Bioinformatic aspects of breeding polyploid crops

Fabian Grandke

München 2016

Bioinformatic aspects of breeding polyploid crops

Fabian Grandke

Dissertation zur Erlangung
des Doktorgrades der Naturwissenschaften
der Fakultät für Biologie
der Ludwig–Maximilians–Universität
München

vorgelegt von
Fabian Grandke

München, den 14.12.2016

Erstgutachter:	Prof. Dr. Dirk Metzler
Zweitgutachter:	Prof. Dr. John Parsch
Tag der Abgabe:	14.12.2016
Tag der mündlichen Prüfung:	09.03.2017

ERKLÄRUNG

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Prof. Dr. Dirk Metzler betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich nicht anderweitig einer Doktorprüfung unterzogen habe.

EIDESSTATTLICHE VERSICHERUNG

Ich versichere ferner hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist.

München, 14.12.2016

Fabian Grandke

DECLARATION OF CO-AUTHOR CONTRIBUTIONS

The study in Chapter 1 (Grandke et al., 2014, appeared in *Journal of Agricultural Science and Technology B*) was designed by Andrzej Czech and myself. I selected the tools and analyzed their limitations with help from Soumya Ranganathan. I wrote the manuscript and incorporated feedback from Dirk Metzler, Jorn R. de Haan, Andrzej Czech and Soumya Ranganathan.

The method in Chapter 2 (Grandke et al., 2016a, appeared in *BMC Genomics*) was designed by Jorn R. de Haan, Henri C. M. Heuven, Priyanka Singh and myself. Jorn R. de Haan developed the idea of using raw genotypes instead of genotype classes. I performed data preprocessing, linear regression analysis and conducted the simulation study. Priyanka Singh calculated the DEBVs and performed the association analysis with PLSR and bayz. Dirk Metzler designed the simulation study and provided feedback on the manuscript. I wrote the manuscript with input from Dirk Metzler, Henri C. M. Heuven and Priyanka Singh.

The method in Chapter 3 (Grandke et al., 2016c, in press at *BMC Bioinformatics*) was designed by Dirk Metzler, Jorn R. de Haan, Nikkie van Bers, Soumya Ranganathan and myself. I developed and implemented the method with input from Dirk Metzler and Soumya Ranganathan. Nikkie van Bers pointed out key aspects about linkage mapping. I performed the simulation study with input from Dirk Metzler. Dirk Metzler advised on validation related aspects and reviewed the manuscript. I wrote the manuscript with input from Nikkie van Bers.

The application in Chapter 4 (Grandke et al., 2016b, in press at *Bioinformatics*) was designed by Birgit Samans and myself. Birgit Samans drafted the basic workflow of CNV detection and performed a preliminary comparison of available methods for segmentation. I developed the final workflow, implemented the methods and designed the R-package. I wrote the manuscript with input from Birgit Samans and Rod Snowdon.

Prof. Dr. Dirk Metzler

Fabian Grandke

Inhaltsverzeichnis

Summary xi

Zusammenfassung xv

General Introduction 1

1 Bioinformatic Tools for Polyploid Crops
– *Journal of Agricultural Science and Technology B* (2014) 4, 593-601 19

2 Continuous Genotype Values for GWAS in Hexaploid Chrysanthemum
–*BMC Genomics* (2016) 17:672 21

3 PERGOLA: Fast and Deterministic Linkage Mapping of Polyploids
–*BMC Bioinformatics* in press 23

4 gsrc - an R package for genome structure rearrangement calling
–*Bioinformatics* in press 25

General Discussion 27

A Supplementary Files for Chapter 2 37

B Supplementary Files for Chapter 3 39

C Supplementary Files for Chapter 4 41

Abbreviations 43

Bibliography 45

List of Publications

1 Bioinformatic Tools for Polyploid Crops
– *Journal of Agricultural Science and Technology B* (2014) 4, 593-601 19
2 Continuous Genotype Values for GWAS in Hexaploid Chrysanthemum
–*BMC Genomics* (2016) 17:672 21
3 PERGOLA: Fast and Deterministic Linkage Mapping of Polyploids
–*BMC Bioinformatics* in press 23
4 gsrc - an R package for genome structure rearrangement calling
–*Bioinformatics* in press 25

Summary

Many important crops are polyploid (e.g. rapeseed, potato, wheat), which is the presence of more than two chromosome copies in one genome. Polyploidy is mainly found in flowering plants, but can also occur in animals and bacteria. The origins and numbers of the additional chromosome sets are diverse and remain a challenge in plant biology. Modern plant breeding requires detailed genetic information, which is unavailable for polyploids because standard methods fail to account for the additional chromosome copies. Therefore, breeding of polyploids is less successful than for diploids. Bioinformatic tools can overcome these limitations by either extending available methods or designing new ones.

The overarching questions of this dissertation are: What are the differences between diploids and polyploids from a bioinformatics point of view? Which currently available plant breeding methods cannot be applied to polyploids? What adaptations to bioinformatic methods are required to account for different ploidy types and levels?

In Chapter 1 (Grandke et al., 2014, appeared in *Journal of Agricultural Science and Technology B*) we describe, compare and discuss available bioinformatic tools for polyploid datasets. We focus on methods which have been developed specifically for polyploids. Our analysis shows that these tools address critical problems, which are unsolvable with existing methods for diploids. However, all tools in our analysis have limitations and cannot be applied to all polyploids, because they are either restricted to particular ploidy types or levels. The conclusion of Chapter 1 serves as motivation for the subsequent chapters: The available polyploid toolbox is incomplete and leaves many research questions unanswered. New methods are required to overcome these limitations and support research in polyploids.

In Chapter 2 (Grandke et al., 2016a, appeared in *BMC Genomics*) we address the problem of genotype calling in higher polyploids (ploidy level > 4) and its consequences for the downstream analysis. Genotype calling is a noise reduction step to extract biologically useful information from raw data (e.g. high-throughput microarrays or genotyping-by-sequencing (GBS)). Genotyping methods developed for diploids and tetraploids fail to call genotypes in higher polyploids, and

there is only one tool which overcomes this limitation, but its results are partially erroneous and misleading. We introduce a new method where we use raw data instead of genotypes calls. It enables us to perform a genome-wide association study (GWAS) with three phenotypic traits in a population of hexaploid chrysanthemum. We use three different regression methods to prevent biased results. A simulation study underpins our findings, and we can identify numerous candidate markers.

In Chapter 3 (Grandke et al., 2016c, in press at *BMC Bioinformatics*) we develop PERGOLA, a new method and publicly available R package for linkage mapping in polyploids. The algorithm uses a heuristic approach for calculating recombination frequencies and hierarchical clustering for linkage grouping. An improved version of optimal leaf ordering (OLO) orders markers remarkably fast. We introduce a new way to represent and compare linkage maps, which is based on dendrograms and supports statistical measures like cophenetic correlation and the Goodman-Kruskal index. We apply our method to simulated and real datasets of varying ploidy levels and show that it calculates correct linkage maps. We compare PERGOLA to available linkage mapping methods for diploids and demonstrate that it outperforms them computationally and provides more accurate maps.

In Chapter 4 (Grandke et al., 2016b, in press at *Bioinformatics*) we develop a new method to detect and visualize genome structure rearrangements in allopolyploids. Allopolyploid genomes consist of at least two subgenomes, which originate from different, but closely related species. The subgenomes are highly similar and lead to errors during meiosis. As a result, regions of one subgenome become substituted by parts of the other subgenome. Based on locus specific markers we developed a tool to find the corresponding deletion and duplication events which we combine with synteny information to find homeologous non-reciprocal translocations (HNRT). Besides the methodology we introduce a novel representation of the results. Our implementation is publicly available as R package.

In summary, we concluded the following to the questions mentioned above: The primary bioinformatics challenge of polyploids are the increased number of genotype classes, which are hardly distinguishable with available technologies and algorithms. We showed that usage of continuous genotype values is a good alternative and avoids genotype classification. Also, the concept of allopolyploid subgenomes originating from different species does not exist in diploids and requires new algorithms, like our method to detect genome structure rearrangements. The standard plant breeding methods of genotype calling, linkage mapping, and haplotype phasing are not readily applicable to polyploid crops. Some available methods for diploids can be adapted to accept more than three genotype classes. Others, like our linkage mapping method, need to be

created from scratch to account for the characteristics of polyploids.

Zusammenfassung

Viele wichtige Kulturpflanzen (z.B. Raps, Kartoffel, Weizen) sind polyploid, was die Präsenz von mehr als zwei Chromosomenkopien im Genom beschreibt. Polyploidie findet man häufig in Blütenpflanzen, aber auch in Tieren und Bakterien. Die Ursprünge und Anzahlen der zusätzlichen Chromosomenkopien sind vielfältig und stellen eine große Herausforderung für die Pflanzenbiologie dar, da sie maßgeschneiderte Analysemethoden erfordern. Moderne Pflanzenzüchtung benötigt detaillierte Informationen über die Genetik der Pflanzen, welche im Falle von Polyplodien nicht zur Verfügung stehen, da Standardmethoden die zusätzlichen Chromosomenkopien nicht berücksichtigen. Darum ist die Züchtung polyploider Pflanzen weniger erfolgreich als die Züchtung diploider Pflanzen. Mit Hilfe von bioinformatischen Anwendungen kann dies ausgeglichen werden, indem bestehende Methoden erweitert oder neue Methoden entwickelt werden.

Die übergreifenden Fragen dieser Dissertation sind: Welches sind die Unterschiede zwischen Diploiden und Polyploiden aus bioinformatischer Sicht? Welche Methoden der Pflanzenzucht können zur Zeit nicht auf polyploide Pflanzen angewendet werden? Welche Adaptionen bioinformatischer Methoden sind notwendig um verschiedene Ploidietypen und -level zu berücksichtigen?

In Kapitel 1 (Grandke et al., 2014, erschienen im *Journal of Agricultural Science and Technology B*) beschreiben, vergleichen und diskutieren wir aktuell verfügbare, bioinformatische Anwendungen für polyploide Datensätze. Wir konzentrieren uns dabei auf Methoden, welche speziell für Polyploide entwickelt wurden. Unsere Analyse zeigt, dass die Anwendungen wichtige Probleme angehen, welche mit den existierenden Methoden für Diploide nicht gelöst werden können. Alle analysierten Anwendungen sind entweder bezüglich der Ploidietypen oder -level beschränkt und können nicht auf alle Polyploiden angewendet werden. Die Zusammenfassung des ersten Kapitels ist gleichzeitig eine Motivation für die folgenden Kapitel: Die verfügbaren Anwendungen für Polyploide sind unvollständig und darum bleiben viele wissenschaftlichen Fragestellungen bisher unbeantwortet. Es bedarf neuer Methoden um diese Beschränkungen zu überwinden und die Erforschung von Polyploiden voranzutreiben.

Im zweiten Kapitel (Grandke et al., 2016a, erschienen in *BMC Genomics*) widmen wir uns

dem Problem der Genotypbestimmung bei Ploidieleveln > 4 und dessen Konsequenzen auf anschließende Analyseschritte. Die Genotypbestimmung dient der Reduktion von Hintergrundrauschen um biologisch relevante Informationen aus Rohdaten (z.B. Hochdurchsatz Microarrays oder Genotypisierung mittels Sequenzierung) zu extrahieren. Mit einer Ausnahme können Genotypisierungsprogramme, welche für diploide und tetraploide Organismen entwickelt wurden, nicht auf höhere Ploidielevel angewendet werden. Leider sind die Ergebnisse dieser Ausnahme teilweise fehlerhaft und können zu falschen Schlussfolgerungen verleiten. Wir stellen eine neue Methode vor, bei der Rohdaten die Genotypklassifikationen ersetzen. Dies erlaubt uns eine genomweite Assoziationsstudie von drei phänotypischen Merkmalen in einer hexaploiden Chrysanthemenpopulation durchzuführen. Um methodenseitigen Bias auszuschließen, verwenden wir drei verschiedene Regressionsmethoden und vergleichen die Ergebnisse, welche zahlreiche Kandidatenmarker enthalten. Abschließend untermauern wir unsere Resultate mittels einer Simulationsstudie, bei der wir das Experiment in silico nachstellen.

In Kapitel 3 (Grandke et al., 2016c, im Druck bei *BMC Bioinformatics*) entwickeln wir PERGOLA, eine neue Methode und R-Paket zur Erstellung von Kopplungskarten für Polyploide. Der Algorithmus basiert auf einem heuristischen Verfahren zur Berechnung von Rekombinationshäufigkeiten und Kopplungsgruppenberechnung durch hierarchisches Clustering. Wir erweitern die Methode der optimalen Blattordnung um die Ordnung von Markern zu beschleunigen. Wir führen eine neue Darstellung und Vergleichsmöglichkeit für Kopplungskarten ein, welche auf Dendrogrammen basiert und statistische Maße wie die kophänetische Korrelation und den Goodman-Kruskal-Index unterstützt. Wir beweisen sowohl mit simulierten als auch mit realen Daten verschiedener Ploidielevels, dass unsere Methode richtige Kopplungskarten berechnet. Wir vergleichen PERGOLA mit Programmen zur Berechnung von Kopplungskarten für diploide Organismen und zeigen, dass unsere Methode nicht nur schneller ist, sondern auch bessere Resultate erzeugt.

Im vierten Kapitel (Grandke et al., 2016b, im Druck bei *Bioinformatics*) entwickeln wir eine Methode zur Erkennung und Darstellung von Genomumstrukturierungen in Allopolyploiden, deren Genom sich aus zwei Untergenomen, welche von unterschiedlichen, aber verwandten Arten stammen, zusammensetzt. Durch die hohe Ähnlichkeit der Subgenome ist die Meiose fehleranfällig, was zum Austausch von Teilbereichen zwischen den Subgenomen führen kann. Wir haben eine Methode entwickelt, welche, basierend auf subgenomspezifischen Markern, gegensätzliche Deletionen und Duplikationen identifiziert und diese mit Syntenieinformationen abgleicht, um homöologe nicht-reziproke Translokationen zu finden. Diese werden in einer neuen Darstellungsform präsentiert. Die Implementierung unserer Methode ist in Form eines R-Pakets öffentlich

verfügbar.

