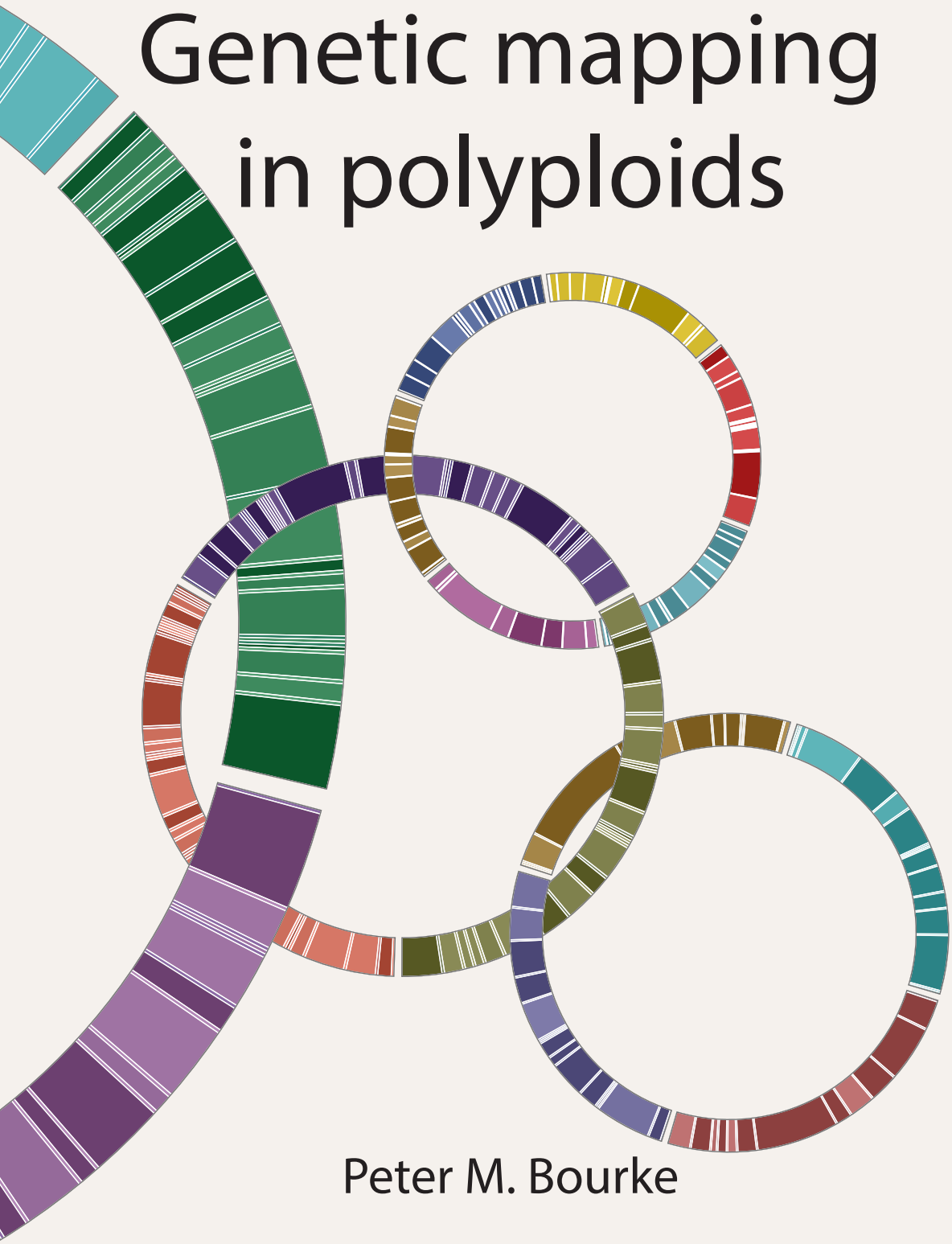
# Genetic mapping in polyploids



Peter M. Bourke

# Propositions

**1.** Ignoring multivalent pairing during polyploid meiosis simplifies and improves subsequent genetic analyses.
(this thesis)

**2.** Classifying polyploids as either autopolyploid or allopolyploid is both inappropriate and imprecise.
(this thesis)

**3.** The environmental credentials of electric cars are more grey than green.

**4.** 'Gene drive' technologies display once again that humans act more like ecosystem terrorists than ecosystem managers.

**5.** Heritability is a concept that promises much but delivers little.

**6.** Awarding patents to cultivars or crop traits is patently wrong.

**7.** The term "air miles" should refer to one's lifetime allowance of fossil-fuelled air travel.

**8.** The Netherlands' most effective educational tool comes on two wheels with a bell.

Propositions belonging to the thesis, entitled

**"Genetic mapping in polyploids"**

Peter M. Bourke
Wageningen, 15th June 2018

# Genetic mapping in polyploids

Peter M. Bourke

**Thesis committee**

**Promotor**
Prof. Dr R.G.F. Visser
Professor of Plant Breeding
Wageningen University & Research

**Co-promotors**
Dr C.A. Maliepaard
Associate professor, Plant Breeding
Wageningen University & Research

Dr R.E. Voorrips
Senior scientist, Plant Breeding
Wageningen University & Research

**Other members**
Dr J. Endelman, University of Wisconsin-Madison, USA
Prof. Dr M.E. Schranz, Wageningen University & Research
Dr J.W. van Ooijen, Kyazma B.V., Wageningen
Prof. Dr B.J. Zwaan, Wageningen University & Research

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences (EPS)

# Genetic mapping in polyploids

Peter M. Bourke

**Thesis**

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 15 June 2018
at 11 a.m. in the Aula.

Peter M. Bourke
Genetic mapping in polyploids,
306 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)
With references, with summary in English

ISBN 978-94-6343-846-9
DOI 10.18174/444415

# Table of contents

# Chapter 1

# General Introduction

# Molecular markers and Mendel's missing caveat

It is only relatively recently that the terms "marker" and "molecular marker" have been taken to be synonymous. In his pioneering experiments with peas, Mendel was already studying the transmission and inheritance of a now-famous set of genetic "markers" (Mendel, 1866), with the physical expression of seven different traits (what we term "phenotypes") being the marker set itself. It took almost fifty years before it was realised that these morphological markers also represented specific locations on chromosomes which were either transmitted together (due to genetic linkage) or independently (due to lack of linkage) (Morgan, 1911). Curiously, Mendel had selected only unlinked traits for his study of peas (whether this was by accident or by design we do not know). His second law, the law of independent assortment, states that alleles of one gene sort into gametes independently of the alleles of another gene. However, because chromosomes and their connection to inheritance had not yet been discovered, Mendel missed one vital caveat to this law, namely that independence only occurs if such genes are located on separate chromosomes (or perhaps at opposite ends of a single long chromosome). Most of what follows in this thesis is based on this non-independent segregation of linked loci. One of the major milestones in genetics was the realisation that a collection of linked markers could be arranged in a linear fashion, with distances between their positions estimated from the counts of co-inherited markers (Sturtevant, 1913). In his demonstration of this fact using six linked morphological markers of the common fruit fly *Drosophila melanogaster*, Sturtevant created the world's first genetic linkage map (Sturtevant, 1913;Van Ooijen and Jansen, 2013).

Nowadays, we generally rely on DNA markers (a.k.a. molecular markers) to identify positions on chromosomes. In this thesis we exclusively present data on single nucleotide polymorphism (SNP) markers. These are nucleotide positions which generally differ in the individuals being screened and are usually selected to be "bi-allelic" (*i.e.* alternating between two possible nucleotides in the material under study). However, the methods we develop are general to any bi-allelic marker system for which marker "dosage" counts can be accurately estimated. The term "dosage" is less commonly used in diploid studies, but is a very important concept in genetic analyses of polyploids. Dosage is generally understood to be the number of copies of the alternative allele carried by an individual at a particular locus. An illustration of the possible marker dosage scores in a tetraploid and hexaploid are shown in Figure 1.

# Linkage mapping and QTL analysis – part I

The principal aim of this thesis is the development and exploration of methods and tools to create linkage maps and perform quantitative trait locus (QTL) analysis in polyploids. Both of these activities combined can conveniently be termed "genetic mapping" as the title of this thesis suggests. These are not new activities – as already mentioned, the first linkage map was constructed in 1913, and the first QTL analysis was arguably conducted a decade later (Sax, 1923). The novelty of this work lies in its application to polyploids, a group of organisms that possess much more complex genomes than their diploid counterparts, and in its use of modern genotyping data which often consists of many tens of thousands of markers.

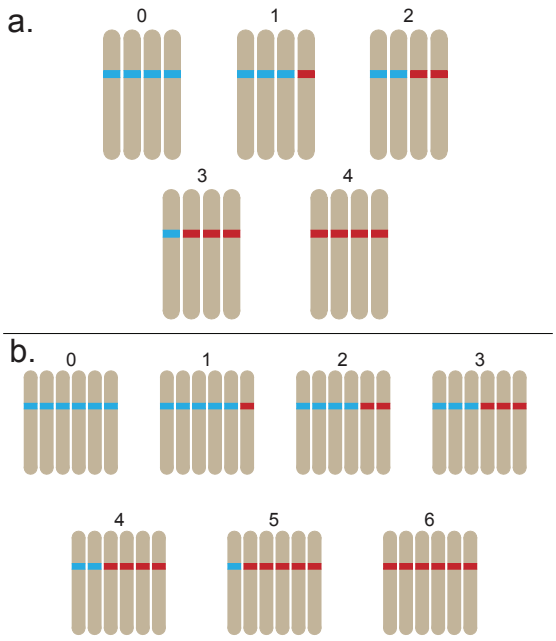



**Figure 1. Representation of the possible marker dosage scores, for a. Tetraploid (*ploidy* = 4) and b. Hexaploid (*ploidy* = 6).** In general, there are *ploidy* + 1 possible dosage classes at a bi-allelic marker position, here shown in blue / red. The convention for assigning dosage scores is to count the number of copies of the alternative allele (coloured red), with the reference allele count (in blue) given by *ploidy* – *dosage*.

Modern plant and animal breeding has benefitted from the use of molecular markers, allowing specific traits of interest to be predicted without having to necessarily grow a plant to maturity and test the trait directly. This can have many advantages, for example by allowing a greater number of traits to be more cost-effectively combined, allowing selection among multiple progenies, or in material for which phenotyping is difficult

or expensive *etc*. At least, that is the intention of marker-assisted selection (MAS). In practice, MAS has met with mixed success in plant breeding programs (Collard and Mackill, 2008;Hospital, 2009). In the autotetraploid crop potato (*Solanum tuberosum* L.), markers have been demonstrated to be useful for certain disease resistances (*e.g.* potato cyst nematode (Schultz et al., 2012)) and have been shown to be a cost-effective proposition if deployed appropriately in a breeding program (Slater et al., 2013). However, in order to use markers effectively, we need to develop models to capture the relationship between a plant's genotype and its phenotype. For polyploids, modelling these relationships is still relatively new and is fraught with added complications not encountered in diploids. Polyploid inheritance patterns are by nature more complex than diploids. Furthermore, many polyploids are outcrossing species that resist the genetic simplification achieved through inbreeding. We will first examine some of the challenges that polyploids pose to the molecular geneticist, and in particular the idiosyncrasies of polyploid meiosis, before returning to the topic of linkage mapping and QTL analysis.

# Polyploidy

A polyploid is any organism that carries more than two copies of each chromosome, from the Greek "*poly*-" meaning *much* or *many*, and "*–ploid*" from "*ploos*", meaning *fold*, thus "*many-fold*". Polyploidy was discovered more than a century ago (Strasburger, 1910) and since then has been a topic of continued interest and debate. Diploidy, the state of possessing two copies of each chromosome, is considered the chromosomal "ground state" or norm for complex organisms. In diploids, parents transmit a single copy of each chromosome to each gamete, thereby re-establishing diploidy in the offspring. In polyploid organisms, more than one copy of each chromosome is transmitted, which is one of the main contributing factors to the complexity of polyploid genetics. There is a tendency for polyploid lineages to return to a diploid conformation over evolutionary timescales, a process termed "diploidisation" or "re-diploidisation" (Ohno, 1970;Le Comber et al., 2010). In the timescales of interest to breeders and researchers however, polyploidy is effectively a permanent condition. Although the definition of polyploidy is quite unequivocal, there can be some confusion over the classification of a species as polyploid or not, particularly as many complex life-forms were polyploid at some point in the past (Van de Peer et al., 2017). "Paleopolyploids" are species that were true polyploids millions of years ago but have since re-diploidised, and the term "neopolyploids" refers to newly-formed polyploids (Ramsey and Schemske, 2002;Lloyd and Bomblies, 2016), possibly artificially generated for research purposes to understand how polyploids deal with the

initial "genomic shock" of having an extra genome (McClintock, 1984). Neopolyploidy may also refer to recently-formed wild polyploid populations such as *Spartina anglica*, an allopolyploid that arose when *Spartina alterniflora* was introduced outside its native range and hybridised with local *Spartina* species (Soltis and Soltis, 2009). Some authors further distinguish "mesopolyploids" as re-diploidised species that underwent whole-genome duplication (WGD) at a less ancient timescale than paleopolyploids and which can be detected by genetic or cytogenetic analyses (Mandáková et al., 2010;Franzke et al., 2011). In this thesis we are principally concerned with extant polyploid species that have not re-diploidised, yet have already passed the (presumed bumpy) early generations *i.e.* they are no longer considered neopolyploid.

Among polyploids, two distinct types are generally recognised – autopolyploids and allopolyploids. These terms can distinguish or emphasise two features, namely the origin of the polyploid (also termed the "taxonomic" definition), or how its chromosomes behave during meiosis (the "genetic" definition) (Ramsey and Schemske, 2002;Doyle and Egan, 2010). Autopolyploids are generally-speaking derived from a single species and exhibit polysomic inheritance. Polysomic inheritance means that all possible combinations of alleles are equally likely to end up in a gamete – although we will return to this question in more detail in the next section as it is one of the fundamental points of interest in this thesis. Allopolyploids on the other hand are derived from at least two species and exhibit disomic inheritance (where disomic means diploid-like inheritance, the result of exclusive pairing and recombination between homologous chromosomes and an absence of pairing and recombination between homoeologous chromosomes). Although also important, they are not the focus of study here. It should be noted that classifying a species as either autopolyploid or allopolyploid is not always straightforward, as demonstrated for example in the debate about the correct classification of the polyploid ancestor of soybean (*Glycine* spp.) (Barker et al., 2016;Doyle and Sherman-Broyles, 2016). In other words, the taxonomic and genetic definitions do not always neatly overlap, particularly in species with a long history of inter-specific hybridisation among progenitor species of varying relatedness. A large body of polyploid research is aimed at understanding how different polyploid lineages arose, and how these newly-wed genomes adapted and evolved to accommodate each other and their changing environment.

There is a third category of polyploid, namely the "segmental allopolyploid" as it was originally termed (Stebbins, 1947). Again this category can be defined from a taxonomic perspective or a genetic perspective – as a hybridisation between two very closely-related species or subspecies, or as a polyploid which demonstrates a meiotic

pairing behaviour that cannot be classified as fully disomic or fully polysomic (recently termed “mixosomic” (Soltis et al., 2016)). Throughout this thesis we rely on the genetic definition, as it is the pairing behaviour that influences how homologues recombine, upon which our methods to study inheritance are ultimately based. Although it is interesting to speculate upon *how* or *why* such differences arose, in the end we are primarily interested in understanding *what* happens, as this is the most solid ground upon which to build a model.

Polyploidy occurs in animals, plants and fungi, with the ancestors of all angiosperms and vertebrates thought to have experienced at least two whole-genome duplications (Putnam et al., 2008;Jiao et al., 2011). In the plant kingdom there are numerous examples of extant polyploids. There are fewer known examples of polyploid animals, which some suggest is due to difficulties in re-establishing a balance in chromosomal sex-determination systems following genome duplication (Muller, 1925;Orr, 1990). However examples do exist, particularly in amphibians and fish (but much less so in other vertebrates) (Mable et al., 2011). Amphibious examples include the Bluespotted-Jefferson salamander complex (Uzzell, 1964), the grey tree-frog *Hyla versicolor* (Ptacek et al., 1994), the American ground frog *Odontophrynus americanus* (Beçak et al., 1966) and the African clawed frog *Xenopus laevis* (Session et al., 2016). Among fish, it is now well-established that a whole genome duplication (WGD) occurred in the ancestor of all salmonids (*e.g.* salmon, trout *etc.*) between 50 and 100 million years ago (Allendorf et al., 2015). Polyploidy is also sometimes artificially induced in animals, for example in Pacific oysters (*Crassostreae gigas*) (Benabdelmouna and Ledu, 2015) or the silkmoth (*Bombyx mori* L.) (Rasmussen and Holm, 1979). There continues to be debate about whether any polyploid mammals exist, with the initial claim that the Argentinian red vizcacha rat (*Tympanoctomys barrerae*) is tetraploid (Gallardo et al., 1999) being more recently challenged in light of new data (Svartman et al., 2005;Evans et al., 2017).

Polyploidy occurs widely among plant species, with recent advances in whole-genome sequencing allowing a detailed analysis of recent and ancient polyploidisation events in an increasingly large number of plant lineages (Vanneste et al., 2014;Van de Peer et al., 2017). In natural populations of plants there is always a small possibility of a new polyploid species arising (usually although not exclusively through unreduced (2n) gametes (Harlan and De Wet, 1975)). In the case of autotetraploids, one possible path for their establishment is through the initial formation of a triploid bridge (from a fusion of n + 2n gametes) (Ramsey and Schemske, 1998;Schinkel et al., 2017). These triploids, although generally infertile, may also produce 2n gametes and hence through

selfing or pollination by 2n gametes from diploids, tetraploids may be formed (Ramsey and Schemske, 1998;Comai, 2005). Note that for a new polyploid lineage to establish and diversify, an even-numbered ploidy is required. An exception to this is when plants exclusively reproduce vegetatively or apomictically, thereby avoiding the disruptions that odd-numbered ploidies pose to balanced meiotic division. However such lineages would be expected to evolve more slowly than sexually-reproducing ones (McDonald et al., 2016). The fusion of unreduced pollen with an unreduced egg cell, both from diploid parents, is also theoretically possible ("one-step" tetraploids (Ramsey and Schemske, 1998)). Induced polyploidy (man-made) can also occur through somatic chromosome doubling, although its status as a means to polyploid formation in natural populations is uncertain (Harlan and De Wet, 1975). An interesting alternative pathway that has only recently been explored is the possibility of polyspermy, where more than one sperm cell fertilises an ovule (Dresselhaus and Johnson, 2018). Interestingly, stressed plants are found to produce more unreduced gametes (such stresses may relate to environmental variables like extreme temperatures, wounding, drought or nutrient deficiency (Ramsey and Schemske, 1998)). Unreduced gametes are thought to arise as a result of defective spindle fibres or cell-plate formation, both of which have been shown to regularly occur at extreme (particularly higher) temperatures (Pécrix et al., 2011;De Storme and Geelen, 2014;Bomblies et al., 2015)).

