.

Morgante F, Huang W, Maltecca C, Mackay TFC. 2018. Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. Heredity **120**, 500–514.

Morinaka Y, Sakamoto T, Inukai Y, Agetsuma M, Kitano H, Ashikari M, Matsuoka M. 2006. Morphological alteration caused by brassinosteroid insensitivity increases the biomass and grain production of rice. Plant Physiology **141**, 924–931.

van Muijen D, Anithakumari AM, Maliepaard C, Visser RGF, van der Linden CG. 2016. Systems genetics reveals key genetic elements of drought induced gene regulation in diploid potato. Plant, Cell & Environment **39**, 1895–1908.

Munns R, James RA, Sirault XRR, Furbank RT, Jones HG. 2010. New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. Journal of Experimental Botany **61**, 3499–3507.

Muñoz PR, Resende MFR, Gezan SA, Resende MDV, de Los Campos G, Kirst M, Huber D, Peter GF. 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics **198**, 1759–1768.

Muraya MM, Chu J, Zhao Y, Junker A, Klukas C, Reif JC, Altmann T. 2017. Genetic variation of growth dynamics in maize (*Zea mays* L.) revealed through automated non-invasive phenotyping. The Plant Journal **89**, 366–380.

Nagaharu U. 1935. Genome Analysis in Brassica with Special Reference to the Experimental Formation of *B. napus* and Peculiar Mode of Fertilization. Japanese Journal of Botany **7**, 389–452.

Nakabayashi K, Okamoto M, Koshiba T, Kamiya Y, Nambara E. 2005. Genome-wide profiling of stored mRNA in *Arabidopsis thaliana* seed germination: epigenetic and genetic regulation of transcription in seed. The Plant Journal **41**, 697–709.

Nakagawa S, Schielzeth H. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biological Reviews of the Cambridge Philosophical Society **85**, 935–956.

Neumann K, Klukas C, Friedel S, Rischbeck P, Chen D, Entzian A, Stein N, Graner A, Kilian B. 2015. Dissecting spatiotemporal biomass accumulation in barley under different water regimes using high-throughput image analysis. Plant, Cell & Environment **38**, 1980–1996.

Neumann K, Zhao Y, Chu J, Keilwagen J, Reif JC, Kilian B, Graner A. 2017. Genetic architecture and temporal patterns of biomass accumulation in spring barley revealed by image analysis. BMC Plant Biology **17**, 137.

Ni Z, Kim E-D, Ha M, Lackey E, Liu J, Zhang Y, Sun Q, Chen ZJ. 2009. Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. Nature **457**, 327–331.

Noh SA, Choi Y-I, Cho J-S, Lee H. 2015. The poplar basic helix-loop-helix transcription factor BEE3 - Like gene affects biomass production by enhancing proliferation of xylem cells in poplar. Biochemical and Biophysical Research Communications **462**, 64–70.

Okushima Y, Mitina I, Quach HL, Theologis A. 2005. AUXIN RESPONSE FACTOR 2 (ARF2): a pleiotropic developmental regulator. The Plant Journal **43**, 29–46.

Oliveira IC, Brears T, Knight TJ, Clark A, Coruzzi GM. 2002. Overexpression of cytosolic glutamine synthetase. Relation to nitrogen, light, and photorespiration. Plant Physiology **129**, 1170–1180.

Pal SK, Liput M, Piques M, *et al.* 2013. Diurnal changes of polysome loading track sucrose content in the rosette of wild-type arabidopsis and the starchless pgm mutant. Plant Physiology **162**, 1246–1265.

Pape J-M, Klukas C. 2015. Utilizing machine learning approaches to improve the prediction of leaf counts and individual leaf segmentation of rosette plant images. Proceedings of the Computer Vision Problems in Plant Phenotyping Workshop 2015. Swansea: British Machine Vision Association, 3.1-3.12.

Park T, Casella G. 2008. The Bayesian Lasso. Journal of the American Statistical Association **103**, 681–686.

Parkin IAP, Koh C, Tang H, *et al.* 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biology **15**, R77.

Parthasarathy A, Savka MA, Hudson AO. 2019. The Synthesis and Role of β-Alanine in Plants. Frontiers in Plant Science **10**, 921.

Patti GJ, Yanes O, Siuzdak G. 2012. Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews. Molecular Cell Biology **13**, 263–269.

Pauli D, Andrade-Sanchez P, Carmo-Silva AE, *et al.* 2016. Field-Based High-Throughput Plant Phenotyping Reveals the Temporal Patterns of Quantitative Trait Loci Associated with Stress-Responsive Traits in Cotton. G3: Genes, Genomes, Genetics **6**, 865–879.

Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF. 2000. B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature **405**, 200–203.

Pflugfelder D, Metzner R, van Dusschoten D, Reichel R, Jahnke S, Koller R. 2017. Non-invasive imaging of plant roots in different soils using magnetic resonance imaging (MRI). Plant Methods **13**, 102.

Pham A-T, Maurer A, Pillen K, Brien C, Dowling K, Berger B, Eglinton JK, March TJ. 2019. Genome-wide association of barley plant growth under drought stress using a nested association mapping population. BMC Plant Biology **19**, 134.

Piepho HP. 2009. Ridge Regression and Extensions for Genomewide Selection in Maize. Crop Science **49**, 1165.

Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, Wishart D. 2019. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites **9**, 76.

Piques M, Schulze WX, Höhne M, Usadel B, Gibon Y, Rohwer J, Stitt M. 2009. Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. Molecular Systems Biology **5**, 314.

Pommerrenig B, Junker A, Abreu I, Bieber A, Fuge J, Willner E, Bienert MD, Altmann T, Bienert GP. 2018. Identification of Rapeseed (*Brassica napus*) Cultivars With a High Tolerance to Boron-Deficient Conditions. Frontiers in Plant Science **9**, 1142.

Poorter H, Bühler J, van Dusschoten D, Climent J, Postma JA. 2012. Pot size matters: a meta-analysis of the effects of rooting volume on plant growth. Functional Plant Biology **39**, 839.

Prashar A, Jones HG. 2016. Assessing Drought Responses Using Thermal Infrared Imaging. Methods in Molecular Biology **1398**, 209–219.

Prem D, Solís M-T, Bárány I, Rodríguez-Sanz H, Risueño MC, Testillano PS. 2012. A new microspore embryogenesis system under low temperature which mimics zygotic embryogenesis initials, expresses auxin and efficiently regenerates doubled-haploid plants in *Brassica napus*. BMC Plant Biology **12**, 127.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics **155**, 945–959.

Pu Y, Liu L, Wu J, *et al.* 2019. Transcriptome Profile Analysis of Winter Rapeseed (*Brassica napus* L.) in Response to Freezing Stress, Reveal Potentially Connected Events to Freezing Stress. International Journal of Molecular Sciences **20**, 2771.

Qian W, Sass O, Meng J, Li M, Frauen M, Jung C. 2007. Heterotic patterns in rapeseed (*Brassica napus* L.): I. Crosses between spring and Chinese semi-winter lines. Theoretical and Applied Genetics **115**, 27–34.

Qian L, Voss-Fels K, Cui Y, Jan HU, Samans B, Obermeier C, Qian W, Snowdon RJ. 2016. Deletion of a Stay-Green Gene Associates with Adaptive Selection in *Brassica napus*. Molecular Plant **9**, 1559–1569.

Qu C, Zhao H, Fu F, Zhang K, Yuan J, Liu L, Wang R, Xu X, Lu K, Li J-N. 2016. Molecular Mapping and QTL for Expression Profiles of Flavonoid Genes in *Brassica napus*. Frontiers in Plant Science **7**, 1691.

R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radoev M, Becker HC, Ecke W. 2008. Genetic Analysis of Heterosis for Yield and Yield Components in Rapeseed (*Brassica napus* L.) by Quantitative Trait Locus Mapping. Genetics **179**, 1547–1558.

Rahaman MM, Chen D, Gillani Z, Klukas C, Chen M. 2015. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. Frontiers in Plant Science **6**, 619.

Raineri J, Ribichich KF, Chan RL. 2015. The sunflower transcription factor HaWRKY76 confers drought and flood tolerance to *Arabidopsis thaliana* plants without yield penalty. Plant Cell Reports **34**, 2065–2080.

Ravelombola W, Shi A, Weng Y, *et al.* 2017. Association analysis of salt tolerance in cowpea (*Vigna unguiculata* (L.) Walp) at germination and seedling stages. Theoretical and Applied Genetics **131**, 79–91.

Rebetzke GJ, Jimenez-Berni J, Fischer RA, Deery DM, Smith DJ. 2019. Review: High-throughput phenotyping to enhance the use of crop genetic resources. Plant Science **282**, 40–48.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences of the United States of America **98**, 11479–11484.

Ren T, Hu Y, Tang Y, Li C, Yan B, Ren Z, Tan F, Tang Z, Fu S, Li Z. 2018. Utilization of a Wheat55K SNP Array for Mapping of Major QTL for Temporal Expression of the Tiller Number. Frontiers in Plant Science **9**, 333.

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE. 2012*a*. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature Genetics **44**, 217–220.

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE. 2012*b*. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proceedings of the National Academy of Sciences of the United States of America **109**, 8872–8877.

Riewe D, Jeon H-J, Lisec J, Heuermann MC, Schmeichel J, Seyfarth M, Meyer RC, Willmitzer L, Altmann T. 2016. A naturally occurring promoter polymorphism of the Arabidopsis FUM2 gene causes expression variation, and is associated with metabolic and growth traits. The Plant Journal **88**, 826–838.

Riewe D, Koohi M, Lisec J, Pfeiffer M, Lippmann R, Schmeichel J, Willmitzer L, Altmann T. 2012. A tyrosine aminotransferase involved in tocopherol synthesis in Arabidopsis. The Plant Journal **71**, 850–859.

Rincent R, Nicolas S, Bouchet S, *et al.* 2014. Dent and Flint maize diversity panels reveal important genetic potential for increasing biomass production. Theoretical and Applied Genetics. **127**, 2313–2331.

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics **16**, 85–97.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**, 139–140.

Roitsch T, Cabrera-Bosquet L, Fournier A, Ghamkhar K, Jiménez-Berni J, Pinto F, Ober ES. 2019. Review: New sensors and data-driven approaches—A path to next generation phenomics. Plant Science **282**, 2–10.

Rojas CA, Eloy NB, Lima M de F, Rodrigues RL, Franco LO, Himanen K, Beemster GTS, Hemerly AS, Ferreira PCG. 2009. Overexpression of the Arabidopsis anaphase promoting complex subunit CDC27a increases growth rate and organ size. Plant Molecular Biology **71**, 307–318.

Rosado-Souza L, Scossa F, Chaves IS, *et al.* 2015. Exploring natural variation of photosynthetic, primary metabolism and growth parameters in a large panel of *Capsicum chinense* accessions. Planta **242**, 677–691.

Saghafi D, Delangiz N, Lajayer BA, Ghorbanpour M. 2019. An overview on improvement of crop productivity in saline soils by halotolerant and halophilic PGPRs. 3 Biotech **9**, 261.

Sahni S, Prasad BD, Liu Q, Grbic V, Sharpe A, Singh SP, Krishna P. 2016. Overexpression of the brassinosteroid biosynthetic gene DWF4 in *Brassica napus* simultaneously increases seed yield and stress tolerance. Scientific Reports **6**, 28298.

Saito K, Matsuda F. 2010. Metabolomics for Functional Genomics, Systems Biology, and Biotechnology. Annual Review of Plant Biology **61**, 463–489.

Salas Fernandez MG, Bao Y, Tang L, Schnable PS. 2017. A High-Throughput, Field-Based Phenotyping Technology for Tall Biomass Crops. Plant Physiology **174**, 2008–2022.

Samans B, Chalhoub B, Snowdon RJ. 2017. Surviving a Genome Collision: Genomic Signatures of Allopolyploidization in the Recent Crop Species *Brassica napus*. The Plant Genome **10**, 1–15.

Sang M, Shi H, Wei K, Ye M, Jiang L, Sun L, Wu R. 2019. A dissection model for mapping complex traits. The Plant Journal **97**, 1168–1182.

Sattler MC, Carvalho CR, Clarindo WR. 2016. The polyploidy and its key role in plant breeding. Planta **243**, 281–296.

Scharr H, Minervini M, French AP, *et al.* 2016. Leaf segmentation in plant phenotyping: a collation study. Machine Vision and Applications **27**, 585–606.

Schauer N, Semel Y, Roessner U, *et al.* 2006. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nature Biotechnology **24**, 447–454.

Schiessl S, Huettel B, Kuehn D, Reinhardt R, Snowdon R. 2017. Post-polyploidisation morphotype diversification associates with gene copy number variation. Scientific Reports **7**, 41845.

Schilling S, Gramzow L, Lobbes D, *et al.* 2015. Non-canonical structure, function and phylogeny of the Bsister MADS-box gene OsMADS30 of rice (*Oryza sativa*). The Plant Journal **84**, 1059–1072.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of *Arabidopsis thaliana* development. Nature Genetics **37**, 501–506.

Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE. 2004. Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics **167**, 485–498.

Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE. 2018. Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. Genetics **208**, 1373–1385.

Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, Weigel D. 2005. Specific effects of microRNAs on the plant transcriptome. Developmental Cell **8**, 517–527.

Seifert F, Thiemann A, Schrag TA, Rybka D, Melchinger AE, Frisch M, Scholten S. 2018. Small RNA-based prediction of hybrid performance in maize. BMC Genomics **19**, 371.

Shah S, Weinholdt C, Jedrusik N, Molina C, Zou J, Große I, Schiessl S, Jung C, Emrani N. 2018. Whole-transcriptome analysis reveals genetic factors underlying flowering time regulation in rapeseed (*Brassica napus* L.). Plant, Cell & Environment **41**, 1935–1947.

Shahid M, Cai G, Zu F, Zhao Q, Qasim MU, Hong Y, Fan C, Zhou Y. 2019. Comparative Transcriptome Analysis of Developing Seeds and Silique Wall Reveals Dynamic Transcription Networks for Effective Oil Production in *Brassica napus* L. International Journal of Molecular Sciences **20**, 1982.

Shang L, Ma L, Wang Y, *et al.* 2016. Main Effect QTL with Dominance Determines Heterosis for Dynamic Plant Height in Upland Cotton. G3: Genes, Genomes, Genetics **6**, 3373–3379.

Shen Y, Sun S, Hua S, Shen E, Ye C-Y, Cai D, Timko MP, Zhu Q-H, Fan L. 2017. Analysis of transcriptional and epigenetic changes in hybrid vigor of allopolyploid *Brassica napus* uncovers key roles for small RNAs. The Plant Journal **91**, 874–893.

Shi J, Li R, Qiu D, Jiang C, Long Y, Morgan C, Bancroft I, Zhao J, Meng J. 2009. Unraveling the complex trait of crop yield with quantitative trait loci mapping in *Brassica napus*. Genetics **182**, 851–861.

Shi J, Li R, Zou J, Long Y, Meng J. 2011. A dynamic and complex network regulates the heterosis of yield-correlated traits in rapeseed (*Brassica napus* L.). PLOS ONE **6**, e21645.

Shi R, Melzer M, Zheng S, Benke A, Stich B, von Wirén N. 2018. Iron Retention in Root Hemicelluloses Causes Genotypic Variability in the Tolerance to Iron Deficiency-Induced Chlorosis in Maize. Frontiers in Plant Science **9**, 557.

Shin J-H, Blay S, Graham J, McNeney B. 2006. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. Journal of Statistical Software **16**.

Shull GH. 1948. What Is 'Heterosis'? Genetics **33**, 439–446.

Snowdon RJ. 2007. Cytogenetics and genome analysis in Brassica crops. Chromosome Research **15**, 85–95.

Sprague GF, Tatum LA. 1942. General vs. Specific Combining Ability in Single Crosses of Corn. Agronomy Journal **34**, 923–932.

Springer NM, Stupar RM. 2007. Allelic variation and heterosis in maize: how do two halves make more than a whole? Genome Research **17**, 264–275.

Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. 2007. pcaMethods--a bioconductor package providing PCA methods for incomplete data. Bioinformatics **23**, 1164–1167.

Stahl A, Pfeifer M, Frisch M, Wittkop B, Snowdon RJ. 2017. Recent Genetic Gains in Nitrogen Use Efficiency in Oilseed Rape. Frontiers in Plant Science **8**, 963.

Stanton RA, Pratley JE, Hudson D, Dill GM. 2010. Herbicide tolerant canola systems and their impact on winter crop rotations. Field Crops Research **117**, 161–166.

Stein A, Coriton O, Rousseau-Gueutin M, Samans B, Schiessl SV, Obermeier C, Parkin IAP, Chèvre A-M, Snowdon RJ. 2017. Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. Plant Biotechnology Journal **15**, 1478–1489.

Steinfath M, Gärtner T, Lisec J, Meyer RC, Altmann T, Willmitzer L, Selbig J. 2010. Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theoretical and Applied Genetics. **120**, 239–247.

Sulpice R, Pyl E-T, Ishihara H, *et al.* 2009. Starch as a major integrator in the regulation of plant growth. Proceedings of the National Academy of Sciences of the United States of America **106**, 10348–10353.

Sun D, Cen H, Weng H, *et al.* 2019. Using hyperspectral analysis as a potential high throughput phenotyping tool in GWAS for protein content of rice quality. Plant Methods **15**, 54.

Sun C, Wang B, Yan L, *et al.* 2016. Genome-Wide Association Study Provides Insight into the Genetic Control of Plant Height in Rapeseed (*Brassica napus* L.). Frontiers in Plant Science **7**, 1102.

Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS. 2006. All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proceedings of the National Academy of Sciences of the United States of America **103**, 6805–6810.

Szała L, Kaczmarek Z, Popławska W, Liersch A, Wójtowicz M, Matuszczak M, Biliński ZR, Sosnowska K, Stefanowicz M, Cegielska-Taras T. 2019. Estimation of seed yield in oilseed rape to identify the potential of semi-resynthesized parents for the development of new hybrid cultivars. PLOS ONE **14**, e0215661.

Takahashi S, Osabe K, Fukushima N, *et al.* 2018. Genome-wide characterization of DNA methylation, small RNA expression, and histone H3 lysine nine di-methylation in *Brassica rapa* L. DNA Research **25**, 511–520.

Tang M-Q, Cheng X-H, Tong C-B, Liu Y-Y, Zhao C-J, Dong C-H, Yu J-Y, Ma X-G, Huang J-Y, Liu S-Y. 2015. Genome-wide Association Analysis of Plant Height in Rapeseed (*Brassica napus*). Acta Agronomica Sinica **41**, 1121–1126.

Tang Y, Liu X, Wang J, *et al.* 2016. GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. The Plant Genome **9**.

Tanger P, Klassen S, Mojica JP, *et al.* 2017. Field-based high throughput phenotyping rapidly identifies genomic regions controlling yield components in rice. Scientific Reports **7**, 42839.

Thompson KF. 1972. Cytoplasmic male-sterility in oil-seed rape. Heredity **29**, 253–257.

Tian C, Wang Y, Yu H, He J, Wang J, Shi B, Du Q, Provart NJ, Meyerowitz EM, Jiao Y. 2019. A gene expression map of shoot domains reveals regulatory mechanisms. Nature Communications **10**, 141.

Tisné S, Serrand Y, Bach L, *et al.* 2013. Phenoscope: an automated large-scale phenotyping platform offering high spatial homogeneity. The Plant Journal **74**, 534–544.

Tschiersch H, Junker A, Meyer RC, Altmann T. 2017. Establishment of integrated protocols for automated high throughput kinetic chlorophyll fluorescence analyses. Plant Methods **13**, 54.

Ubbens JR, Stavness I. 2017. Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. Frontiers in Plant Science **8**, 1190.

Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GTS, Gruissem W, Van de Peer Y, Inzé D, De Veylder L. 2005. Genome-wide identification of potential plant E2F target genes. Plant Physiology **139**, 316–328.

Vanhaeren H, Gonzalez N, Coppens F, De Milde L, Van Daele T, Vermeersch M, Eloy NB, Storme V, Inzé D. 2014. Combining growth-promoting genes leads to positive epistasis in *Arabidopsis thaliana*. eLife **3**, e02252.

Vanhaeren H, Gonzalez N, Inzé D. 2015. A Journey Through a Leaf: Phenomics Analysis of Leaf Growth in *Arabidopsis thaliana*. The Arabidopsis Book **13**, e0181.

Vanhaeren H, Inzé D, Gonzalez N. 2016. Plant Growth Beyond Limits. Trends in Plant Science **21**, 102–109.

Vanhaeren H, Nam Y-J, De Milde L, Chae E, Storme V, Weigel D, Gonzalez N, Inzé D. 2017. Forever Young: The Role of Ubiquitin Receptor DA1 and E3 Ligase BIG BROTHER in Controlling Leaf Growth and Development. Plant Physiology **173**, 1269–1282.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science **91**, 4414–4423.

Vanstraelen M, Benková E. 2012. Hormonal interactions in the regulation of plant development. Annual Review of Cell and Developmental Biology **28**, 463–487.

Vasseur F, Fouqueau L, de Vienne D, Nidelet T, Violle C, Weigel D. 2019. Nonlinear phenotypic variation uncovers the emergence of heterosis in *Arabidopsis thaliana*. PLOS Biology **17**, e3000214.

Verkest A, Weinl C, Inzé D, De Veylder L, Schnittger A. 2005. Switching the cell cycle. Kip-related proteins in plant cell cycle control. Plant Physiology **139**, 1099–1106.

de Vienne D, Bost B, Fiévet J, Zivy M, Dillmann C. 2001. Genetic variability of proteome expression and metabolic control. Plant Physiology and Biochemistry **39**, 271–283.

de Vlaming R, Groenen PJF. 2015. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. BioMed Research International **2015**, 143712.

Voorend W, Nelissen H, Vanholme R, De Vliegher A, Van Breusegem F, Boerjan W, Roldán-Ruiz I, Muylle H, Inzé D. 2016. Overexpression of GA20-OXIDASE1 impacts plant height, biomass allocation and saccharification efficiency in maize. Plant Biotechnology Journal **14**, 997–1007.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences **131**, 281–285.

Wahl V, Brand LH, Guo Y-L, Schmid M. 2010. The FANTASTIC FOUR proteins influence shoot meristem size in *Arabidopsis thaliana*. BMC Plant Biology **10**, 285.

Walter A, Scharr H, Gilmer F, *et al.* 2007. Dynamics of seedling growth acclimation towards altered light conditions can be quantified via GROWSCREEN: a setup and procedure designed for rapid optical phenotyping of different plant species. New Phytologist **174**, 447–455.

Wan H, Cui Y, Ding Y, *et al.* 2017. Time-Series Analyses of Transcriptomes and Proteomes Reveal Molecular Networks Underlying Oil Accumulation in Canola. Frontiers in Plant Science **7**, 2007.

Wang N, Chen B, Xu K, Gao G, Li F, Qiao J, Yan G, Li J, Li H, Wu X. 2016. Association Mapping of Flowering Time QTLs and Insight into Their Contributions to Rapeseed Growth Habits. Frontiers in Plant Science **7**, 338.

Wang Y, Henriksson E, Söderman E, Henriksson KN, Sundberg E, Engström P. 2003. The Arabidopsis homeobox gene, ATHB16, regulates leaf development and the sensitivity to photoperiod in Arabidopsis. Developmental Biology **264**, 228–239.

Wang Y-Y, Li Y-Q, Wu H-Y, *et al.* 2018*a*. Genotyping of Soybean Cultivars With Medium-Density Array Reveals the Population Structure and QTNs Underlying Maturity and Seed Traits. Frontiers in Plant Science **9**, 610.

Wang N, Qian W, Suppanz I, Wei L, Mao B, Long Y, Meng J, Müller AE, Jung C. 2011*a*. Flowering time variation in oilseed rape (*Brassica napus* L.) is associated with allelic variation in the FRIGIDA homologue BnaA.FRI.a. Journal of Experimental Botany **62**, 5641–5658.