Zusammengefasst haben wir folgende Antworten auf die anfänglich genannten Fragen erarbeitet: Die größte bioinformatische Herausforderung von Polyploiden besteht in der erhöhten Anzahl von Genotypklassen, welche mit bestehenden Methoden und Algorithmen schwer zu unterscheiden sind. Die Verwendung kontinuierlicher Genotypen ist eine gute Alternative zu Genotypklassen. Das Konzept von allopolyploiden Untergenomen existiert für diploide Organismen nicht und bedarf neuer Algorithmen, wie zum Beispiel unser Methode zur Erkennung von Genomumstrukturierungen. Standardanwendungen in der Pflanzenzüchtung wie Genotypbestimmung, Kopplungskartenberechnung und Haplotypisierung können nicht ohne weiteres auf polyploide Organismen angewendet werden. Einige verfügbare Methoden müssen lediglich erweitert werden, damit sie mit mehr als drei Genotypklassen funktionieren, andere müssen vollständig ersetzt werden um die Besonderheiten von Polyploiden zu berücksichtigen. Unsere Anwendung PERGOLA ist eine solche Neuentwicklung zur Berechnung von Kopplungskarten für Polyploide.

General Introduction

Overview of this dissertation

This dissertation is written in cumulative style and structured into six chapters as shown in Figure 1. In this chapter, I will introduce the basic concepts of this dissertation: polyploidy, plant breeding, and various computational methods. They will provide the reader with the knowledge required to understand the findings in Chapters 1 - 4, which I will discuss in the final chapter. Chapter 1 investigates the state-of-the-art of bioinformatic tools for polyploid crops and their limitations. It explains our motivation for the subsequent chapters of this dissertation. Chapters 2 to 4 contain the main content of this dissertation in the form of peer-reviewed publications. They address the different research questions of this dissertation and overcome limitations that we detected in Chapter 1. The final chapter is a comprehensive, detailed discussion of the previous chapters. It links the individual projects together and thus, provides answers to the central research questions. Furthermore, I look beyond the context of plant breeding and provide proposals for future studies.

Polyploidy

Polyploidy is the presence of more than two chromosome copies in a genome. It is abundant in flowering plants and has been observed in animals and bacteria, as well (Song et al., 2012). Polyploidy does not include partial genome copy aberrations (e.g. trisomy 21 in humans), which are referred to as aneuploidy. Polyploid genomes form through various ways as shown in Figure 2 and differ in ploidy type and level. In nature, polyploidy is not a steady state, but rather an evolutionary snapshot and intermediate condition after hybridization or genome duplication events (Doyle et al., 2016). In contrast plant breeders often induce polyploidy into diploids to obtain desirable characteristics like seedless crops and higher yield (Sattler et al., 2016).

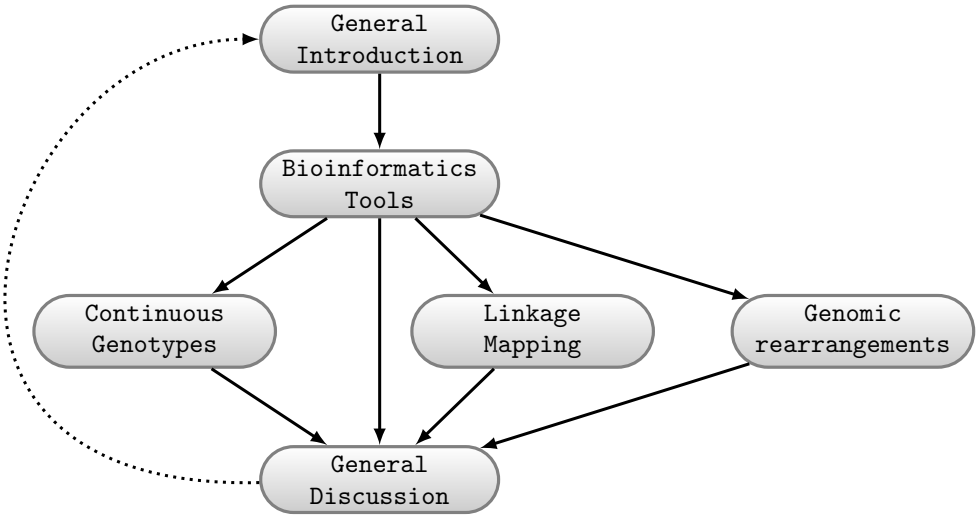

Figure 1: The structure of this dissertation: The general introduction explains basic terms, concepts, and methods that are required to understand this dissertation and raises its central research questions. Bioinformatic Tools (Chapter 1) provides an overview of available bioinformatic methods and their limitations. Continuous Genotypes (Chapter 2) proposes a new solution to the polyploid genotype calling problem for genome-wide association studies, which avoid the shortcomings of existing methods. Linkage Mapping (Chapter 3) describes a new fast and deterministic method for linkage mapping in polyploids. Genomic rearrangements (Chapter 4) introduces a novel application to detect and visualize genomic rearrangements in allopolyploids. The general discussion links the chapters back to the initial research questions and places the findings of this dissertation into a broader context.

Forms of polyploidy

There are two main forms of polyploidy: auto- and allopolyploidy. They describe how additional chromosome copies were introduced into the genomes of formerly diploid ancestors. Combinations of both forms are possible if species underwent multiple polyploidization events.

Autopolyploids originate from diploid gametes of the same species. Usually, diploid organisms produce haploid gametes, which then merge with another gamete to build a new diploid zygote. Diploid gametes develop either through errors in meiosis of diploids or from polyploid organisms. When diploid gametes fuse with haploid gametes they produce triploid zygotes, which are infertile in most species. When two diploid gametes of a species fuse, they build tetraploid zygotes. This variant is called autotetraploid because both gametes originate from the same species (compare left path in Figure 2). Tetraploid zygotes are usually stable and fertile. Potato is an autotetraploid model species of high economic value, and its genome has been well studied. Its close genetic relationship with tomato further contributes to its genomic interest. In Chapter 4 we calculate a

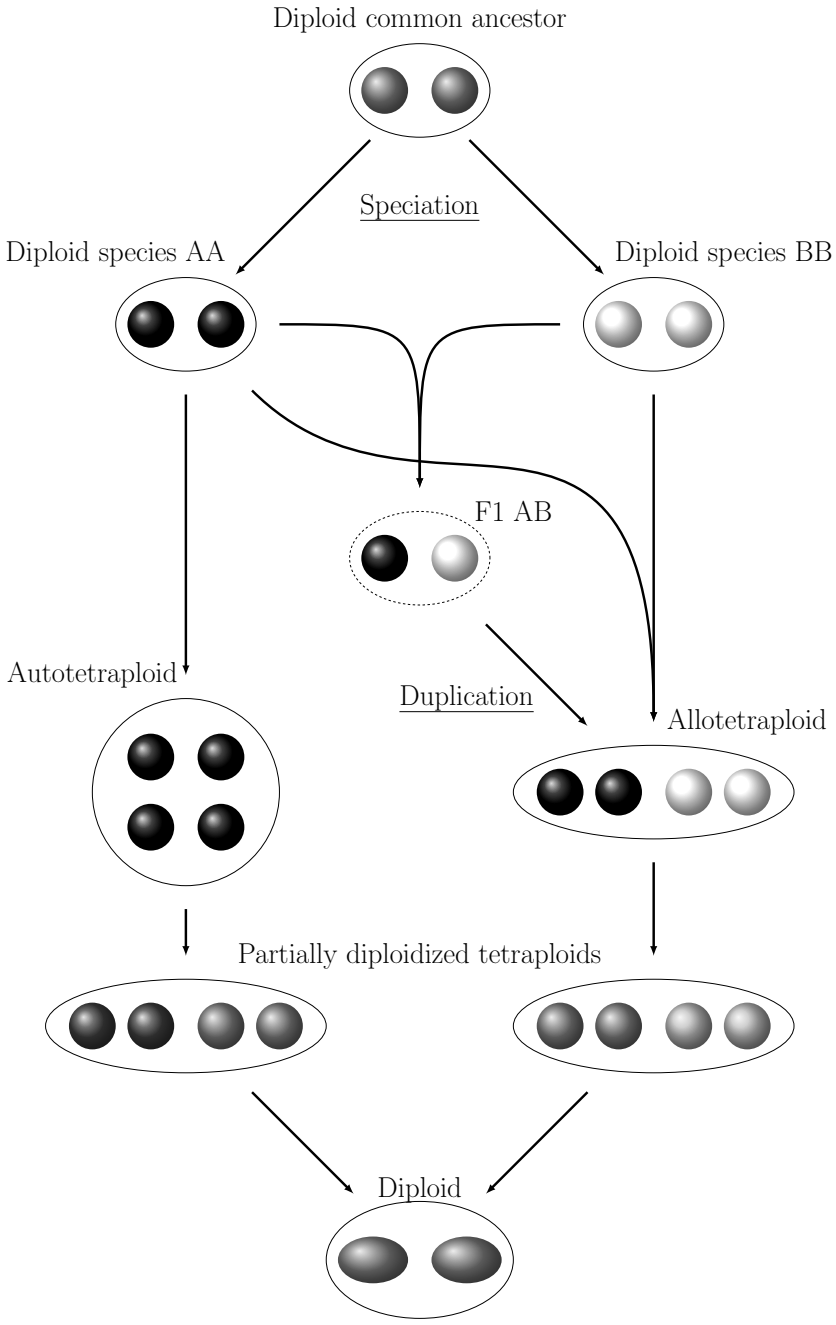

Figure 2: Origins of polyploidy: White ellipses represent different forms of ploidy and greyscaled ellipses indicate genomes within them (the increased ones at the bottom imply genome growth). Two diploid progenitors descended from a common diploid ancestor and formed through speciation. Left path: Autotetraploidy is formed through genome duplication and later returns to a diploid state. Center path: Two diploid species hybridize and form a diploid offspring, whose genome duplicated into an allotetraploid. Right path: Two diploid species hybridize and form a tetraploid offspring, which diploidizes in the subsequent steps. This figure has been adapted from Comai (2005).

linkage map for autotetraploid potato.

In contrast to autotetraploids, allopolyploids derive from gametes of different species. Either a hybridization event took place, and the hybridized diploid genome duplicated, or a diploid gamete of one species fuses with the diploid gamete of another species (compare center and right paths Figure 2). Rapeseed (*Brassica napus L.*) is an allotetraploid model organism. It is the most important oil crop in Europe and the second most important energy crop world-wide (soybean is first). Its economic importance led to intensive research and many publicly available genetic resources. The genome of rapeseed consists of two subgenomes A and C, derived from *Brassica rapa* and *Brassica oleracea*, respectively (compare Figure 3). In Chapter 4 we investigate genomic rearrangements in a *Brassica napus*.

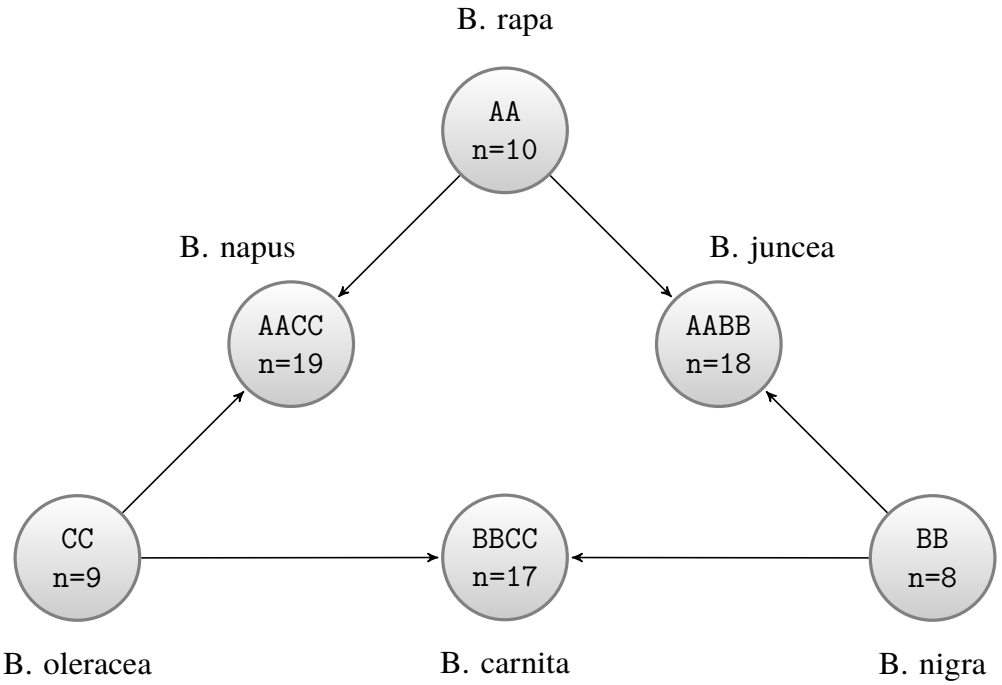

Figure 3: Triangle of Wu, a schematic overview of origins and relations of various *Brassica* (*B.*) species: Each circle represents a species, capital letters indicate (sub-)genomes and numbers show haploid chromosome counts. *B. nigra*, *B. oleracea*, and *B. rapa* are diploid, with 8, 9 and 10 chromosomes, respectively. *B.juncea*, *B. carnita*, and *B. napus* are allotetraploid and arose from spontaneous interspecific hybridization between their respective two diploid progenitors (indicated by arrows). The chromosome count of the diploids is summed up in the tetraploids. This figure has been adapted from Nagaharu (1935).

Synteny is defined as two similar blocks of genes on two sets of chromosomes of different subgenomes. In our example of rapeseed shared genes are mapped to subgenomes A and C and

reveal a conserved synteny structure. The visualizations in Vignette - Synteny Block Calculation in Appendix C show large blocks of synteny (e.g. chromosomes A01 and C01), but also syntenic regions outside the main blocks, which appear as shadows and indicate non-collinear synteny. There are three causes for these data points beyond the general synteny structure . First, the genomes underwent hexaploid stages in the past (Cheng et al., 2013). Hence, some parts of the genomes are highly similar and result in two shadowed regions in other chromosomes(e.g. A01 and A02 show shaded copies of A10 for the synteny region of C09). The orientation can switch due to genome structural rearrangements. The second cause for imperfect synteny blocks is mapping mistakes. Gene positions are obtained by mapping their DNA sequence onto a reference genome sequence, which can lead to multiple hits and only one of them is kept. Also, the reference genome sequence is erroneous and does not reflect the genetic reality of the *Brassica napus L.* genome. The third cause for noisy synteny are mutations, where individual genes translocate into different chromosomes or chromosome positions.

