Dissecting and Modeling the Phenotypic Components of Plant Growth and Drought Responses based on High-Throughput Image Analysis

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftliche Fakultät I
Biowissenschaften
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Herrn Dijun Chen
Geb. am 27.06.1983 in Huangshi, Hubei Province, China

Gutachter:
Prof. Dr. Thomas Altmann
Prof. Dr. Björn Junker
Prof. Dr. Björn Usadel

Verteidigt am: 22.03.2017 in Halle/Saale

"Perhaps a suitable analogy to explain the short-falls of Dawkins's account of evolution is to think of an oil painting. In this analogy Dawkins has explained the nature and range of pigments; how the extraordinary azure colour was obtained, what effect cobalt has, and so on. But the description is quite unable to account for the picture itself. This view of evolution is incomplete and therefore fails in its side-stepping of how information (the genetic code) gives rise to phenotype, and by what mechanisms. Organisms are more than the sum of their parts, and we may also note in passing that the world depicted by Dawkins has lost all sense of transcendence."

Simon Conway Morris, 1998

Dedication

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

Dissecting and Modeling the Phenotypic Components of Plant Growth and Drought Responses based on High-Throughput Image Analysis

Abstract

Recent technological advances and remarkable successes have led to high-throughput phenotyping becoming a tool of choice for quantifying the phenotypic traits or components of plant growth and performance. Efficient analysis and interpretation of huge and high-dimensional phenotypic data sets obtained from such studies remain enormous challenges due to lack of a standard analytical framework. In this thesis, I present a comprehensive framework for high-throughput phenotype data analysis in plants, which enables the extraction and dissection of a high-dimensional list of phenotypic traits from non-destructive plant imaging over time.

As a proof of concept, I first investigate the phenotypic components of the drought responses of 312 plants from 18 different barley cultivars during vegetative growth. I define a core set of 54 phenotypic traits that are highly reproducible and representative, and have greatly extended the trait list from previous studies. I further analyze dynamic properties of trait expression over growth time based on these phenotypic features. I observe that different trait groups show largely different patterns of genotype and environmental effects during plant growth. The data are highly valuable to understand plant development and to further quantify growth and crop performance features. I then test various growth models to predict plant growth patterns and identify several relevant parameters that support biological interpretation of plant growth and stress tolerance. These image-based traits and model-derived parameters are promising for subsequent genetic mapping to uncover the genetic basis of complex agronomic traits. Finally, several models are constructed to predict biomass from image-based features in three consecutive barley experiments. It is observed that plant biomass can be accurately predicted from image-based parameters using a random forest model. The prediction accuracy remains high across experiments. The relative contribution of individual feature from the model reveals new insights into the phenotypic determinants of plant biomass outcome.

Taken together, I anticipate that the analytical framework and analysis results presented in this thesis will be useful to advance our views of phenotypic trait components underlying plant development and their performance, and possess great potential applications in plant breeding under the context of phenomics.

Acknowledgements

It is my great pleasure to thank my supervisor, Dr. Christin Klukas (research group leader of Image Analysis, IPK), for his superb guidance and support throughout my entire PhD research, for entrusting me with great freedom and responsibility, and for his constructive criticism, constant encouragement, patient support and belief in my capabilities, sometimes going beyond my own expectations. I am grateful to Prof. Dr. Thomas Altmann (head of the Department of Molecular Genetics and research group leader of Heterosis, IPK), for being such a great model and his willing to serve as my official supervisor, and for his guidance and criticism on my thesis redaction. I am also grateful to my Thesis Advisory Committee members, Prof. Dr. Thomas Altmann, Prof. Dr. Jörg Degenhardt, Prof. Dr. Klaus Humbeck, Prof. Dr. Björn Junker, Prof. Dr. Wolfgang Sippl and Prof. Dr. Björn Usadel, for their generous interest in my work and for their correction and comments on my thesis.

I would like to thank our experimental collaborators, Dr. Kerstin Neumann and Dr. Benjamin Kilian (from Research Group Genome Diversity), who provided invaluable high-throughput phenotyping data, and without whom the research presented here would not have been possible. I would like to thank Dr. Swetlana Friedel (former research group leader of Data Inspection, IPK) provided assistance on plant growth modeling in Chapter 3. It was a wonderful experience to work with all members of Image Analysis Group, Ingo Mücke, Jean-Michel Pape and Michael Ulrich, who have been helpful and supportive throughout all this time and provided a stimulating and fun atmosphere.

This thesis has been a nice journey that gave me the opportunity to meet many crazy people during these years. They supported me in various way to the realisation of my thesis. I am indebted to all colleagues who have helped with advice and discussions during my degree and to everyone who helped with critical input and proofreading of this thesis.

Thanks to all the co-authors of Chapters 2 and 3: Dr. Kerstin Neumann, Dr. Swetlana Friedel, Dr. Benjamin Kilian, Prof. Dr. Ming Chen (from Zhejiang University, China), Prof. Dr. Thomas Altmann, and Dr. Christian Klukas for making it possible. Thanks to Dr. Rongli Shi (from Research Group Heterosis) and Jean-Michel Pape (from Research Group Image Analysis) for their help in the story of Chapter 4.

Further, I would like to thank the Federal Ministry of Education and Research (BMBF, 0315958A), and the EU funded project EPPN (Grant Agreement No. 284443) for funding my research in IPK, and the Robert Bosch Stiftung (32.5.8003.0116.0) and the Federal Agency for Agriculture and Food (BEL, 15/12-13, 530-06.01-BiKo CHN) for the ongoing financial and academic support during my degree.

I have greatly enjoyed my time in IPK-Gatersleben. The working atmosphere in the Institute is wonderful and I am grateful to everyone here for their warm friendship and company. I thank Zifeng Guo and Dr. Thorsten Schnurbusch (from Research Group Plant Architecture) for collaboration research. At the same time, I would like to thank all my Chinese friends at Gatersleben who make this place as at home. I am particularly grateful to Dr. Britt Leps for her nice efforts and friendly supports that made my daily life

in Gatersleben much more convenient during the past years.

Last, but not least, I am enormously grateful to my family and to all my friends from near and far, for having been understanding and supportive during these exciting but busy times. I am especially grateful to my dear wife for her amazing patience, constant support, love, and willingness to move far away from China while I pursued my degree, and my dear little son Richie whose sweetness makes me a happiest dad in the world. I thank my parents for encouraging me to follow my dreams and for everything they have done for me in the past years.

Afterword: It has been almost two years since I finished my PhD training at IPK. I would like to thank Prof. Dr. Kerstin Kaufmann (from Humboldt University) for providing me a "postdoctoral" position in her flower group (from May 2015 till now) even without my PhD degree. It is a really amazing experience from which I have learned a lot not only about novel knowledge but also about how to do better independent research.

Dijun Chen

March 20^{th}, 2015, Gatersleben

Update on March 6^{th}, 2017, Potsdam

Contents

Abstract . . . . . **iv**

Acknowledgements . . . . . **v**

Abbreviations . . . . . **xiii**

1 Introduction — current knowledge on high-throughput plant phenotyping and its applications **1**

1.1 Aim of the thesis 2

1.2 Structure of the thesis 3

1.3 High-throughput phenotyping in plants 3

1.3.1 High-throughput phenotyping facilities 10

1.3.2 Large-scale image processing and analysis 12

1.3.3 Applications of high-throughput plant phenotyping 15

1.3.4 A proposed general framework for high-throughput phenotyping data analysis . . . 17

1.4 Publications on which this thesis is based 18

2 Dissecting the high-dimensional phenotypic components of plant growth and drought responses **20**

2.1 Introduction 20

2.2 Results 21

2.2.1 Extraction of phenotypic traits from high-throughput image data 21

2.2.2 Image-derived parameters reflect drought stress responses 22

2.2.3 Plant phenomic map and phenotypic similarity 27

2.2.4 Phenotypic profile reflects global population structure 28

2.2.5 Dynamic genotypic and environmental effects on phenotypic variation 31

2.2.6 Change of heritability and trait-trait genetic and phenotypic correlations over growth time 36

2.3 Discussion 38

2.4 Materials and methods 39

2.4.1 Plant materials and growth conditions . . . 40
2.4.2 Image analysis . . . 41
2.4.3 Feature preprocessing . . . 41
2.4.4 Feature selection . . . 42
2.4.5 Hierarchical clustering analysis and PCA . . . 42
2.4.6 Phenotypic similarity tree and Mantel test . . . 43
2.4.7 Plant classification using SVM . . . 43
2.4.8 Analysis of phenotypic variance . . . 44
2.4.9 Broad-sense heritability . . . 45
2.4.10 Estimation of genetic and phenotypic correlations . . . 45

3 Plant growth modeling based on time-lapse image data . . . 46
3.1 Introduction . . . 46
3.2 Results . . . 47
3.2.1 Modeling barley plant growth under normal conditions . . . 47
3.2.2 Modeling barley plant growth under drought stress conditions . . . 52
3.2.3 Model-derived parameters describing plant growth patterns and performance . . . 55
3.2.4 Growth modeling of a worldwide collection of maize plants . . . 57
3.3 Discussion . . . 60
3.4 Materials and methods . . . 63
3.4.1 Plant image data . . . 63
3.4.1.1 High-throughput phenotyping of a worldwide set of maize plants . . . 63
3.4.2 Image analysis . . . 64
3.4.3 Plant growth modeling . . . 64
3.4.4 Trait repeatability . . . 65

4 Prediction of plant biomass accumulation based on image-derived parameters . . . 66
4.1 Introduction . . . 66
4.2 Results . . . 67
4.2.1 Development of statistical models for modeling plant biomass accumulation using image-derived features . . . 67
4.2.2 Coordinate patterns of plant image-based profiles and their biomass output . . . 69
4.2.3 Relating image-based signals to plant biomass output . . . 71
4.2.4 Contribution of different image-based features to predicting plant biomass . . . 73
4.2.5 Image-based features are predictive of plant biomass across experiments with similar conditions or treatments . . . 76

4.3 Discussion . . . 79
4.4 Materials and methods . . . 80
4.4.1 Germplasm and experiments . . . 80
4.4.2 Image analysis . . . 81
4.4.3 Feature selection . . . 81
4.4.4 Data transformation . . . 81
4.4.5 Hierarchical clustering analysis and PCA . . . 81
4.4.6 Models for predicting plant biomass . . . 82
4.4.7 Evaluation of the prediction models . . . 82

5 Summary and outlook . . . 84
5.1 Summary . . . 84
5.2 Outlook . . . 85

References . . . 86

Appendix A Glossary . . . 100

Appendix B Supplemental Tables . . . 103

Appendix C Online Resources . . . 108

Appendix D Curriculum Vitae . . . 109

List of Tables

1.1 Key imaging techniques used in high-throughput plant phenotyping. 6

1.2 Automated or semi-automated plant phenotyping platforms. 7

1.3 Plant phenomics community. 13

2.1 Overview of 18 barley genotypes used in this study. 25

3.1 Mechanistic models used for modeling biomass accumulation in this study. 49

3.2 Calculation of absolute growth rate and relative growth rate. 50

3.3 Summary of growth model-derived parameters. 59

4.1 Overview of three barley experiments. 67

S1 The 54 investigated phenotypic traits in barley. 103

S2 A worldwide collection of maize plants selected from from IPK Genebank. 106

List of Figures

1.1 The genotype-phenotype map (G-P map) . . . 4
1.2 The spectral regions . . . 10
1.3 High-throughput phenotyping infrastructure . . . 11
1.4 A global stronghold of high-throughput phenotyping facilities . . . 12
1.5 The typical workflow of a image-processing pipeline . . . 14
1.6 IAP: integrated analysis platform . . . 16
1.7 A comprehensive framework for high-throughput phenotyping in plants . . . 18

2.1 Experimental design for high-throughput phenotyping in barley . . . 22
2.2 Pipeline for analysis of high-throughput phenotyping data in barley . . . 23
2.3 Reproducibility of phenotypic traits . . . 24
2.4 Assessment of trait reproducibility analysis . . . 25
2.5 Trait similarity . . . 27
2.6 Phenotypic traits revealing the stress symptom . . . 28
2.7 Classification of plants based on the SVM methodology . . . 29
2.8 Phenotypic similarity revealed by genotype similarity . . . 30
2.9 Phenotypic profile reflects global population structures in the temporal scale . . . 32
2.10 PCA performed over time . . . 33
2.11 PCA performed on control and stressed plants, respectively . . . 34
2.12 Dissection of the sources of phenotypic variance . . . 35
2.13 Trait heritability and trait-trait genetic and phenotypic correlations . . . 37

3.1 Correlation analysis of manual measurements with phenotypic traits . . . 48
3.2 Growth modeling of barley plants under normal conditions . . . 51
3.3 Evaluation of the performance of growth curves for control plants . . . 53
3.4 Growth modeling of barley plants under drought stress conditions . . . 54
3.5 Evaluation of the performance of growth curves for stressed plants . . . 55
3.6 Comparison of stress elasticity and several drought tolerance indexes . . . 56
3.7 Experimental design for high-throughput phenotyping of a worldwide collection of maize plants . . . 58

3.8 Evaluation of the performance of growth modeling for maize plants 59
3.9 Growth modeling of maize plants 61

4.1 Modeling pipeline for predicting plant biomass accumulation based on image-derived parameters 68
4.2 Predictability of image-based traits to plant biomass 70
4.3 Quantitative relationship between image-based features and plant biomass 72
4.4 The relative importance of image-based features in prediction of plant biomass 74
4.5 The relative importance of image-based features in prediction of biomass in control plants 75
4.6 The relative importance of image-based features in prediction of biomass in stressed plants 76
4.7 Comparison of prediction accuracy across different experiments 77
4.8 Comparison of prediction accuracy across different treatments 78

A.1 Histogram bin-based feature extraction in different color spaces 102

Abbreviations

AGR	**a**bsolute **g**rowth **r**ate
ANOVA	**an**alysis **o**f **va**riance
cv.	**c**ulti**v**ars
CV	**c**oefficient of **v**ariation
DAS	**d**ays **a**fter **s**owing
DH	**d**ouble **h**aploid
DW	**d**ry **w**eight
FC	**f**ield **c**apacity
FLUO	**fluo**rescence
FW	**f**resh **w**eight
GWAS	**g**enome-**w**ide **a**ssociation **s**tudies
G×E	**g**enotype × **e**nvironment interaction
HCA	**h**ierarchical **c**luster **a**nalysis
HTP	**h**igh-**t**hroughput **p**henotyping
IAP	**i**ntegrated **a**nalysis **p**latform
LMM	**l**inear **m**ixed **m**odel
MARS	**m**ultivariate **a**daptive **r**egression **s**plines
MLR	**m**ultivariate **l**inear **r**egression
NIR	**n**ear-**i**nfra**r**ed
PCA	**p**rinciple **c**omponent **a**nalysis
PCC	**P**earson **c**orrelation **c**oefficient
QTL	**q**uantitative **t**rait **l**oci
REML	**re**sidual **m**aximum **l**ikelihood
RF	**r**andom **f**orest

RGR	**r**elative **g**rowth **r**ate
RMSRE	**r**oot **m**ean **s**quared **r**elative **e**rror
SOM	**s**elf-**o**rganizing **m**ap
SVM/SVR	**s**upport **v**ector **m**achine/**r**egression

Chapter 1

Introduction — current knowledge on high-throughput plant phenotyping and its applications

In the coming decades, crop production must be significantly increased to meet the predicted production demands of the global population that is expected to grow to more than 9 billion by 2050 under changing climates[1] (Tilman et al., 2011). However, achieving this goal will be a tremendous challenge for plant scientists and breeders because the average rate of crop production increase (1.3% per year) cannot keep pace with the expected demands (2.4% per year) (Ray et al., 2013, 2012). But at the same time, extensive breeding and agronomic efforts provide potential to select and breed high yielding and stress-tolerant plants far more rapidly and efficiently than is currently possible (Pingali, 2012). High-throughput genotyping platforms support the discovery and analysis of genome-wide genetic markers (genotypes) in populations in a routine manner (Davey et al., 2011; Edwards et al., 2013), offering the potential to increase the rate of genetic improvement (Phillips, 2010). However, our capabilities for systematic assessment and quantification of plant phenotypes have not kept pace (Furbank and Tester, 2011; Houle et al., 2010), limiting our ability to dissect genetic basis underlying plant growth, yield and adaptation to stress (Araus and Cairns, 2014). Commonly used conventional phenotyping procedures are labor-intensive, time-consuming, lower-throughput and costly, and frequently destructive to plants (e.g. fresh or dry weight determination), whereas measurements are often taken at certain times or at particular developmental stages, a scenario known as the "phenotyping bottleneck" (Furbank and Tester, 2011).

Recently, the introduction of techniques for high-throughput phenotyping (HTP) has boosted the area of plant phenomics, where new technologies such as non-invasive imaging, spectroscopy, robotics and high-performance computing are combined to capture multiple phenotypic values at high resolution, high precision, and in high throughput. This will ultimately enable plant scientists and breeders to conduct numerous phenotypic experiments in an automated format for large plant populations under different environments to monitor non-destructively the performance of plants over time (Eberius and Lima-

[1] http://www.unpopulation.org/

Guerra, 2009). Various automated or semi-automated high-throughput plant phenotyping platforms have been recently developed and are applied to investigate plant performance under different environments (Arvidsson et al., 2011; Biskup et al., 2009; Golzarian et al., 2011; Granier et al., 2006; Jansen et al., 2009; Nagel et al., 2012; Walter et al., 2007). The huge amounts of image data routinely accumulated in these platforms need to be efficiently managed, processed and finally mined and analyzed. Thus, we are now facing the "big data problems" (Schadt et al., 2010) brought about by such real-time imaging technologies in the phenomics era. Consequently, the major challenge for image analysis is the automated extraction of important phenotypic parameters to be used in genetic analyses (such as association mapping), in breeding (efficient phenotypic selection), or in industrial screening (e.g. large collections of transgenic or genetically modified plants).

1.1 Aim of the thesis

HTP has been subjected to development for over ten years and technical advancements in HTP make the system-wide quantifying of plant phenomics feasible. Several studies have been applied to study very specific aspects of plant phenomics based on several well-investigated phenotypic traits from traditional phenotyping approaches. These studies have clearly shown that HTP is an ideal replacement of traditional phenotyping in plants. However, a comprehensive investigation of plant phenomics as well as its dynamics and performance based on an extended list of phenotypic traits is still missing. The general aim of this thesis is to close this gap. More specifically, I aim to investigate the phenotypic components and dynamics of plant growth and drought responses based on high-dimensional phenotypic trait analyses, and to elucidate the relationship between plant biomass and image-derived parameters. The following questions will be addressed in this thesis:

1. How many informative phenotypic traits can be extracted from a HTP experiment?
2. How about the dynamics nature of these informative traits during plant growth?
3. Can HTP data be used to model plant growth?
4. Which parameters are important to determine plant growth?
5. Which image-derived parameters can be used to describe plant performance, such as drought responses?
6. To what extent that image-derived parameters are predictive of complex phenotype, such as plant biomass?
7. How about the heritability of these candidate phenotypic traits?

Answering these questions will definitely enhance our view of HTP application for the dissection of plant growth and performance. For a first impression herein, I have provided and used a general analytical framework for dissecting and modeling of HTP data in plants. I made a comprehensive analysis of a high-dimensional list of phenotypic traits extracted from huge image datasets. I characterized and

compared the growth patterns of different plant species based on time-lapse image data. I showed that plant performance can be solely predicted from image-derived parameters, shedding light on several novel traits of importance underlying plant growth. Overall, the methods and results presented in the thesis will provide new starting points for future works addressing crop improvement.

1.2 Structure of the thesis

In the present thesis, the introductory chapter provides a general overview of the currently developed HTP infrastructures, an introduction of existing image processing pipelines designed for HTP data analysis, a brief outlook of emerging applications based on HTP.

The results section consists of three independent chapters:

- ✔ Chapter 2 describes various strategies used for high-dimensional phenotypic trait analysis in barley (*Hordeum vulgare*);
- ✔ Chapter 3 presents growth modeling of barley and maize (*Zea mays*) plants based on time-lapse image data;
- ✔ Chapter 4 shows how to use image-derived parameters to predict plant performance in barley and maize;
- ✔ Chapter 5 encompasses brief remarks, conclusions and an outlook on future research in HTP data analysis.

1.3 High-throughput phenotyping in plants

Creation of the desirable phenotype is the ultimate goal of crop improvement. The term phenotype includes the ensemble of an organism's observable traits or characteristics such as its morphological, developmental, physiological, pathological or biochemical properties, phenology and behaviour that can be monitored, quantified, and/or visualized by some technical procedure (Mahner and Kary, 1997; Varki et al., 1998). Phenotypes are always results of the expression of genetic constitution under the influence of environmental factors. Phenomics is defined as the study of all the phenotypes of an organism (phenome) that are result of genetic code (G), environmental factors (E) and their interactions (G×E). In contrast to genotypes, which are essentially single one-dimensional as merely determined by the linear DNA code, phenotypes are usually multi-dimensional and are frequently capricious in different spatial and temporal situations. An important field of research today is trying to improve, both qualitatively and quantitatively, the capacity to measure phenomes. We have relatively well developed technologies of measurements, *in vivo* or in destructive manners, of physiological states and other *internal phenotypes* (endophenotypes), such as gene expression, protein and metabolite levels, whereas our ability to measure *external phenotypes* (exophenotypes) is rapidly evolving.

We will never be able to come even close to a complete characterized of the phenome due to its highly dynamic and high-dimensional properties. However, increasing the quantitative information obtained by phenotypic measurements is an important goal for phenomics (Houle et al., 2010). Phenotypic variation, a fundamental prerequisite and the perpetual force for evolution by natural selection, results from the complex interactions between genotype and environment (G×E). Phenome-wide data are essential and necessary for enabling us to trace causal links in the genotype-phenotype map (G-P map Waddington, 1968) as they define the space of all possible phenotypes (P space; Figure 1.1).

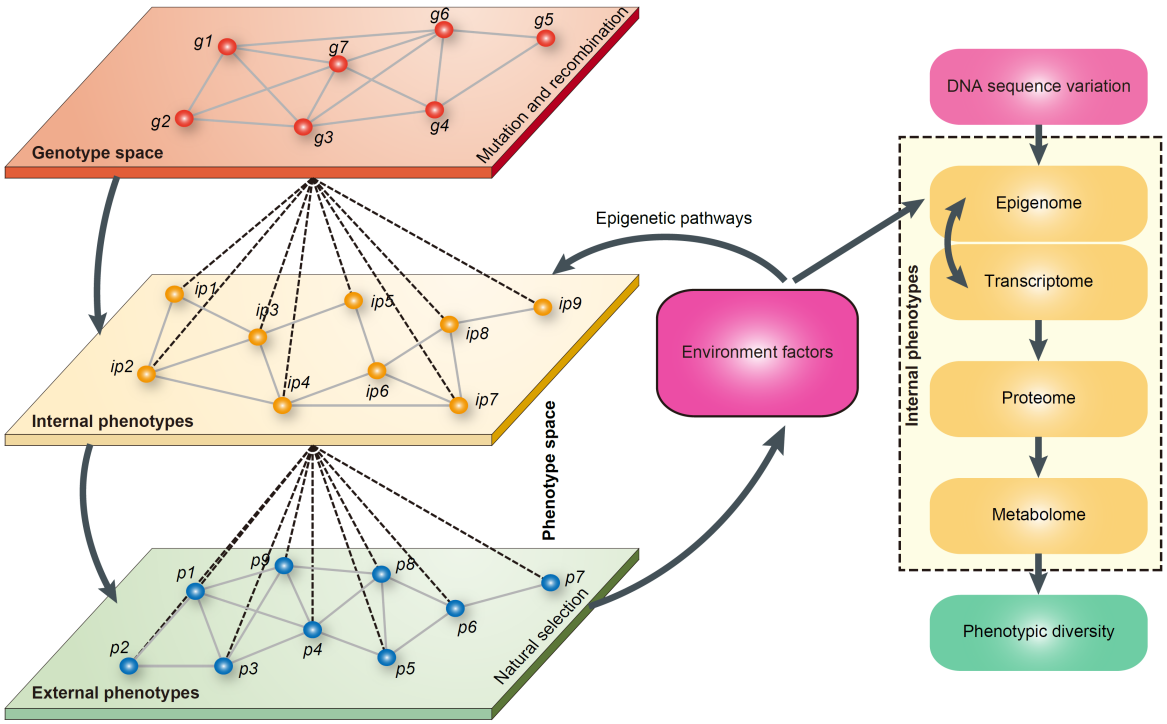

Figure 1.1: The genotype-phenotype map (G-P map)

The left panel shows the relationship of the genotype space (G space) and the phenotype space (P space) (Houle et al., 2010). The corresponding information that transmits from G space to P space is shown in the right panel. Genotypes could gain mutation and recombination over generations. Phenotypes can be broadly classified into internal and external phenotypes. These internal phenotypes include properties from molecular, cellular or tissue levels, which in turn shape external phenotypes such as morphology and behaviour. Upon the environmental stimuli, the epigenetic process creates the phenotypes using genotype information. External phenotypes can in turn shape the environment that an individual occupies, creating complex feedback relationships between genes, environments and phenotypes. Natural selection act in the P space to change the average phenotype of parents away from the average phenotype of the generation. The importance of the environment suggests that we should explicitly broaden the G-P map to the genotype-environment-phenotype (G-E-P) map. *g*: genotype; *p*: phenotype; *ip*: internal phenotype. This figure was taken from Chen et al. (2014a). ■

Plant phenotyping is intended to measure complex traits related to growth, yield and adaptation to stress with a certain accuracy and precision at different scales of organization, from organs to canopies. High-throughput automated imaging is the ideal tool for plant phenomic studies, which enables comprehensive and quantitative measurement of plant phenotypes in terms of extensive phenotyping (measuring

more phenotypic information at the same time) and intensive phenotyping (characterizing a phenotype in great detail, e.g., at population-wide and through plant growth cycle) (Houle et al., 2010). Owing to the recent increased availability of high-precision robotic handling machinery, many imaging-based technologies that span molecular to organismal spatial scales have been or are being established and enable us to extract multiparametric phenotypic information in great details. Generally, these noninvasive methods can be used to measure plant phenotypes related to growth and performance by the way to look over a range of the electromagnetic spectra far beyond human vision (Figure 1.2; Table 1.1; reviewed in Berger et al. (2010), Fiorani and Schurr (2013) and Li et al. (2014)).

For example, visible imaging is used to mimic human perception to provide information regarding plant growth and development features (Fiorani and Schurr, 2013; Li et al., 2014), including shoot biomass and morphology (Arvidsson et al., 2011; Golzarian et al., 2011; Jansen et al., 2009; Leister et al., 1999; Tackenberg, 2007), yield traits (Duan et al., 2011), panicle traits (Ikeda et al., 2010), imbibition and germination rates (Dias et al., 2011), leaf morphology (Bylesjo et al., 2008; Hoyos-Villegas et al., 2014; Weight et al., 2008), seedling growth (Walter et al., 2007, 2012), seed morphology (Chern et al., 2007; Joosen et al., 2012), root architecture (Clark et al., 2011; Iyer-Pascuzzi et al., 2010) and stress tolerance (Berger et al., 2010; Golzarian et al., 2011; Rajendran et al., 2009). Fluorescence imaging offers a rapid way to detect plant photosystem II status *in vivo* (Baker, 2008; Maxwell and Johnson, 2000) and is widely used in monitoring the effects of plant pathogens/disease (Balachandran et al., 1997; Bürling et al., 2010; Chaerle et al., 2004; Lohaus et al., 2000; Rolfe and Scholes, 2010; Scholes and Rolfe, 2009; Swarbrick et al., 2006) and early stress responses to *abiotic* and *biotic* factors (Baker, 2008; Berger et al., 2010; Chaerle et al., 2007a,b; Chen et al., 2014b; Harbinson et al., 2012; Jansen et al., 2009; Konishi et al., 2009; Lenk et al., 2007; Woo et al., 2008), and other physiological phenomena that are related to photosynthesis status. Near-infrared (NIR) imaging (900˜1700 nm spectral range) can be used to study leaf and canopy water status (Seelig et al., 2008, 2009), as water has highly absorbing bands between 1450 and 1550 nm. This technique can thus used to detect drought stress (Berger et al., 2010; Chen et al., 2014b; Harshavardhan et al., 2014; Munns et al., 2010; Saint Pierre et al., 2012), although the exploitation of NIR imaging is still in its infancy. Thermal infrared (IR) imaging (8˜14 μm spectral range) can be used to measure leaf and canopy temperature to study stomatal conductance (Jones et al., 2009), allowing a reliable way to detect changes in the physiological status of plants in response to biotic or abiotic stress (Li et al., 2014). In practice, IR imaging has successfully been used in real breeding programs to select traits for drought resistance in dry environments (Fiorani and Schurr, 2013).

Table 1.1: Key imaging techniques used in high-throughput plant phenotyping.

Imaging techniques†	**Principle**	**Targeted traits**	**Applications**
RGB/visible light [C,F]	The RGB (visible light) camera can be used to measure visible (VIS) reflectance having a wavelength in a range of 390 nanometres (nm) to 750 nm, resulting in gray or colour value images.	Image-based projected area / volume, dynamics growth, colour, shape / architecture / morphology descriptors	This imaging technique can be used to asses plant growth status, biomass accumulation, nutritional status or health status (Camargo et al., 2014; Golzarian et al., 2011; Yang et al., 2014).
Near-infrared [C]	The near-infrared (NIR) sensor uses non-visible light components in the NIR region of the spectrum (900˜1700 nm), resulting in gray images.	Plant characteristics such as moisture content (related to water status, maturity or ripeness)	This imaging technique allow to detect drought stress (Chen et al., 2014b; Harshavardhan et al., 2014), and can also be applied to study water movement in soil (e.g., root's water extraction efficiency).
Fluorescence [C,F]	Through fluorescence cameras, any fluorescence excitable by blue light with sufficient emission (420˜500 nm) can be captured both in 2D and 3D systems under backlight or reflective conditions. It offers a fleet way to probe photosystem II status *in vivo*.	Chlorophyll and other fluorophores signal, plant health/disease status	Chlorophyll fluorescence imaging is used as a diagnostic tool in plant physiology studies, such as detection of photosynthetic activity and stress responses (Chen et al., 2014b; Fiorani and Schurr, 2013; Hairmansis et al., 2014).
Infrared [C,F]	Infrared (IR) cameras use light in the thermal infrared region of the spectrum (8˜14 μm).	Leaf and canopy temperature and insect infestation	IR imaging provide a novel technique to measure the leaf or canopy temperature and thus to assess plant transpiration rate under highly controlled conditions (Jones et al., 2009; Munns et al., 2010).
3D imaging [C,F]	Stereo camera systems; laser scanning instruments with widely different ranges, time-of-light cameras.	Shoot structure,leaf angle distributions, canopy structure, root architecture	3D imaging has been used to measure structural parameters in various plant species (Biskup et al., 2007; Busemeyer et al., 2013a; Klose et al., 2009; van der Heijden et al., 2012).
CT [C]	X-ray computed tomography (CT) and X-ray digital radiography, a technology to produce tomographic images of specific areas of a scanned object, allowing to see inside the object without cutting.	Morphometric parameters in 3D, tillers, and grain quality	CT imaging has been used to measure tiller numbers and grain quality in rice (Yang et al., 2014), and cereal 3D root analysis (Flavel et al., 2012).

Table 1.1 (continued)

MRI [C]	Magnetic resonance imaging is able to visualize plant internal structures and metabolites.	Morphometric parameters in 3D, water content	MRI can be used to study plant physiology and metabolism "in vivo" (Borisjuk et al., 2012; Granier and Vile, 2014), and 3D root analysis (Hillnhütter et al., 2011; Rascher et al., 2011)
PET [C]	Positron emission tomography; positron emission detectors for short-lived isotopes.	Water transport, sectorality, flow velocity	PET is used to visualize distribution and transportation of radionuclide-labelled tracers involved in metabolism related activities (Granier and Vile, 2014; Jahnke et al., 2009).

† Techniques are currently used in controlled (C) or field (F) environments.

This table was adapted from Fiorani and Schurr (2013), Araus and Cairns (2014) and Li et al. (2014).

Table 1.2: Automated or semi-automated plant phenotyping platforms.

Name	Description	Reference
Controlled environment-based phenotyping platforms†		
GlyPh	A low-cost, automatic platform for high-throughput measurement of plant growth and water use in soybean (*Glycine max*). GlyPh allows the evaluation of up to 120 plants growing in individual pots.	(Pereyra-Irujo et al., 2012)
GROWSCREEN	An in-house system used in the Jülich Plant Phenotyping Centre (JPPC) to study leaf growth and fluorescence and root architecture in large plant populations. GROWSCREEN 3D is a pioneered solution developed for 3D analysis of leaves in tobacco (*Nicotiana tabacum*). It enables more accurate measurements of leaf area and extraction of additional volumetric traits.	(Biskup et al., 2009; Jansen et al., 2009; Nagel et al., 2012; Walter et al., 2007); http://www2.fz-juelich.de/icg/icg-3/jppc/growscreen/
GROW Map	Setup for monitoring of leaf/root growth via digital image sequence processing at JPPC	http://www.fz-juelich.de/ibg/ibg-2/EN/methods_jppc/methods_node.html

Table 1.2 (continued)

HRPF	High-throughput rice (*Oryza sativa*) phenotyping facility (HRPF) designed with two main section: rice automatic phenotyping (RAP) and yield trait scorer (YTS). This high-throughput platform developed for automatic screening rice germplasm resources and populations throughput the growth period and after harvest.	(Yang et al., 2014)
LemnaTec Scanalyzer	An robotic greenhouse system that uses non-destructive imaging to monitor plant growth under fully controlled conditions in high-thoughput. The LemnaTec platform aims to visualise and analyse the biology beyond human vision through imaging automatisation.	(Arvidsson et al., 2011; Brien et al., 2013; Camargo et al., 2014; Chen et al., 2014b; Golzarian et al., 2011; Hairmansis et al., 2014; Harshavardhan et al., 2014; Honsdorf et al., 2014; Junker et al., 2015); http://www.lemnatec.com/
Plant Scan	A novel automated screening platform and mesh-based technique developed for high-throughput 3D plant analysis. It was initially used for the analysis of aerial-parts in cotton (*Gossypium hirsutum*) and demonstrated highly accurate when comparing with with manual measurement data.	(Paproki et al., 2012); http://www.csiro.au/Outcomes/Food-and-Agriculture/HRPPC/PlantScan.aspx
Phenodyn	An platform to measures growth rate and transpiration rate every minute, together with environmental conditions (current throughput: 480 plants).	(Sadok et al., 2007); http://bioweb.supagro.inra.fr/phenodyn/
PHENOPSIS	An automated platform developed by Optimalog (France) for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis (*Arabidopsis thaliana*). The PHENOPSIS platform allows to weight, irrigate precisely and take a picture of more than 500 individual plants in rigorously controlled conditions.	(Granier et al., 2006); http://bioweb.supagro.inra.fr/phenopsis/
Phenoscope	This automated phenotyping platform is an integrated device allowing simultaneous culture of 735 individual Arabidopsis plants and high-throughput acquisition, storage and analysis of quality phenotypes.	(Tisne et al., 2013); http://www.observatoirevegetal.inra.fr/observatoirevegetal_eng/Scientific-platforms/Phenoscope
QubitPhenomics	Qubit Systems provides Conveyor and Robotic PlantScreen^{TM} Systems for plant phenomics analysis. The conveyor system can be configured for single pots, multiple pots or trays, providing flexibility of use with numerous different species, or with a single species throughout its growth cycle.	http://qubitphenomics.com/
TraitMill	A high-thoughput gene engineering system developed by CropDesign that enables large-scale plant transformation and automated high resolution phenotypic evaluation of crop performance in rice.	(Reuzeau, 2007; Reuzeau et al., 2005); http://www.cropdesign.com/tech_traitmill.php

Table 1.2 (continued)

WIWAM	Similar to PHENOPSIS, WIWAM is an automated imaging platform handling a large number of plants simultaneously and measuring a variety of plant growth parameters with automatic watering and imaging system at regular time intervals	(Skirycz et al., 2011); http://wiwam.be/
Field-based phenotyping platforms[§]		
BreedVision	A multi-sensor field-based phenotyping platform for small grain cereals. BreedVision has been applied to measure various agronomic traits in triticale.	(Busemeyer et al., 2013a,b; Liu et al., 2014; Würschum et al., 2014)
PhenoField	A mobile multispectral imaging platform for precise field phenotyping. The PhenoField system has been used to study canopies in wheat (*Triticum* spp.).	(Svensgaard et al., 2014) http://www.plantphenomics.org.au/services/phenomobile/
Phenomobile	Phenomobile was developed at the High Resolution Plant Phenomics Centre, Canberra and is a multispectral imaging platform	(Deery et al., 2014)
Pheno-Copter	A high-throughput field-based phenotyping system. Pheno-Copter was applied to study ground cover in sorghum, canopy temperature in sugarcane and three-dimensional measures of crop lodging in wheat.	(Chapman et al., 2014)
NA	A plant phenotyping system during field deployment in Maricopa, Arizona. Three types of sensors were deployed for measuring plant canopy height, temperature and reflectance in cotton.	(Andrade-Sanchez et al., 2014)
NA	A semi-automatic system for high throughput phenotyping wheat cultivars in-field conditions. Four identical spectrometers and two digital cameras were deployed.	(Comar et al., 2012)

[†] This part was adapted from Chen et al. (2014a).

[§] More field-based phenotyping platforms were reviewed in White et al. (2012) and Deery et al. (2014). *NA*: official name is not available.

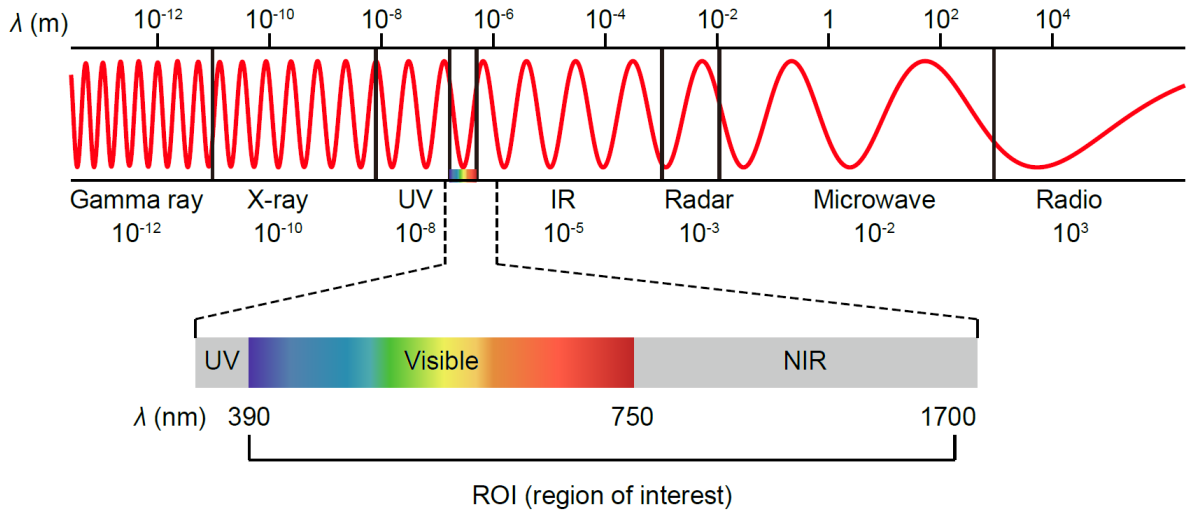

Figure 1.2: The spectral regions

A diagram of the electromagnetic spectrum, showing the range of wavelengths (λ; modified from Wikipedia, http://en.wikipedia.org/). The spectral regions (called "region of interest", ROI) of visible and near-infrared (VNIR), with the wavelengths ranging from 400 nanometers (nm) to 1700 nm, can be detected by LemnaTec system. UV: ultraviolet; IR: infrared; NIR: near infrared. ■

1.3.1 High-throughput phenotyping facilities

Thanks to the developed of robotics and new imaging sensors, various automated or semi-automated HTP systems are being developed and used to examine plant function and performance under controlled conditions or field-based environments (Table 1.2). A HTP infrastructure consists of its "hard" and "soft" parts (Figure 1.3) and is generally implemented for specific plant species due to their different architecture. The hard part of a HTP installation is generally fixed while its soft part is rather flexible for different experimental designs. For example, the same phenotyping system can be used to study either a mapping population or a mutant population of plants, and at the same time, different treatments (e.g., normal watering or drought stress) can be applied to the population.

Fully controlled environment-based phenotyping platforms are deplyed in growth chambers or greenhouses with robotics, precise environmental control and remote sensing techniques to assess plant growth and performance. These platforms are designed for large-scale phenotyping of a limited set of plant species, including small rosette plants such as Arabidopsis (Arvidsson et al., 2011; Granier et al., 2006) and several important cereal crops (e.g. Golzarian et al., 2011; Reuzeau et al., 2005). PHENOPSIS (Granier et al., 2006) is one of the pioneering platforms that was developed to dissect genotype-environment effects on plant growth in Arabidopsis. GROWSCREEN (Biskup et al., 2009; Jansen et al., 2009; Nagel et al., 2012; Walter et al., 2007) was designed for rapid optical phenotyping of different plant species. Among the advancing solutions, the state-of-the-art phenotyping platform developed by LemnaTec (http://www.lemnatec.com/) is a robotic greenhouse system that uses non-destructive imaging to monitor plant growth under controlled environmental conditions (e.g., controlled supply of nutrition,

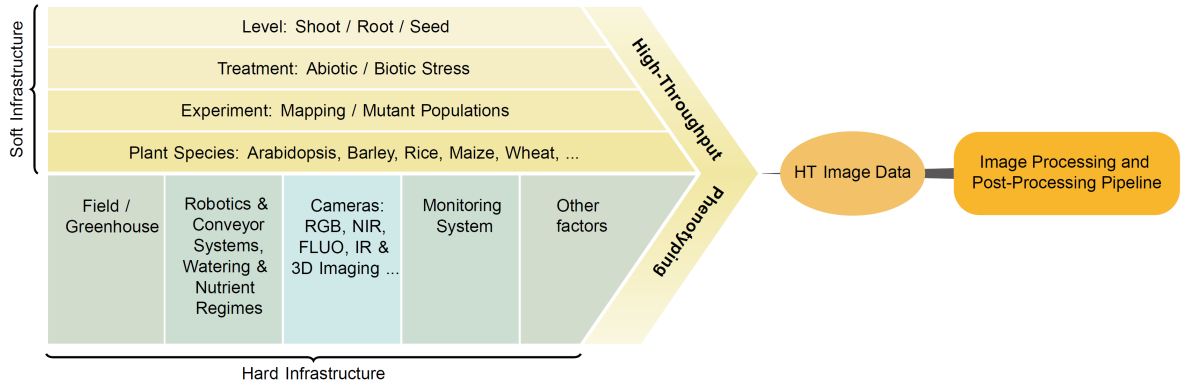

Figure 1.3: High-throughput phenotyping infrastructure

The high-throughput phenotyping infrastructure includes hard and soft parts. The hard infrastructure consits of the "hardware" of the phenotyping system. The soft infrastructure denotes the system capability and experimental design above the system. HT, high-throughput. ■

water availability, irradiation and temperature) over a period of time. LemnaTec Scanalyzer platforms have been deployed in growth chambers or greenhouses at various facilities around the world (Figure 1.4). For example, an increasing number of phenotyping centers with installations of LemnaTec systems are now emerging in Europe, Australia, America, China and India. Owing to its ingenious sensors, such as visible, fluorescence, thermal and near-infrared imaging cameras, The LemnaTec platform can be used to assess a range of phenotypic traits, including the physical and physiological status of plants (such as plant geometric properties, pigment or photosynthetic activity / chlorophyll, canopy temperature and water content). This system was successfully used in the prediction of biomass accumulation for Arabidopsis (Arvidsson et al., 2011) and cereal plants (Golzarian et al., 2011), and the detection of abiotic stress (Chen et al., 2014b; Hairmansis et al., 2014; Harshavardhan et al., 2014; Honsdorf et al., 2014).

Although controlled environment-based phenotyping platforms enable detailed, non-invasive information to be captured throughout the plant life cycle, results from controlled environments are difficult to extrapolate the field (Araus and Cairns, 2014; Fiorani and Schurr, 2013), as field conditions are notoriously heterogeneous. For example, the soil volume, solar radiation, wind speed and evaporation rates are hard to control in the field, making results difficult to interpret. Thus, large-scale phenotyping under field environmental conditions remains a bottleneck for future breeding advances (Araus and Cairns, 2014; Araus et al., 2008; Cabrera-Bosquet et al., 2012; Cobb et al., 2013). Given field-based phenotyping platforms are the only tool to be of use in the selection of genotypes that will perform well in farming practice (White et al., 2012), future efforts on development of high-throughput phenotyping should receive much more attention. In this regard, several custom-designed devices for field phenotyping have been established in the past few years (Table 1.2), including the system designed in Maricopa (Arizona) (Andrade-Sanchez et al., 2014), the Avignon system (France) (Comar et al., 2012), the "BreedVision" system from Osnabrucke (Busemeyer et al., 2013a), and the "Phenomobile" designed at the High Resolution Plant Phenomics Facility in Canberra (Deery et al., 2014). The current technical developments in field-based phenotyping are reviewed in Araus and Cairns (2014), Cobb et al. (2013), Li et al. (2014)

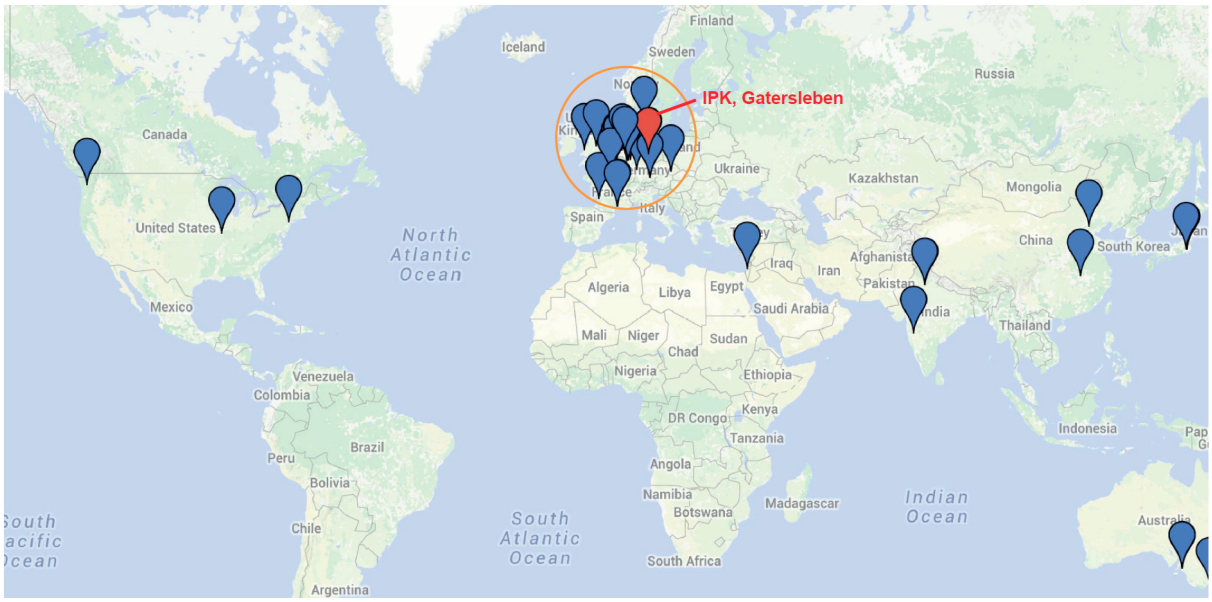

Figure 1.4: A global stronghold of high-throughput phenotyping facilities

This map was built with the Google Maps API (https://goo.gl/Fa6zNo) based on data collected from the websites of IPPN, EPPN, DPPN and LemnaTec. A hotspot is observed in Europe. ■

and White et al. (2012).