Wang D, Salah El-Basyoni I, Stephen Baenziger P, Crossa J, Eskridge KM, Dweikat I. 2012. Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. Heredity **109**, 313–319.

Wang X, Wang H, Long Y, *et al.* 2015. Dynamic and comparative QTL analysis for plant height in different developmental stages of *Brassica napus* L. Theoretical and Applied Genetics. **128**, 1175–1192.

Wang N, Wang Y, Tian F, King GJ, Zhang C, Long Y, Shi L, Meng J. 2008. A functional genomics resource for *Brassica napus*: development of an EMS mutagenized population and discovery of FAE1 point mutations by TILLING. New Phytologist **180**, 751–765.

Wang X, Wang H, Wang J, *et al.* 2011*b*. The genome of the mesopolyploid crop species *Brassica rapa*. Nature Genetics **43**, 1035–1039.

Wang S, Wei J, Li R, Qu H, Chater JM, Ma R, Li Y, Xie W, Jia Z. 2019. Identification of optimal prediction models using multi-omic data for selecting hybrid rice. Heredity **123**, 395–406.

Wang J, Zhou Z, Zhang Z, Li H, Liu D, Zhang Q, Bradbury PJ, Buckler ES, Zhang Z. 2018*b*. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. Heredity **121**, 648–662.

Ward BP, Brown-Guedira G, Kolb FL, Van Sanford DA, Tyagi P, Sneller CH, Griffey CA. 2019. Genome-wide association studies for yield-related traits in soft red winter wheat grown in Virginia. PLOS ONE **14**, e0208217.

Wei C, Wang H, Heng S, Wen J, Yi B, Ma C, Tu J, Shen J, Fu T. 2019. Construction of restorer lines and molecular mapping for restorer gene of hau cytoplasmic male sterility in *Brassica napus*. Theoretical and Applied Genetics. **132**, 2525–2539.

Wen W, Li K, Alseekh S, *et al.* 2015. Genetic Determinants of the Network of Primary Metabolism and Their Relationships to Plant Performance in a Maize Recombinant Inbred Line Population. The Plant Cell **27**, 1839–1856.

Werner CR, Qian L, Voss-Fels KP, Abbadi A, Leckband G, Frisch M, Snowdon RJ. 2017. Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. Theoretical and Applied Genetics. **131**, 299–317.

Werner CR, Voss-Fels KP, Miller CN, Qian W, Hua W, Guan C-Y, Snowdon RJ, Qian L. 2018. Effective Genomic Selection in a Narrow-Genepool Crop with Low-Density Markers: Asian Rapeseed as an Example. The Plant Genome **11**, 170084.

Westhues M, Schrag TA, Heuer C, *et al.* 2017. Omics-based hybrid prediction in maize. Theoretical and Applied Genetics. **130**, 1927–1939.

Whittaker JC, Thompson R, Denham MC. 2000. Marker-assisted selection using ridge regression. Genetical Research **75**, 249–252.

Windhausen VS, Atlin GN, Hickey JM, *et al.* 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3: Genes, Genomes, Genetics **2**, 1427–1436.

Wright S. 1934. Physiological and Evolutionary Theories of Dominance. The American Naturalist **68**, 24–53.

Wu J, Lin L, Xu M, Chen P, Liu D, Sun Q, Ran L, Wang Y. 2018. Homoeolog expression bias and expression level dominance in resynthesized allopolyploid *Brassica napus*. BMC Genomics **19**, 586.

Wu G, Poethig RS. 2006. Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3. Development **133**, 3539–3547.

Wu Z, Wang B, Chen X, Wu J, King GJ, Xiao Y, Liu K. 2016*a*. Evaluation of Linkage Disequilibrium Pattern and Association Study on Seed Oil Content in *Brassica napus* Using ddRAD Sequencing. PLOS ONE **11**, e0146383.

Wu J, Zhao Q, Liu S, Shahid M, Lan L, Cai G, Zhang C, Fan C, Wang Y, Zhou Y. 2016*b*. Genome-wide Association Study Identifies New Loci for Resistance to Sclerotinia Stem Rot in *Brassica napus*. Frontiers in Plant Science **7**, 1418.

Würschum T, Liu W, Alheit KV, Tucker MR, Gowda M, Weissmann EA, Hahn V, Maurer HP. 2014*a*. Adult plant development in triticale (× *triticosecale wittmack*) is controlled by dynamic genetic patterns of regulation. G3: Genes, Genomes, Genetics **4**, 1585–1591.

Würschum T, Liu W, Busemeyer L, Tucker MR, Reif JC, Weissmann EA, Hahn V, Ruckelshausen A, Maurer HP. 2014*b*. Mapping dynamic QTL for plant height in triticale. BMC Genetics **15**, 59.

Würschum T, Maurer HP, Dreyer F, Reif JC. 2013. Effect of inter- and intragenic epistasis on the heritability of oil content in rapeseed (*Brassica napus* L.). Theoretical and Applied Genetics **126**, 435–441.

Xie Q, Frugis G, Colgan D, Chua NH. 2000. Arabidopsis NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. Genes & Development **14**, 3024–3036.

Xu S. 2003. Theoretical basis of the Beavis effect. Genetics **165**, 2259–2268.

Xu S, Xu Y, Gong L, Zhang Q. 2016. Metabolomic prediction of yield in hybrid rice. The Plant Journal **88**, 219–227.

Yang W, Duan L, Chen G, Xiong L, Liu Q. 2013. Plant phenomics and high-throughput phenotyping: accelerating rice functional genomics using multidisciplinary technologies. Current Opinion in Plant Biology **16**, 180–187.

Yang W, Guo Z, Huang C, *et al.* 2014. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. Nature Communications **5**, 5087.

Yang P, Shu C, Chen L, Xu J, Wu J, Liu K. 2012. Identification of a major QTL for silique length and seed weight in oilseed rape (*Brassica napus* L.). Theoretical and Applied Genetics **125**, 285–296.

Yang W, Tempelman RJ. 2012. A Bayesian antedependence model for whole genome prediction. Genetics **190**, 1491–1501.

Yang L, Wen K-S, Ruan X, Zhao Y-X, Wei F, Wang Q. 2018. Response of Plant Secondary Metabolites to Environmental Factors. Molecules **23**, 762.

Yi Chou E, Schuetz M, Hoffmann N, Watanabe Y, Sibout R, Samuels AL. 2018. Distribution, mobility, and anchoring of lignin-related oxidative enzymes in Arabidopsis secondary cell walls. Journal of Experimental Botany **69**, 1849–1859.

Yong H-Y, Wang C, Bancroft I, Li F, Wu X, Kitashiba H, Nishio T. 2015. Identification of a gene controlling variation in the salt tolerance of rapeseed (*Brassica napus* L.). Planta **242**, 313–326.

You Q, Yang X, Peng Z, Xu L, Wang J. 2018. Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array. Frontiers in Plant Science **9**, 104.

Younas M, Xiao Y, Cai D, Yang W, Ye W, Wu J, Liu K. 2012. Molecular characterization of oilseed rape accessions collected from multi continents for exploitation of potential heterotic group through SSR markers. Molecular Biology Reports **39**, 5105–5113.

Yu K, Wang X, Chen F, *et al.* 2018*a*. Quantitative Trait Transcripts Mapping Coupled with Expression Quantitative Trait Loci Mapping Reveal the Molecular Network Regulating the Apetalous Characteristic in *Brassica napus* L. Frontiers in Plant Science **9**, 89.

Yu L-H, Wu J, Tang H, Yuan Y, Wang S-M, Wang Y-P, Zhu Q-S, Li S-G, Xiang C-B. 2016. Overexpression of Arabidopsis NLP7 improves plant growth under both nitrogen-limiting and -sufficient conditions by enhancing nitrogen and carbon assimilation. Scientific Reports **6**, 27795.

Yu J, Yang X-D, Wang Q, Gao L-W, Yang Y, Xiao D, Liu T-K, Li Y, Hou X-L, Zhang C-W. 2018*b*. Efficient virus-induced gene silencing in *Brassica rapa* using a turnip yellow mosaic virus vector. Biologia Plantarum **62**, 826–834.

Zampieri M, Sekar K, Zamboni N, Sauer U. 2017. Frontiers of high-throughput metabolomics. Current Opinion in Chemical Biology **36**, 15–23.

Zeng X, Zhu L, Chen Y, *et al.* 2011. Identification, fine mapping and characterisation of a dwarf mutant (bnaC.dwf) in *Brassica napus*. Theoretical and Applied Genetics **122**, 421–428.

Zenke-Philippi C, Thiemann A, Seifert F, Schrag T, Melchinger AE, Scholten S, Frisch M. 2016. Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles. BMC Genomics **17**, 262.

Zhai Y, Cai S, Hu L, Yang Y, Amoo O, Fan C, Zhou Y. 2019. CRISPR/Cas9-mediated genome editing reveals differences in the contribution of INDEHISCENT homologues to pod shatter resistance in *Brassica napus* L. Theoretical and Applied Genetics. **132**, 2111–2123.

Zhang H, Berger JD, Herrmann C. 2017*a*. Yield stability and adaptability of canola (*Brassica napus* L.) in multiple environment trials. Euphytica **213**, 155.

Zhang T, Hu Y, Jiang W, *et al.* 2015. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nature Biotechnology **33**, 531–537.

Zhang F, Huang J, Tang M, *et al.* 2019*a*. Syntenic quantitative trait loci and genomic divergence for Sclerotinia resistance and flowering time in *Brassica napus*. Journal of Integrative Plant Biology **61**, 75–88.

Zhang X, Huang C, Wu D, *et al.* 2017*b*. High-Throughput Phenotyping and QTL Mapping Reveals the Genetic Architecture of Maize Plant Growth. Plant Physiology **173**, 1554–1564.

Zhang K, Nie L, Cheng Q, *et al.* 2019*b*. Effective editing for lysophosphatidic acid acyltransferase 2/5 in allotetraploid rapeseed (*Brassica napus* L.) using CRISPR-Cas9 system. Biotechnology for Biofuels **12**, 225.

Zhang A, Sun H, Wang P, Han Y, Wang X. 2012. Modern analytical techniques in metabolomics analysis. The Analyst **137**, 293–300.

Zhang Y, Zhang N. 2018. Imaging technologies for plant high-throughput phenotyping: a review. Frontiers of Agricultural Science and Engineering **5**, 406–419.

Zhao J, Becker HC, Zhang D, Zhang Y, Ecke W. 2006. Conditional QTL mapping of oil content in rapeseed with respect to protein content and traits related to plant development and grain yield. Theoretical and Applied Genetics. **113**, 33–38.

Zhao Y, Li Z, Liu G, *et al.* 2015. Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proceedings of the National Academy of Sciences of the United States of America **112**, 15624–15629.

Zhao W, Wang X, Wang H, *et al.* 2016. Genome-Wide Identification of QTL for Seed Yield and Yield-Related Traits and Construction of a High-Density Consensus Map for QTL Comparison in *Brassica napus*. Frontiers in Plant Science **7**, 17.

Zhao Q, Wu J, Cai G, Yang Q, Shahid M, Fan C, Zhang C, Zhou Y. 2019*a*. A novel quantitative trait locus on chromosome A9 controlling oleic acid content in *Brassica napus*. Plant Biotechnology Journal, doi.org/10.1111/pbi.13142.

Zhao C, Zhang Y, Du J, Guo X, Wen W, Gu S, Wang J, Fan J. 2019*b*. Crop Phenomics: Current Status and Perspectives. Frontiers in Plant Science **10**, 714.

Zheng Z-L. 2009. Carbon and nitrogen nutrient balance signaling in plants. Plant Signaling & Behavior **4**, 584–591.

Zheng M, Peng C, Liu H, *et al.* 2017. Genome-Wide Association Study Reveals Candidate Genes for Control of Plant Height, Branch Initiation Height and Branch Number in Rapeseed (*Brassica napus* L.). Frontiers in Plant Science **8**, 1246.

Zheng M, Zhang L, Tang M, Liu J, Liu H, Yang H, Fan S, Terzaghi W, Wang H, Hua W. 2019. Knockout of two BnaMAX1 homologs by CRISPR/Cas9-targeted mutagenesis improves plant architecture and increases yield in rapeseed (*Brassica napus* L.). Plant Biotechnology Journal, doi.org/10.1111/pbi.13228.

Zhou Q, Han D, Mason AS, Zhou C, Zheng W, Li Y, Wu C, Fu D, Huang Y. 2017*a*. Earliness traits in rapeseed (*Brassica napus*): SNP loci and candidate genes identified by genome-wide association analysis. DNA Research **25**, 229–244.

Zhou Q, Zhou C, Zheng W, Mason AS, Fan S, Wu C, Fu D, Huang Y. 2017*b*. Genome-Wide SNP Markers Based on SLAF-Seq Uncover Breeding Traces in Rapeseed (*Brassica napus* L.). Frontiers in Plant Science **8**, 648.

Zhu A, Greaves IK, Liu P-C, Wu L, Dennis ES, Peacock WJ. 2016. Early changes of gene activity in developing seedlings of Arabidopsis hybrids relative to parents may contribute to hybrid vigour. The Plant Journal **88**, 597–607.

Zhu J-Y, Sae-Seaw J, Wang Z-Y. 2013. Brassinosteroid signalling. Development **140**, 1615–1620.

Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE, Schadt EE. 2012. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. PLOS Biology **10**, e1001301.

Zou J, Hu D, Mason AS, *et al.* 2018. Genetic changes in a novel breeding population of *Brassica napus* synthesized from hundreds of crosses between *B. rapa* and *B. carinata*. Plant Biotechnology Journal **16**, 507–519.

8. Supplementary data

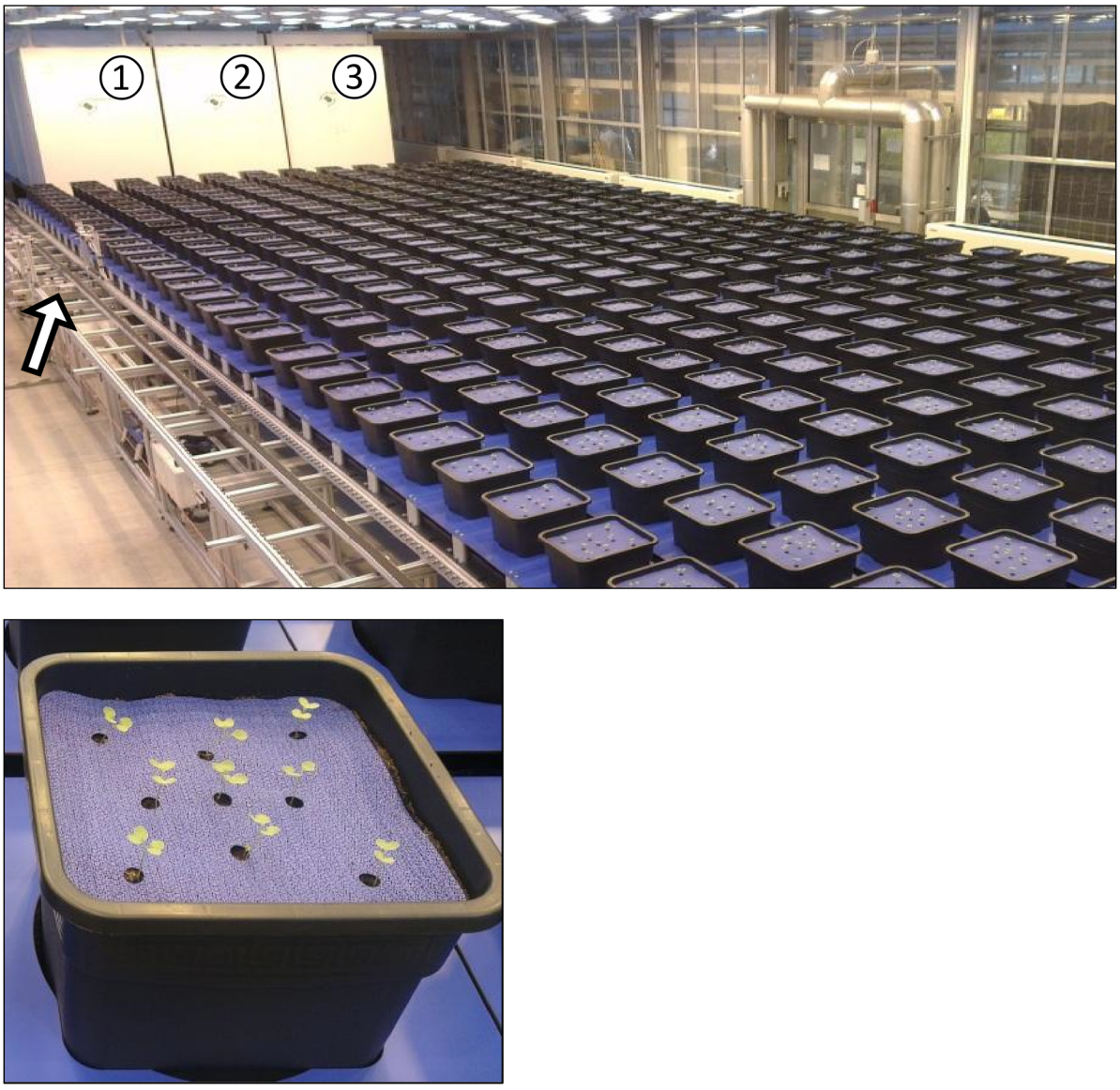


Figure S1. The IPK phenotyping platform for large plants

The upper photo shows the view across the IPK phenotyping facility for large plants with 396 mobile carriers on a conveyor belt system, an automated watering station and balance (arrow), and three imaging chambers. The chambers harbour camera systems for top and side views for static fluorescence (FLUO, ①), visible light (VIS, ②) and near infrared (NIR, ③). The system is located inside a glasshouse with illumination and semi-controlled climate conditioning. The lower photo shows one carrier (replicate) with a 25 l square pot and initially nine plants per genotype. The pot surface was covered by a blue mat for subsequent image background correction.

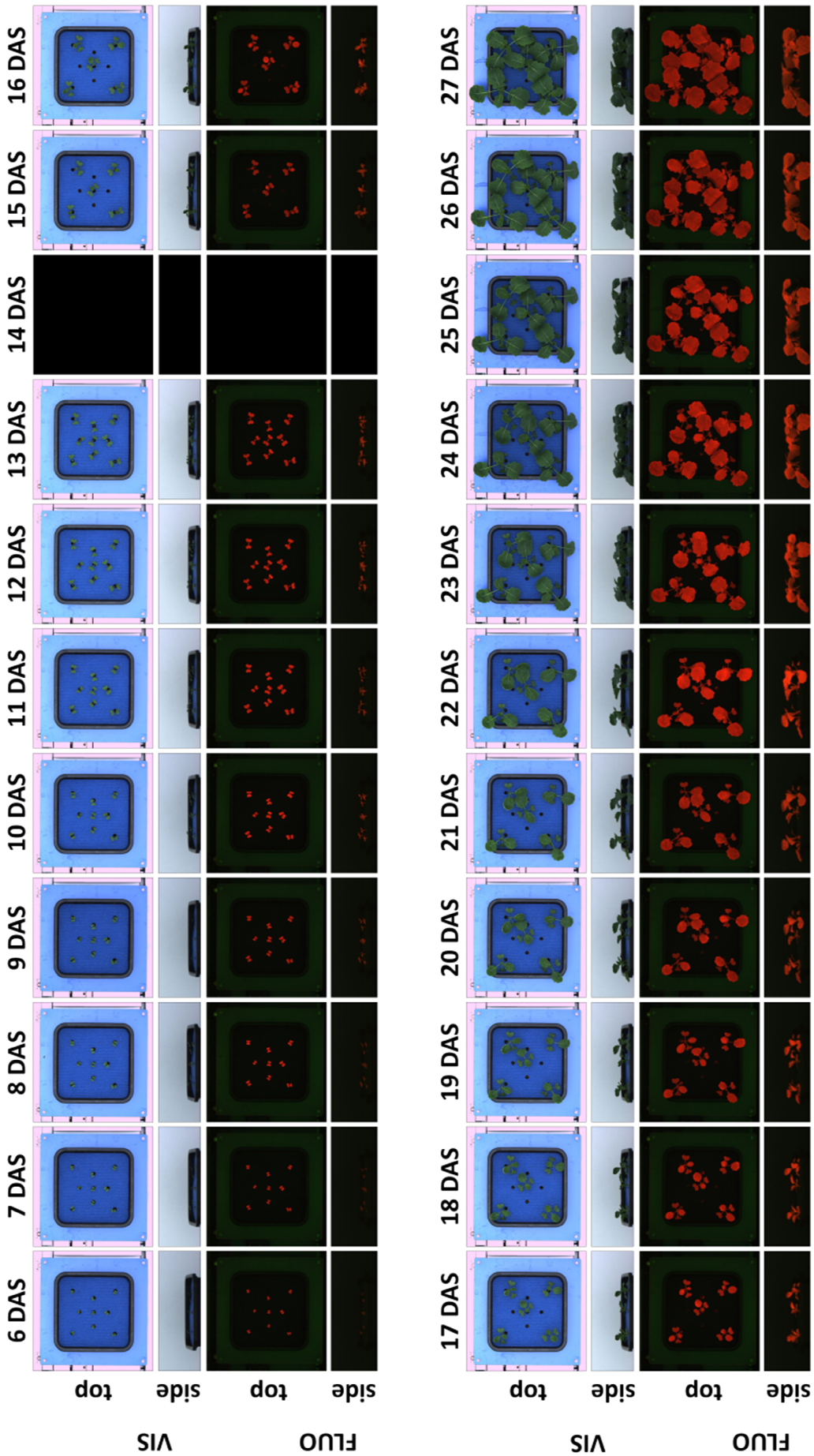


Figure S2. Example of acquired raw image data

Carrier 1087 (Experiment III, Pollinator 083) is shown as an example of the acquired raw imaging data. Plants were phenotyped daily with two camera systems: Visible light (VIS) and static fluorescence (FLUO) image data were recorded as top and side view from 6 to 27 DAS. At 14 DAS no imaging was performed due to sampling of early shoot material for molecular and biochemical analyses.