Current and former polyploids can be categorized based on the time of their polyploidization event(s). Polyploids derived by ancient genome duplications are paleopolyploids, while more recent polyploids are mesopolyploids (e.g. *Brassica napus* is a mesohexaploid). If diploidization completed in a species, but there is evidence for ancient polyploidy the term paleopolyploid is still valid (e.g. *Saccharomyces cerevisiae*).

Ploidy levels

Polyploidy is defined as copies of full haploid chromosome sets larger than two. While genome duplication and unreduced gametes can, in principle, lead to any number of chromosome set, ploidy levels are not distributed uniformly. There is a strong bias towards even numbers of ploidy, and four is the most common ploidy level. Even polyploids produce balanced gametes with full chromosome copies and are stable and fertile. Uneven polyploids cannot build bivalents (where homologous chromosomes pair up during meiosis), and the gametes are infertile in many cases. The higher a ploidy level, the less common it is. Tetraploids are the lowest even polyploids. The development of a tetraploid does not require many steps (e.g. genome duplications) and explains their abundance. In contrast, dodecaploids ($12x$) require several events of genome multiplication or hybridization with other polyploids. However, they exist and are stable (e.g. *Celosia argentea*).

Additional chromosome copies can be beneficial because the redundancy of genetic material leads to increased tolerance towards mutations. These may result in higher fitness (e.g. neofunctionalization) and can be an advantage over diploid relatives. Allopolyploids maintain the same level of heterozygosity because intergenomic recombination is restricted. On an evolutionary

scale, however, polyploidy is disadvantageous. Maintenance of redundant genetic information is inefficient and leads to numerous problems in cell architecture, meiosis, and mitosis. In the long term natural polyploid genomes return to a diploid stage. Duplicated loci differentiate (e.g. sub-functionalization), and eventually, subgenomes become incompatible. At that stage, the organism has become diploid. Auto- and allopolyploidy describe two extreme cases of polyploidy shortly after they developed. During the process of diploidization, these two classifications become less accurate. The transition from a polyploid to a diploid takes many generations and includes stages in which some chromosomes are still compatible, and others are not. These partially polyploid organisms are intermediates that are neither auto- or allopolyploids nor diploids.

Plant breeding

In the past individuals with desirable phenotypic traits were selected and used as progenitors for the next generation. In contrast, modern plant breeding is based on Darwin's and Mendel's discoveries about evolution and inheritance (Borlaug, 1983). Knowledge of genetic laws switched the parent-oriented to an offspring oriented breeding scheme. This development reached its preliminary peak with the *Green Revolution* in the 1960's. It caused development of many new varieties, increased food production in developing countries and was a huge success against the global hunger problem (Hazell, 2009).

Linkage mapping

Linkage mapping creates a genetic map that reveals the relation between markers in a population (see section Population types for details). In contrast to physical maps, which describe the distance between markers in base pairs (bp), genetic maps show how many recombination events occur in centiMorgan (cM). One centiMorgan represents one recombination event per 100 individuals. Two markers can be close on a genetic map and more distant in a physical map and vice versa. Linkage maps can be created with any markers that allow tracking recombination within a population. Ideally, markers are distributed evenly and densely over the whole genome. Figure 4 shows the individual steps of linkage map creation.

Recombination frequency is a pairwise measurement between all markers. It shows how many recombinations events happened between two markers in a population and can be seen as a distance measure. The interpretability of recombination frequency depends on the sample size within the population - the larger, the better. The unit is centiMorgan, where one centiMorgan represents the

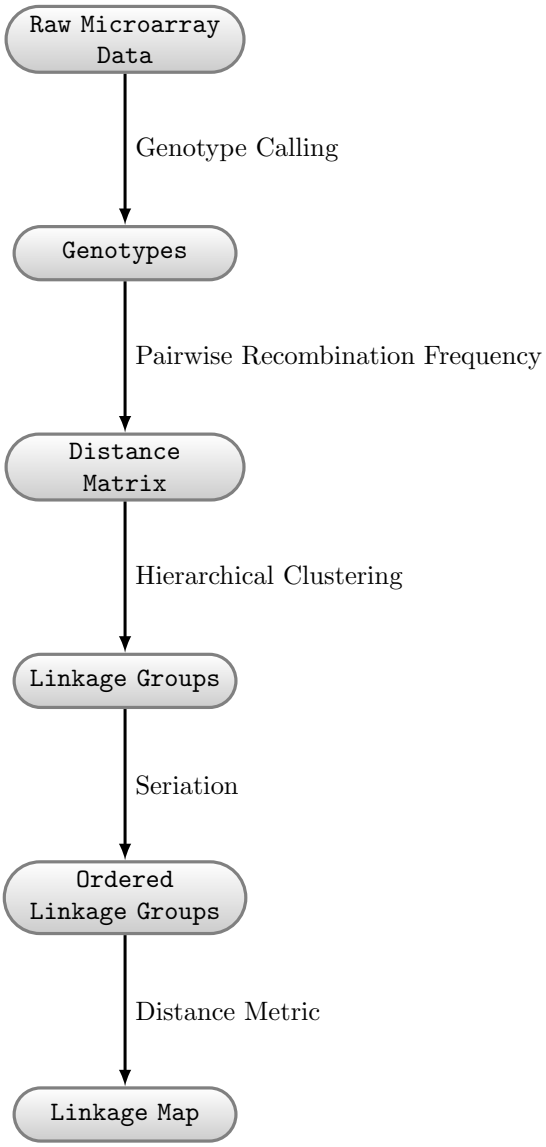

Figure 4: Schematic overview of linkage mapping: All samples of the population are genotyped with a high-throughput microarray. The raw microarray data is transformed into genotype calls, using genotype calling methods. Pairwise calculation of recombination frequencies results in a distance matrix for all markers (e.g. SNPs). Based on these distances, the markers are clustered into linkage groups. The markers within each group are ordered by seriation methods. The spaces between the markers are determined by distance metrics and result in the final linkage map.

distance between two loci where one percent recombination is detected.

Markers are grouped into linkage groups based on recombination frequencies. Ideally, each linkage group represents one (haploid) chromosome, but that cannot always be achieved. Single markers or small numbers of markers end up in their own linkage groups. These need to be filtered out based on a lower threshold for the number of markers in a group. Sometimes one chromosome is represented by two linkage groups because markers are not evenly distributed. If the marker density is particularly low around the centromere, the two groups represent the p- and q-arms of the chromosome.

Once linkage groups are defined, markers within the groups are ordered based on pairwise recombination frequencies. Markers with low recombination frequencies are placed adjacent to each other, while markers with high recombination frequencies are placed distantly. Computationally this is very complex as described in section Seriation on page 14.

The objective of marker spacing is to transform non-additive recombination frequencies r into additive map distances d (Huehn, 2011). Thereby we account for undetected multiple crossing overs between two markers. Interference describes the dependency of crossing-overs in adjacent regions on the same chromosome and can influence spacing in linkage maps. Several mapping functions are available which all have been implemented in Chapter 3.

Haldane Assumes recombinations to be Poisson distributed and excludes interference (Haldane, 1919).

$$d = -\frac{1}{2}\ln\left(1 - 2r\right)$$

Kosambi Assumes positive interference of $1 - 2r$, (Kosambi, 1943)

$$d = \frac{1}{4}\ln\left(\frac{1 + 2r}{1 - 2r}\right)$$

Carter Accounts for higher interference rates than Kosambi's mapping function (Carter et al., 1951)

$$d = \frac{1}{4}\left(\frac{1}{2}\left(\ln\left(1 + 2r\right) - \ln\left(1 - 2r\right)\right) + \arctan(2r)\right)$$

Mapping population types

Mapping populations are used to assess linkage between genetic markers and use this information to find quantitative trait loci. Knowledge about offspring individuals allows to trace back recombination events in the corresponding parental generation. Ideally, each parent underwent multiple

generations of selfing, and its genome is largely homozygous. Alternatively, artificially produced doubled-haploid (DH) lines are suitable for the purpose of mapping.

Segregating F2 populations are generated by selfing one F1 progenitor or randomly crossing multiple F1 progenitors of two parents. The F1 generation is heterozygous for most markers, and the F2 segregates over the full range of genotypes (e.g. AA:AB:BA:BB for a diploid F1 AB). Segregating populations are preferred for linkage mapping because it provides detailed insight into the recombination patterns.

Backcross (BC) populations are generated by backcrossing one F1 progenitor with one of the parents (or a genetically similar individual). The population segregates only over one-half of the possible genotypes. This is a limitation for linkage mapping approaches because recombination in the homozygous parent cannot be observed.

Recombinant inbred lines (RIL) are generated by selfing one F1 progenitor. The selfed F2 generation is than intermated for several generations. The last intermated generation is then selfed for multiple generations. The result is a population with fixed homozygous recombinations. Compared to segregating F2 and BC populations, RILs require much more time to be created and are financially more costly. However, RIL populations include much more recombination events resulting in a better linkage map (more markers and more precise distances).

Genotype calling

Genotype calling describes the determination of an individual's genotypic information through biotechnological techniques and bioinformatic algorithms (Rapley et al., 2004). Genotypic information is substantial for modern plant breeding, which is based on markers. Multiple technologies can be used to detect genotypes, the most popular ones are

- Sanger sequencing (SS) (Clevenger et al., 2015)
- Genotyping by sequencing (GBS) (Scheben et al., 2016; Goodwin et al., 2016)
- SNP microarrays (Kwong et al., 2016; Bianco et al., 2016)
- allele-specific PCR (Semagn et al., 2013; Patil et al., 2016)

The technologies vary in the expenditure of time, financial cost and output type and quality. Sanger sequencing is well established and produces reliable genotypic information. However, it is slow, expensive and impractical for high-throughput genotyping of large populations. Nevertheless, it is well established and reliable and can be used to validate other technologies (Clevenger et al.,

2015). GBS is faster and more economical than SS but is more erroneous. Its reliability depends on the sequencing coverage and library preparations. The output consists of sequence reads that need to be mapped to a reference genome. (Missing) variation at marker positions can be used to predict genotypes. SNP microarrays are the cheapest option for large populations and large numbers of known markers. They require high upfront costs to design the array, but once this is done, they can be reproduced and analyzed at low costs. Allele-specific PCR multiplies DNA at known SNPs or indels and attaches fluorescent labels for two different alleles. It is cheap and very flexible compared to microarrays. The measured SNPs can be modified easily, and researchers are not bound to outdated SNP selections of available microarray designs. The latter two technologies measure two alleles per SNP and provide signal strengths for each of them. Ratios between the two alleles are used to calculate genotype classes (Peiffer et al., 2006).

None of the described technologies results in perfect genotypes for all markers. Low coverage in GBS or technical problem in SNP arrays may lead to unexpected allele ratios. Genotype calling is used to reduce noise in the data and determine genotypes from the ratios between alleles. Usually, one marker at a time is assessed for all samples. Clustering methods are used to distinguish between possible genotype classes. For instance, AA, AB and BB for a diploid with two alleles A and B. The number of genotype classes increases with the ploidy level. In Chapter 2 we show a study where genotype calling is difficult due to the ploidy level of six (hexaploid) (Grandke et al., 2016a). Instead of relying on erroneous genotype calls it is advantageous to use raw genotype values in this case.

Genotype - Phenotype association

The overall aim in plant breeding is the improvement of crops by fixing and improving phenotypic traits. Knowledge about underlying genetics can support this procedure because the best phenotypes are not necessarily the best progenitors for a breeding program. Instead, the combination of less well performing individuals might produce better offspring. In the case of monogenic traits it is easy to select parents who have the desired phenotype, but for quantitative traits, this is not straightforward. Hence, breeders aim to detect quantitative trait loci (QTL), regions of the genome which are associated with a trait.

Marker-assisted selection (MAS) is an established method in plant breeding where polymorphic markers are used to select individuals from a bi-parental population. Markers, which are linked to phenotypic traits, provide information which is difficult to obtain otherwise (Collard et al., 2008). In the past, the performance of MAS was limited by the number of available markers (Heffner et al., 2009).

More recently, genome-wide association studies (GWAS) gained popularity and led to the discovery of new QTL. They were enabled by the availability of large sets of genetic variants (e.g. SNPs) which provided a more detailed insight into the genetic foundation of crops. The SNPs which are distributed over the genome are tested for association with phenotypic traits. Groups of highly associated SNPs reveal QTLs in the entire genome. In contrast to MAS, GWAS are not limited to bi-parental populations.

The latest trend in plant breeding is genomic selection (GS) (Heffner et al., 2009). In contrast to MAS and GWAS, GS does not aim to identify QTL for quantitative traits. Instead, GS uses marker data in combination with pedigree information and phenotypes to build a model that accurately predicts the performance of individuals. These predictions can be used to select progenitors in a breeding program. The disadvantage of this approach is that it only works for highly similar populations and application to other varieties might result in a lower accuracy of the model.

Computational and statistical methods

The chapters of this dissertation use several computational and statistical methods which are briefly described in this section.

Clustering

Clustering aims to identify structures in data based on similarity (Hastie et al., 2013). Each data point is assigned to a group (cluster), which is determined by the clustering method and parameters. Different methods can result in different clustering results. Ideally, each cluster can be matched to a real life condition or category. Clustering describes the general task, rather than a specific method of which there are many different ones (Jain et al., 1999). In this dissertation, I apply cluster analysis in four distinct fields and choose four different methods which I describe here:

Marker grouping is a key step in linkage mapping as described above and in Chapter 3. We use a hierarchical agglomerative clustering (HAC) with single-linkage fusion to define linkage groups (Hastie et al., 2013). In a linkage mapping context, clustering is based on pairwise recombination between all markers, which is represented by a distance matrix. The HAC algorithm works bottom-up, and each marker is treated as a singleton cluster in the beginning. Clusters are successively merged until all markers are in one cluster. The order of merging is decided by distance measures, which determine the two closest clusters. Common measures are:

single-linkage The shortest distance between any markers in each cluster

$$min\{d(a,b) : a \in A, b \in B\}$$

complete-linkage The longest distance between any markers in each cluster

$$max\{d(a,b) : a \in A, b \in B\}$$

average-linkage The average distance between all markers in each cluster

$$\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$$

average-group-linkage The average distance between all markers in the union of both clusters

$$\frac{1}{(|A|+|B|)(|A|+|B|-1)} \sum_{x,y \in A \cup B} d(x,y)$$

Single-linkage is most appropriate for marker grouping because it gives importance to short distances between nearby markers and allows for long distances between markers at reciprocal chromosome ends. The two clusters with the lowest distance are merged into one cluster at the height of their distance. The result is a tree where each marker joins at a specific height. This tree can be split into subtrees based on height or the number of subtrees. When a linkage map is created, the number of chromosomes is usually known, and the tree can be split accordingly. If the markers are distributed equally over the genome, each chromosome is represented by one linkage group.