In most cases, neopolyploid plants are usually at an immediate disadvantage, being un-adapted and reproductively isolated (what is termed "minority cytotype disadvantage" (Levin, 1975;Husband, 2000)). The speed at which newly-established polyploid lineages prosper and diversify varies, with indications that there may be a significant time lag before this occurs (Schranz et al., 2012). One of the fascinating hypotheses that has emerged is the possibility that a disproportionately-high number of WGDs occurred close to the Cretaceous-Paleogene boundary (around 66 million years ago) (Van de Peer et al., 2017). Non-avian dinosaurs are thought to have disappeared around the same time, along with 60-70% of all plant and animal life, coinciding with a number catastrophic phenomena including perturbations in the global climate, increased volcano activity and the impact of a large meteor near Chicxulub, Mexico (Renne et al., 2013). Clearly this was a difficult time to be alive on planet Earth. Nevertheless, in times of severe environmental stress the increased genomic plasticity of polyploids (te Beest et al., 2011) coupled with their propensity to vegetatively propagate (Herben et al., 2017), as well as reduced competition from severely-stressed or dying diploids may have contributed to their success.

Polyploids are particularly common among domesticated crops (Salman-Minkov et al., 2016), a fact that has helped drive interest to better understand these species. In many cases polyploidy is deliberately induced – for example modern ornamental breeding often relies on inter-specific hybrids to create novel varieties, which are often "polyploidised" (through colchicine treatment for mitotic polyploidisation, or through selection of 2n gametes) to overcome sterility in the $F_1$ (Van Tuyl and Lim, 2003). Fruit breeders also generate seedless fruit by crossing parents of different ploidy levels, resulting in sterile fruit-bearing (usually triploid) offspring (Bradshaw, 2016). Many of the native attributes of polyploids may also have endeared them to the early agriculturalists, *e.g.* larger organs (tubers, fruits, flowers *etc.*), also known as the "gigas" effect (Sattler et al., 2016), or their ability to be clonally propagated. We therefore find ourselves in the position of relying on some of the most genetically-complex species to provide us with the basic necessities for life. Examples of some globally-important polyploid crops include allopolyploids such as wheat (*Triticum aestivum* L.), cotton (*Gossypium hirsutum* L.), coffee (*Coffea arabica* L.), oilseed rape (*Brassica napus* L.), oats (*Avena sativa* L.), peanut (*Arachis hypogaea* L.), strawberry (*Fragaria* × *ananassa* L.) and autopolyploids such as potato (*Solanum tuberosum* L.), alfalfa (*Medicago sativa* L.), sweetpotato (*Ipomoea batatas* L.), leek (*Allium ampeloprasum* L.), blueberry (*Vaccinium corymbosum* L.), chrysanthemum (*Chrysanthemum* spp.) and rose (*Rosa* × *hybrida* L.). Despite their importance, polyploids, and in particular autopolyploids, remain an understudied and poorly-understood group. One may attribute this to their genetic complexity and the fact that few software tools are developed to analyse autopolyploid data (noting that due to the diploid-like inheritance of allopolyploids it is possible to use many of the diploid software tools for those crops). It is also only relatively recently that datasets containing sufficient marker numbers to allow investigations into their genetic properties have become available. In this thesis we aim to tackle this deficit by developing tools and methods to analyse autopolyploid genotype and phenotype data. Although the methods we develop are general, we have so far only applied them to plant datasets. Before going into more detail on how this was done, we first need to understand the process of meiosis.

# Meiosis

The topic of meiosis in polyploids is central to this thesis. Indeed, one of our primary interests in markers is that they provide a detailed picture of meiosis when deployed over a population of related individuals. The inheritance information from the population allows us not only to organise these markers into ordered clusters representing chromosomes,

but also to track the pairing and recombination that occurred in each parental meiosis that gave rise to a particular offspring individual. Meiosis can be defined as "two successive nuclear divisions that produce gametes carrying one half of the genetic material of the original cell" (Griffiths et al., 2012). In plants and fungi as opposed to animals, the direct product of meiosis is technically not a gamete, but rather a spore which gives rise to a gametophyte which later produces the gamete cells. However, from the perspective of inheritance, all the interesting phenomena occur during meiosis, and not in any later intermediate (mitotic) cell divisions which carry the haploid chromosomes towards their final destiny as gametes. Therefore, in this thesis the term "gamete" is often loosely used to denote the products of meiosis, when "spore" would have been more appropriate in plant-specific contexts.

## Meiosis in polyploids

In describing polyploid meiosis, one can either choose to give a very detailed picture or to sketch the most important features and hope that in so doing the audience are still following. In this thesis we try to take the middle road, avoiding unnecessary terminology or technical detail where possible. Apart from lucidity, we are also only interested in meiosis from a practical perspective – it is simply a process that we model in order for us to make sense of marker data. We leverage this data to come to a better understanding of the chromosomal composition of our experimental organisms (in the fullest sense – from gene order and position to gene function and effect). One clear example of our (over) simplification is the use of the term "pairing" in meiosis. In the early stages of meiosis I (leptotene) homologues pair up – they physically align. This is followed by the formation of synaptonemal complexes during zygotene. However, in cases where this pairing is aberrant (*e.g.* between homoeologues) any synaptonemal complexes that may have formed are dissolved or corrected in late zygotene and pachytene, and only bivalents between homologues persist into metaphase I. In this thesis we avoid terms such as leptotene, zygotene and pachytene, and use "pairing" to describe the conformation of (recombining) homologues (or homoeologues even) that have persisted into metaphase I. If this pairing behaviour is not random but preferentially occurs between certain homologues, we call this behaviour "preferential pairing", which is one of the main distinguishing characteristics of a segmental allopolyploid. However technically-speaking it would be better to talk about "preferential crossover formation" instead (Lloyd and Bomblies, 2016). In fact, the literature on the subject of preferential pairing appears to be written by two groups – those who are comfortable with terms like

"leptotene" and those who are not.

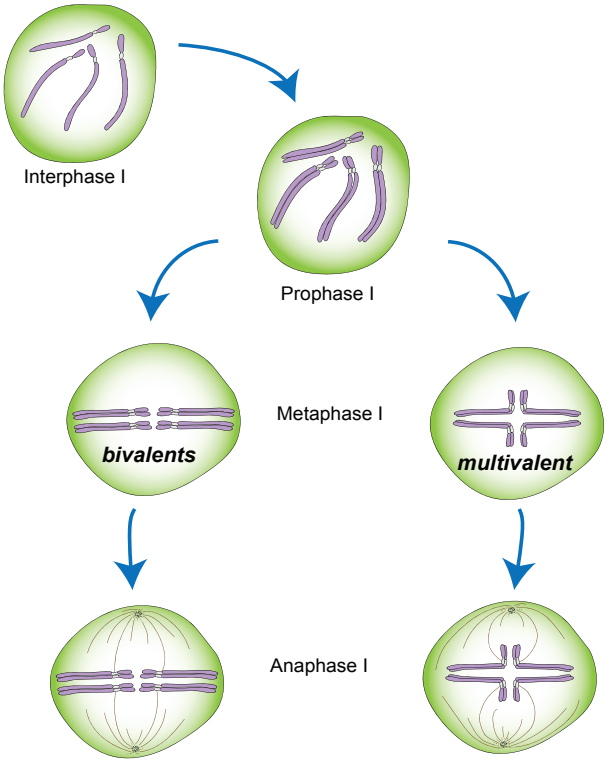



**Figure 2. Simplified scheme of the early stages of polyploid meiosis, in this case for an autotetraploid.** By the beginning of Metaphase I, pairs of sister chromatids will have associated and come to lie on the equatorial plane, either in *bivalent* or *multivalent* pairing structures.

From a simplified perspective, therefore, there are a number of aspects that are relevant for this thesis. The first concerns the early stages of meiosis which lead to the formation of different types of pairing structures (Figure 2). As already mentioned, the precise details of Prophase I are beyond the scope of this introduction, but the interested reader is directed to any number of excellent reviews on the subject of meiosis in general (Harrison et al., 2010;Da Ines et al., 2014;Hunter, 2015;Mercier et al., 2015;Zickler and Kleckner, 2015), and in particular those that look specifically at polyploid meiosis (Cifuentes et al., 2010;Zielinski and Scheid, 2012;Grandont et al., 2013;Moore, 2013;Lloyd and Bomblies, 2016).

At the onset of meiosis each homologous chromosome condenses and is replicated into pairs of sister chromatids (shown in Figure 2, Prophase I). These are held together by cohesion complexes that persist into the later stages of meiosis (Hunter, 2015). These pairs of sister chromatids then associate either as pairs, in which case they are described

as *bivalents*, or in groups of three or more, described as *multivalents* (Figure 2, Metaphase I). It is now known that in order for meiosis to proceed in an orderly fashion, there must be formation of chiasma leading to at least one cross-over per chromosome (the number of cross-overs per chromosome tends to also have an upper bound, with rarely more than three observed) (Mercier et al., 2015) which occur through the formation and subsequent repair of double-strand breaks in the DNA (Henderson and Keeney, 2004). The most commonly-observed multivalents involve four homologues (also called a *quadrivalent*), with on average 27% of all pairing structures found to be quadrivalents in a meta-analysis of autopolyploid meiosis (Ramsey and Schemske, 2002). By contrast, only 2% of trivalent structures were observed. Uneven numbers of pairing homologues are more likely to lead to aneuploid gametes (carrying an unbalanced number of chromosomes) which rarely survive. Multivalents are more likely to be observed in autopolyploids than allopolyploids (because in the latter case there are barriers to pairing and recombination between homoeologues), but even in allopolyploids they may occur (Ramsey and Schemske, 2002).

## Double reduction

It has long been suggested that multivalents are aberrant pairing structures that can often lead to aneuploidy and reduced fertility (Darlington, 1937;Kostoff, 1940;Hazarika and Rees, 1967;Lloyd and Bomblies, 2016), but this view likely stems from a bias towards neopolyploids which have been shown to produce abnormally-high numbers of multivalents in early generations, stabilising to lower frequencies in later generations (Ramsey and Schemske, 1998;2002;Santos et al., 2003;Bomblies et al., 2016). There are numerous reports of multivalents being formed in a wide range of established autopolyploids, without necessarily negatively impacting on fertility (Swaminathan and Howard, 1953;Morrison and Rajhathy, 1960;Jones and Vincent, 1994;Khawaja et al., 1997;Ramsey and Schemske, 2002). As this thesis is concerned with stable, established autopolyploids, we are primarily interested in the genetic consequences of multivalents and in particular the phenomenon of *double reduction* (Figure 3). Double reduction has been known about for almost a century and theories of how it should be modelled have been considered by some of the early pioneers of statistical genetics (Haldane, 1930;Mather, 1935;Fisher, 1947). It is quite common nowadays for authors to assert that double reduction is required for a complete and accurate description of autopolyploid inheritance (and by extension, that it be ignored at great peril) (Luo et al., 2004;Wu et al., 2004;Wu and Ma, 2005;Li et al., 2010;Lu et al., 2012;Xu et al., 2013). On the other

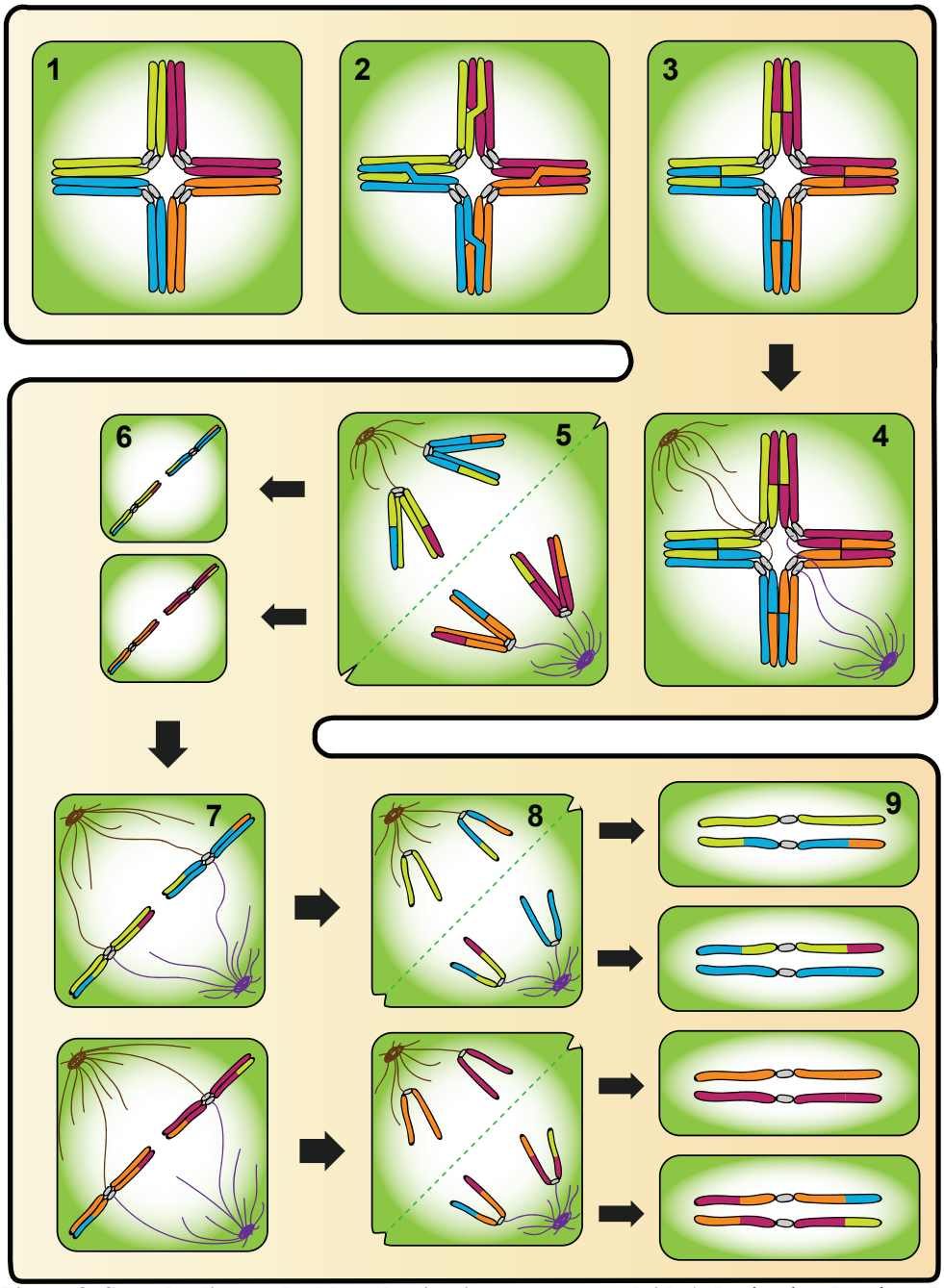



**Figure 3. Steps leading to double reduction in an autotetraploid. 1.** During late Prophase I, pairs of sister chromatids associate. Multivalents involving four pairs of sister chromatids are also called quadrivalents, as shown. **2.** For quadrivalents to be maintained from Prophase I into Metaphase I, cross-overs need to form between at least three of the four pairing partners. Here, four cross-overs are shown between all partners, leading to a ring-quadrivalent. **3.** Resolved double-strand breaks leading to cross-overs, as seen in late Metaphase I. **4.** During Anaphase I,

spindle fibres originating from the centrosomes attach to the centromeres. **5.** For proper chromosomal segregation, two of the four pairing homologues must migrate to each pole. This mechanism appears to be unregulated in neopolyploids but established polyploids often exhibit correct segregation leading to no abnormalities. For double reduction to occur, a homologue and its pairing recombination partner must migrate to the same pole, a situation that is impossible with only bivalent pairing. **6.** Following interkinesis (which includes Telophase I with the division of the cell in two, and Prophase II which is rather uneventful) pairs of sister chromatids carrying recombinant homologues come to lie again at the equatorial plane in Metaphase II. **7.** Anaphase II subsequently follows, with spindle fibres again attaching to the centromeres. At this stage, cohesion complexes that held chromatids together are dissolved. **8.** In late Anaphase II, cell division again occurs. **9.** Based on how sister homologues segregate, it is possible for gametes (spores) carrying two identical copies of part of a homologue to occur. In this example, the upper two gametes carry double reduction products (*e.g.* homozygous blue section), whereas the lower two gametes do not.

hand, a number of recent reviews of allo- and autopolyploid meiosis completely fail to mention it (Bomblies et al., 2016;Lloyd and Bomblies, 2016). In this thesis double reduction is not ignored, but in many cases we do omit it from our models. However, we take particular care that if so doing, we do not introduce unnecessary bias.