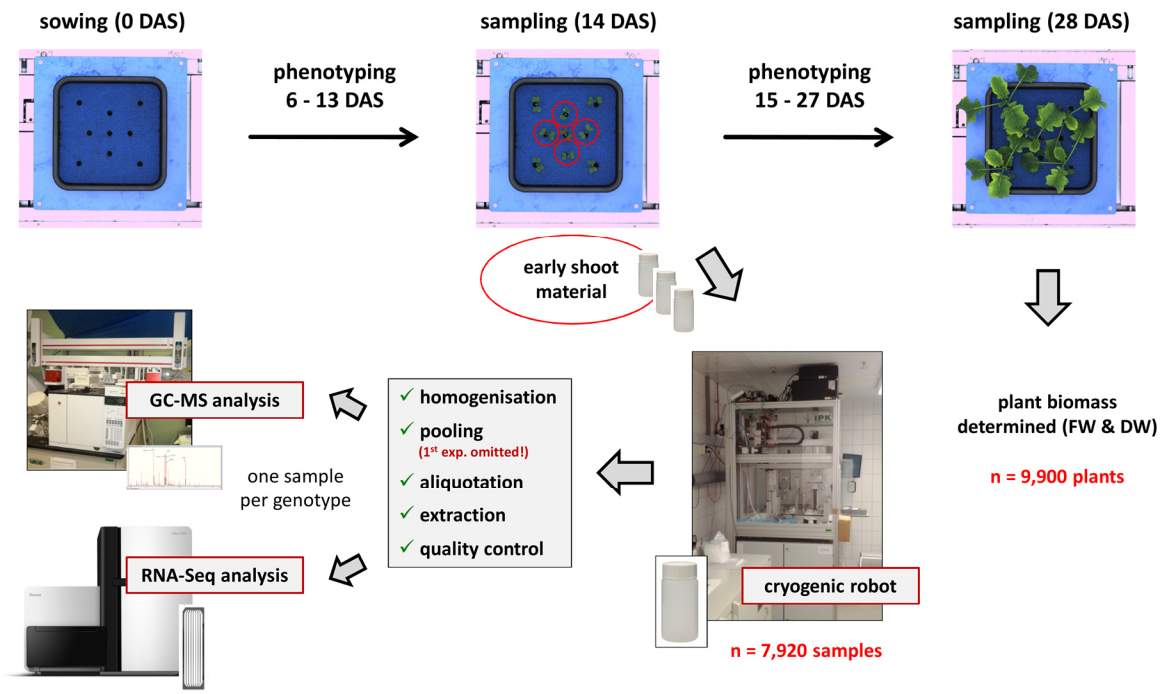


Figure S3. Flow chart of sampling and sample post-processing

The flowchart illustrates the sample collection and processing. Plants were sown and grown until 28 DAS in the phenotyping facility. The four inner plants around the central plant were sampled at 14 DAS, immediately quenched in liquid nitrogen and stored at -80 °C (7,920 individual plant samples). The remaining five plants per pot were grown until 28 DAS and sampled to analyse plant biomass (fresh and dry weight; 9,900 individual plants). The earlier sampled plants were homogenised using a cryogenic grinding robot and the four plants per genotype / pot were pooled. Subsequently, equal amounts of material from the different phenotyping experiments were pooled and mixed. The first phenotyping experiment (1413RCM) was omitted due to technical issues and higher temperatures during the early growth phase compared to the other experiments. Three 15 mg (± max. 1.5 mg) and two 50 mg (± max. 1.5 mg) aliquots of homogenised plant material were generated using the robot and manual weighing, respectively. The 15 mg aliquots were subjected to polar metabolite extraction and GC-MS analysis, while the 50 mg aliquots were used to extract total RNA for whole transcriptome profiling using RNA-sequencing.

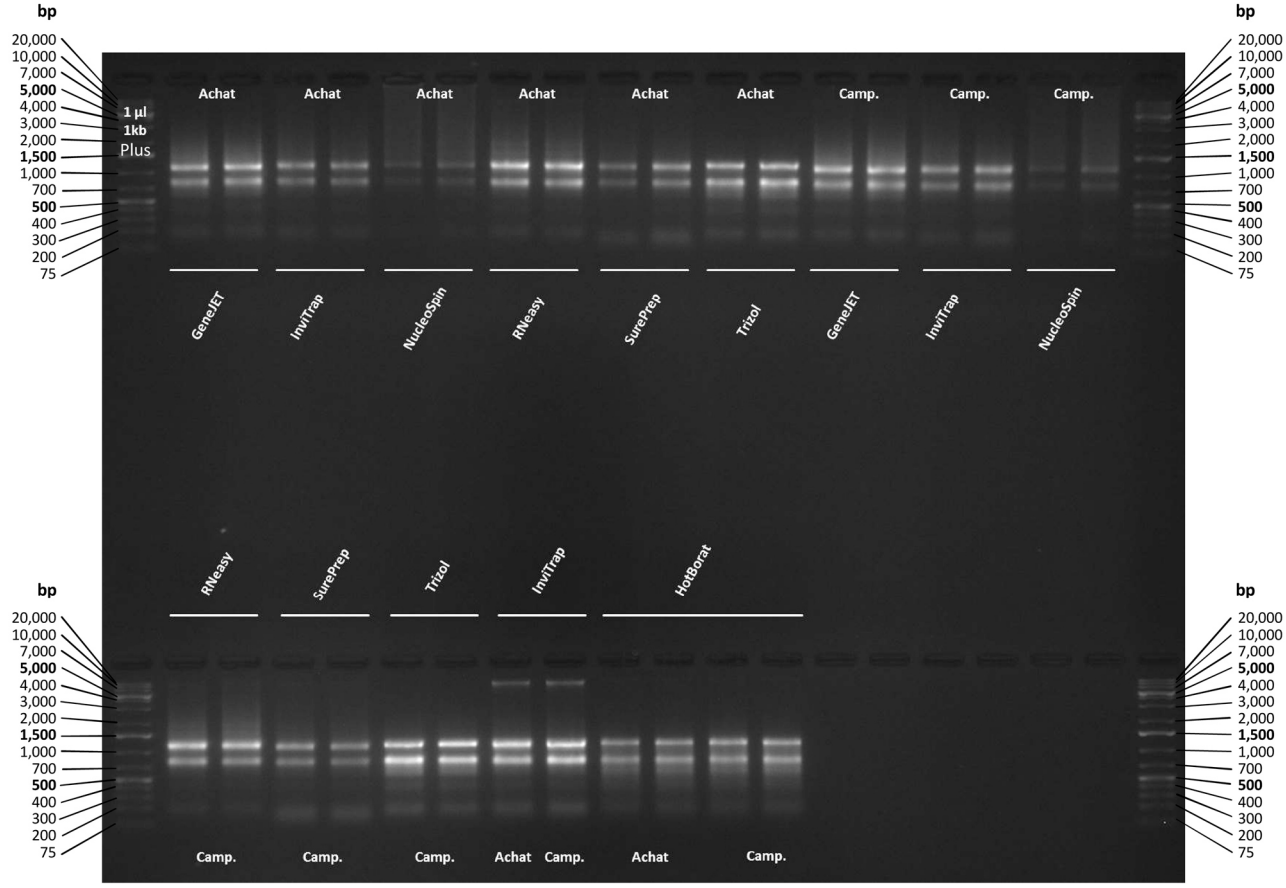


Figure S4. RNA-agarose gel (1.5 %) in Tris-Acetat-EDTA (TAE) buffer

Shown is a 1.5 % agarose electrophoresis gel to analyse the amount and purity of total RNA extracted by different commercial kits and methods from *Brassica napus* shoot material sampled at 14 DAS. Nucleic acids were separated in in Tris-Acetat-EDTA (TAE) buffer, 100 V, 300 mA, 50 W. The gel was stained with ethidium bromide (0.5 µg/ml gel) and the nucleotides visualised under ultraviolet (UV) light. Two genotypes, Achat and Campino (Camp.), were used for these tests. 1 µl of the 1 kb DNA ladder were loaded to the gel on the outermost right and left slots. The used extraction procedures were: the GeneJET RNA Purification Kit, the InviTrap Spin Plant RNA Mini Kit, NucleoSpin RNA Plant Kit, the RNeasy Plant Mini Kit, and the SurePrep Plant/Fungi Total RNA Purification Kit) and two extraction protocols: the TRIzol method and the Hot Borat method, as indicated above the loading slots. Intact total RNA in overall good quality and without visible degradation could be extracted by all methods. The two larger bands correspond to the 28S and 18S rRNAs. Genomic DNA contaminations are only visible for the RNA extracted with the InviTrap Spin Plant RNA Mini Kit (bottom row, 8-9 slots from the left).

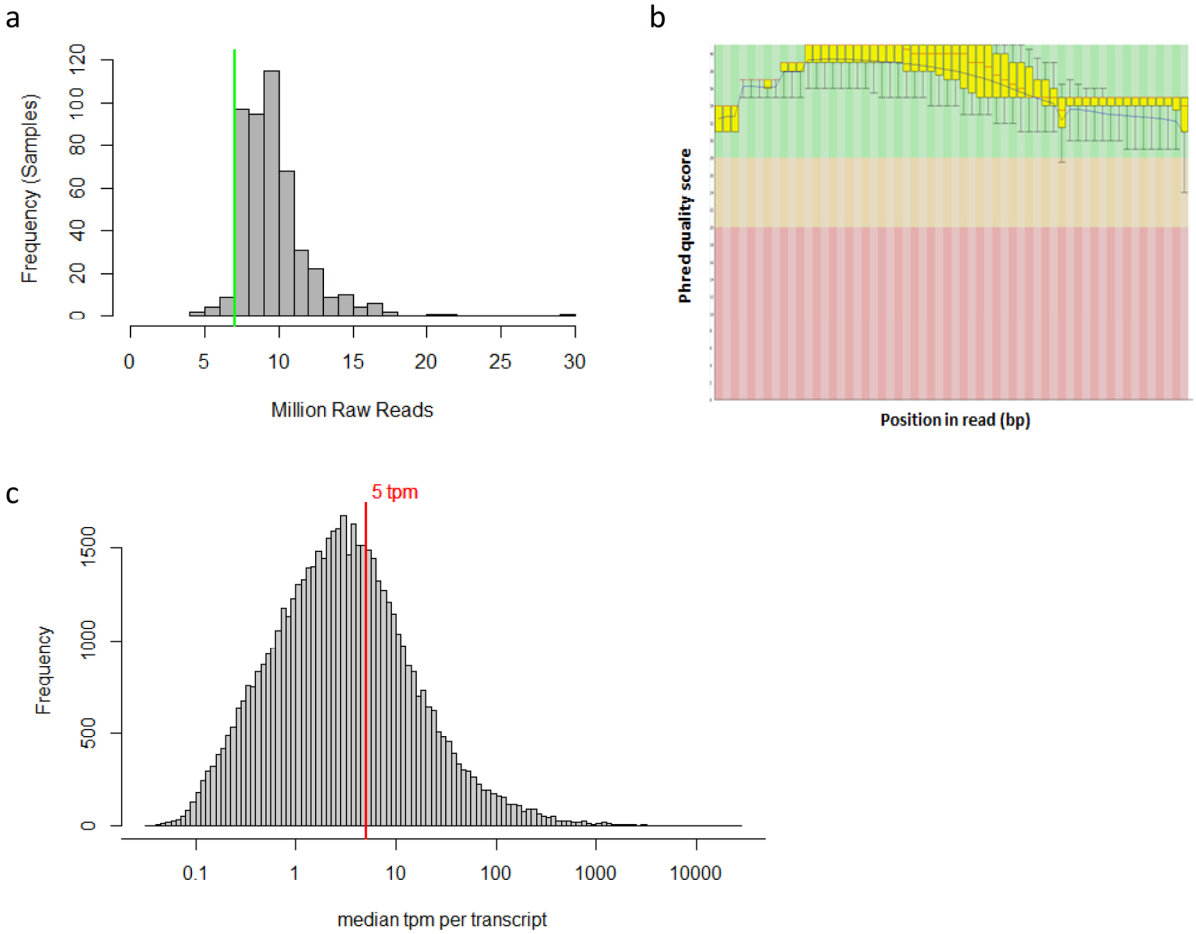


Figure S5. Overview of transcriptome data and quality

Subfigure **a** shows the distribution of the average number of reads generated for each sample (genotype). The y-axis displays the frequency (number of samples) and the x-axis the average number of reads in million reads. The green vertical line indicates the threshold of 7 million reads per sample anticipated for the experiment. Subfigure **b** displays the sequencing read quality of a representative sample extracted with the Lexogen SENSE mRNA-Seq Library Prep Kit V2 kit. The Phred quality score (0 to 40) is indicated on the y-axis, and the base position in the read on the x-axis (left to right: 1 to 108 bp). The green, orange and red colour correspond to Phred quality score ranges of 28-40 (high), 20-28 (medium) and 0-20 (low), respectively. The plot was generated using fastQC tool. Subfigure **c** shows the distribution of the median transcripts per million (tpm) values over all 477 genotypes with the x-axis on a logarithmic scale. The vertical red line represents the applied filtering threshold of 5 tpm. All transcripts expressed ≥ 5 tpm (on the right side of the red line) were used for further analyses.

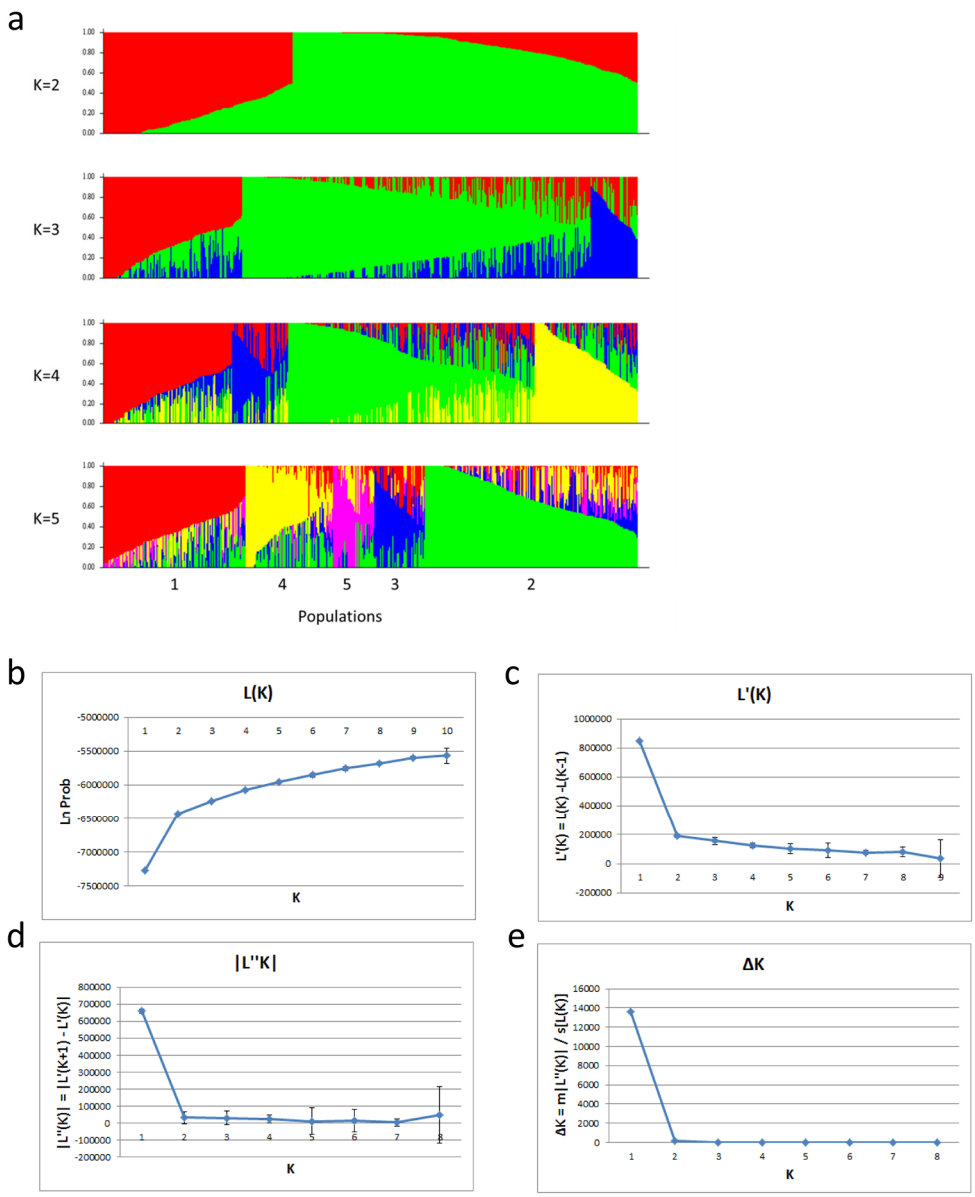


Figure S6. Population structure analysis

Population structure for all 477 *Brassica napus* genotypes subjected to GWAS was analysed using the programme STRUCTURE, version 2.3.4 (Pritchard *et al.*, 2000). Population clustering for *K*= 1 to 10 was performed using the 'admixture' model with a burn-in period of 10.000, 10.000 MCMC replications and three iterations per K. The lambda parameter was set to λ= 0.304. Subfigure **a** shows plots for *K*= 2 to 5. Genotypes were sorted by their ancestry vector (Q). Each genotype is represented by a thin vertical line. Each colour represents a population, and the colour of individual genotypes represents their proportional membership in the different populations. Populations: red, pop1; green, pop2; blue, pop3; yellow, pop4; pink, pop5. In the subfigures **b-e** statistics used to select K are shown as described by Evanno *et al.* (2005): **b** mean Ln probability L(K); **c** mean difference between successive likelihood values of K, L'(K) and standard deviations for K= 1 to 10; **d** absolute value of the difference between successive values of L'(K), |L''(K)| and standard deviations; **e** ΔK as the mean of the absolute values of L''(K) averaged over the three runs divided by the standard deviation of L(K).

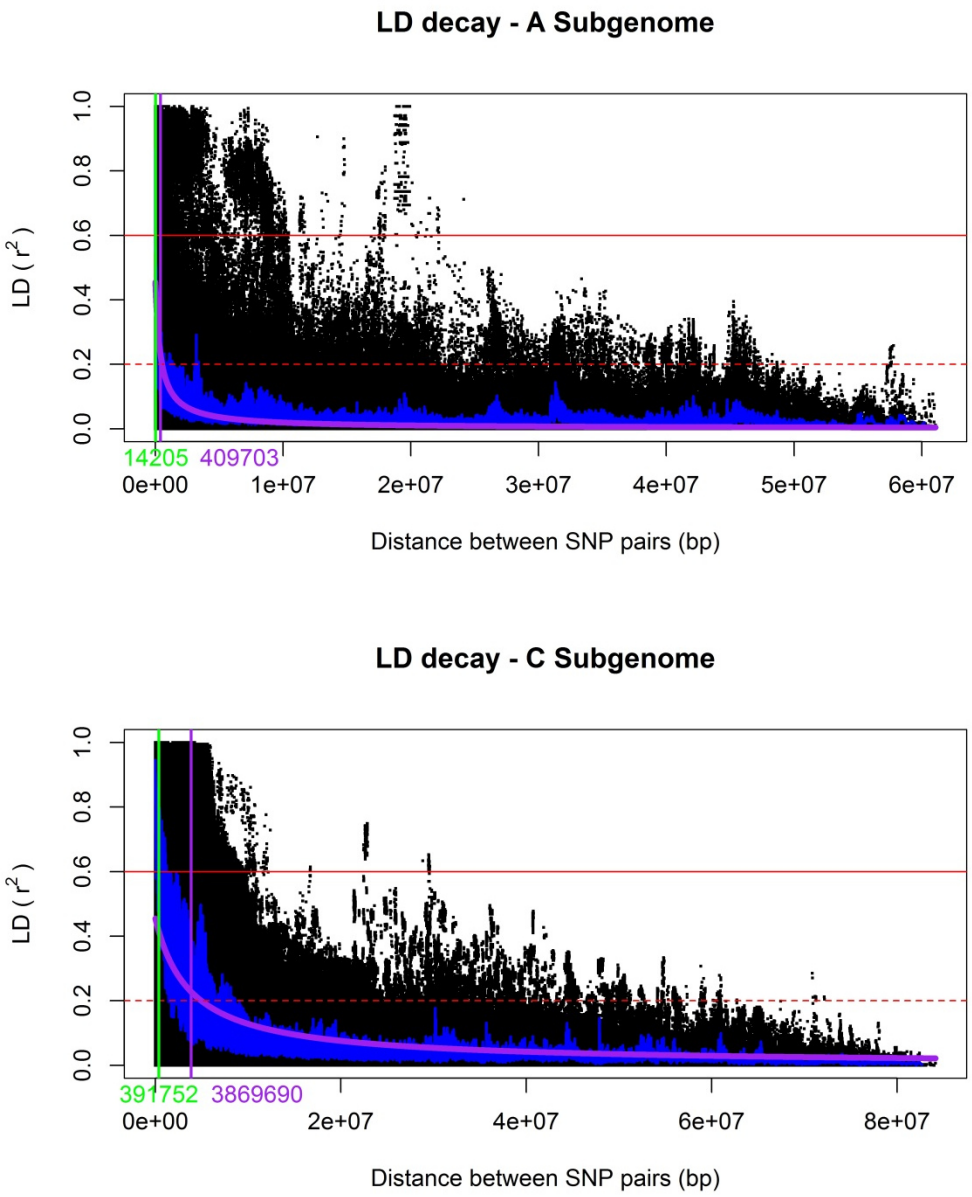


Figure S7. LD-decay in the A and C subgenomes

Pairwise marker linkage disequilibrium (LD) was calculated as r^2 values from the SNP data and plotted against the physical marker distances on the *B. napus* A and C subgenomes, separately. The blue line represents a rolling mean of LD of 100 markers. The purple line shows the LD decay calculated according to Hill and Weir (1988) using a non-linear model. The horizontal solid and dashed lines correspond to an LD of 0.6 and 0.2, respectively. The vertical green line indicates the distance where the rolling mean drops below a LD value of 0.6, and the vertical purple line the half-decay, the distance at which half of the maximum (short range) LD has decayed, respectively. Both the half-decay position and the position at LD 0.6 were substantially higher in the C subgenome than in the A subgenome.

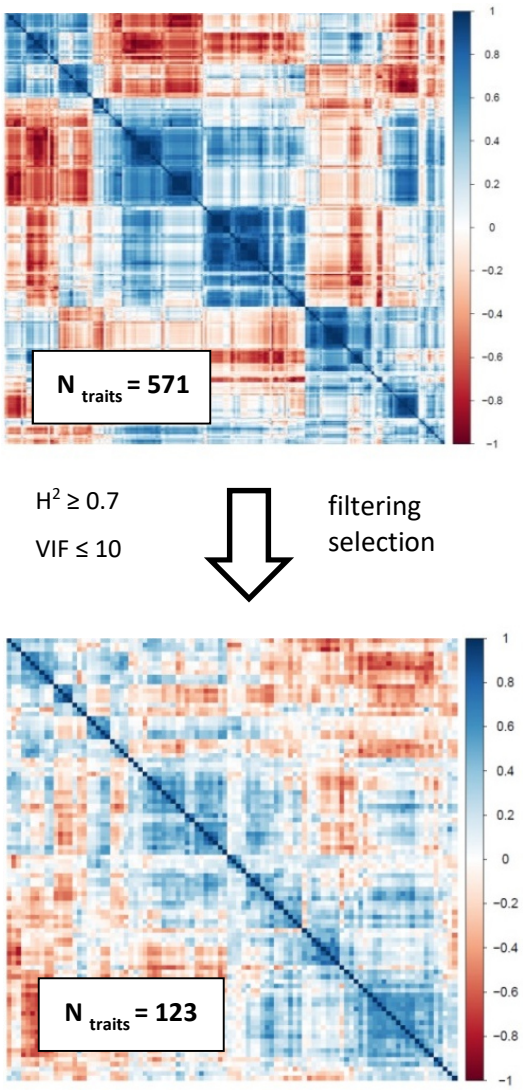


Figure S8. Trait selection and reduction of multi-collinearity

The figure displays the results of the phenotypic trait selection. The upper part of the figure shows the correlogram of the 571 image-derived traits after filtering for broad-sense heritabilities higher than 0.7 for at least one day. The colour scale on the right side corresponds to the Pearson correlation coefficients ranging from dark red (-1) over white (0) to dark blue (1). Several blocks of highly intercorrelated traits are visible. The lower part of the figure shows the correlogram after the stepwise variable selection using variance inflation factors (until VIF ≤ 10) was applied. The number of phenotypic traits was reduced to 123 traits. Traits from highly correlated blocks of traits were removed by the approach, indicated by the brighter colours and the less pronounced blocks.