K-means clustering is a well-established method in data sciences (Macqueen, 1967). It finds, provided a fixed number k of desired clusters, k cluster centers and assigns each data point to one of them so that the within-cluster sum of squares (WCSS) is minimized. The problem itself is computationally difficult (NP-hard), but several heuristic algorithms have been developed that quickly converge (e.g. Lloyd, 2015). That can lead to imperfect solutions (local minima), instead of the global optimum and depends on initial cluster partitions. In Chapter 4 k-means is employed to call genotypes, based on the signal ratio between two alleles. The diploid nature of the subgenome-specific markers reduces the number of potential clusters to one, two or three. The best *k* is determined based on the Bayesian information criterion (BIC) (Schwarz, 1978; Wang et al., 2011). Several software tools apply k-means to classify genotypes (Gidskehaug et al., 2011;

Lin et al., 2008; Shah et al., 2012)

Density-based spatial clustering of applications with noise (DBSCAN) is the third clustering method used in this dissertation (Ester et al., 1996a). In contrast to the previously described methods, it does not require a fixed number of desired clusters and distinguishes between core points, reachable points, and outliers. DBSCAN detects clusters independent of their form and reduces the single-link effect, where to clusters are connected by a thin line of points. It uses two parameters ϵ and $minPts$ and determines the number of clusters based on the data. $minPts$ is the minimum number of points within the neighborhood of a core point, which is defined by its maximum radius ϵ. Points within the neighborhood that have less than $minPts$ points in their ϵ neighborhood are (density) reachable points. Points without neighbors within a ϵ distance are classified as outliers (noise). If a density reachable points is reachable by multiple clusters, it can be assigned to any of them and thus, DBSCAN is not deterministic and depends on the data processing order. The algorithm cannot detect clusters with largely different densities because the same two parameters are used for all data points. DBSCAN is used in Chapter 4 to distinguish between large blocks of synteny and homologous copies which can be found throughout the genome. The syntenic blocks build dense clusters and consist of core points and reachable points. The homologous copies are shorter, less conserved and are classified as outliers by DBSCAN (compare Vignette - Synteny Block Calculation in Appendix C).

Circular binary segmentation (CBS) clusters two-dimensional microarray data points into three classes: decreased, identical or increased DNA copy numbers (number of copies of genomic DNA) (Olshen et al., 2004). The first and second dimensions represent the relation between loci and the signal intensity, respectively. The method recursively divides up each chromosome until it identifies segments which have median signal intensities significantly different from their neighbors. If the segments of adjacent points with similar values exceed an upper or lower threshold, they are labeled as *gain* or *loss* of contiguous segments of the genome. In Chapter 4 CSB is used to detect copy number variations (CNVs) from microarray signal intensities. It can find large deletions or duplications and is robust against noise caused by misplaced or non-hybridized markers. CSB is applied in various software tools for CNV detection in diploids (Miller et al., 2011; Shi et al., 2013; Wiel et al., 2007; Li et al., 2012). Lai et al. (2005) compare CSB to alternative methods like HMMs and expectation maximization algorithms and conclude that it is slow, but performs consistently well.

Seriation

In the context of this dissertation, *seriation* refers to the calculation of a linear order for all points of a dataset (Arabie et al., 1996). The goodness of a particular order is determined by loss or merit functions. For instance the Hamiltonian path length, which interprets the dissimilarity matrix of pairwise recombination frequencies as a finite weighted graph (Caraux et al., 2005; Hubert, 1974). Nodes and weighted edges of the graph represent the data points and the corresponding distances, respectively. The ordering problem is computationally challenging and has an order of $O(n!)$, i.e. the number of possible solutions grows factorially with every additional data point n. For larger datasets, it is infeasible to calculate all possible solutions. Hence, heuristic solutions need to be employed. One of them is hierarchical clustering, which significantly reduces the number of possible combinations by transforming the distance matrix into a dendrogram. Buchta et al. (2008) provide a comprehensive overview about loss/merit functions and seriation methods. In Chapter 3 seriation is used to order markers within linkage groups. The aim is to minimize the distance between adjacent markers.

A simple approach to order data points in a dendrogram would be *OSL1* (Gruvaeus et al., 1972). It is a bottom-up approach, which starts at the leaf level and successively improves orders of subtrees from the leaves to all internal nodes up to the root. When two clusters *c1* and *c2* are merged, the left- and rightmost endpoints *c1l* and *c1r* are compared to *c2l* and *c2r*. Internal orders of clusters (from leftmost to rightmost) remain unchanged and only the node connecting two clusters is affected. The clusters are rotated so that the most similar endpoints are adjacent to each other. Rotation of subtrees does not impact the dendrogram itself because the hierarchical structure remains the same. Instead, it adds information to the dendrogram and the previously random order of leafs becomes a feature. While this approach improves the random order of the hierarchical clustering into a better one, it is vulnerable to local optima and does not improve orders within clusters, once they are built.

A better heuristic is the *optimal leaf ordering (OLO)* algorithm (Bar-Joseph et al., 2001). It minimizes the Hamiltonian path (a path through a graph, where all vertices are visited exactly once) length of the leafs by swapping the dendrogram's subtrees without changing its topology. OLO aims for a global solution and is robust against local optima. The Hamiltonian path length of an OLO solution is always equal to or shorter than the corresponding path length of an OSL1 solution. However, the results are not necessarily unique because multiple orders can have the same Hamiltonian path length. The improved solution comes at the cost of computational performance because the OLO algorithm is more complex and therefore slower than OSL1. Table 1 shows a

	A	B	C	D
B	1			
C	3	2.5		
D	2.5	2	1	
E	4	4	1.2	1.1

Table 1: Pairwise distances between the six markers A-E.

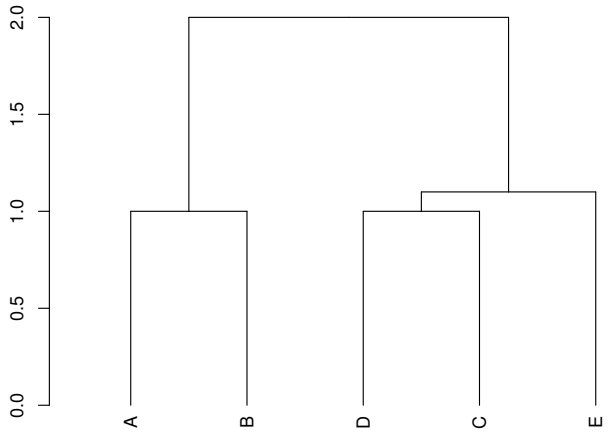

Figure 5: Dendrogram of example markers from Table 1 ordered by OLO.

minimal example were OSL1 and OLO result in Hamiltonian path lengths of 5.6 (AB**CD**E) and 5.2 (AB**DC**E), respectively (compare Figure 5). OSL1 creates the subtree **CD**E (1.1), which is better than **DC**E (1.2). However, CDE is a local optimum and the global path length for of OSL1 is higher because the distance BC (2.5) is larger than BD (2) and invalidates the primary advantage.

Regression analysis

Regression analysis is a statistical concept to determine relationships between variables. In Chapter 2 it is used to find associations between genotypes (independent variables) with a phenotypic trait (dependent variables) (Grandke et al., 2016a). Many different methods are available for regression analysis, but not all can be applied to all datasets. We choose three methods which represent different classes of regression methods and could be applied to our data.

Linear regression assumes a simple relationship between independent and dependent variables x and y (Chambers et al., 1992):

$$Y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

α, β and ϵ represent the intercept, regression coefficient and error term, respectively. i denotes the sample index in the population. The model is fitted using least-squares and the results are the coefficient of determination R^2 and p-values, which indicate the proportion of explained variance and statistical significance, respectively. In Chapter 2 a simple linear regression is used in a GWAS, and each SNP is fitted individually to the phenotypes (Grandke et al., 2016a). p-values are transformed into q-values to account for multiple testing (compare Storey(2003) and Storey

(2015) for further details). Linear regression works well for monogenic traits, but is limited for polygenic traits where each SNP has a low contribution to the phenotype (Freedman, 2009). Instead multivariate methods should be used for GWAS of phenotypes where many collinear SNPs are involved.

Partial least squares (PLS) regression projects observable variables (factors) into a new space to predict the behavior of dependent variables (responses) (Wold, 2004; Helland, 2004). The underlying assumption is that many factors are highly collinear and only a few latent factors can explain most of the responses' variance. Similar methods like principal component analysis (PCA) and maximum redundancy analysis (MRA) maximize factor variance and response variance, respectively (Jolliffe, 2002; Wold et al., 1987; Rao, 1964; Wollenberg, 1977). In contrast, PLS aims to extract latent factors while maintaining variances in factors and responses. In Chapter 2 we use PLS to transform genotypes into latent factors and build a model (Grandke et al., 2016a). The main latent factors are used to predict significant genotypes which are associated with phenotypes. The number of genotypes in the dataset is much larger than the number of phenotypes, which is an ideal situation to apply PLS. Yi et al. (2015) compare PLS to PCA for GWAS and show that both methods outperform linear regression.

Another approach to the problem is Bayesian variable selection (BVS) which simultaneously estimates effects of all genotypes and polygenic effects between them (Schurink et al., 2012). It calculates Bayes factors (BF) for each genotype as the odds ratio between the estimated posterior and prior probabilities. The BF is the ratio of the likelihood probability between two hypotheses and can be used as alternative to p-values, which have many known drawbacks (Good et al., 2003; Goodman, 1999). In Chapter 2 we apply BVS in a genome-wide association study for various traits (Grandke et al., 2016a). O'Hara et al. (2009) provide a comprehensive review about BVS methods.

Aims of the dissertation

This dissertation aims to identify and bridge the gaps of methods and tools that limit polyploid plant breeding. The overarching questions are:

1. What are the differences between diploids and polyploids from a bioinformatics point of view?
2. Which currently available methods cannot be applied to polyploids?
3. What adaptations to bioinformatic methods are required regarding different ploidy types and levels?

Chapter 1

Bioinformatic Tools for Polyploid Crops

Fabian Grandke, Soumya Ranganathan, Andrzej Czech, Jorn R. de Haan and Dirk Metzler
Journal of Agricultural Science and Technology B (2014) 4, 593-601.

The publication is available at `http://www.davidpublisher.org/index.php/Home/Article/index?id=694.html`
DOI: 10.17265/2161-6264/2014.08.001

Chapter 2

Continuous Genotype Values for GWAS in Hexaploid Chrysanthemum

Fabian Grandke, Priyanka Singh, Henri Heuven, Jorn R. de Haan and Dirk Metzler
BMC Genomics (2016) 17:672.

The publication is available at `https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2926-5`
DOI: 10.1186/s12864-016-2926-5

Chapter 3

PERGOLA: Fast and Deterministic Linkage Mapping of Polyploids

Fabian Grandke, Soumya Ranganathan, Nikkie van Bers, Jorn R. de Haan and Dirk Metzler
BMC Bioinformatics in press
Accepted for publication on December 8, 2016

The publication is available at `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1416-8`
DOI: 10.1186/s12859-016-1416-8

Chapter 4

gsrc - an R package for genome structure rearrangement calling

Fabian Grandke, Birgit Samans, Rod Snowdon
Bioinformatics in press
Accepted for publication: October 8, 2016
Advance Access version published online: October 22, 2016

The publication is available at `https://academic.oup.com/bioinformatics/article-abstract/33/4/545/2593902/gsrc-an-R-package-for-genome-structure`
DOI: 10.1093/bioinformatics/btw648

General Discussion

In this chapter, I will discuss the findings of Chapters 1 to 4 and answer the overarching questions raised in the general introduction.

- What are the differences between diploids and polyploids from a bioinformatics point of view?
- Which currently available methods cannot be applied to polyploids?
- What adaptations to bioinformatic methods are required regarding different ploidy types and levels?

Besides, I will look at the bigger picture and discuss applications for our findings outside the context of plant breeding. I will close this chapter with an outlook at remaining challenges and a conclusion.

Bioinformatic differences between diploids and polyploids

In the general introduction, I defined polyploids and elaborated the different origins. In the chapters of this dissertation, we investigated unsolved problems which arose uniquely for polyploids and developed solutions to them. In this section, I want to generalize the particular findings and discuss them from a broader perspective. The main difference between diploids and polyploids are the additional genotype classes. While diploid loci are limited to two different alleles per individual, polyploids can have multiple ones. The same situation arises for duplicated regions in diploids but is rather rare and affected regions can be excluded from the analysis. In polyploids, this is the default condition and needs to be accounted for. Most natural ploidy levels are even, except triploid tardigrades (Bertolani, 2001). Lower ploidy levels are also more common than higher ones due to diploidization. Even if only two alleles are involved in a polyploid locus, it remains a problem for many bioinformatic methods because they were developed for diploids

only (Dufresne et al., 2014; Hollister, 2015). The increased number of chromosome sets results in more than three genotype configurations (Troggio et al., 2013). In Chapter 1 we investigated approaches to overcome limitations of available methods for diploids. Three of them address the topic of polyploid genotype calling, which is described in the general introduction. There is no difference between auto- and allopolyploids in this step of the analysis pipeline, except that allopolyploid markers, can either be subgenome-specific or not. Subgenome specific markers have diploid genotypes as shown in the B-allele frequency distributions in Chapter 4, while unspecific markers can have more than three genotype classes (Gidskehaug et al., 2011). Two of the three genotype calling methods are limited to tetraploids, which is the most common ploidy level. Both work well for datasets from one specific platform but underperform for datasets produced with another technology. The other method, SuperMASSA, is more generic and can be applied to different datasets, independent of ploidy levels and technology. However, we show in Chapter 2 that its output is erroneous and misclassifications lead to wrong predictions in a GWAS. We identified several cases where genotype classes were incorrect. Hence, we developed the method of continuous genotype association and showed that it outperforms available genotype calling methods for polyploids. Our findings do not imply that genotype calling is useless because it is a valuable noise reduction step. Instead, they show that the fundamental problem of genotype calling in polyploids can be solved if an algorithm is tailored to a specific dataset for one ploidy level and one technology. On the contrary, we disprove this with the hexaploid chrysanthemum dataset, where neither available genotype calling tool nor any customized approach worked. It is not clear to what extent bioinformatic approaches can solve this problem. The current setup requires filtering of many SNPs per dataset, due to low coverage or insufficient signal strength. Instead, genotyping technologies should be chosen with the additional chromosome sets in mind to provide a higher resolution which reflects polyploid genotypes better. For instance, GBS with high coverage or microarrays with increased numbers of beads per SNP. The latter one could easily be achieved by using multiple arrays per sample. These approaches would be more expensive for the same number of SNPs and samples, but provide better information and lead to more reliable results. Alternatively, chips with fewer markers could be used, and the final number of informative SNPs would remain the same because fewer markers would be filtered out if the resolution is better. Thus, the costs would be unchanged, but the remaining markers could be scanned with a significantly higher resolution. Besides the technical difficulties of genotype calling the increased number of alleles comes with high performance costs for some methods. In Chapter 1 we observe computational times of more than a day for haplotype phasing a small tetraploid dataset. Large datasets or higher ploidy levels take significantly longer because the number of possible haplotypes

increases exponentially for some of the methods.