However, it is notable that it is still generally too expensive to set up automated phenotyping facilities, especially when the hardare required (robotics, camers, conveyor system, monitoring systems) (Fiorani and Schurr, 2013). To meet the demand of data access, exchange and sharing existing phenotyping installations, several international/local communities in the context of consortia (Table 1.3), such as the International Plant Phenotyping Network (IPPN; http://www.plantphenomics.com/), European Plant Phenotyping Network (EPPN; http://www.plant-phenotyping-network.eu/), the German Plant Phenotyping Network (DPPN; http://www.dppn.de/) and the Australian Plant Phenomics Facility (APPF; http://www.plantphenomics.org.au/), have been established by forming network of facilities.

1.3.2 Large-scale image processing and analysis

Raw data acquired from HTP systems are subjected to storage and subsequent image analysis. Image data can be either analyzed immediately after imaging or analyzed at later time for all plants when a phenotyping experiment is completed, or even reanalyzed in future when new request arises. To aviod time-consuming performance problems and to ensure an optimal configuration adjusted for the whole dataset in the image processing software, image storage and analysis are often separated. Images generated from various cameras in different imaging compartments are generally analyzed in parallel to extract up to hundreds or thousands of parameters per image. Furthermore, additional parameters (for example, projected area and digital volume) can be derived from image-based parameters (Klukas et al., 2014).

Image processing and analysis plays a significant role in plant phenotyping. Image processing is a

Table 1.3: Plant phenomics community.

Project	Description	URLs
IPPN	International Plant Phenomics Network. IPPN is an international consortium that will boost plant phenotyping science by developing novel technologies and concepts used for the application of plant production and the analysis of ecosystem performance.	http://www.plant-phenotyping.org/
EPPN	European Plant Phenotyping Network. This project will establish the network that integrates European plant phenotyping efforts and builds a competitive community to the goal of the understanding of the link between genotype and phenotype as well as their interaction with the environment.	http://www.plant-phenotyping-network.eu/
DPPN	German Plant Phenotyping Network. DPPN is a Germany funded project that partners undertake a joint research program and share their phenotyping infrastructure within networking activities.	http://www.dppn.de/
JPPC	The Jlich Plant Phenotyping Centre. This project is with aims to elucidate the functional role of gene networks under natural conditions with the aid of the development of non-invasive phenotyping tools and methods as well as the existing genetic resources.	http://www2.fz-juelich.de/icg/icg-3/jppc/phenotyping/
PHENOME	PHEOME, launched in 2012, is a project funded by French investment for the future. It will provide France with an up-to-date, versatile, high throughput infrastructure and suite of methods allowing characterization of panels of genotypes of different species (important crop species) under scenarios associated with climate changes.	http://urgi.versailles.inra.fr/Projects/PHENOME/
APPF	The Australian Plant Phenomics Facility. APPF is developed to alleviate the "phenotyping bottleneck" by utilizing high throughput plant phenotyping and "reverse phenomics" approaches with aims to probe and improve plant function and performance.	http://www.plantphenomics.org.au/

This table was adapted from Chen et al. (2014a).

form of signal processing that transforms a digital image into a set of characteristics or parameters related to the image. In plant phenotyping, the extracted image-based parameters can be considered as proxies of a set of plant phenotypes for direct use. A typical image processing pipeline consists of four key steps: (1) pre-processing, (2) segmentation, (3) feature extraction and (4) post-processing (Figure 1.5).

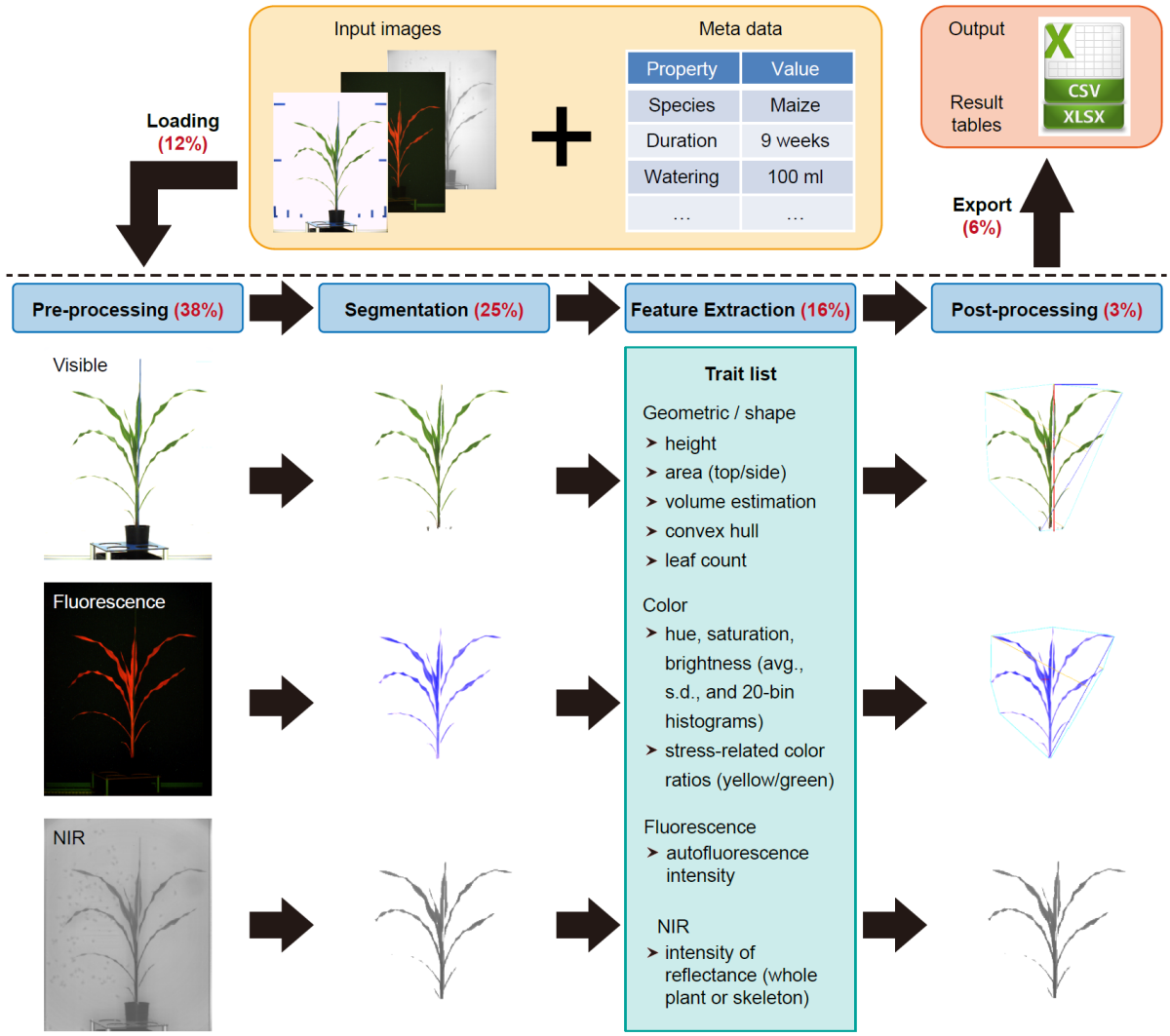

Figure 1.5: The typical workflow of a image-processing pipeline

Figure shows the IAP image processing pipeline applied to a maize dataset generated by the LemnaTec system. Image and metadata are imported via IAP functionalities (the above panel) and subjected to image processing, including (1) pre-processing prepare the images for segmentation, (2) segmentation divide the image in different parts which have a different meaning (foreground plant, background imaging chamber and machinery), (3) feature extraction classify the segmentation result and get a trait list. Examples include images from visible-light, fluorescence and near-infrared (NIR) cameras and (4) post-processing summarize calculated results for each plant, optionally analysis results can be marked in the images. Finally, result images are exported. Numbers in parentheses indicate the percentage of overall processing time for each analysis step. ■

With the rapid advances of HTP, a massive list of software tools (reviewed in Lobet et al., 2013, `http://www.plant-image-analysis.org/`) for plant image analysis are being developed to extract a wide

range of measurements, such as plant height, leaf length, width, shape, projected area, digital volume, compactness, relative growth rate and colorimetric analysis. These developments enable the phenotyping of specific organs (e.g., leaf, root and shoot) or of whole plants, and are even used for three-dimensional plant analysis. However, the trait information gained from these tools is still very limited. In addition, these analytical tools are individually designed to address specific questions (Sozzani and Benfey, 2011) and software tools that are capable of processing multispectral images are still underdeveloped. LemnaTec offers its own software solution called LemnaGrid (http://www.lemnatec.com/product/lemnagrid), which is based on the visual programming concept (Burnett, 2001), to analyze plant images from Scanalyzer 3D system with different cameras. LemnaGrid is quite handy for rapid prototyping and was successfully used in the prediction of biomass accumulation for Arabidopsis (Arvidsson et al., 2011; Camargo et al., 2014) and cereal plants (Golzarian et al., 2011; Hairmansis et al., 2014). However, as a commercial solution, LemnaGrid has the limitations of extensibility and automatization and is not intended to be further developed or significantly modified by the user (Berger et al., 2012). Thus only predefined functionalities are accessible. To meet these challenges, our IAP software (Figure 1.6; Klukas et al., 2014) has been developed to support a broad set of functionalities including data management, image processing and possible extensions via plugins and add-ons. Importantly, several essential yet tightly interdependent components of the IAP system have been implemented: (1) elaborate bioimage toolkits (such as ImageJ, Schneider et al., 2012) used to extract comprehensive and quantitative measurement from imaging datasets; (2) reusability and extension of algorithms into analysis workflows; (3) flexibility and interoperability of data management tools; (4) automated pipelines for data analysis; (5) seamless integration of other data visualization and analysis systems like (VANTED, Junker et al., 2006); and (6) specific graphical user interfaces (GUIs) for end users regardless of their scientific background and programming skills. These highlighted features make IAP as a full and extendable image-analysis framework for high-throughput phenotyping.

1.3.3 Applications of high-throughput plant phenotyping

The applications of HTP can be broadly categorized at two different levels: to gain deep insight into plant phenotypes and to dissect genetics underlying of these phenotypic traits by using genetic mapping approaches. In the first case, HTP is being applied to measure diverse phenotypic traits and their dynamics that are related to plant growth and performance. On the other hand, increase in the genetic information now puts more pressure on plant scientists and breeders for providing ample and accurate phenotypic data, with the goal of developing new variety or hybrid superior to existing one. Most breeding techniques, such as genetic mapping (including marker assisted selection, linkage-based QTL mapping and association mapping) and analysis of mutant populations, require proper phenotypic analysis. Manually collecting massive phenotypic data is time consuming and labor intensive. HTP is the ideal tool to alleviate this phenotyping bottleneck by dissecting the phenotypic components of complex traits. For example, several recent GWAS studies (Meijon et al., 2014; Slovak et al., 2014; Topp et al., 2013; Yang et al., 2014) were performed by using pheotypic traits derived from HTP data, revealing that HTP can

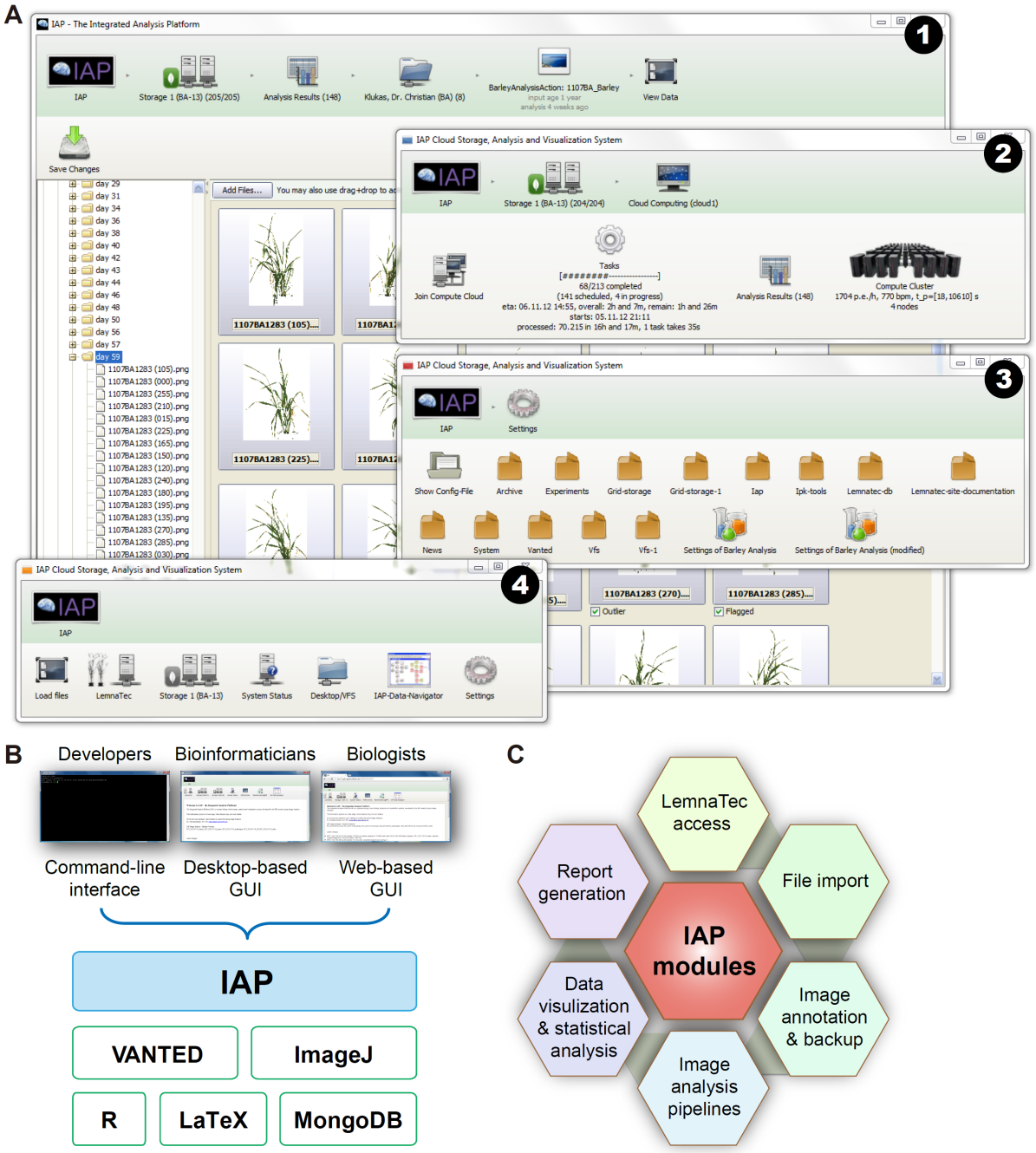

Figure 1.6: IAP: integrated analysis platform
(legend on next page).

replace traditional phenotyping techniques for gene identification. In the near future, integrating HTP and genetic mapping will bring on the revolution in the rate of trait discovery and the vast improvement of phenotypic prediction (Brown et al., 2014).

1.3.4 A proposed general framework for high-throughput phenotyping data analysis

Although there is an explosion of HTP systems developed for plant phenomics, the phenotypic components underlying dynamic processes in plants such as growth, development, or responses to environmental challenges and their properties remain unexplored. For these reasons, there is increasing demand for software tools that are capable of efficiently analyzing large image data sets and subsequent statistical methods to investigate comprehensively collected phenotypic data.

In this thesis, I present a general framework for high-throughput plant image data analysis (Figure 1.7), which was developed alongside currently available high-throughput image processing pipelines, such our IAP system (Klukas et al., 2014), and was extended from our published post-processing pipeline for high-throughput image analysis (Chen et al., 2014b). The core components of this framework consist of five parts: sample preparation, image acquisition, data management, image processing and data mining (Figure 1.7). Briefly, experimental setup can be controlled and optimized to minimize the influence of external environment in the robotic greenhouse system. The intensity of stress, the level of irrigation and the content of nutrient can be defined and controlled during a phenotyping experiment. Various types of image data, such as near-infrared (NIR)-, visible (color)- and fluorescence (FLUO)-images, can be acquired daily/hourly from different views (top view and side views from different angles) in the phenotyping platform (reviewed in Chapter 1.3). Consequently, timely retrieved data from imaging system are organized into data management system and subjected to the automated image processing pipeline (reviewed in Chapter 1.3.2) that extracts a large number of phenotypic trait values. Finally, by

► **Figure 1.6** (continued). **(A)** The graphical user interface (GUI) of the IAP system. Several windows can be opened by the user in parallel as shown in the screenshot: (1) the main window showing the overview of experiment data (browsing and processing images), (2) monitoring status of analysis jobs and grid-computing nodes, (3) the panel of system settings, and (4) the buttons of the main menu. This figure part was taken from Klukas et al. (2014). **(B)** Architecture and design of IAP. IAP uses the flexible and high-performance Mongo database (MongoDB; `http://www.mongodb.org/`) for image storage and management, ImageJ toolkit (Schneider et al., 2012) for image processing, R software (`http://www.r-project.org/`) for comprehensive statistical analysis, and the data structures of VANTED (Junker et al., 2006) for manipulating the experiment data. The IAP project provides three types of interfaces for a broad range of end users including developers, bioinformaticians and biologists. Developers can modify and extend existing software libraries provided by IAP, or implement and integrate new algorithm to meet specific requirements. Bioinformaticians can conduct image-processing pipelines and adjust some parameters under specific situations when necessary (using desktop-based GUI). Biologists can retrieve the analysis results from the web-based GUI. **(C)** IAP pipeline consisting of several sequential analysis modules, which enable an automated workflow for phenotyping data analysis. IAP automatically extracts phenotypic features from the images, only requiring users to modify default values of a few parameters. The parametric adjustment steps could be simply done though a GUI. This automated analysis workflow enable detecting plant growth in time and to change experimental conditions if needed based on the real-time observation. Besides, due to the huge amount of imaging data daily generated, IAP was implemented as a distributed storage and computing platform to speed up analysis. ■

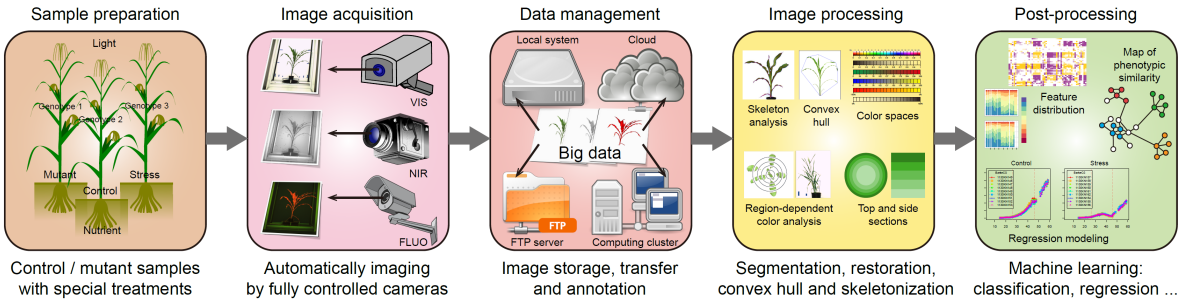

Figure 1.7: A comprehensive framework for high-throughput phenotyping in plants

The framework consists of five main steps (indicated in boxes). Firstly, plants are cultured under controlled environmental conditions in robotic greenhouse systems (sample preparation). Each plant with special treatments (such as abiotic stress and/or induced genetic mutation) is located in a container with controlled nutrient supply which is retrieved as needed by the conveyor belt. Secondly, different types of digital cameras (for example, imaging cameras in near infrared (NIR)-, fluorescence (FLUO)- and visible (VIS)-spectra) can be adopted to capture images in real time from different perspectives (for example, from the top and side views; image acquisition). During the imaging, plants are subjected to watering and weighting to ensure phenotyping in a non-invasive way. Next, "big data" acquired from the imaging system should be efficiently managed (such as image storage, annotation and backup) and transferred when needed (data management). Finally, image-processing methods are used to derive a representative set of phenotypic traits from image data (image processing) and data mining methods are used to decide the values of the extracted features or to mathematically model phenotypic data (data mining). Note that the first two steps have been implemented in automated phenotyping systems such as LemnaTec (http://www.lemnatec.com/), and the next two steps have the solutions in our IAP system (http://iap.ipk-gatersleben.de/, Klukas et al., 2014). This paper focuses on the last step to develop efficient post-processing methodology to interpret high-throughput plant phenotyping data. ■

applying well-established statistical models (for example, Chen et al., 2014b), the extracted phenotypic traits can be used to assess plant growth and performance features. Furthermore, by integrating data from other domains, these imaged-based traits and model-derived parameters are promising for subsequent genetic mapping (in mapping populations) and functional analysis (in large collections of transgenic or genetically modified plants).

1.4 Publications on which this thesis is based

Parts of this thesis include results from the following publications that are the result of my work conducted as Doctoral Student at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK).

- ✔ Klukas, C., Chen, D., and Pape, J. M. (2014). Integrated analysis platform: An open-source information system for high-throughput plant phenotyping. *Plant Physiol*, 165(2):506–518 (Chapter 1)
- ✔ Chen, D., Chen, M., Altmann, T., and Klukas, C. (2014a). *Bridging Genomics and Phenomics*, chapter 11, pages 299–333. Springer Berlin Heidelberg (Chapter 1)

- ✔ Chen, D., Neumann, K., Friedel, S., Kilian, B., Chen, M., Altmann, T., and Klukas, C. (2014b). Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. *Plant Cell*, 26:4636–4655 (Chapters 2 and 3)
- ✔ Chen, D. (2016). Htpmod: an r package for modeling plant growth and its phenotypic components in the era of plant phenomics. *in preparation* (Chapter 3)
- ✔ Chen, D., Shi, R., Pape, J.-M., and Klukas, C. (2015). Predicting plant biomass accumulation from image-derived parameters. *submitted (preprint doi: 10.1101/046656)* (Chapter 4)

Chapter 2

Dissecting the high-dimensional phenotypic components of plant growth and drought responses

2.1 Introduction

Plant breeding is currently meeting the tremendous challenge for crop improvement in the face of a growing human population and global environmental change. While recently developed genotyping methods promise to identify additional genes and variants of interest used in agronomic improvement (Takeda and Matsuoka, 2008), plant breeders are seeking efficient phenotyping approaches to select traits with the greatest potential for yield improvement to speed up the crop breeding progress (Tester and Langridge, 2010). The "phenotyping bottleneck" (Furbank and Tester, 2011) — our ability of depiction and quantification of plant phenotypes largely lagging behind that of genotypes — can now be alleviated by the introduction of high-throughput phenotyping (HTP or phenomics) using non-invasive image technologies as well as high-performance computing. Several structural and functional imaging techniques (such as visible, infrared, hyperspectral and chlorophyll fluorescence imaging) are employed to study plant architecture, growth and physiological status (Berger et al., 2010; Yang et al., 2013; Zhu et al., 2011). Such multifunctional phenotyping tools enable us to accurately measure increasingly large numbers of plants and phenotypic traits over a long period of plant growth. Altogether, these advances have made it possible to deeply investigate the phenotypic components of complex traits and to study their influence on crop yield.

Automated non-invasive precise HTP is especially interesting in the context of dissecting the complex genetic architecture of biomass development and of drought stress tolerance. Impact of stress, such as drought, depends heavily on timing and intensity of the dry period and on environmental conditions (Araus et al., 2002; Calderini et al., 2001) hampering heritability as a pre-requisite for genetic mapping of quantitative trait loci (QTL) (Painawadee et al., 2009; Ribaut et al., 1997; Sellammal et al., 2014).

Drought tolerance has been investigated in various QTL studies since the start of the molecular marker age (Lilley et al., 1996; Nezhad et al., 2012; Szira et al., 2008; Xiong et al., 2006). Adequate controlled phenotyping and daily phenotypic observation of drought stress development has a huge potential to boost the understanding of the genetics of drought tolerance.

Here, several algorithms were implemented in a pipeline for efficient analysis and interpretation of huge and high-dimensional phenotypic data sets to support understanding plant growth and performance. The pipeline was applied to a core set of 18 different barley cultivars, which were daily imaged under well-watered and drought-stress conditions. A list of representative phenotypic traits were extracted and quantified from the digital imaging data. Linear mixed models were used to dissect variance components of phenotypic traits and showed that the traits revealed variable genotypic and environmental effects and their interactions over time. Key parameters such as trait heritability and genetic trait correlations were assessed, indicating image-derived traits are valuable in genetic association studies.

2.2 Results

2.2.1 Extraction of phenotypic traits from high-throughput image data

I applied the methodology to a compendium of ˜50,400 images (˜100 GB of data) collected for 18 barley genotypes from four agronomic groups (Table 2.1), with six (for double haploid [DH] lines) or nine (for non-DH lines) replicated plants per genotype per treatment. Over a course of seven weeks plants were monitored in a noninvasive way under control and drought-stress conditions using an automated plant transport and imaging system (Figures 2.1 and 2.2; see Chapter 2.4.1). Three types of image data, near-infrared (NIR)-, visible (color)- and fluorescence (FLUO)-images, were acquired daily from different views (top view and side views from different angles) in the phenotyping system, and were used for trait extraction (reviewed in Chapter 1.3.2). Data retrieved from the imaging platform were organized into the IAP system (Klukas et al., 2014) and processed through an analysis pipeline specifically adjusted for mid-sized important crop species such as barley, resulting in values of nearly 400 phenotypic traits extracted from images of each individual plant (Figure 2.2A,C).

These phenotypic measurements can be classified broadly into four categories: plant geometric traits (measuring shape descriptors of plants), color-related properties, NIR-signals and FLUO-based traits (Figure 2.2C). Quantitative traits were first evaluated based on their reproducibility among replicated plants (see Chapter 2.4.3; Figure 2.3A-B) against random plant pairs, to avoid introducing low quality or weak phenotypic traits into the analysis. 173 (44.6%) traits showed high reproducibility among replicate samples after removing outliers (Pearson correlation coefficient $r > 0.8$ and one-sided Welchs t-test $P < 0.001$; Figure 2.2A). It was found that 87.0% of traits that showed genotypic effects or 93.1% of traits that showed treatment effects (adjusted $P < 0.01$; see below) passed this filtering (Figure 2.4), indicating that most of the informative traits were still covered though the stringent applied criteria. Clustering analysis of these highly reproducible traits showed that large sets of traits were excessively

correlated with each other (Figure 2.5), indicating that these traits might be highly redundant descriptors of plant properties within the investigated cultivar set. To get an optimal set of phenotypic traits for a statistical model, the indicator of variance inflation factors (O'Brien, 2007) ($VIF > 5$) was applied to remove redundant and non-informative features (see Chapter 2.4.4). After manual checking, 54 (31.2%) traits were selected from the entire set of reproducible measures and used them in the remaining analysis (Figures 2.2A and 2.3C, and Supplemental Table S1). However, it is notable that this barley collection is relatively small, and some of the excluded phenotypic traits might be considered when applying the model to larger plant populations.

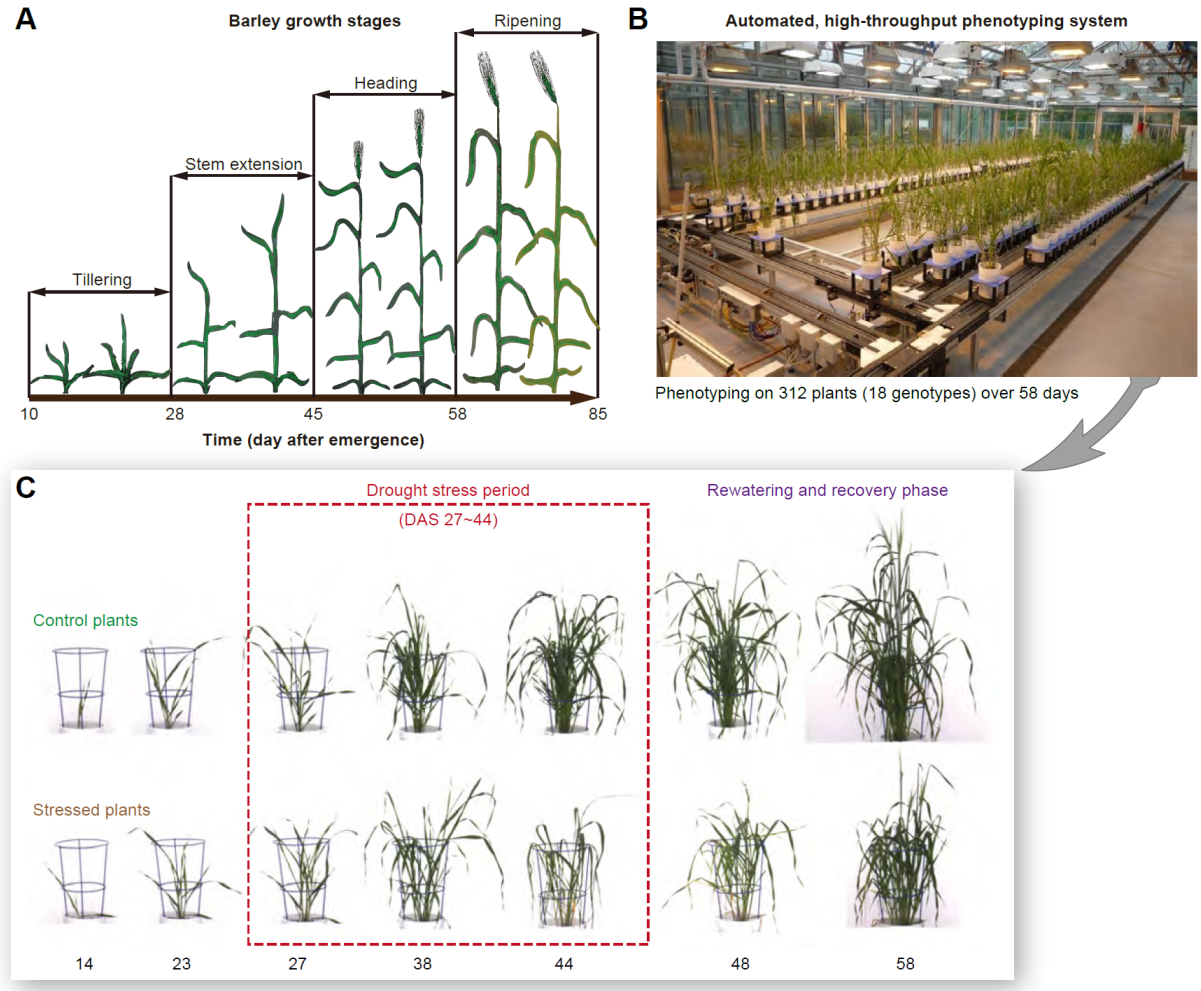

Figure 2.1: Experimental design for high-throughput phenotyping in barley

(A) The growth stages of spring barley. **(B)** High-throughput phenotyping of barley plants in a LemnaTec system (http://www.lemnatec.com/). **(C)** Plants were monitored in a noninvasive way under control and drought-stress conditions. Drought stress (in dash box) was treated at the stage of "stem extension" as indicated in **(A)**. This figure was taken from Chen et al. (2014b). ■

2.2.2 Image-derived parameters reflect drought stress responses

Many of the phenotypic changes (such as changes of biomass) were readily detectable upon stress treatment, whereas others (such as dynamics of water content) were less obvious or too subtle to be discerned

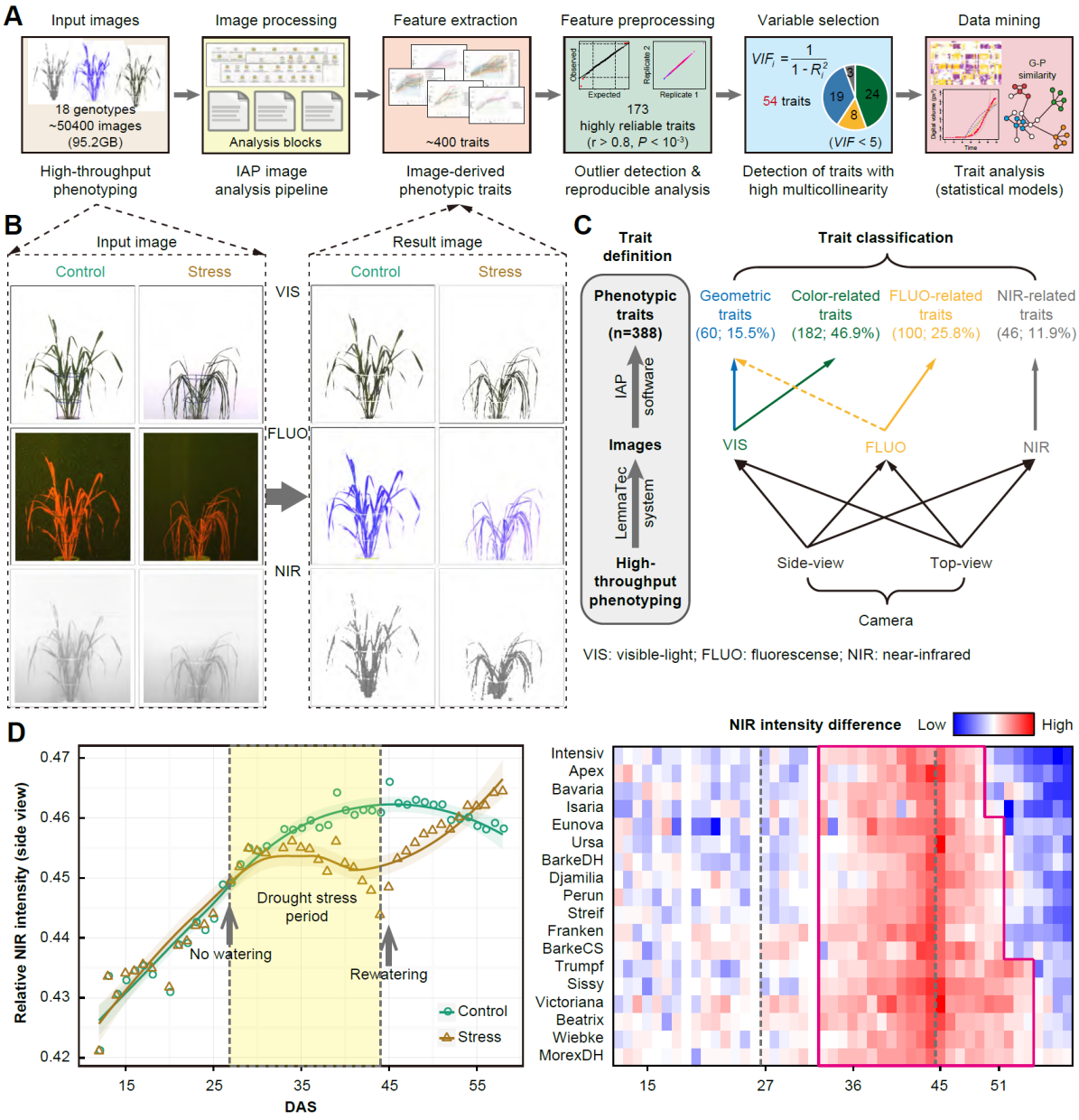

Figure 2.2: Pipeline for analysis of high-throughput phenotyping data in barley

(A) The workflow used for barley phenotyping data analysis. High-throughput imaging data from the LemnaTec system were imported and processed using the barley analysis pipeline in the IAP system. The extracted phenotypic traits were further processed and evaluated (see Chapters 2.4.3 and 2.4.4). **(B)** Input (left) and result (right) images in the analysis pipeline. Shown are images from 44-day old of plants (the last day of stress phase) captured by VIS-. FLUO- and NIR-cameras from the side view. **(C)** Classification of phenotypic traits. Traits are classified into four categories: color-related, NIR-related, FLUO-related and geometric features, based on images obtained from three types of cameras and two views. **(D)** Phenotypic traits revealing the stress symptom. Left: An example shows a NIR-related trait over time. Right: heatmap shows NIR intensity difference, measured by the ratio value between control and stress plants. Blue indicates low difference, whereas red indicates high difference. Note that plants from different genotypes show different patterns, indicating their different stress tolerance. This figure was taken from Chen et al. (2014b). ■

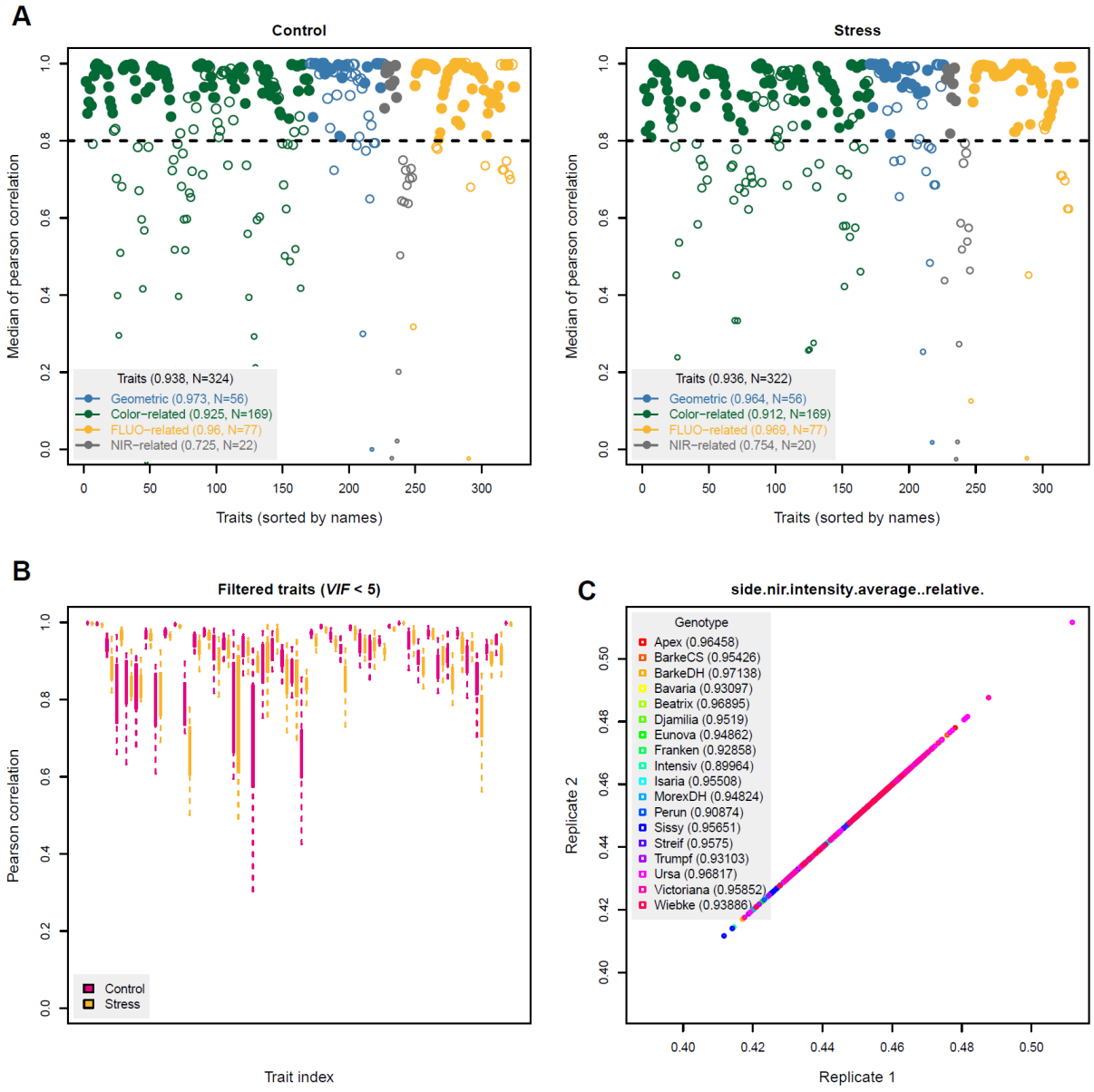

Figure 2.3: Reproducibility of phenotypic traits

(A) Highly reproducible analysis of phenotypic traits based on control (left) and stressed plants (right). Replicate plants with the same genotype and treatment were used to assess the reproducibility of traits based on Pearsons correlation ($r > 0.8$). Besides, the correlation in replicate plants is considered to be significant higher than by chance ($P < 0.001$; see Chapter 2.4.3). Filled dots represent the filtered traits with high reproducibility ($r > 0.8$ and $P < 0.001$). The median values of Pearsons correlation for each trait categories are indicated. The number of traits with non-empty values is provided as well. The trait reproducibility is consistence between control and stressed plants. **(B)** The reproducibility of the 54 filtered traits. **(C)** An example of highly reproducible traits: the NIR-intensity trait. This figure was taken from Chen et al. (2014b). ■

by eye (Figures 2.2B and 2.6). Previous studies have suggested that plant water stress might be monitored effectively using NIR imaging (Berger et al., 2010; Knipling, 1970; Munns et al., 2010; Tucker, 1980). It was shown that the highly reproducible NIR-intensity trait is an effective feature for monitoring plant responses to drought (Figures 2.2D and 2.3C). Plants showed a rapid decrease of the NIR signal

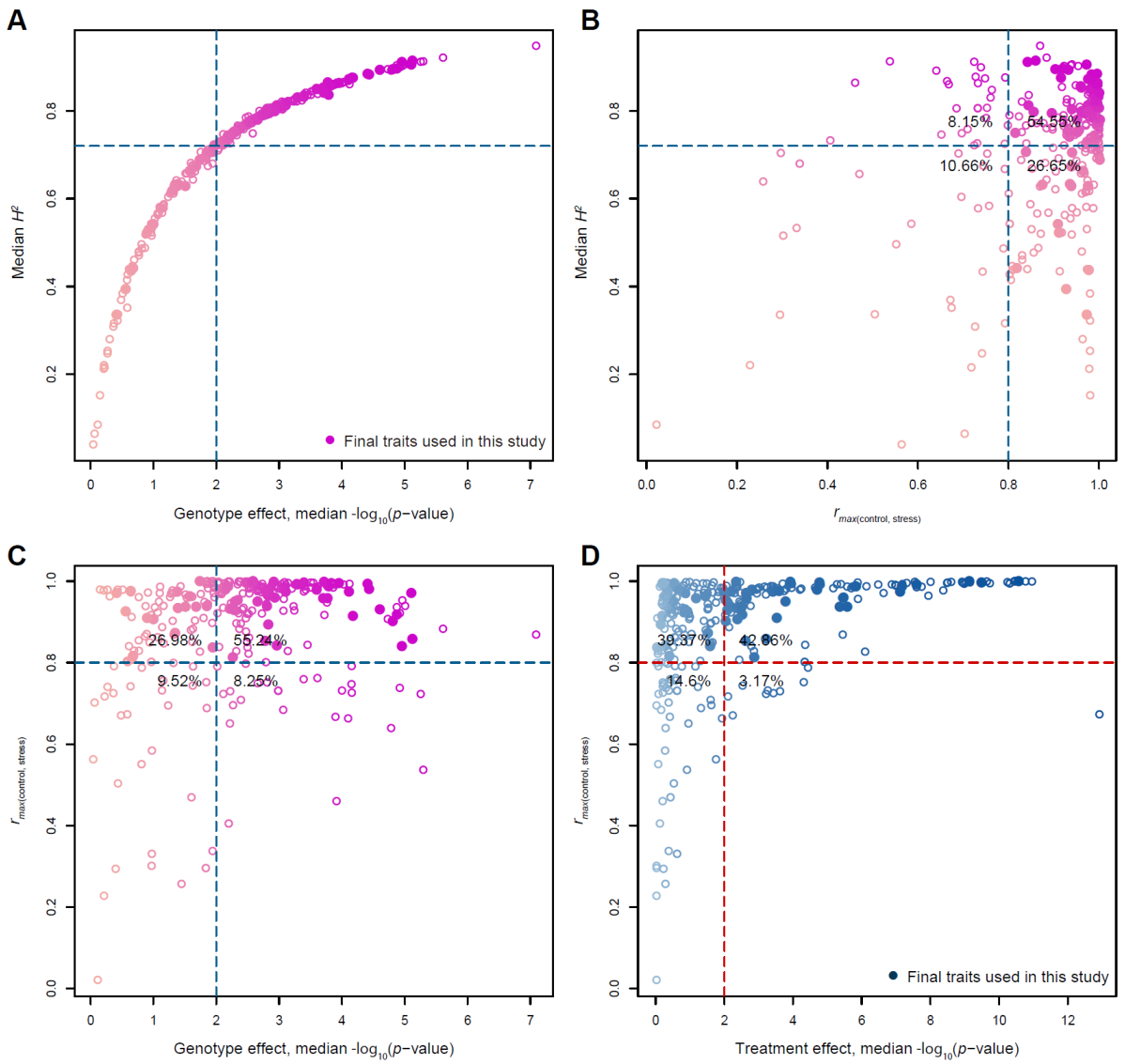

Figure 2.4: Assessment of trait reproducibility analysis

For all the plots, each point represents one trait. Filled circles indicate the final 54 traits used in this study. The dash lines indicate the corresponding cut-offs. **(A)** Scatter plot showing genotype effect (median negative log-transformed p-values) versus heritability (H^2). **(B)** Scatter plot of trait reproducibility (the maximum correlation value in either control or stress treatments) and H^2. **(C)** Scatter plot of genotype effect and trait reproducibility. **(D)** Scatter plot of treatment effect and trait reproducibility. This figure was taken from Chen et al. (2014b). ■

Table 2.1: Overview of 18 barley genotypes used in this study.