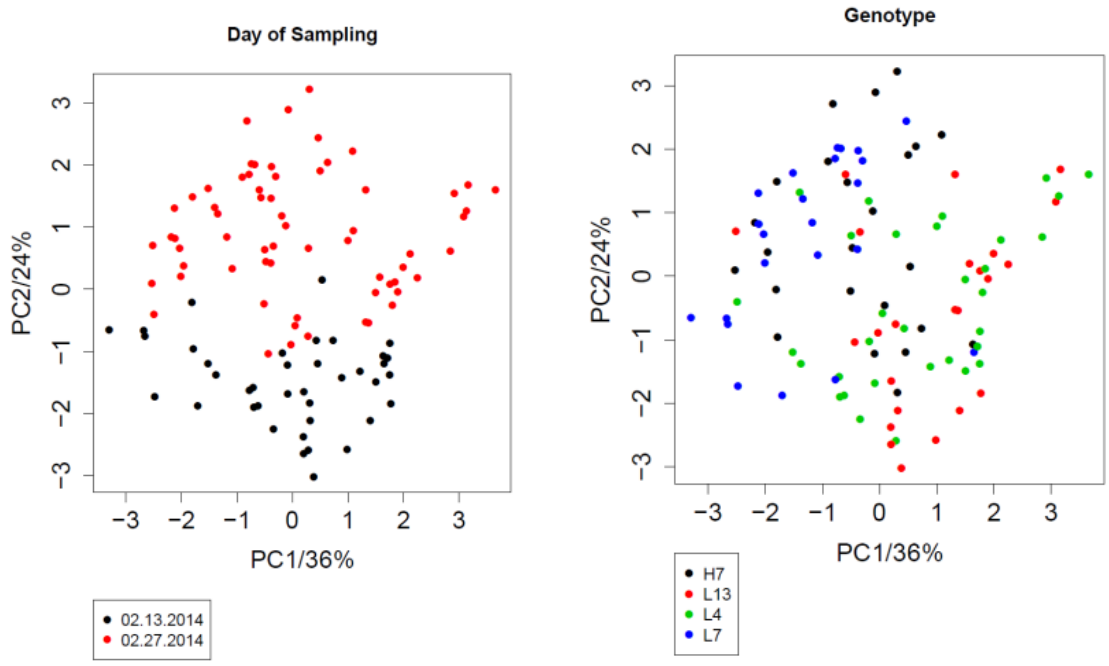


Figure S9. Optimization of experimental design – metabolite profiling

Principal component analysis was performed on normalised and outlier-corrected (median ± 2x SD) polar primary metabolite data obtained from a GC-MS analysis. In total, 190 metabolites, 81 of known and 109 of unknown chemical structure were quantified relatively (ion counts). Metabolite levels were Pareto scaled and centred prior to analysis. PCA calculation was performed by an iterative method using a Bayesian model to handle missing values. The first four principal components were calculated and scatter plots of PC1 and PC2 are shown with the proportion of explained variance, 36 % for PC1 and 24 % for PC2, given on the corresponding axes. Samples in the left figure are coloured by sampling day: 28 DAS (red) and 14 DAS (black). At least eight replicates per genotype / time point combination were analysed. Samples in the right figure are coloured by genotypes, as indicated in the legend.

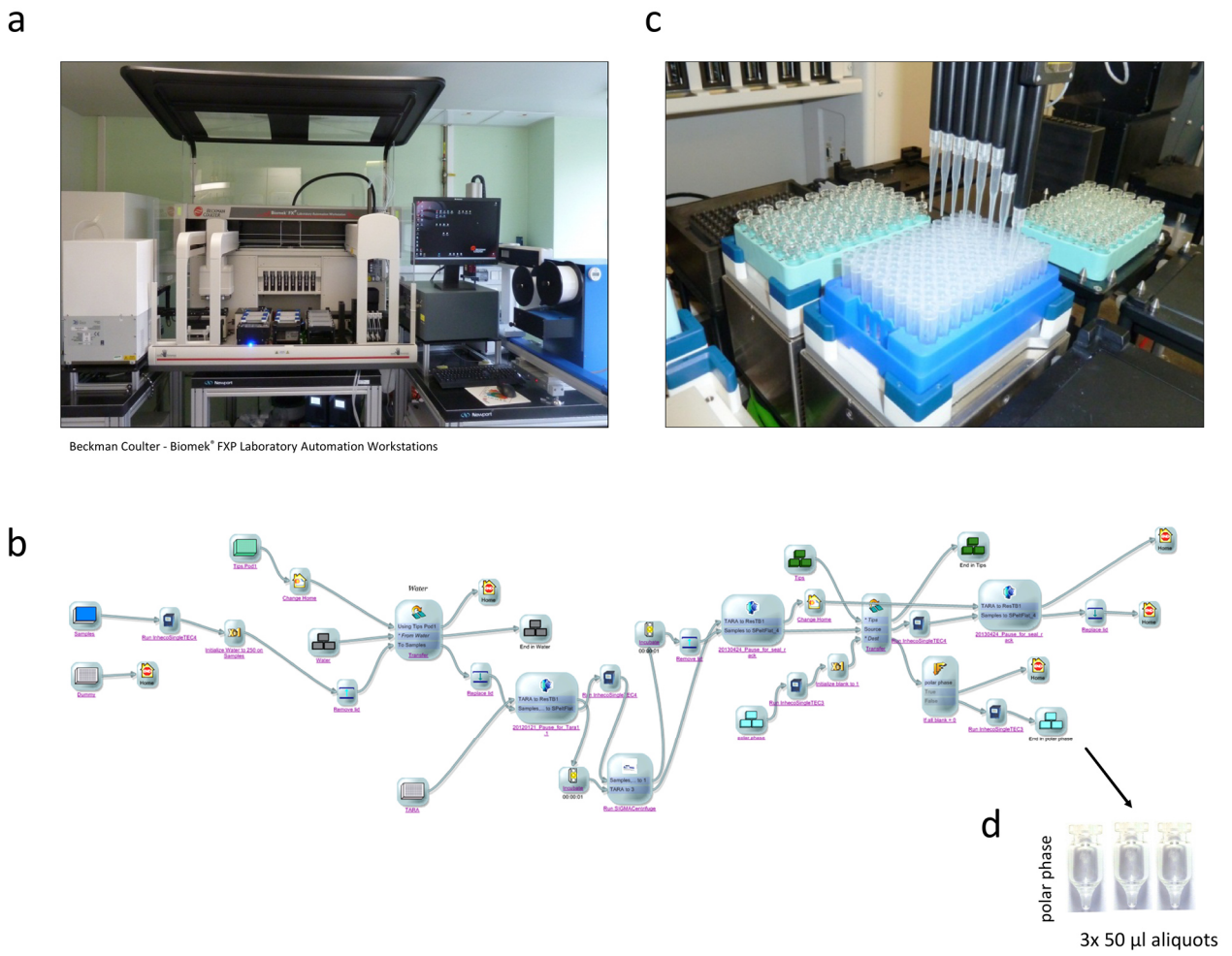


Figure S10. Automated extraction of polar metabolites using a liquid handling system

The figure shows the automated liquid handling system used to establish the (semi-)automated extraction of polar primary metabolites. **a** The system is composed of a Biomek FXP laboratory automation workstation, with four temperate shakers, an automated rack capper/decapper, a labware storage module, a plate centrifuge, self-filling reservoirs, a plate sealer and a TECAN plate reader Infinite 200 PRO. **b** The extraction procedure, as described in materials and methods, was implemented using the Biomek software and the SAMI EX 4.0 software. **c** Liquid handling and labware movement was realised with the two multichannel pipettors (96 channel head plus gripper and a flexible 8 channel head). The polar phase of each sample was transferred from the Micronic screw cap rack with 96 individual 1.4 ml tubes into GC-MS glass vials. **d** A total of three 50 µl aliquots were generated for each sample.

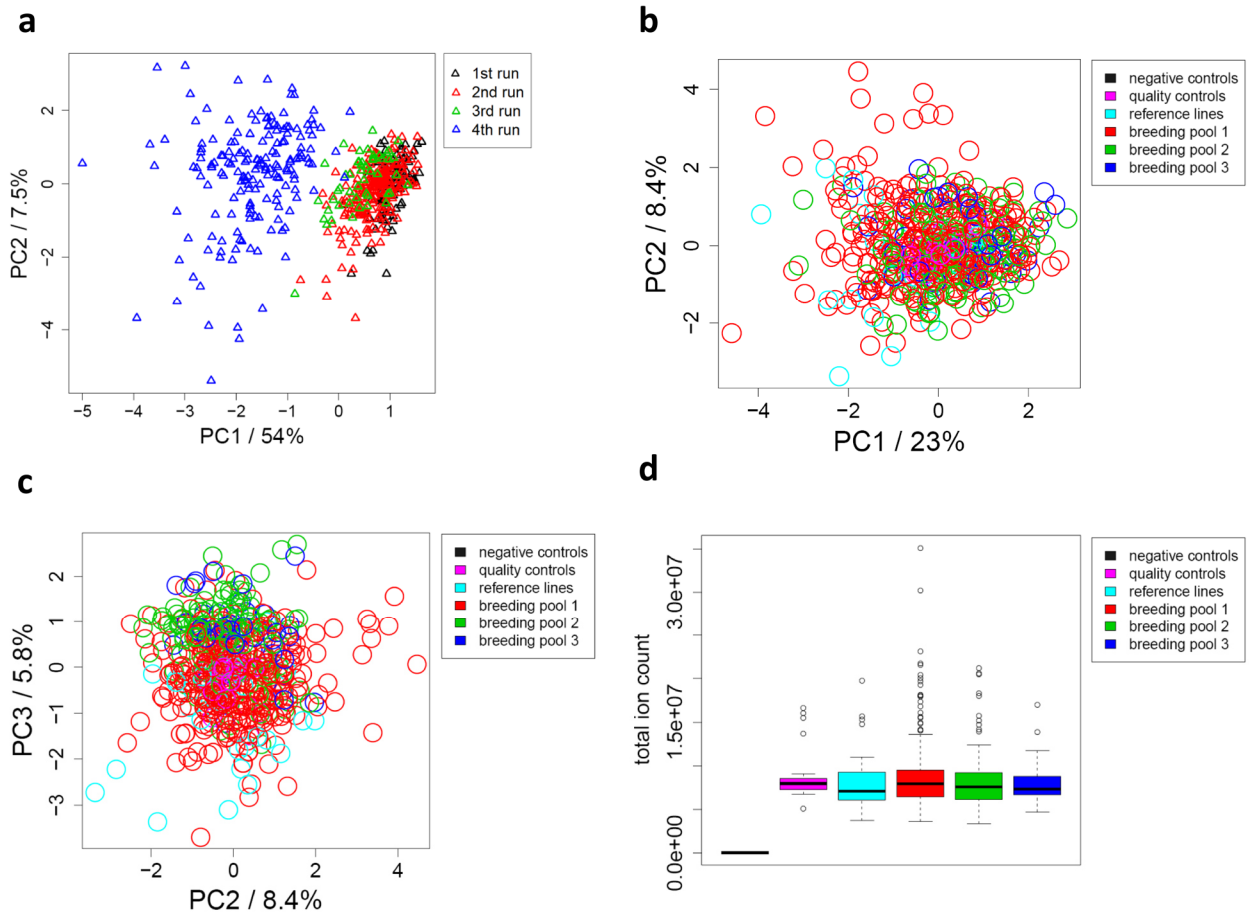


Figure S11. Quality control of polar primary metabolite data after normalisation

Principal component analyses were performed on weight (and measurement day) normalised polar primary metabolite data obtained from a GC-MS analysis. Metabolites were centred and scaled and z-scores for each metabolite were generated (negative controls were excluded from the PCA). PCA calculations were performed by an iterative method using a Bayesian model to handle missing values. **a** Scatter plot of PC1 and PC2 before 'measurement day' (GC-MS run; see colour key) normalisation with the proportions of explained variance given on the axes. Scatter plots of **b** PC1 and PC2, and **c** PC2 and PC3 after data normalisation (see materials and methods) with the proportions of explained variance given on the axes. **d** Boxplots of total ion counts for each sample. The colour key refers to negative controls (n= 8); the quality control pools of all samples (n= 27); the four reference lines ('Achat', 'Campino', 'MS1', 'MS2', each n= 7), and the 477 pollinators grouped by the three breeding pools ('breeding pool 1, 'breeding pool 2, and 'breeding pool 3), respectively.

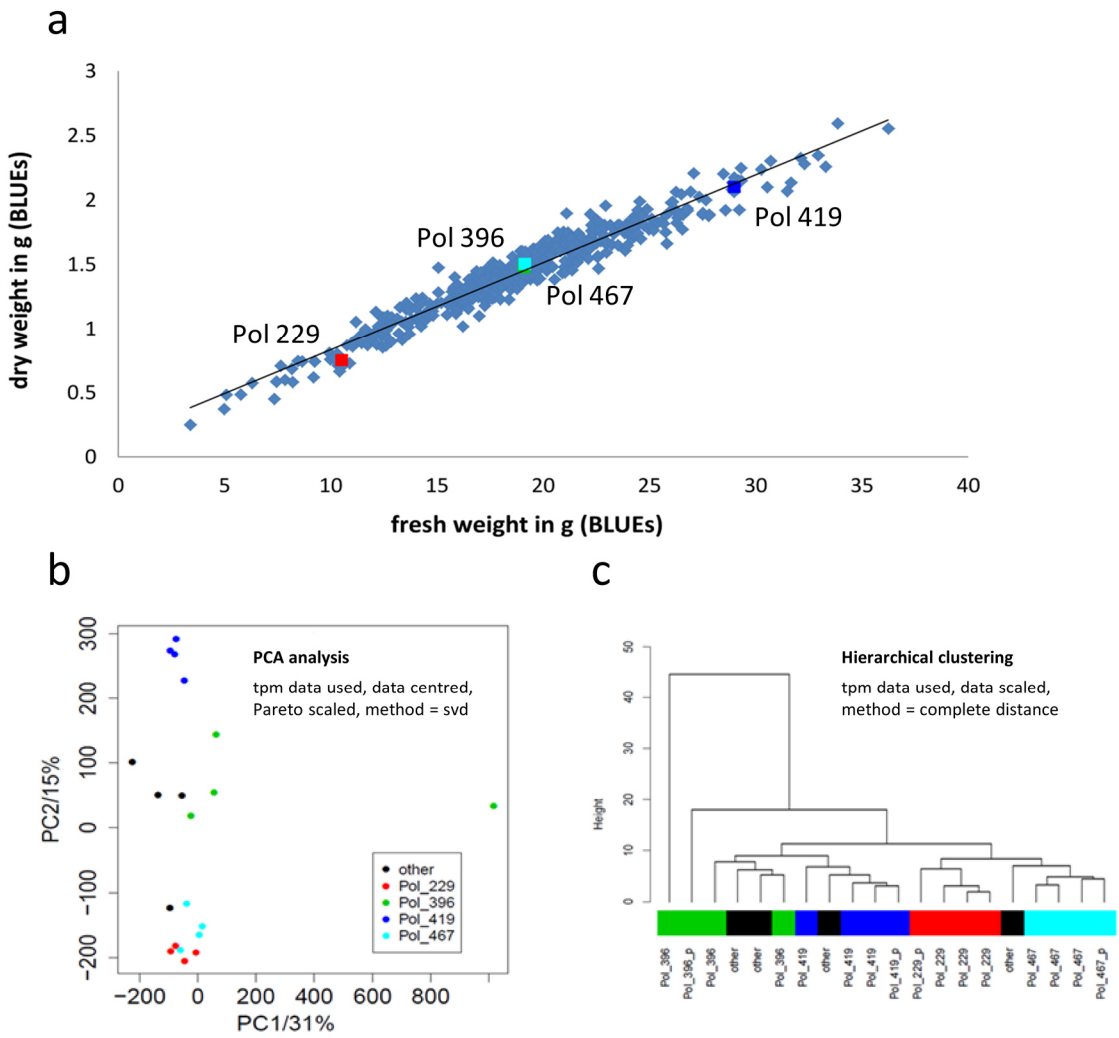


Figure S12. RNA-Seq pilot experiment

Subfigure **a** illustrates the behaviour of the four genotypes selected for the pilot experiment: Pol 229 (red, low biomass), Pol 396 (cyan, medium biomass), Pol 469 (green, medium biomass), and Pol 419 (dark blue, high biomass) in the context of biomass production of all 477 genotypes. Subfigure **b** shows the scatterplot of a principal component analysis using centred and Pareto scaled transcript data (tpm) with the first two principal components explaining 31 % and 15 % of the variance, respectively. The PCA calculation was done by singular value decomposition (svd) of the data matrix with all 38,590 annotated transcripts expressed > 0 tpm in all samples. Each genotype was replicated four times: three samples (a mix of four plants from one single pot from each respective phenotyping experiment) and one combined pool of plants from all phenotyping experiments. Genotypes were coloured as indicated in the figure legend. Note: four randomly selected other genotypes (black colour) were also included in the analysis. Subfigure **c** shows the results of a hierarchical clustering analysis (HCA) of scaled transcript data (tpm) with Euclidean distance and method 'complete distance'. The pools (denoted by a '_p' suffix in the sample name) group near or within the constituent samples.

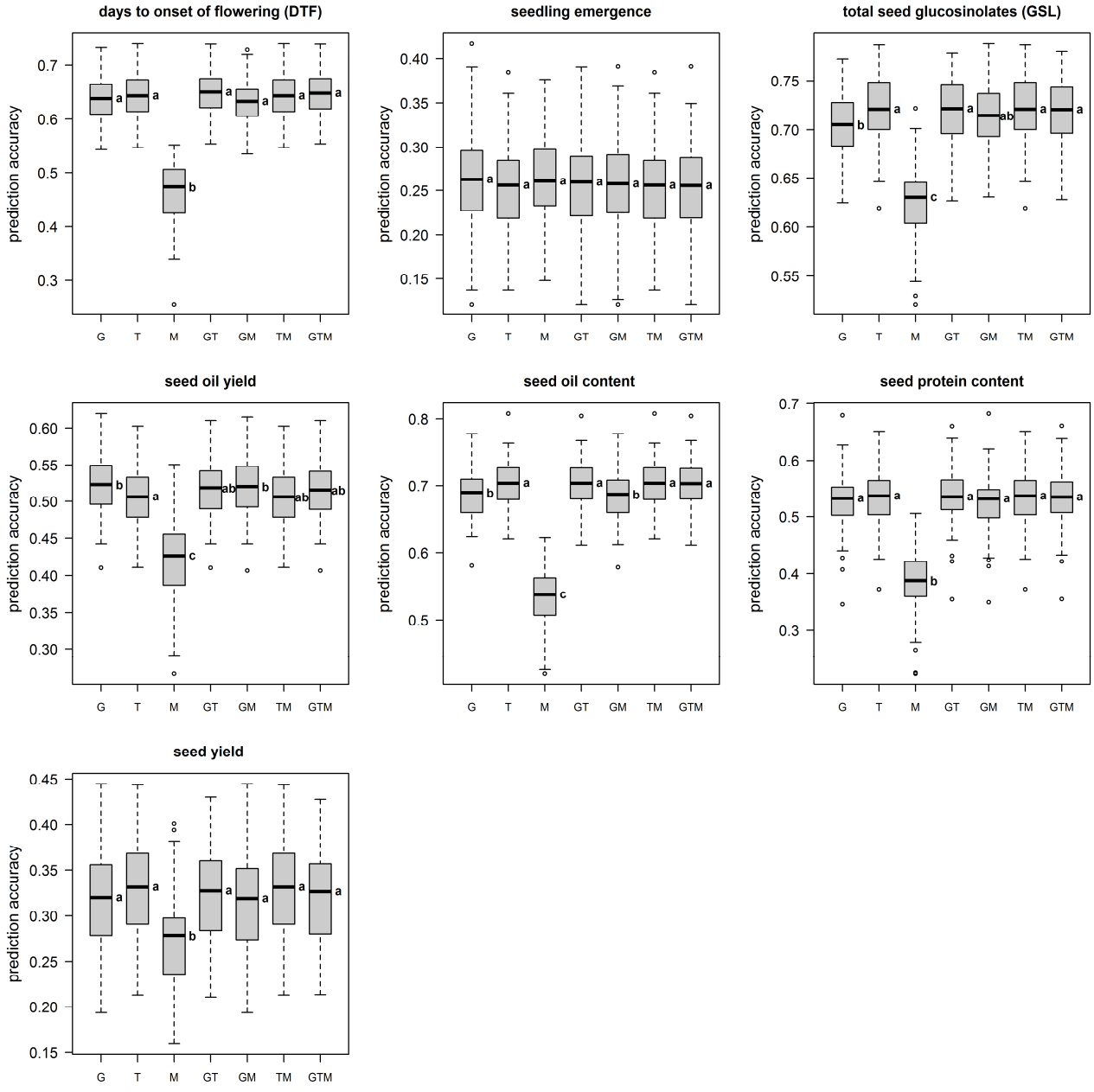


Figure S13. Prediction accuracies for RKHS models

Predictions based on reproducing kernel Hilbert space regression (RKHS) models are summarised for the seven agronomic traits as boxplots. The prediction accuracies were defined as the correlation between the true and the predicted phenotypic values. The different -omics data sets as predictors are denoted as: G, genomic data; T, transcriptomic data; M, metabolite data and their respective combinations G+T, G+M, M+T and G+T+M. Letters beside the boxes indicate significant differences between predictor sets determined by a one-way ANOVA followed by a post-hoc Tukey's multiple comparison test.

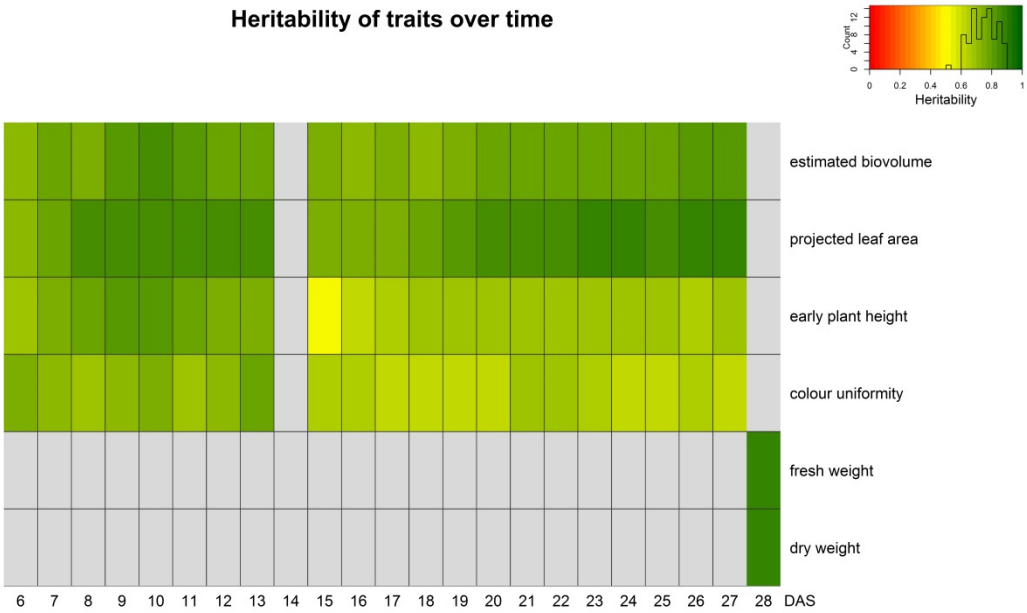


Figure S14. Heritability of selected phenotypic traits over time

Broad-sense heritability values (H^2) of the four image-derived traits (estimated biovolume, projected leaf area, early plant height, colour uniformity) and end-point biomass are shown as a heatmap over time (6 to 28 DAS). H^2 for end-point biomass was estimated based on 15 individual plants grown in an incomplete randomised block design. H^2 values for the four image-derived traits were estimated for each time point individually, based on three replicates (carriers). Grey colour indicates missing values.