The second main difference between diploids and polyploids are the ploidy types, allo- and autopolyploidy, as explained in the general introduction. Knowledge about the origin of polyploidy of the species is required because the two ploidy types need to be distinguished in some analysis steps. In some cases, autopolyploids can be treated as diploids with increased allele count (e.g. linkage mapping in Chapter 3). In contrast, for genotype simulation polyploid meiotic characteristics are important and need to be considered. For instance, PedigreeSim takes double reduction into account, a phenomenon only present in polyploids (compare Chapter 1) (Voorrips et al., 2012). Allopolyploids can be treated as diploids with increased chromosome count in some cases and are also referred to as amphidiploids. In Chapter 4 most SNPs are locus specific, i.e. are present in either subgenome A or subgenome C, but not both. Hence, genotypes are diploid, and the upstream pipeline of gsrc works similar to diploid alternatives. Only the final part about synteny and HNRTs needs to take polyploidy aspects into account. On the contrary, in Chapter 1 we show that four out of ten methods for polyploids cannot be applied to allotetraploids. Segmental alloploidy, as seen in Atlantic salmon is a particular challenge because the ploidy level varies along the genome. In Chapter 1 we discuss the R-package beadarrayMSV, which determines ploidy types for each SNP individually. A similar setup arises for allopolyploids where some loci are subgenome-specific, and others are not. Specific ones can be used for diploid-like linkage mapping and the general ones to assess synteny between the subgenomes as shown in Chapter 4.

Disadvantages and limitations

Now that we understand the main differences between diploids and polyploids from a bioinformatics perspective, we can analyze how these differences limit research and breeding of polyploid crops. The individual steps of modern plant breeding are organized in a workflow as described in the general introduction. Usually, it starts with either sequencing or array-based technologies to determine genotypes. Sequencing is followed up by assemblies which map sequence reads to a reference (mapping assembly) or build a new genome (de-novo assembly). The assembled reads are then scanned for polymorphisms, insertions or deletions (Clevenger et al., 2015). Sequencing reads are prone to errors, and strict filtering methods are applied to separate errors from variants. In higher polyploids, a variant can be present in one to p alleles, where p is the ploidy level. Thus, variant detection in polyploids is challenging and requires a trade-off between error removal and variant detection. Knowledge about population-wide variants or pedigree information can increase confidence for rare variants. The next step for remaining variants and array-based data is genotype

calling to reduce noise, as described in the general introduction. In Chapters 1 and 2 we showed that this is highly challenging and despite many efforts could not be successfully completed for some species, due to limitations of available methods. Thus, genotype calling remains an important task in bioinformatics for polyploids data from microarray technology. However, we demonstrated that continuous genotype values are good alternatives for GWAS in higher polyploids. On the contrary, a recent study on unidirectional diploid-tetraploid introgression among British birch trees constructed a novel pipeline to call variants from targeted resequencing data (Zohren et al., 2016). It takes tri- and tetraallelic variants into account, accepts di- and triploid SNPs and, is tolerant towards missing data. More recently, Blischak et al.(2016) developed a method to call genotypes from sequencing data using a Bayesian model. The authors state that it is limited to autopolyploids and oversimplifies the biological reality.

The next step is the creation of a linkage map for a polyploid population. In the past, this was impossible for most polyploid datasets because available methods required special marker setups and were very limited (compare Chapter 1). In Chapter 3 we developed PERGOLA, a linkage mapping tool (publicly available R package)that works independently of ploidy types and levels. Hence, linkage mapping is no longer a general limitation for research of polyploid crops. We validated the algorithm through simulation studies and demonstrated that it calculates accurate linkage maps for various datasets, including errors and missing data. We further compared it to currently available methods for diploids and showed that, again, it not only produces good linkage maps but also outperformed all available tools computationally. Nevertheless, PERGOLA was developed for populations where both parents are DH or inbred lines and homozygous at each marker. For higher polyploids, this is a condition, which is hard to obtain because it takes many generations of inbreeding until all loci are homozygous. Heterozygous loci can be excluded from linkage mapping, but that reduces the number of markers on the map and subsequently the accuracy of the method. Alternatively, a likelihood-based approach could determine the correct recombination frequency between markers, even if the parents were not homozygous at all positions (Hackett et al., 2013). It would reduce the computational performance of linkage mapping, but require fewer generations of inbreeding for the parents, which would more than balance the costs. For autotetraploids, such a likelihood-based approach has already been developed but has not been implemented in a publicly available tool (Hackett et al., 2013). A similar approach for higher ploidy levels has not yet been developed and remains an open challenge.

The map could then be used to determine haplotypes of a population, but available tools are not satisfying as discussed in Chapter 1. It remains a big challenge, and new bioinformatic methods are required. Eventually, the cost of sequencing-based methods with long reads will drop to a price

that allows haplotyping by sequencing. However, this will be difficult because multiple highly similar chromosome copies are hard to distinguish. Furthermore, currently available long read sequencing methods have significantly higher error rates than short read methods (Laver et al., 2015). A recent study compared three sequence-based haplotyping methods and their ability to find determine haplotypes in polyploids of varying levels (Motazedi et al., 2016). The authors conclude that all methods fail to calculate proper haplotypes for higher polyploids and there is much room for improvement. Their findings require high sequencing depths and cannot be transferred to data originating from array technology. Taken together, haplotype calling remains a problem for research of polyploids, which has not been addressed in the context of this dissertation.

A new approach combines the previous topics of linkage mapping and haplotyping in a potato study (Bourke et al., 2016). It phases pairs of SNPs to calculate the correct recombination frequency, similarly to Hackett et al. (2013) and uses the haplotype-supported values to calculate a linkage map. Linkage maps also allow calculation of quantitative trait loci (QTL), which are the main aim of bioinformatic analyses in the context of plant breeding (Collard et al., 2005). SNPs which lay within a QTL are used to scan large populations for individuals with a particular combination of desirable traits. These selected markers can cheaply be measured in large quantities with systems like competitive allele-specific PCR (KASP^{TM}) or customized microarrays (Semagn et al., 2013).

Insertion/deletion (indel) polymorphisms and CNVs are other types of genetic markers which are used for association studies (Väli et al., 2008; Imprialou et al., 2016). Again, various methods and tools were available for diploids, but not for polyploids. Homeologous non-reciprocal translocations (HNRT) are stretches of one subgenome which are translocated into the other one and are frequent in allopolyploids, due to high similarity between the subgenomes. The impact of HNRT is not well understood as they could only be identified manually based on sequence coverage data (Samans, 2015). We developed gsrc, a publicly available R package to detect CNVs and HNRTs in allopolyploids from microarray data, as described in Chapter 4. It allows automatic analysis and visualization of genomic rearrangements in large populations. We demonstrate how synteny blocks can be calculated from either genome sequences or mapped gene sequences and provide detailed examples for allotetraploid rapeseed (*Brassica napus* L.) and cotton (*Gossypium hirsutum* L.) in Appendix C. Our method requires precise marker positions to find stretches of adjacent markers with similar signal intensities. Often these positions are determined based on a reference genome (Bancroft et al., 2015). Mistakes in the reference genome or differences between the reference genome and the actual genome of the investigated samples lead to misplaced markers, especially in resynthesized samples. These can disturb the detection of CNVs and subsequently

HNRTs. A recent study compared physical and genetic SNP positions in rapeseed and found that only 20,138 of 52,157 could be mapped definitively (Clarke et al., 2016). Another difficulty is standardization of the SNPs in bi-parental populations. In the current version of our tool, each SNP is standardized within the population. This approach works well for natural populations or diversity sets where variations and indels are limited to a small subset of individuals. However, in bi-parental populations genomic rearrangements and genetic variants in any of the parents are inherited by nearly 50 percent of the offspring. This can bias the standardization process and markers which are not present in one-half of the population are shifted towards the average signal value and appear as duplication in the other half. Hence, samples from bi-parental populations need to be standardized separately with diversity sets to account for marker specific variations without falsifying the signal intensity.

Most of our analyses in Chapters 2 to 4 are based on high-throughput microarray data. SNPs on arrays are usually biallelic, i.e. capture only two allelic variants at each position, which is targeted by sequences of flanking regions. The same applies for competitive allele-specific PCR and follows from the fact that most SNPs only have two variants. Polyploids can have more than two alleles per SNP locus and can be tri- or even quadriallelic (Bassil et al., 2015). This limitation can be overcome with sequencing technology but remains for microarrays and PCR-based genotyping technologies. The detection of multiallelic SNPs results in challenges for the downstream analysis. We neglected tri- and quadriallelic SNPs during the developed of our GWAS, linkage mapping and translocation detection methods because they are in general less frequent than biallelic SNPs (Hodgkinson et al., 2010). The exact frequency of multiallelic SNPs for the investigated species is not known. We excluded valuable information from our analyses, and multiallelic SNP-tolerant versions of our methods may lead to better results in the future.

A current trend in plant biology is the development of methods to calculate and investigate pan-genomes, which represent the genomic variation of a species rather than the genome of one individual (Medini et al., 2005). A reference genome is usually represented by one nucleotide sequence per chromosome. Known genetic variations (e.g. SNPs and CNVs) are stored separately indicating the differences between individual genomes and the reference sequence. Pan-genomes only recently became feasible due to decreasing sequencing costs. While the creation of a pan-genome is already a challenging task, this becomes even more difficult for the complex genomes of polyploids. For instance, the pan-genome of *Brassica oleracea* (e.g. cabbage and broccoli) is available, but the economically more important rapeseed (*Brassica napus*) remains unknown (Golicz et al., 2016). The highly similar subgenomes of allopolyploids are hard to distinguish based on short sequence reads. A common workaround in allopolyploid assembly projects is the

inclusion of related diploid genomes into the analysis to support the mapping decision. However, modern genomes differ from their ancestral genomes in many aspects, and the diploid relatives do not represent the allopolyploid subgenomes very well (Cheung et al., 2009). Calculation of pan-genomes is sensitive to variation in every individual, and thus the diploid genomes are not reliable references. A similar challenge applies for pan-transcriptomes, which are used as an intermediate step towards the pan-genomes because RNA-Seq is cheaper and does not include highly repetitive sequences (Hirsch et al., 2014). Nevertheless, for transcriptomic data, alternative splicing and varying expression levels between different tissues add more layers of difficulty to the problem.

Beyond plant breeding

All four chapters of this dissertation were written in a plant breeding context. However, their findings can be transferred to other areas where polyploidy is relevant. Recent research underpins the great impact of polyploidy in many biological processes (Schoenfelder et al., 2015).

Many diploid species have polyploid ancestors, and the polyploid history can still be observed today. The polyploid footprints in the genome are an excellent source of information to understand the evolution of a species (Soltis et al., 2012). Genome duplications caused genetic variety which was advantageous for ancient autopolyploids. Other species formed through hybridization and were temporarily allopolyploid. The long-term disadvantages of polyploidy where overcome by diploidization as explained in the general introduction. To understand these developments and the evolution behind it, detailed knowledge about polyploidy and polyploidization is crucial. gsrc, the tool we developed in Chapter 4, can be used to investigate CNVs and HNRTs in resynthesized allopolyploids, which usually have many rearrangements and sometimes loose entire chromosomes (Mason et al., 2015; Gaeta et al., 2007). The results could lead to a better understanding of rearrangement tolerance and requirements for successful hybridization. Besides the general interest in the evolutionary background of a species, the understanding of underlying mechanisms can be linked back to plant breeding and used to improve crosses and develop new hybrids (e.g trigenomic hexaploid Brassica from a triploid hybrid of *B.napus L.* and *B. nigra*) (Mason, 2016; Pradhan et al., 2010).

Polyploidy also occurs in bacteria and archaea (Soppa, 2014). Among them the species with the largest known ploidy level, *Epulopiscium sp. type B* (Mendell et al., 2008). Our methods from Chapters 2 and 3 can be used for GWAS and linkage mapping in polyploid bacteria and archaea. Also, artificial polyploidy is a promising approach to sequence bacterial genomes (Dichosa et al.,

2012). Single cell sequencing suffers from amplification bias and breakages of genomic DNA. Parts of the genome remain unknown and limit the research in the field. A promising approach is the artificial polyploidization of bacterial cells by inhibition of the bacterial cytoskeleton protein *FtsZ* to block cell division (Dichosa et al., 2012). The polyploid cells have more DNA, which is easier to amplify by qPCR. This leads to improved results of sequencing and a provides better insight into the bacterial genomes.

Animals can also be polyploid, and research in this field can benefit from the findings of this dissertation (Song et al., 2012). Known ploidy levels range up to the dodecaploid Uganda clawed frog (Pasquier, 2009). Most polyploid animals are not subject to any breeding program but are of general research interest. However, the Atlantic salmon, which has a high economic value and is mainly cultivated in aquaculture, is segmentally polyploid. In Chapter 1 we investigated the limitations of beadarrayMSV, which has been developed for a dataset of Atlantic salmon. The methods from Chapters 2 to 4 can be used in this context, as well. Particularly the association of continuous genotypes could be a solution to the problem of varying ploidy levels along the genome. PERGOLA allows to create linkage maps without genotype classification and could be applied for salmon, as well.

Also, there are various fields in human medicine where polyploidy is important. Mammalian polyploidy occurs either naturally (e.g. in hepatocytes), due to stress/aging or in the context of cancer (Davoli et al., 2011; Storchova et al., 2004). In all cases the cells are autopolyploid and, similarly to plants, tetraploidy is most common. Understanding the processes of polyploidization in mammalian organisms could lead to new targets for disease treatments. For instance, polyploid cancer cells are thought to facilitate rapid tumor evolution and prohibition of polyploidization could reduce therapy resistance (Coward et al., 2014). The methods and tools developed in this dissertation could lead to a better general understanding of polyploidy and thus indirectly support the development of novel disease treatments. Furthermore, usage of continuous genotype values as suggested in Chapter 2 could be useful not only for GWAS but also in other research steps. The method is independent of ploidy levels, which is an advantage of tissues which partly consist of diploids and polyploids or where the ploidy level varies.