Group†	Genotype§	Release	Breeder	Pedigree
DH	**BarkeDH**	1996	Breun	Libelle × Alexis
DH	**MorexDH**	1978	MN AES	Cree × Bonanza
1	Ackermanns **Bavaria**	1910	Ackermann	Selection from Bavarian landrace

Table 2.1 (continued)

1	**Heils Franken**	1895	Heil	Selection from Franconian landrace
1	**Isaria**	1924	Ackermann	Danubia × Bavaria
1	Pflugs **Intensiv**	1921	Pflug	Selection from Bavarian landrace
2	**Apex**	1983	Lochow	Aramir × (Ceb.6721 × Julia × Volla × L100)
2	**Perun**	1988	Hrubcice/NKGNord	HE 1728 × Karat
2	**Sissy**	1990	Streng	(Frankengold × Mona) × Trumpf
2	**Trumpf**	1973	Hadmersleben	Diamant × 14029/64/6 ((Alsa × S3170/Abyss) × 11719/59) × Union
3	**Barke**	1996	Breun	Libelle × Alexis
3	**Beatrix**	2004	Nordsaat	Viskosa × Pasadena
3	**Djamila**	2003	Nordsaat	(Annabell × Si 4) × Thuringia
3	**Eunova**	2000	Probstdorf	H 53 D × CF 79
3	**Streif**	2007	Saatzucht Streng GmbH & Co. KG	Pasadena × Aspen
3	**Ursa**	2001	Nordsaat	(Thuringia × Hanka) × Annabell
3	**Victoriana**	2007	Probstdorfer/ Saatzucht	(LP 1008.5.98 × LP 5191) × Saloon
3	**Wiebke**	2000	Nordsaat	Unknown

† Agronomic group. The DH-population parents are indicated with "DH". Cultivar Morex is a six-rowed, spring barley from US. All other cultivars are two-rowed spring barleys released in Germany. Genotypes are grouped according to the year of release (except for DH lines): Group 1 (< 1950), Group 2 (1950-1990) and Group 3 (> 1990).

§ **bold** indicates the short name used in all the figures. This table was taken from Chen et al. (2014b).

after about six days of drought stress. Restoration of the NIR signal was seen after re-watering. The NIR-based indicator also provides a measure of the different abilities to recover among different genotypes (Figure 2.2D).

To explore more comprehensively the ability of these traits to reflect the responses to the external treatment, a support vector machine (SVM)-based approach (Iyer-Pascuzzi et al., 2010; Loo et al., 2007) was used, in which "optimal" hyperplanes separate treated and untreated samples (Figure 2.7A). It was found that accuracy in distinguishing between stressed and control plants reached over 90% after one week of drought stress and nearly 100% separability after ten days of stress (Figure 2.7B). Besides, the "phenotypic direction" (the normal vector of the hyperplane in SVM) of greatest separation between the two groups of plants revealed three grouped patterns over time, corresponding to the three different treatment periods: growth before onset of drought treatment, during drought stress, and in the recovery phases (Figure 2.7C). These results suggest that the treatment effects of these traits changed dynamically according to the external treatment and growth stage (see below).

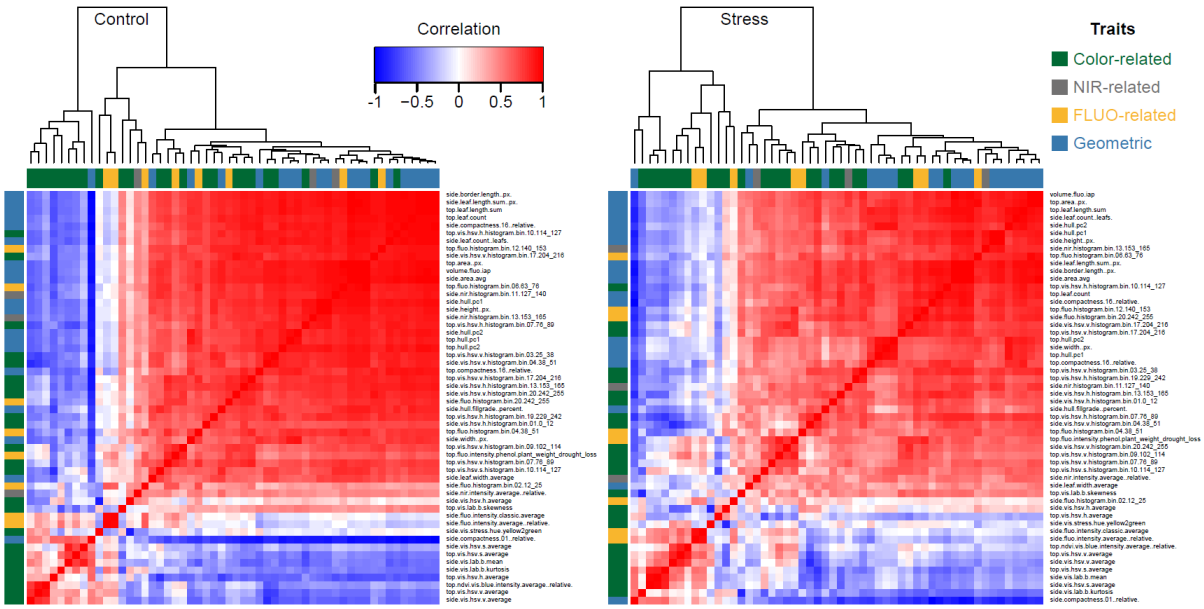

Figure 2.5: Trait similarity

Canonical correlation analysis of phenotypic traits based on control (left) and stressed plants (right). Heatmap plot is organized by hierarchical clustering with the tree (top). Traits are listed on the right. This figure was taken from Chen et al. (2014b). ■

2.2.3 Plant phenomic map and phenotypic similarity

To gain a global plant phenotypic map across the entire cultivar set, clustering approaches were performed on the comprehensive phenome-wide data (Figure 2.8A-B). This map provides important information regarding plant phenotypic similarity or dissimilarity and supports further evaluation of the defined traits. From a cluster analysis with complete linkage applied to the normalized dataset, it was found that stressed plants were clearly distinguished from controls plants irrespective of genotype, but plants of the same genotype or among agronomic groups tended to be grouped together (Figure 2.8A, upper panel), supporting the idea that similar genotypes lead to similar phenotypes. For the 54 investigated traits, correlation coefficients of trait profiles between pairs of genotypes of the same agronomic groups were significantly higher than pairs of different groups ($P < 2.2 \times 10^{-16}$, one-sided Mann-Whitney U-test; Figure 2.8A, lower panel). Similar results were observed in a large genome-wide association study (GWAS) mapping population, in which 34 traits were investigated across 413 diverse rice accessions in the field (Zhao et al., 2011). To fine visualize phenotypic similarity revealed by genotype similarity, a self-organizing map (SOM) (Kohonen, 1990) clustering analysis was performed on the dataset (Figure 2.8B). The SOM plot showed that plants from the same genotype were concentrated at certain locations in the map, and stressed plants were clearly separated from the control plants.

Next, a neighbor-joining tree (termed "phenotypic similarity tree") was deduced based on the 54 informative traits to reveal the phenotypic similarity of plants of different origins (see Chapter 2.4.6). The phenotypic similarity trees were constructed for plants cultivated under control and stress conditions, respectively (Figure 2.8C). It was observed that members of the same agronomic groups belonged to closed

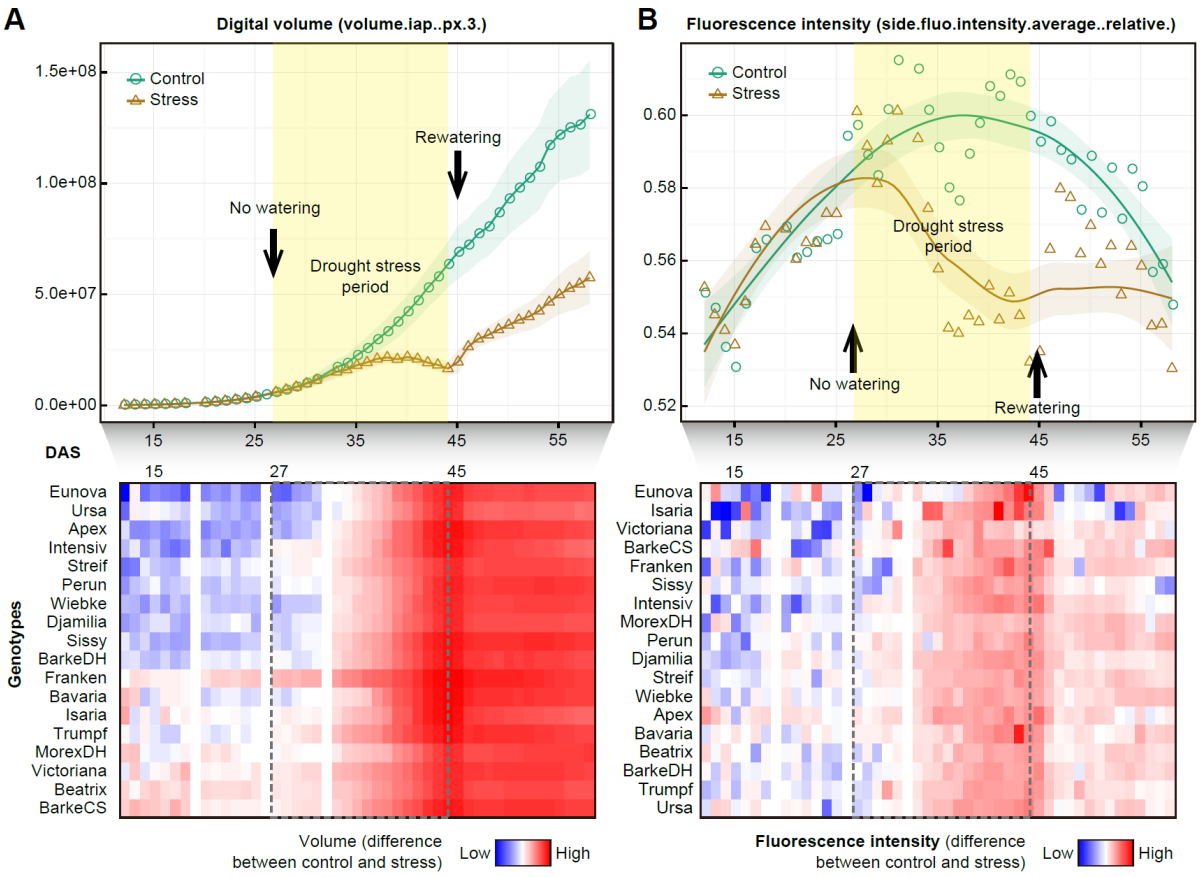

Figure 2.6: Phenotypic traits revealing the stress symptom

Related to Figure 2.1D. Top: Examples show the trait of **(A)** "digital volume" and **(B)** "fluorescence intensity" over time. Bottom: heatmap shows the difference in traits, measured by the ratio values, between control and stressed plants. Blue indicates low difference, whereas red indicates high difference. Note: plants from different genotypes show different patterns. This figure was taken from Chen et al. (2014b). ■

branches of the tree (Figure 2.8C, left), reflecting the domestication and breeding history of these cultivars. The phenotypic similarity tree reshaped following the drought stress although the relative relationship of most cultivars within the same groups was unchanged (Figure 2.8C). Consistent with this observation, the phenotypic distance matrices of these two trees are positively associated (Pearsons coefficient $r = 0.71$ and $P < 0.001$, Mantel test; Figure 2.8D). However, it was observed that barley cultivars such as Apex, Djamilia, and Heils Franken showed least robustness in maintaining their phenotypic relationship when they were exposed to drought stress (Figure 2.8C-D), suggesting that the phenotypic plasticity of these cultivars in response to stress treatment is different.

2.2.4 Phenotypic profile reflects global population structure

To further explore the phenotypic relationships of these plants, principal component analyses (PCA) were carried out to capture global phenotypic variation in the whole population and to extract specific phenotypic traits relevant for the discrimination of agronomic groups (Figures 2.9 and 2.10). The top

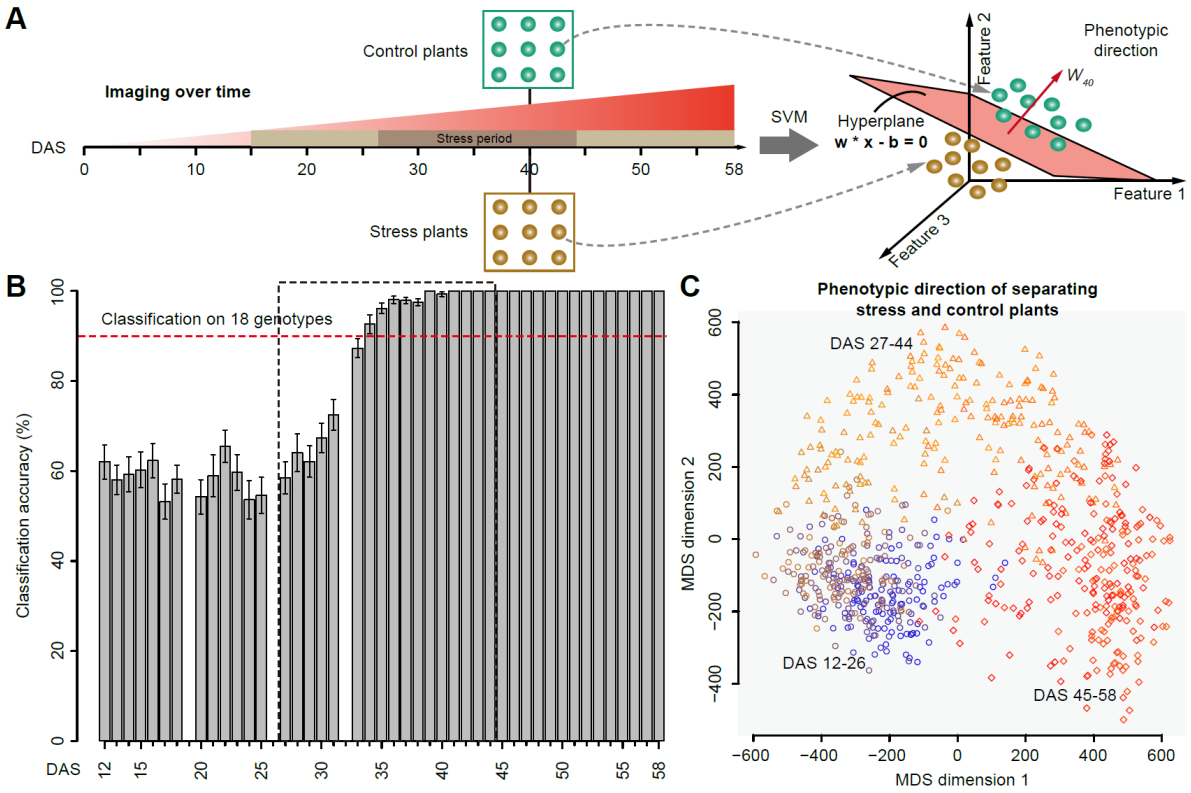

Figure 2.7: Classification of plants based on the SVM methodology

(A) An SVM-based methodology used for classification of plants with different treatments. Measurements of multidimensional traits (the highly reproducible traits) were used to represent plants in a high-dimensional feature space. An SVM-based classifier was used to determine the optimum hyperplane which separated plants into control and stress groups for each genotype from every daily imaging data (for example, in 40 days after sowing, DAS 40). Hyperplane orientation represented its weight vector (W_{40} for DAS 40), indicating the "phenotypic direction" of greatest separation between the two plant groups. **(B)** Classification accuracy to evaluate the performance of classification. Dashed line indicates classification accuracy of 90%. Stress period is indicated. Error bars, s.e.m. ($n = 18$). (C) Multidimensional scaling (MDS) plot showing the patterns of phenotypic direction over time. Each point represents the phenotypic direction for a specific genotype from a specific time point. Three distinct patters were observed, corresponding to three different phases to the experiment. This figure was taken from Chen et al. (2014b). ■

six principal components (PCs) explain at least 60% of the total phenotypic variation (Figure 2.9A). Notably, the accumulative variance explained by these PCs increases with plant growth, having a slight peak at the end of stress phase, accounting for 83.3% of total variation. The increasing accumulative variance over time was observed for control and stressed plants, respectively (Figure 2.11A-B), indicating that plants showed more phenotypic differences at the later growth stage.

At the end of the stress period, the first PC (PC1) explains more than half (52.9%) of the phenotypic variation, which perfectly separated stressed plants from control plants (Figure 2.9B). Accordingly, geometric and NIR-intensity traits are the main factors in the trait space separating these two groups of plants. Meanwhile, PC1 gradually increases along the stress phase while it decreases when plants recovered with watering, suggesting that more phenotypic variance can be observed between control and

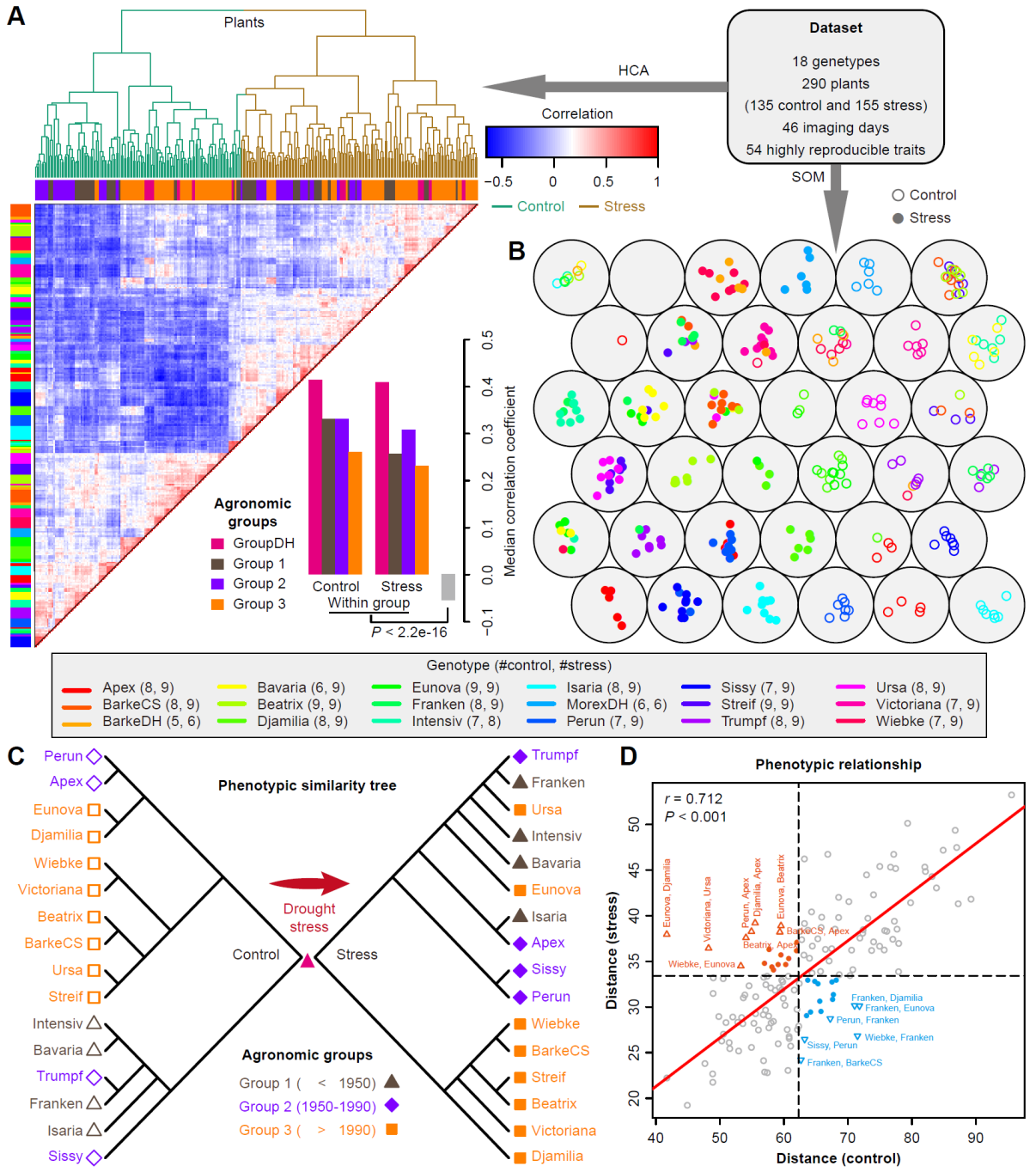

Figure 2.8: Phenotypic similarity revealed by genotype similarity

This figure was taken from Chen et al. (2014b) (legend on next page).

stressed plants under more serious stress. Other PCs with smaller proportions of explained variance generally distinguish plants of different agronomic groups from each other. For example, PC2 was mainly driven by the phenotypic difference in groups 2 (released before 1990) and 3 (released after 1990) (Figure 2.9B), corresponding to the main PCs as observed in control (Figure 2.11C) and stressed plants (Figure 2.11D). Interestingly, more diversity in color-related traits was observed in plants of agronomic group 2, likely revealing the human selection of breeding of these cultivars. The third principal component (PC3) mainly distinguishes plants of agronomic group 1 from the DH group (Figures 2.9B and 2.11C-D).

However, the different patterns in the PCA from control and stress plants (Figure 2.11A-B) can be explained in part by complex genotype-treatment interactions. Overall, the observations that the first PC separates control and stress plants and that the other PCs separate agronomic and genotype groups are in agreement with the results of the clustering analysis, which showed that plants had larger phenotypic dissimilarity between treatments than between genotype groups (Figure 2.8A), further indicating that the environment (drought stress treatment) shows dramatic effects on plant growth and development.

2.2.5 Dynamic genotypic and environmental effects on phenotypic variation

A linear mixed model was used to decompose phenotypic variance (P) into different causal agents: genetic (G) and environmental (E) sources, and their interaction effects (G×E). The mixed-effects model was fitted using a restricted maximum likelihood approach and the statistical significance of variance components was estimated by the log-likelihood ratio test (log-LR test; see Chapter 2.4.8). It was found temporal dynamics of genotypic and environmental influences on overall trait development (Figure 2.12A-B). In the early growth phase, phenotypic variance was mostly the result of unknown environmental effects (residual effects). As plants grew, genotypic factors became more important. The increasing genetic effect on phenotypic variance was observed up to about six days after the onset of stress treatment, after which the environmental factors (e.g. drought stress) became progressively more important, while the genetic effect became relatively less important. Although less obvious, the opposite pattern was seen in the recovery phase (Figure 2.12A), likely due to the decline in phenotypic differences between control and stressed plants. The decline in error variance and increase in environmental variance are reflected by a dynamic change of the total experimental coefficient of variation (CV) over time based on the investigation of geometric traits (Figure 2.12B). The total experimental CV increased as the drought stress

► **Figure 2.8** (continued). **(A)** and **(B)** Clustering analysis of phenomic profiling data. **(A)** Hierarchical clustering analysis (HCA) and **(B)** a six-by-six self-organizing map (SOM) were used to reveal the phenotypic similarity of all the investigated barley plants based on the highly reproducible traits. In **(A)**, Coloured bars along the top of the heatmap reflect the sampled agronomic group assignment (group 1-3 and DH) as labeled. Coloured bars along the left indicated the corresponding genotypes of individuals as listed in the key. The lower panel shows the median correlation values among individual plants from the same agronomic groups and different groups. In **(B)**, plants with similar genotypes or treatments tend to be at nearby map locations. Control and stress plants are coloured and indicated in blank and filled points, respectively. The numbers in the key show the number of plants from the same genotypes belonging to the control or stress group. **(C)** Phylophenetic trees showing the phenotypic relationship of plants from agronomic groups 1-3 under control (left; blank shapes) and stress (right; filled shapes) conditions. The trees were constructed from overall phenotypic distance matrices (see Chapter 2.4.6). **(D)** Scatter plot indicating the degree of correlation of phenotypic distance between genotypes under both control (x axis) and stress conditions (y axis). Mantel test was performed to examine whether the phenotypic distances in the two conditions correlate with each other. p-value was calculated with Monte-Carlo simulation (with 10,000 permutations). Genotype pairs that are far away from the regressed line (red) are labeled and coloured (orange, small distances in control and large distances in stress; blue, otherwise). ■

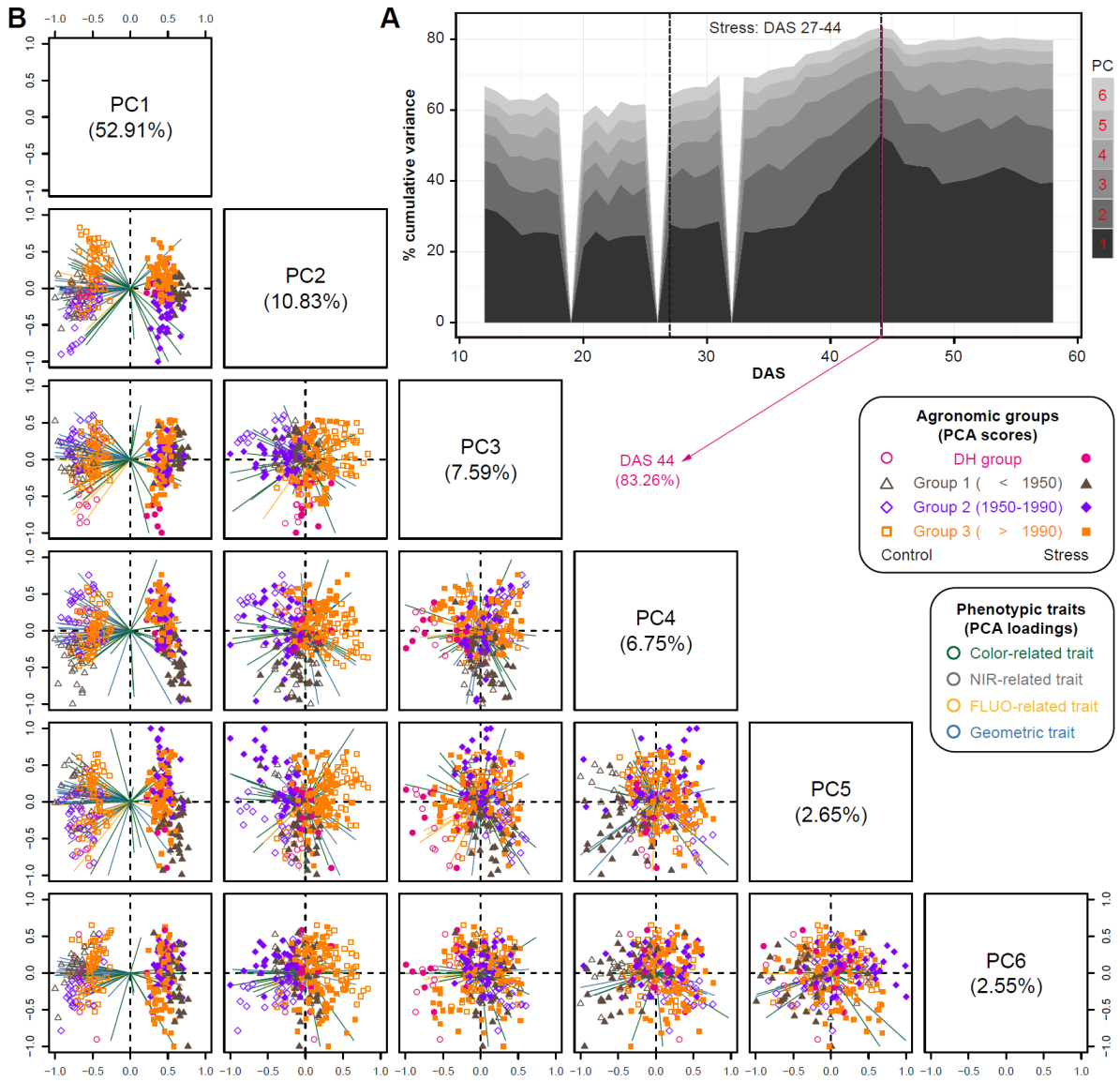

Figure 2.9: Phenotypic profile reflects global population structures in the temporal scale

(A) Projections of top six principal components (PCs) based on principal component analysis (PCA) of phenotypic variance over time. The percentage of total explained variance is shown. The stress period is indicated by the dashed box. **(B)** Scatter plots showing the PCA results on DAS 44 (explained the largest variance). The first six PCs display 83.3% of the total phenotypic variance. The component scores (shown in points) are coloured and shaped according to the agronomic groups (as legend listed in the box). The component loading vectors (represented in lines) of each variable (traits as coloured according to their categories) were superimposed proportionally to their contribution. See also Figures 2.10 and 2.11. DAS, day after sowing. This figure was taken from Chen et al. (2014b). ■

became more severe and declined during the recovery phase. However, the genetic CV across the cultivars was relatively constant upon drought treatment. The genetic CV in stressed plants became less than that in control plants after the onset of treatment (Figure 2.12B), indicating that plants showed more phenotypical diversity under normal growth conditions than in stressed conditions. Genetic CV peaked at the beginning of plant growth, revealing heterogeneity of plant growth at the initial growth stage. A moderate level of G×E interaction effects (with the proportion of explained phenotypic variance ranging

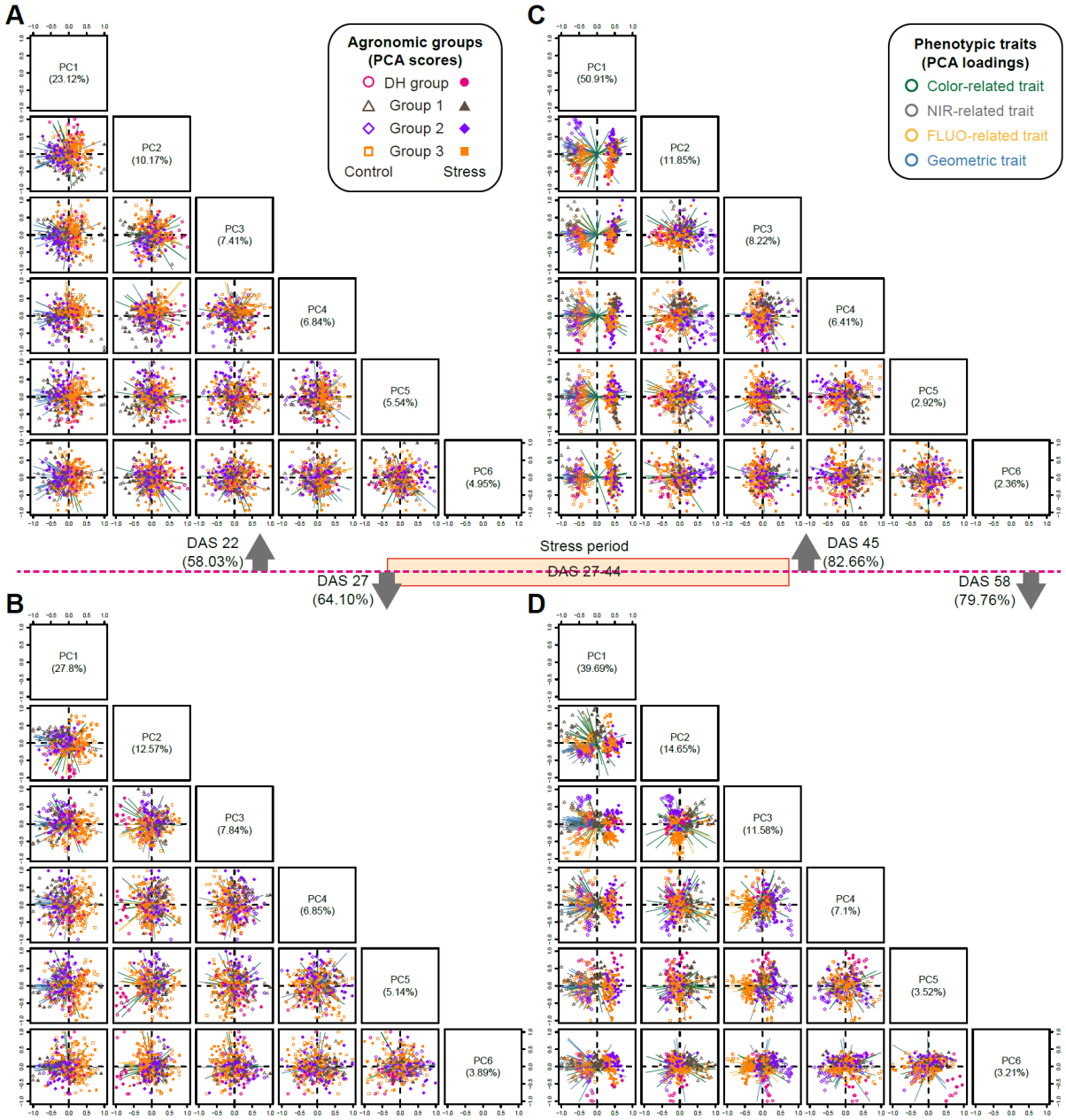

Figure 2.10: PCA performed over time

Related to Figure 2.9B. Scatter plots show the top six PCs on **(A)** DAS 22, **(B)** 27, **(C)** 45 and **(D)** 58. The proportion of variance explained by the PCs is shown in parentheses. The component scores (dots) are coloured and shaped according to agronomic groups of plants. The loadings (lines) of each variable (traits) are coloured according to their categories. This figure was taken from Chen et al. (2014b). ■

from 2.6% ˜15.4%; Figure 2.12A) was also observed, indicating that there are genetic differences in the response to drought among different cultivars. It was found that the G×E effects progressively increased with plant development, independent from external environment changes.

To gain a deeper insight into traits that could shed light on the genotype and treatment effects as well as their interaction, the likelihood estimation (the LOD score Joosen et al., 2013) was calculated from the linear mixed models to determine whether the G, E, and G×E effects have statistical significance on

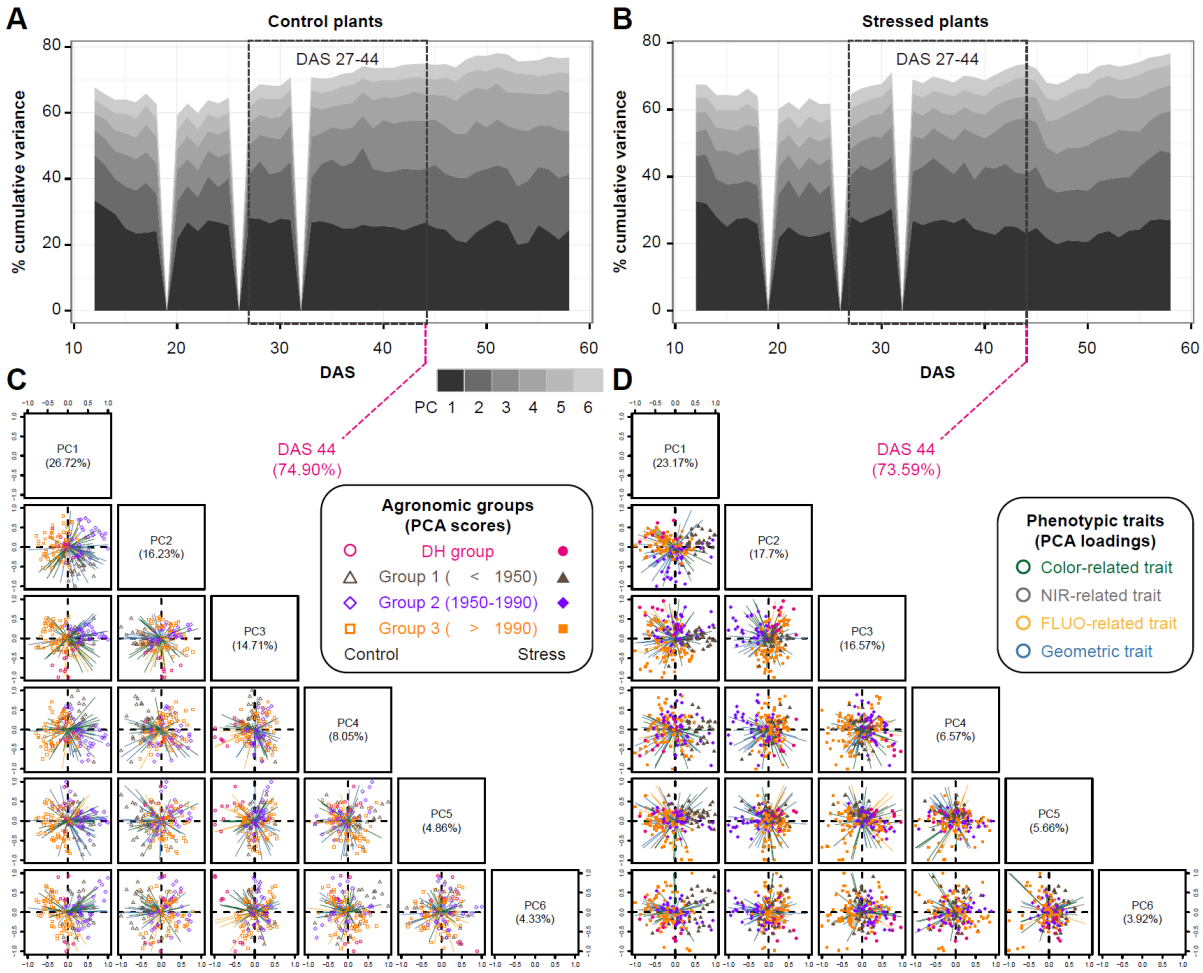

Figure 2.11: PCA performed on control and stressed plants, respectively

Related to Figure 2.9. Principal component analysis (PCA) of phenotypic variance over time for control **(A)** and stressed plants **(B)**. The percentage of total variance explained by the top six principal components is shown. The stress period is indicated by the dashed box. **(C)** and **(D)** Scatter plots showing the PCA results on DAS 44 (to compare the results of Figure 2.9). The first six PCs display 74.9% and 73.6% of the total phenotypic variance for control **(C)** and stressed plants **(D)**, respectively. The component scores (shown in points) are coloured and shaped according to the agronomic groups (as legend listed in the box). The component loading vectors (represented in lines) of each variable (traits as coloured according to their categories) were superimposed proportionally to their contribution. This figure was taken from Chen et al. (2014b). ■

phenotypic variance for each trait. The G effect showed dynamic behavior during plant growth (Figure 2.12C). In general, color and FLUO-related traits revealed strong G effects with high LOD scores over time. In contrast, geometric and NIR-related traits displayed strong G effects mostly in the middle stage of plant development. However, most of the phenotypic traits exhibited the E effects with significant LOD scores at the late period of drought stress or/and after the stress (Figure 2.12C). For example, traits such as fluorescence intensity, NIR intensity, area and volume were strongly affected by the E effects, agreeing with the known observations of decreased photosynthetic activity (Baker, 2008; Jansen et al., 2009; Woo et al., 2008), leaf water content (Seelig et al., 2008, 2009) and biomass accumulation (Berger et al., 2010; Rajendran et al., 2009) for plants under drought. In general, geometric traits, such as leaf length, plant

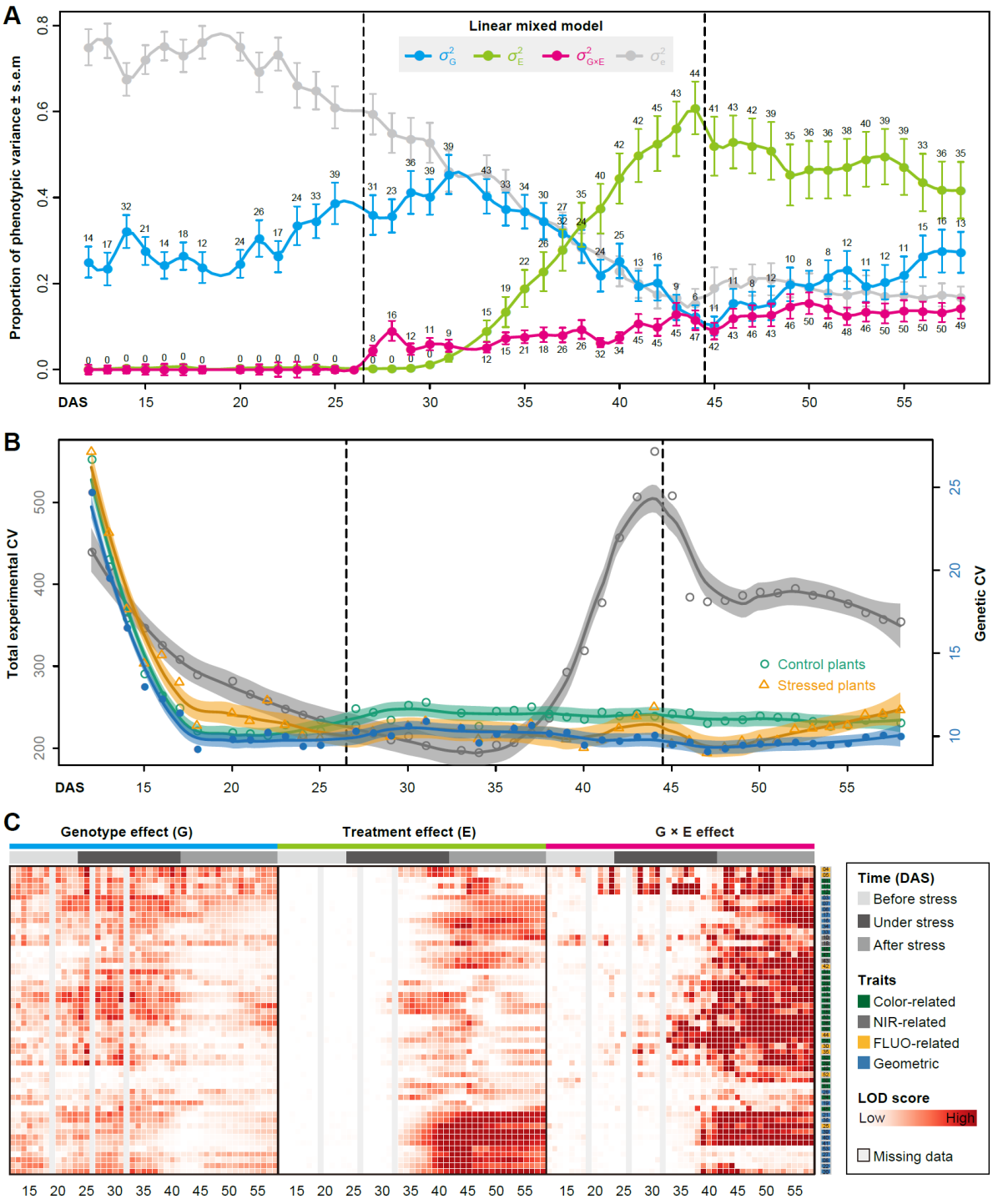

Figure 2.12: Dissection of the sources of phenotypic variance

This figure was taken from Chen et al. (2014b) (legend on next page).

height and projected area, showed strong and durable E effects, while only earlier E effects were seen for color-related traits. Nearly all traits were observed to have significant G×E effects ($P < 0.001$, log-LR test) at the recovery stage (Figure 2.12C), indicating that the impact of genetic factors for most traits is highly influenced by drought stress.

2.2.6 Change of heritability and trait-trait genetic and phenotypic correlations over growth time

Heritability of a trait and genetic correlations among traits are two key parameters that are used in plant breeding for making decisions concerning the design and selection of breeding schemes (Chen and Lubberstedt, 2010; Holland et al., 2003). It has been speculated that the dynamic change of heritability over time for a population is a consequence of changes in the magnitude of G and E effects (Visscher et al., 2008). However, most estimates of heritability are based on very few measures taken within specific growth stages (Busemeyer et al., 2013b; El-Lithy et al., 2004; Van Poecke et al., 2007). Recently, Zhang et al. (2012) used a HTP approach to document dynamic patterns of heritability of growth-related traits over growth time in Arabidopsis. Here, the change of broad-sense heritability (H^2) (Nyquist, 1991) was first investigated over barley growth time and with treatment. Consistent with the results of Zhang et al. (2012), the investigated traits showed dynamic changes in heritability during the entire plant growth stage (Figure 2.13A, left), as exemplified in the growth-related trait digital volume (Figure 2.13A, bottom right). Traits from different categories showed distinct patterns of heritability over time. It was found that heritability of E-sensitive traits, such as height, projected area, digital volume, leaf length and leaf numbers, decreased during drought stress, in agreement with previous findings that quantitative traits reflecting the performance of crops under drought conditions tend to have low to modest heritability (Tuberosa, 2012). Furthermore, it was found that geometric traits showed significantly higher heritability than physiological traits such as FLUO- and NIR-related traits ($P < 2.2 \times 10^{-16}$, Welch's t-test; Figure 2.13A, top right), indicating that variation in morphological traits during plant growth is governed in large part by genetic factors, rather than environmental factors.