## Preferential pairing

One other aspect of polyploid meiosis that requires our attention is the phenomenon of "preferential pairing" (or as pointed out already, perhaps more accurately termed "preferential crossover formation" (Lloyd and Bomblies, 2016)). Allopolyploids consist of both homologues (closely related chromosomes presumed to have derived from the same species) and homoeologues (less closely-related chromosomes, presumed to be from distinct species). During meiosis, homologues are found to pair and recombine whereas homoeologues are not. This is often under the direct control of specific genes or loci, such as the *Ph1* locus in wheat (*Triticum aestivum* L.) (Okamoto, 1957;Riley and Chapman, 1958) or the *PrBn* locus in oilseed rape (*Brassica napus* L.) (Jenczewski et al., 2003;Nicolas et al., 2009). In such instances we refer to "fully preferential pairing" between homologues, leading to disomic inheritance. In autopolyploids on the other hand there is an equal chance of pairing and recombination between all homologues during meiosis, leading to polysomic inheritance. However, these states represent the two extremes of a supposed continuum between random and non-random pairing, with the term "segmental allopolyploid" (Stebbins, 1947) used to gather together everything in between (including the possibility of certain chromosomes behaving disomically and others polysomically). Examples where a mixture of disomic and polysomic inheritance have been observed are rainbow trout (*Oncorhynchus mykiss*) (Allendorf and Danzmann, 1997), peanut (*Arachis hypogaea* L.) (Leal-Bertioli et al., 2015;Nguepjop et al., 2016),

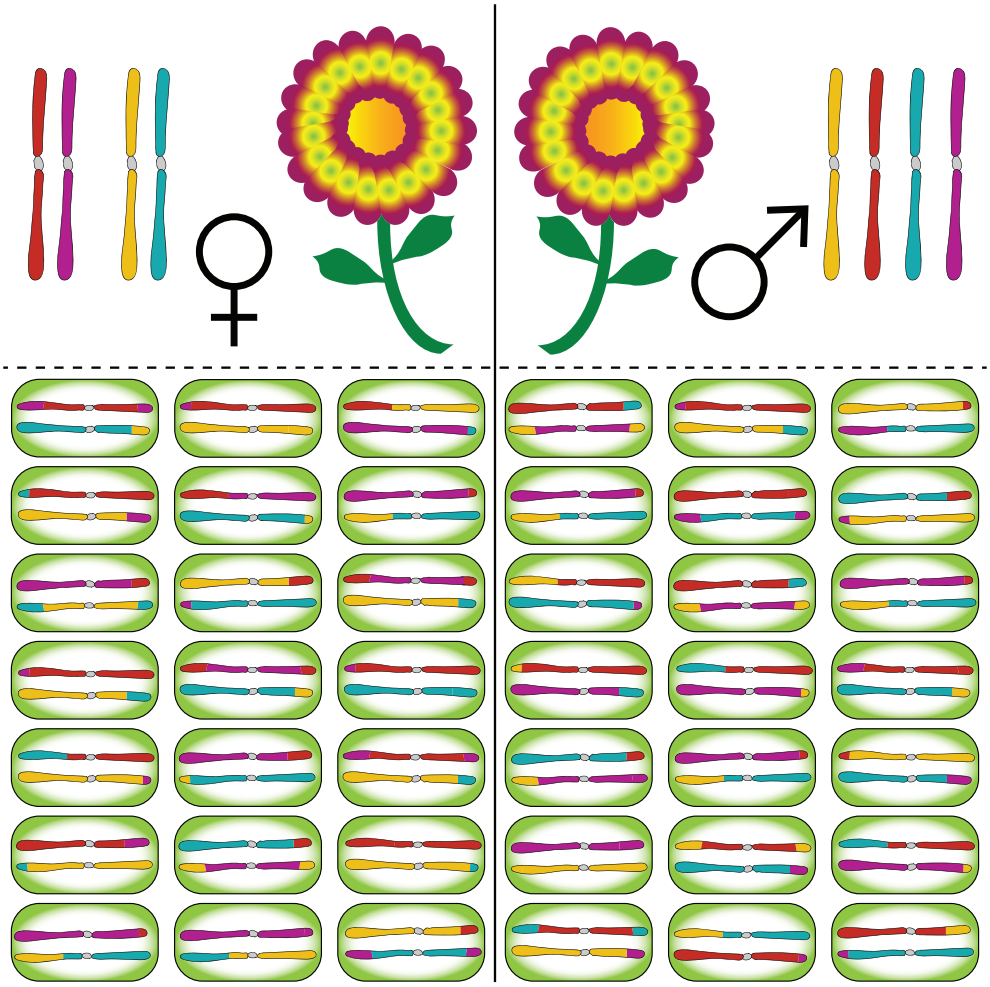


**Figure 4. Gametes of two hypothetical tetraploid parents (displaying homologues of a single chromosome).** In the maternal meiosis (left), the red and violet-coloured homologues (and the yellow and blue homologues) pair and recombine more often than would be expected by chance. Almost all gametes sampled carry red/violet and yellow/blue homologue mosaics, with far fewer red/yellow & violet/blue or red/blue & violet/yellow combinations. In this case, we say there is preferential pairing between red and violet homologues (and also yellow and blue homologues – one pairing implies the other). In the paternal meiosis (right), no such preference exists leading to balanced numbers of gametes from each of the three possible (bivalent) pairing combinations.

sugarcane (*Saccharum officinarum* × *S. spontaneum*) (Jannoo et al., 2004) and birdsfoot trefoil (*Lotus corniculatus* L.) (Fjellstrom et al., 2001). In these cases it is not always clear whether we are dealing with an autopolyploid that is gradually becoming re-diploidised, or an allopolyploid from two very closely-related species (Soltis et al., 2016). Diagnosing the mode of inheritance in polyploids (disomic / polysomic / mixosomic) is a challenging endeavour, but the task has been made considerably easier by the use of

markers. By following the inheritance of specific types of markers in a population, it is possible to detect whether there are statistically significant deviations from the expected proportions of marker alleles. These can be tested against an initial hypothesis of disomy or polysomy, usually informed by a study of the existing literature on the subject. Suppose we cross two heterozygous tetraploid plants and consider the segregation of markers in the resulting $F_1$. It is possible for us to uncover the parental origin of each of the homologues inherited in the population, and in this way reconstruct the parental gametes that contributed to each offspring (Figure 4). This gives us even greater clarity regarding the parental meiosis as we are in a position to map all the recombination events (and therefore also know between which pairing partners they occurred).

# Linkage mapping and QTL analysis – part II

We return to the topic of linkage maps and QTL analyses now that some of the more important eccentricities surrounding autopolyploids have been laid bare. At its core, the process of linkage mapping in polyploids is analogous to that of a diploid (Figure 5.a). The first step is to genotype a set of related individuals, either using a pre-defined set of markers or by sequencing the individuals directly and determining the marker set subsequently. Based on the co-segregation of markers in the population, the markers can be clustered together into linkage groups (ideally corresponding to chromosomes), ordered within these linkage groups and spaced according to the genetic distances between them (Figure 5.a).

However, polyploid linkage mapping, and in particular autopolyploid linkage mapping, adds a further level of complexity by distinguishing between homologue linkage groups and chromosomal linkage groups. The terminology often used to describe the process of classifying a marker across homologues is *phasing*, that is determining whether the segregating marker alleles are on the same homologue (for which we say markers are linked in *coupling* phase) or whether they reside on different homologues of the same chromosome (for this we say they are linked in *repulsion* phase). Phase considerations in diploids are trivial if inbred parents are used; an exception is when parents are out-breeding (so-called cross-pollinating or CP populations (Van Ooijen, 2006)), for which linkage phase must also be carefully considered (Maliepaard et al., 1997). The first main challenge of polyploid linkage mapping therefore is to assign a marker not just to a linkage group and a position, but to determine precisely how its alleles are distributed across the parental genomes in relation to all the other markers (Figure 5.b). The second

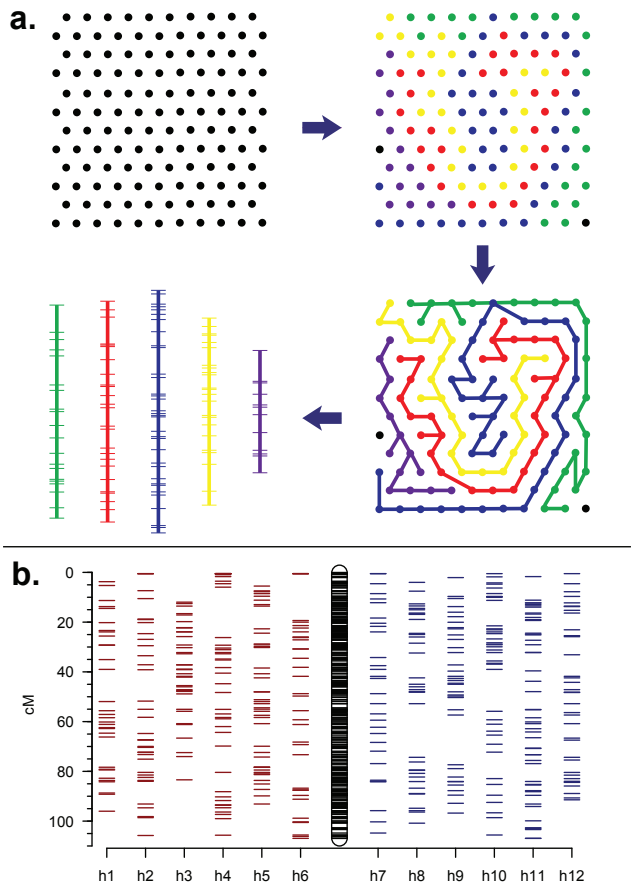



**Figure 5. Linkage mapping in polyploid species. a.** Principal stages in linkage map construction. Firstly, genotype (marker) data is generated for a population. These markers are clustered into groups (this can be at the level of both chromosome and homologue, not shown here). Markers are subsequently ordering within these clusters. In the final step, map positions are assigned to the markers, usually on a centiMorgan scale. Markers may be lost in the clustering stage due to lack of linkage, possibly due to poor data quality. Some markers may also map to the same position, as indicated by the side branches in the right-hand figure. **b.** Example of an integrated autohexaploid linkage map (centre), with phasing of the markers on the maternal homologues h1 – h6 in red, and paternal homologues h7 – h12 in blue. Marker positions in centiMorgans (cM) are shown on the y-axis.

challenge is how to best order the markers. In this thesis we do not tackle marker ordering but rely on the work of others and in particular, we apply a recent algorithm that applies multi-dimensional scaling to efficiently solve the problem (Preedy and Hackett, 2016). As with diploid CP populations, recombination frequency estimates can be of varying precision based on the informativeness of a particular marker combination (Maliepaard et al., 1997). This non-constant variance is reflected in variable LOD scores, resulting

in the need for a marker ordering algorithm that performs weighted optimisation (Stam, 1993;Preedy and Hackett, 2016).

Once a linkage map has been created it is possible to test for associations between genetic positions on the parental homologues and phenotypic traits of interest. Similar to the central challenge of linkage mapping, one of the main challenges in polyploid QTL analysis is to correctly predict the parental origin of QTL alleles (Figure 6). In the case of a QTL with a single allele of positive effect, this can usually be achieved by single-marker approaches (given sufficient marker coverage in the QTL region). However, in situations where QTL are influenced by multiple allelic variants at the same locus it may not be possible to track all positive alleles using single-marker approaches. Instead,

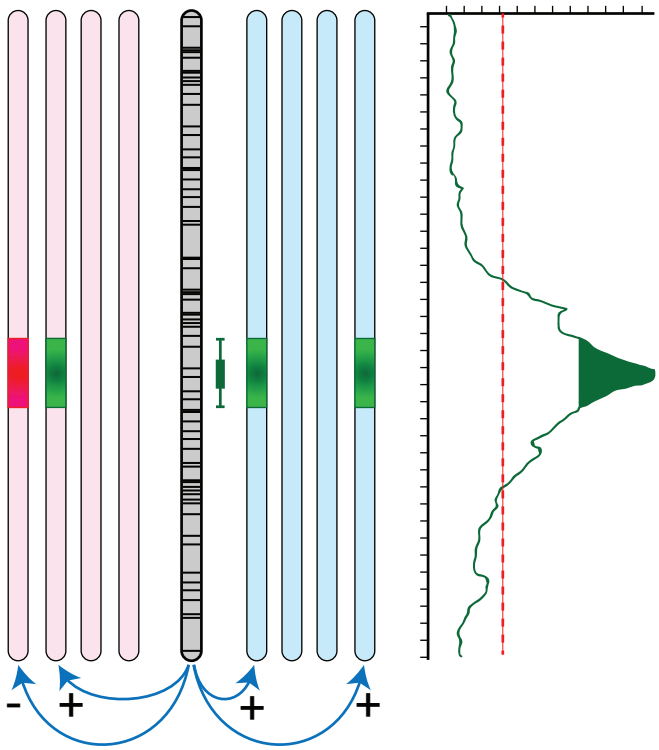



**Figure 6. Locating and phasing QTL in an autotetraploid.** As with diploids, a logarithm of odds (LOD) profile gives an indication of the strength of association between a genetic position and a particular trait. Significant associations are usually taken as those that exceed a certain threshold, shown here as a dashed red line. QTL positions are often displayed beside a linkage map with various support intervals (e.g. a LOD-1 and LOD-2 support interval) using software such as MapChart (Voorrips, 2002). However, in a polyploid we are also interested in the configuration of a QTL. In this example, two alleles with positive (+) effect originated from parent 2, and both a positive and negative (-) allele originated from parent 1. In this way, offspring carrying optimal combinations of alleles, potentially at multiple loci, can be identified.

QTL methods that move away from single marker information to instead use multi-locus identity-by-descent (IBD) probabilities (Hackett et al., 2013;Zheng et al., 2016) may provide additional diagnostic power (Kempthorne, 1957;Hackett et al., 2014).

# The polyploid revolution

Up until relatively recently, breeding in polyploid crops has relied completely on phenotypic selection. For simple traits this may well have been sufficient, but the effectiveness of phenotypic selection for complex traits like yield is questionable (Jansky, 2009). There is increased interest among polyploid breeders in the use of molecular and genomic tools to assist with their breeding programs. The reasons why this has not so far happened were technological, methodological and financial. The technology to be able to efficiently and cheaply dissect polyploid genomes in detail is only now becoming available. The methodology to interpret this data and the software tools to run these analyses (such as described in this thesis) are following suit. We therefore stand at the dawn of a revolution in polyploid genomics and breeding, brought about by a favourable combination of these three factors.

# Layout of this thesis

The central aim of this thesis is to develop and test methods to perform genetic mapping in polyploid populations. This can constitute a significant computational challenge, given the complexity and sheer size of modern genotyping datasets. We begin therefore with a review of the available software tools for the genetic analysis of polyploids in **Chapter 2**. This chapter explores the current options for assigning marker dosages and assembling haplotypes, performing linkage mapping and QTL analyses, genome-wide association studies and genomic selection, as well as looking at the availability of physical maps and tools for simulating polyploid populations.

The rest of the thesis can be broken down into two main subjects – linkage mapping and QTL analysis, with linkage mapping methods developed in the first half of the thesis, and QTL mapping in the second half. However, we also are interested in understanding the meiotic behaviour of polyploids (which linkage mapping can help to clarify). In **Chapter 3** we generate homologue-specific linkage maps in an autotetraploid potato population and use them to investigate the frequency of double reduction events across the population. A denser set of linkage maps, consisting of all marker segregation types,

is integrated in **Chapter 4** using information from bridging markers (markers with alleles which reside on multiple homologues). We also perform a simulation study to test the effects of double reduction and partial preferential chromosome pairing on the accuracy of recombination frequency estimates upon which our maps are based.

In **Chapter 5** we go one step further by developing a high-density SNP map for a tetraploid rose population. We perform an in-depth exploration of the meiotic pairing and recombination behaviour and in the process, develop a method to correct for mixosomic inheritance in the creation of autotetraploid linkage maps. These initial chapters culminate in **Chapter 6**, where we present a new software package for the creation of integrated and phased linkage maps in polyploid species, called *polymapR*.

In **Chapter 7**, we apply our mapping methodology to a large population of the popular autohexaploid cut-flower *Chrysanthemum* × *morifolium*. The integrated linkage map and marker phase information is used to detect QTL for a number of flower-related traits. **Chapter 8** explores the power of QTL mapping methods in autopolyploids, comparing models which allow for double reduction with those that do not. In **Chapter 9**, these methods are used to perform a QTL analysis in tetraploid rose, exploring the genetic architecture of a number of economically-important physiological traits measured across multiple growing environments. In **Chapter 10**, we present *polyqtlR*, a novel software package for performing QTL analyses in autopolyploid populations, building on the work of the previous chapters. Finally, the thesis is rounded off with a discussion in **Chapter 11** on how the findings presented in this thesis can contribute to improvements in polyploid breeding as well deepening our understanding of these fascinating species.

# Chapter 2

# Tools for genetic studies in experimental populations of polyploids

Peter M. Bourke[1], Roeland E. Voorrips[1], Richard G. F. Visser[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

*Published (with modifications) as* Bourke, P.M., Voorrips, R.E., Visser, R.G.F. and Maliepaard, C. (2018). "Tools for genetic studies in experimental populations of polyploids", **Frontiers in Plant Science, 9:513. doi: 10.3389/fpls.2018.00513**

# Abstract

Polyploid organisms carry more than two copies of each chromosome, a condition rarely tolerated in animals but which occurs relatively frequently in the plant kingdom. One of the principal challenges faced by polyploid organisms is to evolve stable meiotic mechanisms to faithfully transmit genetic information to the next generation upon which the study of inheritance is based. In this review we look at the tools available to the research community to better understand polyploid inheritance, many of which have only recently been developed. Most of these tools are intended for experimental populations (rather than natural populations), facilitating genomics-assisted crop improvement and plant breeding. This is hardly surprising given that a large proportion of domesticated plant species are polyploid. The current polyploid analytic toolbox includes software for assigning marker genotypes (and in particular, estimating the dosage of marker alleles in the heterozygous condition), establishing chromosome-scale linkage phase among marker alleles, constructing (short-range) haplotypes, generating linkage maps, performing genome-wide association studies (GWAS) and quantitative trait locus (QTL) analyses, and simulating polyploid populations. These tools can also help elucidate the mode of inheritance (disomic, polysomic or a mixture of both as in segmental allopolyploids) or reveal whether double reduction and multivalent chromosomal pairing occur. An increasing number of polyploids (or associated diploids) are being sequenced, leading to publicly-available reference genome assemblies. Much work remains in order to keep pace with developments in genomic technologies. However, such technologies also offer the promise of understanding polyploid genomes at a level which hitherto has remained elusive.