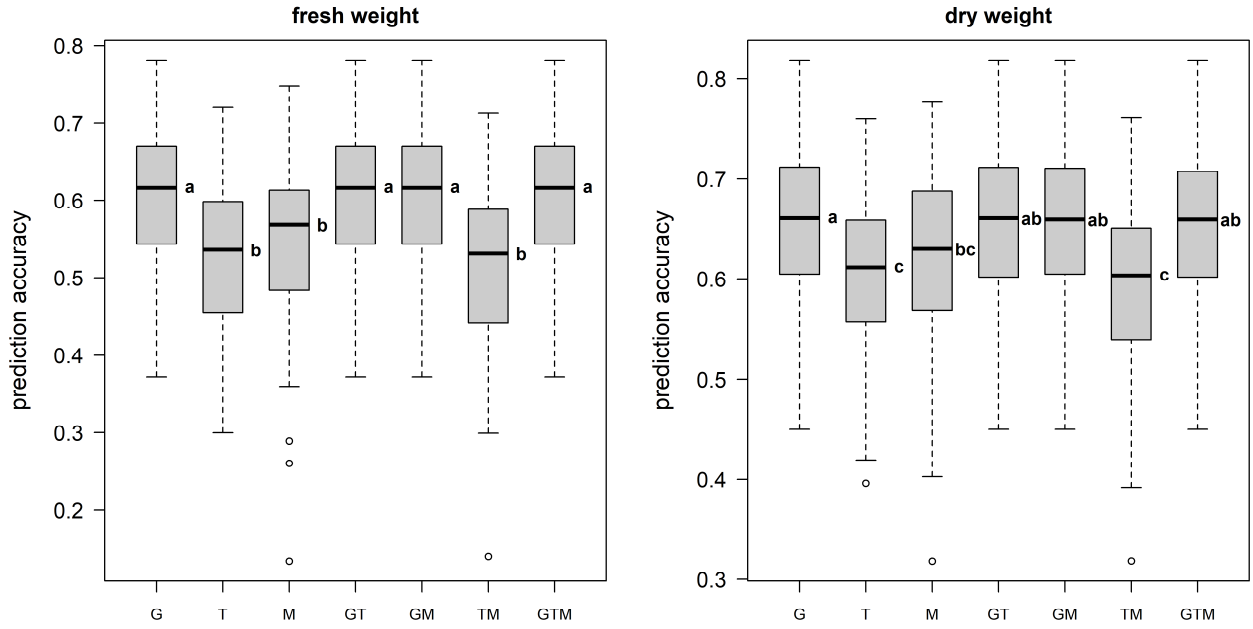


Figure S15. Prediction accuracies for hybrid biomass (FW & DW) in the glasshouse

The summary of best linear unbiased predictions (BLUPs) for early plant biomass of the 120 hybrids is given as boxplots: fresh weight (left) and dry weight (right). The biomass data was obtained from the 5th glasshouse phenotyping experiment at 28 DAS. The prediction accuracies were defined as the correlation between the true and the predicted phenotypic values. The different -omics data sets as predictors are denoted as: G, genomic data; T, transcriptomic data; M, metabolite data and their respective combinations G+T, G+M, M+T and G+T+M. Letters beside the boxes indicate significant differences between predictor sets determined by a one-way ANOVA followed by a post-hoc Tukey's multiple comparison test.

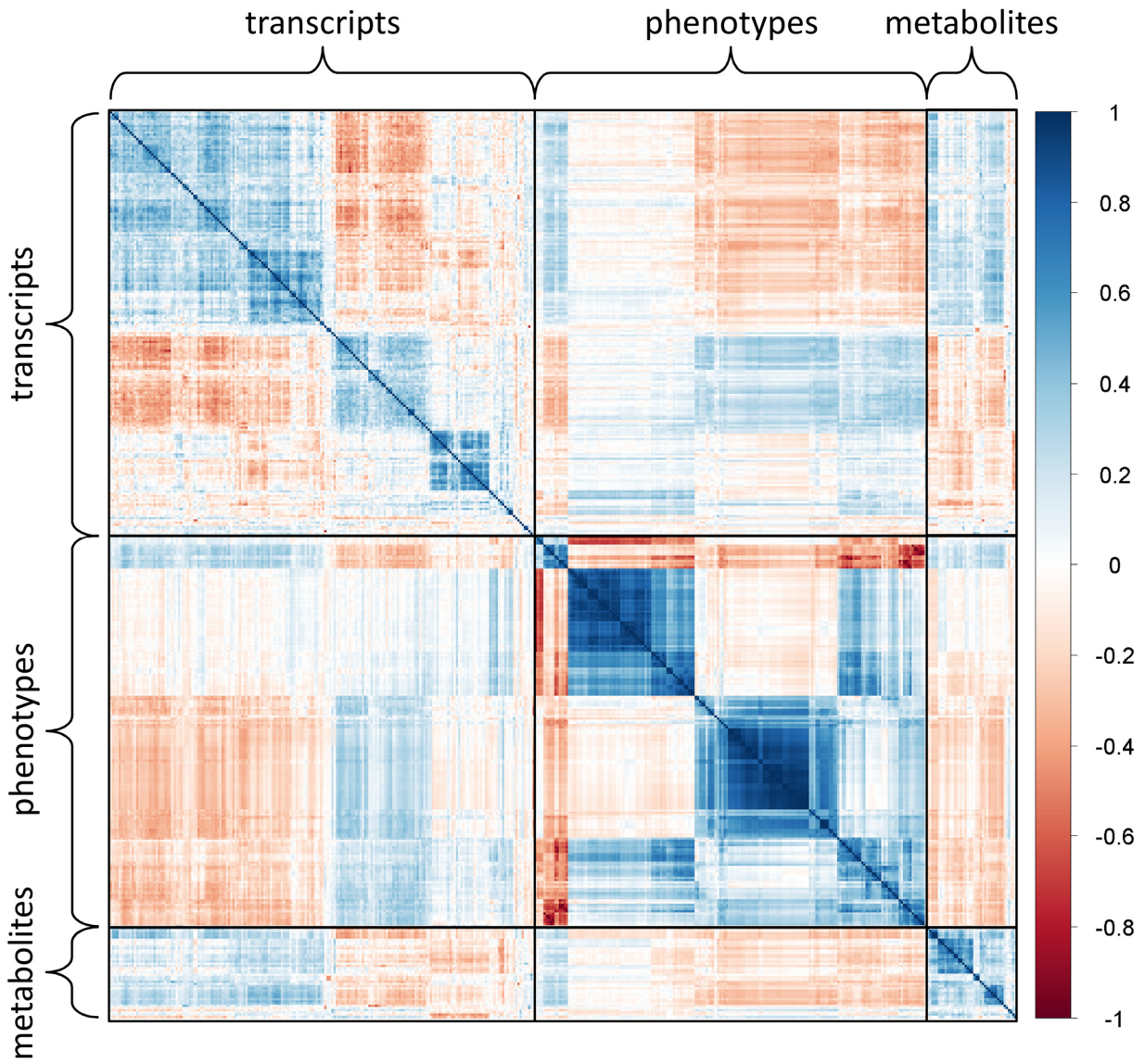


Figure S16. Correlation analyses within and between the -omics data sets

The correlogram displays the Pearson correlation matrix between the three -omics data sets. Only features with significant correlations ($|r| \geq 0.4$; p-value $_{FDR} \leq 0.05$) between sets are plotted ($n_{transcripts}= 193$, $n_{phenotypes}= 179$, $n_{metabolites}= 42$). Features were clustered for each data set individually using a hierarchical clustering analysis (HCA) with the agglomeration method 'complete linkage'. Black boxes separate the individual sets of correlations. The correlation matrix and raw values are given in Data S12.

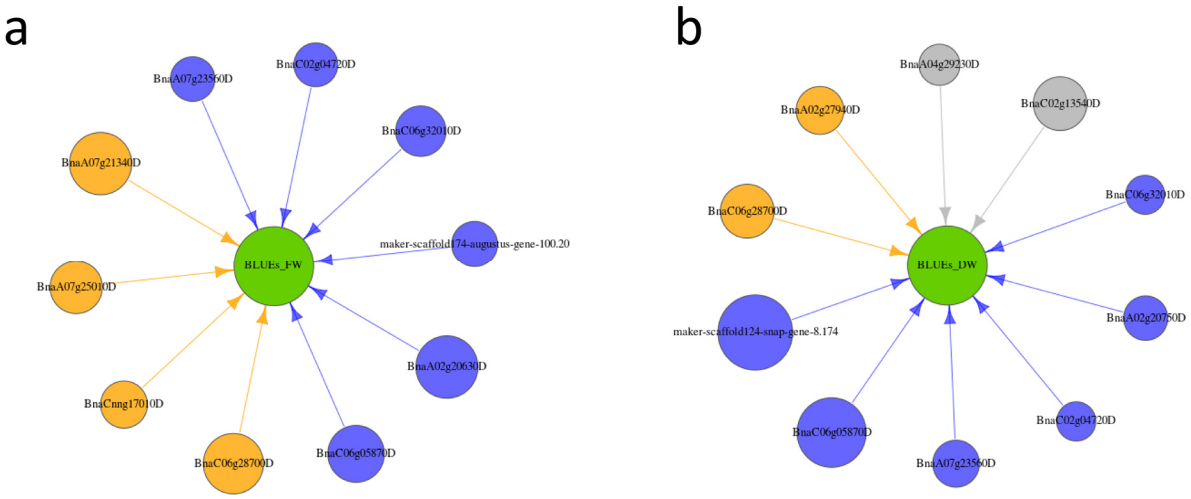


Figure S17. GENIE3 network analysis for biomass using transcript data

The figure shows candidate genes with effects on early plant biomass identified by a gene network inference analysis using tree-based ensemble methods in GENIE3. **a** The graph shows the top ten candidate genes associated with fresh weight. **b** The graph shows the top ten candidate genes associated with dry weight. The colours indicate if the transcripts are positively (orange) or negatively (blue) correlated with biomass. Grey colour indicates non-significant Pearson correlation. The sizes of the vertices correspond to the weight of the association. Five genes, including 'BnaC06g28700D', encoding the 72 kDa signal recognition particle, were found among the top ten candidates for both, fresh and dry weight. Gene annotation details can be obtained from Table 3.

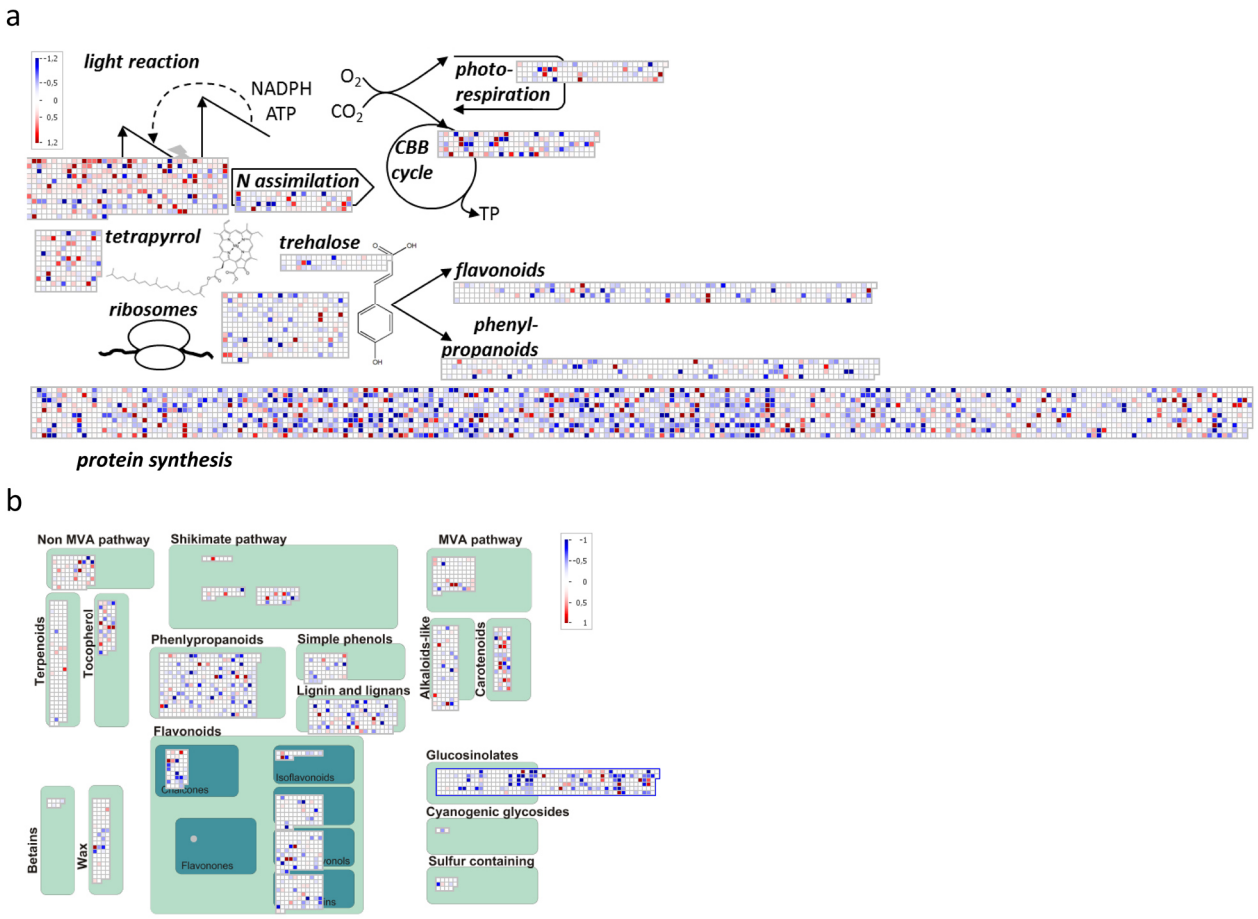
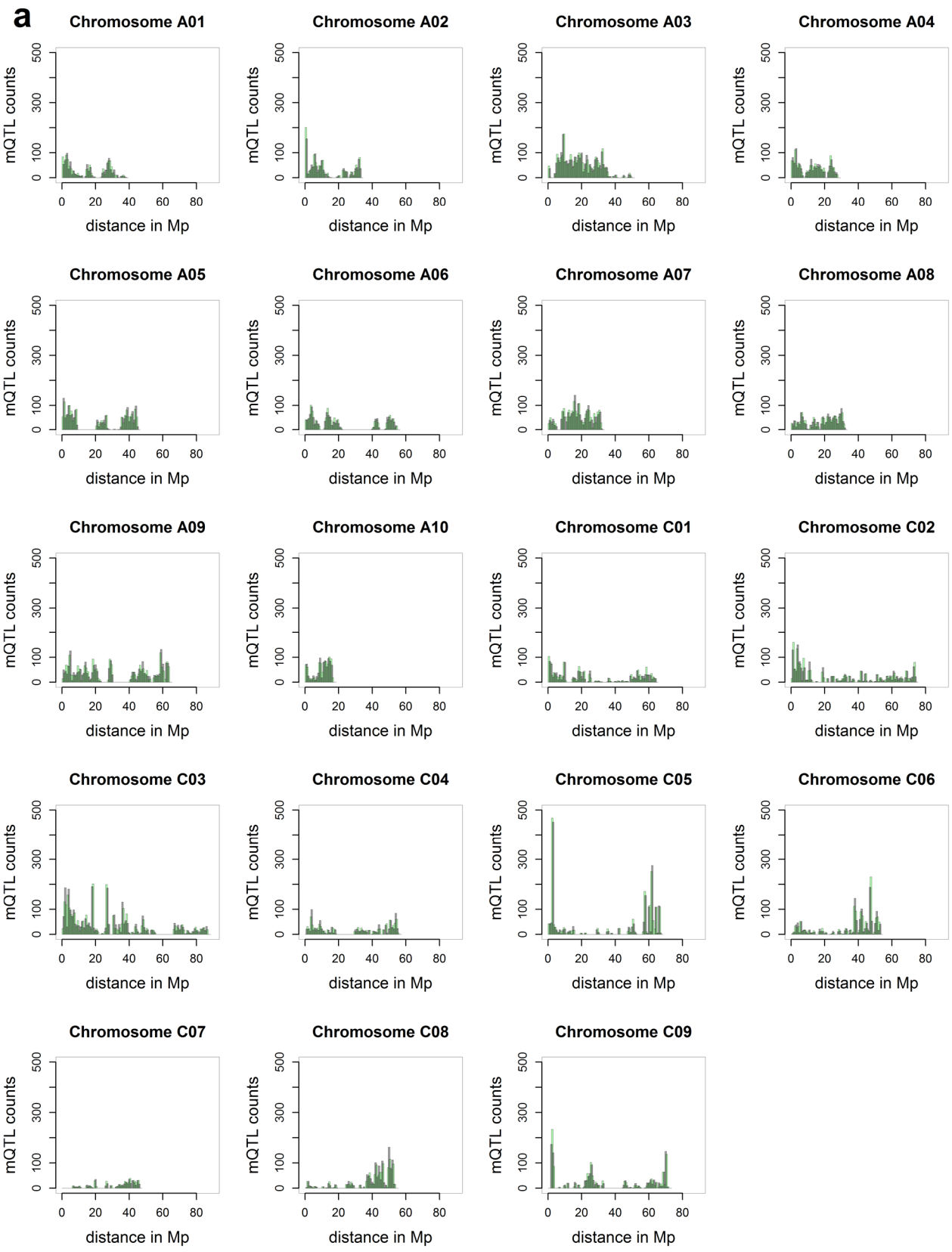
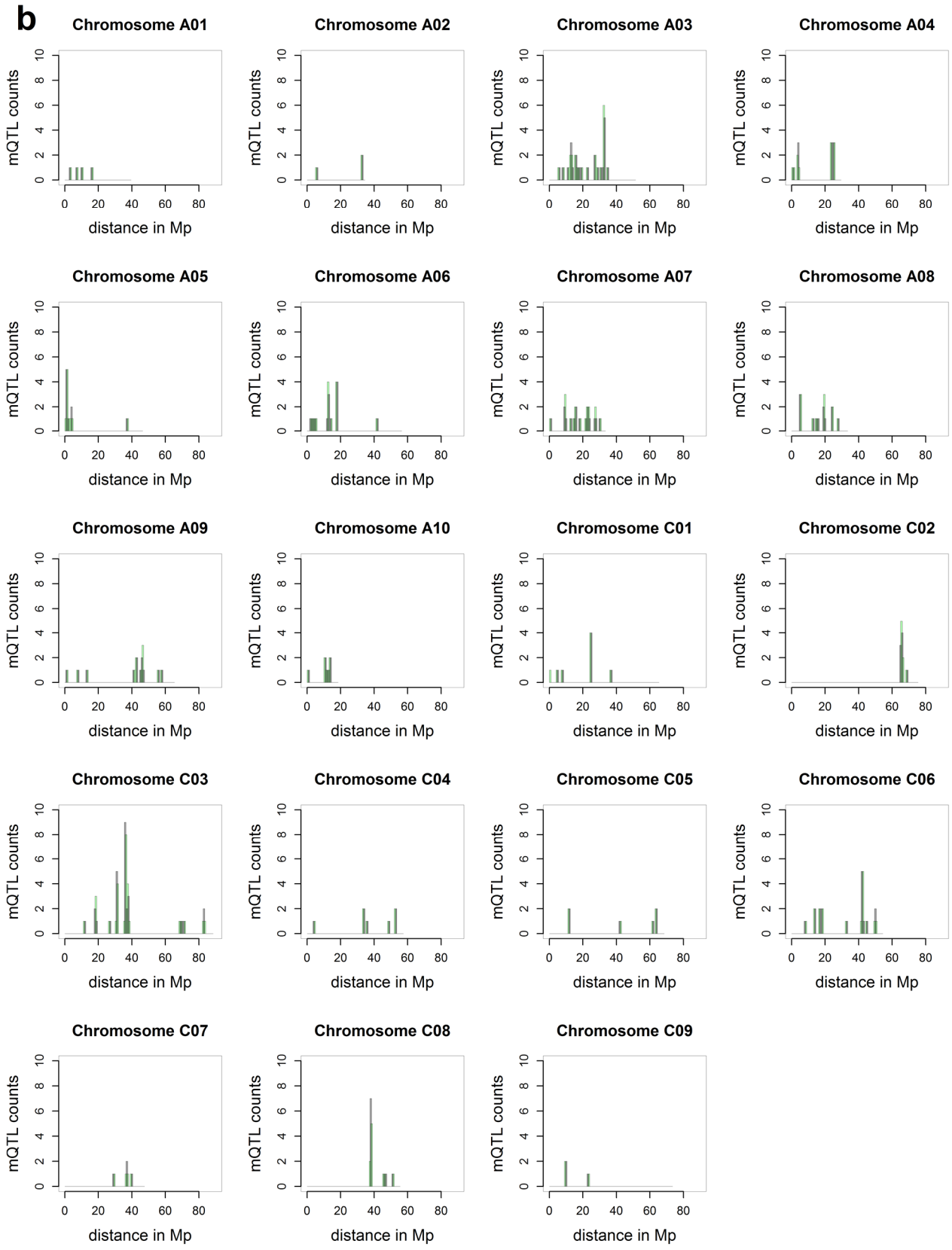


Figure S18. Pathway analysis in lines with contrasting biomass

Subfigure **a** shows the Mapman representation of all coding sequences (CDS, including non-significant) in the selected pathways (customised overview). Pollinator 229 (low biomass) is enriched in protein synthesis, as indicated by the high number of more blue dots in bottom box; Wilcoxon sum rank test, BH corrected: *p*-value < 10^{-20}). Pollinator 419 (high biomass) is enriched in photosynthesis / light reaction, as indicated by the red dots in the top left box (Wilcoxon sum rank test, BH corrected: *p*-value < 10^{-20}). Subfigure **b** shows the Mapman representation of all CDS (including non-significant) in selected pathways of the secondary metabolism. Pollinator 229 (low biomass) is enriched in glucosinolate synthesis (Wilcoxon sum rank test, BH corrected: *p*-value < 10^{-8}).


a
Chromosome A01
Chromosome A02
Chromosome A03
Chromosome A04
Chromosome A05
Chromosome A06
Chromosome A07
Chromosome A08
Chromosome A09
Chromosome A10
Chromosome C01
Chromosome C02
Chromosome C03
Chromosome C04
Chromosome C05
Chromosome C06
Chromosome C07
Chromosome C08
Chromosome C09
mQTL counts
0
100
300
500
0
20
40
60
80
distance in Mp


b
Chromosome A01
Chromosome A02
Chromosome A03
Chromosome A04
Chromosome A05
Chromosome A06
Chromosome A07
Chromosome A08
Chromosome A09
Chromosome A10
Chromosome C01
Chromosome C02
Chromosome C03
Chromosome C04
Chromosome C05
Chromosome C06
Chromosome C07
Chromosome C08
Chromosome C09
mQTL counts
distance in Mp

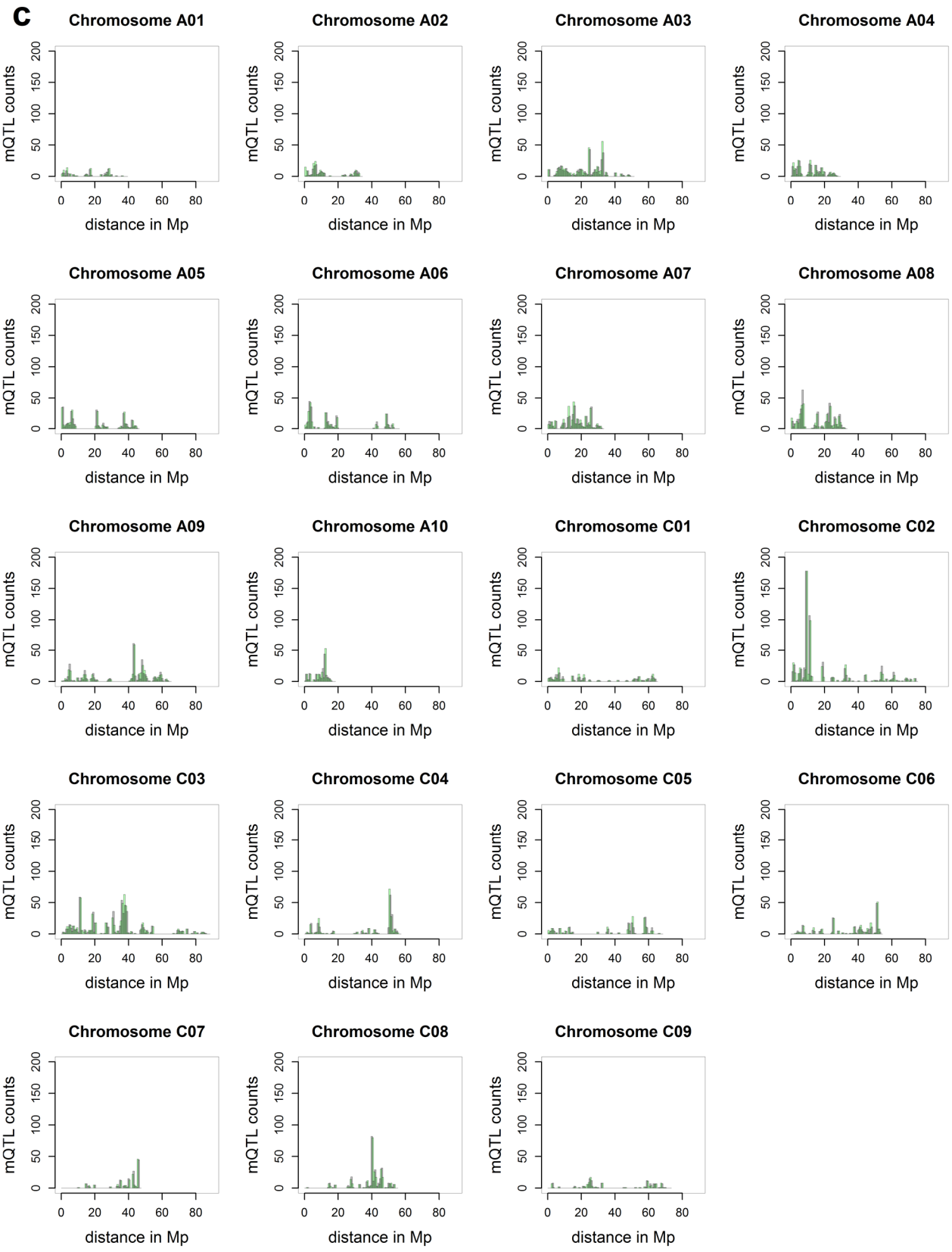

c
Chromosome A01
Chromosome A02
Chromosome A03
Chromosome A04
Chromosome A05
Chromosome A06
Chromosome A07
Chromosome A08
Chromosome A09
Chromosome A10
Chromosome C01
Chromosome C02
Chromosome C03
Chromosome C04
Chromosome C05
Chromosome C06
Chromosome C07
Chromosome C08
Chromosome C09
mQTL counts
distance in Mp
0
50
100
150
200
0
20
40
60
80

Figure S19. QTL-hotspots for eQTL, mQTL and phenotypic QTL

The composite figures display **a** the expression QTL (eQTL) distribution **b** the metabolite QTL (mQTL) distribution and **c** the phenotypic QTL distribution on all 19 *Brassica napus* chromosomes. The QTL were binned into overlapping 1 Mb intervals (transparent green and grey colour, e.g. 0-500 kb and 250-750 kb, respectively) for representation. The number of QTL per bin is indicated on the vertical axis, the chromosomal position in Mb is shown on the horizontal axis. The axes were locked at maximum values of n_{eQTL}= 500, n_{mQTL}= 10, n_{QTL}= 200 and 90 Mb, respectively, to allow cross-comparison between chromosomes within the sets.