Next, trait-trait genetic (r_g) and phenotypic correlations (r_p) were calculated during plant growth. The genetic correlations were calculated from a bivariate model (see Chapter 2.4.10) which allows testing of the genetic overlap between different traits, while the phenotypic correlations measure the observed phenotypic similarity of different traits. A correlation network was used to visualize the structure of ge-

► **Figure 2.12** (continued). **(A)** Dissecting the phenotypic variance over time by linear mixed models. For phenotypic data before stress treatment, $\sigma^2_{G\times E}$ is confounded with σ^2_e. Filled circles represent average variance of each component computed over all traits and solid lines represent a smoothing spline fit to the supplied data. Error bars represent the s.e.m. with 95% confidence intervals. The numbers of traits with significance at $P < 0.001$ are indicated above the bars. The stress period is indicated in dashed box. **(B)** The total experimental coefficient of variation (CV; coloured in grey) and genetic CV across lines (green for control, orange for stressed and blue for the whole set of plants) over time. Data points denote the average CV value over all geometric traits. Solid lines denote the loess smoothing curves and shadow represents the estimated standard error. **(C)** Statistical significance of genotype effect (left), treatment effect (middle) and their interaction effect (right), as detected by linear mixed models. The shading plot indicates the significance level (Bonferroni corrected p-values) in terms of LOD scores (-log probability or log of the odds score). Traits are sorted according to their overall effect patterns. Trait identifiers are listed on the right, which are given according to Figure 2.13A. G, genotype; E, environment (treatment); DAS, day after sowing; FLUO, fluorescence; NIR, near-infrared. ■

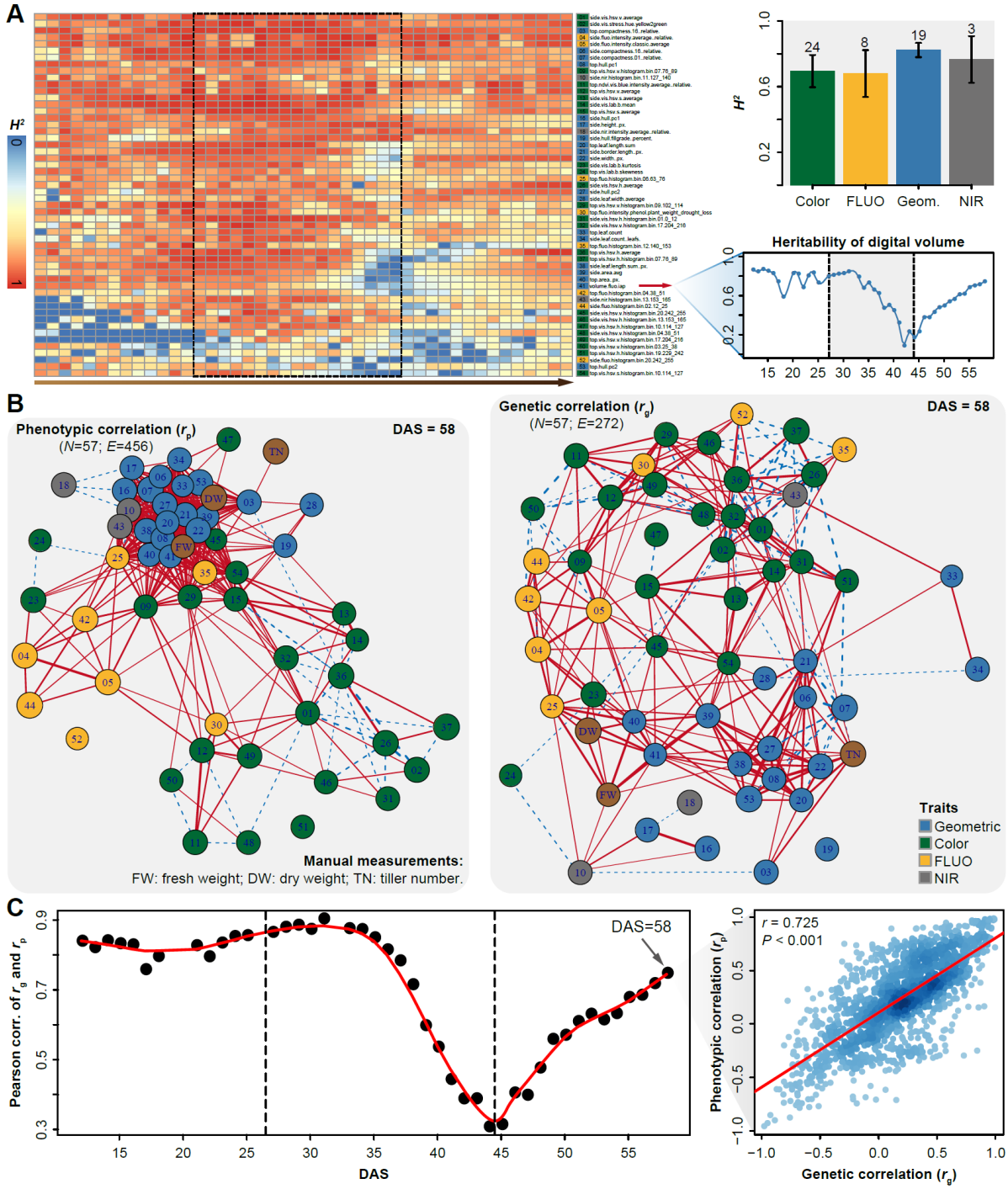

Figure 2.13: Trait heritability and trait-trait genetic and phenotypic correlations

This figure was taken from Chen et al. (2014b) (legend on next page).

netic and phenotypic correlations at the harvesting period (DAS 58/59), where the manual measurements (such as fresh weight [FW], dry weight [DW] and tiller number [TN]) were included as well (Figure 2.13B). As expected, these two correlation matrices correlated well with each other ($r = 0.73$ and $P < 0.001$, Mantel test; Figure 2.13C). Traits of the same category showed strong and positive genetic and phenotypic correlations. However, color-related traits were either not correlated or negatively correlated with other traits (Figure 2.13B), indicating that the variation in these traits has an independent genetic basis

from other traits. FW and DW showed the highest correlation with the predicted volume trait, both genetically and phenotypically ($r_g = 0.94$ and $r_p = 0.97$ for FW; $r_g = 0.79$ and $r_p = 0.95$ for DW), suggesting that the volume trait is a good image-derived estimate of plant biomass. Intriguingly, TN and plant compactness detected from top view images showed significant genetic and phenotypic correlations ($r_g = 0.77$ and $r_p = 0.52$), suggesting pleiotropy between barley TN and compactness. Finally, genetic and phenotypic correlations were computed over time (Figure 2.13C). The correlation pattern dynamically changed according to the intensity of the external stress, with decreasing correlation during the drought period and the lowest correlation ($r = 0.31$) at the end of stress period. This observation indicates that the extent of genetic influence on most traits was low when plants faced serious stress, thus supporting the hypothesis that plants exhibit extensive phenotypic plasticity in response to environmental stress (Sultan, 2000).

2.3 Discussion

High-throughput, automated digital imaging is a powerful tool to help alleviate the phenotyping bottleneck in plants (Furbank and Tester, 2011), as demonstrated by recent studies of plant/root growth and development using a variety of HTP systems (Meijon et al., 2014; Moore et al., 2013; Slovak et al., 2014; Yang et al., 2014; Zhang et al., 2012). In the emerging era of plant phenomics, we urgently need automated, rapid and robust analytical methods for large-scale processing of image data and extraction of extended features, as well as appropriate analysis frameworks for data interpretation (Fiorani and Schurr, 2013). A general framework was herein developed to meet these requirements, both in terms of image processing and post-processing of phenotypic data. As proof of concept, the methodology was validated by using phenotypic data of barley cultivars collected in an automated plant transport and imaging platform. This framework is readily extensible to the analysis of other plant species (such as Arabidopsis, maize and wheat) and other sensors (such as visible, NIR, FLUO cameras).

Plants reveal complex phenotypic traits which are expected to be extremely highly dimensional (Dhondt et al., 2013; Houle et al., 2010). Increasing the number of phenotypic measurements by image feature extraction is an important goal in phenomics. As reported here, the pipeline presented here is capable of parallel processing of image data from multiple sensors, and supports the extraction of a large

► **Figure 2.13** (continued). **(A)** Heatmap showing broad-sense heritability (H^2) of the investigated phenotypic traits over time (left), as exemplified by the digital volume (bottom right). Box plot (top right) shows the average heritability of phenotypic traits from the four categories (right). Error bars, s.e.m. with 95% confidence intervals. **(B)** Network visualizing significant phenotypic (r_p; left) and genetic (r_g; right) correlations among the 54 image-derived traits and three manual measurements (brown nodes). For visualization purpose, only significant correlations are shown ($P < 0.01$ for r_g and r_p, and $r_p > 0.5$). Trait identifiers are given as in **(A)** and coloured according to their classification as indicated. Positive correlations are shown by solid lines in red and negative correlations are shown by dashed lines in blue. **(C)** Pearsons correlation of r_g and r_p over time. The test of relationship between matrices of r_g and r_p was performed using Mantels test, as exemplifying on the right panel. ■

number of relevant traits (Klukas et al., 2014). The number of traits, including image-based features and model-derived parameters, extracted from the pipeline greatly exceeds existing pipelines (Camargo et al., 2014; De Vylder et al., 2012; Green et al., 2012; Hartmann et al., 2011; Paproki et al., 2012; Wang et al., 2009; Zhang et al., 2012). Sophisticated methods were applied to select a list of representative traits that are powerful in revealing descriptive phenotypic patterns of plants. It was shown that (1) there are clearly different patterns of phenotypic profiles for plants from different treatments (Figure 2.8A), individual genotypes (Figure 2.8B) and also from different agronomic groups (Figure 2.9 and 2.11); and (2) most of the traits reflected variable treatment effects (Figure 2.12) and even individual traits revealed genotypic differences in the response to drought and in the recovery process (Figure 2.2D).

Furthermore, the dynamic patterns of various phenotypic traits provided a snapshot of the complex dynamic process of plant growth (Figure 2.13), implying dynamic genetic control underlying phenotypic plasticity of plant development. The time-lapse phenotypic data provides a solid basis for functional mapping of dynamic QTLs underlying trait formation, by incorporating development features (estimated from mathematical models; see Chapter 3) of trait formation into the statistical framework for QTL mapping (Wu and Lin, 2006). Indeed, the pipeline is flexible enough to use in large panels of mapping populations and is easy to integrate into existing pipelines (as developed in R) for association mapping (Aulchenko et al., 2007; Kang et al., 2008; Lipka et al., 2012).

Dissecting phenotypic components of complex agronomic traits such as those associated with crop yield and stress tolerance can be achieved by model-assisted methods (called "the dissection approach"), in which complex phenotypes are dissected into more simple and heritable traits (Tardieu and Tuberosa, 2010). Such attempts have been made previously to dissect the sensitivity of flowering time to environmental conditions (Reymond et al., 2003; Yin et al., 2005a,b). In this study, as a further step towards biological insights from such image-derived parameters, genetic correlations were calculated between traits, such as might be considered for selection of desired phenotypic trait combinations in breeding programs (Chen and Lubberstedt, 2010; Porth et al., 2013; Stackpole et al., 2011). The identification of a concerted negative genetic correlation of an indicator of water content/drought tolerance (NIR signal; Figure 2.2D) with plant height (Figure 2.13B) appears to be highly advantageous for breeding strategies: breeding for higher drought tolerance could simultaneously select lower plant height, and vice versa. From a practical perspective, genetically correlated traits can be considered as proxies of the target trait in association genetic analyses, when measurements of the target trait are more time and/or labor intensive. In this case, the image-derived parameters plant volume and compactness are potential proxies for biomass and tiller numbers, respectively (Figure 2.13B).

Altogether, the analysis framework presented here will help to bridge the gap between plant phenomics and genomics aiming at a methodology to efficiently unravel genes controlling complex traits.

2.4 Materials and methods

2.4.1 Plant materials and growth conditions

The methodology was applied on a barley panel and produced a phenotypic map for barley plants from 18 genotypes (Table 2.1) under control and drought-stress conditions over time. A LemnaTec HTS-Scanalyzer 3D platform was used to screen 16 German two-rowed spring barley cultivars (cv.) and two parents of a Double Haploid (DH)-mapping population (cv. Morex and cv. Barke) for vegetative drought tolerance. The 16 genotypes can be divided into three agronomic groups according to their breeding history: group 1 (released before 1950), group 2 (released between 1950 and 1990) and group 3 (released after 1990). The parental cultivars are considered as an independent group (DH group). Nine plants per genotype and treatment for the 16 German cultivars and 6 plants for the DH-parents were investigated during one experiment from May to July 2011. Plants grew under controlled greenhouse conditions and were phenotyped on a daily basis over the entire experimental phase using the fully automated system consisting of conveyer belts, a weighing and watering station and three imaging sensors. The growth conditions in the greenhouse were set to 18°C during day time and 16°C at night. The day light period lasted about 13 hours starting from 7 am.

Two seeds of each cultivar were sown per pot (two litre in volume; 19.5 cm in height; 14.5 cm in diameter; http://www.berryplastics.com), and the pots were kept at a field capacity (FC) level of 90% by using the automated target-weight watering option of the system. After seven days, plants were thinned out to one plant per pot. Subsequently, 200 g of blue coloured quartz sand was added to each pot as a cover layer, reducing the evaporation and providing a uniform blue background for image analysis. Blue-coloured supports were used to stabilize plants and prevent leaf damage during automatic shunting of the pots. The FC was determined by filling 10 pots with 970 g of substrate ("Klasmann Substrate no. 2", http://www.klasmann-deilmann.com) and watering carefully to saturation and weighing 2 days after saturation. Substrate of each pot was then dried for one week at 80°C and weighed again, thus representing the weight of soil alone. Field capacity was calculated as the difference in weight between dry and soaked soil.

Drought stress was applied four weeks after sowing by withholding water. Control plants remained well-watered at a FC of 90%. After a stress period of 18 days plants were re-watered to 90% FC and kept well-watered again for another two weeks. For each plant, top and side cameras were used to capture images daily at three different wavelength bands: visible light, fluorescence and near-infrared (Figure 2.2B-C). In this manner, thousands of images were acquired for each genotype and treatment during the whole phenotyping period.

In addition, several manual measurements were collected for each plant. Above-ground biomass of each plant was measured as plant fresh weight and dry weight at DAS 58. Tiller number was counted manually for each plant at three time points: DAS 27, DAS 45, and DAS 58.

2.4.2 Image analysis

The barley analysis pipeline that was implemented in the IAP software (v0.94; Klukas et al., 2014) was used to perform the image processing operations (Figure 2.2A). Briefly, image datasets and the corresponding metadata were automatically loaded into the IAP system from the LemnaTec database by using the built-in IAP functionality. The structured image data analysis was performed using the barley analysis pipeline with optimized parameters. Image processing included four main steps: (1) pre-processing – to prepare the images for segmentation, (2) segmentation – to divide the image into different parts which have different meanings (for example, foreground the plant part; background – imaging chamber and machinery), (3) feature extraction – to classify the segmentation result and get a trait list, and (4) post-processing – to summarize calculated results for each plant. The analysis was performed in a grid-computing mode to speed up image processing. Analyzed results were exported in csv file format via IAP functionalities, which can be used for further data inspection (see Online Data Set 1 in Appendix C). The resulting spreadsheet includes columns for different phenotypic traits and rows for data from different time points. The corresponding metadata is included in the result table as well. Depending on the computing resource available, IAP can process large-scale image data in a reasonable time ranging from a few hours to a few days (Klukas et al., 2014). An image dataset of the size used in this study can be processed within three days on a local PC with 6 gigabytes (GB) of system memory using four central processing unit (CPU) cores.

Each plant was characterized by a set of 388 phenotypic traits, also referred to as features, which were grouped into four categories: 60 geometric features, 100 fluorescence-related (FLUO-related) features, 182 color-related features, and 46 near-infrared-related (NIR-related) features. These traits were defined by considering image information from different cameras (visible light, fluorescence and near infrared) and imaging views (side and top views). See the IAP online documentation for details about the trait definition.

2.4.3 Feature preprocessing

The preprocessing of phenotypic data involves outlier detection and trait reproducibility assessment. Defects may be introduced during the imaging period or in the image processing steps. Grubbs test (Grubbs, 1950) was first adopted to detect outliers based on the assumption of normal distribution of phenotypic data points for repeated measures on replicated plants of a single genotype for each trait. Grubbs test can be used to detect if a particular sample contains one outlier ($P < 0.01$) at a time. The outlier was expunged from the dataset and the test was iterated until no outliers were detected.

Next, it was reasoned that phenotypic information should be robust and informative enough (rather than noise) to infer differences in genotype or treatment in terms of higher reproducibility over replicated plants in comparison to random samples of plants. The reproducibility of phenotypic traits was evaluated by the Pearson correlation coefficient. The correlation coefficient values were computed over each pair of replicated plants (from the same genotype) for each treatment. For comparison, correlation values over

two sets of plants (with the same size) were calculated from two randomly selected genotypes. The traits were considered as highly reproducible if (1) the median correlation coefficient over genotypes was larger than 0.8, and (2) the coefficients were significantly higher in replicates than in random plant pairs (Welch's t-test $P < 0.001$). The above criteria should be satisfied in at least one treatment condition. Therefore, the original 388 traits were reduced to 217 highly reproducible ones. After removing redundancy, 173 high-quality traits (Figure 2.2A) were retained and used for further analyses.

Plants with empty values were discarded for analysis. A phenotypic matrix (whose rows represented phenotyped plants over time and whose columns indicated highly reproducible traits) was obtained. The phenotypic profile was further normalized (if necessary) to zero mean and unit variance, computed for all phenotyped plants over time.

2.4.4 Feature selection

The resulting datasets may contain many redundant features (phenotypic traits) which are correlated with each other. To reduce the excessive correlation among explanatory variables, the so-called "multi-collinearity", a method was implemented to select an optimal set of explanatory variables for a statistical model. This process is accomplished with stepwise variable selection using variance inflation factors (*VIF*s), which is defined as

$$VIF_i = \frac{1}{1 - R_i^2}$$

where the *VIF* for variable X_i is obtained using the coefficient of determination (R^2) of the regression of that variable against all other explanatory variables. Specifically, a *VIF* value is first calculated for each variable using the full set of explanatory variables,and the variable with the highest value is removed. Next, all *VIF* values with the new set of variables are recalculated, and the variable with the next highest value is removed, and so on. The above procedure is repeated until all values are below the desired threshold. As a general rule, $VIF < 5$ was considered as a cut off value for the high multicollinearity problem. The "VIF" function was implemented in the *fmsb* R package to calculate *VIF*.

2.4.5 Hierarchical clustering analysis and PCA

Hierarchical clustering analysis (HCA) and principle component analysis (PCA) were carried out to visualize the data globally. HCA builds a hierarchy from individuals by progressively merging clusters, while PCA is a technique used to reduce dimensionality of the data by finding linear combinations (dimensions; in this case, the number of traits) of the original data.

To identify plants from the same genotype or agronomic groups with similar phenotypic composition, HCA was performed with the normalized data based on the list of highly reproducible traits. All analyses were conducted with the complete linkage hierarchical clustering method and Euclidean distances and were visualized as a heatmap with a dendrogram by using the "heatmap.2" function of the corresponding R package.

PCA was carried out to characterize each plant based on phenotypic composition and to indicate the affiliations within the phenotypic diversity of four agronomic groups. PCA was performed using Bayesian principal component analysis (the "bpca" function) as implemented in the R package *pcaMethods* (Stacklies et al., 2007). The first six principal components (PCs 1-6) and the corresponding component loading vectors PCs 1-6) were visualized and summarized in scatter plots, in which principal components are coded in color and in shape according to genotypes of origin (control plants in blank points and stressed plants in filled points) and component loadings (indicated in lines) are coloured according to phenotypic classification. PCA was performed for control, stress, and the total list of plants, respectively (Figures 2.9, 2.10 and 2.11).

2.4.6 Phenotypic similarity tree and Mantel test

As phenotypic traits are derived from heritable characters, the influence of environmental factors and their interactions, it is possible to measure the phenotypic relationship of different genotypes based on the available traits. A "phenotypic similarity tree" was constructed to show the phenotypic relationship from a global perspective. Phenotypic similarity trees can be used to quantitatively describe the relationship of genotypes and phenotypes and to compare the differences of phenotypes under different conditions (Zhao et al., 2011). Genotypes from the DH groups were excluded from the phenotypic similarity tree analysis.

First, a phenotypic profile for each genotype was calculated as the average value from replicated plants. Next, a phenotypic distance (based on the Euclidean measure) matrix of pairwise comparisons between genotypes was estimated based on the normalized phenotypic profile. The above analysis was performed for control and stressed plants, respectively. For stressed plants, only data after DAS 34 were taken into consideration because from that time point stressed plants showed differences in their phenotypes from control plants (see the below SVM method). Finally, the phenotypic similarity trees were generated based on the distance matrices using the function "plot.phylo" implemented in the R package *ape* (Paradis et al., 2004).

A Mantel test (Mantel, 1967) was performed to examine the extent of correlation of the phenotypic distances between the control and stress plant sets. A positive correlation would be expected in the case that plants maintain their phenotypic similarity in different environments. The phenotypic distance matrixes from above was used to conduct the analysis. The Mantel test was computed using the function "mantel" in the corresponding R package with 10,000 permutations (Monte-Carlo simulation) and selecting Pearsons correlation method.

2.4.7 Plant classification using SVM

Based on their phenotypic traits (features), plants from the same genotype were classified into control and stress groups (Figure 2.7A), using the pairwise classification strategy of the support vector machine (SVM) algorithm as provided by the libsvm library (Chang and Lin, 2011) via the R package *e1071*.

The SVM classifier was used to find "optimal" hyperplanes separating two groups of plants in the multi-dimensional feature space. Using a linear kernel, the SVM parameters were optimized through 2-fold cross-validation to maximize the accuracy rate for classification and to minimize the mean squared error for regression. Specifically, a classifier was trained on a randomly chosen subset of half of the images (˜9 images) from one specific genotype or treatment from one specific day (the training set) and then used the classifier to validate the other half of the images (the validation set).

2.4.8 Analysis of phenotypic variance

The observed variance in a particular phenotypic variable (trait) can be partitioned into components attributable to different sources of variation, for example, the variation of genotype (G), environment (E) and their interaction (G×E). The analysis of variance was performed by using linear mixed model (LMM) for each phenotype trait measured in each day, as defined:

$$y = X\beta + Z\mu + \varepsilon$$

where y denotes a vector of individual plant observations of a given trait; X and Z are incidence matrices associating observations with fixed effects (in vector β) and random effects (in vector μ), respectively; ε is the vector of random residuals assuming $\varepsilon \sim (0, I\sigma_\varepsilon^2)$ (I is the identity matrix). Variance components for each trait, such as genotypic effect $g \sim (0, I\sigma_G^2)$, environment effect $e \sim (0, I\sigma_E^2)$ and their interaction effect $ge \sim (0, I\sigma_{GE}^2)$, were estimated in the LMM using residual maximum likelihood (REML), as implemented in ASReml-R v.3.0 (Gilmour et al., 2009). The statistical significance of variance components was estimated by the log-likelihood ratio test (log-LR test). The statistic for the log-LR test (denoted by D) is twice the difference in the log-likelihoods of two models:

$$D = 2\left(\log(L_{alt}) - \log(L_{null})\right)$$

where $\log(L_{alt}$ is log-likelihood of the alternative model (with more parameters) and $\log(L_{null}$ is log-likelihood of the null model, and both log-likelihoods can be calculated from the ASReml mixed model. Under the null hypothesis of zero correlation, the test statistic was assumed to be χ^2-distributed with degrees of freedom equal the difference in number of covariance parameters estimated in the alternative versus null models. Resulting p-values from LMM were corrected for multiple comparisons with the Benjamini-Hochberg false discovery rate (FDR) method (Benjamini and Hochberg, 1995). The LOD (log of odds) scores were further calculated as the log probability (corrected p-value) (Joosen et al., 2013). Hierarchical clustering was applied to the matrix of LOD scores consisting traits as rows and imaging days as columns.

As a relative indicator of dispersion, the coefficient of genetic variance (CV_g) was calculated as the ratio of the standard deviation (square root of the among-genotype variance) to the mean of the corresponding trait value across all genotypes. This analysis was respectively performed for control plants, stress plants and the whole set of plants (based on the mean value of control and stress plants). Similarly, the total experimental CV (CV_e) was calculated as the sum of the square root of the experimental variance, including controlled (i.e., treatment effect) and uncontrolled variation, to the mean of trait

value for one specific genotype. Since CV is only reasonable to be calculated for data measured on a ratio scale (rather an interval scale), only geometric traits were considered in this calculation.

2.4.9 Broad-sense heritability

The broad-sense heritability (H^2) of a trait is the proportion of the total (phenotypic) variance (σ_P^2) that is explained by the total genotypic variance (σ_G^2) (Nyquist, 1991), which was calculated as follows:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2/2 + \sigma_e^2/2r}$$

where r is the average number of replications.

2.4.10 Estimation of genetic and phenotypic correlations

A bivariate LMM was used to estimate genetic correlations between each pair to traits (the proportion of variance that two traits share due to genetic causes) in each day. Assuming $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$ as the response vector for the subject i with Y_i^k the vector of measurement of the trait k ($k = 1, 2$), the bivariate model is defined as follows:

$$y_i = X_i\beta + Z_i\mu_i + \varepsilon_i \text{ with } \begin{cases} \mu_i \sim N(0, G) \\ \varepsilon_i \sim N(0, R) \end{cases}$$

where the genetic covariate matrix $G = \begin{bmatrix} \sigma_{g^1}^2 & cov_{g^1g^2} \\ cov_{g^1g^2} & \sigma_{g^2}^2 \end{bmatrix}$ and the covariance matrix of measurement errors $R = \begin{bmatrix} \sigma_{\varepsilon^1}^2 & cov_{\varepsilon^1\varepsilon^2} \\ cov_{\varepsilon^1\varepsilon^2} & \sigma_{\varepsilon^2}^2 \end{bmatrix}$. With the assumption that μ_i and ε_i are mutually independent, it is apparent that $Var(Y_i) = Z_iG_iZ_i^T + R$. The genetic correlation between pairs of traits was estimated as $r_g = \frac{cov_{g^1g^2}}{\sqrt{\sigma_{g^1}^2\sigma_{g^2}^2}}$. The significance of the genetic correlation was estimated using the log-LR test by comparing the likelihood of the model allowing genetic co-variance between the two traits to vary and the likelihood of the model with the genetic co-variance fixed to zero. The above analyses were performed in ASReml-R v.3.0 (Gilmour et al., 2009).

Phenotypic correlations r_p among different traits were calculated by Pearson correlation. The significance of the correlations was tested using the "cor.test" function in R.

To test the relationship between matrices of genetic and phenotypic correlations, a Mantel test (Mantel, 1967) was performed for the correlations in each day. The genetic and phenotypic correlations were visualized in networks. For visualization purpose, only significant correlations were shown ($P < 0.01$).

Chapter 3

Plant growth modeling based on time-lapse image data

3.1 Introduction

The most attractive advantage of non-destructive automated imaging techniques is the possibility to repeatedly measure the same plants over time, allowing novel insights into the high dynamics of plant growth (Schunk and Eberius, 2012). In this way, plants can be phenotyped extensively (towards comprehensive measurement) and intensively (e.g., population-wide, time-lapse). In high-throughput phenotyping (HTP), each plant is measured repeatedly. Through image analysis, various image-derived parameters (phenotypic traits) can be obtained for a single plant at one time point, allowing the detection of significant phenotypic differences among plants with varied genetic background or under different environmental conditions. From these measurements, distinct, biologically relevant parameters may be determined. For example, it is now possible to study plant biomass accumulation (e.g., dry weight) and growth rate in various growth phases. The advances in HTP in turn help in the development and application of growth models for plants by taking the environmental influence into consideration.

The complexity of plant growth has been long recognized (Blackman, 1919; Erickson, 1976; Gompertz, 1825; Hunt, 1982; Karkach, 2006). Many mechanistic growth models have been established to model the laws of plant growth (Archontoulis and Miguez, 2013; Karkach, 2006; Paine et al., 2012; Thornley and France, 2007), which aim to provide the simplest description that accurately captures the growth dynamics of individuals. It is well known that plant growth follows a sigmoidal growth curve (Damgaard and Weiner, 2008; Hunt, 1982; Vanclay, 1994). Several sigmoidal growth models, such as the logistic and Gompertz models (Karadavut et al., 2008, 2010), with biologically interpretable parameters have been proposed to probe the growth of individual plants. These advances in plant growth modeling have allowed a deeper understanding of relationships between plants and their abiotic environment (Paine et al., 2012).

In the following chapter, several linear and nonlinear functions were used to model biomass accumulation for barley and maize plants under control and/or stressed conditions. The established growth

models allow biological interpretation of parameters. Model-based parameters revealed several important aspects regarding plant development, and provide a solid basis for subsequent QTL (quantitative trait locus) analysis aimed at understanding the genetic control of plant growth.

3.2 Results

3.2.1 Modeling barley plant growth under normal conditions

Time-lapse phenotypic data generated by HTP provide valuable information to study plant growth. Of all the phenotypic traits investigated, the image-based volume, which combined information from both side and top views of cameras, had the best correlation with manual measurements of biomass, such as fresh weight (FW) and dry weight (DW; Figure 3.1). The image-derived volume estimate was thus used to model plant growth and considered it as a proxy measure of plant above-ground biomass.

Firstly, time-lapse phenotypic data were used to model and predict plant growth under normal growth conditions in barley. It has been shown that the growth of Arabidopsis plants follows the logistic model (Paul-Victor et al., 2010; Tessmer et al., 2013; Zust et al., 2011), while the growth of maize kernels prefers to fit the Gompertz model (Meade et al., 2013). However, the pattern of barley growth is poorly investigated. In order to determine a suitable growth curve of biomass accumulation for barley plants under control conditions, five nonlinear mechanistic models including exponential, monomolecular (Richards, 1959), logistic (Verhulst, 1977), Gompertz (Gompertz, 1825) and Weibull (Weibull, 1951) curves (Table 3.1) were compared. Of these models, logistic, Gompertz and Weibull models are sigmoid functions (with an S shape and an inflection point) and are often applied to describe plant growth as a function of time (Archontoulis and Miguez, 2013). To implement these models in an efficient way, the nonlinear relationship of the models was transformed into linearized forms (Table 3.1) and fitted these linearized models based on the existing linear regression approach. The fitting quality of models was determined based on the criteria of (1) the coefficient of determination (R^2; based on the linearized model), (2) the root mean squared relative error ($RMSRE$), and (3) the Pearson correlation coefficient (r) between the predicted values and the observed values (see Chapter 3.4.3).

The results indicated that the Weibull model $y = K - (K - y_0)\, e^{-rt^m}$ has performed better than the other models to simulate biomass accumulation over time (Figure 3.2A), in terms of the lowest $RMSRE$, the highest R^2 as well as r (Figure 3.3A-C), and the best predictability of real biomass (Figure 3.3D). It was found that the predicted digital biomass from the Weibull model and the image-based digital volume showed about the same correlations with FW ($r = 0.891$ versus 0.892; Figure 3.2B).

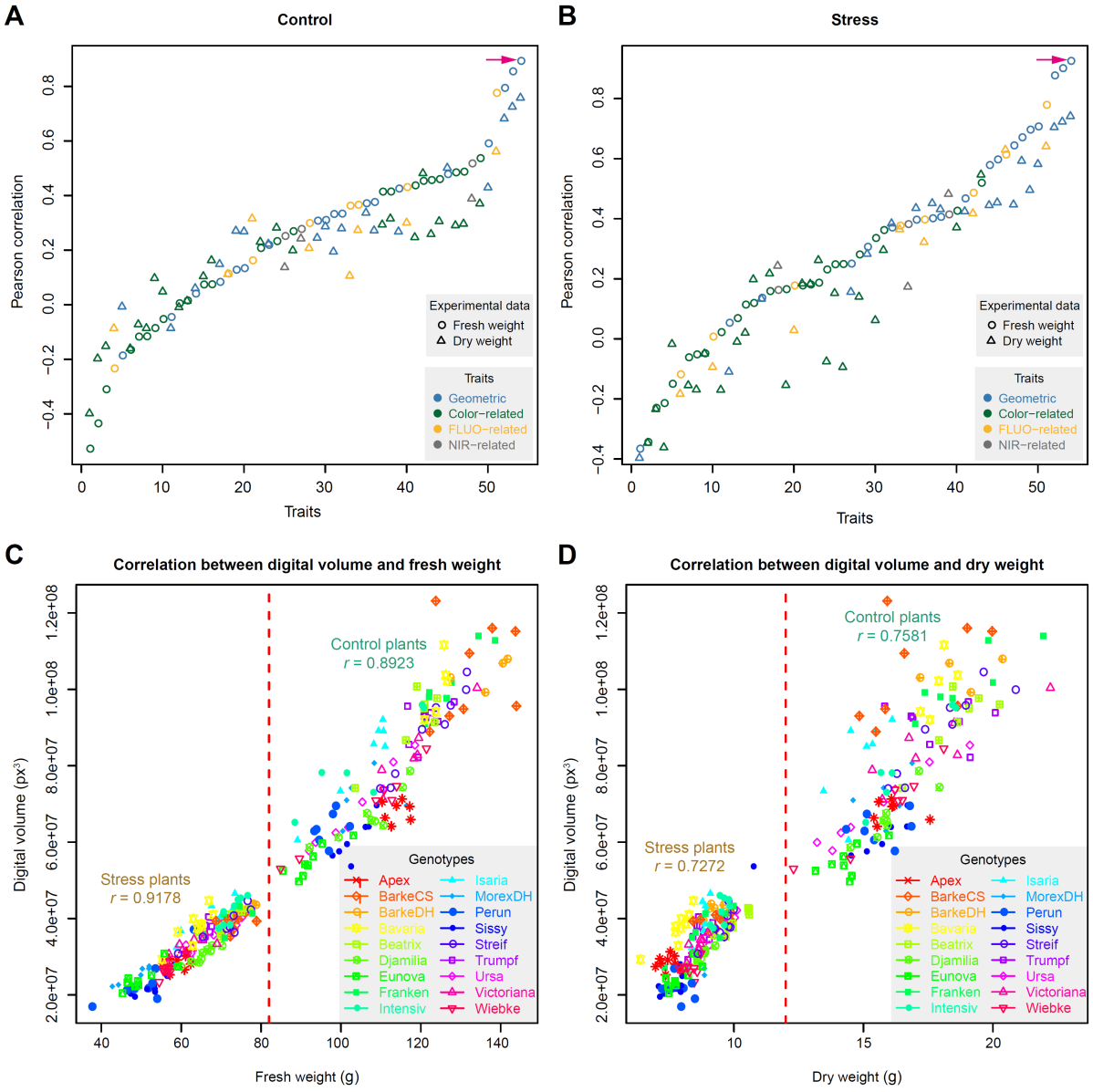

Figure 3.1: Correlation analysis of manual measurements with phenotypic traits

Correlation of all phenotypic traits with fresh weight and dry weight for control plants **(A)** and stressed plants **(B)**. The digital volume has the best correlation with manually measured biomass. For **(A)** and **(B)**, traits are coloured according their classification. Correlation analysis of digital volume with biomass for control plants **(C)** and stressed plants **(D)**. For **(C)** and **(D)**, data points are coloured according to the genotype origin of plants. The correlation coefficients are indicated for stress and control plant, respectively. This figure was taken from Chen et al. (2014b). ■

Table 3.1: Mechanistic models used for modeling biomass accumulation in this study.

	Model	Differential equation†	Analytical solution†	Linearized form†
Control plants	**Exponential**	$\frac{\mathrm{d}y}{\mathrm{d}t} = ry(t)$	$y = y_0 e^{rt}$	$\ln(y) = \ln(y_0) + rt$
	Monomolecular	$\frac{\mathrm{d}y}{\mathrm{d}t} = r\left(K - y(t)\right)$	$y = K - (K - y_0)\, e^{-rt}$	$\ln \frac{1}{K-y} = \ln \frac{1}{K-y_0} + rt$
	Gompertz	$\frac{\mathrm{d}y}{\mathrm{d}t} = ry(t) \ln \frac{K}{y(t)}$	$y = K\left(\frac{y_0}{K}\right)^{e^{-rt}}$	$-\ln\left(-\ln \frac{y}{K}\right) = -\ln\left(-\ln \frac{y_0}{K}\right) + rt$
	Logistic§	$\frac{\mathrm{d}y}{\mathrm{d}t} = ry(t)\left(1 - \frac{y(t)}{K}\right)$	$y = \frac{Ky_0}{y_0 + (K - y_0)\, e^{-rt}}$	$\ln \frac{y}{K-y} = \ln \frac{y_0}{K-y_0} + rt$
	Weibull¶	$\frac{\mathrm{d}y}{\mathrm{d}t} = rmt^{m-1}\left(K - y(t)\right)$	$y = K - (K - y_0)\, e^{-rt^m}$	$\ln\left(\ln \frac{K-y_0}{K-y}\right) = \ln(r) + m \ln(t)$
Stressed plants	**Quadratic**	$\frac{\mathrm{d}y}{\mathrm{d}t} = b - 2at$	$y = c + bt - at^2$	$y = c + bt - at^2$
	Bell-shaped 1	$\frac{\mathrm{d}y}{\mathrm{d}t} = 2Aa\,(t - t_{max})\, e^{a(t-t_{max})^2}$	$y = Ae^{a(t-t_{max})^2}$	$\ln(y) = \ln(A) + a\,(t - t_{max})^2$
	Bell-shaped 2	$\frac{\mathrm{d}y}{\mathrm{d}t} = A\,(b/t - a)\, t^b e^{-at}$	$y = At^b e^{-at}$	$\ln(y) = \ln(A) + b \ln(t) - at$
	Bell-shaped 3	$\frac{\mathrm{d}y}{\mathrm{d}t} = A(b - 2at)e^{bt-at^2}$	$y = Ae^{bt-at^2}$	$\ln(y) = \ln(A) + bt - at^2$
	Linear‡	$\frac{\mathrm{d}y}{\mathrm{d}t} = r$	$y = y_0 + rt$	$y = y_0 + rt$

† y is biomass; t denotes time; r is intrinsic growth rate for control plants or re-growth rate for stressed plants; K is upper asymptote of biomass for control plants in monomolecular, Gompertz, logistic and Weibull models; m determines the slope of growth in Weibull model; $t_{max} = \frac{b}{2a}$ is the time point (the center of the peak in bell-shaped curves) at which plant under stress shows the asymptotic maximum biomass (determined by A). Other parameters are constants.

§ Only the three-parameter version of logistic model was considered. In this model, the lower asymptote is fixed at 0 and the inflection point falls strictly at $y = K/2$.

¶ Weibull model with three parameters was considered, where $y_0 = 0$. The model can thus be simplified as $y = K\left(1 - e^{-rt^m}\right)$. It is reasonable in most cases. For example, at planting, the plant biomass is very close to zero (Archontoulis and Miguez, 2013).

‡ Linear growth for stressed plants is only modeled in the recovery phase.

Table 3.2: Calculation of absolute growth rate and relative growth rate.

Model	AGR, time basis†	RGR, time basis†	RGR, biomass basis†
Control plants			
Exponential	ry_0e^{rt}	r	r
Monomolecular	$r\left(K-y_0\right)e^{-rt}$	$\frac{r\left(K-y_0\right)}{y_0+K\left(e^{rt}-1\right)}$	$\frac{r(K-y)}{y}$
Gompertz	$rK\ln\left(\frac{K}{y_0}\right)e^{-rt}\left(\frac{y_0}{K}\right)^{e^{-rt}}$	$r\ln\left(\frac{K}{y_0}\right)e^{-rt}$	$r\ln\left(\frac{K}{y}\right)$
Logistic	$\frac{ry_0K\left(K-y_0\right)e^{-rt}}{\left(y_0+\left(K-y_0\right)e^{-rt}\right)^2}$	$\frac{r\left(K-y_0\right)e^{-rt}}{y_0+\left(K-y_0\right)e^{-rt}}$	$r\left(1-\frac{y}{K}\right)$
Weibull	$rm\left(K-y_0\right)t^{m-1}e^{-rt^m}$	$\frac{rm\left(K-y_0\right)t^{m-1}}{Ke^{rt^m}-\left(K-y_0\right)}$	$rm\left(\frac{1}{r}\ln\frac{K-y_0}{K-y}\right)^{\frac{m-1}{m}}\frac{K-y}{y}$
Stressed plants			
Quadratic	$b-2at$	$\frac{b-2at}{c+bt-at^2}$	$\pm\sqrt{b^2+4a(c-y)}$
Bell-shaped 1	$2Aa\left(t-t_{max}\right)e^{a(t-t_{max})^2}$	$2a\left(t-t_{max}\right)$	$\pm2\sqrt{a\ln\frac{y}{A}}$
Bell-shaped 2	$A\left(b/t-a\right)t^be^{-at}$	$b/t-a$	NA
Bell-shaped 3	$A(b-2at)e^{bt-at^2}$	$b-2at$	$\pm\sqrt{b^2-4a\ln\frac{y}{A}}$
Linear	r	$\frac{r}{y_0+rt}$	$\frac{r}{y}$

† AGR: absolute growth rate ($\mathrm{d}y/\mathrm{d}t$); RGR: relative growth rate ($\mathrm{d}y/\mathrm{d}t/y$). RGR can be expressed either as a function of biomass or as a function of time. For Quadratic and Bell-shaped models, the sign of RGR (biomass basis) is determined by "+" when $t \leq t_{max}$ and "-" when $t > t_{max}$. Refer to Table 3.1 for explanations of other symbols. *NA*: not available.

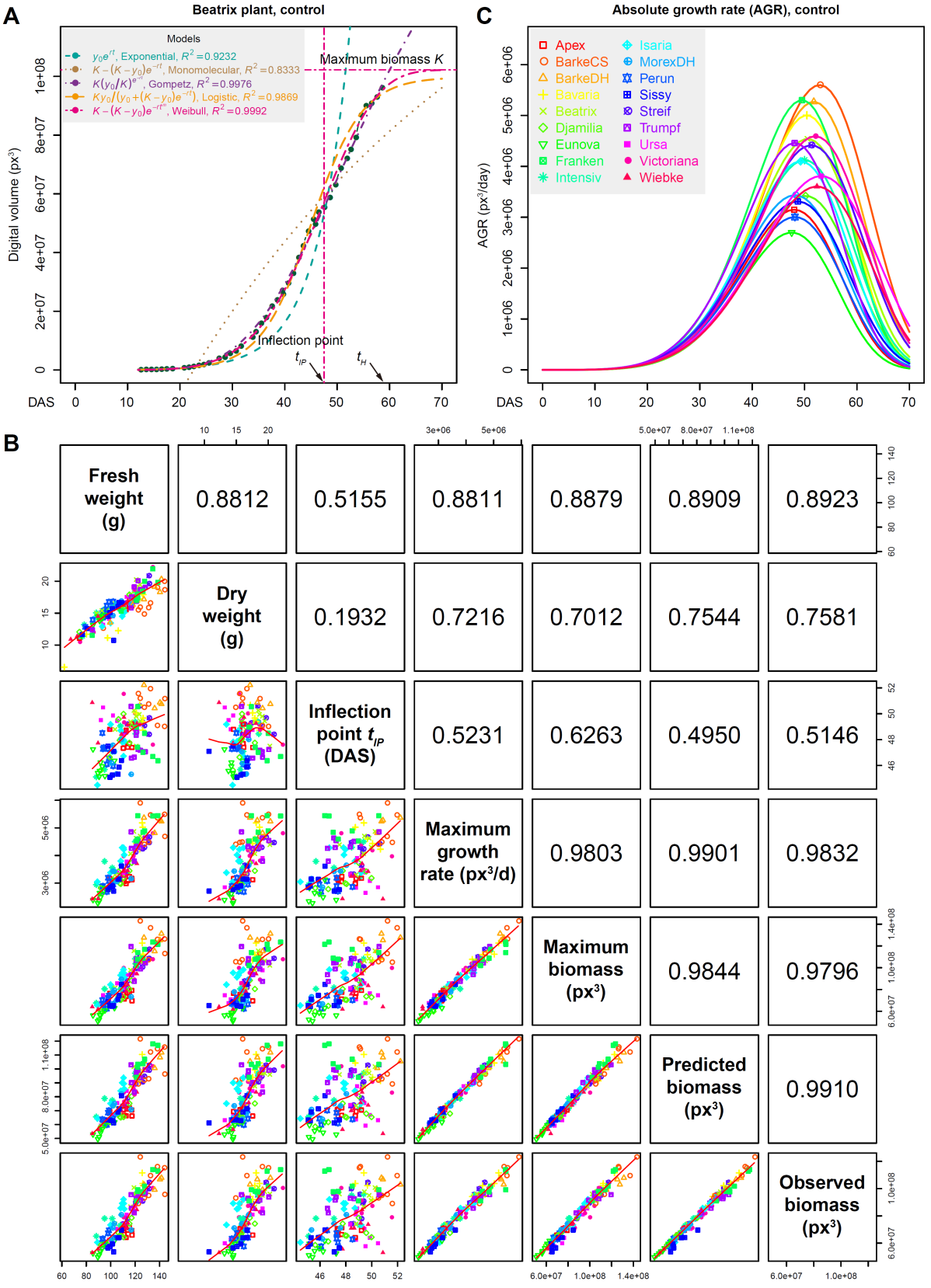

Figure 3.2: Growth modeling of barley plants under normal conditions

(legend on next page).

Once the best growth curve has been selected, two forms of growth rates, absolute growth rate (AGR)

and relative growth rate (RGR), can be calculated on the basis of time or biomass (Table 3.2). In the investigation of barley growth, estimating plant growth rate as a free parameter from the Weibull model seems biologically reasonable since there is no general accepted approach that measures the plant growth rate over time. Furthermore, it is much easier to compare plant growth rates (whether AGR or RGR) among different genotypes by comparisons of time (or biomass)-specific functions, rather than comparing time-point estimates of growth rates (Paine et al., 2012). Using function-derived growth rates, we can test the degree to which plants differ in terms of the timing and magnitude of AGR or RGR peaks. When plotting AGR as functions of time, it was found that genotypes show distinct patterns of AGR along plant growth and their differences in timing and magnitude are significant (Figure 3.2C).

The Weibull model can also be used to determine the inflection point ($t_{IP} = \left(\frac{m-1}{rm}\right)^{\frac{1}{m}}$) at which individuals exhibit their maximum AGR (R_{IP}). The mean values of R_{IP} within genotypes ranged from $2.59 \times 10^5\, px^3/day$ (Eunova) to $5.17 \times 10^5\, px^3/day$ (BarkeDH). The inflection point splits the growth curve into two stages with opposite growth dynamics, initially exponential growth and gradually reduced relative growth rate as plants reach their asymptotic maximum growth capacity (Figure 3.2C; Zeide, 1993). Notably, it was observed that the maximum growth rate is highly correlated with FW ($r = 0.88$; Figure 3.2B), indicating its significant impact on crop biomass yield. However, the exact inflection time-point has less impact on the biomass accumulation ($r = 0.52$).