# Key words

Polyploid genetics, polyploid software tools, autopolyploid, allopolyploid, segmental allopolyploid

# Introduction

One of the most fundamental descriptions of any organism is its ploidy level and chromosome number, generally written in the form 2n = 2x = 10 (here, for the ubiquitous model plant species *Arabidopsis thaliana* L.). Plant scientists in particular will be familiar with this representation of the chromosomal constitution of the sporophyte generation (*i.e.* the adult plant). The second term in this seemingly simple equation describes the normal complement of chromosomal copies possessed by a member of that species, which is generally 2x ("two times") for diploids. Species where this number exceeds two are collectively referred to as polyploids. Not unexpectedly, each polyploid individual is the product of the fusion of gametes from two parents, just like their diploid counterparts. In other words, polyploids can also be defined as individuals derived from non-haploid gametes (in the case of triploids derived from diploid × tetraploid crosses, only one gamete satisfies this condition). The transmission of non-haploid gametes is one of the main "complexifying" features of polyploidy, leading to a whole range of implications for the genetic analysis of these "hopeful monsters" (Goldschmidt, 1933).

The ongoing genomics revolution can be seen as a rising tide which has also lifted the polyploid genetics boat, although not quite to the same level as for diploids. Most genetic advances are made in model organisms, among which self-fertilising diploid species predominate. It is therefore not surprising that most tools and techniques for molecular-genetic studies are specific to diploids. However, polyploid species are particularly important to mankind in the provision of food, fuel, feed and fibre (not to mention "flowers", if ornamental plant species are also included), making the genetic analysis of polyploid species an important avenue of research for crop improvement.

Although a collective term such as "polyploidy" has its uses, it tends to obscure some fundamental differences between its members. For example, polyploids are generally subdivided into autopolyploids and allopolyploids (Kihara and Ono, 1926). Autopolyploids arise through genomic duplication within a single species, generally through the production of unreduced gametes (Harlan and De Wet, 1975) and exhibit polysomic inheritance, meaning pairing and recombination can occur between all homologous copies of each chromosome during meiosis. One of the most well-studied examples is autotetraploid potato (*Solanum tuberosum* L.). Allopolyploids, on the other hand, are the product of genomic duplication between species (usually through hybridisation involving unreduced gametes (Harlan and De Wet, 1975)) and display disomic inheritance, where more-related chromosome copies ("homologues") may pair and recombine during meiosis, whilst less-related chromosome copies ("homoeologues", also spelled "homeologues" (Glover et al., 2016)) do not. Among allopolyploids, allohexaploid wheat (*Triticum aestivum* L.) is probably the most well-

studied. If pairing and recombination between homoeologues occurs to a limited extent, the species may be referred to as "segmental allopolyploid" (Stebbins, 1947), traditionally deemed to have arisen from hybridisation between very closely-related species (Stebbins, 1947;Chester et al., 2012) but which may also be the result of partially-diploidised autopolyploidy (Soltis et al., 2016). In many cases, a species cannot be clearly designated as one type or another, leading to uncertainty or debate on the subject (Barker et al., 2016;Doyle and Sherman-Broyles, 2016). From the perspective of genetics and inheritance, allopolyploids behave much like diploid species and therefore many of the tools developed for diploids can be directly applied. The main challenge that faces allopolyploid geneticists is in distinguishing between homoeologous gene copies carried by sub-genomes within an individual (Kaur et al., 2012;van Dijk et al., 2012;Rothfels et al., 2017). Autopolyploids (and segmental allopolyploids) do not behave like diploids, and are therefore in most need of specialised methods and tools for subsequent genetic studies. In this review we focus primarily on the availability of tools and resources amenable to polysomic (and "mixosomic" (Soltis et al., 2016)) species, with less emphasis on allopolyploid-specific solutions. Although the development of novel methodologies for the genetic analysis of polyploids are interesting, without translation into a software tool for use by the research community they remain purely conceptual and with limited impact. We therefore try to limit our attention to the tools currently available rather than cataloguing descriptions of unimplemented methods.

Experimental populations, in use since Mendel's ground-breaking work (Mendel, 1866), are traditionally derived from a controlled cross between two parental lines of interest (either directly studying the $F_1$ or some later generation). We use the term here to distinguish our subject matter from "wild" or "natural" populations, which would necessitate sampling individuals from an extant population in the wild. Quantitative genetics, particularly the genetics of human pathology, has greatly benefitted from the use of large panels of individuals to perform so-called "genome-wide association studies" (GWAS). The use of such panels offers to complement the experimental toolbox of polyploid geneticists as well, and although perhaps not strictly-speaking an "experimental" population, we consider them relevant to the current discussion.

Here, we review the current possibilities for polyploid genotyping, including the scoring of marker dosage (allele counts) and generation of haplotypes, the availability of reference sequences, the possibilities for linkage mapping, quantitative trait locus (QTL) analysis and GWAS, as well as tools to simulate polyploid organisms for *in silico* studies. We also reflect on current and future developments, and the tools that will be needed to keep pace with the innovations we are witnessing in genomic technologies.

# Polyploid genotyping

One of the most crucial aspects in the study of polyploid genetics is the generation of accurate genotypic data. However it is also fraught with difficulties, not least the detection of multiple loci when only a single locus is targeted (Mason, 2015;Limborg et al., 2016). Various technologies exist, with almost all current applications aimed at identifying single nucleotide polymorphisms (SNPs). Although many genomic "service-providers" (*e.g.* companies or institutes that offer DNA sequencing) have their own tools to analyse and interpret raw data, these tools are not always suitable for use with polyploid datasets.

## Genotyping technologies

Although gel-based marker technologies continue to be used and have certain advantages (*e.g.* low costs associated with small marker numbers, requiring only basic laboratory facilities, multi-allelism *etc.*), most studies now rely on SNP markers for genotyping due to their great abundance over the genome, their high-throughput capacity and their low cost per data point. Targeted genotyping such as SNP arrays (a.k.a. "SNP chips") rely on previously-identified and selected polymorphisms, usually identified from a panel of individuals chosen to represent the gene pool under investigation. In contrast, untargeted genotyping generally uses direct sequencing of individuals, albeit after some procedure to reduce the amount of DNA to be sequenced (*e.g.* by exome sequencing (Ng et al., 2009) or target enrichment (Mamanova et al., 2010)). The disadvantages of targeted approaches have been well explored (particularly regarding ascertainment bias, where the set of targeted SNPs on an array poorly represents the diversity in the samples under investigation due to biased methods of SNP discovery) (Albrechtsen et al., 2010;Moragues et al., 2010;Didion et al., 2012;Lachance and Tishkoff, 2013), although there are advantages and disadvantages to both methods (Mason et al., 2017). Apart from costs, differences exist in the ease of data analysis following genotyping, with sequencing data requiring greater curation and bioinformatics skills (Spindel et al., 2013;Bajgain et al., 2016) as well as potentially containing more erroneous and missing data (Spindel et al., 2013;Jones et al., 2017). In polyploids, SNP arrays have been developed in numerous species, which include both autopolyploid (or predominantly polysomic polyploids) and allopolyploid species. Examples of the former include alfalfa (Li et al., 2014b), chrysanthemum (van Geest et al., 2017c), potato (Hamilton et al., 2011;Felcher et al., 2012;Vos et al., 2015), rose (Koning-Boucoiran et al., 2015) and sour cherry (Peace et al., 2012). Examples of allopolyploid SNP arrays include cotton (Hulse-Kemp et al., 2015), oat (Tinker et al., 2014), oilseed rape (Dalton-Morgan et al., 2014;Clarke et al., 2016), peanut (Pandey et al., 2017), strawberry (Bassil et al., 2015) and wheat (Akhunov et al., 2009;Cavanagh et al., 2013;Wang et al., 2014b;Winfield et

al., 2016). Untargeted approaches such as genotyping-by-sequencing have also been applied, for example in autopolyploids such as alfalfa (Zhang et al., 2015;Yu et al., 2017), blueberry (McCallum et al., 2016), bluestem prairie grass (*Andropogon gerardii*) (McAllister and Miller, 2016), cocksfoot (*Dactylis glomerata*) (Bushman et al., 2016), potato (Uitdewilligen et al., 2013;Sverrisdóttir et al., 2017), sugarcane (Balsalobre et al., 2017;Yang et al., 2017b) and sweet potato (Shirasawa et al., 2017), and in allopolyploids such as coffee (Moncada et al., 2016), cotton (Islam et al., 2015;Reddy et al., 2017), intermediate wheatgrass (*Thinopyrum intermedium*) (Kantarski et al., 2017), oat (Chaffin et al., 2016), prairie cordgrass (*Spartina pectinata*) (Crawford et al., 2016), shepherd's purse (*Capsella bursa-pastoris*) (Cornille et al., 2016), wheat (Poland et al., 2012;Edae et al., 2015) and zoysiagrass (*Zoysia japonica*) (McCamy et al., 2018) (noting that the precise classification of some of these species as auto- or allopolyploids has yet to be conclusively determined). Whatever the technology used, it is clear that we are currently witnessing an explosion of interest in polyploid genomics. However, the critical issue of how to make sense of this data remains, starting with the assignment of marker dosage, a.k.a. "genotype calling".

## Assignment of dosage

One of the key distinguishing features of polysomic polyploidy is the fact that there are multiple heterozygous conditions possible in genotyping data. We use the term marker "dosage" to denote the minor allele count of a marker; a species of ploidy $q$ possesses $q + 1$ distinct dosage classes in the range 0 to $q$ (Figure 1). Of course the concept of marker dosage could also be used in diploid species, but coding systems such as the lm x ll / nn x np / hk x hk system (Van Ooijen, 2006) predominate. Marker dosage is generally understood to apply to bi-allelic markers (such as single SNPs), although it is conceivable to score marker dosage at multi-allelic loci. If marker dosage cannot be accurately assessed, genotypes would likely have to be dominantly-scored (*i.e.* all heterozygous classes would be grouped with one of the homozygous classes), resulting in a loss of information (Piepho and Koch, 2000).

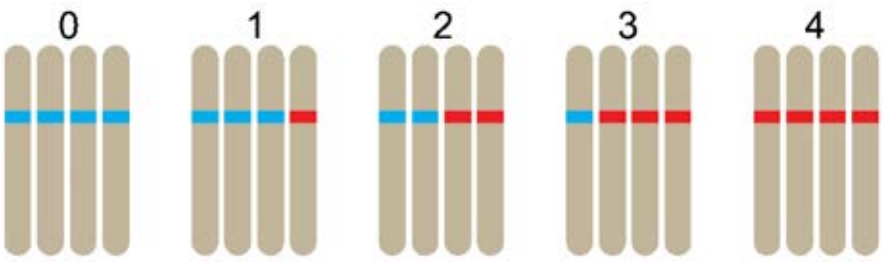



**Figure 1. Dosage possibilities in an autotetraploid.** In a tetraploid, five distinct dosages are possible at a bi-allelic marker positions, ranging from 0 copies of the alternative allele through to 4 copies. Here, the alternative allele is coloured red, with the reference allele coloured blue.

All available dosage-calling tools rely on a population in order to determine marker dosage. In other words, calibration between the various dosage classes is performed across the population (for which we are not implying any degree of relatedness in the population other than coming from the same species). All current tools are designed to process genotyping data from SNP arrays, in other words using the relative strength of two allele-specific (fluorescent) signals to assign a discrete dosage value. With increasing interest in genotyping-by-sequencing (GBS) data, we anticipate that tools which use read-counts of potentially multiple SNPs (or multi-SNP haplotypes) will soon be developed, although these have yet to appear. One of the current challenges under investigation regarding GBS-based genotype calling is the accurate determination of dosage (Kim et al., 2016), which may require relatively deep sequencing (*e.g.* 60-80x coverage estimated in autotetraploid potato (Uitdewilligen et al., 2013)).

Returning to the SNP array-based tools, the two main service providers for high-density SNP arrays, Illumina and Affymetrix, both offer proprietary software solutions for analysing polyploid datasets. Affymetrix's Power Tools and Illumina's GenomeStudio (with its Polyploid Genotyping Module) were developed with both diploid and polyploid datasets in mind. However, there have also been a number of genotyping tools that have been put into the public domain. One of the first of these to be released was fitTetra (Voorrips et al., 2011), a freely-available R package (R Core Team, 2016) designed to assign genotypes to autotetraploids that were genotyped on either Illumina's Infinium or Affymetrix's Axiom arrays. fitTetra fits mixture models to bi-allelic SNP intensity ratios either under the constraint of Hardy-Weinberg equilibrium within the population, or as an unconstrained fit, using an expectation-maximisation (EM) algorithm in fitting. This can have the drawback of requiring significant computational resources for high-density marker datasets, although it is automated and can therefore process large datasets in a single run. The original release was specific to tetraploid data only. However, an updated version (fitPoly) can process genotyping data of all ploidy levels and is soon to be made available as a separate R package (R. E. Voorrips, personal communication). The SuperMASSA application (Serang et al., 2012) can also process data from all ploidy levels (as it was initially developed to dosage-score sugarcane data, notorious for its cytogenetic complexity) and is currently hosted online by the Statistical Genetics Laboratory in the University of São Paulo, Brazil. One of the interesting features of SuperMASSA is that prior knowledge of the exact ploidy level is not needed (useful for a crop like sugarcane). Instead, the genotype configuration which maximises the posterior probability across all specified ploidy levels is chosen. In practice, most researchers will already know the ploidy of their samples (although aneuploid progeny in some species may occur) and can constrain the model search. A major draw-back of the current implementation is that markers are analysed one-by-one, and results need to

be copied from the webpage each time (there is currently no downloadable version of the software available). For high-density datasets with tens of thousands of markers, this is clearly impractical.

The R package polysegRatioMM (Baker et al., 2010) generates marker dosages for dominantly-scored markers using the JAGS software (Plummer, 2003) for Markov Chain Monte Carlo (MCMC) generation. Fully polysomic behaviour is assumed, and segregation ratios of marker data are used to derive the most likely parental scores. Although able to process data from all even ploidy levels, the software only considers a subset of marker types (marker that are nulliplex in one parent or simplex in both parents). Nowadays, there is a move away from dominantly-scored markers to co-dominant marker technologies like SNPs, and parental samples are usually included in multiple replicates (and so can be genotyped directly with offspring, rather than imputed from the offspring). The package is therefore of questionable use for modern genotyping datasets. An unrelated R package, beadarrayMSV (Gidskehaug et al., 2010), was developed to handle Illumina Infinium SNP array data from "diploidising" tetraploid species such as the Atlantic salmon. The software was designed to score markers which target multiple loci (so-called multi-site variants, or MSVs), as well as single-locus markers displaying disomic inheritance. In a comparison with fitTetra, beadarrayMSV was unable to accurately genotype autotetraploid data from potato, although conversely fitTetra performed poorly on salmon data (Voorrips et al., 2011). This demonstrates that appropriate software is needed for specific situations (indeed, in many cases specific scenarios have motivated the development of specialised software).

Having prior knowledge about the expected meiotic behaviour of the species is always advantageous when it comes to analysing any polyploid data. This is especially true for the latest dosage-calling software to be released, the ClusterCall package for R (Schmitz Carley et al., 2017). Here, prior knowledge of the meiotic behaviour of the species is required, since the expected segregation ratios of an $F_1$ autotetraploid population are used to assign dosage scores to the clusters identified through hierarchical clustering. In well-behaved autotetraploids such as potato (Swaminathan and Howard, 1953;Bourke et al., 2015) this is arguably not a problem (as long as skewed segregation does not occur), and indeed can lead to increased accuracy in genotype calling (Schmitz Carley et al., 2017). However, in less well-characterised species such as leek, alfalfa or many ornamental species, the precise meiotic behaviour may not always follow the expected tetrasomic model, causing potential problems with fitting. The authors are aware of this and suggest that alternatives like fitTetra or SuperMASSA be used in circumstances where a tetrasomic model no longer holds. Unfortunately, such prior knowledge is not always available before genotyping takes place – meiotic behaviour can even differ between

individuals of a species that was thought to display meiotic homogeneity (*e.g.* complete tetrasomy) (Bourke et al., 2017).

## Haplotype assembly

Although bi-allelic SNP markers have many practical advantages, they carry less inheritance information than multi-allelic markers. Crop researchers and breeders often wish to develop a simple diagnostic marker test for a trait of interest. Unfortunately, the chances of having a single SNP in complete linkage disequilibrium with a favourable or causative allele of a gene of interest is very small. Markers which have been found to uniquely "tag" a favourable allele in one population may not do so in another. For more than a decade, the increased power of haplotype-based associations have been known and reported in human genetic studies (Zhang et al., 2002;de Bakker et al., 2005), with the term "haplotype" denoting a unique stretch of sequence. Translating haplotyping approaches from diploid to polyploid species has been a non-trivial exercise, requiring novel algorithms to handle the overwhelming range of possibilities that can arise (especially when allowing for sequencing errors and (possible) recombinations). Multi-SNP haplotypes can be assembled from single dosage-scored SNPs (originating from SNP array data), although haplotypes are more commonly generated using overlapping sequence reads (Figure 2). A number of different polyploid haplotyping tools (for sequence reads) have been developed in recent years, including polyHap (Su et al., 2008), SATlotyper (Neigenfind et al., 2008), HapCompass (Aguiar and Istrail, 2013), HapTree (Berger et al., 2014), SDhaP (Das and Vikalo, 2015), SHEsisplus (Shen et al., 2016) and TriPoly (Motazedi et al., 2017a). Three of these tools (HapCompass, HapTree and SDhaP) were recently compared and evaluated over a range of different simulated read depths, ploidy levels and insert sizes for paired-end reads (Motazedi et al., 2017b). The authors found that each of these software programs had particular advantages, for example HapTree was found to produce more accurate haplotypes for triploid and tetraploid data, whilst HapCompass performed best at higher ploidies (6x and higher) (Motazedi et al., 2017b). Both SHEsisplus and TriPoly have yet to be independently tested. For allopolyploid species, the user-friendly Haplotag software has been designed to identify both single SNPs and multi-SNP haplotypes from GBS data (Tinker et al., 2016). An interesting feature is the use of a simple "heterozygosity filter" that excludes haplotypes with higher than expected heterozygosity across a population (suggesting paralogous loci). Currently however, data from outcrossing or autopolyploid species is not suitable for this software.