Outlook

Based on the findings in this dissertation I see three major challenges for the future. Further, I listed remaining constraints in the field of bioinformatics for polyploids in the section Disadvantages and limitations.

Linkage mapping

We showed that our R package PERGOLA could produce accurate linkage maps for di- and polyploids, but it relies on homozygous parents (Grandke et al., 2016c). Obtaining genomes which are largely homozygous becomes increasingly challenging for higher polyploids because it requires more generations of selfing. A valid workaround is the exclusion of non-homozygous markers from the analysis. Nevertheless, this is not ideal, and in the case of high ploidy levels, less than 50 percent of the markers might be available for linkage mapping. A promising approach is to classify each marker based on the parents' genotypes and use maximum likelihood to assess recombination frequencies (Hackett et al., 2013; Bourke et al., 2016). The method increases computational times, but includes more markers and thus, improve the accuracy of linkage mapping. Currently, it is limited to autotetraploid crops and needs to be extended to account for higher ploidy levels.

Haplotyping

Haplotypes improve genomic predictions and are of great interest in the context of polyploids. The methods in this dissertation are based on genotypes (raw data or genotype calls) and not haplotypes. The haplotyping methods presented in Chapter 1 are of limited use because they are not computationally feasible for large datasets and higher ploidy levels (Grandke et al., 2014). The slow computational performance of available methods results from the large number of possible haplotypes. There is an urgent need for faster methods to identify haplotype blocks in polyploids (Motazedi et al., 2016). One possible solution would be a heuristic approach, where not all possible combinations of genotypes are taken into account, but only the most likely ones.

Sequencing

The development of our methods was based on high-throughput microarray data. The current costs of genotyping-by-sequencing based methods exceed the costs of microarrays, once a microarray has been developed. In the future, this is expected to change, and sequencing will be the preferred technique (Thomson, 2014). While in principle, the methods presented in this dissertation can be applied to sequencing data, this needs to be validated and may require some changes. The raw data values in Chapter 2 might be replaced by read count ratios at each SNP position as described in Zohren et al. (2016). Similarly, the intensities in Chapter 4 might be replaced by read counts to detect CNVs (Ji et al., 2015). However, sequencing data provides more information than the number of reads at a specific position in the genome. De-novo assemblies might show HNRTs

without the need for synteny regions, but will be challenging for allopolyploids with highly similar subgenomes (Michael et al., 2015; Yang et al., 2016). The same problem arises for new methods like CRISPR/Cas, which are about to revolutionize plant biology (Osório, 2015). Addressing unique loci in one of the subgenomes might be challenging if they are (at least partially) highly similar. New bioinformatic methods are required to design sgRNA sequences which either target one subgenome or both.

Conclusions

Analyzing polyploid datasets is crucial to breeders and researchers working on various important crops. We analyzed a broad spectrum of bioinformatic applications designed for research and modern plant breeding. Only a few of them can handle polyploid datasets, but have been designed with only one particular species in mind and thus cannot easily be applied to others due to ploidy types and levels. Data analysis workflows that were established for diploid species cannot be applied to polyploids because available tools require diploid genotype classes. We identified genotype classification as a key process, which becomes increasingly difficult with rising ploidy levels. High-throughput microarrays and other technologies have limited signal accuracies and thus, raise a challenge for the downstream analysis of higher polyploids. We developed a series of methods and software tools which do not require genotype classifications and work with continuous values instead. GWAS results become even better because genotype classifications in higher polyploids are erroneous and lead to misclassifications. Our linkage mapping tool creates maps independently of ploidy type and level. Further, it outperforms available tools for diploids regarding computational time. We developed an application to detect and visualize genomic rearrangements in allopolyploid species. Both tools are publicly available R packages and provide access to our methods for both expert and non-expert users. Our findings show that the limitations of polyploid data analysis can be overcome by bioinformatic methods. If polyploidy is taken into account during the planning of an experiment, it can even be advantageous. Future research on polyploid bioinformatics should focus on faster haplotyping methods and data originating from sequencing. Otherwise, the field of plant breeding moves on to sequencing-based methods, and tools will be designed exclusively for diploids and polyploids will stay behind - again. Taken together, our methods provide new functionalities for research on polyploid crops and enable scientists to work on polyploids as if they were diploids.

Appendix A

Supplementary Files for Chapter 2

Supplementary files are available online at https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2926-5 (DOI: 10.1186/s12864-016-2926-5).

Appendix B

Supplementary Files for Chapter 3

Supplementary files are available online at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1416-8 (DOI: 10.1186/s12859-016-1416-8).

Appendix C

Supplementary Files for Chapter 4

Supplementary files are available online at https://academic.oup.com/bioinformatics/article-abstract/33/4/545/2593902/gsrc-an-R-package-for-genome-structure (DOI: 10.1093/bioinformatics/btw648).

Abbreviations

1DKM one dimensional k-means

BAF B-allele frequency

BC backcross

BF Bayes factors

BIC Bayesian information criterion

bp base pairs

BVS Bayesian variable selection

cM centiMorgan

CBS circular binary segmentation

CNV copy number variation

DBSCAN density-base spatial clustering of applications with noise

DH doubled haploid

GBS genotyping by sequencing

GWAS genome-wide association study

GS genomic selection

GSNAP genomic short-read nucleotide alignment program

HAC hierarchical agglomerative clustering

HIPP haplotype interference by pure parsimony

HMM hidden Markov models

HNRT homeologous non-reciprocal translocation

LR linear regression

LRR Log R ratio

MAS marker-assisted selection

MCMC Markov Chain Monte Carlo

MSV multi-side variants

OLO optimal lead ordering

PCA principle component analysis

PCR polymerase chain reaction

PCT Polar coordinate transformation

PLS partial least squares

PLSR partial least squares regression

QTL quantitative trait loci

RIL recombinant inbred line

SAT boolean satisfiability problem

SARF sum of adjacent recombination frequencies

SNP single nucleotide polymorphism

SPLS sparse partial least squares

SS Sanger sequencing

Bibliography

Acquaah, G. (2007). Principles of Plant Genetics and Breeding. 2nd ed. BLACKWELL PUBLISHING.

Affymetrix Power Tools (2015 (accessed July 25, 2015)). `http://www.affymetrix.com/estore/partners_programs/\programs/developer/tools/powertools.affx`.

Ankerst, M., M. M. Breunig, H.-P. Kriegel and J. Sander (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. Philadelphia, Pennsylvania, USA: ACM Press, pp. 49–60.

Arabie, P, L. J. Hubert and G De Soete (1996). Clustering and Classification. WORLD SCIENTIFIC, pp. 5–63.

Baker, F. B. (1974). Stability of Two Hierarchical Grouping Techniques Case I: Sensitivity to Data Errors. *Journal of the American Statistical Association* 69.346, pp. 440–445.

Baker, P. (2008). polySegratio: An R library for autopolyploid segregation analysis.

Bancroft, I., F. Fraser, C. Morgan and M. Trick (2015). Collinearity analysis of Brassica A and C genomes based on an updated inferred unigene order. *Data in Brief* 3, pp. 51–55.

Bar-Joseph, Z., D. K. Gifford and T. S. Jaakkola (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* 17 Suppl 1, S22–29.

Bassil, N. V., T. M. Davis, H. Zhang, S. Ficklin, M. Mittmann, T. Webster, L. Mahoney, D. Wood, E. S. Alperin, U. R. Rosyara, H. Koehorst-vanc Putten, A. Monfort, D. J. Sargent, I. Amaya, B. Denoyes, L. Bianco, T. van Dijk, A. Pirani, A. Iezzoni, D. Main, C. Peace, Y. Yang, V. Whitaker, S. Verma, L. Bellon, F. Brew, R. Herrera and E. van de Weg (2015). Development and preliminary evaluation of a 90 K Axiom®{SNP} array for the allo-octoploid cultivated strawberry Fragaria x ananassa. *BMC Genomics* 16, p. 155.

Bertioli, D. J., P. Ozias-Akins, Y. Chu, K. M. Dantas, S. P. Santos, E. Gouvea, P. M. Guimaraes, S. C. M. Leal-Bertioli, S. J. Knapp and M. C. Moretzsohn (2013). The Use of SNP Markers for Linkage Mapping in Diploid and Tetraploid Peanuts. *G3: Genes|Genomes|Genetics* 4.1, pp. 89–96.

Bertolani, R. (2001). Evolution of the Reproductive Mechanisms in Tardigrades - A Review. *Zoologischer Anzeiger - A Journal of Comparative Zoology* 240.3, pp. 247–252.

Bianco, L., A. Cestaro, G. Linsmith, H. Muranty, C. Denancé, A. Théron, C. Poncet, D. Micheletti, E. Kerschbamer, E. A. Di Pierro, S. Larger, M. Pindo, E. Van de Weg, A. Davassi, F. Laurens, R. Velasco, C.-E. Durel and M. Troggio (2016). Development and validation of the Axiom®Apple 480K {SNP} genotyping array. *The Plant Journal* 86.1, pp. 62–74.

Blischak, P. D., L. S. Kubatko and A. D. Wolfe (2016). Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Molecular Ecology Resources* 16.3, pp. 742–754.

Borlaug, N. E. (1983). Contributions of Conventional Plant Breeding to Food Production. *Science* 219.4585, pp. 689–693.

Bourke, P. M., R. E. Voorrips, T. Kranenburg, J. Jansen, R. G. F. Visser and C. Maliepaard (2016). Integrating haplotype-specific linkage maps in tetraploid species using SNP markers. *Theoretical and Applied Genetics* 129.11, pp. 2211–2226.

Broman, K. W., H. Wu, S. Sen and G. A. Churchill (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19.7, pp. 889–890.

Buchta, C., K. Hornik and M. Hahsler (2008). Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software* 25.3, pp. 1–34.

Cai, G., Q. Yang, B. Yi, C. Fan, D. Edwards, J. Batley and Y. Zhou (2014). A Complex Recombination Pattern in the Genome of Allotetraploid Brassica napus as Revealed by a High-Density Genetic Map. *PLOS ONE* 9.10, e109910.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden (2009). BLAST+: architecture and applications. *BMC bioinformatics* 10, p. 421.

Caraux, G. and S. Pinloche (2005). PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics* 21.7, pp. 1280–1281.

Carter, T. C. and D. S. Falconer (1951). Stocks for detecting linkage in the mouse, and the theory of their design. *Journal of Genetics* 50.2, pp. 307–323.

Carvalho, B., H. Bengtsson, T. P. Speed and R. A. Irizarry (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* 8.2, pp. 485–499.

Carvalho, B. S., T. A. Louis and R. A. Irizarry (2010). Quantifying uncertainty in genotype calls. eng. *Bioinformatics* 26.2, pp. 242–249.

Casci, T. (2010). Population genetics: SNPs that come in threes. *Nature Reviews Genetics* 11.1, pp. 8–8.

Chambers, J. and T. Hastie (1992). Chapter 4: linear models. *Statistical Models in S. Wadsworth & Brooks/Cole*.

Cheema, J. and J. Dicks (2009). Computational approaches and software tools for genetic linkage map estimation in plants. *Briefings in Bioinformatics* 10.6, pp. 595–608.

Chen, Z. (2013). Statistical Methods for QTL Mapping. CRC Press. 310 pp.

Cheng, F., T. MandÃąkovÃą, J. Wu, Q. Xie, M. A. Lysak and X. Wang (2013). Deciphering the Diploid Ancestral Genome of the Mesohexaploid Brassica rapa. *The Plant Cell Online*, tpc.113.110486.

Cheung, F., M. Trick, N. Drou, Y. P. Lim, J.-Y. Park, S.-J. Kwon, J.-A. Kim, R. Scott, J. C. Pires, A. H. Paterson, C. Town and I. Bancroft (2009). Comparative Analysis between Homoeologous Genome Segments of Brassica napus and Its Progenitor Species Reveals Extensive Sequence-Level Divergence. *The Plant Cell* 21.7, pp. 1912–1928.

Chistiakov, D. A., B. Hellemans and F. A. M. Volckaert (2006). Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. *Aquaculture* 255.1-4, pp. 1–29.

Clarke, W. E., E. E. Higgins, J. Plieske, R. Wieseke, C. Sidebottom, Y. Khedikar, J. Batley, D. Edwards, J. Meng, R. Li, C. T. Lawley, J. Pauquet, B. Laga, W. Cheung, F. Iniguez-Luy, E. Dyrszka, S. Rae, B. Stich, R. J. Snowdon, A. G. Sharpe, M. W. Ganal and I. A. P. Parkin (2016). A high-density SNP genotyping array for Brassica napus and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theoretical and Applied Genetics* 129.10, pp. 1887–1899.

Clevenger, J., C. Chavarro, S. Pearl, P. Ozias-Akins and S. Jackson (2015). Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. *Molecular Plant* 8.6, pp. 831–846.

Collard, B., M. Jahufer, J. Brouwer and E. Pang (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* 142.1-2, pp. 169–196.

Collard, B. C. Y. and D. J. Mackill (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363.1491, pp. 557–572.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6.11, pp. 836–846.

Compton, M. E., D. J. Gray and G. W. Elmstrom (1996). Identification of tetraploid regenerants from cotyledons of diploid watermelon cultured in vitro. *Euphytica* 87.3, pp. 165–172.

Coward, J. and A. Harding (2014). Size Does Matter: Why Polyploid Tumor Cells are Critical Drug Targets in the War on Cancer. *Frontiers in Oncology* 4.

Davoli, T. and T. de Lange (2011). The Causes and Consequences of Polyploidy in Normal Development and Cancer. *Annual Review of Cell and Developmental Biology* 27.1, pp. 585–610.

Dichosa, A. E. K., M. S. Fitzsimons, C.-C. Lo, L. L. Weston, L. G. Preteska, J. P. Snook, X. Zhang, W. Gu, K. McMurry, L. D. Green, P. S. Chain, J. C. Detter and C. S. Han (2012). Artificial Polyploidy Improves Bacterial Single Cell Genome Recovery. *PLOS ONE* 7.5, e37387.

Doyle, J. J. and S. Sherman-Broyles (2016). Double trouble: taxonomy and definitions of polyploidy. *New Phytologist*.