3.2.2 Modeling barley plant growth under drought stress conditions

Modeling plant growth under stress conditions is more complex. To my knowledge, there are no previous studies attempting to model stressed plant growth. It would be very attractive to study plant stress response based on growth model-derived parameters. According to the observations of plant growth patterns from image data, plant growth can be divided into two parts describing the stress period (bell-shaped growth curve) and the recovery phase (linear re-growth model) (Tables 3.1 and 3.2). Of the three tested bell-shaped curves, the bell-shaped model $y = Ae^{bt-at^2}$ (model 3) fitted best for stressed plants that underwent wilting with a concomitant decrease in estimated volume (Figures 3.4A and 3.5).

► **Figure 3.2** (continued). **(A)** Plant growth prediction based on fitting of the digital volume by using five different mechanistic models. The best-fitted model — Weibull model — can be considered as the growth curve of barley plants. Several Weibull-model derived parameters such as the "inflection point" (t_{IP}, a time-point with the maximum absolute growth rate) and "maximum biomass" (the maximum growth capacity; parameter K) are indicated. Dots in green represent data points derived from images and curves represent the least-squares fit to the observed data. Shown is the result of fitting for a Beatrix plant. See also Online Data Set 2 in Appendix C for growth modeling for all plants. t_H: the time point for harvesting. **(B)** Pairwise comparison of model-derived parameters, image-derived data and manually determined fresh weight or dry weight for control plants. Each point in the dot plots (bottom-left quadrants) represents one plant from a specific genotype as coloured and labeled in **(C)**. Pearson's correlation coefficients are indicated in top-right quadrants. **(C)** The absolute growth rate (AGR) derived from the Weibull models, which were fitted at the genotype level. The t_{IP} time points are indicated by dots. Different genotypes are indicated by different colors. ■

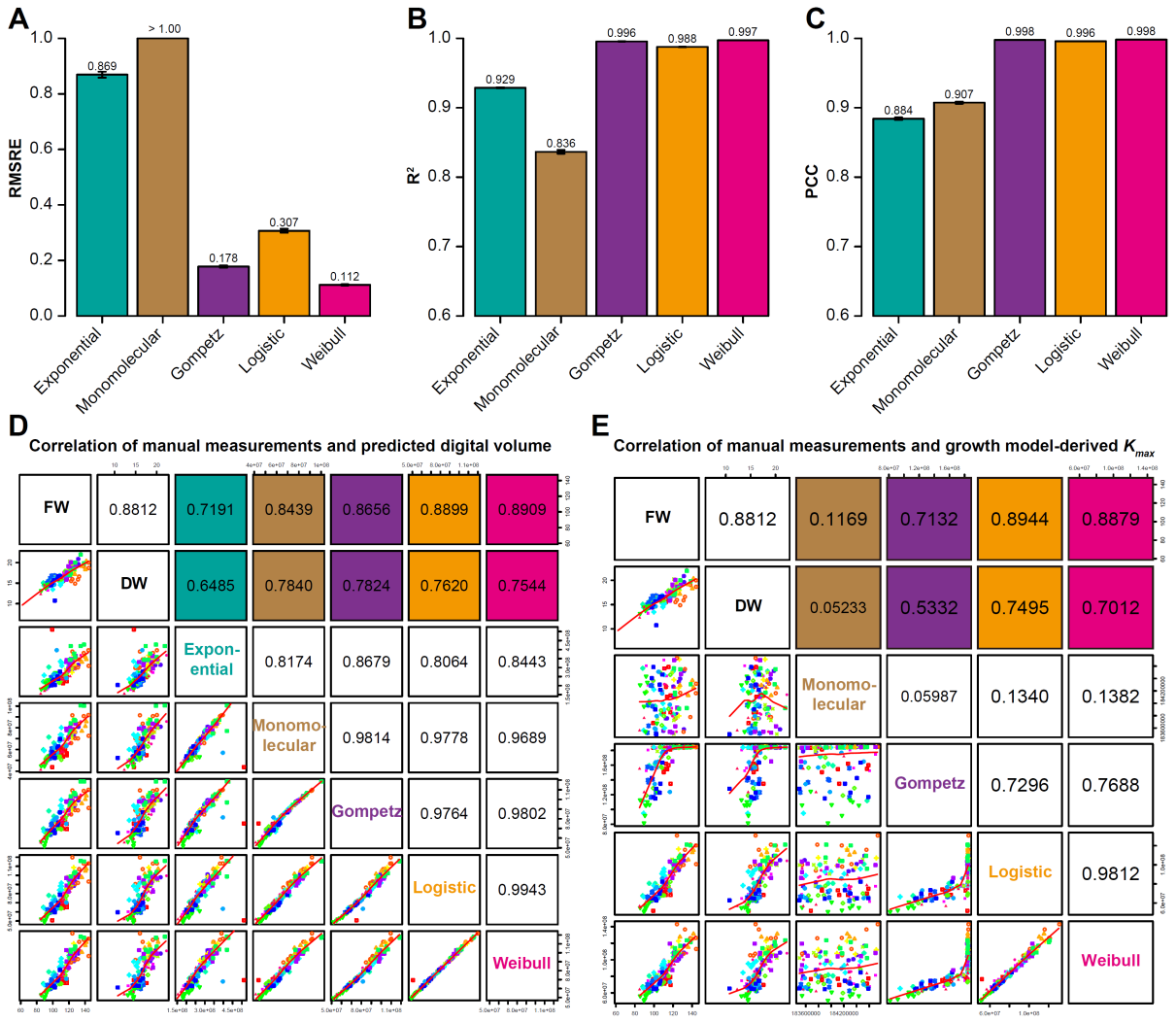

Figure 3.3: Evaluation of the performance of growth curves for control plants

The quality of fit for each model is evaluated by **(A)** the root mean squared relative error of prediction ($RMSRE$; the lower value, the better prediction), **(B)** the adjusted the coefficient of determination (R^2; the higher value, the better prediction), and **(C)** the Pearson correlation coefficient (PCC; r) between the predicted values and the observed values (the higher value, the better prediction). In **(A)**, **(B)** and **(C)**, bar height denotes the average value for all control plants; error bars denote s.e.m.. **(D)** Scatter plot representing the predicted biomass and manual measurements (fresh weight [FW] and dry weight [DW]). **(E)** Scatter plot showing pairwise comparison of the maximum growth capacity (K_{max}) derived four asymptotic growth models and the manual measurements (FW and DW) when plants were harvested. ■

This bell-shaped curve reveals a time point ($t_{max} = \frac{b}{2a}$) when plants showed the maximum estimated volume under stress and two inflection points (t_{IP1} and t_{IP2}) aside t_{max} (Figure 3.4A). These parameters may be indicative for plant stress responses. However, the volume at t_{max} was not a good indicator of final biomass ($r = 0.27$; Figure 3.4B). Plants showed rapid growth after re-watering in a relatively short recovery phase, which could be quantified with a simple linear model ($y = y_0 + rt$; median adjusted $R^2 = 0.98$). The re-growth rate ($R_{rec} = r$) was determined from the model to show the speed of recovery in different individuals (Figure 3.4C), with mean values over genotypes ranging from $8.72 \times 10^4\, px^3/day$ (MorexDH) to $2.44 \times 10^5\, px^3/day$ (Isaria). Interestingly, the recovery growth rate was strongly correlated

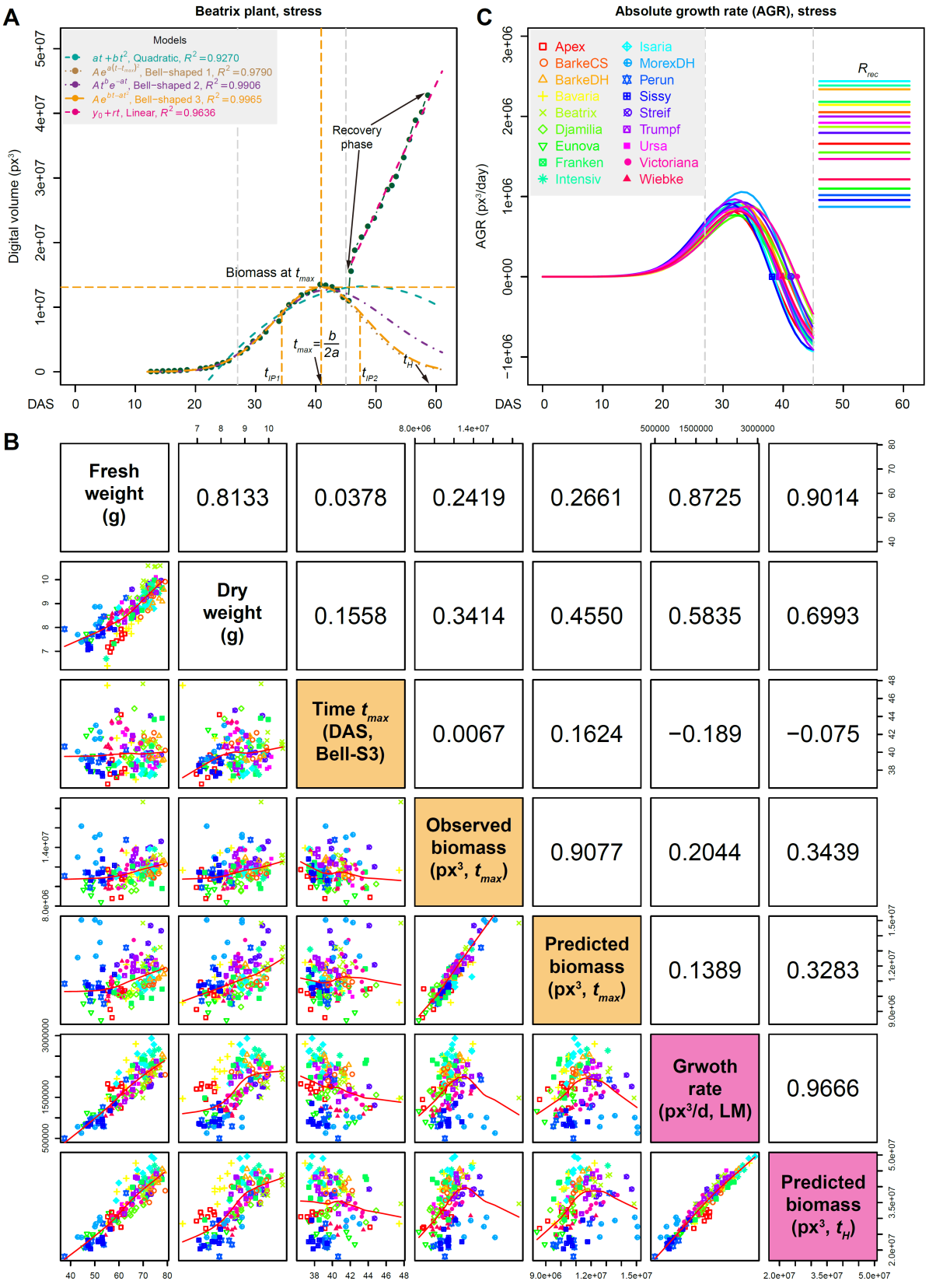

Figure 3.4: Growth modeling of barley plants under drought stress conditions

(legend on next page).

► **Figure 3.4** (continued). **(A)** Plant growth prediction based on fitting of the image-derived volume in drought stress conditions. Plant growth before re-watering is modeled by one quadratic function and three different bell-shaped functions. Growth in recovery phase is modeled by a linear function. The quality of fit (R^2) of each model is given. Five vertical lines from left to right: the start of stress period (grey), the left inflection point (t_{IP1} estimated from the best-fitted bell-shaped model 3, orange), the time of maximum biomass under stress ($t_{max} = \frac{b}{2a}$, orange), the end of stress period (grey) and the right inflection point (t_{IP2}, orange). Shown is the result of fitting for a Beatrix plant. See also Online Data Set 3 in Appendix C. t_H, time for harvesting. **(B)** Pairwise comparison of model-derived parameters, image-derived data and manual measurements for stressed plants. **(C)** The absolute growth rate (AGR) derived from the bell-shaped model 3 models (under stress) and linear models (in recovery phase). The time t_{max} is indicated by dots. Each genotypes were fitted independently and differently coloured. ■

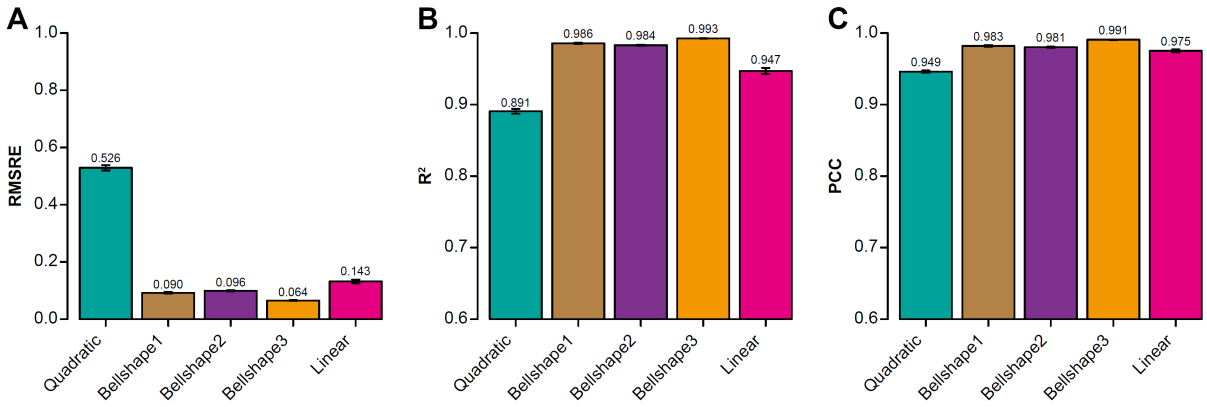

Figure 3.5: Evaluation of the performance of growth curves for stressed plants

The quality of fit for each model is evaluated by **(A)** the root mean squared relative error of prediction ($RMSRE$; the lower value, the better prediction), **(B)** the adjusted the coefficient of determination (R^2; the higher value, the better prediction), and **(C)** the Pearson correlation coefficient (PCC; r) between the predicted values and the observed values (the higher value, the better prediction). Bar height denotes the average value for all stressed plants; error bars denote s.e.m.. ■

with FW ($r = 0.87$; Figure 3.4B).

3.2.3 Model-derived parameters describing plant growth patterns and performance

Model parameters are the intrinsic factors that determine the shapes of growth curves. Therefore, model-derived parameters can be used to describe plant growth patterns when a specific growth model is applied to fit plant growth. Namely, plants differ from each other in their growth patterns by those model-derived parameters. Here, the Weibull model-derived parameters were first calculated for control plants based on the above mentioned core set of barley cultivars (Table 3.3). In Weibull model, parameter K is the limiting value of growth potential (i.e., the final biomass K_{max}), r is a shape parameters that determines the spread of the curve along the time and governs the rate at which plant growth approaches its potential maximum K_{max}, while m is the allometric constant (Fekedulegn et al., 1999). It was found that K_{max}

is highly correlated with harvested biomass ($r = 0.89$; Figure 3.2B), revealing a reasonable estimation of plant growth potential. Each parameters were evaluated in terms of repeatabilities. Interestingly, all these parameters showed relatively high repeatabilities. Meanwhile, The bell-shaped model 3 and linear model-based parameters were also derived for stressed plants (Table 3.3).

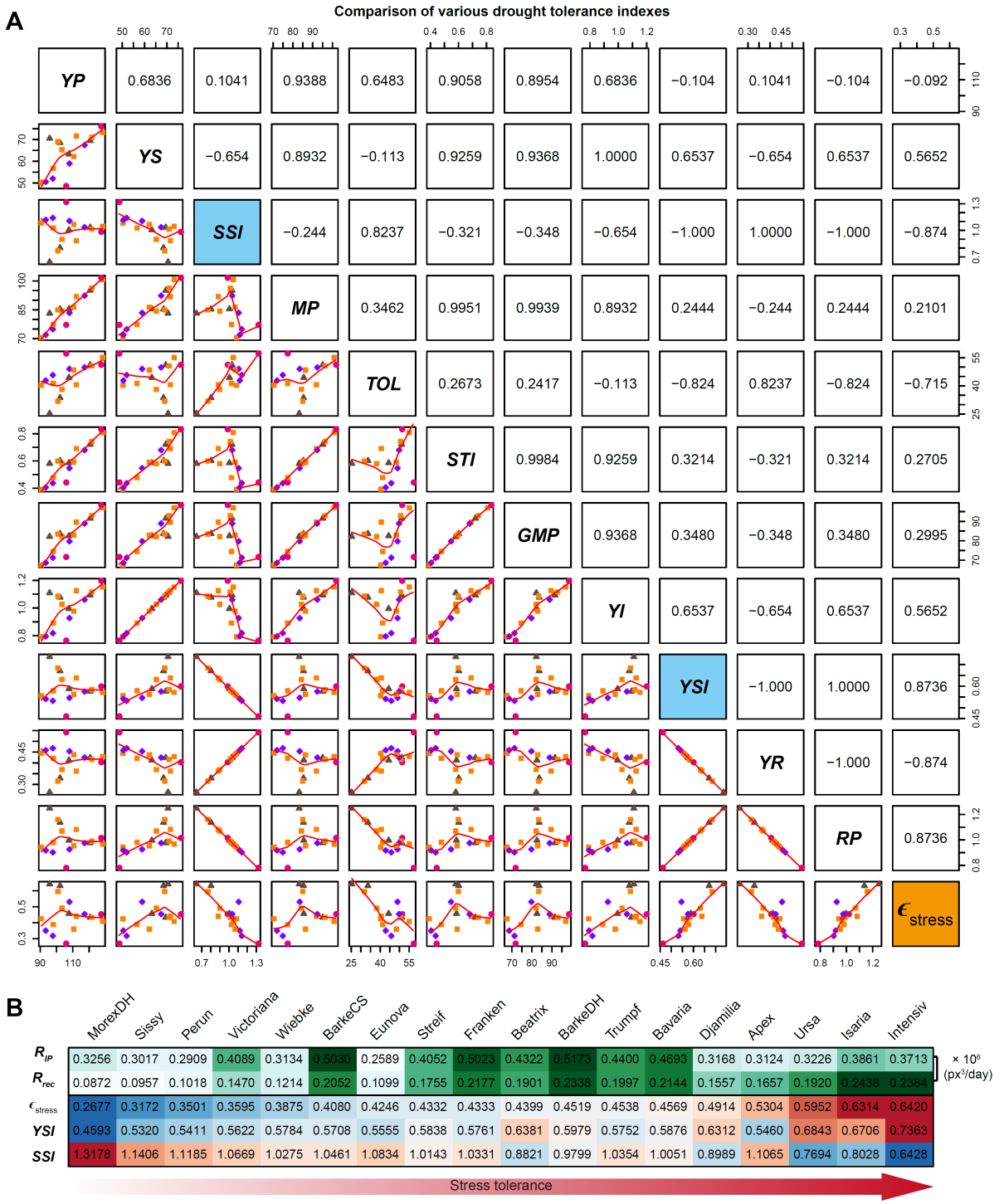

Figure 3.6: Comparison of stress elasticity and several drought tolerance indexes (legend on next page).

As a further step toward interpretation of the model-derived parameters, combined parameters could

► **Figure 3.6** (continued). **(A)** Scatter plot showing pairwise comparison of stress elasticity (ϵ_{stress}) and eleven drought tolerance indexes. The drought tolerance indexes were calculated based on biomass of fresh weight. Pearson's correlation coefficients of these indexes are indicated in top-right quadrants. YP: biomass under control condition (Y_p); YS: biomass under stress condition (Y_s); stress susceptibility index $SSI = 1 - (Y_s/Y_p)/SI$, where $SI = 1 - (\overline{Y}_s/\overline{Y}_p)$, $\overline{Y}_p$ and $\overline{Y}_s$ are the means of Y_p and Y_s, respectively (Fischer and Maurer, 1978); mean productivity $MP = (Y_s + Y_p)/2$ (Hossain et al., 1990; Rosielle and Hamblin, 1981); stress tolerance $TOL = Y_p - Y_s$ (Hossain et al., 1990; Rosielle and Hamblin, 1981); stress tolerance index $STI = (Y_p \times Y_s)/(\overline{Y}_p)^2$ (Fernandez, 1992); geometric mean productivity $GMP = \sqrt{Y_p \times Y_s}$ (Fernandez, 1992); yield index $YI = Y_s/\overline{Y}_s$ (Gavuzzi et al., 1997; Lin et al., 1986); yield stability index $YSI = Y_s/Y_p$ (Bouslama and Schapaugh, 1984; Fereres et al., 1986); yield reduction ratio $YR = 1 - YSI = 1 - Y_s/Y_p$ (Araghi and Assad, 1998); relative performance $RP = (Y_s/\overline{Y}_s)/(Y_p/\overline{Y}_p)$ (Abo-Elwafa and Bakheit, 1999). **(B)** Comparison of plant growth rates between control and stress conditions. R_{IP} represents the growth rate (px^3/day) at the inflection point of control plants. R_{rec} denotes the recovered growth rate (px^3/day) in recovery phase of stress plants. stress, referred to "stress elasticity" and calculated as the ratio of R_{rec} and R_{IP}. Two drought tolerance indexes, YSI and SSI, are provided for comparison. This figure was adapted from Chen et al. (2014b). ■

be derived from the normal plant growth and stressed plant growth models. Since R_{IP} (denoting the maximum growth rate for plants under control conditions) and R_{rec} (indicating the maximum growth rate for plants in recovery phase) are strong correlated with final biomass of control and stressed plants, respectively, their ratio was defined for each genotype as "stress elasticity" as:

$$\epsilon_{stress} = \frac{R_{rec}}{R_{IP}}$$

ϵ_{stress} showed high correlation ($r > 0.5$) with several drought tolerance indexes of different genotypes (Figure 3.6A), such as yield stability index (Bouslama and Schapaugh, 1984) and stress susceptibility index (Fischer and Maurer, 1978). It was found that cultivars MorexDH, Perun and Sissy showed the lowest tolerance to drought stress, while Ursa, Isaria and Pflugs Intensiv showed the highest tolerance (Figure 3.6B).

3.2.4 Growth modeling of a worldwide collection of maize plants

The established methodologies of plant growth modeling were then applied to a worldwide collection of maize plants, which were collected for 36 genotypes (including two high performance [HP] lines) with origin from 20 countries (Supplemental Table S2). Over a course of five weeks, 223 maize plants were monitored in a LemnaTec phenotyping system under three different conditions: wet, normal and drought stress (Figure 3.7A-C). The phenotyping period covers the most vegetative stage until the tassel emerges when plants attain their maximum height. Image data were subjected to trait extraction using the IAP software (Klukas et al., 2014). The trait of projected volume (Online Data Set 4 in Appendix C) was used to predict plant growth over time, as this trait is highly correlated to manually measured plant biomass (Figure 3.7D). Due to the limitations the phenotyping system, data points of plants with height more than $2.3m$ were excluded from analysis because of out of the range of imaging.

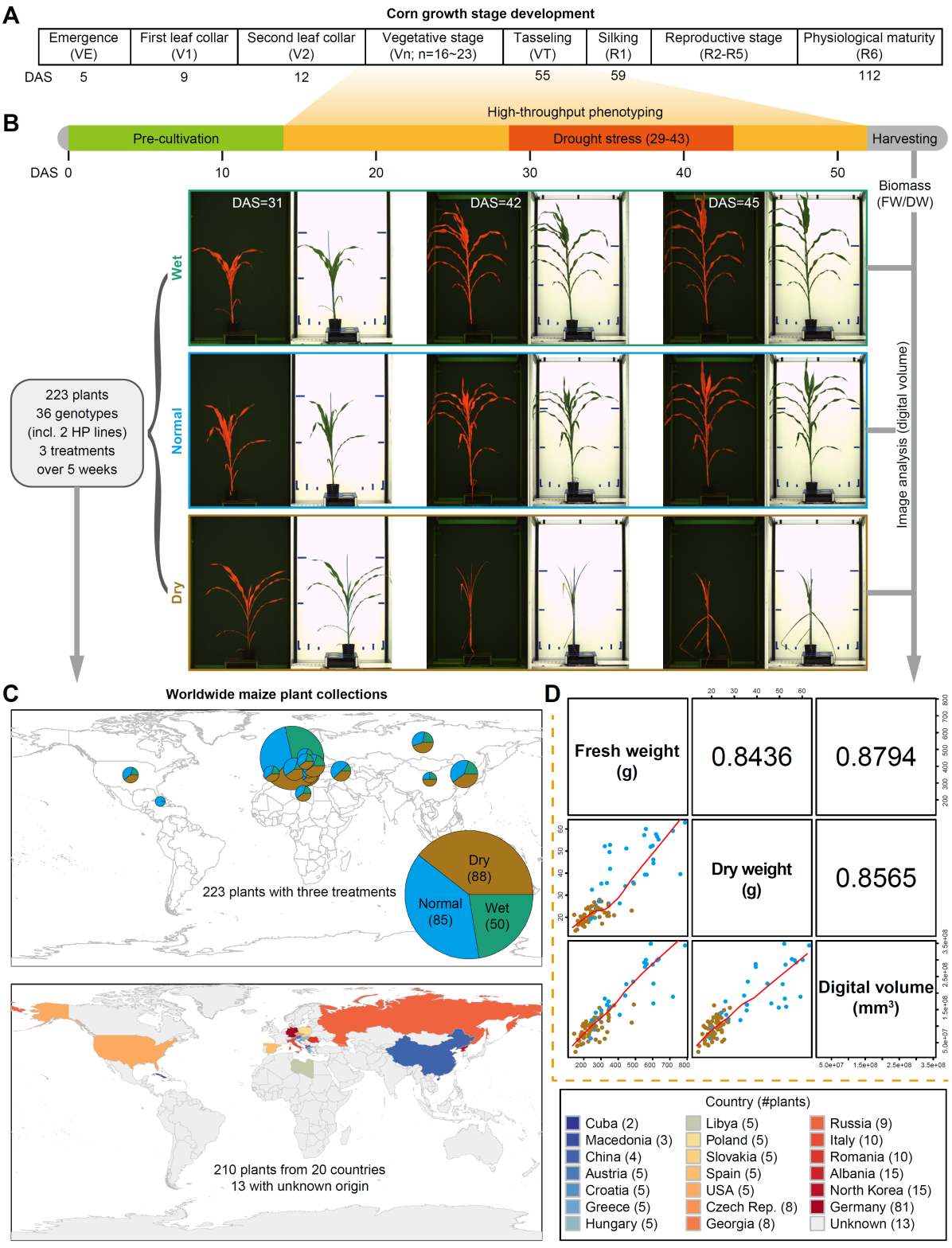

Figure 3.7: Experimental design for high-throughput phenotyping of a worldwide collection of maize plants

(legend on next page).

Plants with wet and normal treatments were fitted using the same set of nonlinear models as they were used in barley plant growth modeling under normal conditions (see Chapter 3.2.1). It was found that

► **Figure 3.7** (continued). **(A)** The growth stage of maize. **(B)** The strategy of high-throughput phenotyping (HTP) of a diverse set of maize plants with different treatments. HTP was performed in a LemnaTec system. Plants were monitored in a noninvasive manner under wet, normal and drought stress conditions. Two types of images (visible and fluorescence) from both side view (shown) and top view (not shown) were used for plant image-based volume construction. **(C)** Distribution of the worldwide maize collection. 223 plants in total were collected from 36 genotypes with 20 country origin. Top: pie charts indicate the distribution of plants with three treatments. Bottom: countries are coloured according the number of plants with their origin (indicated in the parentheses). **(D)** Correlation of image-based volume and manually measured plant biomass. ■

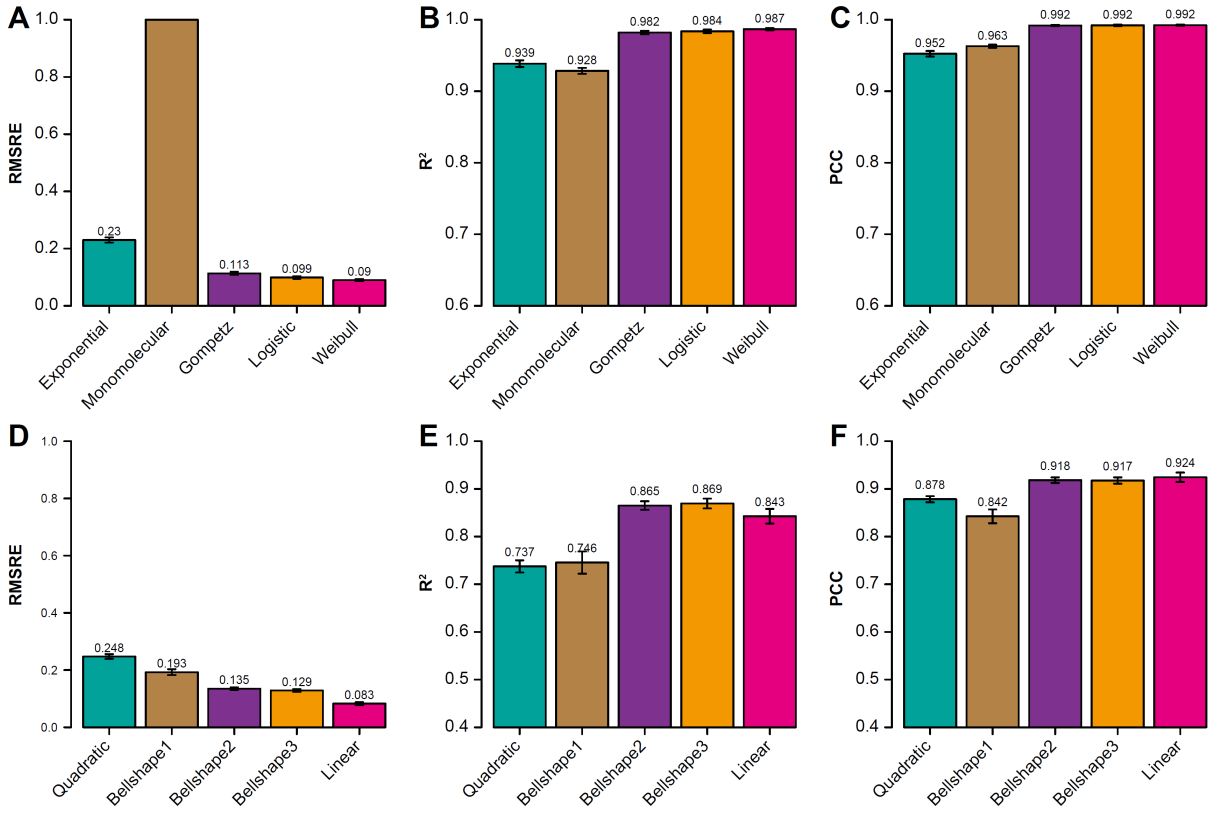

Figure 3.8: Evaluation of the performance of growth modeling for maize plants

Refer to Figures 3.3A-C and 3.5 for legends. ■

Table 3.3: Summary of growth model-derived parameters.

Model†	Parameter§	mean	s.d.	w^2¶
Weibull model	m	5.656	0.219	0.791
	r	3.378×10^{-10}	2.517×10^{-10}	0.849
	$K_{max} = K$	8.923×10^{7}	2.076×10^{7}	0.971
	t_{IP}	48.277	1.728	0.932
	R_{IP}	3.773×10^{6}	8.659×10^{5}	0.970

Bell-shaped model 3	a	8.921×10^{-3}	1.368×10^{-3}	0.721
	b	0.709	0.080	0.712
	t_{max}	40.031	1.990	0.710
	A_{max}	1.118×10^{7}	1.264×10^{6}	0.934
	$\lvert t_{IP2} - t_{IP1} \rvert$	15.118	1.266	0.660
Linear model	$R_{rec} = r$	1.719×10^{6}	5.502×10^{5}	0.971
Combined	$\epsilon_{stress} = R_{rec}/R_{IP}$	0.452	0.129	0.911

† The parameters of these model can be found from Table 3.1.

§ K_{max}: the maximum growth capacity, determined by parameter K; t_{IP}: the inflection time point; R_{IP}: the maximum growth rate at time t_{IP}; t_{max}: the time point that plant shows the maximum biomass (A_{max}) under stress; t_{IP1} and t_{IP2} are two inflection time points determined by the bell-shaped curve. $|t_{IP2} - t_{IP1}|$ is the time range between these two inflection points.

¶ w^2: repeatability of the corresponding parameter.

Weibull curve is the model of choice for describing maize plant growth under wet and normal conditions (Figures 3.9A and 3.8A-C). It was also observed that stressed maize plants follow the similar growth patterns as observed in barley: bell-shaped curve of growth under stress and linear-like re-growth in recovery stage (Figures 3.9A and 3.8D-F). However, maize plants showed sharper peaks and more narrow width in "bell" curves than barley plants, suggesting that maize plants are more sensitive to drought stress. Accordingly, it was observed the re-growth rates R_{rec} for stressed plants are largely lower than the maximum growth rates R_{IP} for plants with wet and normal treatments based on evaluation of a set of HP plants (Figure 3.9B).

Furthermore, the genotype-level growth rates (under normal conditions) were calculated for plants collected from different countries. Plants originated from different countries showed clear distinct patterns of growth rates (Figure 3.9C). Figure 3.9D showed the distribution of the maximum growth rates R_{IP} over the world according to the country origin of plants. For example, plants originated from Germany had the maximum R_{IP}, while plants originated from Russia had the minimum R_{IP}. Besides, it was found that R_{IP} is highly correlated with the maximum growth capability (K_{max}; Figure 3.9E), which is consistent with the previous observations in barley plants (Figure 3.2B).

3.3 Discussion

HTP is an ideal tool to study plant growth dynamics due to its noninvasive way of phenotyping protocol. In this regard, plant growth for the same plant can be investigated over time, thus to relieve the common problems — such as autocorrelation and heteroscedasticity of the residuals (Thornley and France, 2007) — caused in traditional growth modeling based on heterogeneous data collected from different individuals. In this chapter, linear and various nonlinear growth models were applied for modeling plant growth in two important crop species — barley and maize — under both normal growth and drought stress conditions.

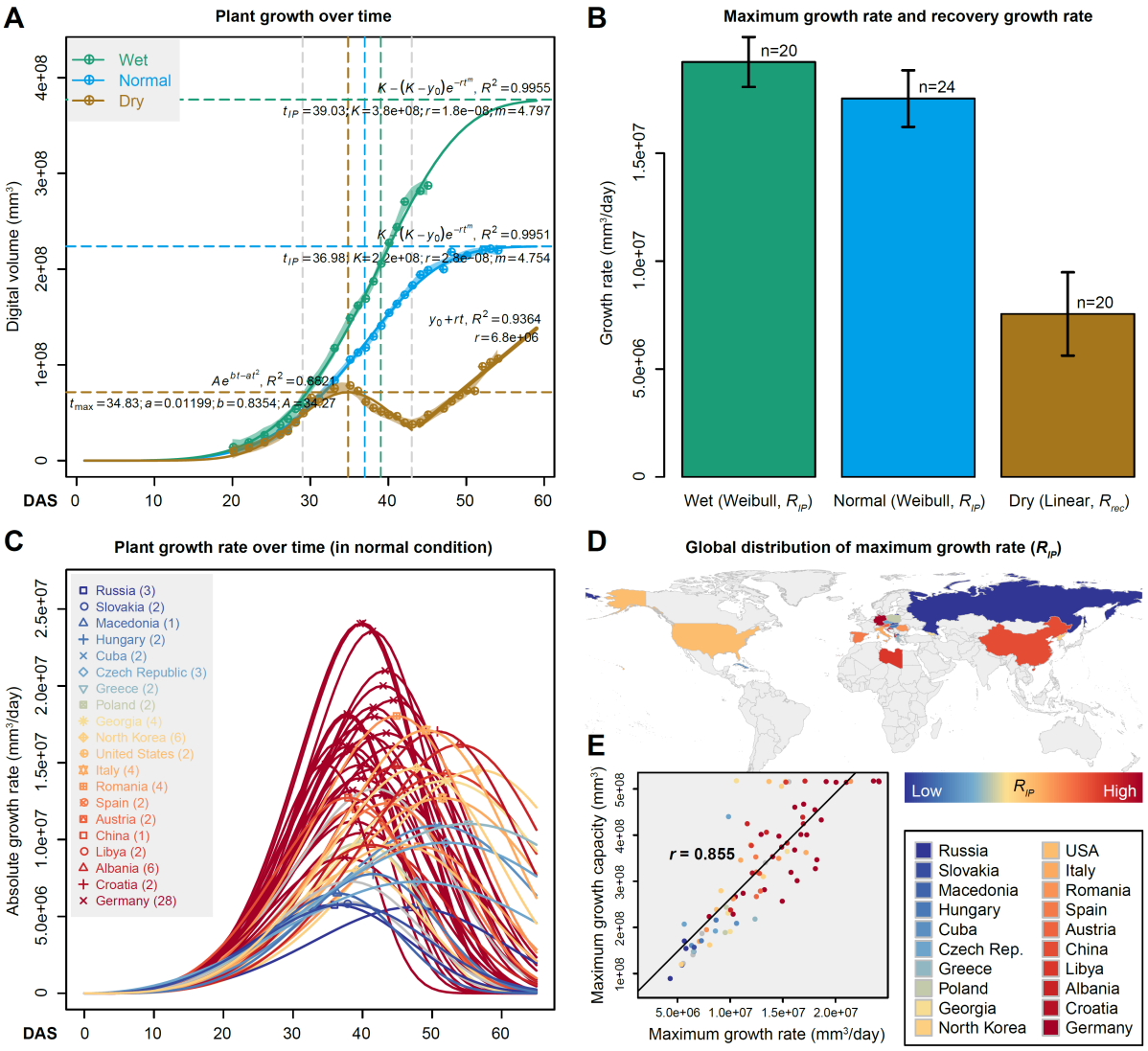

Figure 3.9: Growth modeling of maize plants
(legend on next page).

In a nutshell, it was found that both crops share the similar growth patterns under normal or water-limited conditions (Figures 3.2, 3.4 and 3.9), implying that the methodologies herein can be applicable to other crops in the future when HTP data are available.

Due to limiting belowground resources and ontogenetic changes such as the onset of flowering, crop plant growth biomass will finally approach an asymptote (Paine et al., 2012). In this study, since the time scale of phenotyping for barley and maize plants covers the whole vegetative stage, it was reasoned that asymptotic growth models are appropriate for analyses. Indeed, asymptotic models (i.e., logistic, Gompertz and Weibull models) show better performance in modeling crop plant growth than non-asymptotic models (i.e., exponential model; Figures 3.3A-C and 3.8A-C). In addition, logistic, Gompertz and Weibull models are almost equally good in prediction. For example, the logistic model can be considered as an alternative model for modeling barley plant growth (Chen et al., 2014b) as it shows similar performance with the Weibull model (Figure 3.3). However, each model has its advantages and disadvantages and it

is desirable to select an appropriate one to fit the experimental data and with biologically interpretable parameters (Archontoulis and Miguez, 2013; Paine et al., 2012). The logistic function reveals symmetric growth with an inflection point at half the final biomass ($K_{max}/2$), while the Weibull model is more robust to deal with asymmetric growth as the inflection point can be flexible over time. From this aspect, it can be concluded that the Weibull model is preferred to the logistic model for describing plant growth in crops like barley and maize.

On the other hand, non-asymptotic models can be appropriate for modeling the initial stages or partial stages of the lifespan of plant growth, even though the underlying assumption that growth continues indefinitely is somehow implicit (Paine et al., 2012). For example, a recent study showed that sorghum (*Sorghum bicolor* L. Moench) plants followed the non-asymptotic power law model in a "nitrogen" experiment based on the investigation of projected leaf area over four weeks (Neilson et al., 2015).

The methodology developed here for describing plant growth can be used for a further purpose of applying it to a large genetic mapping population. The selected nonlinear Weibull function could be used within a genetic mapping model to determine the underlying genetic bases associated with crop growth. In practice, complex agronomic phenotypes, such as plant growth and stress tolerance, can be dissected into more simple and heritable traits by model-assisted methods (Tardieu and Tuberosa, 2010). Herein a set of mathematical parameters of the growth models that define the shape of plant growth for different genotypes were identified (Table 3.3), such as the inflection time-points (t_{IP}, repeatability $w^2 = 0.93$), the maximum growth rate (R_{IP}, $w^2 = 0.97$), the maximum growth capacity (K_{max}, $w^2 = 0.97$) and the stress elasticity (ϵ_{stress}, $w^2 = 0.91$), which showed very high repeatabilities and are explicitly related to plant growth and drought tolerance, thereby permitting identification of stable QTLs controlling their expression through dynamic QTL mapping approaches (Wu and Lin, 2006). Notably, such traits in the dissection approach typically are not measurable via traditional phenotyping approaches.

► **Figure 3.9** (continued). **(A)** Growth modeling of maize plants under three different conditions as indicated in different colors. Plants with wet and normal treatments are fitted with Weibull curves. Stressed plants were fitted with bell-shaped curve (model 3) under stress condition and with linear model in recovery stage. The formulas of model functions, goodness of fit (adjusted R^2) and model-derived parameters are provided. Dots denote the average values over all plants with a specific treatment, shadow represents the estimated standard error, and curves represent the least-squares fit to the average data. Vertical lines represent the infection points for Weibull curves and the position of the peak for bell-shaped curve. Horizontal lines represent the values of K_{max} for Weibull models and the peak for bell-shaped curve. See Online Data Set 5 in Appendix C for growth modeling for individual plants. **(B)** Comparison of the maximum of growth rate (R_{IP} from Weibull model) for plants with wet and normal treatments and recovery growth rate (R_{rec} from linear model) for stressed plant. Evaluation is based on the common set of high performance lines. **(C)** The absolute growth rate (AGR) of normal plants derived from the Weibull models, which were fitted at the genotype level. The number of plants used for modeling is given in parentheses. **(D)** Distribution of the maximum growth rate (R_{IP}). Countries are coloured according the median values of R_{IP}. **(E)** Correlation of the maximum growth capacity (K_{max}) and the maximum growth rate (R_{IP}). ■

3.4 Materials and methods

3.4.1 Plant image data

Barley plant image data were obtained as described in the previous chapter (see Chapter 2.4.1; (Chen et al., 2014b)). Maize plant image data are described as below.

3.4.1.1 High-throughput phenotyping of a worldwide set of maize plants

A collection of a worldwide varieties of maize genotypes was provided by the IPK Genebank (Supplemental Table S2). The selection includes 34 genotypes from 20 different countries and two high performance (HP) lines (with German origin) from KWS Company[1].

On average four replicates per genotype were grown, except from the two high-performance lines, which includes 26 replicates each. The overall cultivation time was 55 days. The plants were seeded in five litre pots (filled with 4 kg of an IPK soil mixture composed of 40% IPK made compost [composed of 9% organic matter, pH 6.9, with 153 mg/l N, 731 mg/l P_2O_5, 1259 mg/l K_2O, 272 mg/l Mg], 40% substrate 2 [Klasmann-Deilmann GmbH, Geeste, Germany] and 20% sand) and pre-cultivated in an external greenhouse for 16 days in a controlled environment. Within this duration the plant are large enough for imaging on the high-throughput imaging system. After, the plants were randomly placed for 39 days on the phenotyping system. The plants grew in a climate controlled glass house at 25/20°C day/night, 65% relative air humidity, and 205−245 μmol m^2s^1 PAR supplemental illumination using SonT Agro high pressure sodium lamp (Philips, Amsterdam, The Netherlands) with the light period set to 16 h (06:00−22:00 h). The watering were handled by the system through a peristaltic pump, each plant were watered daily by adding equivalent volume of water that was lost from the soil. By a daily re-arrangement of the plants, growth effects affected by positioning effects should be avoided. After 14 days of growth, a stress phase were initiated for 15 days. For every genotype a control-line were established which were unaffected from changing conditions. Additionally, for the two KWS lines the watering were also extended over the normal distribution to simulate effects of over-watering of plants. An detailed overview about the cultivation duration and the exact harvest times is shown in Figure 3.7B.

The plants were cultivated in an automated greenhouse equipped with a HTP system from the LemnaTec company which allows the automatic cultivation of plants. Each plant was placed on a separate carrier which can be moved by an conveying system. The system enables an automatic imaging using three different imaging sensors. Images were acquired in the visible light spectra (˜390-750 nm) using a Basler[2] Pilot piA2400-17gc with resolution of 2454 × 2056 pixels. This images are useful for many analysis task, e.g. extracting architectural or color features. An fluorescence imaging system, performing an excitation of blue light with an average intensity of 450 nm[3] and an image capturing in a range of

[1] http://www.kws.com/
[2] Basler AG, Ahrensburg, Germany.
[3] minimum intensity at 400 nm and maximum intensity at 500 nm.

~520-750 nm by using a Basler Scout scA1400-17gc with resolution of 1390 × 1038 pixels, a relative prediction for fluorescence activity in plants is possible. Finally a near-infrared imaging in a range of ~1420-1480 nm using a Nir 300[4] camera with a resolution of 320 × 256 pixels is performed. The camera has an maximum intensity at 1452 nm with and full width at half maximum (FWHM) of 27 nm. During the experiment three different zoom configuration were used to get the maximum detail in respect to the plant development. Also blank reference images (images including only the empty imaging chamber), for each imaging system, are created before each imaging run. This images could later be used for the image processing. The images were saved as uncompressed .png (portable network graphic) files. The created dataset include 80133 images with an overall size of 281 GB (uncompressed). Images were acquired from different positions including a top-view and several side-views. For side-view images two angles at zero and 90 degree were considered. On days on which the manual harvests were performed additional images from 24 side angles were taken in steps of 15 degree. The harvested plants were randomly chosen from the whole set. Weighting and watering data were collected automatically by the system, during the watering procedure.