The input data of haplotyping software can be grouped into two types. Individual SNP genotyping data (with a known marker order) was used by the first wave of polyploid haplotyping implementations such as polyHap and SATlotyper. More recently,

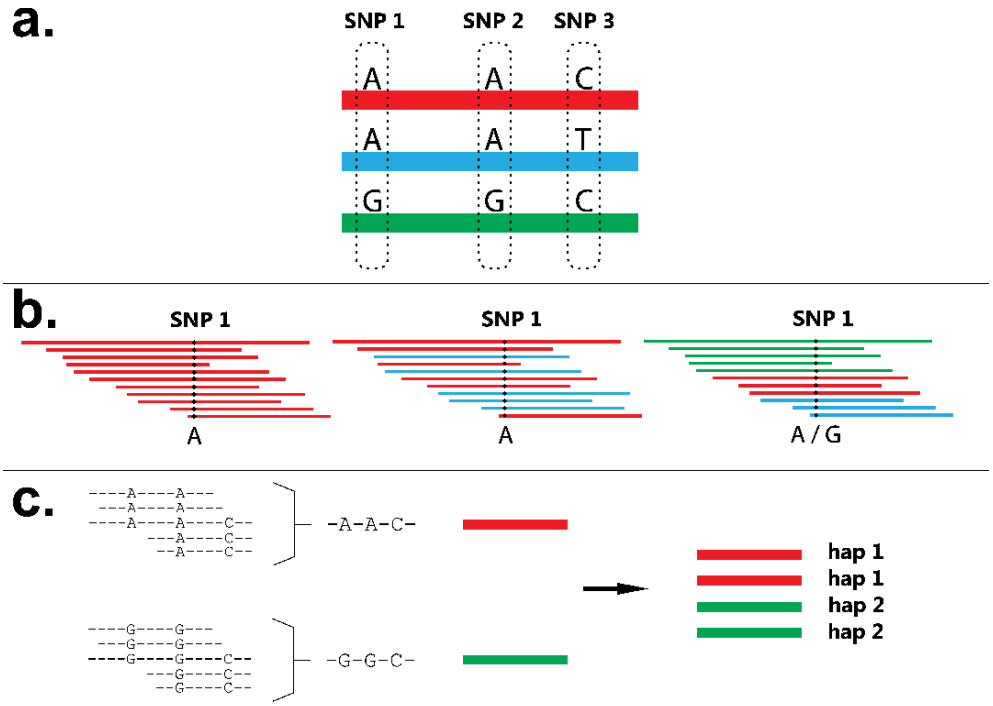



**Figure 2. Generation of multi-SNP haplotypes. a.** In this example, three possible haplotypes exist spanning polymorphic positions SNP 1, 2 and 3. **b.** Single-SNP genotyping cannot distinguish between the "A" allele originating from different haplotypes, combining them into a single allele as illustrated in the second SNP call. **c.** In a haplotyping approach, overlapping reads are used to re-assemble and phase single SNP genotypes. Here, the known ploidy level of the species (4x) is used to impute the dosage of the two haplotypes identified in this individual, given a 1:1 ratio between the assembled haplotype read-depths.

haplotyping tools use sequence reads as their input, although some pre-processing is required: reads must first be aligned followed by extraction of their SNPs (*i.e.* masking of non-polymorphic sites) to generate a SNP-fragment matrix with individual reads as rows and SNP positions as columns (as described for HapCompass (Aguiar and Istrail, 2013)). In other words, all haplotyping tools (apart perhaps from Haplotag (Tinker et al., 2016)) require that users possess a certain level of bioinformatics skills. In terms of applications and usage, there appears to be some hesitation among the polyploid research community to take up these tools (Figure 3), perhaps because of the required bioinformatics qualifications. Although we expect polyploid haplotypes to become increasingly used in the future, the development of user-friendly and computationally-efficient tools is first needed before haplotype-based genotypes become truly mainstream.

One interesting development is the application of haplotyping to whole genome assemblies (as opposed to genotyping a population). This has recently been attempted in the tuberous hexaploid crop sweet potato (*Ipomoea batatas*) (Yang et al., 2017a). The authors first produced a consensus assembly to which reads were re-mapped for variant calling, followed by a phasing algorithm which resolved the six haplotypes of the sequenced cultivar for about 30% of the assembly (Yang et al., 2017a). Ultimately, about half of the assembled genome could be haplotype-resolved. Future sequencing (or re-sequencing) efforts in polyploid species should produce more phased genomes, which will no doubt be useful for haplotyping applications (for example in validating predicted haplotypes).

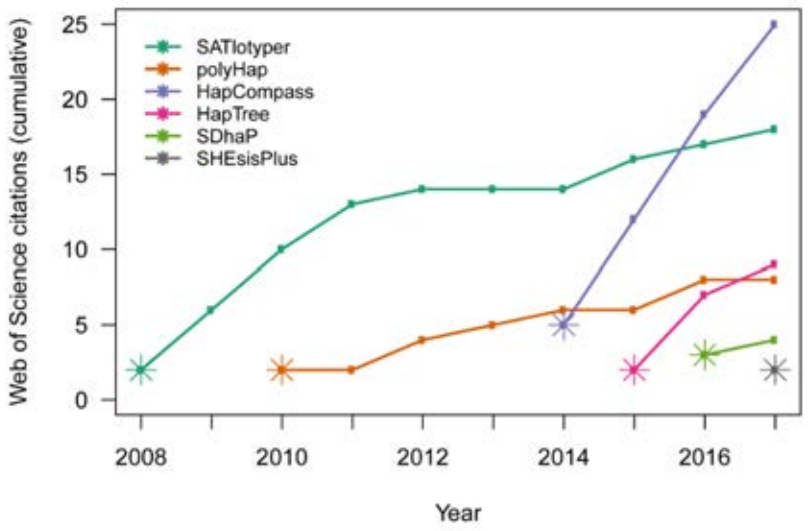



**Figure 3. Cumulative number of Web of Science citations (core collection) by January 1st 2018 for the most popular polyploid haplotyping tools.** Year of first citation is highlighted by a star.

# Physical maps

Arguably, one of the most important "tools" in current genomics studies is access to a high-quality reference genome assembly. Species for which a reference genome assembly exists have even been classified as "model organisms" (Seeb et al., 2011), such is the importance and impact a genome can bring to research on that species. Without a reference sequence available, the scope of genomic research remains limited. For example, genome-wide association studies rely on knowledge of the relative position of SNP markers (usually on a physical map), and many sequencing applications rely on a reference assembly on which to map reads. A reference genome also facilitates the development of molecular markers (*e.g.* primer development), the comparison of results

between different genetic studies (by providing a single reference map), as well as allowing comparisons of specific sequences such as genes, enabling prediction of gene function across related species.

Polyploid genomes are by definition more complex than diploid genomes, having multiple copies of each homologous chromosome. Many polyploid species are also outbreeding, leading to increased heterozygosity which is problematic in *de novo* assemblies and necessitates specialised approaches (Kajitani et al., 2014). The most common solution until now has been to sequence a representative diploid species. For example in highly-heterozygous autotetraploid potato, a completely homozygous doubled monoploid (*Solanum tuberosum* group *Phureja* DM1-3) was sequenced (Potato Genome Sequencing Consortium, 2011) which still represents the primary reference sequence today (http://solanaceae.plantbiology.msu.edu/). In the case of allopolyploids, multiple diploid progenitor species are often sequenced instead (*e.g.* peanut (Bertioli et al., 2016)). The emergence of the pan-genome concept, originally proposed for microbial species (Tettelin et al., 2005), has interesting implications for how highly-heterozygous polyploid genomes will be presented in future. We have already mentioned the arrival of phased genomics with the sweet potato genome, which aimed to generate six chromosome-length phased assemblies for each of its 15 chromosomes (Yang et al., 2017a). In future, both pan-genomes and phased genomes are likely to play a bigger role in polyploid reference genomics. Example of polyploid species that have so far been "sequenced" are listed in Table 1. This is by no means an exhaustive list, nor does it describe all developments for the listed species. For example, the sequence of allotetraploid *Coffea arabica* (which accounts for roughly 70% of all coffee production) has recently been assembled, with a draft assembly (*Coffea arabica* UCDv0.5) available on the Phytozome database (www.phytozome.net). What Table 1 highlights is that at the time of writing, there were already a wide range of polyploid crop species that have well-developed genomic resources, despite the fact that in many cases these are from closely-related or progenitor diploid species. In time, just like for coffee, we predict that direct sequencing of polyploid species themselves will gradually replace the haploidised reference sequences in importance and application, leading to more insights of direct relevance to polyploids.

# Linkage maps

Although the first genetic linkage map was developed over a hundred years ago (Sturtevant, 1913), their use in genetic and genomic studies has persisted into the "next-generation" era. This can be attributed to a number of factors. A linkage map is a description of the recombination landscape within a species, usually from a single experimental cross of interest. For breeders, knowledge of genetic distance is arguably

more important than physical distance, as it reflects the recombination frequencies in inheritance studies as well as describing the extent of linkage drag around loci of interest. Many software for performing quantitative trait locus (QTL) analysis require linkage maps of the markers, not physical maps. This is because co-inheritance of markers and phenotypes within a population are assumed to be coupled – a physical map gives less precise information about the co-inheritance of markers than a linkage map does since physical distances do not directly translate to recombination frequencies (particularly in the pericentromeric regions). Another reason why linkage maps continue to be developed is that they are often the first genomic representation of a species, upon which more advanced representations can be built. They provide useful long-range linkage information over the whole chromosome, which is often missing from assemblies of short sequence reads. This fact has been repeatedly exploited in efforts at connecting and correctly orientating scaffolds during genome assembly projects (Bartholomé et al., 2015;Fierst, 2015).

As mentioned in the Introduction, polyploids can be divided into disomic or polysomic species, with the additional possibility of a mixture of both inheritance types in the case of segmental allopolyploids. Many linkage maps in polyploids have been based exclusively on 1:1 segregating markers, also known as simplex markers (because the segregating allele is in simplex condition (one copy) in one of the parents only). These markers possess a number of advantages over other marker segregation types, but also some distinct disadvantages. In their favour, coupling-phase simplex markers in polyploid species behave just like they would in diploid species, regardless of the mode of inheritance involved (repulsion-phase recombination frequency estimates are not invariant across ploidy levels or modes of inheritance, but exert less influence on map construction due to lower LOD scores). The advantage of this is clear: in unexplored polyploid species for which the mode of inheritance is uncertain, simplex markers allow an "assumption-free" linkage map to be created, following which the mode of inheritance can be further explored. The only exception to this is if double reduction occurs, *i.e.* when a segment of a single chromosome gets transmitted with its sister chromatid copy to an offspring, a consequence of multivalent pairing and a particular sequence of chromatid segregation and division during meiosis (Haldane, 1930;Mather, 1935). Double reduction occurs randomly in polysomic species and only introduces a small bias into recombination frequency estimates (Bourke et al., 2015). This means that, ignoring the possible influence of double reduction, diploid mapping software can generally be used for simplex marker sets at any ploidy level and for any type of meiotic pairing behaviour (Figure 4), opening up a very wide range of diploid-specific software options (Cheema and Dicks, 2009).

**Table 1. Some examples of currently-available reference sequences for polyploid species.**

| **AUTOPOLYPLOIDS** | | | |
|---|---|---|---|
| **Target species** | **Sequenced species (ploidy)** | **Genome browser** | **References** |
| Alfalfa, *Medicago sativa* (4x) | *Medicago truncatula* (2x) | medicagogenome.org \| plants.ensembl.org | (Young et al., 2011) (Tang et al., 2014) |
| Kiwifruit, *Actinidia chinensis* (6x) | *Actinidia chinensis* (2x) | bdg.hfut.edu.cn/kir \| bioinfo.bti.cornell.edu/cgi-bin/kiwi/home.cgi | (Huang et al., 2013) |
| Potato, *Solanum tuberosum* (4x) | *Solanum tuberosum* (2x) | solanaceae.plantbiology.msu.edu \| plants.ensembl.org | (Potato Genome Sequencing Consortium, 2011) |
| Sweet potato, *Ipomoea batatus* (6x) | *Ipomoea batatus* (6x) | public-genomes-ngs.molgen.mpg.de/SweetPotato \| ipomoea-genome.org | (Yang et al., 2017a) |
| **ALLOPOLYPLOIDS** | | | |
| Banana, *Musa acuminata* (3x) | *Musa acuminata* (2x) | banana-genome-hub.southgreen.fr \| plants.ensembl.org | (D'Hont et al., 2012) |
| Coffee, *Coffea arabica* (4x) | *Coffea canephora* (2x) | coffee-genome.org | (Denoeud et al., 2014) |
| Cotton, *Gossypium hirsutum* (4x) | *Gossypium arboretum* (2x) *Gossypium raimondii* (2x) | cottongen.org | (Li et al., 2014a) (Wang et al., 2012) |
| Oilseed rape, *Brassica napus* (4x) | *Brassica napus* (4x) | genoscope.cns.fr/brassicanapus \| plants.ensembl.org | (Chalhoub et al., 2014) |
| Peanut, *Arachis hypogaea* (4x) | *Arachis duranensis* (2x) *Arachis ipaensis* (2x) | peanutbase.org | (Bertioli et al., 2016) |
| Quinoa, *Chenopodium quinoa* (4x) | *Chenopodium quinoa* (4x) | cbrc.kaust.edu.sa/chenopodiumdb | (Jarvis et al., 2017) |
| Strawberry, *Fragaria* × *ananassa* (8x) | *Fragaria vesca* (2x) | rosaceae.org | (Shulaev et al., 2011) |
| Wheat, *Triticum aestivum* (6x) | *Triticum aestivum* (6x) | wheat-urgi.versailles.inra.fr \| plants.ensembl.org | (International Wheat Genome Sequencing Consortium, 2014) |

However, simplex marker sets have some limitations. Firstly, in selecting only simplex markers, a large proportion of markers with different segregation patterns are not used. This usually reduces the map coverage (while increasing the per-marker costs of the final set of mapped markers). More importantly, simplex markers give limited information about linkage in repulsion phase, particularly at higher ploidy levels (van Geest et al., 2017a). This means that homologue-specific maps can be produced, but they are unlikely

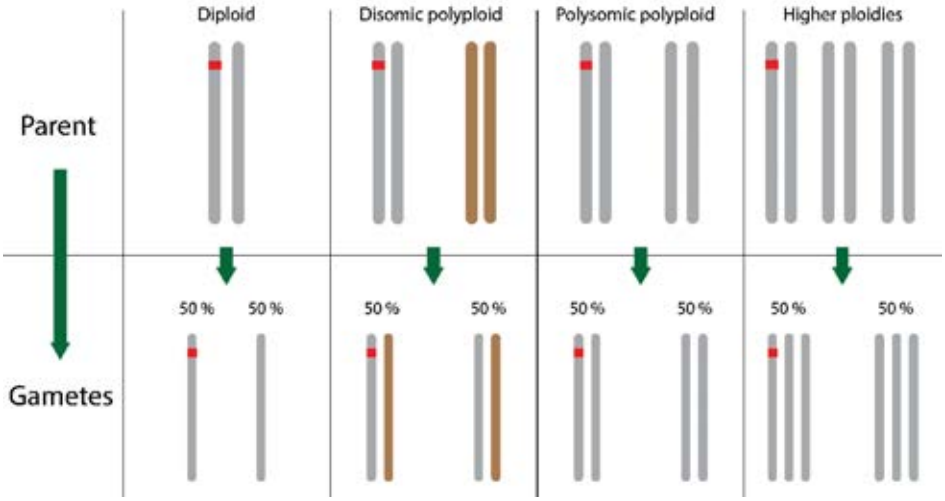



**Figure 4. Simplex markers.** These markers carry a single copy of the segregating marker allele and inherit similarly across all ploidy levels and pairing behaviours, allowing diploid mapping software to be used. Here, the (simplex) SNP allele is coloured red.

to be well-integrated between homologues in a single parent, and impossible to integrate across parents. In other words, the chromosomal numbering will most likely be inconsistent between parental maps if only simplex markers are used. Producing a consensus or fully integrated map is desirable for many reasons, including being able to detect and model more complex QTL configurations than just simplex QTL. Therefore, a truly polyploid linkage mapping tool should be able to include all marker segregation types, not just 1:1 segregating markers.

## Polyploid linkage mapping software

Linkage mapping can be broken into three steps – linkage analysis, marker clustering and marker ordering. There are still relatively few software tools that can perform all three of these steps for polysomic species. Perhaps the most well-known and widely-used software tool is TetraploidMap for Windows (Hackett and Luo, 2003;Hackett et al., 2007). As well as producing linkage maps for autotetraploid species, this software also

performs QTL interval mapping (returned to later). Recently, TetraploidMap was updated to enable the use of dosage-scored SNP data (Hackett et al., 2013). The updated version, TetraploidSNPMap (Hackett et al., 2017), is freely available to download from the Scottish BioSS website (bioss.ac.uk/knowledge/tetramap.html), and possesses a sophisticated graphical user interface (GUI) which will be extremely welcome for users in both the research and breeding community. Apart from its dependency on the Windows platform, the main drawback of TetraploidSNPMap (TSNPM) is that it is programmed to analyse autotetraploid data only, and there is no indication when or if it will be expanded to other ploidy levels or modes of inheritance. However, tetraploidy is the most common polyploid condition (Comai, 2005) and therefore this software is still relevant for a broad range of species.

Recently, an alternative linkage mapping package called polymapR was released, which is described in a pre-print manuscript (Chapter 6). Like TSNPM, polymapR uses dosage-scored marker information from $F_1$ populations to estimate recombination frequencies by maximum likelihood in a two-point linkage analysis. It can perform linkage analysis for polysomic triploids, tetraploids and hexaploids as well as segmental allotetraploid populations. As an R-based package it requires some level of user familiarity with R, but comes with a descriptive vignette which should make it accessible even to novice R users. It uses the same high-speed map ordering algorithm as TSNPM, namely MDSMap (Preedy and Hackett, 2016), and produces both integrated and phased linkage maps (*i.e.* separate maps for each parental homologue that are also integrated into a single consensus map). So far, developmental versions of this software have been used to generate high-density linkage maps in tetraploid potato (Bourke et al., 2016), tetraploid rose (Bourke et al., 2017) and hexaploid chrysanthemum (van Geest et al., 2017a).