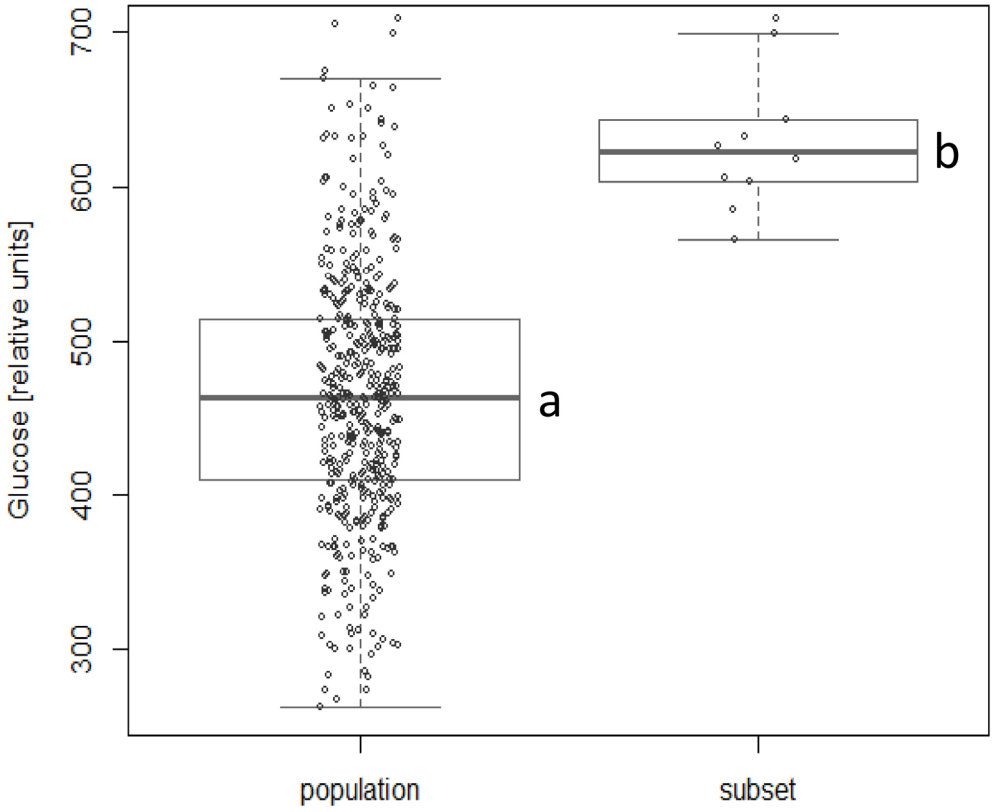
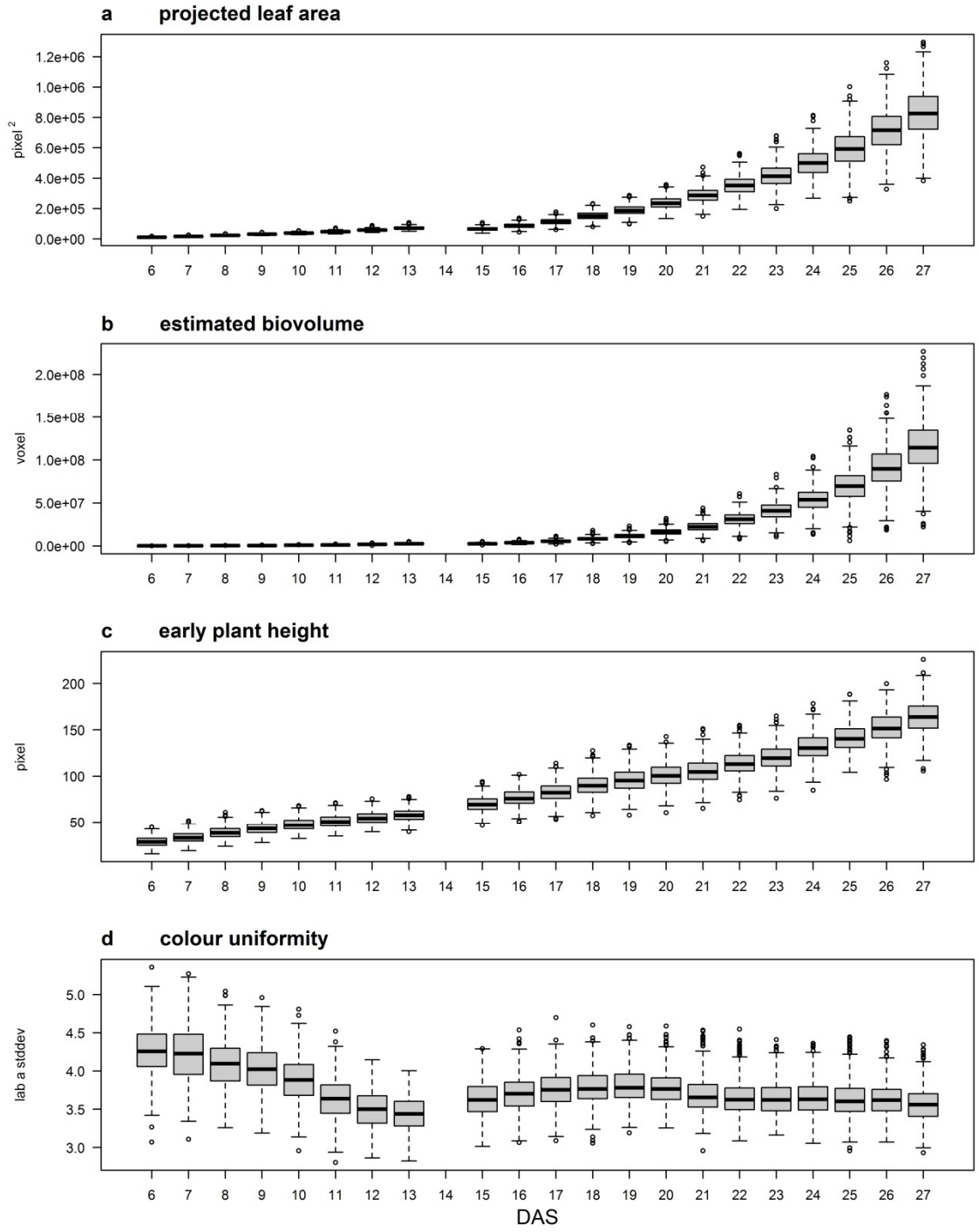


Figure S20. Differences in glucose content in lines with deletion on chromosome C03

The figure shows the glucose content relatively quantified by GC-MS. The left boxplot represents the overall distribution of glucose content in the 477 analysed genotypes. The right boxplot shows a subset of ten genotypes carrying a deletion on chromosome C03. These genotypes display a significantly higher glucose content compared to the population mean (Welch Two Sample *t*-test; *p*-value= 4.668e-7).


a projected leaf area
pixel 2
1.2e+06
1.0e+06
8.0e+05
6.0e+05
4.0e+05
2.0e+05
0.0e+00
b estimated biovolume
voxel
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
c early plant height
pixel
200
150
100
50
d colour uniformity
lab a stddev
5.0
4.5
4.0
3.5
3.0
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
DAS

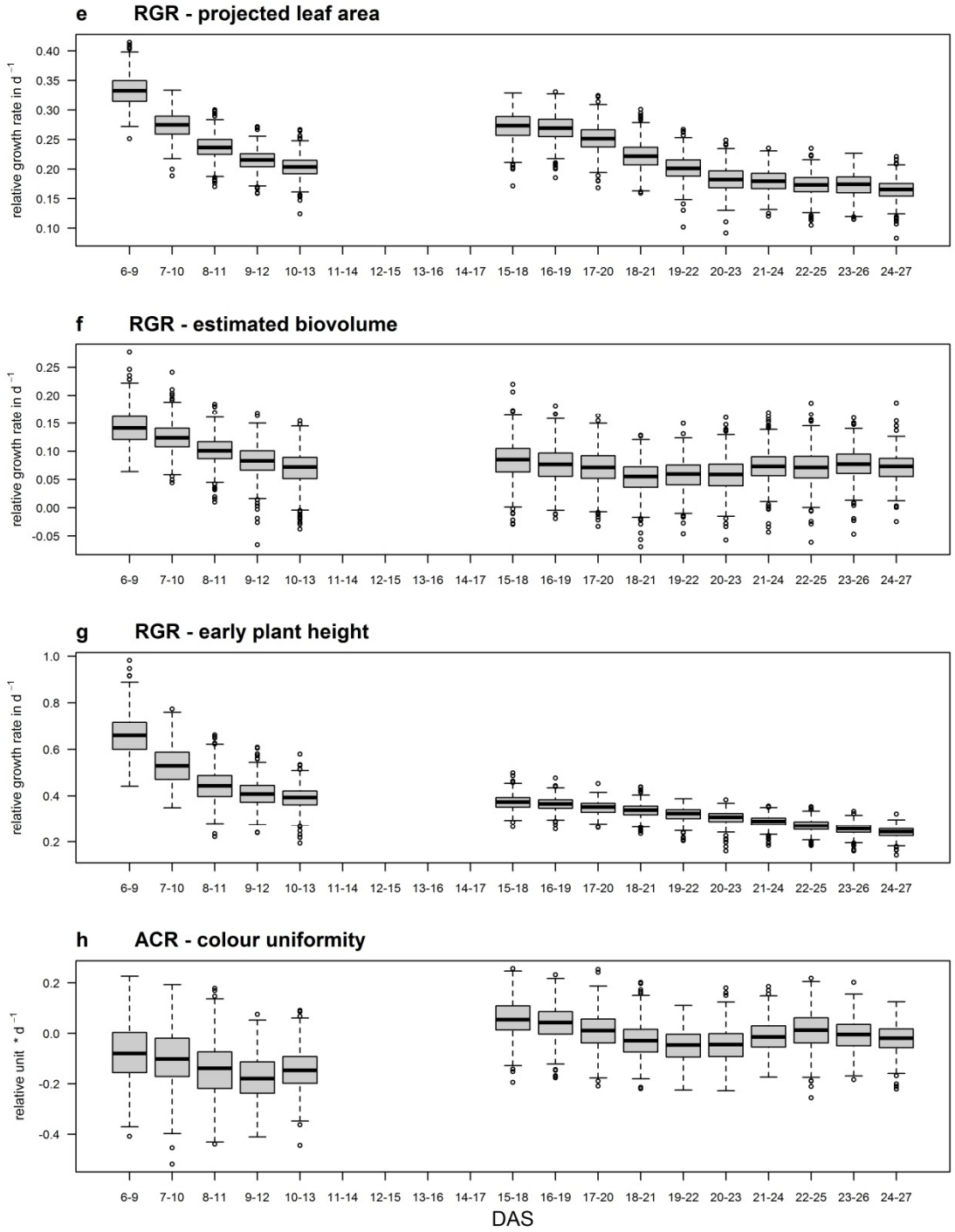


Figure S21. Overview of selected phenotypic data

Boxplots represent BLUEs across experiments of **a** projected leaf area, **b** estimated biovolume, **c** early plant height and **d** plant colour uniformity over time. Colour uniformity is given as the standard deviation of the a-values in the L*a*b* colour space of the plant pixels. The lower this value, the more uniform is the plant colour. BLUEs for relative growth rates (RGRs) of **e** projected leaf area, **f** estimated biovolume, **g** early plant height and absolute change rates (ACRs) of **h** colour uniformity were calculated over three days.

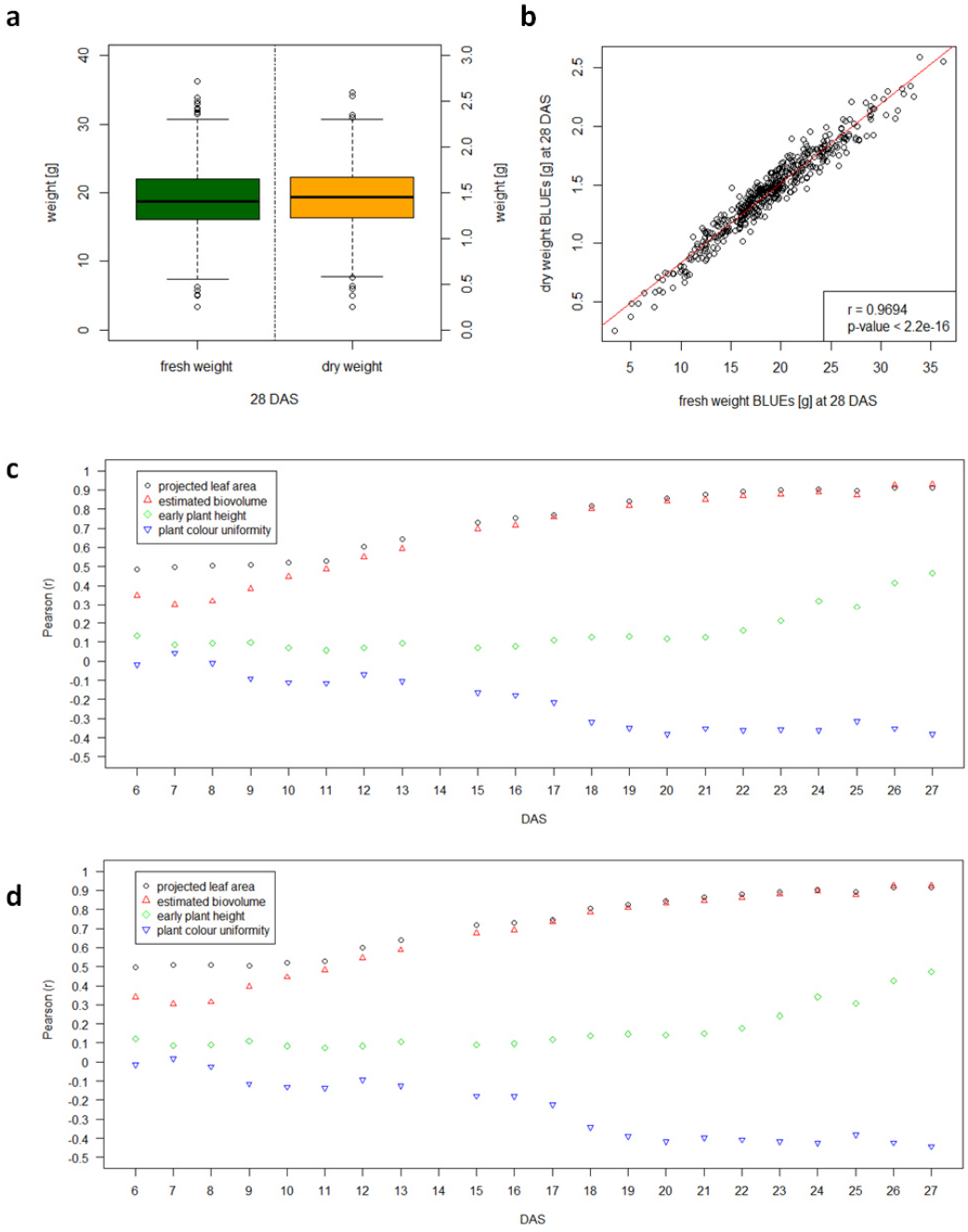


Figure S22. Biomass distribution and correlation with image-derived traits

a End-point biomass data as fresh weight (FW, in green) and dry weight (DW, in orange) were manually determined at 28 DAS. BLUEs for both traits were estimated based on 15 individual plants grown in an incomplete randomised block design over four experiments. Data are displayed as boxplots. **b** Pearson correlation of fresh weight (FW) and dry weight (DW) **c** Pearson correlations of fresh weight (FW) with image-derived traits. **d** Pearson correlations of dry weight (DW) with image-derived traits.

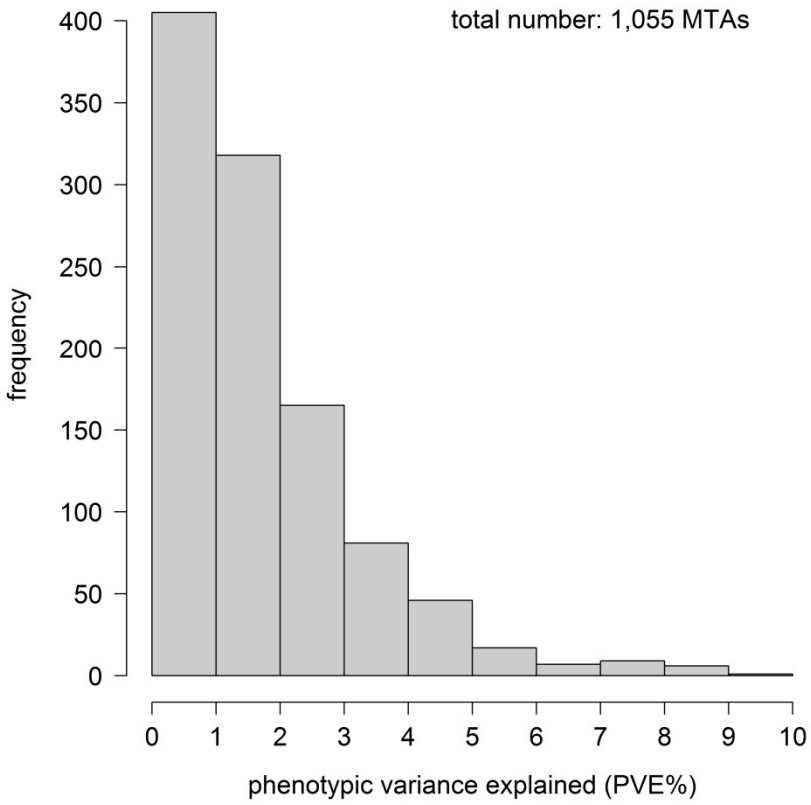


Figure S23. Phenotypic variance explained (PVE%) by detected MTAs

Histogram of phenotypic variance explained (PVE%) of all 1,055 marker-trait associations (MTAs) for growth and biomass-related traits with a *p*-value $_{(FDR)} \leq 0.1$ detected in this study for projected leaf area (n=200), estimated biovolume (n=191), early plant height (n=182), colour uniformity (n=192), end-point fresh (n=13) and dry weight (n=9), as well as relative growth rates for area (n=76), volume (n=100) and height (n=73), and absolute change rates for colour uniformity (n=19). On average markers explained 1.72 PVE%. Individual markers explained up to 9.05 PVE% of particular traits. A comprehensive list of all MTAs is provided as Data S11.