3.4.2 Image analysis

Image analysis was peformed by using the IAP software (Klukas et al., 2014), as described in Chapter 2.4.2. Parameter were adjusted and optimized according to the experiment configuration and imaging conditions. One of the main features used in this chapter, the estimated plant "digital volume", is implemented inside IAP and calculated as

$$V_{IAP} = \sqrt{A_{side.view}^2 \times A_{top.view}}$$

where $A_{side.view}$ and $A_{top.view}$ are the mean values of projected areas from side-view (at different angles) and top-view images, respectively. While there are different ways to estimate plant volume, V_{IAP} shows the best performance in biomass estimation (Klukas et al. (2014)).

3.4.3 Plant growth modeling

Of the investigated traits, image-based volume showed the best correlation with manually measured fresh weight and dry weight (Figures 3.1 and 3.7) and thus was considered to represent the digital biomass of plants. The plant growth was modeled using digital biomass for control and stressed plants, respectively.

Growth in control conditions was modeled with five different mechanistic models: exponential, monomolecular, logistic, Gompertz and Weibull models (Table 3.1). To fit these models using the linear regression function "lm" in R, the nonlinear relationship of the models were first transformed into linearized forms. These linearized models were fitted using the estimated volume of each control plant. To fit models in high quality, it was required that the input data include at least one data point from the first one-fourth and the last one-fourth of the whole growth stage. The fitting quality of models was assessed and compared based on the following criteria:

[4]Camera from Allied Vision Technologies GmbH former VDS Vosskühler GmbH, Stadtroda, Germany.

(i) the Pearson correlation coefficient (PCC; r) between the predicted values and the observed values;

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \widetilde{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(\widehat{y}_i - \widetilde{y})^2}}$$

(ii) the adjusted coefficient of determination (R^2), namely, the fraction of variance explained by the linear-transformed model;

$$R^2 = \left(1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} - \frac{p}{n-1}\right)\left(\frac{n-1}{n-p-1}\right)$$

(iii) the root mean squared relative error of prediction, defined as

$$RMSRE = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{y_i}\right)^2}{n}}$$

where $\widehat{y}_i$ is the predicted and y_i is the observed biomass at the ith time point for a specific plant, $\widetilde{y}$ is the mean value of the predicted biomass and $\overline{y}$ is the mean value of the observed biomass, n denotes the number of data points used for growth modeling and p is the number of model parameters.

Of the five models, the Weibull model fitted best (Figures 3.3 for barley and 3.8A-C for maize). Several useful parameters (derived traits) can be derived from the Weibull model: (1) the intrinsic growth rate (R) which measures the speed of growth; (2) the inflection point (IP) which represents the time point when plant reaches the maximal speed of growth; and (3) the maximum final vegetative biomass (K_{max}), which was estimated for each plant on the basis that the model could fit the data with the largest R^2. To this end, K_{max} was initially assigned to the image-based volume at the last day and the corresponding R^2 is calculated. The process was iterated with 1% increment of K_{max} at each step, and the iteration was stopped when there was no increment of R^2.

Modeling of growth in stress conditions is divided into two parts: (1) growth before and during the stress phase and (2) re-growth during recovery phase. In the first phase, three different bell-shaped curves and a quadratic curve were fitted to the data, while in the recovery phase a simple linear model was used to characterize re-growth (Table 3.1). The bell-shaped models were first linearized and then fitted using the linear regression function. The bell-shaped model $y = Ae^{bt-at^2}$ fitted best and was used for parameter extraction. Parameters estimated from this bell-shaped model included: time point of maximum biomass ($t_{max} = b/2a$) and biomass at t_{max} (Table 3.3). After stress, the linear model revealed the speed of re-growth (R_{rec}).

3.4.4 Trait repeatability

Repeatability (w^2) is the proportion of phenotypic variance attributable to differences in repeated measures of the same genotype (in terms of replicated plants). Repeatabilities were calculated as $w^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_e^2/r}$, where r is the number of of replicated plants. Genotypic variance σ_G^2 was estimated in the linear mixed model (see Chapter 2.4.8) by residual maximum likelihood (REML) assuming that $G_i \sim N(0, \sigma_G^2)$.

Chapter 4

Prediction of plant biomass accumulation based on image-derived parameters

4.1 Introduction

Biomass accumulation is an import indicator of crop final product and plant performance. It is thus considered as a key trait in plant breeding and agriculture improvement. The conventional means of measuring plant biomass is very time consuming and labor intensive because plants need to be harvested and dried before measuring their fresh or dry weights. Moreover, owing to its destructive measurements, this method is impossible to measure the same plant over time. Digital image analysis has been proposed as a fast alternative way to accurately infer plant biomass.

Recently, plant biomass has been subject to intensive investigation by using high-throughput phenotyping (HTP) approaches in both controlled growth chambers (Feng et al., 2013; Golzarian et al., 2011; Tackenberg, 2007) and field environments (Busemeyer et al., 2013b; Cao et al., 2013; Ehlert et al., 2010, 2008; Erdle et al., 2011), demonstrating that the ability of imaging-based methods to infer plant biomass accumulation. For example, Golzarian et al. (2011) modeled the plant biomass (dry weight) in wheat (*Triticum aestivum* L.) as a linear function of projected area, assuming plant density is constant. However, this method under-estimated dry weight of salt stressed plants while over-estimated that of control plants. Although the authors argued that the bias was largely related to plant age and the model can be improved by including the factor of plant age (Golzarian et al., 2011), the differences in plant density between stressed and control plants may be due to different physiological properties of plants rather than plant age. In another study, Busemeyer et al. (2013b) developed a calibrated biomass determination model for triticale (x *Triticosecale* Wittmack L.) under field conditions based on multiple linear regression analysis of a diverse set of parameters with selectivity to both the volume of the plants and their density. Indeed, this model largely improved the prediction accuracy of the calibration models based on a single type of parameters and can precisely predict biomass accumulation across environments (Busemeyer et al., 2013b). Nonetheless, this analysis was performed at plot level rather than at single

plant level, making the results difficult to interpret. In such studies, it is often difficult to accurately determine the biomass and image-based trait values of each plant (e.g., temperature differences resulting in imaging errors and differences in competition as an attribute of a certain genotype), which limits the predictive power of these models. Besides, the number of traits used in these studies were quite limited and perhaps not representative.

In this chapter, I endeavour to develop a general framework to study the relationship between plant biomass (refer to shoot biomass hereafter) and image-derived parameters. A multitude of supervised and unsupervised statistical methods were applied to investigate different aspects of biomass determinant by a list of representative phenotypic traits in three consecutive experiments in barley. It is shown that image-based features can accurately predict plant biomass output and collectively account for large proportion of the variation in biomass accumulation. I also investigate the relative importance of different feature categories and of individual features in prediction of biomass accumulation. Furthermore, I compare the contribution difference of the image-based features in prediction of two types of biomass measurements, fresh weight and dry weight. In addition, I test the possibility of the models in prediction of plant biomass in different experiments with different treatments. As high-throughput plant phenotyping is going to be the technique of choice for automated phenotype measurements in plant breeding in the near future, I anticipate that the methodologies proposed in this work have various potential applications.

4.2 Results

4.2.1 Development of statistical models for modeling plant biomass accumulation using image-derived features

Table 4.1: Overview of three barley experiments.

Experiment	#plants/#genotypes†	Date of sowing	Date of harvesting	Biomass§
Exp. 1	312/18	27.05.2011	24.07.2011	FW & DW
Exp. 2	312/18	22.07.2011	18.09.2011	FW
Exp. 3	312/18	16.09.2011	13.11.2011	FW & DW

† Number of plants or genotypes. The genotype information refers to Table 2.1.

§ Types of biomass measurement. FW: fresh weight; DW: dry weight.

In the previous chapter, it has been shown that a single phenotypic trait — the three-dimensional digital volume, which is a derived feature from projected side and top areas — can be reasonably predictive of plant biomass accumulation (Figures 3.1 for barley plants and 3.7D for maize plants). I argue that the predictive power can be improved when multiple phenotypic traits are combined in a prediction model, as plant biomass is determined not only by their structure features but also by their density (physiological properties). To further investigate the relationship between image-derived parameters and plant biomass accumulation, I took advantage of deep phenotyping data which include both structural (e.g., geometric

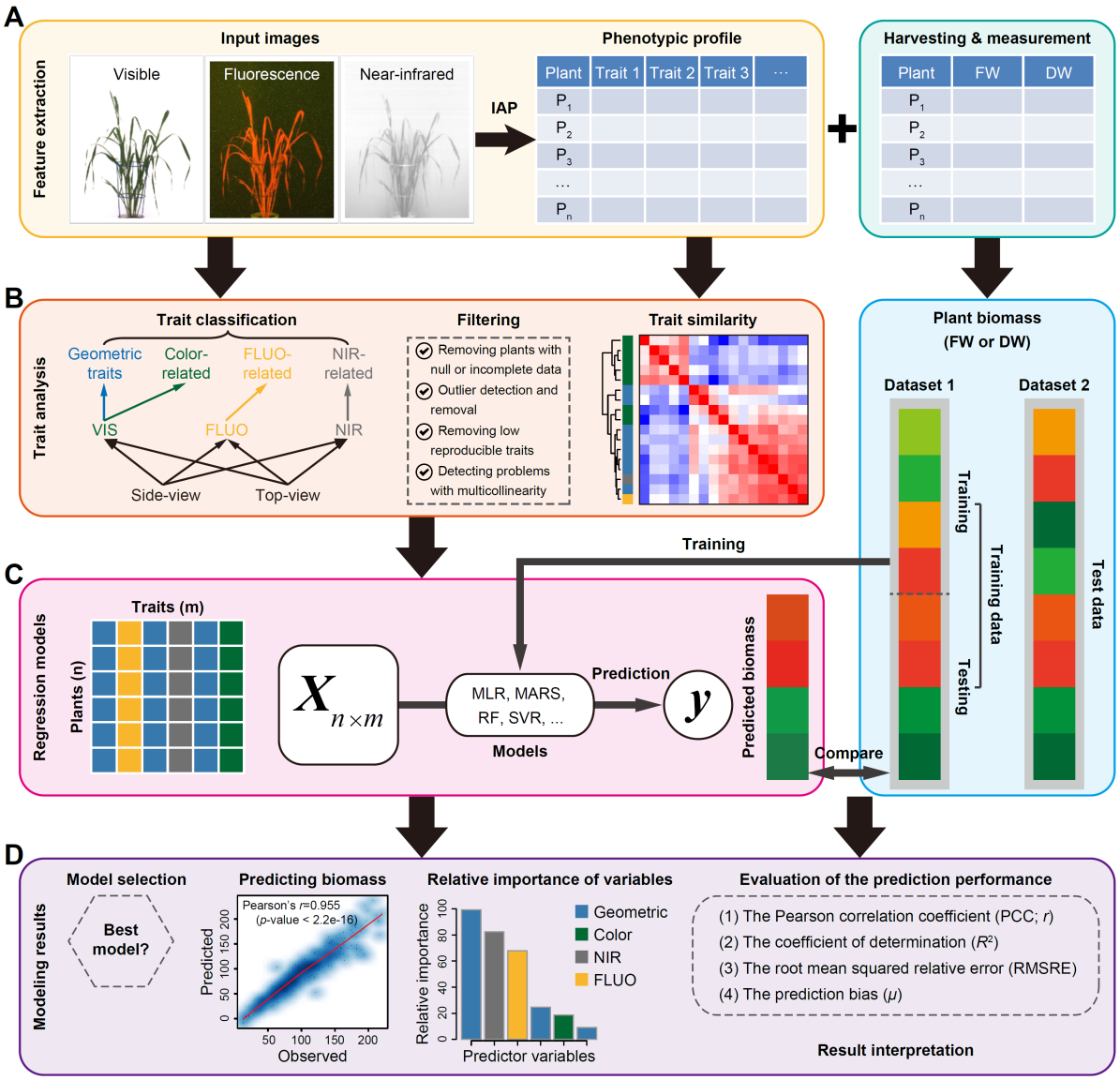

Figure 4.1: Modeling pipeline for predicting plant biomass accumulation based on image-derived parameters

(legend on next page).

traits) and physiological traits (e.g., plant moisture content; Figure 4.1A-B).

I constructed models to quantify the ability of image-derived features to statistically predict the biomass accumulation. I developed models using four widely used machine-learning methods (Figure 4.1C): multivariate linear regression (MLR), multivariate adaptive regression splines (MARS), random forest (RF) and support vector regression (SVR), which have extensively been used in accurate prediction of gene expression (Cheng et al., 2012; Cheng and Gerstein, 2012; Cheng et al., 2011; Dong et al., 2012; Karlić et al., 2010) and DNA methylation levels (Das et al., 2006; Ma et al., 2014; Zhang et al., 2015; Zheng et al., 2013). I combined the biomass measurements (fresh weight [FW] and/or dry weight [DW]) with image-based features and then divided them into a training data set and a test data set. A model was trained on the training data set and then was applied to the test data set to predict the plant biomass.

The relationship between plant biomass accumulation and image-derived features was assessed based on the criterion of the Pearson correlation coefficient (r) between the predicted values and the actual values, or the coefficient of determination (R^2; the percentage of variance of biomass explained by the model; Figure 4.1D).

I applied the methodology to three consecutive experiments (Figure 4.2A; Table 4.1), which were designed to investigate vegetative biomass accumulation in response to two different watering regimes under semi-controlled greenhouse conditions in a core set of barley cultivars by non-invasive phenotyping (Chen et al., 2014b; Neumann et al., 2015). There are 312 plants with 18 genotype origin for each experiment. Plants were monitored using three types of sensors (visible, fluorescence [FLUO] and near-infrared [NIR]) in an imaging system LemnaTec-Scanalyzer 3D. An extensive list of phenotypic traits ranging from geometric (shape descriptors) to physiological properties (i.e., color-, FLUO- and NIR-related traits) can be extracted from these image data (Figure 4.1B) using the image processing pipeline IAP (Klukas et al., 2014). A representative list of traits for each plant in the last growth day were selected to test their predictability of plant biomass.

4.2.2 Coordinate patterns of plant image-based profiles and their biomass output

A list of representative and non-redundant phenotypic traits were extracted for each plant from image datasets for each experiments (see Chapter 4.4.3; Figure 4.1B). Thirty-six high-quality traits in common were obtained to describe plant growth status in the last growth day. As a result, each dataset was assigned a matrix whose elements are the signals of different features in different plants (Figure 4.1C). I applied unsupervised methods, such as hierarchical clustering (HCA; Figure 4.2B) and principal component analysis (PCA; Figure 4.2C), on these datasets and found that plants from different experiments

► **Figure 4.1** (continued). **(A)** Input data, including high-throughput image data and manually measured biomass data. Plants were phenotyped using various cameras such as visible (or color), fluorescence (FLUO) and near-infrared (NIR) sensors. Image analysis was performed with IAP software (Klukas et al., 2014) for feature extraction. The same plants were harvested and measured at the end of growth stage. Generally, two types of biomass was measured: fresh weight (FW) and dry weight (DW). **(B)** Trait processing. All the phenotypic traits are grouped into four categories: geometric, color-related, FLUO-related and NIR-related traits. Phenotypic data were subjected to quality check to remove low-quality data. **(C)** Each plant was described by a list of traits, resulting in a predictor matrix whose rows represent plants and columns represent image-based traits. This matrix was used to predicted plant biomass accumulation by MLR (multivariate linear regression), MARS (multivariate adaptive regression splines), RF (random forest) and SVR (support vector regression) models. The right panel represents the schema of model validation. In the first schema, a dataset (Dataset 1) was divided into training set and testing set in a ten-fold cross-validation manner. In the second schema, the whole of one dataset (Dataset 1) was used for training and another dataset (Dataset 2) was used for testing. **(D)** Model selection, evaluation and result interpretation. The correlation of the predicted values and measured values was used to assess the overall performance of the model. ■

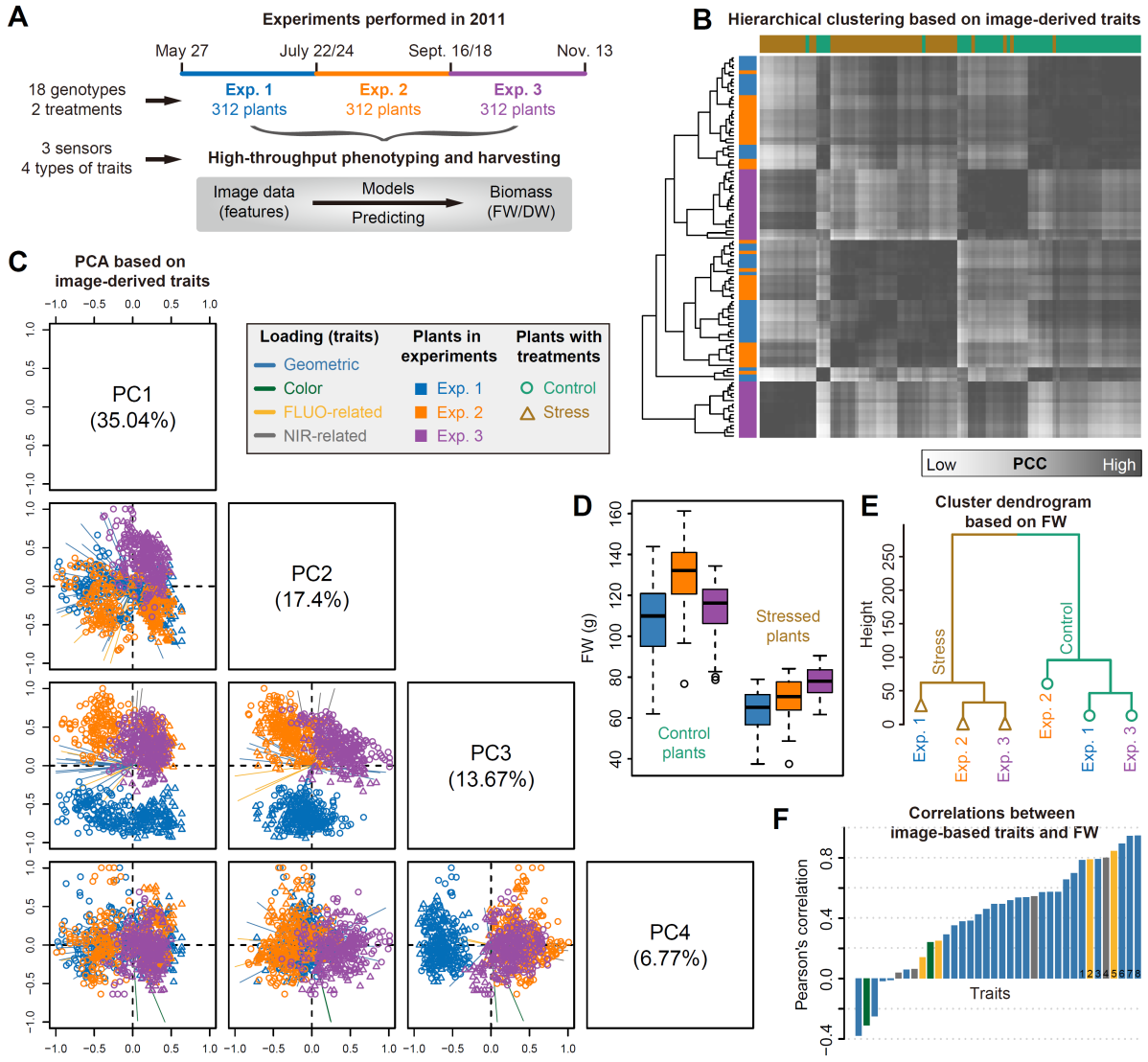

Figure 4.2: Predictability of image-based traits to plant biomass (legend on next page).

with different treatments showed clearly distinct patterns of phenotypic profiles. For instance, stressed plants and control plants were separated by the first principal component (PC1) in PCA or the top clusters in HCA, while plants from different experiments were distinguished by PC2 and PC3 in PCA or subordinate clusters in HCA. Accordingly, it was observed that biomass (e.g., FW) of plants from different experiments with different treatments was significantly different (two-way ANOVA, $P <$ 2e-16; Figure 4.2D). The relationship was reflected by a dendrogram from clustering analysis based on the means of FW over genotypes (Figure 4.2E). Furthermore, it was found that the overall phenotypic patterns of these plants were similar to their biomass output (Figure 4.2B-E), revealing that these image-based features are potential factors determining the accumulation of plant biomass. I thus explored the relationship between the signals of these image-based features and the level of plant biomass output. The correlation coefficients in each dataset were calculated. The correlation patterns were consistent in different datasets and more than half of the features revealed high correlation coefficients ($r > 0.5$; Figure 4.2F). Interest-

ingly, the top ranked features include both structure features (such as digital volume, projected area and plant area border length) and density-related features (such as NIR and FLUO intensities).

4.2.3 Relating image-based signals to plant biomass output

The above analyses suggest that plant biomass can be at least partially inferred from image-based features. I then applied the regression models (Figure 4.1C) to predict plant biomass using image-based features. To examine which model has the best performance and to select an appropriate model for biomass prediction, I focused on the analyses in the first experiment (i.e., experiment 1), since the phenotypic traits of the corresponding dataset have been intensively investigated in the previous study (Chen et al., 2014b, as presented in Chapter 2). In this experiment, plant biomass was quantified in two forms: FW and DW (Table 4.1). I selected a collection of 45 image-derived parameters from this dataset that were non-redundant and highly representative.

I next tried to predict FW (Figure 4.3A) and DW (Figure 4.3C) based on this set of image-derived features using four different regression models. The models were respectively tested on control plants, stressed plants and the whole set of plants. I compared and evaluated the performance of these models. Although the performance of these models was roughly comparable, RF, SVR and MARS methods had better performance than that of MLR method for prediction of both FW (Figure 4.3B) and DW (Figure 4.3D), implying a nonlinear relationship between image-based phenotypic profiles and biomass output. The RF model largely outperformed other models especially in predicting biomass of control plants, accounting for the most variance ($R^2 = 85\%$ for FW and $R^2 = 62\%$ for DW; Figure 4.3B,D, left panels) and showing the best prediction accuracy (Pearson's correlation $r = 0.93$ for FW and $r = 0.80$ for DW; Figure 4.3B,D, middle panels). I also compared the prediction accuracy of the models (the correlation coefficients between the predicted biomass and the actual biomass) with the best predictability of individual feature (here, the "digital volume"; Figure 4.3B,D, middle panels). It was found that

► **Figure 4.2** (continued). **(A)** Schema depicting three consecutive high-throughput phenotyping experiments in barley. Plants in each experiment were harvested for biomass measurements: fresh weight (FW; for all experiments) and dry weight (DW; only for experiment 1). **(B)** Heatmap of Pearson's correlations between plants. Pearson's correlation coefficient (PCC) was calculated based on image-derived traits. Cluster dendrograms for experiments (left) and treatments (top) are shown. **(C)** Scatter plots showing projections of top four Principal components (PCs) based on PCA of image-based data. The component scores (shown in points) are colored and shaped according to the experiments (as legend listed in the box). The component loading vectors (represented in lines) of each traits (as colored according to their categories) were superimposed proportionally to their contribution. **(D)** Boxplot showing the distribution of FW across different experiments. **(E)** A dendrogram from cluster analysis based on the means of FW data over genotypes. **(F)** Pearson's correlation (mean values in the three datasets) between image-based traits and FW. Traits with the largest mean correlations values are labeled: 1 — sum of leaf length (side view), 2 — sum of FLUO intensity (side), 3 — plant area border length (side), 4 — sum of NIR intensity (top), 5 — sum of FLUO intensity (top), 6 — projected area (top), 7 — projected area (side) and 8 — digital volume. ■

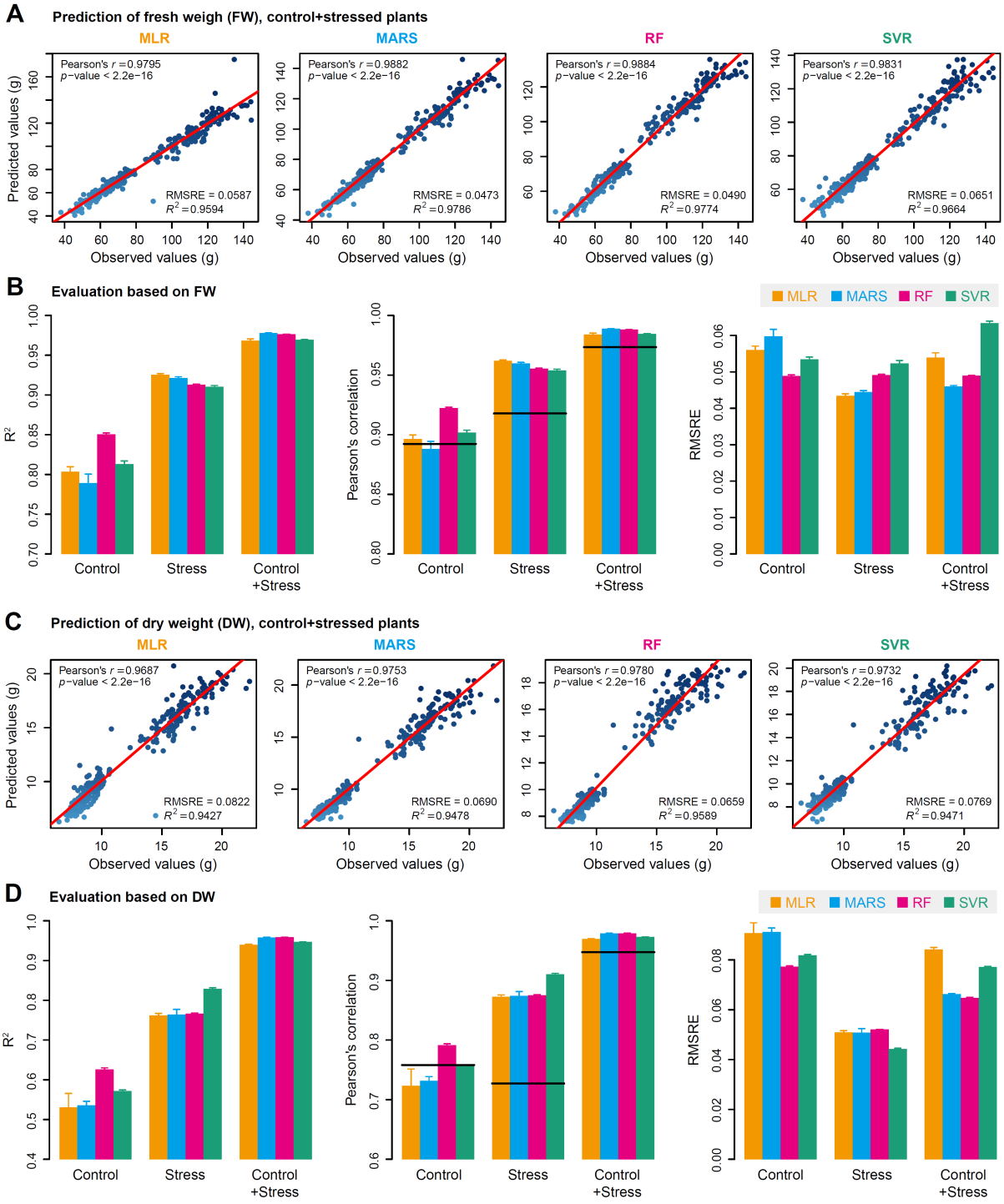

Figure 4.3: Quantitative relationship between image-based features and plant biomass (legend on next page).

the models generally showed better prediction power than the single digital volume-based prediction, indicating that additional features improved the predictive power. In this study, I focus on results from the RF method in the rest of analysis, although results from different methods are highly consistent and lead to the same conclusions.

4.2.4 Contribution of different image-based features to predicting plant biomass

As mentioned above, the image-based features can be classified broadly into four categories: plant structure properties, color-related features, NIR signals, and FLUO-based traits (Figure 4.1B). The last three types of features reflect plant physiological properties and can be considered as plant density-related traits and are thus related to their fresh or dry matter content. For each individual feature or each type of features, I constructed a degenerate model of biomass prediction using the corresponding feature(s) as the predictor(s). I compared the capability of each individual or type of features for predicting biomass accumulation in the first experiment (i.e., experiment 1). Geometric features showed the most predictive power among the four categories for prediction of both FW and DW, but were slightly less predictive than all features in a full model (Figure 4.4A-B). Strikingly, the predictability of other types of features (such as color-related and FLUO-based traits) was substantial, indicating that these traits may act as unforeseen factors in biomass prediction. In addition, the NIR-based features showed higher predictive capability for FW than for DW in control and stressed plants, revealing NIR signals are import factors determining FW accumulation.

Next, I investigated the relative importance (RI) of each feature for predicting biomass using a full model in the whole set of plants (i.e., "control + stressed plants"; Figure 4.4C-D, upper panels). In a RF model, the RI of a feature is calculated as the increase of prediction error (%IncMSE) when phenotypic data for this feature is permuted (Breiman, 2001), and thus indicates the contribution of the feature after considering its intercorrelation in a model. It was found that the top ten most important features in the full model for predicting FW and DW included both structure and density-related traits. As expected, projected area (from side or top view) and digital volume were the top ranked features, which have individually been considered as proxies of shoot biomass in previous studies (Arvidsson et al., 2011; Chen et al., 2014b; Dietz and Steinlein, 1996; Golzarian et al., 2011; Hairmansis et al., 2014; Leister et al., 1999; Neilson et al., 2015; Paruelo et al., 2000; Walter et al., 2007).

In principle, I would expect that the more importance of a feature in the full model, the more predictive power of this feature in a degenerate model. Surprisingly, there was no clear correlation observed between the feature importance and their predictive power (Figure 4.4C-D). For example, several color-

► **Figure 4.3** (continued). **(A)** and **(C)** Scatter plots of manually measured plant biomass (fresh weight [FW] and dry weight [DW]) versus predicted biomass values using four prediction models: multivariate linear regression (MLR), multivariate adaptive regression splines (MARS), random forest (RF) and support vector regression (SVR). The red line indicates the expected prediction ($y = x$). The quantitative relationship between image-based features and biomass is evaluated by Pearson's correlation coefficient (PCC r and its corresponding p-value), $RMSRE$ (root mean squared relative error) and the percentage of variance explained by the models (the coefficient of determination R^2). **(B)** and **(D)** Summary of the predictive power of each regression models. The results are based on ten-fold cross-validation with ten trials. Models were evaluated based on control plants, stressed plants and the whole set of plants. The solid lines in the middle panel represent PCC between digital volume and biomass for specific datasets. ■

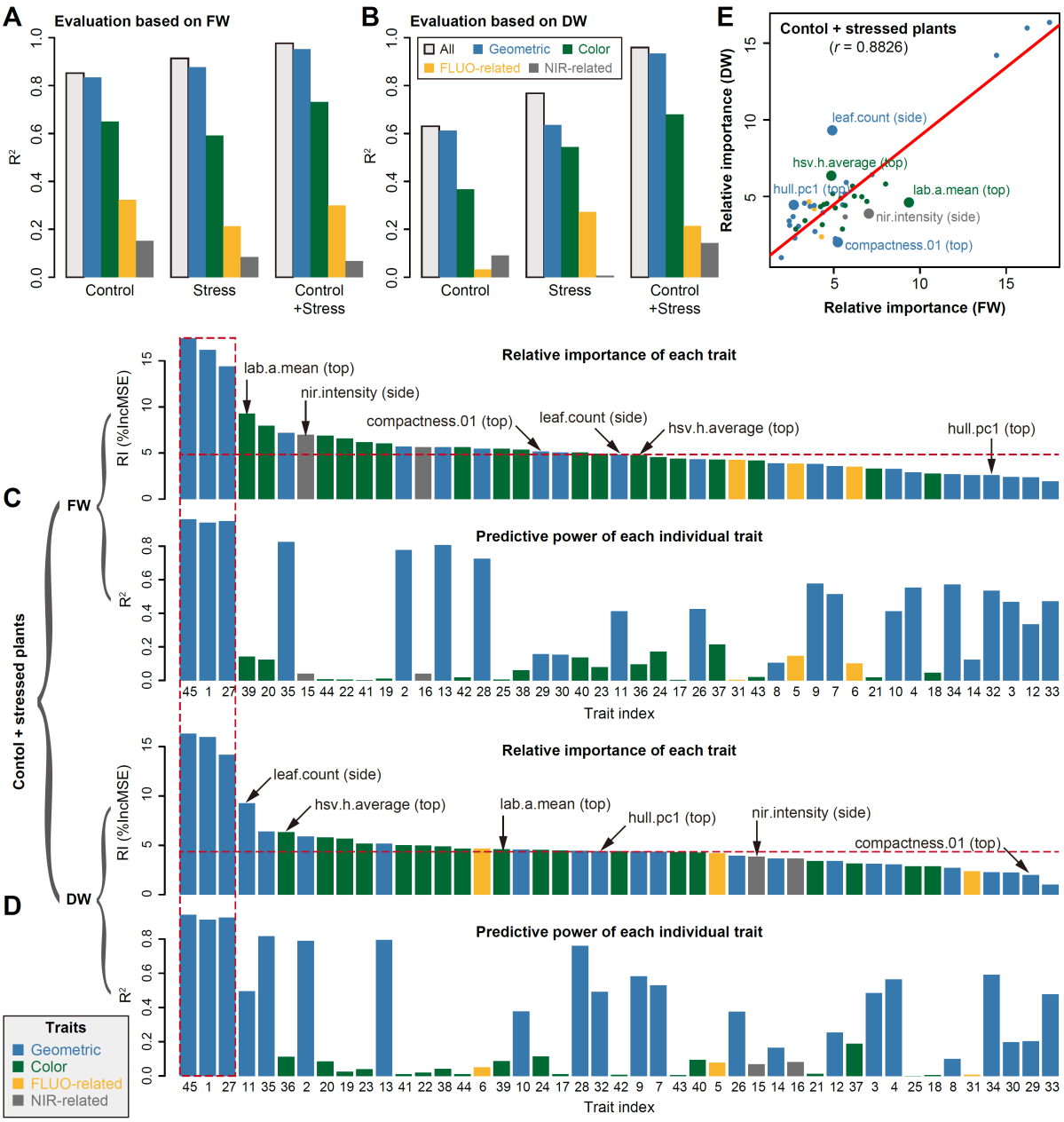

Figure 4.4: The relative importance of image-based features in prediction of plant biomass (legend on next page).

related and NIR-based features that are in the list of top ten most important ones revealed insubstantial predictive power in individual models. This observation implies that the underlying biomass determinant is extremely complex rather a linear combinations of the investigated features.

Furthermore, I compared the relative importance of each feature in predicting FW and DW (Figure 4.4E). Although I observed a positive correlation ($r = 0.88$) between the feature importance for FW and DW, there are several features showed largely different, including "nir.intensity" (derived from side view images), "compactness.01" (top), "hull.pc1" (top), "leaf.count" (side), "hsv.h.average" (top) and "lab.a.mean" (top). For instance, NIR intensity and plant compactness (top view) may be important for predicting FW but not for DW. Meanwhile, I performed the above analyses using only control (Figure

4.5) or stressed plants (Figure 4.6), respectively. It was found that the patterns of feature importance were distinct between these two groups of plants. For example, NIR intensity was ranked in the top fifth feature for predicting FW for stressed plants but not substantial for control plants. These findings indicate that the difference in plant biomass determinant is reflected by their image-based phenotypic traits.

► **Figure 4.4** (continued). The capabilities of different types of image-based features to predict plant biomass based on evaluation of either fresh weight (FW) **(A)** or dry weight (DW) **(B)**. The overall predictive accuracies of each types of features are indicated. Grey bar denotes the predictive accuracy using all features. The relative importance of each feature in the Random Forest model (upper panel) and the predictive accuracy of each individual feature as the single predictor (lower panel) based on investigation of either FW **(C)** or DW **(D)**. The calculation is based on the whole set of plants (control and stressed plants). Note that feature labels are shared in the upper and lower panels. Features are shown in numbers as ordering by their names. Three features highlighted in red dash box are digital volume, projected side area and projected top area. **(E)** Comparison of the relative importance of features in prediction of FW and DW. The top six most different features are highlighted and labeled. ■

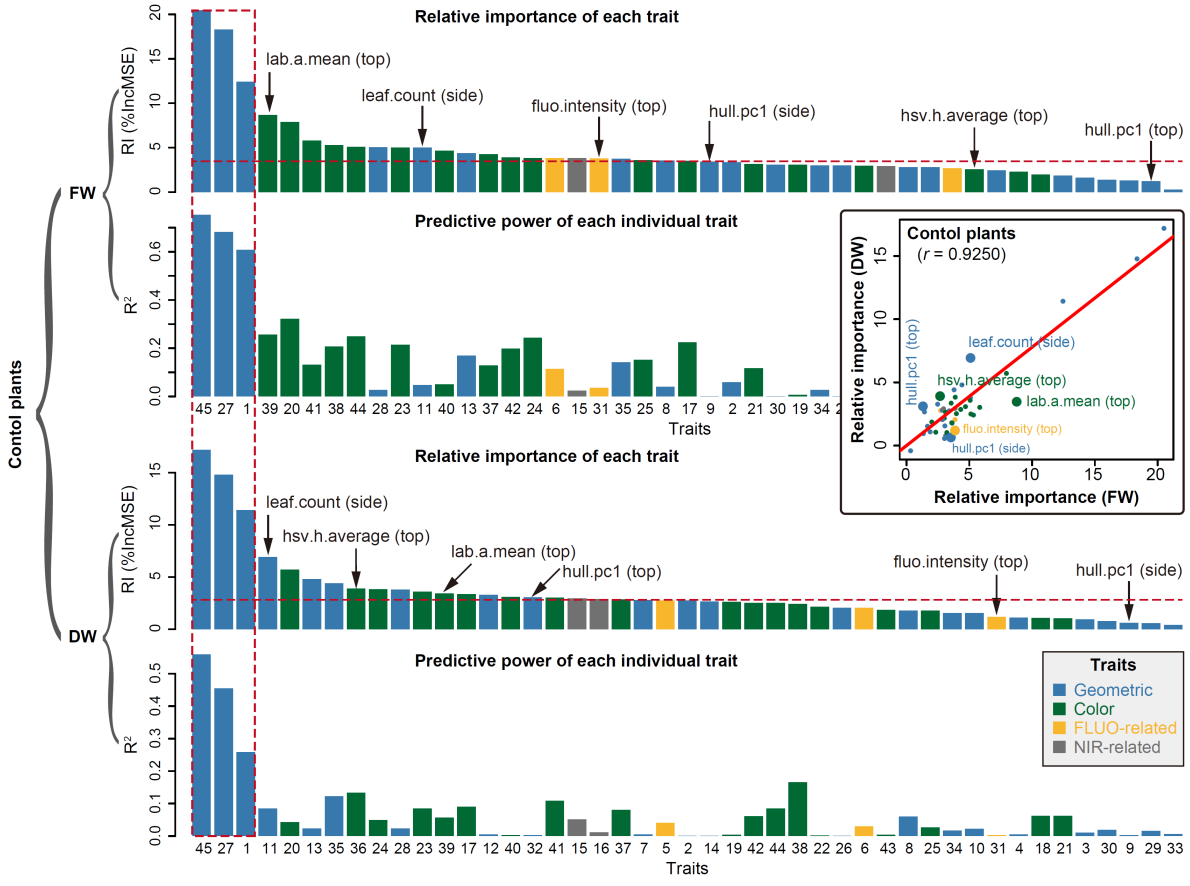

Figure 4.5: The relative importance of image-based features in prediction of biomass in control plants Refer to Figure 4.4 for legend. The calculation is based on control plants. ■

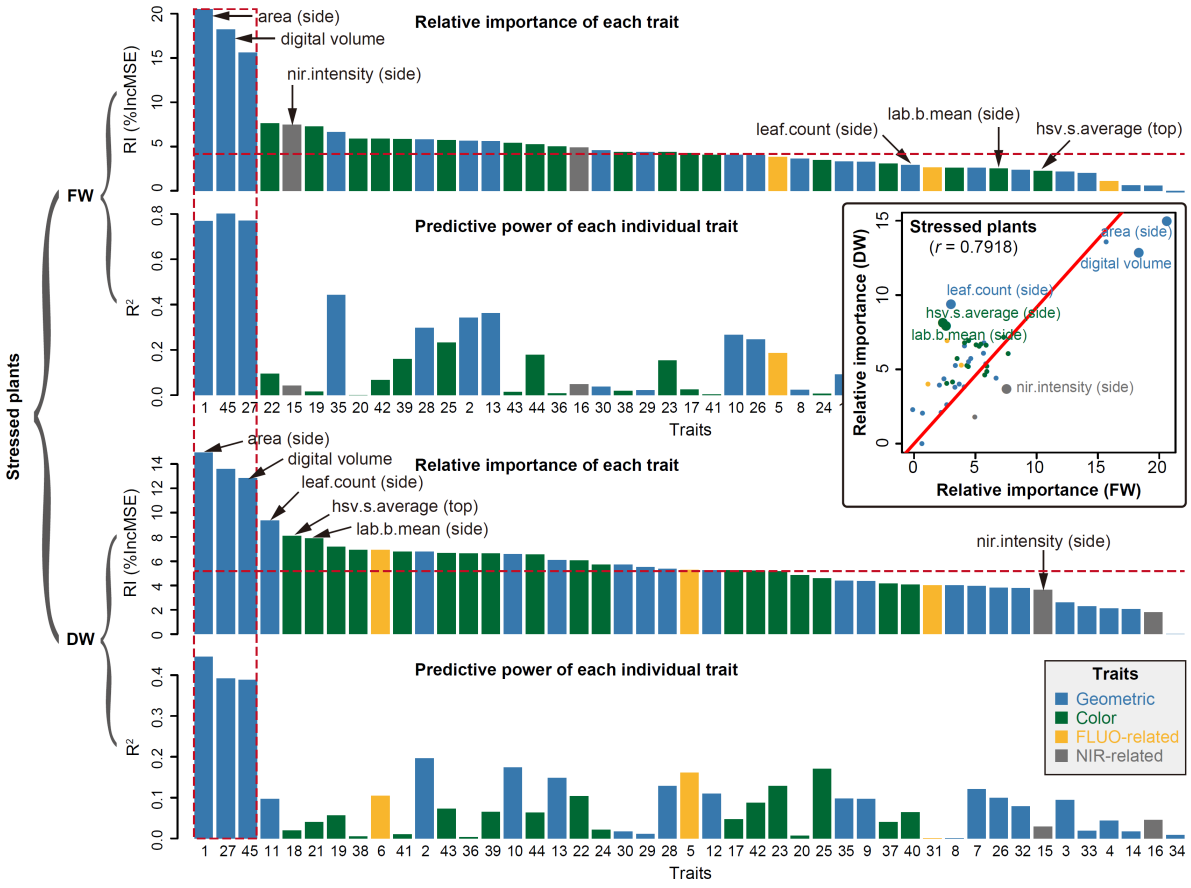

Figure 4.6: The relative importance of image-based features in prediction of biomass in stressed plants Refer to Figure 4.4 for legend. The calculation is based on stressed plants. ■

4.2.5 Image-based features are predictive of plant biomass across experiments with similar conditions or treatments

In this section, I set out to explore whether the models are generalizable across different experiments. I applied the models trained in one experiment to predict biomass (herein FW) in other experiments using a common set of features. Examples of such cross-experiment prediction are shown in Figure 4.7A, wherein I tested all possibility of cross prediction using the whole set of plants in the corresponding experiment. In general, the prediction accuracy within individual experiments remains high ($r > 0.97$ and $R^2 > 0.93$ for all three experiments; Figure 4.7B), revealing that the models are effective at predicting plant biomass by image-based feature signals among different experiments. Moreover, the prediction accuracy of cross-experiment prediction is still relatively high, with $r > 0.81$ and $R^2 > 0.65$, implying that the models accurately captured the relationships among the various image-based features. However, it was observed that the third experiment has relative weaker correlations with other two experiments for predicting biomass, while the first two experiments showed strong correlations or even identical with each other (Figure 4.7A). This may be mainly due to plants in experiment 3 behaving very differently from plants in experiments 1 and 2 (Neumann et al., 2015).

At the same time, I tested cross predicability of the models using treatment-specific data in the

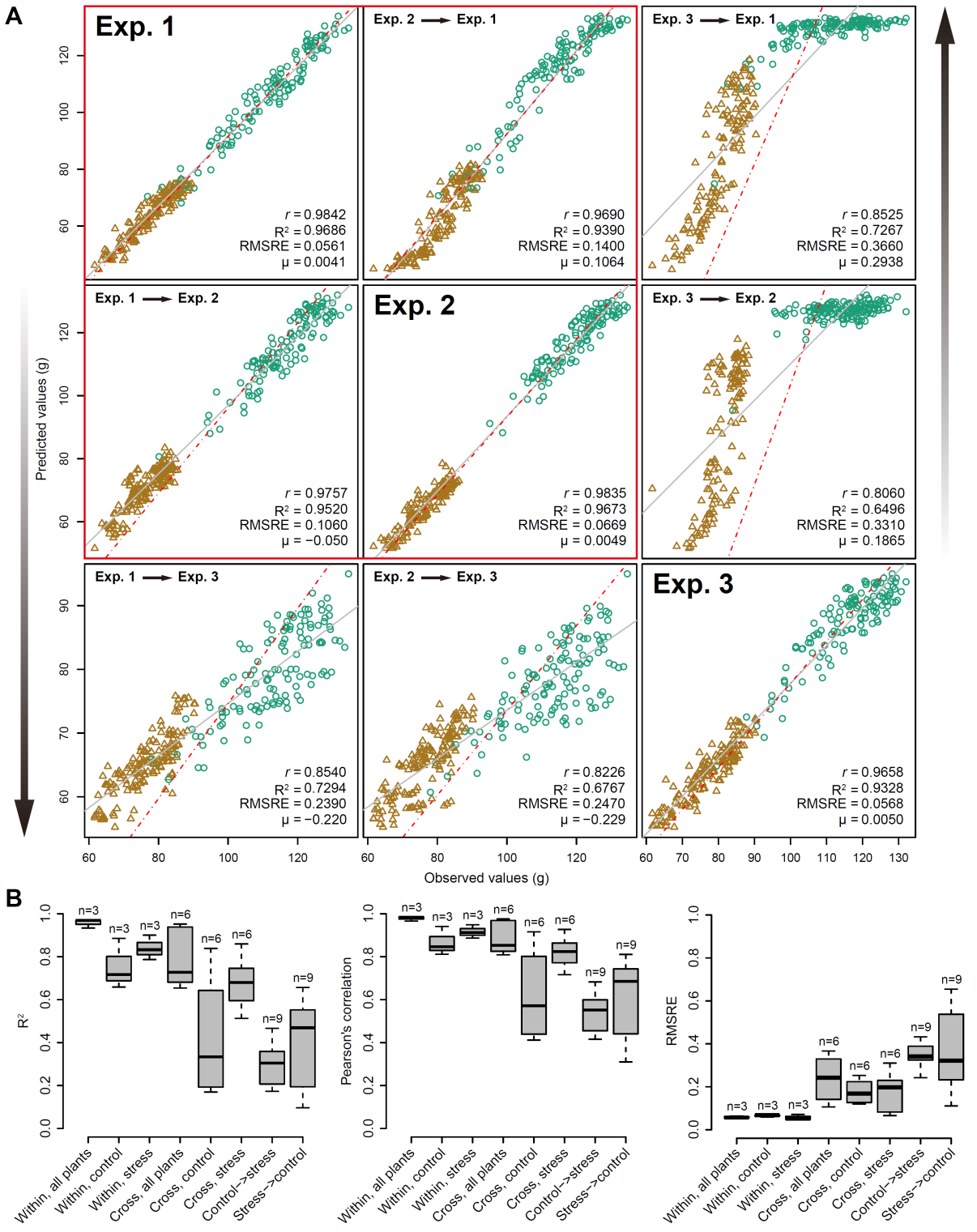

Figure 4.7: Comparison of prediction accuracy across different experiments

(legend on next page).

experiments (Figure 4.8). Similar results were observed as above using the whole dataset (Figure 4.7B). The weak predictive power of cross prediction for control plants resulted from the third experiment which showed powerless in predicting biomass of other experiments or predicted by other experiments. Generally, control and stressed plants were found to have very weak predictive power with each other

(Figure 4.8), as supported by the distinct patterns of relative feature importance between these two plant groups (Figures 4.5 and 4.6).