Another recently-released R package that can perform linkage map construction is the netgwas package, also described in a pre-print manuscript (Behrouzi and Wit, 2017a). netgwas claims to be able to construct maps at any ploidy level in both inbred and outbred bi-parental populations, and rather than computing recombination frequencies and LOD scores, it uses conditional dependence relationships between markers based on discrete graphical models. The algorithm automatically detects linkage groups (which are traditionally identified by a user-specified LOD threshold) and does not rely on knowledge of parental dosage scores (which should offer robustness against parental genotyping errors). The output of netgwas is clustered and ordered marker names, but without assigning genetic positions (centiMorgans) or marker phasing, which are part of the TSNPM and polymapR output. The lack of marker phasing in particular is a major drawback, as phase considerations are crucial in polyploid genetic analyses. However,

given its novel and computationally-efficient approach to map construction, it appears to be a very interesting addition to the current range of polyploid mapping tools.

Another software program that is able to perform all three major steps in polyploid linkage mapping is the PERGOLA package in R (Grandke et al., 2017). This software can analyse marker data from all ploidy levels and modes of inheritance,

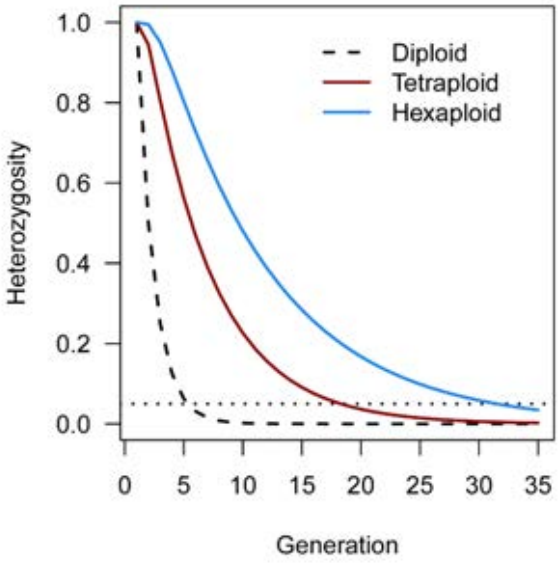



**Figure 5. Theoretical rate of decrease in heterozygosity in polyploid species from repeated rounds of inbreeding / selfing, using expressions derived by Haldane (1930).** For autotetraploids (red line), 95% homozygosity (horizontal dotted line) is achieved after on average 19 generations of selfing, while for a hexaploid (blue line) 95% homozygosity is reached after approximately 32 generations. By contrast, a diploid reaches 95% homozygosity after approximately 5 generations of selfing (black dashed line).

but is limited to populations derived from completely inbred (homozygous) founder parents, such as $F_2$ or $BC_1$ populations. While these sorts of experimental population are common in diploid plant species, they are much less common in polyploids due to the difficulty in reaching homozygosity through selfing (Haldane, 1930). Generally speaking, polyploids are more heterozygous than diploids (Soltis and Soltis, 2000) although there is no general consensus regarding their tolerance of inbreeding (Krebs and Hancock, 1990;Soltis and Soltis, 2000;Galloway et al., 2003;Galloway and Etterson, 2007). There are indications that polyploid plant species self-fertilise more often than their diploid relatives (Barringer, 2007). However, regardless of whether polyploids tolerate some levels of inbreeding or not, heterozygosity is maintained for many more generations in repeatedly-selfed polyploids than in selfed diploids (Figure 5). It therefore appears likely that PERGOLA was developed for newly-formed polyploids

derived from inbred diploid lines. The complexities facing extant (or heterozygous) polyploid species such as unknown marker phasing, or variable marker information contents are ignored by PERGOLA, making it doubtful that this tool will have a wide impact on linkage mapping in existing polyploid populations.

One other software that should be mentioned is PolyGembler, recently described in a pre-print manuscript (Zhou et al., 2017). It proposes a novel approach to the creation of linkage maps in outcrossing polyploids, and is also suitable for diploid mapping. Interestingly, it combines a haplotyping algorithm (derived from the polyHap algorithm (Su et al., 2008)) to first generate phased multi-marker scaffolds or haplotypes. These are then used to calculate recombination frequencies by counting recombination events both within and between these scaffolds, leading to an extremely simple estimate of *r* which has no corresponding LOD score. Scaffolds are clustered using a graph partitioning algorithm, and thereafter, the computationally-efficient CONCORDE travelling-salesman solver is employed to order markers (as is done for example in TSPmap (Monroe et al., 2017)). This assumes that the variance of all *r* estimates is equal and that weights are not required – which may well be the case if the haplotype scaffolds are correctly constructed. PolyGembler claims to be able to handle the high levels of missing data and genotyping errors associated with GBS data. Although it is applicable to multiple ploidy levels, the authors point out that mapping at the hexaploid level becomes computationally difficult due to the huge number of possible combinations in the formation of haplotypes. However, it appears to be a very promising tool which combines both genetic and bioinformatic approaches in a single pipeline.

Apart from those tools which constitute a complete linkage mapping pipeline, there have been some specific tools recently developed which we predict will have an important impact on future polyploid mapping applications. One of the most significant of these is the MDSMap package in R (Preedy and Hackett, 2016), a novel approach for determining a map order using multi-dimensional scaling. Marker data in polyploid species possesses variable information content, a fact that can be appreciated by considering the haplotype origin of markers of dosage 1 from a duplex marker in a tetraploid species. Certain combinations of markers provide very unambiguous information about co-inheritance, whereas others do not. Therefore, weights are required to prevent imprecise combinations from exerting a large influence on the map order. Before MDSMap was developed, the only reliable algorithm for ordering weighted recombination frequencies was the weighted regression algorithm from JoinMap (Stam, 1993;Van Ooijen, 2006). However, this has the disadvantage of being very slow for higher numbers of marker and is therefore of limited use with current high-density marker datasets. The MDSMap approach can achieve similar results in a fraction of the

time, and takes as its input the same information as JoinMap does, the pairwise recombination frequency estimates and logarithm of odds (LOD) scores, making this tool suitable for linkage map construction at any ploidy level, provided pairwise linkage analysis can be performed.

One final tool that has also proven useful for polyploid linkage map construction is the LPmerge package in R (Endelman and Plomion, 2014). LPmerge uses linear programming to remove the minimum number of constraints in marker order in order to create a conflict-free consensus map. It was originally developed to create integrated genetic maps from multiple (diploid) populations. That said, polyploids contain multiple copies of each chromosome and therefore also present a similar challenge if we consider each homologue map as originating from a different population, with non-simplex markers as bridging markers (mapped in more than one population). Homologue-specific maps are still regularly generated in polyploid mapping studies (*e.g.* in potato (Bourke et al., 2015;Bourke et al., 2016), rose (Vukosavljev et al., 2016) or sweet potato (Shirasawa et al., 2017)), for which LPmerge (or a similarly-efficient integration algorithm) could then be used to generate chromosomally-integrated maps.

# Genome-wide association studies

Genome-wide association studies (GWAS) have emerged as a powerful tool for detecting causative loci underlying phenotypic traits. They have been particularly popular in species where the generation of experimental populations is problematic (such as humans). GWAS has been readily adopted across a broad spectrum of species since then, due to the promise of increased mapping resolution, a more diverse sampling of alleles and a simplicity in population creation (no crossing required) (Bernardo, 2016). There are certain disadvantages though, particularly in how rare (and potentially important) variants can be missed (Ott et al., 2015) and the confounding effect of population structure on results (Korte and Farlow, 2013). Nevertheless, GWAS continues to be an important analytical option to help shed greater light on genotype – phenotype associations.

## Polyploid GWAS

The application of GWAS in polyploid species is relatively new, although there have already been a number of studies published in various crop species, for example in potato, oilseed rape, wheat, and oats (Uitdewilligen et al., 2013;Gajardo et al., 2015;Sukumaran et al., 2015;Tumino et al., 2016;Tumino et al., 2017). GWAS studies usually need to account for population structure and relatedness to prevent spurious

associations, often in the context of linear mixed models (Yu et al., 2006;Bradbury et al., 2007;Zhang et al., 2010b). One challenge in applying GWAS to polyploid species is how to define a relatedness metric between polyploid individuals (*i.e.* how to generate the kinship matrix, ***K***). So far, there have been two software tools released for polyploid GWAS, namely the R package GWASpoly (Rosyara et al., 2016) and the previously-mentioned SHEsisPlus (Shen et al., 2016). Of these, only GWASpoly looks critically at the form of the kinship matrix ***K***. Three different forms of ***K*** were tested in the development of the package, with the canonical relationship matrix (VanRaden, 2008) (termed the realised relationship matrix by the authors (Rosyara et al., 2016)) found to best control against inflation of significance values. This is also the default ***K*** provided in the GWASpoly package. An alternative approach to GWAS mapping for polyploids is provided by the netgwas package (Behrouzi and Wit, 2017b), previously mentioned for its linkage mapping capacity. Again, graphical models form the basis of the approach, which goes beyond single-marker association mapping to investigate genotype-phenotype interactions using all markers simultaneously in a graph structure. There is almost no discussion on how confoundedness between population structure and phenotypes are handled, but the authors claim the detection of false positive associations is not problematic.

One final aspect worth considering is the issue of deploying an adequate number of markers in a polyploid GWAS, which potentially represents a much larger genomic space. In *Arabidopsis thaliana*, it was estimated that between 140K and 250K SNPs would be needed to fully cover the genome based on a study of linkage disequilibrium in that species (Kim et al., 2007). Modelling the decay of linkage disequilibrium in polyploid species is a more complex exercise. It was previously suggested that estimates of linkage disequilibrium may be inflated in polyploid species (Jannoo et al., 1999;Flint-Garcia et al., 2003). A more recent survey of linkage disequilibrium in autotetraploid potato using SNP dosages estimated that at most 40K SNPs would be needed for QTL discovery in potato (Vos et al., 2017), a much lower estimate than for *Arabidopsis* (Kim et al., 2007). The discrepancy comes in part from the differences in how these figures were estimated, using a 'hide-the-SNP' simulation for *Arabidopsis* versus a 'rule of thumb' calculation for potato, but mainly from the difference in the extent of LD between the two species (estimated at ~10 Kb in *A. thaliana* versus ~2 Mb in *S. tuberosum* (Kim et al., 2007;Vos et al., 2017)). Detecting or even defining linkage disequilibrium between markers linked in repulsion phase is non-trivial in autopolyploids (Vos et al., 2017), which is analogous to the problem of detecting and estimating recombination frequency between such markers in a linkage mapping study. So far, we are not aware of any software tool that has been developed to estimate the extent of linkage disequilibrium in

polyploids, which would complement the design of future GWAS studies in polyploid species.

# QTL analysis

The term "QTL analysis" usually refers to studies that aim to detect regions of the genome (so-called quantitative trait loci (Geldermann, 1975)) that have a significant statistical association with a trait in specifically-constructed experimental populations. These populations are most often created by crossing two contrasting parental lines ("bi-parental" populations), although there is increasing interest in using more complex population designs in order to increase the range of alleles and genetic backgrounds being studied (*e.g.* "MAGIC" populations (Huang et al., 2015)). As already discussed, there is great difficulty in developing inbred lines by repeatedly selfing polyploids due to the sampling of alleles during polyploid gamete formation (in a diploid this sampling generates $\binom{2}{1} = 2$ combinations; for a tetraploid this rises to $\binom{4}{2} = 6$ and in a hexaploid $\binom{6}{3} = 20$ combinations, resulting in protracted heterozygosity (Figure 5)), not to mention the problem of inbreeding depression associated with many outcrossing polyploid species. Therefore, most QTL analyses in polyploid species have been performed using the directly-segregating $F_1$ progeny of a cross between heterozygous parents (a "full sib" population). This leads to poor resolution of QTL positions when compared to the more popular diploid inbred populations like RILs *etc.*, as well as the fact that populations must be vegetatively propagated if replication over years or different growing environments is desired. For many polyploid species, vegetative propagation is indeed possible (Herben et al., 2017) and $F_1$ populations have the added advantage of being relatively quick and simple to develop, while, because of a generally high level of heterozygosity, many loci will be segregating in the $F_1$. Therefore despite their drawbacks, $F_1$ populations remain the bi-parental population of choice for mapping studies.

The methods for QTL analysis in diploid species have become increasingly convoluted (van Eeuwijk et al., 2010); in polyploid species such theoretical complexities have yet to be attempted, given the more immediate difficulties in accurately genotyping as well as modelling polyploid inheritance. Just like for linkage mapping and GWAS, the range of software tools available for QTL analysis in polyploids remains rather limited, although there are a number of recent developments that are helping transform the field.

One of the only dedicated software for tetraploid QTL analysis is the already-mentioned TetraploidMap software (Hackett et al., 2007). This software enables interval mapping to be performed in autotetraploid $F_1$ populations (as well as a simple single-marker

ANOVA test), using a restricted range of markers (1x0, 2x0 and 1x1 markers only, where 1x0 denotes a marker dosage of 1 in one parent and 0 in the other, *etc.*). Although still available, it has been superseded by the TetraploidSNPMap software (Hackett et al., 2017). TetraploidSNPMap (TSNPM) uses SNP dosage data to either construct a linkage map (as already described) or perform QTL interval mapping. In contrast to its predecessor, TSNPM can analyse all marker segregation types, and allows the user to explore different QTL models at detected peaks. At its core is an algorithm to determine identity-by-descent (IBD) probabilities for the offspring of the population, which are then used in a weighted regression performed across the genome.

An independent software tool that has been developed to determine IBD probabilities in tetraploids is TetraOrigin (Zheng et al., 2016), implemented in the Mathematica programming language. TetraOrigin relaxes the assumption of random bivalent pairing during meiosis (which TSNPM employs) to allow for both preferential chromosomal pairing as well as multivalent formation and the possibility of double reduction. Although not programmed in a user-friendly format like TSNPM, it is relatively straightforward to use, taking an integrated linkage map and marker dosage matrix as input. It does not perform QTL analysis directly, but the resulting IBD probabilities can then be used to model genotype effects in a QTL scan either using a weighted regression approach like TSNPM, or in a linear mixed model setting. IBD probabilities allow interval mapping since they can be interpolated at any desired intervals on the linkage map.

For ploidy levels other than tetraploid, there are currently no dedicated software tools available for QTL analysis or IBD probability estimation. Single-marker approaches such as ANOVA on the marker dosages (assuming additivity – various dominant models could also be explored; see *e.g.* (Rosyara et al., 2016)) are of course possible and require access to basic statistical software packages such as R (or even Excel). However, such approaches are not ideal – they are only effective if marker alleles are closely linked in coupling with QTL alleles, and offer no ability to predict the QTL segregation type or mode of gene action as is done for example in TSNPM (Hackett et al., 2017). As interest increases in the genetic dissection of important traits in polyploid species, we anticipate that it is only a matter of time before more flexible cross-ploidy solutions are developed. Methodologies developed for tetraploid species often claim that “extension to higher ploidy levels is straightforward”. These sorts of disingenuous claims attempt to mark new research territory as already solved. If extensions to higher ploidy levels were indeed straightforward we would already be reporting on a wider range of tools available for them – as far as we can tell, so far there are none.

Returning to the topic of population types, we also anticipate that more powerful QTL analyses can be performed by combining information over multiple populations. Approaches such as pedigree-informed analyses, implemented for diploids in the FlexQTL software (Bink et al., 2008), could overcome some of the limitations imposed by the restrictions on population types in software for polyploids. However, it may take some time before such tools become translated to the polyploid level.

# Genomic prediction and genomic selection

There has been much attention given to the advantages of using *all* marker data to help predict phenotypic performance, rather than focusing on single markers (or haplotypes) that are linked to QTL as was previously advocated. The motivation behind this is clear – many of the most important traits in domesticated animal and plant species are highly quantitative, with far too many small-effect loci present to be able to tag them all with single markers (Bernardo, 2008). One of the most important traits in any breeding program is also a famously quantitative trait: yield. It has been suggested that despite many years of phenotypic selection, crop yield in tetraploid potato has essentially remained unchanged (Jansky, 2009;Slater et al., 2016). This is a remarkable indictment of traditional selection methods, yet offers much-needed impetus for the development and deployment of new paradigms in breeding for quantitative traits.

Genomic prediction first arose in animal breeding circles (Meuwissen et al., 2001), where the concept of estimating breeding values from known pedigrees was already well-established. However, the estimation of breeding values in polyploid species requires special consideration due to the complexity of polysomic inheritance and the possibility of double reduction. In practice, breeding values are usually estimated using restricted maximum likelihood (REML) to solve mixed model equations, requiring the generation of an inverse additive relationship matrix $\boldsymbol{A^{-1}}$, also called the numerator relationship matrix. The form of $\boldsymbol{A^{-1}}$ depends on, among other things, whether the inheritance is polysomic or disomic, and whether double reduction occurs (Kerr et al., 2012;Hamilton and Kerr, 2017). Recently, the R package polyAinv was released which computes the appropriate $\boldsymbol{A^{-1}}$ as well as the kinship matrix $\boldsymbol{K}$ and the inbreeding coefficients $F$ (Hamilton and Kerr, 2017). However, in one study of nine common traits in autotetraploid potato, the inclusion of double reduction, or even the adoption of an autotetraploid-appropriate relationship matrix was found to have a minimal impact on the results (Slater et al., 2014). Studies which ignore the specific complexities of autopolyploids may still benefit from genomic prediction and selection, as for example was demonstrated in tetraploid potato (Sverrisdóttir et al., 2017). Commonly-used

software tools for estimating breeding values at the diploid level include ProGeno (Maenhout, 2018) and ASreml (V.S.N. International, 2018) which could be suitable for polyploid breeding programs, although this has yet to be conclusively demonstrated.