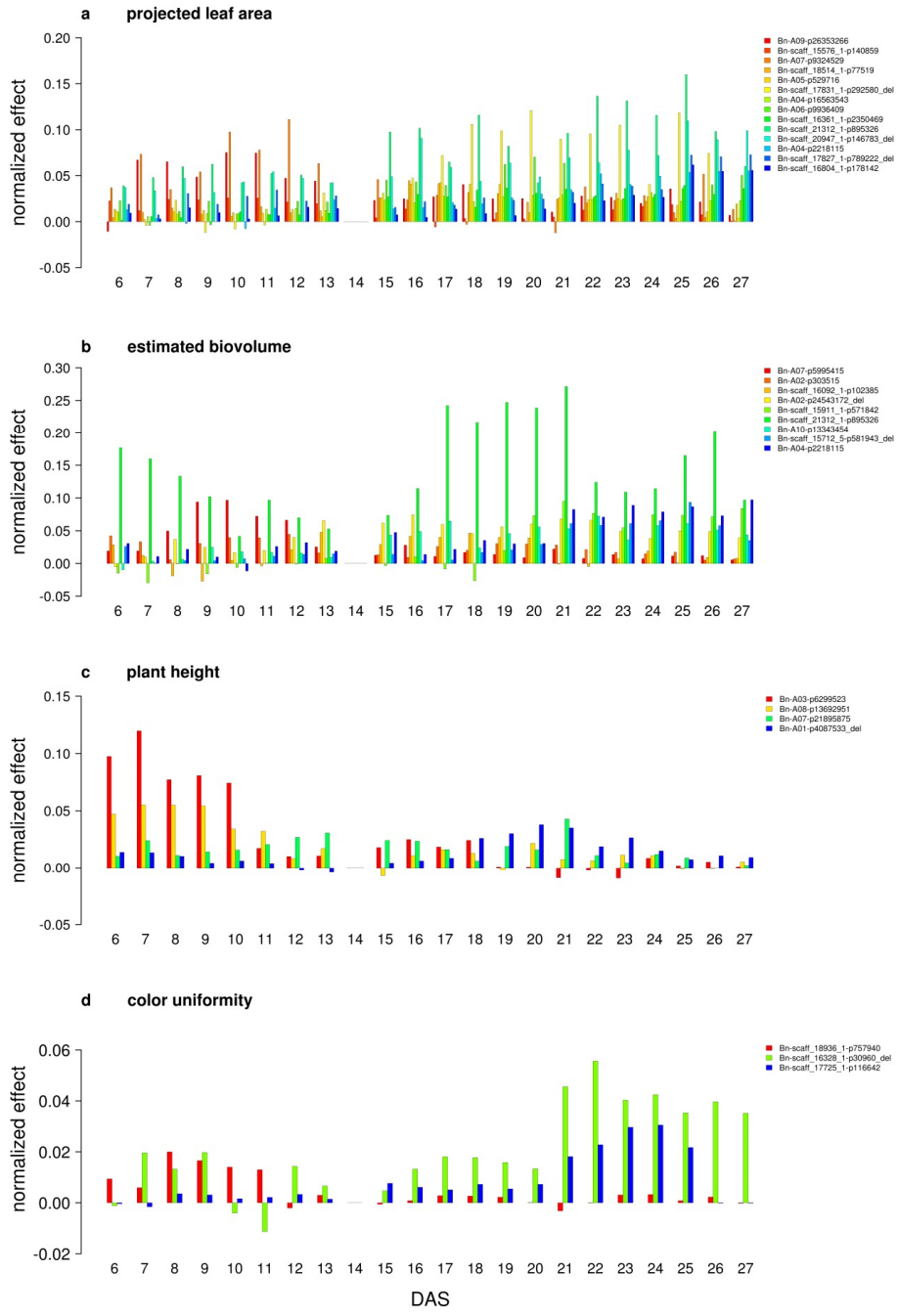


Figure S24. Allele effects of dynamic associations for growth-related traits

The figures display the allele effects of dynamic associations for the traits **a** projected leaf area, **b** estimated biovolume, **c** early plant height and **d** plant colour uniformity. GWAS analyses were performed on BLUEs in R / package 'FarmCPU'. Different colours indicate markers with *p*-value $_{(FDR)}$ ≤0.1 at three consecutive days, with the colour gradient corresponding to the temporal pattern. DAS denotes days after sowing. BLUEs were estimated using three replicates (carriers) with nine and five plants for 6 to 13 DAS and 15 to 27 DAS, respectively. No data were recorded at 14 DAS due to sampling of shoot material. Allele effects were normalised by dividing the effects by the median of the phenotypic trait of each day. For simplification, predominantly negative allele effects were inversed.

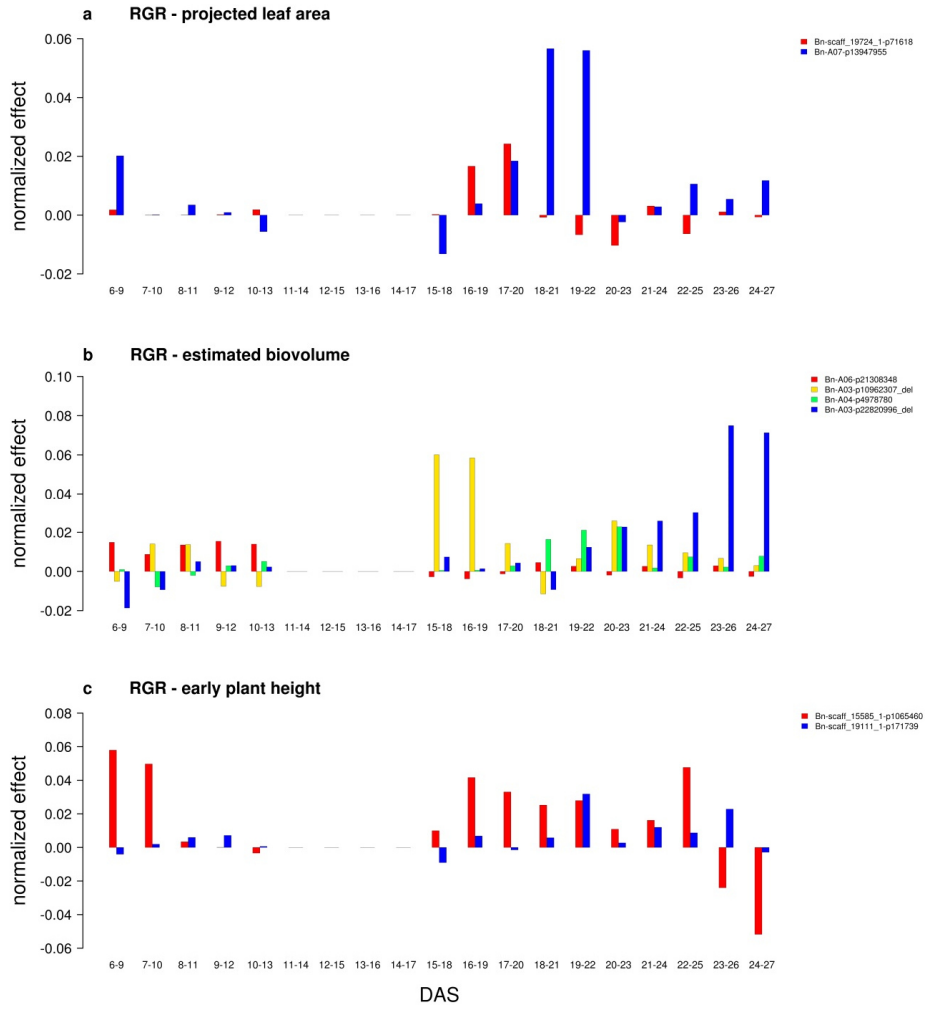
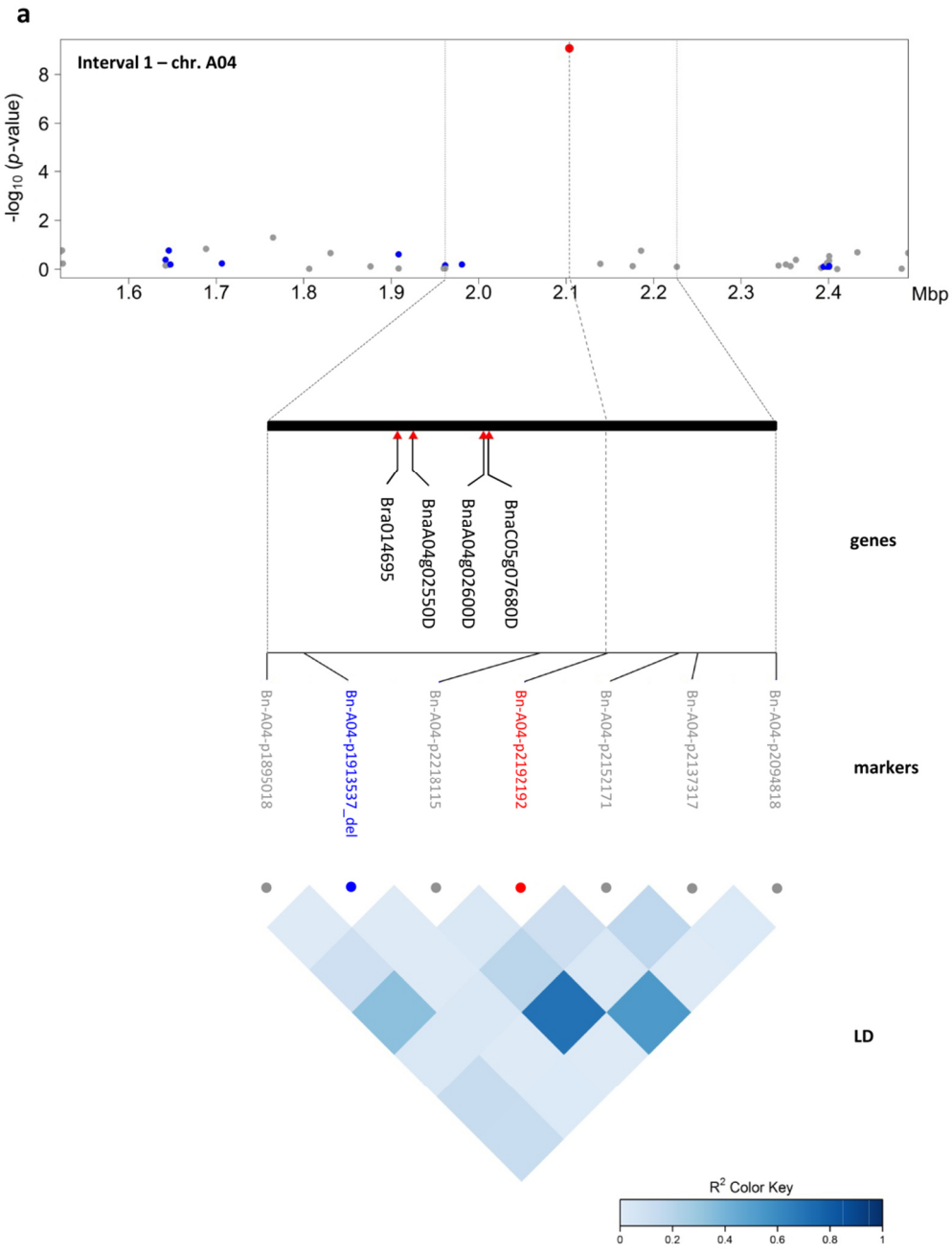
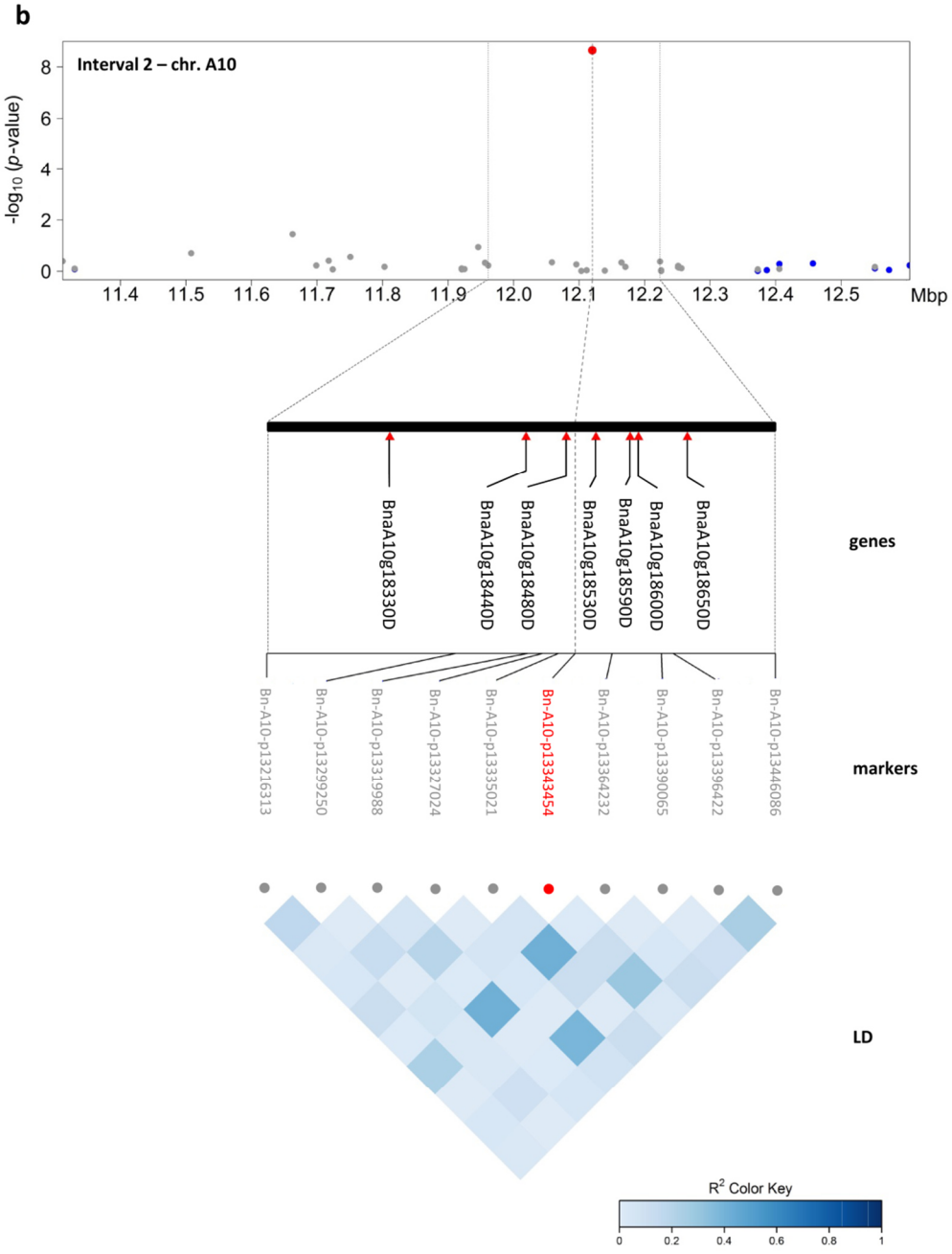
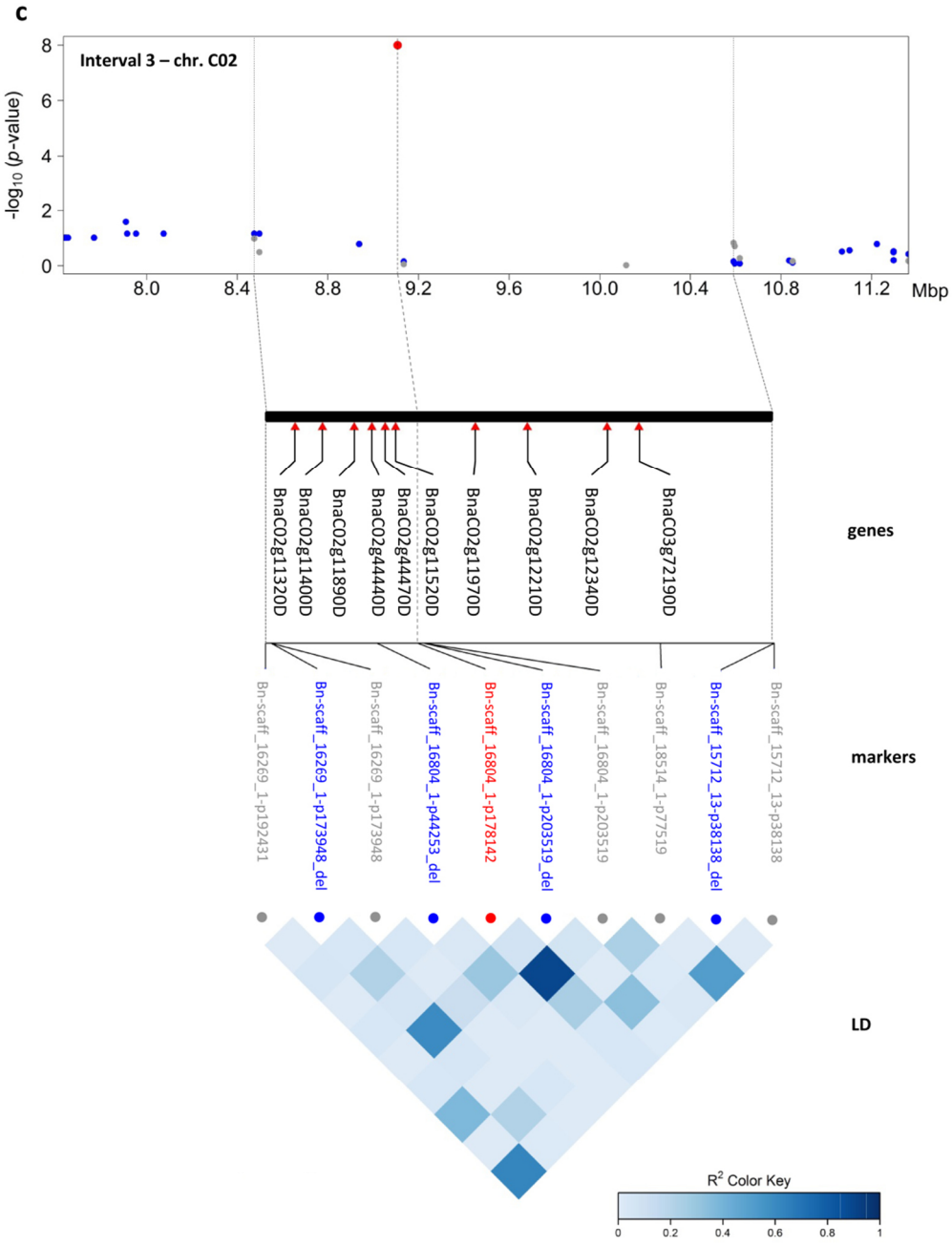


Figure S25. Allele effects of dynamic associations for relative growth rates

GWAS was performed on BLUEs of **a** RGR - projected leaf area, **b** RGR - estimated biovolume and **c** RGR - early plant height in R / package 'FarmCPU'. Different colours indicate markers with *p*-value $_{(FDR)}$ ≤0.1 at two consecutive intervals. DAS denotes days after sowing. BLUEs were estimated using three replicates (carriers) with nine and five plants for 6 to 13 DAS and 15 to 27 DAS, respectively. No data were recorded at 14 DAS due to sampling of shoot material. Allele effects were normalised by dividing the effects by the median of the phenotypic trait of each day. For simplification, predominantly negative allele effects were inversed.


a
Interval 1 – chr. A04
-log10 (p-value)
8
6
4
2
0
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
Mbp
Bra014695
BnaA04g02550D
BnaA04g02600D
BnaC05g07680D
genes
Bn-A04-p1895018
Bn-A04-p1913537_del
Bn-A04-p2218115
Bn-A04-p2192192
Bn-A04-p2152171
Bn-A04-p2137317
Bn-A04-p2094818
markers
LD
R2 Color Key
0
0.2
0.4
0.6
0.8
1


b
Interval 2 – chr. A10
-log10 (p-value)
8
6
4
2
0
11.4
11.5
11.6
11.7
11.8
11.9
12.0
12.1
12.2
12.3
12.4
12.5
Mbp
BnaA10g18330D
BnaA10g18440D
BnaA10g18480D
BnaA10g18530D
BnaA10g18590D
BnaA10g18600D
BnaA10g18650D
genes
Bn-A10-p13216313
Bn-A10-p13299250
Bn-A10-p13319988
Bn-A10-p13327024
Bn-A10-p13335021
Bn-A10-p13343454
Bn-A10-p13364232
Bn-A10-p13390065
Bn-A10-p13396422
Bn-A10-p13446086
markers
LD
R2 Color Key
0
0.2
0.4
0.6
0.8
1


c
Interval 3 – chr. C02
-log10 (p-value)
8
6
4
2
0
8.0
8.4
8.8
9.2
9.6
10.0
10.4
10.8
11.2
Mbp
BnaC02g11320D
BnaC02g11400D
BnaC02g11890D
BnaC02g44440D
BnaC02g44470D
BnaC02g11520D
BnaC02g11970D
BnaC02g12210D
BnaC02g12340D
BnaC03g72190D
genes
Bn-scaff_16269_1-p192431
Bn-scaff_16269_1-p173948_del
Bn-scaff_16269_1-p173948
Bn-scaff_16804_1-p44253_del
Bn-scaff_16804_1-p178142
Bn-scaff_16804_1-p203519_del
Bn-scaff_16804_1-p203519
Bn-scaff_18514_1-p77519
Bn-scaff_15712_13-p38138_del
Bn-scaff_15712_13-p38138
markers
LD
R2 Color Key
0
0.2
0.4
0.6
0.8
1

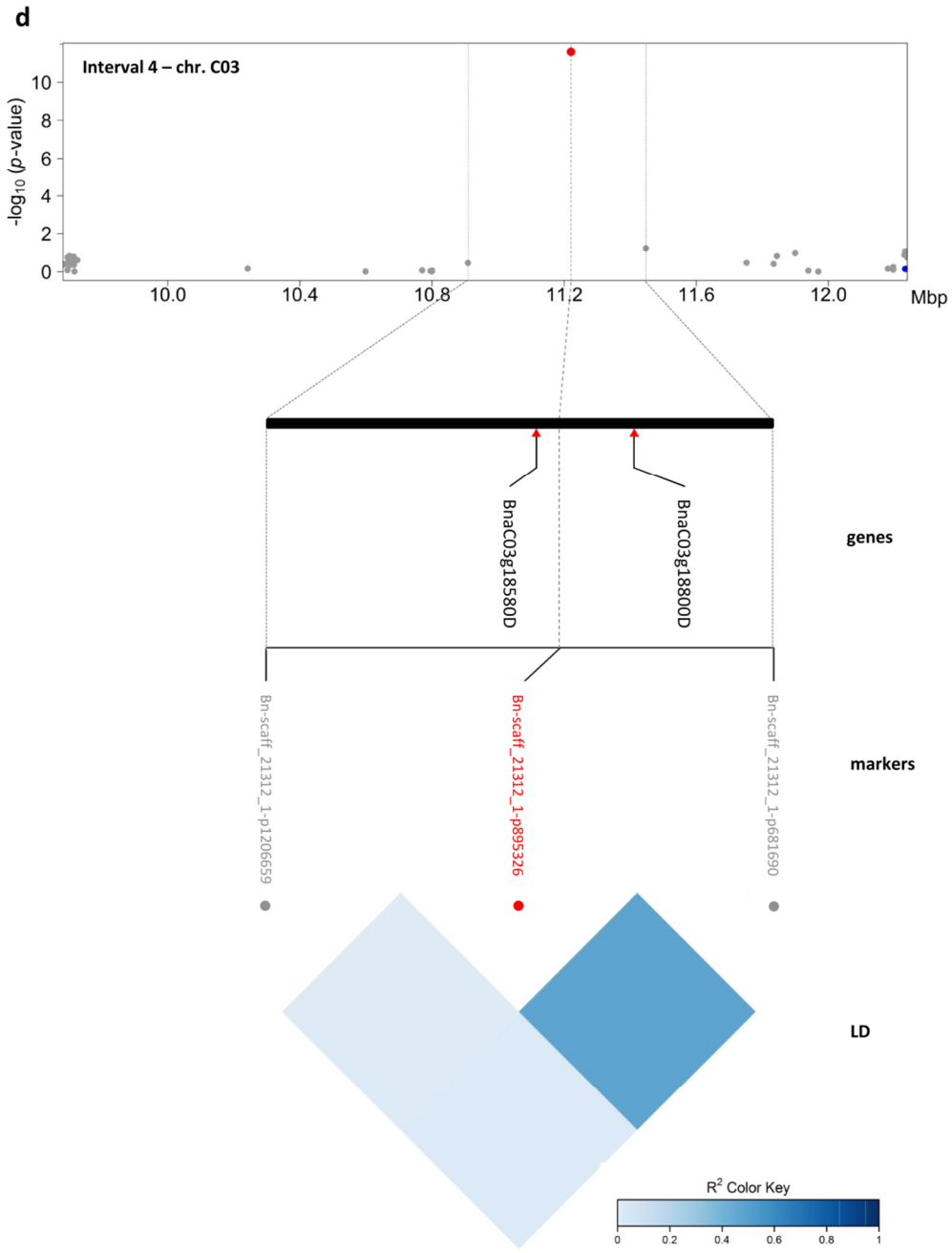

d
Interval 4 – chr. C03
-log10 (p-value)
10
8
6
4
2
0
10.0
10.4
10.8
11.2
11.6
12.0
Mbp
BnaC03g18580D
BnaC03g18800D
genes
Bn-scaff_21312_1-p1206659
Bn-scaff_21312_1-p895326
Bn-scaff_21312_1-p681690
markers
LD
R2 Color Key
0
0.2
0.4
0.6
0.8
1

Figure S26. Manhattan plots for representative associations in the candidate regions

The Manhattan plots describe genome-wide marker-trait associations for four of the five candidate regions (Table 6). **a** Interval 1 on chromosome A04. The trait 'estimated biovolume at 27 DAS' is shown as a representative trait for the 14 traits associated with the marker 'Bn-A04-p2218115'. **b** Interval 2 on chromosome A10. The trait 'estimated biovolume at 22 DAS' is shown as a representative trait for the 16 traits associated with the marker 'Bn-A10-p13343454'. **c** Interval 3 on chromosome C02. The trait 'colour uniformity at 27 DAS' is shown as a representative trait for the 9 traits associated with the marker 'Bn-scaff_16804_1-p178142'. The CNV markers are usually not in LD with the SNP markers and therefore disrupt the structure of the LD blocks. Hence, the two SNP markers 'Bn-scaff_16804_1-p178142' and 'Bn-scaff_16804_1-p203519' should be regarded as LD block. **d** Interval 4 on chromosome C03. The trait 'estimated biovolume at 21 DAS' is shown as a representative trait for the 16 traits associated with the marker 'Bn-scaff_21312_1-p895326'. The significantly associated SNPs are indicated by red dots. Grey and blue dots represent surrounding non-significant SNP and CNV markers, respectively. Please note that the FarmCPU GWAS method, which iteratively uses fixed and random effect models and pseudo QTN as covariates, results in a different appearance of the Manhattan plots. Significant associations are illustrated by 'helicopters' rather than 'skyscrapers', see materials and methods. For reasons of clarity and comprehensibility, the zoom-in of the candidate regions was extended to the next flanking SNP markers. Red triangles indicate the positions of selected candidate genes (Table 6). The LD heatmaps in the bottom sections show the correlations (r^2) between surrounding SNP markers.