► **Figure 4.7** (continued). **(A)** Application of the model learned from one experiment to other experiments. **(B)** Boxplots of coefficient determination (R^2, left) Pearson's correlation coefficients (r, middle) and the root mean squared relative error ($RMSRE$, right) for different comparisons. "Within" denotes a model trained and tested on data from the same dataset with specific treatments (control, stress or both), and "Cross" represents a model trained on one dataset and tested on another dataset. "Control → stress" denotes a model trained on data with control treatment and tested on data with stress treatment, and vice versa for "stress → control". ■

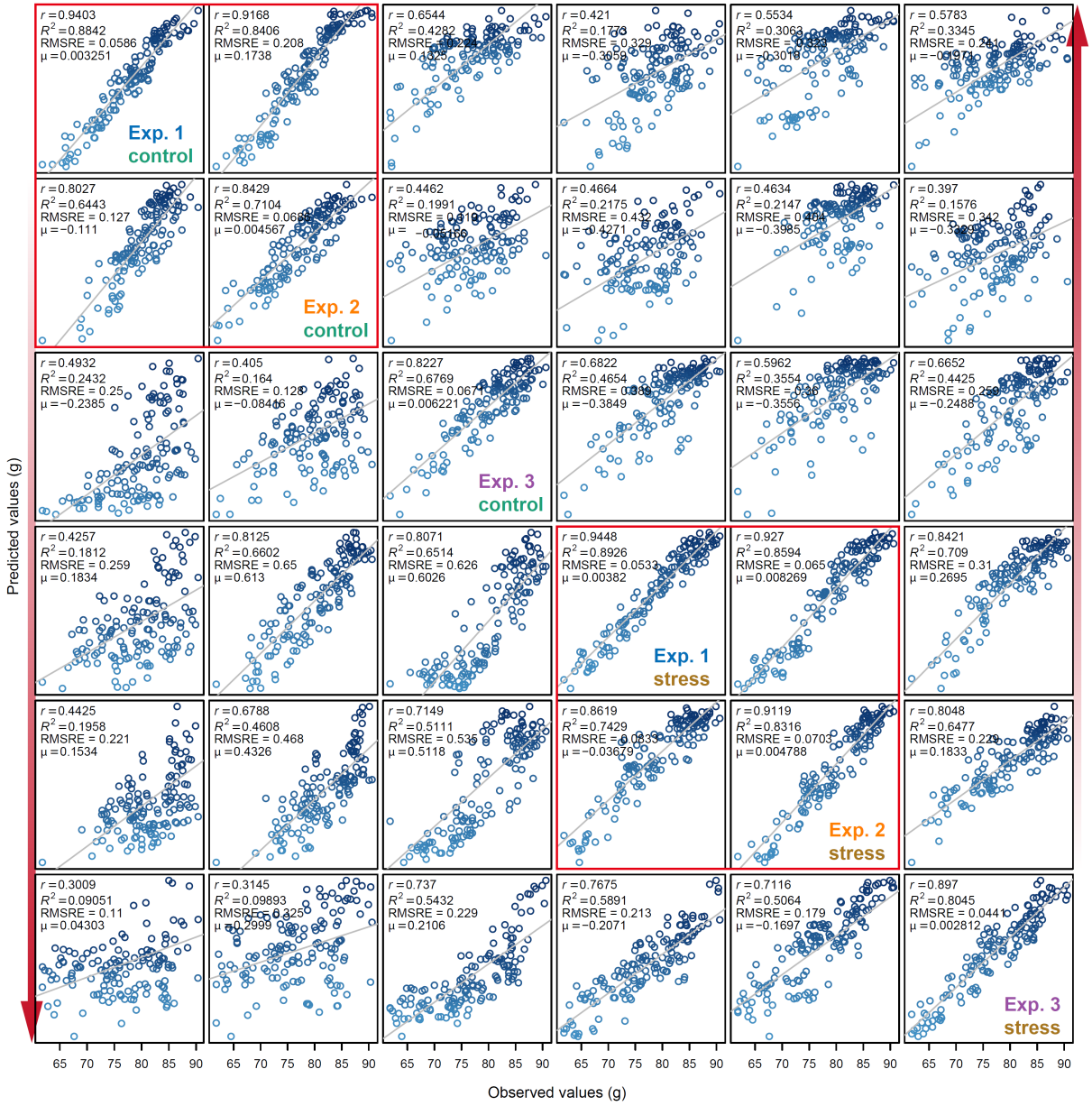

Figure 4.8: Comparison of prediction accuracy across different treatments

Refer to Figure 4.7B for legend. ■

4.3 Discussion

Biomass is a complex but an important trait in functional ecology and agronomy to study plant growth, crop productive potential and regeneration. Many different techniques, in either destructive or non-destructive manners, have been used to estimate biomass (Catchpole and Wheeler, 1992). The typical destructive methods for measuring biomass are very time consuming and labor intensive when investigating many individuals at the same time. While non-destructive imaging methods do not have such limitations, accurately predicting biomass from image data requires efficient mathematical models as well as representative image-derived features. Although previous attempts have been made to estimate plant biomass from image data, most of these studies consider only a single image-based feature or very few features in their models which are often linear-based, ignoring the fact that the phenotypic components underlying biomass accumulation are presumably complex.

In this study, I have presented a systematic analysis of relationship between plant biomass accumulation and image-derived signals, to confirm the assumption that biomass can be accurately predicted from image-based parameters. I built a random forest model of biomass accumulation using a comprehensive list of representative image-based features. In the comparison between the RF model and alternative regression models, it was found that the RF model outperforms other models in terms of (1) better predictive power – especially in comparison with the linear model, confirming the complex phenotypic architecture of biomass, and (2) feasible biological interpretability — the ability to readily extract information about the importance of each feature in prediction. The high prediction accuracy based on this model, in particular the cross-experiment performance, is promising to relieve the phenotyping bottleneck in biomass measurement in breeding applications. For example, based on an established small reference dataset to train a RF model, it is possible to predict biomass in several large plant populations within one experiment or across several experiments using image data by taking advantage of high-throughput phenotyping technologies. However, because of environmental effects on biomass accumulation, the application of the model will require the testing experiments showing similar conducted conditions with that of the reference experiment. Alternatively, the model can be trained from a much larger reference panel of plants that are grown in diverse environmental conditions and then is applied to a diverse set of experiments. This notion is first evidenced from the observation that the model shows more predictive power in plants with two treatments than with single treatment (Figure 4.3). Indeed, when applying the model to the combined dataset from all the three experiments, it was found the prediction accuracy remains very high ($R^2 = 0.96$ and $r = 0.98$, average values from ten times of ten-fold cross-validation).

In contrast to previous studies (Arvidsson et al., 2011; Dietz and Steinlein, 1996; Feng et al., 2013; Golzarian et al., 2011; Hairmansis et al., 2014; Leister et al., 1999; Neilson et al., 2015; Paruelo et al., 2000; Tackenberg, 2007; Walter et al., 2007), in which biomass was investigated using only single image-derived parameter (such as projected area) or several geometric parameters, the analyses extend these studies by incorporating more representative features that cover both structural and physiological-related properties into a more sophistic model. Although the predictive power of the model is roughly higher

than that of single feature-based prediction, such as the digital volume (Figure 4.3; Chen et al., 2014b), the model reveals the relative contribution of individual feature in prediction of biomass. The information regarding the importance of each feature will offer new insights into the phenotypic determinants of plant biomass outcome. Interestingly, it was found that several top ranked features, such as digital volume and NIR intensity, show genetic correlations with biomass of fresh weight (Figures 4.4C and 2.13; Chen et al., 2014b), implying these top ranked features may represent the main "phenotypic components" of biomass outcome and can be further used to dissect genetic components underlying biomass accumulation. However, as the current ability to characterize plant physiological-related properties from image data is still poor, I believe that the model can be further improved when new types of cameras and/or newly defined features are available.

In summary, I have developed a quantitative model for dissecting the phenotypic components of biomass accumulation based on image data. Apart from predicting biomass outcome, the methods can be used to determine the most important image-based features related to plant biomass accumulation, which are promising for subsequent genetic mapping to uncover the genetic basis of biomass. I anticipate that these statistical methods will be broadly used in plant breeding in the context of phenomics.

4.4 Materials and methods

4.4.1 Germplasm and experiments

Barley plant image data were obtained as described in Chapter 2.4.1 and were recently published elsewhere (Chen et al., 2014b; Neumann et al., 2015). Briefly, a core set of 16 two-rowed spring barley cultivars (*Hordeum vulgare* L.) and two parental cultivars of a double haploid (DH) were monitored for vegetative biomass accumulation (Table 2.1). Three independent experiments with identical setup were performed in a (semi-) controlled greenhouse at IPK by using the automated phenotyping and imaging platform LemnaTec-Scanalyzer 3D. Experiments were performed consecutively from May to November 2011 over a period of 58 days each (Table 4.1). The greenhouse setup enabled sowing for the next experiment already 2 days before the old experiment ended. For this, new pots were placed in the middle of the greenhouse, while the old experiment was still on the conveyer belts.

Each experiment consisted of two treatments: well-watered (control treatment) and water limited (drought stress treatment). In each treatment, nine plants per core set cultivar as well as six plants per DH parent were tested. This resulted in a total of 312 plants per experiment, corresponding to the maximal capacity of the phenotyping platform. Watering and imaging were performed daily. Drought stress was imposed by intercepting water supply from 27 days after sowing (DAS 27) until DAS 44. Stressed plants were re-watered at DAS 45. In total, each of the experiments was accumulating about 100 GB of raw data. At the end of experiments (DAS 58), plants were harvested to measure above-ground biomass in form of plant fresh weight (FW; for all experiments) and/or dry weight (DW; for experiment 1).

4.4.2 Image analysis

Image datasets were processed by the barley analysis pipelines in the IAP software. Analysed results were exported in the csv file format via IAP functionalities, which can be used for further data inspection. The result table includes columns as different phenotypic traits and rows as imaged plants over time. The corresponding metadata is included in the result table as well.

Each plant was characterized by a set of phenotypic traits also referred to as features, which were grouped into four categories: geometric features, fluorescence-related (FLUO-related) features, color-related features and near-infrared-related (NIR-related) features. These traits were defined by considering image information from different cameras (visible light, fluorescence and near infrared) and imaging views (side and top views). See the IAP online documentation (http://iapg2p.sourceforge.net/documentation.pdf) for details about trait definition.

4.4.3 Feature selection

Feature selection was performed with the same procedure as described in Chapter 2.4.4. I applied the feature selection technique to each dataset. Generally, almost identical subset features were captured from different datasets. I manually added several representative traits due to removal by variance inflation factors. For example, the digital volume and projected area are highly correlated with each other but both of them were kept, because I would investigate the predictive power of both features. Moreover, the regression models are insensitive to collinear features. I thus kept as much representative features as possible. To apply the prediction models among different datasets, a common set of features supported by all the datasets were used.

4.4.4 Data transformation

Each plant can be presented by a representative list of phenotypic traits, resulting in a matrix $X_{n\times m}$ for each experiment, where n is the number of plants and m is the number of phenotypic traits. Missing values were filled by mean values of other replicated plants. To make the image-derived parameters from diverse sources comparable, the columns of X was normalized by dividing by the maximum value of each column across all plants. Plants with empty values of manual measurements (FW and DW) were discarded for analysis. These transformed data were subjected to regression models.

4.4.5 Hierarchical clustering analysis and PCA

Hierarchical clustering analysis (HCA) and principle component analysis (PCA) were performed on the transformed data matrix $X_{n\times m}$ in the same way as described in Chapter 2.4.5.HCA was also performed using the genotype-level mean value of FW data to check the similarity of overall plant growth patterns in different experiments.

4.4.6 Models for predicting plant biomass

To understand the underlying relationship between image-derived parameters and the accumulated biomass (such as FW and DW), I constructed predictive models based on four different machine-learning methods: multivariate linear regression (MLR), multivariate adaptive regression splines (MARS), random forest (RF) and support vector regression (SVR). In these models, the normalized phenotypic profile matrix $X_{n\times m}$ for a representative list of phenotypic traits were used as the predictors (explanatory variables) and the measured DW/FW as the response variable Y.

All these models were implemented in R (http://www.r-project.org/; release 2.15.2). To assess the relative contribution of each phenotypic trait to predicting the biomass, the relative feature importance for each model was also calculated . Specifically, for the MLR model, I used "lm" function in the base installation packages. The relative importance of predictor variables in the MLR model were estimated by a heuristic method (Johnson, 2000) which decomposes the proportionate contribution of each predictor variable to R^2. For MARS, I used the "earth" function in the *earth* R package. The "number of subsets (nsubsets)" criterion (counting the number of model subsets that include the variable) was used to calculate the variable importance, which is implemented in the "evimp" function. For the RF model, I used the *randomForest* R package which implements Breiman's random forest algorithm (Breiman, 2001). I chose the "%IncMSE" (increase of mean squared error) to represent the criteria of relative importance measure. For SVR, I utilized the *e1071* R package which provides functionalities to use the libsvm library (Chang and Lin, 2011). The absolute values of the coefficients of the normal vector to the "optimal" hyperplane can be considered as the relative importance of each predictor variable contributing to regression (Iyer-Pascuzzi et al., 2010; Loo et al., 2007).

4.4.7 Evaluation of the prediction models

To evaluate the performance of the predictive models, a 10-fold cross-validation strategy was adopted to check the prediction power of each regression model. Specifically, each dataset was randomly divided into a training set (90% of plants) and a testing set (10% of plants). I trained a model on the training data and then applied it to predict biomass in the testing data. Afterwards, the predicted biomass in the testing set was compared with the manually measured biomass. The predictive accuracy of the model can be measured by

(i) the Pearson correlation coefficient (PCC; r) between the predicted values and the observed values;

(ii) the coefficient of determination (R^2) which equals to the fraction of variance of biomass explained by the model, defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

where SS_{res} and SS_{tot} are the sum of squares for residuals and the total sum of squares, respectively, $\widehat{y}_i$ the predicted and y_i the observed biomass of the ith plant, $\overline{y}$ is the mean value of the observed biomass; and

(iii) the root mean squared relative error of cross-validation, defined as

$$RMSRE = \sqrt{\frac{\sum_{i=1}^{s} (\frac{y_i - \widehat{y}_i}{y_i})^2}{s}}$$

where s denotes the sample size of the testing dataset.

I repeated the cross-validation procedure ten times. The mean and standard deviation of the resulting R^2 and $RMSRE$ values were calculated across runs.

To illustrate the broad utility of the methods across seasons (thus different growth environments) and treatments (e.g., control versus drought stress) in the same season, I applied the models in different contexts with cohort validation. Specifically, I trained the biomass prediction models under one specific context and predicted biomass in another different context and *vice versa*. The predictive accuracy of the model was evaluated based on the measures R^2 and $RMSRE$ as described above. Furthermore, the predictive power was reflected by the bias μ between the predicted and observed values, defined as

$$\mu = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\widehat{y}_i - y_i}{y_i}$$

where n denotes the sample size of the dataset. This bias indicates over- ($\mu > 0$) or under-estimation ($\mu < 0$) of biomass.

Chapter 5

Summary and outlook

5.1 Summary

Significantly improved crop varieties are urgently needed to feed the rapidly growing human population under changing climates. While genome sequence information and excellent genomic tools are in place for major crop species, the systematic quantification of phenotypic traits or components thereof in a high-throughput fashion remains an enormous challenge. In order to help bridge the genotype to phenotype gap, a comprehensive framework for high-throughput phenotype data analysis in plants was developed herein.

Within this framework, an extensive list of phenotypic traits can be extracted from non-destructive plant imaging over time. A series of supervised and unsupervised methods have been presented for efficient analysis and interpretation of huge and high-dimensional phenotypic data sets to support understanding plant growth and performance. As a proof of concept, I investigate the phenotypic components of the drought responses of 18 different barley cultivars during vegetative growth. I analyze dynamic properties of trait expression over growth time based on 54 representative phenotypic features. I use linear mixed models to dissect variance components of phenotypic traits and show that the traits revealed variable genotypic and environmental effects and their interactions over time. Key parameters such as trait heritability and genetic trait correlations are assessed, indicating image-derived traits are valuable in genetic association studies. These data are highly valuable to understand plant development and to further quantify growth and crop performance features.

I next test various growth models to predict plant biomass accumulation based on the image-derived parameter "digital volume" in both barley and maize under normal and stress conditions. It is found that barley and maize plants share simliar growth patterns as described by Weibull models. Several relevant parameters that support biological interpretation of plant growth and stress tolerance has been identified. These model-derived parameters reveal several important aspects regarding plant development and provide a solid basis for subsequent genetic mapping uncover the genetic basis of plant growth.

Finally, I construct several models to examine the quantitative relationship between image-based fea-

tures and plant biomass accumulation. I apply the methodology to three consecutive barley experiments with control and stress treatments. It is observed that plant biomass can be accurately predicted from image-based parameters using a random forest model. The high prediction accuracy based on this model, in particular the cross-experiment performance, is promising to relieve the phenotyping bottleneck in biomass measurement in breeding applications. I further quantify the relative contribution of individual feature for predicting biomass, revealing new insights into the phenotypic determinants of plant biomass outcome.

Taken together, I anticipate that the analytical framework and analysis results presented in this thesis will be useful to advance our views of phenotypic trait components underlying plant development and their responses to environmental cues, and will have broad applications in plant breeding under the context of phenomics.

5.2 Outlook

In this thesis, although I mainly validate the methodology using phenotypic data of barley cultivars collected by three different cameras, the framework is readily extensible to the analysis of other plant species (such as Arabidopsis, maize and wheat) and other newly developed sensors. In the near future, high-throughput plant phenotyping will receive a flood of applications in genetic mapping and mutant screening. In this regard, the analytical framework provides the starting point on the journey towards systematic plant phenotyping.

Bibliography

Abo-Elwafa, A. and Bakheit, B. (1999). Performance, correlation and path coefficient analysis in faba bean. *Assiut J Agric Sci*, (30):77–91.

Andrade-Sanchez, P., Gore, M. A., Heun, J. T., Thorp, K. R., Carmo-Silva, A. E., French, A. N., Salvucci, M. E., and White, J. W. (2014). Development and evaluation of a field-based high-throughput phenotyping platform. *Funct Plant Biol*, 41(1):68–79.

Araghi, S. G. and Assad, M. T. (1998). Evaluation of four screening techniques for drought resistance and their relationship to yield reduction ratio in wheat. *Euphytica*, 103(3):293–299.

Araus, J., Slafer, G., Reynolds, M., and Royo, C. (2002). Plant breeding and drought in c3 cereals: what should we breed for? *Ann Bot*, 89(7):925–940.

Araus, J. L. and Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends Plant Sci*, 19(1):52–61.

Araus, J. L., Slafer, G. A., Royo, C., and Serret, M. D. (2008). Breeding for yield potential and stress adaptation in cereals. *Crit Rev Plant Sci*, 27(6):377–412.

Archontoulis, S. V. and Miguez, F. E. (2013). Nonlinear regression models and applications in agricultural research. *Agron J*.

Arvidsson, S., Perez-Rodriguez, P., and Mueller-Roeber, B. (2011). A growth phenotyping pipeline for arabidopsis thaliana integrating image analysis and rosette area modeling for robust quantification of genotype effects. *New Phytol*, 191(3):895–907.

Aulchenko, Y. S., Ripke, S., Isaacs, A., and van Duijn, C. M. (2007). Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–6.

Baker, N. R. (2008). Chlorophyll fluorescence: a probe of photosynthesis in vivo. *Annu Rev Plant Biol*, 59:89–113.

Balachandran, S., Hurry, V., Kelley, S., Osmond, C., Robinson, S., Rohozinski, J., Seaton, G., and Sims, D. (1997). Concepts of plant biotic stress. some insights into the stress physiology of virus-infected plants, from the perspective of photosynthesis. *Physiologia Plantarum*, 100(2):203–213.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300.

Berger, B., de Regt, B., and Tester, M. (2012). High-throughput phenotyping of plant shoots. In *High-Throughput Phenotyping in Plants*, pages 9–20. Springer.

Berger, B., Parent, B., and Tester, M. (2010). High-throughput shoot imaging to study drought responses. *J Exp Bot*, 61(13):3519–28.

Biskup, B., Scharr, H., Fischbach, A., Wiese-Klinkenberg, A., Schurr, U., and Walter, A. (2009). Diel growth cycle of isolated leaf discs analyzed with a novel, high-throughput three-dimensional imaging method is identical to that of intact leaves. *Plant Physiol*, 149(3):1452–61.

Biskup, B., Scharr, H., Schurr, U., and Rascher, U. (2007). A stereo imaging system for measuring structural parameters of plant canopies. *Plant Cell Environ*, 30(10):1299–1308.

Blackman, V. (1919). The compound interest law and plant growth. *Ann Bot*, (3):353–360.

Borisjuk, L., Rolletschek, H., and Neuberger, T. (2012). Surveying the plant's world by magnetic resonance imaging. *Plant J*, 70(1):129–46.

Bouslama, M. and Schapaugh, W. T. (1984). Stress tolerance in soybeans. i. evaluation of three screening techniques for heat and drought tolerance1. *Crop Sci*, 24(5):933.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brien, C. J., Berger, B., Rabie, H., and Tester, M. (2013). Accounting for variation in designing greenhouse experiments with special reference to greenhouses containing plants on conveyor systems. *Plant Methods*, 9(1):1–22.

Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., Furbank, R. T., Badger, M., Pogson, B. J., and Borevitz, J. O. (2014). Traitcapture: genomic and environment modelling of plant phenomic data. *Curr Opin Plant Biol*, 18:73–79.

Bürling, K., Hunsche, M., and Noga, G. (2010). Quantum yield of non-regulated energy dissipation in psii (y (no)) for early detection of leaf rust (puccinia triticina) infection in susceptible and resistant wheat (triticum aestivum l.) cultivars. *Precision Agriculture*, 11(6):703–716.

Burnett, M. M. (2001). *Visual Programming*. John Wiley & Sons, Inc.

Busemeyer, L., Mentrup, D., Möller, K., Wunder, E., Alheit, K., Hahn, V., Maurer, H. P., Reif, J. C., Würschum, T., Müller, J., et al. (2013a). Breedvisiona multi-sensor platform for non-destructive field-based phenotyping in plant breeding. *Sensors*, 13(3):2830–2847.

Busemeyer, L., Ruckelshausen, A., Moller, K., Melchinger, A. E., Alheit, K. V., Maurer, H. P., Hahn, V., Weissmann, E. A., Reif, J. C., and Wurschum, T. (2013b). Precision phenotyping of biomass accumulation in triticale reveals temporal genetic patterns of regulation. *Sci Rep*, 3:2442.

Bylesjo, M., Segura, V., Soolanayakanahally, R. Y., Rae, A. M., Trygg, J., Gustafsson, P., Jansson, S., and Street, N. R. (2008). Lamina: a tool for rapid quantification of leaf size and shape parameters. *BMC Plant Biol*, 8:82.

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., and Luis Araus, J. (2012). High-throughput phenotyping and genomic selection: The frontiers of crop breeding convergef. *J Integr Plant Biol*, 54(5):312–320.

Calderini, D., Savin, R., Abeledo, L., Reynolds, M., and Slafer, G. (2001). *The importance of the period immediately preceding anthesis for grain weight determination in wheat*, pages 503–509. Springer.

Camargo, A., Papadopoulou, D., Spyropoulou, Z., Vlachonasios, K., Doonan, J. H., and Gay, A. P.

(2014). Objective definition of rosette shape variation using a combined computer vision and data mining approach. *PLoS ONE*, 9(5):e96889.

Cao, Q., Miao, Y., Wang, H., Huang, S., Cheng, S., Khosla, R., and Jiang, R. (2013). Non-destructive estimation of rice plant nitrogen status with crop circle multispectral active canopy sensor. *Field Crops Res*, 154:133–144.

Catchpole, W. and Wheeler, C. (1992). Estimating plant biomass: a review of techniques. *Australian J Ecol*, 17(2):121–131.

Chaerle, L., Hagenbeek, D., De Bruyne, E., Valcke, R., and Van Der Straeten, D. (2004). Thermal and chlorophyll-fluorescence imaging distinguish plant-pathogen interactions at an early stage. *Plant Cell Physiol*, 45(7):887–896.

Chaerle, L., Hagenbeek, D., De Bruyne, E., and Van Der Straeten, D. (2007a). Chlorophyll fluorescence imaging for disease-resistance screening of sugar beet. *Plant Cell, Tissue and Organ Culture*, 91(2):97–106.

Chaerle, L., Leinonen, I., Jones, H. G., and Van Der Straeten, D. (2007b). Monitoring and screening plant populations with combined thermal and chlorophyll fluorescence imaging. *J Exp Bot*, 58(4):773–784.

Chang, C. C. and Lin, C. J. (2011). Libsvm: A library for support vector machines. *Acm Transactions on Intelligent Systems and Technology*, 2(3).

Chapman, S. C., Merz, T., Chan, A., Jackway, P., Hrabar, S., Dreccer, M. F., Holland, E., Zheng, B., Ling, T. J., and Jimenez-Berni, J. (2014). Pheno-copter: a low-altitude, autonomous remote-sensing robotic helicopter for high-throughput field-based phenotyping. *Agron J*, 4(2):279–301.

Chen, D. (2016). Htpmod: an r package for modeling plant growth and its phenotypic components in the era of plant phenomics. *in preparation*.

Chen, D., Chen, M., Altmann, T., and Klukas, C. (2014a). *Bridging Genomics and Phenomics*, chapter 11, pages 299–333. Springer Berlin Heidelberg.

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

Chen, D., Shi, R., Pape, J.-M., and Klukas, C. (2015). Predicting plant biomass accumulation from image-derived parameters. *submitted (preprint doi: 10.1101/046656)*.

Chen, Y. and Lubberstedt, T. (2010). Molecular basis of trait correlations. *Trends Plant Sci*, 15(8):454–61.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*, 22(9):1658–1667.

Cheng, C. and Gerstein, M. (2012). Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res*, 40(2):553–568.

Cheng, C., Yan, K.-K., Yip, K. Y., Rozowsky, J., Alexander, R., Shou, C., Gerstein, M., et al. (2011).

A statistical framework for modeling gene expression using chromatin features and application to modencode datasets. *Genome Biol*, 12(2):R15.

Chern, C.-G., Fan, M.-J., Yu, S.-M., Hour, A.-L., Lu, P.-C., Lin, Y.-C., Wei, F.-J., Huang, S.-C., Chen, S., Lai, M.-H., et al. (2007). A rice phenomics studyphenotype scoring and seed propagation of a t-dna insertion-induced rice mutant population. *Plant Mol Biol*, 65(4):427–438.

Clark, R. T., MacCurdy, R. B., Jung, J. K., Shaff, J. E., McCouch, S. R., Aneshansley, D. J., and Kochian, L. V. (2011). Three-dimensional root phenotyping with a novel imaging and software platform. *Plant Physiol*, 156(2):455–65.

Cobb, J. N., DeClerck, G., Greenberg, A., Clark, R., and McCouch, S. (2013). Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theor Appl Genet*, 126(4):867–887.

Comar, A., Burger, P., de Solan, B., Baret, F., Daumard, F., and Hanocq, J.-F. (2012). A semi-automatic system for high throughput phenotyping wheat cultivars in-field conditions: description and first results. *Funct Plant Biol*, 39(11):914–924.

Damgaard, C. and Weiner, J. (2008). Modeling the growth of individuals in crowded plant populations. *Journal of Plant Ecology*, 1(2):111–116.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H., and Zhang, M. Q. (2006). Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A*, 103(28):10713–10716.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*, 12(7):499–510.

De Vylder, J., Vandenbussche, F., Hu, Y., Philips, W., and Van Der Straeten, D. (2012). Rosette tracker: an open source image analysis tool for automatic quantification of genotype effects. *Plant Physiol*, 160(3):1149–59.

Deery, D., Jimenez-Berni, J., Jones, H., Sirault, X., and Furbank, R. (2014). Proximal remote sensing buggies and potential applications for field-based phenotyping. *Agron J*, 4(3):349–379.

Dhondt, S., Wuyts, N., and Inze, D. (2013). Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci*, 18(8):428–39.

Dias, P. M. B., Brunel-Muguet, S., Dürr, C., Huguet, T., Demilly, D., Wagner, M.-H., and Teulat-Merah, B. (2011). Qtl analysis of seed germination and pre-emergence growth at extreme temperatures in medicago truncatula. *Theor Appl Genet*, 122(2):429–444.

Dietz, H. and Steinlein, T. (1996). Determination of plant species cover by means of image analysis. *Journal of Vegetation Science*, 7(1):131–136.

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigó, R., Birney, E., et al. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol*, 13(9):R53.

Duan, L., Yang, W., Huang, C., and Liu, Q. (2011). A novel machine-vision-based facility for the automatic evaluation of yield-related traits in rice. *Plant Methods*, 7:44.

Eberius, M. and Lima-Guerra, J. (2009). *High-Throughput plant phenotyping-data acquisition, transformation, and analysis*. Bioinformatics.

Edwards, D., Batley, J., and Snowdon, R. J. (2013). Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet*, 126(1):1–11.

Ehlert, D., Heisig, M., and Adamek, R. (2010). Suitability of a laser rangefinder to characterize winter wheat. *Precision Agriculture*, 11(6):650–663.

Ehlert, D., Horn, H.-J., and Adamek, R. (2008). Measuring crop biomass density by laser triangulation. *Comput Electron Agric*, 61(2):117–125.

El-Lithy, M. E., Clerkx, E. J., Ruys, G. J., Koornneef, M., and Vreugdenhil, D. (2004). Quantitative trait locus analysis of growth-related traits in a new arabidopsis recombinant inbred population. *Plant Physiol*, 135(1):444–58.

Erdle, K., Mistele, B., and Schmidhalter, U. (2011). Comparison of active and passive spectral sensors in discriminating biomass parameters and nitrogen status in wheat cultivars. *Field Crops Res*, 124(1):74–84.

Erickson, R. O. (1976). Modeling of plant-growth. *Annual Review of Plant Physiology and Plant Mol Biol*, 27:407–434.

Fekedulegn, D., Mac Siurtain, M. P., and Colbert, J. J. (1999). Parameter estimation of nonlinear growth models in forestry. *J Exp Bot*, 33(4):327–336.

Feng, H., Jiang, N., Huang, C., Fang, W., Yang, W., Chen, G., Xiong, L., and Liu, Q. (2013). A hyperspectral imaging system for an accurate prediction of the above-ground biomass of individual rice plants. *Review of Scientific Instruments*, 84(9):095107.

Fereres, E., Gimenez, C., and Fernandez, J. M. (1986). Genetic-variability in sunflower cultivars under drought .1. yield relationships. *Aust J Biol Sci*, 37(6):573–582.

Fernandez, G. (1992). *Effective selection criteria for assessing stress tolerance*. Proceedings of the International Symposium on Adaptation of Vegetables and Other Food Crops in Temperature and Water Stress. Taiwan.

Fiorani, F. and Schurr, U. (2013). Future scenarios for plant phenotyping. *Annu Rev Plant Biol*, 64:267–91.

Fischer, R. A. and Maurer, R. (1978). Drought resistance in spring wheat cultivars .1. grain-yield responses. *Aust J Biol Sci*, 29(5):897–912.

Flavel, R. J., Guppy, C. N., Tighe, M., Watt, M., McNeill, A., and Young, I. M. (2012). Non-destructive quantification of cereal roots in soil using high-resolution x-ray tomography. *J Exp Bot*, 63(7):2503–11.

Furbank, R. T. and Tester, M. (2011). Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci*, 16(12):635–44.

Gavuzzi, P., Rizza, F., Palumbo, M., Campanile, R. G., Ricciardi, G. L., and Borghi, B. (1997). Evaluation of field and laboratory predictors of drought and heat tolerance in winter cereals. *Canadian Journal of Plant Science*, 77(4):523–531.

Gilmour, A. R., Gogel, B., Cullis, B., and Thompson, R. (2009). Asreml user guide release 3.0.

Golzarian, M. R., Frick, R. A., Rajendran, K., Berger, B., Roy, S., Tester, M., and Lun, D. S. (2011). Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods*, 7:2.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, pages 513–583.

Granier, C., Aguirrezabal, L., Chenu, K., Cookson, S. J., Dauzat, M., Hamard, P., Thioux, J. J., Rolland, G., Bouchier-Combaud, S., Lebaudy, A., Muller, B., Simonneau, T., and Tardieu, F. (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 Phytol*, 169(3):623–35.

Granier, C. and Vile, D. (2014). Phenotyping and beyond: modelling the relationships between traits. *Curr Opin Plant Biol*, 18:96–102.

Green, J. M., Appel, H., Rehrig, E. M., Harnsomburana, J., Chang, J. F., Balint-Kurti, P., and Shyu, C. R. (2012). Phenophyte: a flexible affordable method to quantify 2d phenotypes from imagery. *Plant Methods*, 8(1):45.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21(1):27–58.

Hairmansis, A., Berger, B., Tester, M., and Roy, S. J. (2014). Image-based phenotyping for non-destructive screening of different salinity tolerance traits in rice. *Rice*, 7(1):16.

Harbinson, J., Prinzenberg, A. E., Kruijer, W., and Aarts, M. G. (2012). High throughput screening with chlorophyll fluorescence imaging and its use in crop improvement. *Curr Opin Biotechnol*, 23(2):221–226.

Harshavardhan, V. T., Van Son, L., Seiler, C., Junker, A., Weigelt-Fischer, K., Klukas, C., Altmann, T., Sreenivasulu, N., Baumlein, H., and Kuhlmann, M. (2014). Atrd22 and atuspl1, members of the plant-specific burp domain family involved in arabidopsis thaliana drought tolerance. *PLoS ONE*, 9(10):e110065.

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

Hillnhütter, C., Sikora, R., Oerke, E.-C., and Van Dusschoten, D. (2011). Nuclear magnetic resonance: a tool for imaging belowground damage caused by heterodera schachtii and rhizoctonia solani on sugar beet. *J Exp Bot*, page err273.

Holland, J. B., Nyquist, W. E., and Cervantes-Martnez, C. T. (2003). Estimating and interpreting heritability for plant breeding: An update. *Plant breeding reviews*, 22:9–112.

Honsdorf, N., March, T. J., Berger, B., Tester, M., and Pillen, K. (2014). High-throughput phenotyping to detect drought tolerance qtl in wild barley introgression lines. *PLoS ONE*, 9(5):e97047.

Hossain, A. B. S., Sears, R. G., Cox, T. S., and Paulsen, G. M. (1990). Desiccation tolerance and its relationship to assimilate partitioning in winter-wheat. *Crop Sci*, 30(3):622–627.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nat Rev Genet*, 11(12):855–66.

Hoyos-Villegas, V., Houx, J., Singh, S., and Fritschi, F. (2014). Ground-based digital imaging as a tool to assess soybean growth and yield. *Crop Sci*, 54(4):1756–1768.

Hunt, R. (1982). *Plant Growth Curves: The Functional Approach to Plant Growth*. London.

Ikeda, M., Hirose, Y., Takashi, T., Shibata, Y., Yamamura, T., Komura, T., Doi, K., Ashikari, M., Matsuoka, M., and Kitano, H. (2010). Analysis of rice panicle traits and detection of qtls using an image analyzing method. *Breeding Science*, 60(1):55–64.

Iyer-Pascuzzi, A. S., Symonova, O., Mileyko, Y., Hao, Y., Belcher, H., Harer, J., Weitz, J. S., and Benfey, P. N. (2010). Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems. *Plant Physiol*, 152(3):1148–57.

Jahnke, S., Menzel, M. I., van Dusschoten, D., Roeb, G. W., Buhler, J., Minwuyelet, S., Blumler, P., Temperton, V. M., Hombach, T., Streun, M., Beer, S., Khodaverdi, M., Ziemons, K., Coenen, H. H., and Schurr, U. (2009). Combined mri-pet dissects dynamic changes in plant structures and functions. *Plant J*, 59(4):634–44.

Jansen, M., Gilmer, F., Biskup, B., Nagel, K. A., Rascher, U., Fischbach, A., Briem, S., Dreissen, G., Tittmann, S., Braun, S., De Jaeger, I., Metzlaff, M., Schurr, U., Scharr, H., and Walter, A. (2009). Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via growscreen fluoro allows detection of stress tolerance in arabidopsis thaliana and other rosette plants. *Funct Plant Biol*, 36(10-11):902–914.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35(1):1–19.

Jones, H. G., Serraj, R., Loveys, B. R., Xiong, L. Z., Wheaton, A., and Price, A. H. (2009). Thermal infrared imaging of crop canopies for the remote diagnosis and quantification of plant responses to water stress in the field. *Funct Plant Biol*, 36(10-11):978–989.

Joosen, R. V., Arends, D., Li, Y., Willems, L. A., Keurentjes, J. J., Ligterink, W., Jansen, R. C., and Hilhorst, H. W. (2013). Identifying genotype-by-environment interactions in the metabolism of germinating arabidopsis seeds using generalized genetical genomics. *Plant Physiol*, 162(2):553–66.

Joosen, R. V. L., Arends, D., Willems, L. A. J., Ligterink, W., Jansen, R. C., and Hilhorst, H. W. (2012). Visualizing the genetic landscape of arabidopsis seed performance. *Plant Physiol*, 158(2):570–589.

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

Junker, B. H., Klukas, C., and Schreiber, F. (2006). Vanted: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7:109.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23.

Karadavut, U., A.Kayis, S., Palta, ., and Okur, O. (2008). A growth curve application to compare

plant heights and dry weights of some wheat varieties. *American-Eurasian J. Agric. & Environ. Sci.*, 3(6):888–892.

Karadavut, U., Palta, ., Kokten, K., and Bakoglu, A. (2010). Comparative study on some non-linear growth models for describing leaf growth of maize. *Int. J. Agric. Biol.*, 12(2):227–230.

Karkach, A. (2006). Trajectories and models of individual growth. *Demographic Research*, 15(12):347–400.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A*, 107(7):2926–2931.

Klose, R., Penlington, J., and Ruckelshausen, A. (2009). Usability study of 3d time-of-flight cameras for automatic plant phenotyping. *Bornimer Agrartechnische Berichte*, 69:93–105.

Klukas, C., Chen, D., and Pape, J. M. (2014). Integrated analysis platform: An open-source information system for high-throughput plant phenotyping. *Plant Physiol*, 165(2):506–518.

Knipling, E. B. (1970). Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. *Remote Sensing of Environment*, 1(3):155–159.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the Ieee*, 78(9):1464–1480.

Konishi, A., Eguchi, A., Hosoi, F., and Omasa, K. (2009). 3d monitoring spatio–temporal effects of herbicide on a whole plant using combined range and chlorophyll a fluorescence imaging. *Funct Plant Biol*, 36(11):874–879.

Leister, D., Varotto, C., Pesaresi, P., Niwergall, A., and Salamini, F. (1999). Large-scale evaluation of plant growth in arabidopsis thaliana by non-invasive image analysis. *Plant Physiology and Biochemistry*, 37(9):671–678.

Lenk, S., Chaerle, L., Pfündel, E. E., Langsdorf, G., Hagenbeek, D., Lichtenthaler, H. K., Van Der Straeten, D., and Buschmann, C. (2007). Multispectral fluorescence and reflectance imaging at the leaf level and its possible applications. *J Exp Bot*, 58(4):807–814.

Li, L., Zhang, Q., and Huang, D. (2014). A review of imaging techniques for plant phenotyping. *Sensors*, 14(11):20078–20111.

Lilley, J. M., Ludlow, M. M., McCouch, S. R., and OToole, J. C. (1996). Locating qtl for osmotic adjustment and dehydration tolerance in rice. *J Exp Bot*, 47(302):1427–1436.

Lin, C. S., Binns, M. R., and Lefkovitch, L. P. (1986). Stability analysis - where do we stand. *Crop Sci*, 26(5):894–900.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012). Gapit: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–9.

Liu, W., Gowda, M., Reif, J. C., Hahn, V., Ruckelshausen, A., Weissmann, E. A., Maurer, H. P., and Würschum, T. (2014). Genetic dynamics underlying phenotypic development of biomass yield in triticale. *BMC Genomics*, 15(1):458.

Lobet, G., Draye, X., and Perilleux, C. (2013). An online database for plant image analysis software tools. *Plant Methods*, 9(1):38.

Lohaus, G., Heldt, H., and Osmond, C. (2000). Infection with phloem limited abutilon mosaic virus causes localized carbohydrate accumulation in leaves of abutilon striatum: relationships to symptom

development and effects on chlorophyll fluorescence quenching during photosynthetic induction. *Plant Biology*, 2(2):161–167.

Loo, L. H., Wu, L. F., and Altschuler, S. J. (2007). Image-based multivariate profiling of drug responses from single cells. *Nat Methods*, 4(5):445–53.

Ma, B., Wilker, E. H., Willis-Owen, S. A., Byun, H.-M., Wong, K. C., Motta, V., Baccarelli, A. A., Schwartz, J., Cookson, W. O., Khabbaz, K., et al. (2014). Predicting dna methylation level across human tissues. *Nucleic Acids Res*, 42(6):3515–3528.

Mahner, M. and Kary, M. (1997). What exactly are genomes, genotypes and phenotypes? and what about phenomes? *Journal of Theoretical Biology*, 186(1):55–63.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res*, 27(2):209–20.

Maxwell, K. and Johnson, G. N. (2000). Chlorophyll fluorescencea practical guide. *J Exp Bot*, 51(345):659–668.

Meade, K. A., Cooper, M., and Beavis, W. D. (2013). Modeling biomass accumulation in maize kernels. *Field Crops Res*, 151:92–100.

Meijon, M., Satbhai, S. B., Tsuchimatsu, T., and Busch, W. (2014). Genome-wide association study using cellular traits identifies a new regulator of root development in arabidopsis. *Nat Genet*, 46(1):77–81.

Moore, C. R., Johnson, L. S., Kwak, I. Y., Livny, M., Broman, K. W., and Spalding, E. P. (2013). High-throughput computer vision introduces the time axis to a quantitative trait map of a plant growth response. *Genetics*, 195(3):1077–86.

Munns, R., James, R. A., Sirault, X. R., Furbank, R. T., and Jones, H. G. (2010). New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. *J Exp Bot*, 61(13):3499–507.

Nagel, K. A., Putz, A., Gilmer, F., Heinz, K., Fischbach, A., Pfeifer, J., Faget, M., Blossfeld, S., Ernst, M., Dimaki, C., Kastenholz, B., Kleinert, A. K., Galinski, A., Scharr, H., Fiorani, F., and Schurr, U. (2012). Growscreen-rhizo is a novel phenotyping robot enabling simultaneous measurements of root and shoot growth for plants grown in soil-filled rhizotrons. *Funct Plant Biol*, 39(10-11):891–904.

Neilson, E., Edwards, A., Blomstedt, C., Berger, B., Mller, B. L., and Gleadow, R. (2015). Utilization of a high-throughput shoot imaging system to examine the dynamic phenotypic responses of a c4 cereal crop plant to nitrogen and water deficiency over time. *J Exp Bot*.

Neumann, K., Klukas, C., Friedel, S., Rischbeck, P., Chen, D., Entzian, A., Stein, N., Graner, A., and Kilian, B. (2015). Dissecting spatio-temporal biomass accumulation in barley under different water regimes using high-throughput image analysis. *Plant Cell Environ*.

Nezhad, K. Z., Weber, W. E., Roder, M. S., Sharma, S., Lohwasser, U., Meyer, R. C., Saal, B., and Borner, A. (2012). Qtl analysis for thousand-grain weight under terminal drought stress in bread wheat (triticum aestivum l.). *Euphytica*, 186(1):127–138.

Nyquist, W. E. (1991). Estimation of heritability and prediction of selection response in plant-populations. *Crit Rev Plant Scis*, 10(3):235–322.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

Painawadee, M., Jogloy, S., Kesmala, T., Akkasaeng, C., and Patanothai, A. (2009). Heritability and correlation of drought resistance traits and agronomic traits in peanut (arachis hypogaea l.). *Asian J Plant Sci*, 8(5):325.

Paine, C. E. T., Marthews, T. R., Vogt, D. R., Purves, D., Rees, M., Hector, A., and Turnbull, L. A. (2012). How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists. *Methods Ecol Evol*, 3(2):245–256.

Paproki, A., Sirault, X., Berry, S., Furbank, R., and Fripp, J. (2012). A novel mesh processing based technique for 3d plant analysis. *BMC Plant Biol*, 12:63.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–90.