# Mode of inheritance

The term "mode of inheritance" refers to the randomness of meiotic pairing processes that give rise to gametes, and is often used to distinguish between disomic (diploid-like) inheritance, and polysomic (all allele combinations equally possible) inheritance. As alluded to already, intermediate modes of inheritance are theoretically possible if partially-preferential pairing occurs between homologues, resulting in on average more recombinations between certain homologues, and less between others (putative homoeologues). This intermediate inheritance pattern, originally termed segmental allopolyploidy (Stebbins, 1947) and more recently termed mixosomy (Soltis et al., 2016), poses additional challenges over those of purely polysomic or disomic behaviour. One of the main complications is the lack of fixed segregation ratios to test markers against (Allendorf and Danzmann, 1997), which is often used as a measure of marker quality (Stringham and Boehnke, 1996;Pompanon et al., 2005). Currently there are no dedicated tools available to ascertain the most likely mode of inheritance in polyploids. Some "traditional" approaches to predict the mode of inheritance are summarised in (Bourke et al., 2017), many of which are relatively straightforward to implement using a statistical programming environment like R (R Core Team, 2016). In that study, TetraOrigin (Zheng et al., 2016) was used to estimate the most likely pairing configuration that gave rise to each offspring in an $F_1$ tetraploid population. This enabled the authors to test whether there were deviations from the expected patterns of homologue pairing under a tetrasomic model (Bourke et al., 2017). A simple alternative using closely-linked repulsion-phase simplex marker pairs was also proposed and has been implemented in the polymapR package (Chapter 6). Apart from preferential pairing, TetraOrigin can also predict whether marker data arose from bivalent or multivalent pairing during meiosis, facilitating an analysis of the distribution of double reduction products. However, apart from its restriction to tetraploid data, an integrated linkage map is required before TetraOrigin can be employed. In severe cases of mixosomy, it is not obvious how a reliable linkage map should be generated. Corrections for mixosomy in a tetraploid linkage analysis are possible in polymapR, but in extreme cases marker clustering will also be affected, making map construction quite challenging. A confounding complication is the possibility of variable chromosome counts (aneuploidy), as for example encountered in sugarcane (Grivet et al., 1996;Grivet

and Arruda, 2002) or in ornamentals such as *Alstroemeria* (Buitendijk et al., 1997), which makes the diagnosis of the mode of inheritance even more difficult. As more polyploid species begin to be genotyped, the issue of unknown mode of inheritance will likely exert more influence, further necessitating the development of software tools that can provide an accurate assessment of the inheritance mode using marker data, and that can accommodate the full spectrum of polyploid meiotic behaviours.

# Simulation software

As with any software tool, developing standards and scenarios upon which the performance of the tool can be judged is vital to ensure reliable results. In this final section we consider the range of simulation tools currently available for polyploids. Probably the most widely-used polyploid simulation software currently available is PedigreeSim (Voorrips and Maliepaard, 2012). Originally developed to generate diploid and tetraploid populations, the current release (PedigreeSim V2.0) can simulate populations of any even ploidy level (2, 4, 6, ...). What makes PedigreeSim particularly attractive is its ability to simulate a diversity of meiotic pairing conditions, including quadrivalents (which can result in double reduction) or preferential chromosome pairing. It takes four input files (which are relatively simple to generate) that provide a description of the desired simulation parameters and the input marker data. The software then creates (dosage-scored) genotype data for any pedigreed population, *e.g.* an $F_1$ population of specified size (Voorrips and Maliepaard, 2012). Some authors have used PedigreeSim to simulate multiple generations of random mating, allowing an investigation of population structure and linkage disequilibrium in polyploid species (*e.g.* (Rosyara et al., 2016;Vos et al., 2017)), which can be implemented quite easily with some basic programming knowledge. PedigreeSim is written in Java and can run on all major operating systems. A Windows-based software Polylink, which originally performed two-point linkage analysis and simulation of tetraploid populations (He et al., 2001), is no longer available. The R package polySegratio (Baker, 2014) simulates dominantly-scored marker data in autopolyploids of any even ploidy level. Generating the dosage data is straightforward: only the expected proportion of marker types (simplex, duplex, triplex,...) as well as the ploidy is required. However, the markers are essentially completely random, with no connection to any linkage map, which is arguably of limited use for any application that requires some degree of linkage between markers. The simulation capacities of polysegRatio therefore appear to be most useful for testing functions within the package itself, namely those designed to impute parental dosages given the observed segregation ratios in offspring scores.

A final polyploid simulation tool that has recently been developed is the HaploSim pipeline which includes the HaploGenerator function (Motazedi et al., 2017b). HaploGenerator is designed to generate sequence-based haplotypes in a polyploid of any even ploidy, taking the fasta file it is provided with as a reference from which haplotypes are built. The software generates random SNP mutations at a specified distribution before simulating next-generation sequencing (NGS) reads in formats corresponding to a number of current sequencing technologies such as Illumina or Pacific Biosystems (PacBio). The pipeline was originally developed to compare the performance of a number of haplotype assembly algorithms (Motazedi et al., 2017b), but could also be useful for testing the performance of any other tool which uses NGS reads as genotypes.

# Future perspectives

In this review we have attempted to describe the most important software tools that are currently available to the polyploid genetics community. There are likely to be tools that were missed and tools that have subsequently been released – this is the danger of such a review. However, we have tried where possible to also discuss the gaps that are apparent in the current set of available tools which will hopefully help guide their development in future. Polyploid genotyping arguably remains the most critical step, as without accurate genotype data there is little point in building models for polyploid inheritance. However, we are now witnessing the slow emergence of tools that take polyploid genotypes and use them to make inferences on the transmission of alleles and the effects of such alleles in polyploid populations. As genotyping technologies continue to evolve, so too should the suite of tools developed to analyse those genotypes. Tools for analysing SNP dosage data from SNP arrays are well-established, with extensions of current tools to higher ploidy levels planned (*i.e.* fitPoly). The coming decade will likely see a move away from SNP array-based genotyping to the use of sequence-read based genotypes, although this will require that all tools heretofore developed be updated to accommodate the new type of data. Information on the mode of inheritance from marker data is also needed for each population studied, which deserves more attention than it currently receives. A move from diploid-based reference genomes to fully polyploid (and haplotype-resolved) reference genomes would also help broaden the boundaries of polyploid genetics away from the diplo-centric view of genomics which currently dominates. Although there have been many exciting discoveries and developments in polyploid genetics in the past decade or more, we feel its golden age has yet to arrive, an age which will be heralded all the sooner by the provision of robust and user-friendly tools for the genetic dissection of these fascinating group of organisms.

# Chapter 3

# The double reduction landscape in tetraploid potato as revealed by a high-density linkage map

Peter M. Bourke[1], Roeland E. Voorrips[1], Richard G. F. Visser[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

*Published as* Bourke, P. M., Voorrips, R. E., Visser, R. G. F. and Maliepaard, C. (2015). "The Double Reduction Landscape in Tetraploid Potato as Revealed by a High-Density Linkage Map", **Genetics 201 (3), 853-863** *and reproduced here with permission of the Genetics Society of America*

# Abstract

The creation of genetic linkage maps in polyploid species has been a long-standing problem for which various approaches have been proposed. In the case of autopolyploids, a commonly-used simplification is that random bivalents form during meiosis. This leads to relatively straightforward estimation of recombination frequencies using maximum likelihood from which a genetic map can be derived. However, autopolyploids such as tetraploid potato (*Solanum tuberosum* L.) may exhibit additional features such as double reduction, not normally encountered in diploid or allopolyploid species. In this study we produced a high-density linkage map of tetraploid potato and used it to identify regions of double reduction in a bi-parental mapping population. The frequency of multivalents required to produce this degree of double reduction was determined through simulation. We also determined the effect that multivalents or preferential pairing between homologous chromosomes have on linkage mapping. Low levels of multivalents or preferential pairing do not adversely affect map construction when highly-informative marker types and phases are used. We reveal the double reduction landscape in tetraploid potato, clearly showing that this phenomenon increases with distance from the centromeres.



# Key words

Linkage mapping, tetraploid, double reduction, potato, multivalents.

# Introduction

Polyploid species constitute a very important group among cultivated crops. Polyploids themselves can be further divided into auto- and allo-polyploids, with autopolyploids showing random association between homologous chromosomes and allopolyploids showing non-random or preferential pairing during meiosis. Linkage mapping in autopolyploid species remains a challenging exercise despite recent advances in genotyping technology and mapping methodology. Breeding work in many autopolyploid crops has yet to benefit from the use of markers in breeding programs. This is partly due to the lack of software to perform linkage mapping and QTL analysis in polyploids, but is also due to the complicated nature of autopolyploid genomes and genetics. The software program TetraploidMap (Hackett and Luo, 2003) is a notable exception to this, but is constrained by the relatively low numbers of markers it can handle (currently 800 is the maximum) and the need to manually assign marker phase which may become infeasible with large datasets.

One autopolyploid species where large advances in genetic analysis have been made is tetraploid potato (*Solanum tuberosum* L.), in terms of the availability of a high-quality reference sequence (Potato Genome Sequencing Consortium, 2011), many published linkage maps (Meyer et al., 1998;van Os et al., 2006;Felcher et al., 2012;Hackett et al., 2013) as well as methods for performing linkage mapping at the polyploid level (Luo et al., 2001;Bradshaw et al., 2004;Hackett et al., 2013). In comparison to other economically-important autotetraploid species such as alfalfa, rose or leek, the pairing behavior of potato is thought to be relatively well-understood, with random bivalent pairing during prophase I of meiosis being generally assumed (Swaminathan and Howard, 1953;Milbourne et al., 2008). Although a certain proportion of multivalents is known to occur, these are not deemed to occur at a sufficient frequency to merit their inclusion in a pairing model (Bradshaw, 2007).

The simplest marker segregation type to map in a tetraploid cross are simplex x nulliplex marker types which are expected to segregate in a 1:1 fashion. In a tetraploid, we employ the term simplex x nulliplex to collectively refer to 1x0, 3x0, 3x4 and 1x4 markers (with 0x1, 0x3, 4x3 and 4x1 markers being nulliplex x simplex). A relabeling of allele dosages is sufficient to convert all these markers to their simpler form. These have traditionally been the markers most favored in tetraploid mapping because of their simple segregation, reliability in genotype-calling and high information content in coupling-phase. One important practical advantage is that these markers can be mapped using advanced mapping software developed for diploids such as JoinMap (Van Ooijen, 2006) which can efficiently map large numbers of markers as well as providing many checks on map and data quality. Simplex x nulliplex markers also provide the clearest linkage

information to cluster markers into separate homologous chromosomes, forming the basis of homologue maps. In our population, simplex x nulliplex markers were also the most abundant marker segregation type. We therefore restricted our analysis to simplex x nulliplex markers, which nevertheless allowed us to map a total of 3273 markers across both parents.

Simplex x nulliplex markers are also the most useful markers to provide direct evidence of one of the observable consequences of multivalent formation, namely double reduction (DR). In autopolyploid species, pairing may occur between all homologous chromosomes which can lead to complicated pairing structures during the first meiotic division (Milbourne et al., 2008). In cases where a cross-over occurs between two sets of sister chromatids which subsequently migrate to the same pole, it is possible for a chromatid and its recombinant copy (segment) to end up in the same gamete, a situation which can never occur in diploids. For a simplex x nulliplex marker with the segregating allele on the recombinant segment in question, this can lead to a duplex score in that offspring. By simulating comparable mapping populations genotyped with the same mapped markers we were able to estimate the rate of multivalent formation that would account for the observed levels of DR. We also performed a simulation study using populations with different rates of multivalent formation and preferential pairing to investigate the effect that the assumption of random bivalent formation has on the estimation of recombination frequency and marker phase.

# Materials and Methods

## Plant material

An $F_1$ mapping population of 237 individuals was created from the cross between two tetraploid potato varieties, cultivars '*Altus*' (hereafter referred to as parent one, P1) and '*Colomba*' (P2).

## DNA extraction and genotyping

DNA was extracted from leaf material using KingFisher Flex according to the manufacturer's instructions (Thermo Scientific). The concentration of DNA was measured using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific) and the DNA concentration was adjusted to ~50 ng $\mu l^{-1}$ (Vos et al., 2015). For DNA concentrations in the range of 25 – 50 ng $\mu l^{-1}$ the sample was also used; samples having concentrations lower than 25 ng $\mu l^{-1}$ were discarded and DNA isolation was performed again. The samples were genotyped on the SolSTW Infinium SNP array which assayed 17,987 SNPs as described by (Vos et al., 2015). Of these SNPs, 4179 also form part of the SolCap SNP array (Felcher et al., 2012). The arrays were processed according to the

manufacturer's protocol at ServiceXS, Leiden, the Netherlands. Each parent was genotyped in duplicate using two biological replicates. 1662 other tetraploid accessions were sampled in a similar fashion, as well as 516 diploid accessions (for use in another study as well as helping marker dosage fitting).

## Assignment of dosages

The X and Y allele signal intensities were imported from the Illumina data output into the R programming environment (R Core Team, 2016). SNPs were initially filtered so that the average of their total signal intensity (the sum of the X and Y allele signal intensities) over all samples was greater than 0.2. The marker intensities were converted into allele dosages using the fitTetra package for R (Voorrips et al., 2011). Changes to the default settings of the *saveMarkerModels* function of fitTetra were as follows: *p.threshold* was decreased from 0.99 to 0.95, *peak.threshold* was increased from 0.85 to 0.99 and *sd.target* was set to 0.04, where *p.threshold* is the "minimum P-value required to assign a genotype to a sample", *peak.threshold* is the "maximum allowed fraction of the scored samples that are in one peak" and *sd.target* is used to specify the maximum non-penalised standard deviation of the fit on a transformed scale (Voorrips et al., 2011). All diploid and tetraploid samples were included in the fitting because this generally results in a better fit of the dosage classes. Following fitting with fitTetra, the marker dosage scores were screened to ensure consistency between parental and offspring genotypes. Markers with up to 3% invalid scores (scores that were not expected based on the parental genotypes and bivalent chromosome pairing) were allowed. A high frequency of many invalid scores suggests that either the marker performed poorly, there was some consistent error in dosage assignment, or one or both of the parents had been incorrectly genotyped. Highly-skewed markers ($p < 0.001$) were also removed at this stage.

**Table 1. Breakdown of SNP marker numbers after quality filtering**

| Steps in SNP filtering | #SNPs | % |
|---|---|---|
| SolSTW Infinium array total # SNPs | 17987 | 100.0 |
| Dosages assigned by fitTetra [a] | 15266 | 84.9 |
| Both parents ok | 15137 | 84.2 |
| $F_1$ pattern acceptable [b] | 13767 | 76.4 |
| $F_1$ monomorphic | 6553 | 36.4 |
| $F_1$ polymorphic | 7214 | 40.0 |

[a] Markers not scored were either monomorphic or not clearly resolved

[b] Criteria for lack of $F_1$ fit: presence of null alleles, > 3% invalid scores, highly-skewed segregation ($P < 0.001$)

## Marker conversion

Markers that segregated in a 1:1 fashion were re-labelled as simplex x nulliplex (or nulliplex x simplex) for mapping and double reduction analysis. Considering markers whose segregating allele is inherited from P1, these consisted of triplex x nulliplex, triplex x quadruplex and simplex x quadruplex markers. For example, a triplex x nulliplex marker is expected to produce 50% dosage '1' and 50% dosage '2' among the offspring, with observable double reduction scores appearing as dosage '0' (a double-copy of the '0' allele from P1). Re-labelling '2' as '0' and '0' as '2' (with the parents re-labelled as simplex and nulliplex) achieves the desired result of marker conversion.

**Table 2. Tetraploid marker segregation types by number**

| Parental dosage | Segregation | #SNP [a] |
|---|---|---|
| Simplex x nulliplex (SxN) | 1:1 | 1549 |
| Nulliplex x simplex | 1:1 | 1733 |
| Duplex x nulliplex (DxN) | 1:4:1 | 466 |
| Nulliplex x duplex | 1:4:1 | 421 |
| Simplex x simplex (SxS) | 1:2:1 | 949 |
| Simplex x triplex (SxT) | 1:2:1 | 441 |
| Duplex x simplex (DxS) | 1:5:5:1 | 714 |
| Simplex x duplex (SxD) | 1:5:5:1 | 640 |
| Duplex x duplex (DxD) | 1:8:18:8:1 | 303 |
| | | **7214** |

[a] Number of SNP markers after simplifying marker conversions have been performed.

## Linkage map construction

Simplex x nulliplex marker data were recoded to JoinMap 4.1 cross-pollinator format (lm x ll). 'Impossible' genotypes (invalid scores) were made missing before importation into JoinMap. One pair of identical individuals was identified in the dataset (similarity of 0.9922), therefore we removed individual #202. Markers were assigned to linkage groups with a minimum LOD of 4 (a higher LOD was used if clusters broke into large sub-clusters at a higher LOD). Marker clusters were assigned to physical chromosomes based on the position of markers on the physical sequence (Potato Genome Sequencing Consortium, 2011). Mapping was first performed using the groupings from the Groupings Tree using maximum likelihood (ML). Homologues were then identified by large gaps in the estimated map distances (≥ 60 cM), which was also often accompanied by a transition in estimated marker phase. Marker data for separate homologues was exported from JoinMap in .loc files and re-imported for creation of the homologue maps. After an initial mapping of the homologues, individual #067 was found to contribute unrealistic numbers of recombinations in many linkage groups across both parents and

was therefore removed, resulting in a final mapping population of 235 individuals. Mapping was performed using ML with three rounds of map optimization using the default settings for spatial sampling thresholds. Haldane's mapping function was used to convert recombination frequency estimates to map distances as has previously been used for linkage map construction in tetraploid potato (Meyer et al., 1998;Hackett et al., 2013). In a number of cases, we used linkage information from the duplex x nulliplex and simplex x simplex markers to connect sub-homologue linkage groups that had poor internal linkage among simplex x nulliplex markers. Map data was exported from JoinMap as text files and imported into MapChart 2.3 (Voorrips, 2002) for further plotting.

## Comparison of genetic and physical maps

The genetic positions of markers were compared with their physical positions as defined in (Vos et al., 2015). It was found that some markers did not map to the same chromosome as expected from the physical map; a list of such markers is included in Supplementary table 1. The physical position of the centromere boundaries was initially adopted from previously-published values (Sharma et al., 2013). These were not found to coincide precisely with the points of inflection on the genetic-physical map, following which the approximate centromere bounds were re-defined by examination of the aligned genetic-physical plots (also referred to as Marey Maps (Chakravarti, 1991)) and calculating an approximate physical position between marker pairs flanking the points of inflection on these plots. The order of the genetic map was reversed in cases where the genetic maps were found to be inversely ordered with respect to the physical map.