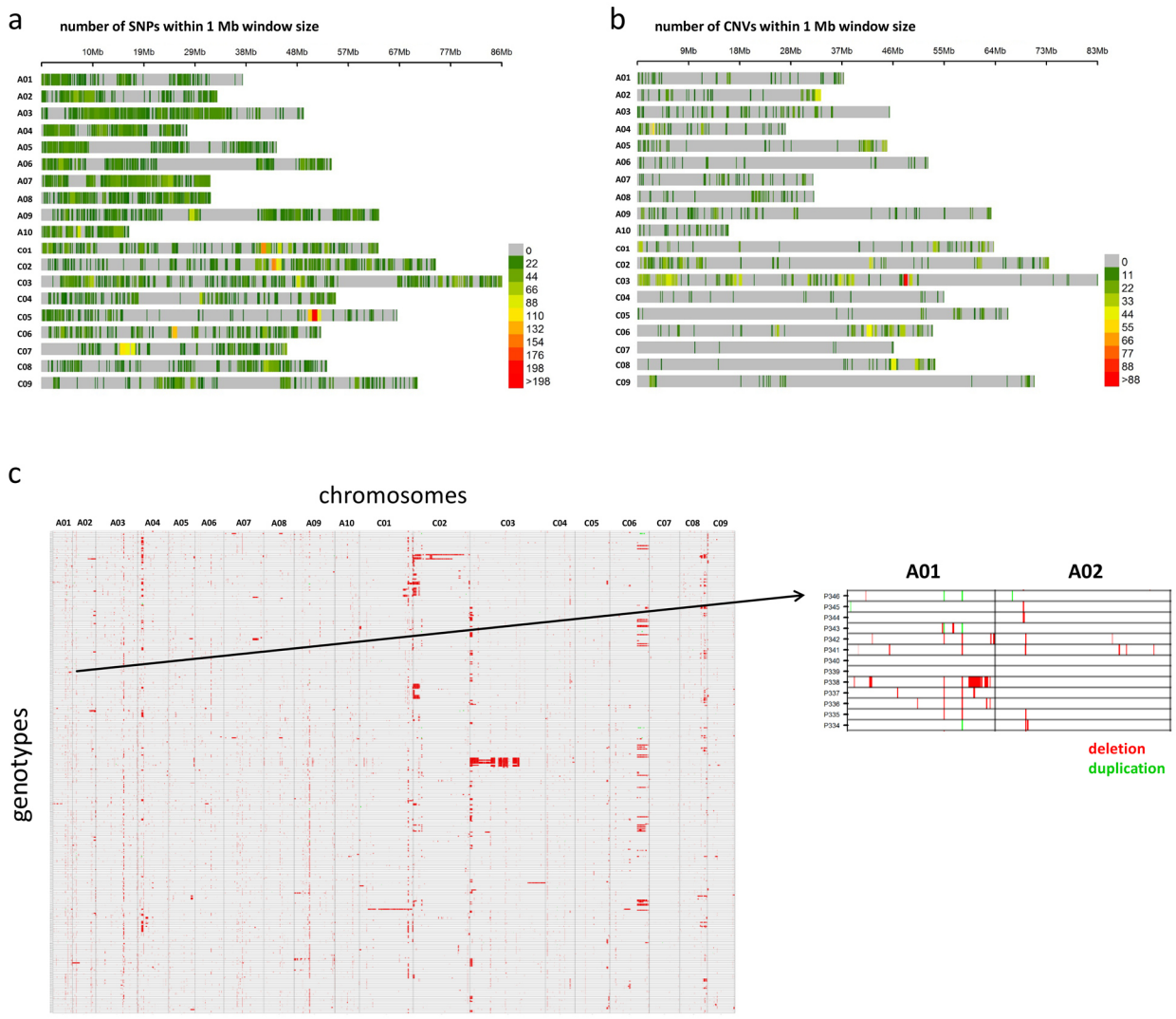


Figure S27. Overview of copy number variation polymorphisms (CNVs)

Subfigure **a** shows the genome-wide SNP marker distribution across the 19 *B. napus* chromosomes. 13,201 unique, single-copy SNPs were binned in 1 Mb intervals. Subfigure **b** shows the genome-wide CNV marker distribution. A total of 3,106 deletions and 4 duplications were binned in 1 Mb intervals. The marker density is indicated by the colour legend (green to red) on the right side. Grey colour indicates regions without SNPs. Subfigure **c** shows all detected deletions and duplications for the 477 spring-type canola genotypes, with an exemplary detailed zoomed-in region on the right side. Rows refer to the genotypes. Columns show chromosome-wise sorted markers. The colours red and green indicate deletion and duplication events, respectively.

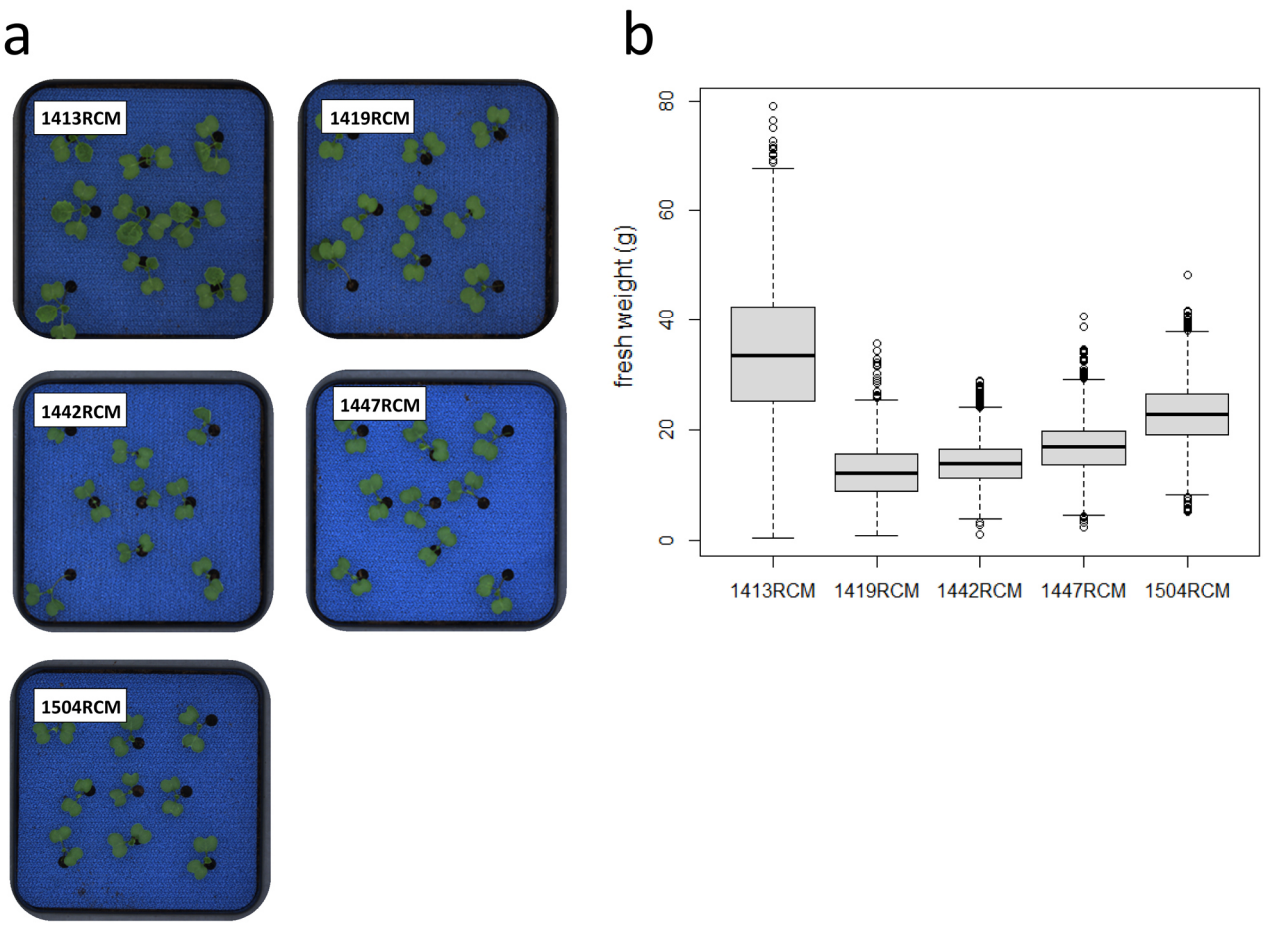


Figure S28. Differences in growth speed due to breakdown of cooling system

A technical failure of the cooling system in the glasshouse occurred during the first phenotyping experiment (1413RCM). The higher temperatures during the first days of plant growth resulted in a substantially increased developmental speed and higher end-point biomass of the plants in comparison to the other phenotyping experiments. This is indicated in the four photos in subfigure **a** showing typical images of the reference genotype 'Campino' from the four phenotyping experiments with the parental lines (1413RCM, 1419RCM, 1442RCM and 1447RCM) and the fifth experiment with the selected hybrids (1504RCM) at 13 DAS. Subfigure **b** displays the overall distribution of biomass (fresh weight; raw values in g) for all genotypes determined at 28 DAS for the five corresponding phenotyping experiment. As the plants in the fifth experiment comprise hybrids, they were expected to display larger biomass compared to plants of the other four experiments.

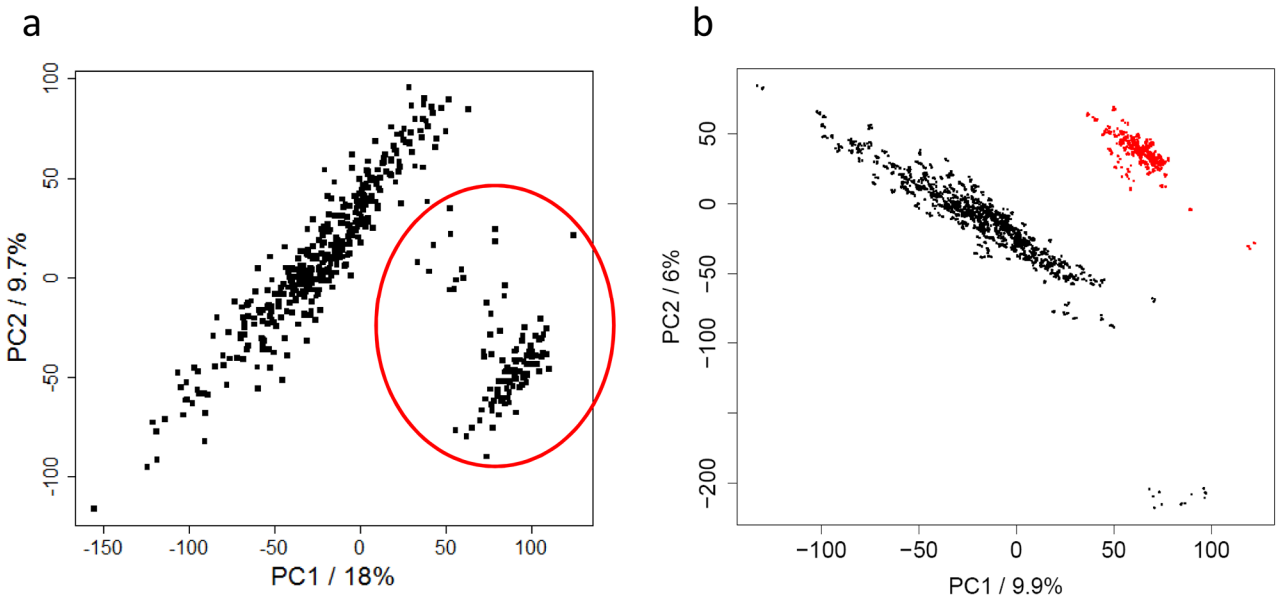


Figure S29. Bias in transcript data due to library preparation

Subfigure **a** shows the principal component analysis scatter plot for all 477 genotypes with concatenated transcriptome data from the individual sequencing runs. PCA was performed on filtered transcript data (median tpm ≥ 5). Transcript data were centred and scaled (z-scores). The PCA calculation was done by singular value decomposition (svd) of the data matrix. PC1 and PC2 explained 18 % and 9.7 % of variance, respectively. The red circle frames samples that separate from the main cluster of samples. The separation was not immediately explainable by any technical or biological grouping. Consequently, a second mapping and PCA analysis was performed using all generated data files from the different sequencing runs separately. As indicated in subfigure **b**, the separation of samples could be explained by systematic differences between batches during library preparation. The red coloured samples belong to the affected libraries (LG6 and LG7).

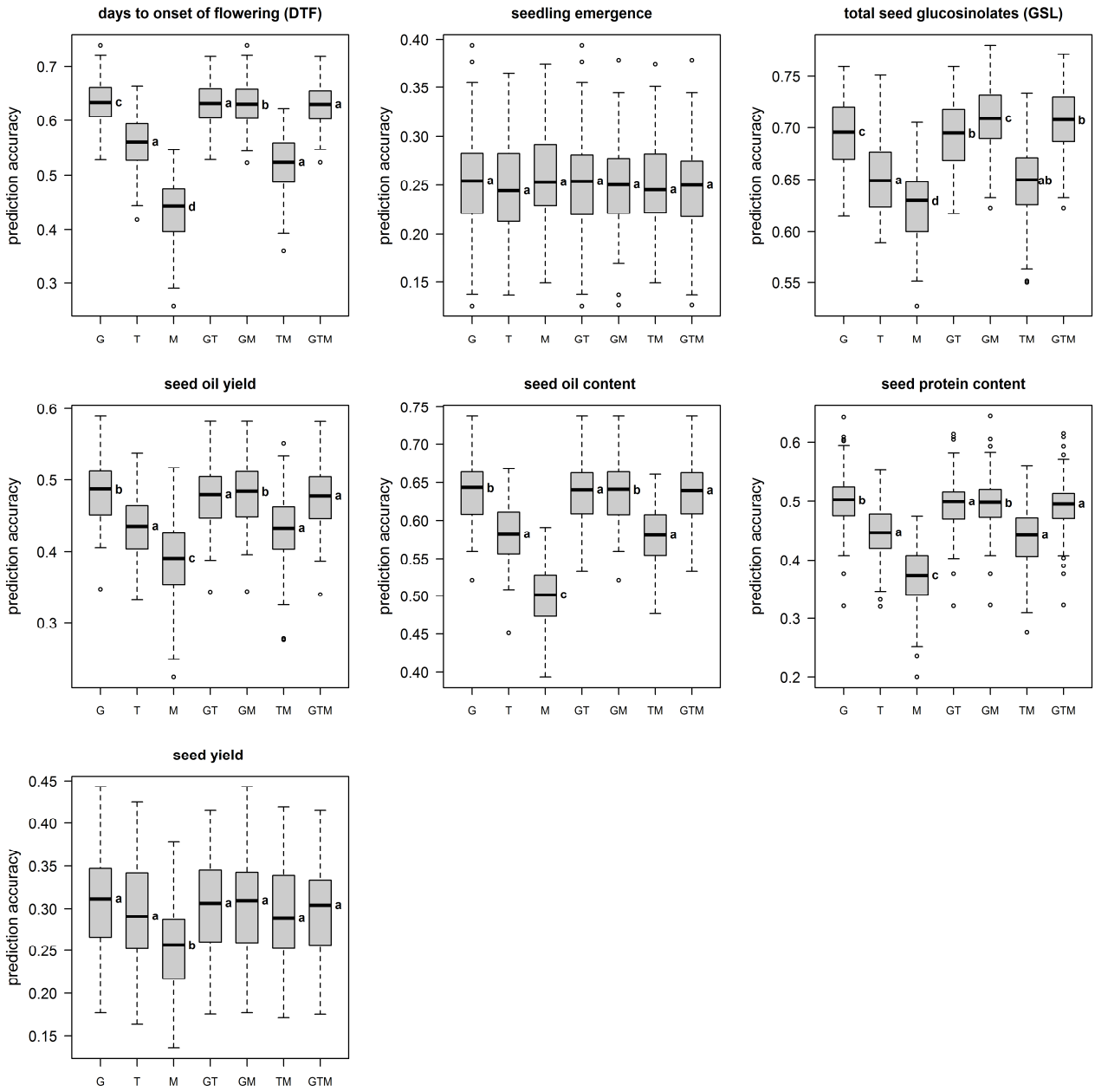


Figure S30. Prediction accuracies for the reduced data set

A summary of (genomic) best linear unbiased predictions (gBLUP) for the seven agronomic traits is given as boxplots. The analysis war performed with a reduced transcriptomics data set, excluding samples from two library batches (LG6 and LG7) and only 388 genotypes. The prediction accuracies were defined as the correlation between the true and the predicted phenotypic values. The different -omics data sets as predictors are denoted as: G, genomic data; T, transcriptomic data; M, metabolite data and their respective combinations G+T, G+M, M+T and G+T+M. Letters beside the boxes indicate significant differences between predictor sets determined by a one-way ANOVA followed by a post-hoc Tukey's multiple comparison test.

9. Acknowledgments

I am grateful to Andrea Apelt, Sibille Bettermann, Sandra Drießlein, Iris Fischer, Monika Gottowik, Beatrice Knüpfer, Marion Michaelis, Ingo Mücke, Alexandra Rech and Gunda Wehrstedt for excellent technical assistance. I thank Jean-Michel Pape for helpful discussions regarding optimisation of the IAP image analysis pipeline. I thank the ´NPZ Innovation GmbH´ and ´Deutsche Saatveredelung AG´, in particular Dr. Amine Abbadi, for providing seed material, population information and agronomic field data. I am grateful to my project partners from the Justus-Liebig University (JLU) Gießen, Dr. Fabian Grandke, Dr. Birgit Samans and Prof. Dr. Rod J. Snowdon who provided me the genotype data, the enhanced *Brassica napus* reference genome assembly and gene annotations. I thank Dr. Axel Himmelbach and the IPK sequencing facility for their support and Prof. Dr. Andrea Bräutigam for her helpful comments regarding the analysis of RNA-Seq data. My special thanks go to Dr. David Riewe for the help during the analytical work and support regarding the analysis of metabolites and to Dr. Christian R. Werner for help with the hybrid prediction analyses. I appreciate the inspiring discussions with Dr. Yong Jiang and Dr. Yusheng Zhao regarding statistics and Dr. Sebastian Beier regarding programming issues. Moreover, I thank in particular my colleagues Marc C. Heuermann and Dr. Rhonda C. Meyer for their helpful comments and discussions. I am especially thanking Dr. Renate H. Schmidt for the countless helpful comments and discussions. Finally, I thank Prof. Dr. Thomas Altman for the opportunity to work on such an interesting project, our discussions and the time and efforts he invested in my supervision and of cause the whole 'Heterosis' research group for the time together and the nice working atmosphere.

This work was conceived as a research project supported by the ´Deutsche Forschungs-gemeinschaft´ (DFG). DFG-GEPRIS: 'PREDICT: Omics-based models for prediction of hybrid performance in oilseed rape' (project number 234585441).

10. Curriculum vitae

Contact information:	Last name:	Knoch
	First name:	Dominic
	Address:	Lange Straße 13,
		06466 Seeland, OT Gatersleben
	Phone:	039482 5 809
	Email:	knochd@ipk-gatersleben.de
	ORCID:	0000-0002-9362-3105

Personal data:	Date of birth:	31.10.1989
	Place of birth:	06128 Halle (Saale)
	Citizenship:	German
	Sex:	male

Formal education:	1996-2000	Grundschule ‘Geschwister Scholl‘, Mücheln
	2000-2002	Sekundarschule ‘Adolf Holst‘ Mücheln/Krumpa
	2002-2005	Geiseltalgymnasium Mücheln
	2005-2008	Gymnasium Querfurt
	2008	Abitur

Scientific education:

2008-2011	Bachelor of Science Biologie (180 CP) degree course (Martin-Luther-University Halle-Wittenberg)
B.Sc. Thesis:	'Expression, Reinigung und Charakterisierung von Spinnenseiden-proteinmultimeren aus Tabak'
2011-2014	Master of Science Biology (120 CP) degree course (Martin-Luther-University Halle-Wittenberg)
M.Sc. Thesis	'Metabolic analysis of mature *Arabidopsis thaliana* seeds'
Since 2014	PhD Student at IPK Gatersleben

Publications:

Hauptmann V, Weichert N, Menzel M, Knoch D, Paege N, Scheller J, Spohn U, Conrad U and Gils M. (2013). Native-sized spider silk proteins synthesized in planta via intein-based multimerization. Transgenic Research **22**, 369–377.

Knoch D, Riewe D, Meyer RC, Boudichevskaia A, Schmidt R and Altmann T. (2017). Genetic dissection of metabolite variation in Arabidopsis seeds: evidence for mQTL hotspots and a master regulatory locus of seed metabolism. Journal of Experimental Botany **68**, 1655–1667.

Knoch D, Abbadi A, Grandke F, Meyer RC, Samans B, Werner CR, Snowdon RJ, Altmann T. (2019). Strong temporal dynamics of QTL action on plant growth progression revealed through high-throughput phenotyping in canola. Plant Biotechnology Journal. (**Epub ahead of print**)

_____ _____

Datum / *Date* Unterschrift des Antragstellers / *Signature of the applicant*

11. Declarations of academic integrity

Eidesstattliche Erklärung / *Declaration under Oath*

Hiermit erkläre ich, dass ich die Arbeit, welche ich einreiche, selbstständig und ohne andere als die angegebenen Quellen verfasst habe. Alle Gedanken, die direkt oder indirekt von externen Quellen stammen, werden ordnungsgemäß als solche gekennzeichnet. Diese Dissertation wurde weder an der Martin-Luther-Universität Halle-Wittenberg noch an einer anderen Universität eingereicht.

Hereby, I declare that I have composed the work that I will submit independently on my own and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such. This Dissertation has neither been previously submitted to the Martin-Luther-University Halle-Wittenberg nor to any other university before.

_____ _____

Datum / *Date* Unterschrift des Antragstellers / *Signature of the applicant*

Erklärung über bestehende Vorstrafen und anhängige Ermittlungsverfahren /
Declaration concerning Criminal Record and Pending Investigations

Hiermit erkläre ich, dass ich weder vorbestraft bin noch, dass gegen mich Ermittlungsverfahren anhängig sind.

I hereby declare that I have no criminal record and that no preliminary investigations are pending against me.

_____ _____

Datum / *Date* Unterschrift des Antragstellers / *Signature of the applicant*