Dufresne, F., M. Stift, R. Vergilino and B. K. Mable (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology* 23.1, pp. 40–69.

Edwards, D., J. Batley and R. J. Snowdon (2013). Accessing complex crop genomes with next-generation sequencing. *Theoretical and applied genetics* 126.1, pp. 1–11.

Ekine, C. C., S. J. Rowe, S. C. Bishop and D.-J. de Koning (2013). Why Breeding Values Estimated Using Familial Data Should Not Be Used for Genome-Wide Association Studies. *G3: Genes—Genomes—Genetics* 4.2, pp. 341–347.

Ester, M., H.-p. Kriegel, J. S and X. Xu (1996a). A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, pp. 226–231.

Ester, M., H.-p. Kriegel, J. S and X. Xu (1996b). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Portland, OR: AAAI Press, pp. 226–231.

FAOSTAT (2012). Food and Agriculture Organization of the United Nations. `http://faostat3.fao.org/home/index.html`.

Fisher, R. A. (1947). The Theory of Linkage in Polysomic Inheritance. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 233.594, pp. 55–87.

Floyd, R. W. (1967). Nondeterministic Algorithms. *Journal of the ACM* 14.4, pp. 636–644.

Freedman, D. A. (2009). Statistical models: theory and practice. cambridge university press.

Gaeta, R. T., J. C. Pires, F. Iniguez-Luy, E. Leon and T. C. Osborn (2007). Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. *The Plant Cell* 19.11, pp. 3403–3417.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31.22, pp. 3718–3720.

Galliot, C., M. E. Hoballah, C. Kuhlemeier and J. Stuurman (2006). Genetics of flower size and nectar volume in Petunia pollination syndromes. *Planta* 225.1, pp. 203–212.

Gar, O., D. J. Sargent, C.-J. Tsai, T. Pleban, G. Shalev, D. H. Byrne and D. Zamir (2011). An Autotetraploid Linkage Map of Rose (Rosa hybrida) Validated Using the Strawberry (Fragaria vesca) Genome Sequence. *PLoS ONE* 6.5, e20463.

Garcia, A. A. F., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva, M. L. C. Vieira, R. Vicentini, E. A. Costa, M. C. Mancini, M. O. S. Garcia, M. M. Pastina, R. Gazaffi, E. R. F. Martins, N. Dahmer, D. A. Sforça, C. B. C. Silva, P. Bundock, R. J. Henry, G. M. Souza, M.-A. van Sluys, M. G. A. Landell, M. S. Carneiro, M. A. G. Vincentz, L. R. Pinto, R. Vencovsky

and A. P. Souza (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Scientific Reports* 3.

Garrick, D. J., J. F. Taylor and R. L. Fernando (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution* 41.1, p. 55.

Gidskehaug, L., M. Kent, B. J. Hayes and S. Lien (2011). Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics* 27.3, pp. 303–310.

Gilmour, A. R., R. Thompson and B. R. Cullis (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* 51.4, pp. 1440–1450.

Golicz, A. A., P. E. Bayer, G. C. Barker, P. P. Edger, H. Kim, P. A. Martinez, C. K. K. Chan, A. Severn-Ellis, W. R. McCombie, I. A. P. Parkin, A. H. Paterson, J. C. Pires, A. G. Sharpe, H. Tang, G. R. Teakle, C. D. Town, J. Batley and D. Edwards (2016). The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications* 7, p. 13390.

Good, P. I. and J. W. Hardin (2003). Common errors in statistics (and how to avoid them). Hoboken, NJ: Wiley-Interscience. 221 pp.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine* 130.12, pp. 995–1004.

Goodwin, S., J. D. McPherson and W. R. McCombie (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17.6, pp. 333–351.

Grandke, F., S. Ranganathan, A. Czech, J. R. de Haan and D. Metzler (2014). Bioinformatic Tools for Polyploid Crops. *Journal of Agricultural Science and Technology B* 4, pp. 593–601.

Grandke, F., P. Singh, H. M. C. Heuven, J. R. de Haan and D. Metzler (2016a). Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: a comparative study in hexaploid chrysanthemum. *BMC Genomics* 17, p. 672.

Grandke, F., R. Snowdon and B. Samans (2016b). gsrc - an R package for genome structure rearrangement calling. *Bioinformatics*, in press.

Grandke, F., S. Ranganathan, N. van Bers, J. R. de Haan and D. Metzler (2016c). PERGOLA: Fast and Deterministic Linkage Mapping of Polyploids. *BMC Bioinformatics*, in press.

Gruvaeus, G. and H. Wainer (1972). Two Additions to Hierarchical Cluster Analysis. *British Journal of Mathematical and Statistical Psychology* 25.2, pp. 200–206.

Günther, T., I. Gawenda and K. J. Schmid (2011). phenosim - A software to simulate phenotypes for testing in genome-wide association studies. en. *BMC Bioinformatics* 12.1, p. 265.

Gusfield, D. (2003). Haplotype Inference by Pure Parsimony. *Combinatorial Pattern Matching*. Lecture Notes in Computer Science 2676. Springer Berlin Heidelberg, pp. 144–155.

Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick (2011). Extension of the bayesian alphabet for genomic selection. en. *BMC Bioinformatics* 12.1, p. 186.

Hackett, C. A. and Z. W. Luo (2003). TetraploidMap: Construction of a Linkage Map in Autotetraploid Species. *Journal of Heredity* 94.4, pp. 358–359.

Hackett, C. A., I. Milne, J. E. Bradshaw and Z. Luo (2007). TetraploidMap for Windows: Linkage Map Construction and QTL Mapping in Autotetraploid Species. *Journal of Heredity* 98.7, pp. 727–729.

Hackett, C. A., K. McLean and G. J. Bryan (2013). Linkage Analysis and QTL Mapping Using SNP Dosage Data in a Tetraploid Potato Mapping Population. *PLoS ONE* 8.5, e63939.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet* 8.29, pp. 299–309.

Hastie, T., R. Tibshirani and J. Friedman (2013). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media. 545 pp.

Hazell, P. B. R. (2009). The Asian Green Revolution. Intl Food Policy Res Inst. 40 pp.

He, Y., X. Xu, K. R. Tobutt and M. S. Ridout (2001). Polylink: to support two-point linkage analysis in autotetraploids. *Bioinformatics* 17.8, pp. 740–741.

Heffner, E. L., M. E. Sorrells and J.-L. Jannink (2009). Genomic Selection for Crop Improvement. *Crop Science* 49.1, p. 1.

Helland, I. (2004). Partial Least Squares Regression. *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc.

Heuven, H. C. M. and L. L. G. Janss (2010). Bayesian multi-QTL mapping for growth curve parameters. *BMC Proceedings* 4 (Suppl 1), S12.

Hill, W. G., M. E. Goddard and P. M. Visscher (2008). Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genet* 4.2, e1000008.

Hirsch, C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni, B. Vaillancourt, F. PeÃśagaricano, E. Lindquist, M. A. Pedraza, K. Barry, N. d. Leon, S. M. Kaeppler and C. R. Buell (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell Online*, tpc.113.119982.

Hirschhorn, J. N. and M. J. Daly (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6.2, pp. 95–108.

Hodgkinson, A. and A. Eyre-Walker (2010). Human Triallelic Sites: Evidence for a New Mutational Mechanism? *Genetics* 184.1, pp. 233–241.

Hollister, J. D. (2015). Polyploidy: adaptation to the genomic environment. *New Phytologist* 205.3, pp. 1034–1039.

Hubert, L. (1974). Some Applications of Graph Theory and Related Non-Metric Techniques to Problems of Approximate Seriation: The Case of Symmetric Proximity Measures. *British Journal of Mathematical and Statistical Psychology* 27.2, pp. 133–153.

Huehn, M. (2011). On the bias of recombination fractions, Kosambi's and Haldane's distances based on frequencies of gametes. *Genome / National Research Council Canada = Genome / Conseil National De Recherches Canada* 54.3, pp. 196–201.

Imprialou, M., A. Kahles, J. B. Steffen, E. J. Osborne, X. Gan, J. Lempe, A. Bhomra, E. J. Belfield, A. Visscher, R. Greenhalgh, N. P. Harberd, R. Goram, J. J. Hein, A. Robert-Seilaniantz, J. J. Jones, O. Stegle, P. X. Kover, M. Tsiantis, M. Nordborg, G. Rätsch, R. Clark and R. Mott (2016). Genomic Rearrangements Considered as Quantitative Traits. *bioRxiv*, p. 087387.

Jain, A. K., M. N. Murty and P. J. Flynn (1999). Data Clustering: A Review. *ACM Comput. Surv.* 31.3, pp. 264–323.

Ji, T. and J. Chen (2015). Modeling the next generation sequencing read count data for DNA copy number variant study. *Statistical applications in genetics and molecular biology* 14.4, pp. 361–374.

Jolliffe, I. (2002). Principal component analysis. Wiley Online Library.

Jöreskog, K. G. and H. O. A. Wold (1982). Systems Under Indirect Observation: Causality, Structure, Prediction. Amsterdam: North-Holland. 360 pp.

Kapell, D. N., D. Sorensen, G. Su, L. L. Janss, C. J. Ashworth and R. Roehe (2012). Efficiency of genomic selection using Bayesian multi-marker models for traits selected to reflect a wide range of heritabilities and frequencies of detected quantitative traits loci in mice. *BMC Genetics* 13.1, p. 42.

Körber, N., B. Wittkop, A. Bus, W. Friedt, R. J. Snowdon and B. Stich (2012). Seedling development in a Brassica napus diversity set and its relationship to agronomic performance. *Theoretical and Applied Genetics* 125.6, pp. 1275–1287.

Kosambi, D. D. (1943). The Estimation of Map Distances from Recombination Values. *Annals of Eugenics* 12.1, pp. 172–175.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer and A. Engelhardt (2012). caret: Classification and Regression Training. R package version 5.15-044.

Kuhn, M. and K. Johnson (2013). Linear Regression and Its Cousins. *Applied Predictive Modeling*. New York, NY: Springer New York, pp. 112–121.

Kwong, Q., C. Teh, A. Ong, H. Heng, H. Lee, M. Mohamed, J.-B. Low, S. Apparow, F. Chew, S. Mayes, H. Kulaveerasingam, M. Tammi and D. Appleton (2016). Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. *Molecular Plant* 9.8, pp. 1132–1141.

Lai, W. R., M. D. Johnson, R. Kucherlapati and P. J. Park (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21.19, pp. 3763–3770.

Lamy, P., J. Grove and C. Wiuf (2011). A review of software for microarray genotyping. *Human Genomics* 5.4, pp. 304–309.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln and L. Newburg (1987). MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1.2, pp. 174–181.

Langham, R. J., J. Walsh, M. Dunn, C. Ko, S. A. Goff and M. Freeling (2004). Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics* 166.2, pp. 935–945.

Laver, T., J. Harrison, P. A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz and D. J. Studholme (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification* 3, pp. 1–8.

Lavia, G. I. (2000). Chromosome studies in wild Arachis(Leguminosae). *Caryologia* 53.3-4, pp. 277–281.

Leitch, A. R. and I. J. Leitch (2008). Genomic plasticity and the diversity of polyploid plants. *Science* 320.5875, pp. 481–483.

Li, J., R. Lupat, K. C. Amarasinghe, E. R. Thompson, M. A. Doyle, G. L. Ryland, R. W. Tothill, S. K. Halgamuge, I. G. Campbell and K. L. Gorringe (2012). CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 28.10, pp. 1307–1313.

Lin, Y., G. C. Tseng, S. Y. Cheong, L. J. H. Bean, S. L. Sherman and E. Feingold (2008). Smarter clustering methods for SNP genotype calling. *Bioinformatics* 24.23, pp. 2665–2671.

Liu, B. H. (1998). Statistical genomics: linkage, mapping, and QTL analysis. CRC Press LLC, xxix + 611 pp.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory* 28.2, pp. 129–137.

Luo, Z. W. (2005). Commentary on Wu and Ma. *Genetics* 171.4, pp. 2149–2150.

Luo, Z. W., R. M. Zhang and M. J. Kearsey (2004). Theoretical basis for genetic linkage analysis in autotetraploid species. *Proceedings of the National Academy of Sciences of the United States of America* 101.18, pp. 7040–7045.

Lynce, I. and J. a. Marques-Silva (2006). Efficient Haplotype Inference with Boolean Satisfiability. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*. AAAI'06. Boston, Massachusetts: AAAI Press, pp. 104–109.

Lynch, M. and B. Walsh (1997). Genetics and analysis of quantitative traits.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Mailund, T., S. Besenbacher and M. H. Schierup (2006). Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics* 7.1, p. 454.

Mammadov, J., R. Aggarwal, R. Buyyarapu and S. Kumpatla (2012). SNP Markers and Their Impact on Plant Breeding. *International Journal of Plant Genomics* 2012, pp. 1–11.

Mason, A. S. (2016). Polyploidy and Hybridization for Crop Improvement. CRC Press.

Mason, A. S. and J. Batley (2015). Creating new interspecific hybrid and polyploid crops. *Trends in Biotechnology* 33.8, pp. 436–441.

Massa, A. N., N. C. Manrique-Carpintero, J. J. Coombs, D. G. Zarka, A. E. Boone, W. W. Kirk, C. A. Hackett, G. J. Bryan and D. S. Douches (2015). Genetic Linkage Mapping of Economically Important Traits in Cultivated Tetraploid Potato (Solanum tuberosum L.) *G3: Genes|Genomes|Genetics* 5.11, pp. 2357–2364.

Mather, K. (1936). Segregation and linkage in autotetraploids. *Journal of Genetics* 32.2, pp. 287–314.

Medini, D., C. Donati, H. Tettelin, V. Masignani and R. Rappuoli (2005). The microbial pangenome. *Current Opinion in Genetics & Development*. Genomes and evolution 15.6, pp. 589–594.

Mehmood, T., K. H. Liland, L. Snipen and S. Sæbø (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69.

Mendell, J. E., K. D. Clements, J. H. Choat and E. R. Angert (2008). Extreme polyploidy in a large bacterium. *Proceedings of the National Academy of Sciences* 105.18, pp. 6730–6734.

Michael, T. P. and R. VanBuren (2015). Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology* 24, pp. 71–81.

Miller, C. A., O. Hampton, C. Coarfa and A. Milosavljevic (2011). ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. *PLOS ONE* 6.1, e16327.