Paruelo, J. M., Lauenroth, W. K., and Roset, P. A. (2000). Estimating aboveground plant biomass using a photographic technique. *Journal of Range Management*, pages 190–193.

Paul-Victor, C., Zust, T., Rees, M., Kliebenstein, D. J., and Turnbull, L. A. (2010). A new method for measuring relative growth rate can uncover the costs of defensive compounds in arabidopsis thaliana. *New Phytol*, 187(4):1102–11.

Pereyra-Irujo, G. A., Gasco, E. D., Peirone, L. S., and Aguirrezábal, L. A. (2012). Glyph: a low-cost platform for phenotyping plant growth and water use. *Funct Plant Biol*, 39(11):905–913.

Phillips, R. L. (2010). Mobilizing science to break yield barriers. *Crop Sci*, 50(Supplement_1):S–99.

Pingali, P. L. (2012). Green revolution: Impacts, limits, and the path ahead. *Proc Natl Acad Sci U S A*, 109(31):12302–12308.

Porth, I., Klapste, J., Skyba, O., Lai, B. S., Geraldes, A., Muchero, W., Tuskan, G. A., Douglas, C. J., El-Kassaby, Y. A., and Mansfield, S. D. (2013). Populus trichocarpa cell wall chemistry and ultrastructure trait variation, genetic control and genetic correlations. *New Phytol*, 197(3):777–90.

Rajendran, K., Tester, M., and Roy, S. J. (2009). Quantifying the three main components of salinity tolerance in cereals. *Plant Cell Environ*, 32(3):237–49.

Rascher, U., Blossfeld, S., Fiorani, F., Jahnke, S., Jansen, M., Kuhn, A. J., Matsubara, S., Märtin, L. L., Merchant, A., Metzner, R., et al. (2011). Non-invasive approaches for phenotyping of enhanced performance traits in bean. *Funct Plant Biol*, 38(12):968–983.

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLoS ONE*, 8(6):e66428.

Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., and Foley, J. A. (2012). Recent patterns of crop yield growth and stagnation. *Nat Commun*, 3:1293.

Reuzeau, C. (2007). Traitmill (tm): A high throughput functional genomics platform for the phenotypic analysis of cereals. *In Vitro Cellular & Developmental Biology-Animal*, 43:S4–S4.

Reuzeau, C., Pen, J., Frankard, V., de Wolf, J., Peerbolte, R., and Broekaert, W. (2005). Traitmill: a discovery engine for identifying yield-enhancement genes in cereals. *Fenzi Zhiwu Yuzhong (Mol Plant Breeding)*, 3:7534.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., and Tardieu, F. (2003). Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. *Plant Physiol*, 131(2):664–75.

Ribaut, J. M., Jiang, C., GonzalezdeLeon, D., Edmeades, G. O., and Hoisington, D. A. (1997). Identification of quantitative trait loci under drought conditions in tropical maize .2. yield components and marker-assisted selection strategies. *Theor Appl Genet*, 94(6-7):887–896.

Richards, F. (1959). A flexible growth function for empirical use. *J Exp Bot*, 10(2):290–301.

Rolfe, S. A. and Scholes, J. D. (2010). Chlorophyll fluorescence imaging of plant–pathogen interactions. *Protoplasma*, 247(3-4):163–175.

Rosielle, A. A. and Hamblin, J. (1981). Theoretical aspects of selection for yield in stress and non-stress environments. *Crop Sci*, 21(6):943–946.

Sadok, W., Naudin, P., Boussuge, B., Muller, B., Welcker, C., and Tardieu, F. (2007). Leaf growth rate per unit thermal time follows qtl-dependent daily patterns in hundreds of maize lines under naturally fluctuating conditions. *Plant Cell Environ*, 30(2):135–46.

Saint Pierre, C., Crossa, J. L., Bonnett, D., Yamaguchi-Shinozaki, K., and Reynolds, M. P. (2012). Phenotyping transgenic wheat for drought resistance. *J Exp Bot*, page err385.

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., and Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nat Rev Genet*, 11(9):647–57.

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). Nih image to imagej: 25 years of image analysis. *Nat Methods*, 9(7):671–5.

Scholes, J. D. and Rolfe, S. A. (2009). Chlorophyll fluorescence imaging as tool for understanding the impact of fungal diseases on plant performance: a phenomics perspective. *Funct Plant Biol*, 36(11):880–892.

Schunk, C. and Eberius, M. (2012). *Phenomics in plant biological research and mutation breeding*, pages 535–560. CABI.

Seelig, H. D., Hoehn, A., Stodieck, L. S., Klaus, D. M., Adams, W. W., and Emery, W. J. (2008). The assessment of leaf water content using leaf reflectance ratios in the visible, near-, and short-wave-infrared. *Int J Remote Sens*, 29(13):3701–3713.

Seelig, H. D., Hoehn, A., Stodieck, L. S., Klaus, D. M., Adams, W. W., and Emery, W. J. (2009). Plant water parameters and the remote sensing r (1300)/r (1450) leaf water index: controlled condition dynamics during the development of water deficit stress. *Irrigation Sci*, 27(5):357–365.

Sellammal, R., Robin, S., and Raveendran, M. (2014). Association and heritability studies for drought resistance under varied moisture stress regimes in backcross inbred population of rice. *Rice Sci*, 21(3):150–161.

Skirycz, A., Vandenbroucke, K., Clauw, P., Maleux, K., De Meyer, B., Dhondt, S., Pucci, A., Gonzalez, N., Hoeberichts, F., Tognetti, V. B., Galbiati, M., Tonelli, C., Van Breusegem, F., Vuylsteke, M., and Inze, D. (2011). Survival and growth of arabidopsis plants given limited water are not equal. *Nat Biotechnol*, 29(3):212–4.

Slovak, R., Goschl, C., Su, X., Shimotani, K., Shiina, T., and Busch, W. (2014). A scalable open-source pipeline for large-scale root phenotyping of arabidopsis. *Plant Cell*, 26(6):2390–2403.

Sozzani, R. and Benfey, P. N. (2011). High-throughput phenotyping of multicellular organisms: finding the link between genotype and phenotype. *Genome Biol*, 12(3):219.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcamethods–a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23(9):1164–7.

Stackpole, D. J., Vaillancourt, R. E., Alves, A., Rodrigues, J., and Potts, B. M. (2011). Genetic variation in the chemical components of eucalyptus globulus wood. *G3 (Bethesda)*, 1(2):151–9.

Sultan, S. E. (2000). Phenotypic plasticity for plant development, function and life history. *Trends Plant Sci*, 5(12):537–42.

Svensgaard, J., Roitsch, T., and Christensen, S. (2014). Development of a mobile multispectral imaging platform for precise field phenotyping. *Agron J*, 4(3):322–336.

Swarbrick, P. J., SCHULZE-LEFERT, P., and Scholes, J. D. (2006). Metabolic consequences of susceptibility and resistance (race-specific and broad-spectrum) in barley leaves challenged with powdery mildew. *Plant Cell Environ*, 29(6):1061–1076.

Szira, F., Bálint, A., Börner, A., and Galiba, G. (2008). Evaluation of drought-related traits and screening methods at different developmental stages in spring barley. *Journal of Agronomy and Crop Sci*, 194(5):334–342.

Tackenberg, O. (2007). A new method for non-destructive measurement of biomass, growth rates, vertical biomass distribution and dry matter content based on digital image analysis. *Ann Bot*, 99(4):777–783.

Takeda, S. and Matsuoka, M. (2008). Genetic approaches to crop improvement: responding to environmental and population changes. *Nat Rev Genet*, 9(6):444–457.

Tardieu, F. and Tuberosa, R. (2010). Dissection and modelling of abiotic stress tolerance in plants. *Curr Opin Plant Biol*, 13(2):206–12.

Tessmer, O. L., Jiao, Y., Cruz, J. A., Kramer, D. M., and Chen, J. (2013). Functional approach to high-throughput plant growth analysis. *BMC Syst Biol*, 7 Suppl 6:S17.

Tester, M. and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. *Science*, 327(5967):818–22.

Thornley, J. H. and France, J. (2007). *Mathematical models in agriculture: quantitative methods for the plant, animal and ecological sciences*. Cabi.

Tilman, D., Balzer, C., Hill, J., and Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proc Natl Acad Sci U S A*, 108(50):20260–20264.

Tisne, S., Serrand, Y., Bach, L., Gilbault, E., Ben Ameur, R., Balasse, H., Voisin, R., Bouchez, D., Durand-Tardif, M., Guerche, P., Chareyron, G., Da Rugna, J., Camilleri, C., and Loudet, O. (2013). Phenoscope: an automated large-scale phenotyping platform offering high spatial homogeneity. *Plant J*, 74(3):534–44.

Topp, C. N., Iyer-Pascuzzi, A. S., Anderson, J. T., Lee, C.-R., Zurek, P. R., Symonova, O., Zheng, Y., Bucksch, A., Mileyko, Y., Galkovskyi, T., et al. (2013). 3d phenotyping and quantitative trait locus

mapping identify core regions of the rice genome controlling root architecture. *Proc Natl Acad Sci U S A*, 110(18):E1695–E1704.

Tuberosa, R. (2012). Phenotyping for drought tolerance of crops in the genomics era. *Front Physiol*, 3:347.

Tucker, C. J. (1980). Remote sensing of leaf water content in the near infrared. *Remote Sensing of Environment*, 10(1):23–32.

van der Heijden, G., Song, Y., Horgan, G., Polder, G., Dieleman, A., Bink, M., Palloix, A., van Eeuwijk, F., and Glasbey, C. (2012). Spicy: towards automated phenotyping of large pepper plants in the greenhouse. *Funct Plant Biol*, 39(11):870–877.

Van Poecke, R. M., Sato, M., Lenarz-Wyatt, L., Weisberg, S., and Katagiri, F. (2007). Natural variation in rps2-mediated resistance among arabidopsis accessions: correlation between gene expression profiles and phenotypic responses. *Plant Cell*, 19(12):4046–60.

Vanclay, J. (1994). *Modelling Forest Growth and Yield: Applications to Mixed Tropical Forests*. CAB International, Wallingford.

Varki, A., Wills, C., Perlmutter, D., Woodruff, D., Gage, F., Moore, J., Semendeferi, K., Bernirschke, K., Katzman, R., Doolittle, R., et al. (1998). Great ape phenome project? *Science*, 282(5387):239–240.

Verhulst, P.-F. (1977). A note on the law of population growth. In *Mathematical Demography*, pages 333–339. Springer.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet*, 9(4):255–66.

Waddington, C. H. (1968). Towards a theoretical biology. *Nature*, 218:525–527.

Walter, A., Scharr, H., Gilmer, F., Zierer, R., Nagel, K. A., Ernst, M., Wiese, A., Virnich, O., Christ, M. M., Uhlig, B., Junger, S., and Schurr, U. (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 Phytol*, 174(2):447–55.

Walter, A., Studer, B., and Kölliker, R. (2012). Advanced phenotyping offers opportunities for improved breeding of forage and turf species. *Ann Bot*, page mcs026.

Wang, H. X., Zhang, W. M., Zhou, G. Q., Yan, G. J., and Clinton, N. (2009). Image-based 3d corn reconstruction for retrieval of geometrical structural parameters. *Int J Remote Sens*, 30(20):5505–5513.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18(1):293297.

Weight, C., Parnham, D., and Waites, R. (2008). Leafanalyser: a computational method for rapid and large-scale analyses of leaf shape variation. *Plant J*, 53(3):578–86.

White, J. W., Andrade-Sanchez, P., Gore, M. A., Bronson, K. F., Coffelt, T. A., Conley, M. M., Feldmann, K. A., French, A. N., Heun, J. T., Hunsaker, D. J., et al. (2012). Field-based phenomics for plant genetics research. *Field Crops Res*, 133:101–112.

Woo, N. S., Badger, M. R., and Pogson, B. J. (2008). A rapid, non-invasive procedure for quantitative assessment of drought survival using chlorophyll fluorescence. *Plant Methods*, 4:27.

Wu, R. and Lin, M. (2006). Functional mapping - how to map and study the genetic architecture of dynamic complex traits. *Nat Rev Genet*, 7(3):229–37.

Würschum, T., Liu, W., Busemeyer, L., Tucker, M. R., Reif, J. C., Weissmann, E. A., Hahn, V., Ruckelshausen, A., and Maurer, H. P. (2014). Mapping dynamic qtl for plant height in triticale. *BMC Genet*, 15(1):59.

Xiong, L., Wang, R. G., Mao, G., and Koczan, J. M. (2006). Identification of drought tolerance determinants by genetic analysis of root response to drought stress and abscisic acid. *Plant Physiol*, 142(3):1065–74.

Yang, W., Duan, L., Chen, G., Xiong, L., and Liu, Q. (2013). Plant phenomics and high-throughput phenotyping: accelerating rice functional genomics using multidisciplinary technologies. *Curr Opin Plant Biol*, 16(2):180–187.

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., Fang, W., Feng, H., Xie, W., Lian, X., Wang, G., Luo, Q., Zhang, Q., Liu, Q., and Xiong, L. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat Commun*, 5:5087.

Yin, X., Struik, P. C., Tang, J., Qi, C., and Liu, T. (2005a). Model analysis of flowering phenology in recombinant inbred lines of barley. *J Exp Bot*, 56(413):959–65.

Yin, X., Struik, P. C., van Eeuwijk, F. A., Stam, P., and Tang, J. (2005b). Qtl analysis and qtl-based prediction of flowering phenology in recombinant inbred lines of barley. *J Exp Bot*, 56(413):967–76.

Zeide, B. (1993). Analysis of growth equations. *Forest Science*, 39(3):594–616.

Zhang, W., Spector, T., Deloukas, P., Bell, J., and Engelhardt, B. (2015). Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. *Genome Biol*, 16(1):14.

Zhang, X., Hause, R. J., J., and Borevitz, J. O. (2012). Natural genetic variation for growth and development revealed by high-throughput phenotyping in arabidopsis thaliana. *G3 (Bethesda)*, 2(1):29–34.

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., Mezey, J., McClung, A. M., Bustamante, C. D., and McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nat Commun*, 2:467.

Zheng, H., Wu, H., Li, J., and Jiang, S.-W. (2013). Cpgimethpred: computational model for predicting methylation status of cpg islands in human genome. *BMC Med Genomics*, 6(Suppl 1):S13.

Zhu, J., Ingram, P. A., Benfey, P. N., and Elich, T. (2011). From lab to field, new approaches to phenotyping root system architecture. *Curr Opin Plant Biol*, 14(3):310–317.

Zust, T., Joseph, B., Shimizu, K. K., Kliebenstein, D. J., and Turnbull, L. A. (2011). Using knockout mutants to reveal the growth costs of defensive traits. *Proc Biol Sci*, 278(1718):2598–603.

Appendix A

Glossary

Color space

Color. Color is the visual perceptual property of the spectrum of light (distribution of light power versus wavelength) received by the human eye with the categories called red, green, blue, and others. Note that the light (called "visible light") which excites the human visual system is a very small portion of the whole electromagnetic spectrum. LemnaTec system employs various controlled cameras, such as RGB/visible, fluorescence and NIR cameras, for hyperspectral reflectance measurement in the spectral regions (called "region of interest", ROI) of visible and near-infrared (VNIR), with the wavelengths ranging from 400 nanometers (nm) to 1700 nm (Figure 1.2). Colors can be described and defined numerically by their coordinates in the color space. Several color models have been mathematically described, such as RGB, HSB, HSL and L*a*b* color spaces (Figure A.1). Appropriate color spaces are important for image processing (for example, image segmentation). For example, we used the combined color spaces for image analysis in the IAP system (Klukas et al., 2014).

RGB color space. RGB stands for three primary colors: red (R), green (G) and blue (B). It is an additive color model in which the color perception is stimulated by the additive mixing of the primary colors. Particularly, the RGB model can be represented by a cube using non-negative values within a range of 0 and 255, and stores individual values for red, green and blue in a triplet (r, g, b) as the three-dimensional coordinate of the point of a given color. Black represents the origin of the cube at the vertex (0, 0, 0) and white at the diagonally opposite vertex (255, 255, 255). The human visual system works in a way that is quite similar to the RGB color space, which captures the largest portion of the human color space. However, due to the high correlation among the three primary colors in RGB space, *in silico* segmentation of color image poses a big challenge.

HSB color space. HSB stands for hue, saturation and brightness, also known as HSV (hue, saturation and value). HSB is a cylindrical-coordinate representation of points in an RGB color

model. The components and colorimetry of HSB color space are derived from the RGB color space. Hue stands for color shade and represents the main wavelength of the color within the visible light. The saturation is the intensity of the color it is the relative bandwidth of the visible output. Brightness is a relative expression of the intensity of the energy output of a visible light source. However, the HSB color space is not sufficient enough for image segmentation.

HSL color space. HSL, short for hue, saturation, lightness/luminance and also known as HSI (hue, saturation and intensity), is quite similar to HSV, with "lightness" replacing "brightness". The difference is that the brightness of a pure color is equal to the brightness of white, while the lightness of a pure color is equal to the lightness of a medium gray.

L*a*b* color space. L*a*b* color space is a color-opponent space with dimension L for lightness (always positive) and a and b for the color-opponent dimensions. A value larger than zero for "A" represents the red component, all negative values stand for the green part. A "B" value above zero represent the yellow component, the negative values stand for the blue component. The L*a*b* color space describes all the colors visible to the human eye and is created to serve as a device independent model to be used as a reference. It is a perceptually uniform color space in which a change of the same amount in a color value should produce a change of about the same visual importance. In L*a*b*, colorimetric determination of color coordinates and color differences.

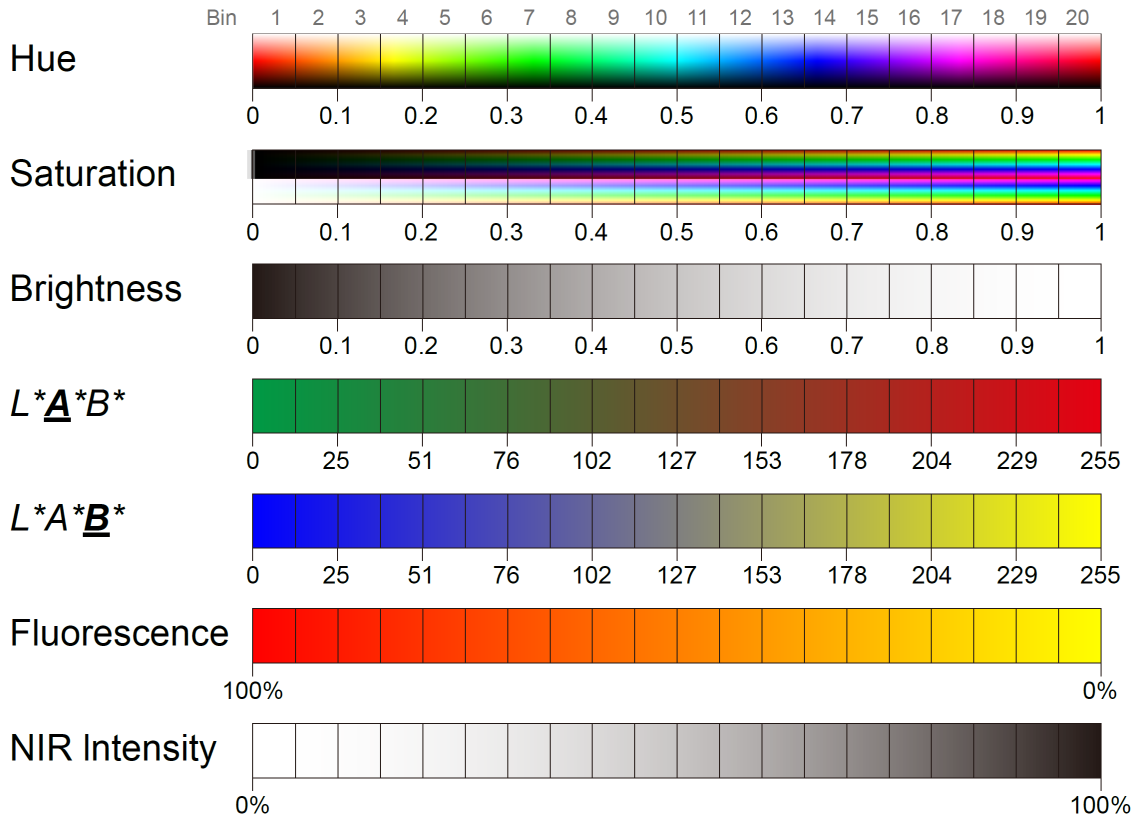

Figure A.1: Histogram bin-based feature extraction in different color spaces

Color analysis is based on histograms with 20-bins for visible light color and brightness investigation (based on HSB and L*a*b* color spaces), for fluorescence activity and NIR intensity investigations. In addition to the histogram values, for each property the average, the standard deviation and the skewness of the pixel values is calculated. ■

Appendix B

Supplemental Tables

Table S1: The 54 investigated phenotypic traits in barley.

Trait	Description	Category†	View§	Camera¶
side.area	projected area from side (filled pixels)	geometric	side	VIS
side.border.length	plant area border length	geometric	side	VIS
side.compactness.01.relative	geometric measure of plant compactness; 4 * π / (whole border length from side 2 / projected side area)	geometric	side	VIS
side.compactness.16.relative	geometric measure of plant compactness; whole border length from side 2 / projected side area	geometric	side	VIS
side.fluo.histogram.bin.02.12_25	number of pixels in intensity bin 2/20 (low)	FLUO	side	FLUO
side.fluo.histogram.bin.20.242_255	number of pixels in intensity bin 20/20 (high)	FLUO	side	FLUO
side.fluo.intensity.average..relative	average intensity of the fluorescence reflection based on the color (pure red highest intensity, yellow lowest intensity)	FLUO	side	FLUO
side.fluo.intensity.classic.average	average intensity of the fluorescence reflection based on the color and brightness (pure red highest intensity, yellow lowest intensity, value is scaled by the brightness)	FLUO	side	FLUO

Table S1 (continued)

side.height	plant height (px or mm)	geometric	side	VIS
side.hull.fillgrade..percent	projected area / area of convex hull	geometric	side	VIS
side.hull.pc1	largest distance between any plant pixels from side view (these pixels are the base for the 'maximum distance line')	geometric	side	VIS
side.hull.pc2	sum of the maximum distance of pixels left and right to the 'maximum distance line'	geometric	side	VIS
side.leaf.count..leaves	estimated leaf count, based on the number of end points of the plant skeleton	geometric	side	VIS
side.leaf.length.sum	length of the plant skeleton	geometric	side	VIS
side.leaf.width.average	average distance of plant pixels to the nearest skeleton pixel	geometric	side	VIS
side.nir.histogram.bin.11.127_140	number of pixels with NIR intensity in the rage of 127-140 (0 no intensity, 255 highest intensity)	NIR	side	NIR
side.nir.histogram.bin.13.153_165	number of pixels with NIR intensity in the rage of 153-165 (0 no intensity, 255 highest intensity)	NIR	side	NIR
side.nir.intensity.average..relative	average near-infrared intensity of plant pixels (0..255, 0 no intensity, 255 highest intensity)	NIR	side	NIR
side.vis.hsv.h.average	average hue of plant pixels	color	side	VIS
side.vis.hsv.h.histogram.bin.01.0_12	number of plant pixels within specific hue range	color	side	VIS
side.vis.hsv.h.histogram.bin.13.153_165	number of plant pixels within specific hue range	color	side	VIS
side.vis.hsv.s.average	average plant pixels color saturation	color	side	VIS
side.vis.hsv.v.average	average brightness	color	side	VIS
side.vis.hsv.v.histogram.bin.04.38_51	number of plant pixels with brightness in specific range	color	side	VIS
side.vis.hsv.v.histogram.bin.17.204_216	number of plant pixels with brightness in specific range	color	side	VIS
side.vis.hsv.v.histogram.bin.20.242_255	number of plant pixels with brightness in specific range	color	side	VIS
side.vis.lab.b.kurtosis	"peakedness" of the b* (blue to yellow) values of the plant color histogram, calculated in the l*a*b*-color space	color	side	VIS
side.vis.lab.b.mean	average color in the b* range of the L*a*b* color space (blue to yellow)	color	side	VIS
side.vis.stress.hue.yellow2green	yellow to green ratio of plant pixels (based on according hsv hue classes)	color	side	VIS
side.width	horizontal extend of the plant	geometric	side	VIS
top.area	projected area from top view (filled pixels)	geometric	top	VIS

Table S1 (continued)

top.compactness.16..relative	geometric measure of plant compactness (whole border length from side) 2 / projected side area	geometric	top	VIS
top.fluo.histogram.bin.04.38_51	number of pixels in intensity bin 4/20 (rel. low)	FLUO	top	FLUO
top.fluo.histogram.bin.06.63_76	number of pixels in intensity bin 6/20 (rel. middle)	FLUO	top	FLUO
top.fluo.histogram.bin.12.140_153	number of pixels in intensity bin 4/20 (rel. high)	FLUO	top	FLUO
top.fluo.intensity.phenol.plant_weight_drought_loss	top plant area reduced by a penalty term for yellowish plant parts (drought stressed leaf areas appear yellowish)	FLUO	top	FLUO
top.hull.pc1	largest distance between any plant pixels from top view (these pixels are the base for the 'maximum distance line')	geometric	top	VIS
top.hull.pc2	sum of the maximum distance of pixels left and right to the 'maximum distance line'	geometric	top	VIS
top.leaf.count	estimated number of leafs (skeleton based)	geometric	top	VIS
top.leaf.length.sum	length of plant pixel skeleton	geometric	top	VIS
top.ndvi.vis.blue.intensity.average..relative	average blue intensity in the rgb color space	color	top	VIS
top.vis.hsv.h.average	average hue in the hsv color space	color	top	VIS
top.vis.hsv.h.histogram.bin.07.76_89	number of pixels in bin 7/20 of the hue histogram (yellow to green)	color	top	VIS
top.vis.hsv.h.histogram.bin.10.114_127	number of pixels in bin 10/20 of the hue histogram (green to blue)	color	top	VIS
top.vis.hsv.h.histogram.bin.19.229_242	number of pixels in bin 19/20 of the hue histogram (red)	color	top	VIS
top.vis.hsv.s.average	average plant pixel color saturation	color	top	VIS
top.vis.hsv.s.histogram.bin.10.114_127	number of pixels in bin 10/20 of the saturation histogram (middle)	color	top	VIS
top.vis.hsv.v.average	average brightness	color	top	VIS
top.vis.hsv.v.histogram.bin.03.25_38	number of pixels with brightness in a specific range (bin 3/20, low)	color	top	VIS
top.vis.hsv.v.histogram.bin.07.76_89	number of pixels with brightness in a specific range (bin 7/20, middle)	color	top	VIS
top.vis.hsv.v.histogram.bin.09.102_114	number of pixels with brightness in a specific range (bin 9/20, middle)	color	top	VIS
top.vis.hsv.v.histogram.bin.17.204_216	number of pixels with brightness in a specific range (bin 17/20, high)	color	top	VIS
top.vis.lab.b.skewness	asymmetry of the b* (blue to yellow) values of the plant color histogram, calculated within the l*a*b*-color space	color	top	VIS

Table S1 (continued)

volume.fluo.iap	estimated digital volume (px^3 or mm^3)	geometric	both	FLUO

† the category of a defined phenotypic trait belonged to

§ a trait is defined based on images from side view, top view or both

¶ the type of image data used to define the trait; VIS, visible-light; FLUO, fluorescence; NIR, near-infrared. This table was taken from Chen et al. (2014b).

Table S2: A worldwide collection of maize plants selected from from IPK Genebank.

Accession No†	**Accession Name**	**Scientific Name**	**Country of Origin**
ZEA 3	Gelber Badischer Landmais	Zea mays L. convar. mays var. mays	Germany
ZEA 16	Breslau II	Zea mays L. convar. mays var. rubra Bonaf.	Poland
ZEA 249	Lester Phister	Zea mays L. convar. dentiformis Körn. var. flavorubra Körn.	Romania
ZEA 323	Risovaja 645	Zea mays L. convar. microsperma Körn. var. oryzoides Körn.	Russia
ZEA 333	Cukrova Cervena	Zea mays L. convar. saccharata Körn.	*NA*
ZEA 384	Aromatnaja	Zea mays L. convar. saccharata Körn. var. flavodulcis Körn.	Russia
ZEA 472	*NA*	Zea mays L. convar. aorista Greb.	Greece
ZEA 668	*NA*	Zea mays L. convar. dentiformis Körn. var. xantodon Alef.	Macedonia
ZEA 701	*NA*	Zea mays L. convar. mays var. mays	Hungary
ZEA 710	*NA*	Zea mays L. convar. dentiformis Körn. var. xantodon Alef.	Czech Republic
ZEA 712	*NA*	Zea mays L. convar. dentiformis Körn. var. xantodon Alef.	Czech Republic
ZEA 719	*NA*	Zea mays L. convar. mays var. mays	Slovakia
ZEA 852	Col/Chung Nam Jusan/2620	Zea mays L.	North Korea
ZEA 1008	*NA*	Zea mays L. convar. mays var. mays	Libya
ZEA 1066	*NA*	Zea mays L. convar. dentiformis Körn. var. flavorubra Körn.	North Korea
ZEA 1129	*NA*	Zea mays L. convar. mays var. alba Alef.	Austria
ZEA 1181	Rainbow Amerindian	Zea mays L. convar. microsperma Körn.	*NA*

Table S2 (continued)

ZEA 3240	Weiße Zarin	Zea mays L. convar. dentiformis Körn. var. leucodon Alef.	Georgia
ZEA 3303	*NA*	Zea mays L.	Italy
ZEA 3327	*NA*	Zea mays L.	Albania
ZEA 3338	Lazuti	Zea mays L. convar. aorista Greb.	Georgia
ZEA 3348	*NA*	Zea mays L. convar. mays var. aurantiaca Kuleshov & Kozhukhov	Romania
ZEA 3361	*NA*	Zea mays L. convar. aorista Greb.	Croatia
ZEA 3425	Strenzfelder	Zea mays L. convar. mays var. mays	Germany
ZEA 3426	Schindelmeiser	Zea mays L. convar. mays var. mays	Germany
ZEA 3434	Brona	Zea mays L. convar. dentiformis Körn. var. flavorubra Körn.	Spain
ZEA 3455	*NA*	Zea mays L	Cuba
ZEA 3528	*NA*	Zea mays L. convar. mays var. aurantiaca Kuleshov & Kozhukhov	North Korea
ZEA 3548	Taos Pueblo Blue	Zea mays L. convar. mays var. caesia Alef.	USA
ZEA 3554	*NA*	Zea mays L. convar. dentiformis Körn.	Albania
ZEA 3555	Miser I bardeti	Zea mays L. convar. mays var. alba Alef.	Albania
ZEA 3572	*NA*	Zea mays L. convar. aorista Greb.	Italy
ZEA 3606	Inrafrueh	Zea mays L. convar. dentiformis Körn. var. flavorubra Körn.	*NA*
ZEA 3651	Meirenhuang	Zea sp.	China
Athletico[§]	*NA*	Zea sp.	Germany
Fernandez[§]	*NA*	Zea sp.	Germany

[†] Detailed information is accessible through the GBIS/I Genebank Information System (http://gbis.ipk-gatersleben.de/GBIS_I/) in IPK Gatersleben. *NA*: not available.

[§] Two high performance (HP) lines (with German origin) from KWS Company (http://www.kws.com/)

Appendix C

Online Resources

Additional online data sets related to this thesis are availabe from the following link https://github.com/htpmod/HTPdata:

- ✔ **Online Data Set 1.** High-throughput phenotyping data in barley.
- ✔ **Online Data Set 2.** Growth modeling of control plants in barley.
- ✔ **Online Data Set 3.** Growth modeling of stressed plants in barley.
- ✔ **Online Data Set 4.** Image-derived data set used for growth modeling in maize.
- ✔ **Online Data Set 5.** Growth modeling of individual plants in maize.

Additional information related to this thesis is availabe from our website:

- ✔ The IAP software can be downloaded from this link: http://iapg2p.sourceforge.net/.
- ✔ Relevant R code and corresponding document are provided at the website of https://github.com/htpmod/HTPmod.

Appendix D

Curriculum Vitae

Dijun Chen

❑ **Date of birth: 27 June, 1983**

❑ **Nationality: Chinese**

❑ **Website: http://goo.gl/G4cKQx**

☎ **+49-(331)-977-6147**

✉ **chendijun2012@gmail.com**

"Live simply with great ambition; fare serenely toward high goal."

University Education / Research Experience

2016.10-Present	Wissenschaftlicher Mitarbeiter, Plant Cell and Molecular Biology, Humboldt University
	- Supervisor: Prof. Dr. Kerstin Kaufmann
2015.05-2016.09	Postdoc / Bioinformatician, Plant Developmental Biology, Potsdam University
	- Supervisor: Dr. Kerstin Kaufmann
2012.07-2015.04	PhD student, Bioinformatics, Department of Molecular Genetics, IPK
	- Supervisors: Dr. Christian Klukas & Prof. Dr. Thomas Altmann
2010.09-2012.06	Master student, Bioinformatics, Huazhong Agricultural University
	- Supervisors: Prof. Ling-ling Chen & Prof. Ming Chen
	- Received a LEIAO Scholarship Award; Graduated in Bioinformatics with an Outstanding Graduation Thesis
2008.08-2010.08	Research assistant, Bioinformatics, Zhejiang University
	- Supervisor: Prof. Ming Chen
2003.09-2008.06	Undergraduate student, Bioinformatics, Harbin Medical University

- Received Best and First Scholarship Awards; Graduated with an Outstanding Graduate (in Heilongjiang Province)

Research Interests

❑ I have broad research interests, including various topics related to bioinformatics or computational biology (either in plants or in human):

- ✔ Functional & integrative genomics
- ✔ Comparative genomics / epigenomics
- ✔ Transcriptomics and non-coding RNAs (incl. miRNAs and lincRNAs)
- ✔ Gene regulation networks
- ✔ High-throughput phenotyping/image (HTP) data analysis
- ✔ Genome-wide association studies (GWAS)
- ✔ Large data or big data exploration and visualization
- ✔ Biological database construction, data mining, integration and standardization
- ✔ Developing methodologies on NGS (RNA-seq, ChIP-seq, DNase-seq, Hi-C and ChIA-PET) data analysis

Publications

2016

✓ Guo Z, **Chen D***, Alqudah A, Röder M, Ganal M, Schnurbusch T*. Genome-wide association analyses of 54 traits identified multiple loci for the determination of floret fertility in wheat. ***New Phytologist*** 2016 Dec 5. doi:10.1111/nph.14342.

✓ Ćwiek-Kupczyńska H, Altmann T, Arend D, Arnaud E, **Chen D**, Cornut G, Fiorani F, Frohmberg W, Junker A, Klukas C, Lange M, Mazurek C, Nafissi A, Neveu P, van Oeveren J, Pommier C, Poorter H, Rocca-Serra P, Sansone S-A, Scholz U, van Schriek M, Seren , Usadel B, Weise S, Kersey P, Krajewski P*. Measures for interoperability of phenotypic data: minimum information requirements and formatting. ***Plant Methods*** 2016;12:44. doi:10.1186/s13007-016-0144-4.

✓ Yan W, **Chen D**, Kaufmann K*. Molecular mechanisms of floral organ specification by MADS domain proteins. ***Curr. Opin. Plant Biol.*** 2016;29. 154162.

2015

✓ Rahaman MM, **Chen D**, Gillani Z, Klukas C, Chen M*. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. ***Front. Plant Sci.*** 2015;6. doi:10.3389/fpls.2015.00619.

†: equal contribution; *: corresponding

✓ Neumann K*, Klukas C, Friedel S, Rischbeck P, **Chen D**, Entzian A, Stein N, Graner A, Kilian B*. Dissecting spatio-temporal biomass accumulation in barley under different water regimes using high-throughput image analysis. ***Plant Cell Environ.*** 2015;38(10). doi:10.1111/pce.12516.

✓ Guo Z, **Chen D**, Schnurbusch T*. Variance components, heritability and correlation analysis of anther and ovary size during the floral development of bread wheat. ***J. Exp. Bot.*** 2015;66(11):3099-3111. doi: 10.1093/jxb/erv117.

✓ Yuan C, Wang J, Andrew PH, Meng X, **Chen D**, Chen M*. Genome-wide View of Natural Antisense Transcripts in *Arabidopsis thaliana*. ***DNA Res.*** 2015;22(3). doi:10.1093/dnares/dsv008.

✓ Krajewski P*, **Chen D**, Cwiek H, van Dijk ADJ, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B, and Weise S. Towards recommendations for metadata and data handling in plant phenotypic experiments. ***J. Exp. Bot.*** 2015;66(18):5417-5427.

2014

✓ **Chen D***, Neumann K, Friedel S, Kilian B, Chen M, Altmann T, Klukas C*. Dissecting the phenotypic components of plant growth and drought responses based on high-throughput image analysis. ***Plant Cell***. 2014;26(12):4636-55.

✓ **Chen D**, Fu LY, Zhang Z, Li G, Zhang H, Jiang L, Harrison AP, Shanahan HP, Klukas C, Zhang HY, Ruan Y*, Chen LL*, Chen M*. Dissecting the chromatin interactome of microRNA genes. ***Nucleic Acids Res.*** 2014;42(5):3028-43.

✓ Klukas C*, **Chen D**, and Pape JM. Integrated Analysis Platform: An Open-Source Information System for High-Throughput Plant Phenotyping. ***Plant Physiol.*** 2014;165(2):506-518.

✓ Ding YD, Chang JW, Guo J, **Chen D**, Li S, Xu Q, Deng XX, Cheng YJ, Chen LL*. Prediction and functional analysis of the sweet orange protein-protein interaction network. ***BMC Plant Biol.*** 2014;14(1):213; doi:10.1186/s12870-014-0213-7.

✓ Liu Y, Wang L, **Chen D**, Wu X, Huang D, Chen LL, Li L, Deng XX, Xu Q*. Genome-wide comparison of microRNAs and their targeted transcripts among leaf, flower and fruit of sweet orange. ***BMC Genomics***. 2014;15(1):695; doi:10.1186/1471-2164-15-695.

✓ Wang J†, **Chen D**†, Lei Y†, Chang JW, Hao BH, Xing F, Li S, Xu Q, Deng XX, Chen LL*. *Citrus sinensis* annotation project (CAP): a comprehensive database for sweet orange genome. ***PLoS ONE***. 2014;9(1):e87723.

2013

✓ Xu Q, Chen LL, Ruan X, **Chen D**, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP, Chen J, Gao S, Xing F, Lan H, Chang JW, Ge X, Lei Y, Hu Q, Miao Y, Wang L, Xiao S, Biswas MK, Zeng W, Guo F, Cao H, Yang X, Xu XW, Cheng YJ, Xu J, Liu JH, Luo OJ, Tang Z, Guo WW, Kuang H, Zhang HY, Roose ML, Nagarajan N, Deng XX*, Ruan Y*. The draft genome of sweet orange (*Citrus sinensis*). ***Nat. Genet.*** 2013;45(1):59-66.

✓ Lu X†, **Chen D**†, Shu D, Zhang Z, Wang W, Klukas C, Chen LL, Fan Y, Chen M*, Zhang C*. The differential transcription network between embryo and endosperm in the early developing maize seed. *Plant Physiol.* 2013;162(1):440-55.

✓ Zou W†, **Chen D**†, Xiong M†, Zhu J, Lin X, Wang L, Zhang J, Chen LL, Zhang H, Chen H, Chen M, Jin M*. Insights into the increasing virulence of the swine-origin pandemic H1N1/2009 influenza virus. ***Sci. Rep.*** 2013;3:1601. doi: 10.1038/srep01601.

2012

✓ **Chen D**, Yuan C, Zhang J, Zhang Z, Bai L, Meng Y, Chen LL*, Chen M*. PlantNATsDB: a comprehensive database of plant natural antisense transcripts. ***Nucleic Acids Res.*** 2012; 40(1):D1187-93.

✓ He Z†, **Chen D**†, Wang K, Chen M*. miRPreditor: a Novel MiRNA Target Predictor Based on SVM with Feature Analysis. ***eJBio***, 2012;8(4):79-89.

✓ Wu Y†, **Chen D**†, He H†, Chen DS, Chen LL, Chen H, Liu Z*. Pseudorabies Virus Infected Porcine Epithelial 1 Cell line Generates a Diverse Set of Host MicroRNAs and a Special Cluster of Viral MicroRNAs. ***PLoS ONE***. 2012; 7(1):e30988.

2011

✓ **Chen D**, Meng Y, Yuan C, Bai L, Huang D, Lv S, Wu P, Chen LL, Chen M*. Plant siRNAs from introns mediate DNA methylation of host genes. ***RNA***. 2011;17(6):1012-24.

✓ Huang D, Huang Y, Bai Y, **Chen D**, Hofestädt R, Klukas C, Chen M*. MyBioNet: interactively visualize, edit and merge biological networks on the Web. ***Bioinformatics***. 2011;27(23):3321-3322.

✓ Meng Y, Gou L, **Chen D**, Mao C, Jin Y, Wu P, Chen M*. PmiRKB: a plant microRNA knowledge base. ***Nucleic Acids Res.*** 2011;39(Database issue):D181-7.

2010

✓ **Chen D**, Meng Y, Ma X, Mao C, Bai Y, Cao J, Gu H, Wu P, Chen M*. Small RNAs in angiosperms: sequence characteristics, distribution and generation. ***Bioinformatics***. 2010;26(11):1391-4.

✓ Chen M*, Meng Y, Gu H, **Chen D**. Functional characterization of plant small RNAs based on next-generation sequencing data. ***Comput. Biol. Chem.*** 2010;34(5-6):308-12.

✓ Chen M*, Meng Y, Mao C, **Chen D**, Wu P*. Methodological framework for functional characterization of plant microRNAs. ***J. Exp. Bot.*** 2010;61(9):2271-80.

✓ **Chen D**, Zhang F, Yuan C, Lu J, Li X, Chen M*. A web-based platform for rice microarray annotation and data analysis. ***Sci. China Life Sci.*** 2010;53(12):1467-73.

✓ Meng Y, Ma X, **Chen D**, Wu P*, Chen M*. MicroRNA-mediated signaling involved in plant root development. ***Biochem. Biophys. Res. Commun.*** 2010;393(3):345-9.

✓ Meng Y, Gou L, **Chen D**, Wu P, Chen M*. High-throughput degradome sequencing can be used to gain insights into microRNA precursor metabolism. ***J. Exp. Bot.*** 2010;61(14):3833-7.

✓ Meng Y, **Chen D**, Ma X, Mao C, Cao J, Wu P, Chen M*. Mechanisms of microRNA-mediated auxin signaling inferred from the rice mutant osaxr. ***Plant Signal Behav.*** 2010;5(3):252-4.

✓ Meng Y, **Chen D**, Jin Y, Mao C, Wu P, Chen M*. RNA editing of nuclear transcripts in *Arabidopsis thaliana*. ***BMC Genomics***. 2010;11 Suppl 4:S12.

✓ Shi Q†, Meng Y†, **Chen D**†, He F, Gu H, Wu P, Chen M*. OsCAS: a comprehensive web-based annotation platform for rice microarray data. ***BioChip J.*** 2010;4(1), 9-15.

2009

✓ Meng Y, Huang F, Shi Q, Cao J, **Chen D**, Zhang J, Ni J, Wu P*, Chen M*. Genome-wide survey of rice microRNAs and microRNA-target pairs in the root of a novel auxin-resistant mutant. ***Planta***. 2009;230(5):883-98.

Book Chapters

2014

✓ **Chen D**, Chen M, Altmann T, Klukas C* (2014). Bridging Genomics and Phenomics. *In* Approaches in Integrative Bioinformatics (pp. 299-333). Springer Berlin Heidelberg.

Conference Posters

2013

High-throughput plant phenomic data analysis using the Integrated Analysis Platform (I-AP). Chen D, Neumann K, Entzian A, Kilian B, Friedel S and Klukas C. *International Symposium on Integrative Bioinformatics* (IB2013; http://www.imbio.de/ib2013/); IPK Gatersleben, Germany

High-throughput Plant Phenomic Data Analysis Using the Integrated Analysis Platform (I-AP). Chen D, Neumann K, Kilian B, Friedel S, Altmann T and Klukas C. *The 14th International Conference on Systems Biology* (ICSB; http://www.icsb2013.dk/); Copenhagen, Denmark

Databases & Websites

2012-2013

✓ The website of **Phenomics.cn**: http://www.phenomics.cn/ [Apache & PHP]

2011-2012	✓ The website of **CAP** (*Citrus sinensis* Annotation Project): http://citrus.hzau.edu.cn/orange/ [Wang *et al.* *PLoS ONE* 9.1 (2014): e87723] [Apache & PHP & MySQL]
2010-2011	✓ The database of **PlantNATsDB** (Plant Natural Antisense Transcripts DataBase): http://bis.zju.edu.cn/pnatdb/ [Chen *et al.* *Nucleic acids research* 40.D1 (2012): D1187-D1193] [Tomcat & Java/JSP & MySQL]
2010-2011	✓ The website of **miRPredictor**: http://bis.zju.edu.cn/mirpredictor/ [He *et al.* *Electronic Journal of Biology* 8.4 (2012): 79-89] [Tomcat & Java/JSP]
2009-2010	✓ The database of **OsCAS** (*Oryza sativa* Chips Annotation System): http://bis.zju.edu.cn/oscas/ [Shi *et al.* *BioChip Journal* 4.1 (2010): 9-15] [Apache & Perl & MySQL]

Presentations/talks

✔ **When 3C meets 3D: an emerging theme of gene regulation.** *IPK PhD seminar.* Gatersleben, Germany (2014)

✔ **Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis.** *IPK PhD seminar.* Gatersleben, Germany (2013)

✔ **The chromatin interactome of human miRNAome.** *University of Essex.* Colchester, Essex, UK (2012)

✔ **The big world of small RNA molecules.** *International Symposium on Integrative Bioinformatics.* Hangzhou, China (2012)

Travel History

✔ **06.05.2012 - 17.05.2012, UK:** visited the *University of Essex* and *University of London*; supported by Henry Lester Trust

✔ **07.12.2012 - 15.12.2012, Australia:** visited the *University of Adelaide*; supported by DAAD

Last updated: June 8, 2017