Motamayor, J. C., K. Mockaitis, J. Schmutz, N. Haiminen, D. L. III, O. Cornejo, S. D. Findley, P. Zheng, F. Utro, S. Royaert, C. Saski, J. Jenkins, R. Podicheti, M. Zhao, B. E. Scheffler, J. C. Stack, F. A. Feltus, G. M. Mustiga, F. Amores, W. Phillips, J. P. Marelli, G. D. May, H. Shapiro, J. Ma, C. D. Bustamante, R. J. Schnell, D. Main, D. Gilbert, L. Parida and D. N. Kuhn (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* 14.6, r53.

Motazedi, E., R. Finkers, C. Maliepaard and D. d. Ridder (2016). Exploiting Next Generation Sequencing to solve the Haplotyping puzzle in Polyploids: a Simulation study. *bioRxiv*, p. 088112.

Nagaharu, U (1935). Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. *Jap J Bot* 7, pp. 389–452.

Neigenfind, J., G. Gyetvai, R. Basekow, S. Diehl, U. Achenbach, C. Gebhardt, J. Selbig and B. Kersten (2008). Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC Genomics* 9.1, p. 356.

O'Hara, R. B., M. J. Sillanpää and others (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis* 4.1, pp. 85–117.

Olshen, A. B., E. S. Venkatraman, R. Lucito and M. Wigler (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5.4, pp. 557–572.

Ooijen, G. v., G. Mayr, M. M. A. Kasiem, M. Albrecht, B. J. C. Cornelissen and F. L. W. Takken (2008). Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *Journal of Experimental Botany* 59.6, pp. 1383–1397.

Osório, J. (2015). Functional genomics: A novel CRISPR-Cas system for easier genome editing? *Nature Reviews Genetics*.

Page, J. T., A. R. Gingle and J. A. Udall (2013). PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. *G3: Genes|Genomes|Genetics* 3.3, pp. 517–525.

Pasaniuc, B., N. Rohland, P. J. McLaren, K. Garimella, N. Zaitlen, H. Li, N. Gupta, B. M. Neale, M. J. Daly, P. Sklar, P. F. Sullivan, S. Bergen, J. L. Moran, C. M. Hultman, P. Lichtenstein, P. Magnusson, S. M. Purcell, D. W. Haas, L. Liang, S. Sunyaev, N. Patterson, P. I. W. de Bakker, D. Reich and A. L. Price (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics* 44.6, pp. 631–635.

Pasquier, L. D. (2009). The fate of duplicated immunity genes in the dodecaploid Xenopus ruwenzoriensis. *Frontiers in Bioscience* Volume.14, p. 177.

Paterson, A. H., J. E. Bowers and B. A. Chapman (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America* 101.26, pp. 9903–9908.

Patil, G., T. Do, T. D. Vuong, B. Valliyodan, J.-D. Lee, J. Chaudhary, J. G. Shannon and H. T. Nguyen (2016). Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. *Scientific Reports* 6, p. 19199.

Peiffer, D. A., J. M. Le, F. J. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. A. Shaw, J. Belmont, S. W. Cheung, R. M. Shen, D. L. Barker and K. L. Gunderson (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research* 16.9, pp. 1136–1148.

Phillips, C., J. Amigo, A Carracedo and M. V. Lareu (2015). Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data. *Forensic Science International. Genetics* 19, pp. 100–106.

Pompanon, F., A. Bonin, E. Bellemain and P. Taberlet (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* 6.11, pp. 847–846.

Pradhan, A., J. A. Plummer, M. N. Nelson, W. A. Cowling and G. Yan (2010). Successful induction of trigenomic hexaploid Brassica from a triploid hybrid of B.napus L. and B. nigra (L.) Koch. *Euphytica* 176.1, pp. 87–98.

R Core Team (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Ramsey, J. and D. W. Schemske (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics* 29.1, pp. 467–501.

Rao, C. R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4, pp. 329–358.

Rapley, R. and S. Harbron (2004). Molecular Analysis and Genome Discovery. *Molecular Analysis and Genome Discovery*. John Wiley & Sons, Ltd, pp. i–xv.

Rehmsmeier, M. (2013). A Computational Approach to Developing Mathematical Models of Polyploid Meiosis. *Genetics* 193.4, pp. 1083–1094.

Ritchie, M. E., R. Liu, B. S. Carvalho and R. A. Irizarry (2011). Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. *BMC Bioinformatics* 12, p. 68.

Salas Fernandez, M. G., P. W. Becraft, Y. Yin and T. Lübberstedt (2009). From dwarves to giants? Plant height manipulation for biomass yield. *Trends in Plant Science* 14.8, pp. 454–461.

Samans, B. (2015). Homeologous Non-Reciprocal Translocations (HNRT) Induce Selectable Genetic Variation in Brassica napus. *Plant and Animal Genome XXIII Conference*. Plant and Animal Genome.

Sattler, M. C., C. R. Carvalho and W. R. Clarindo (2016). The polyploidy and its key role in plant breeding. *Planta* 243.2, pp. 281–296.

Scheben, A., J. Batley and D. Edwards (2016). Genotyping by sequencing approaches to characterise crop genomes: choosing the right tool for the right application. *Plant Biotechnology Journal*.

Schoenfelder, K. P. and D. T. Fox (2015). The expanding implications of polyploidy. *The Journal of Cell Biology* 209.4, pp. 485–491.

Schurink, A., L. L. Janss and H. C. Heuven (2012). Bayesian Variable Selection to identify QTL affecting a simulated quantitative trait. *BMC Proceedings* 6 (Suppl 2), S8.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* 6.2, pp. 461–464.

Semagn, K., R. Babu, S. Hearne and M. Olsen (2013). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Molecular Breeding* 33.1, pp. 1–14.

Serang, O., M. Mollinari and A. A. F. Garcia (2012). Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. *PLoS ONE* 7.2, e30906.

Shah, T. S., J. Z. Liu, J. a. B. Floyd, J. A. Morris, N. Wirth, J. C. Barrett and C. A. Anderson (2012). optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* 28.12, pp. 1598–1603.

Shi, Y. and J. Majewski (2013). FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data. *Bioinformatics* 29.11, pp. 1461–1462.

Sokal, R. R. and F. J. Rohlf (1962). The Comparison of Dendrograms by Objective Methods. *Taxon* 11.2, pp. 33–40.

Soltis, D. E., P. S. Soltis and J. A. Tate (2003). Advances in the study of polyploidy since plant speciation. *New Phytologist* 161.1, pp. 173–191.

Soltis, D. E., R. J. A. Buggs, J. J. Doyle and P. S. Soltis (2010). What we still don't know about polyploidy. *Taxon* 59.5, pp. 1387–1403.

Soltis, P. and D. E. Soltis (2012). Polyploidy and Genome Evolution. Springer Science & Business Media. 416 pp.

Song, C., S. Liu, J. Xiao, W. He, Y. Zhou, Q. Qin, C. Zhang and Y. Liu (2012). Polyploid organisms. *Science China Life Sciences* 55.4, pp. 301–311.

Soppa, J. (2014). Polyploidy in archaea and bacteria: about desiccation resistance, giant cell size, long-term survival, enforcement by a eukaryotic host and additional aspects. *Journal of Molecular Microbiology and Biotechnology* 24.5, pp. 409–419.

Stephen Milborrow (2015). Notes on the earth package.

Storchova, Z. and D. Pellman (2004). From polyploidy to aneuploidy, genome instability and cancer. *Nature Reviews Molecular Cell Biology* 5.1, pp. 45–54.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of statistics*, pp. 2013–2035.

Storey, J. D. (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.0.0.

Su, S.-Y., J. White, D. J. Balding and L. J. Coin (2008). Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics* 9.1, p. 513.

Suzek, B. E., Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu and t. U. Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31.6, pp. 926–932.

Syvänen, A.-C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2.12, pp. 930–942.

Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding and Biotechnology* 2.3, pp. 195–212.

Troggio, M., N. Šurbanovski, L. Bianco, M. Moretto, L. Giongo, E. Banchi, R. Viola, F. F. Fernández, F. Costa, R. Velasco, A. Cestaro and D. J. Sargent (2013). Evaluation of SNP Data from the Malus Infinium Array Identifies Challenges for Genetic Analysis of Complex Genomes of Polyploid Origin. *PLOS ONE* 8.6, e67407.

Uitdewilligen, J. G. A. M. L., A.-M. A. Wolters, B. B. D'hoop, T. J. A. Borm, R. G. F. Visser and H. J. van Eck (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLoS ONE* 8.5, e62355.

Usadel, B., R. Schwacke, A. Nagel and B. Kersten (2012). GabiPD - the GABI Primary Database integrates plant proteomic data with gene-centric information. *Plant Proteomics* 3, p. 154.

Väli, U., M. Brandström, M. Johansson and H. Ellegren (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics* 9, p. 8.

Van Ooijen, J. W. (2006). JoinMap 4 Manual. Manual. Wageningen.

Voorrips, R. E., G. Gort and B. Vosman (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12.1, p. 172.

Voorrips, R. E. and C. A. Maliepaard (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics* 13.1, p. 248.

Wang, H. and M. Song (2011). Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. *The R Journal* 3.2. OCLC: ocn190786122.

Wang, S., D. Wong, K. Forrest, A. Allen, S. Chao, B. E. Huang, M. Maccaferri, S. Salvi, S. G. Milner, L. Cattivelli, A. M. Mastrangelo, A. Whan, S. Stephen, G. Barker, R. Wieseke, J. Plieske, International Wheat Genome Sequencing Consortium, M. Lillemo, D. Mather, R. Appels, R. Dolferus, G. Brown-Guedira, A. Korol, A. R. Akhunova, C. Feuillet, J. Salse, M. Morgante, C. Pozniak, M.-C. Luo, J. Dvorak, M. Morell, J. Dubcovsky, M. Ganal, R. Tuberosa, C. Lawley, I. Mikoulitch, C. Cavanagh, K. J. Edwards, M. Hayden and E. Akhunov (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal* 12.6, pp. 787–796.

Wang, X., X. Shi, B. Hao, S. Ge and J. Luo (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytologist* 165.3, pp. 937–946.

Wiel, M. A. v. d., K. I. Kim, S. J. Vosse, W. N. v. Wieringen, S. M. Wilting and B. Ylstra (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics* 23.7, pp. 892–894.

Wold, H. (2004). Partial Least Squares. *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc.

Wold, S., K. Esbensen and P. Geladi (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems* 2.1, pp. 37–52.

Wollenberg, A. L. v. d. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* 42.2, pp. 207–219.

Wu, J.-H., A. R. Ferguson, B. G. Murray, Y. Jia, P. M. Datson and J. Zhang (2012). Induced polyploidy dramatically increases the size and alters the shape of fruit in Actinidia chinensis. *Annals of Botany* 109.1, pp. 169–179.

Wu, R., M. Gallo-Meagher, R. C. Littell and Z.-B. Zeng (2001a). A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. *Genetics* 159.2, pp. 869–882.

Wu, R. and C.-X. Ma (2005). A General Framework for Statistical Linkage Analysis in Multivalent Tetraploids. *Genetics* 170.2, pp. 899–907.

Wu, S. S., R. Wu, C.-X. Ma, Z.-B. Zeng, M. C. K. Yang and G. Casella (2001b). A Multivalent Pairing Model of Linkage Analysis in Autotetraploids. *Genetics* 159.3, pp. 1339–1350.

Wu, T. D. and S. Nacu (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26.7, pp. 873–881.

Xu, S. (2013). Principles of Statistical Genomics. New York, NY: Springer New York.

Yang, J., D. Liu, X. Wang, C. Ji, F. Cheng, B. Liu, Z. Hu, S. Chen, D. Pental, Y. Ju, P. Yao, X. Li, K. Xie, J. Zhang, J. Wang, F. Liu, W. Ma, J. Shopan, H. Zheng, S. A. Mackenzie and M. Zhang (2016). The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nature Genetics* 48.10, pp. 1225–1232.

Yi, H., H. Wo, Y. Zhao, R. Zhang, J. Dai, G. Jin, H. Ma, T. Wu, Z. Hu, D. Lin, H. Shen and F. Chen (2015). Comparison of dimension reduction-based logistic regression models for case-control genome-wide association study: principal components analysis vs. partial least squares. *Journal of Biomedical Research* 29.4, pp. 298–307.

Zohren, J., N. Wang, I. Kardailsky, J. S. Borrell, A. Joecker, R. A. Nichols and R. J. A. Buggs (2016). Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Molecular Ecology* 25.11, pp. 2413–2426.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Dirk Metzler who gave me the opportunity to be his PhD student. He gave me intellectual freedom in my work, supported my participation in public outreach activities, engaged me in developing new ideas and demanded a high quality of work throughout all my endeavors. His guidance helped me all the time during my research and the writing of this dissertation. I thank all colleagues from the statistical genetics group and department of evolutionary biology for supporting me and making my placements not only productive but also very enjoyable. I am grateful to my examination committee for evaluating this dissertation.

I am also thankful to my former colleagues at Genetwister Technologies B.V. for providing me with interesting research topics and a friendly working environment. Jorn de Haan, Andrzej Czech and Inge Matthies supervised me during the project and mastered many bureaucratic challenges. I thank Nikkie van Bers and Henri Heuven for scientific discussions, manuscript proofreading and a great trip to the PAG conference. I am grateful to Carlos Villacorta and Priyanka Singh for long nights of great food, music and discussions on bioinformatics and entrepreneurship. I thank the personeelsvereniging and all GT colleagues for memorable outings, entertaining activities and enjoyable lunches.

My sincere thanks go to Richard A. Nichols who not only initiated and coordinated INTERCROSSING, but devoted himself towards the project and its participants. I am grateful to my project partner Soumya Ranganathan, who accompanied me during all stages of the INTERCROSSING adventure and helped to unravel my mind during the development of PERGOLA many times. Further, I am grateful to all PIs and ESRs of INTERCROSSING, who made the project not only a successful training network but also a great experience that will impact the rest of my life. In particular, I want to thank Lizzy Sollars for proofreading this dissertation. I am very grateful to the European Research Council for generously funding the INTERCROSSING ITN.

Thanks to Rod Snowdon, Birgit Samans, Christian Obermeier and the other members of the plant breeding group at JLU Gießen for providing me with an inspiring research environment and

deepening my understanding of plant breeding and allopolyploidy.

My deep appreciation goes to my parents, grandparents, Anna, Sebastian, the rest of my family and friends, who supported me during all stages of my academic life and motivated me to go my own way. Finally, I am grateful to Lisa for her support, encouragement, quiet patience and love.