## Conversion rate of physical to genetic distance

The conversion rate between genetic and physical distance was determined by regressing the genetic positions on their physical positions per homologue arm. The slopes of the regression lines for each homologue arm were tested for equality in an analysis of covariance by introducing, where necessary, up to three dummy variables (to code for the presence or absence of a homologue) per chromosome arm per parent (Andrade and Estévez-Pérez, 2014). An average genome-wide estimation of the genetic to physical conversion rate was calculated after excluding a single outlying value from the northern arm of homologue 2 of chromosome 1 in parent 1. This genome-wide recombination rate was used to convert the physical map to a pseudo-integrated genetic map for use in the simulation studies.

## Rates of double reduction

After recoding the 1:1 segregating marker data, duplex marker scores in the offspring were taken as possible evidence for double reduction. Duplex scores can also arise as a

result of genotyping errors. Therefore, we used a relatively strict criterion to decide if such scores were evidence of double reduction: a string of three consecutive duplex scored markers on a homologue map was required in order to be considered strong enough evidence for double reduction. This could theoretically lead to some under-estimation of the rates of double reduction, but the simplex marker density was sufficient that in most cases a double reduction region would contain at least three (segregating) simplex x nulliplex markers.

A routine was written in R to identify strings of three or more duplex scores. The rate of double reduction was determined for each marker by counting the number of times it formed part of a double reduction segment and dividing this by the number of non-missing values scored for that marker across the population. We then derived the average rate of double reduction per homologue for 1Mb windows north and south of the centromeric bounds by calculating the mean rate of double reduction over all markers within that window. These means were aggregated to give a single average rate of double reduction per homologue for each 1Mb window distance from the centromeres across all chromosomes and both parents. The average rate per chromosome was estimated by multiplying the homologue rates by a factor of four.

## Simulation of double reduction and prediction of quadrivalent formation

An approximate “integrated” genetic linkage map was produced using the average cM to Mb conversion rate and physical positions of the simplex markers. Only markers for which the assigned linkage group and physical chromosome corresponded were considered. Marker phase was determined according to the homologue assignment of all markers. Phased marker genotypes and a consensus genetic map position are the basic input for the simulation software PedigreeSim (Voorrips and Maliepaard, 2012), which simulates (diploid or) polyploid populations with specified levels of multivalents and / or preferential pairing. One thousand separate populations of 235 individuals were generated using the same simplex marker data and approximated map under a range of different fractions of quadrivalents. The algorithm for estimating double reduction was applied to the simulated datasets, allowing us to deduce the relationship between the rate of double reduction and the frequency of multivalents underlying meiosis.

## Estimation of the rate of preferential pairing

Repulsion-phase simplex marker data can be used to investigate whether preferential pairing occurs, as the estimates for recombination frequency in repulsion are expected to differ under disomic and tetrasomic inheritance (Qu and Hancock, 2001). We have adapted the approach of (Qu and Hancock, 2001) to correct for multiple-testing using

the false-discovery rate (FDR) (Benjamini and Hochberg, 1995), confining our analysis to within chromosomes to reduce the overall number of tests (coupling or repulsion linkage have no meaning when marker pairs from separate linkage groups are considered).

For two markers *A* and *B* we define $n_{00}$ as the number of individuals with dosage 0 at both markers, $n_{01}$ as the number of individuals with dosage 0 at marker *A* and dosage 1 at marker *B* and so on. The explicit ML estimator for the recombination frequency (*r*) in coupling phase under both disomic and tetrasomic inheritance is invariant ($\frac{n_{01}+n_{10}}{n_{00}+n_{01}+n_{10}+n_{11}}$), whereas in repulsion phase the ML estimator under disomic inheritance is $r_{disom} = \frac{n_{00}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$ and under tetrasomic inheritance, $r_{tetra} = \frac{2(n_{00}+n_{11})-(n_{01}+n_{10})}{n_{00}+n_{01}+n_{10}+n_{11}}$. If the mode of inheritance is tetrasomic, $r_{disom}$ should never fall below the value of 1/3, whereas in the case of disomic inheritance, $r_{disom} \in [0,0.5)$. This forms the basis of an exact Binomial test, with $H_0$: $r_{disom} = 1/3$ and $H_1$: $r_{disom} < 1/3$. Correction for multiple testing was performed using the FDR procedure with $\alpha = 0.05$, as described in (Benjamini and Hochberg, 1995).

## Simulation of mapping under different rates of quadrivalent formation and preferential pairing

One of the hypotheses we wanted to test was whether bivalent formation predominates in tetraploid potato as is commonly assumed. We also wanted to see the effect that deviations from this assumption could have on recombination frequency estimates that are based on a bivalent model. In this study we limited our focus to 1:1 segregating markers. We used PedigreeSim to simulate new mapping populations of 250 individuals with the fraction quadrivalents varying from zero to one in increments of 0.1. For each setting, one thousand simulated populations were generated. The simulated genome had a single chromosome of 100 cM with 51 simplex x nulliplex markers randomly distributed at positions no closer than 0.1 cM apart and the centromere at 25 cM. The true and estimated recombination frequencies between the first marker and the other 50 markers on the chromosome were recorded, as well as the LOD and assigned phase ("coupling" or "repulsion"). Recombination frequencies between marker pairs were estimated using ML, for which explicit estimators can be derived in the case of simplex marker pairs (*c.f.* previous section). Phase was determined by choosing the lowest estimate for the recombination frequency in the range [0,0.5) which we term phasing by the minimum recombination frequency (MINR). This differs from previous studies, where the maximum of the log-likelihood (MLL) was used to assign the most likely phase (Luo et al., 2001;Hackett et al., 2013). Negative estimates for *r* can occur due to

Mendelian sampling variation under weak repulsion linkage. For strongly-negative values ($r < -0.05$), a recombination frequency of 0.499, LOD of zero and phase "Unknown" were assigned and in the case $-0.05 \leq r < 0$, the recombination frequency was set to zero and the LOD and phase were left unchanged. The recombination frequency estimates were regressed on their true values for both coupling and repulsion phase in order to evaluate how close to the true value the estimates fell for each pairing scenario. The proportion of correctly-assigned phases for coupling and repulsion phase markers was also recorded.

# Results

## Genotyping and dosage assignment

Of the 17,987 SNPs assayed, only 40% were found to be acceptable and segregating in this population (Table 1). Acceptable markers were those for which dosages could be assigned by fitTetra, for which parental dosages were scored consistently between replicates and for which parental dosages and offspring segregation patterns were consistent. Approximately 85% of the markers could be assigned dosages by fitTetra, after which a further 5% were rejected for having inconsistent parental – offspring dosages, or for being too highly-skewed ($\chi^2$ test with $p < 0.001$). 1:1 segregating markers formed the largest group among the 7214 segregating markers in our population (Table 2), accounting for over 45% of useable markers.

## Mapping of the 1:1 segregating markers

Almost no simplex x nulliplex markers dropped out during the mapping stage. Of the 1549 simplex x nulliplex markers in P1, 1544 were mapped, and 1729 out of the 1733 P2 markers were mapped (Table 3). The unmapped markers were lost due to poor linkage (either no chromosome assignment or extremely weak linkage within a linkage group) or large numbers of missing values. Marker coverage over all chromosomes was well spaced with on average over 270 markers per chromosome. Only chromosomes 10 and 12 had fewer than 200 mapped markers (126 and 168 markers respectively) with chromosomes 2 and 5 having the highest marker coverage (390 and 388 markers mapped respectively). A number of homologues were split up over more than one linkage group due to insufficient linkage information. In these cases, DxN and SxS markers were used to provide linkage information between homologue fragments. An example of the four homologue maps of chromosome 1 in parent 2 is shown in Figure 1.

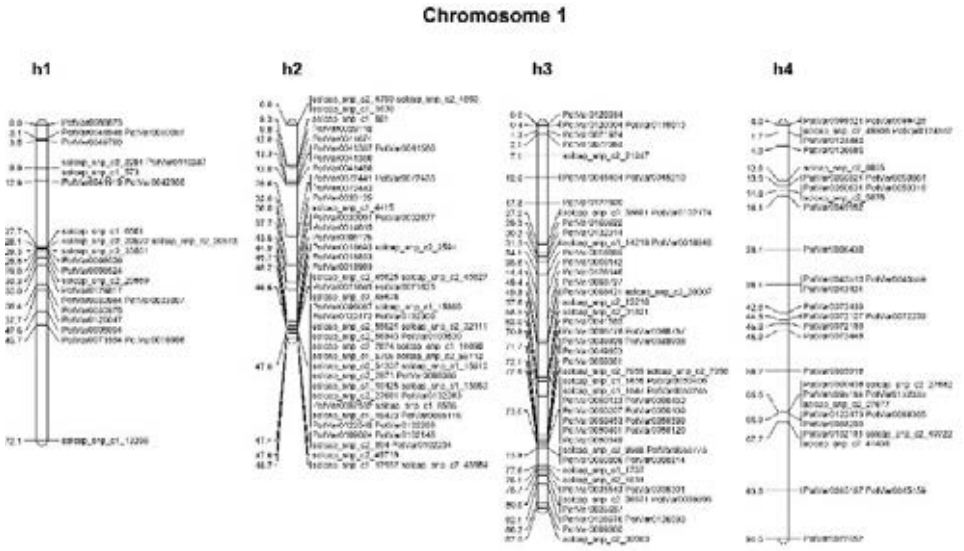



**Figure 1. Homologue linkage maps of potato chromosome 1 for parent 2**

In total, 30 mapped markers were found to have a discrepancy between their assignment to a linkage group in this population and their assigned chromosome on the physical sequence (Felcher et al., 2012;Vos et al., 2015). Of these, two solcap markers (solcap_snp_c2_42265 and solcap_snp_c2_32337) were found to have positions at two physical locations but mapped to a single genetic position. A further 25 mapped markers

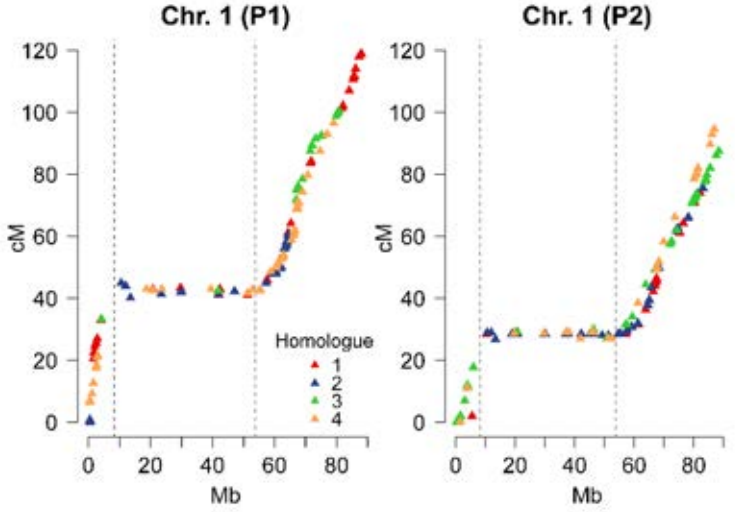



**Figure 2. Comparison of genetic to physical distance with homologue maps of potato chromosome 1.** Approximate centromere bounds are shown as dashed lines, corresponding to the inflection points in the curve (averaged over P1 and P2). Homologue maps were aligned prior to graphing by re-defining the 0 cM positions if necessary.

**Table 3. Composition of parental homologue maps**

| | Parent 1 | | | | | Parent 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Chm** | **h1**[a] | **h2** | **h3** | **h4** | **Total**[b] | **h1** | **h2** | **h3** | **h4** | **Total**[b] |
| **1** | 98.4 (44) | 60.8 (34) | 67.3 (26) | 89.9 (54) | *158* | 72.1 (25) | 48.7 (61) | 87.5 (60) | 94.5 (37) | *183* |
| **2** | 71.5 (44) | 76.7 (34) | 56.0 (31) | 46.0 (46) | *155* | 76.6 (74) | 71.7 (17) | 76.7 (55) | 62.9 (89) | *235* |
| **3** | 91.5 (17) | 59.0 (23) | 56.4 (12) | 86.0 (57) | *109* | 53.7 (110) | 26.6 (26) | 60.2 (47) | 62.0 (42) | *225* |
| **4** | 20.3 (5) | 95.1 (98) | 91.9 (21) | 69.4 (20) | *144* | 52.9 (32) | 70.6 (37) | 113.2 (19) | 66.3 (46) | *134* |
| **5** | 75.1 (32) | 74.3 (101) | 114.7 (9) | 66.3 (50) | *192* | 60.4 (69) | 68.6 (34) | 74.0 (78) | 83.7 (15) | *196* |
| **6** | 67.7 (12) | 76.6 (35) | 75.4 (15) | 69.9 (14) | *76* | 61.3 (24) | 59.8 (51) | 53.2 (28) | 66.1 (50) | *153* |
| **7** | 72.4 (21) | 60.0 (36) | 57.3 (13) | 55.4 (35) | *105* | 51.2 (12) | 55.8 (48) | 66.8 (28) | 50.5 (46) | *134* |
| **8** | 61.2 (40) | 58.1 (99) | 58.6 (20) | 56.1 (27) | *186* | 60 (12) | 69.9 (37) | 48.8 (31) | 66.3 (24) | *104* |
| **9** | 97.0 (8) | 78.8 (62) | 86.9 (24) | 101.8 (23) | *117* | 72.6 (15) | 71.8 (39) | 70.9 (12) | 68.1 (41) | *107* |
| **10** | 66.4 (22) | 64.8 (18) | 58.2 (13) | 64.0 (30) | *83* | 45.7 (5) | 45.3 (4) | 75.1 (11) | 55.9 (23) | *43* |
| **11** | 59.7 (34) | 50.3 (37) | 61.1 (22) | 56.4 (44) | *137* | 44.5 (23) | 59.0 (34) | 77.5 (21) | 53.4 (51) | *129* |
| **12** | 33.9 (15) | 77.6 (27) | 73.1 (20) | 52.2 (20) | *82* | 54.1 (8) | 61.1 (39) | 11.9 (19) | 23.6 (20) | *86* |
| | | | | | **1544** | | | | | **1729** |

h1, homolog 1; h2, homolog 2; *etc.*

[a] Homolog map lengths in centiMorgans using Haldane's mapping function, with number of mapped markers in brackets.

[b] Total number of mapped markers.

were found to have an unknown physical position from the published datasets of marker positions (Felcher et al., 2012;Vos et al., 2015). We provide a list of these markers with their mapped positions in Supplementary table 1. None of the 30 markers which showed linkage-group discrepancies were included in the analysis of cM / Mb conversion rates or double reduction, but they were included on the final genetic maps due to their unambiguous genetic position.

## Position of the centromeres

A graphical comparison of the aligned genetic and physical maps allowed an estimation of the centromeric bounds (Figure 2). When compared to previously-published centromere boundaries (Sharma et al., 2013) the results do not correspond precisely for chromosomes 4, 5, 7, 9, 10, 11 and 12. It is possible that the discrepancies are due to the fact that our estimates are based on a tetraploid population rather than a diploid one (Felcher et al., 2012;Sharma et al., 2013) since the method used to determine the boundaries was essentially the same. Supplementary table 2 provides our estimates for the centromere bounds, used in the calculation of relative distance from the centromere for the double reduction analysis.

## Conversion rate between genetic and physical distance

The cM / Mb conversion rate was determined per homologue arm across all chromosomes in both parents by linear regression of genetic distance on the physical distance (Figure 3). Apart from one clearly outlying value (due to insufficient marker coverage) the recombination rate was found to be relatively constant across all chromosomes, with an average value of 3.07 ± 0.09 (standard error of the mean).

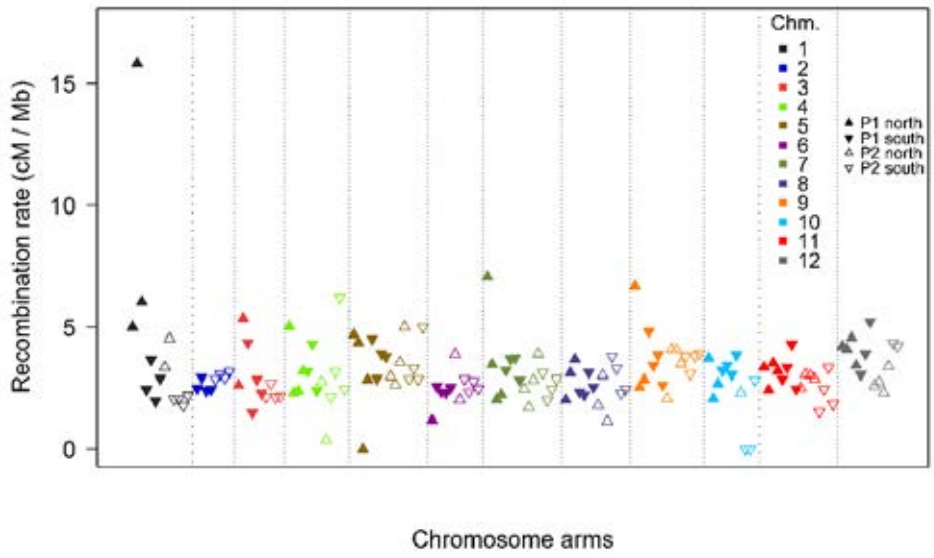



**Figure 3. Average recombination rate across homologous chromosome arms**. Rates calculated per homologue arm (north or south of the centromere) by linear regression of marker positions on the physical versus genetic distance plots. Points are colored by chromosome, with upward-pointing triangles denoting north (p) arms and downward denoting south (q) arms. P1 data is shown by filled triangles, P2 data by empty triangles.

## Double reduction

Double reduction events were identified on all twelve chromosomes, suggesting that multivalent pairing structures can form among all potato chromosomes. Of the 235 individuals in the mapping population, 112 (47.7%) showed evidence of double reduction coming from P1 meioses and 89 (37.9%) showed double reduction segments from P2. Forty-six individuals showed evidence of having inherited a double reduction segment from both parents (but not necessarily from the same chromosome), which corresponds well with the 42.5 individuals expected under independence of parental meioses. The distribution of duplex string lengths shows that singleton duplex scores predominate in this dataset (Supplementary Figure 1). Here we have chosen to consider singleton duplex scores as unsupported evidence for double reduction which cannot be distinguished from errors in dosage estimation.

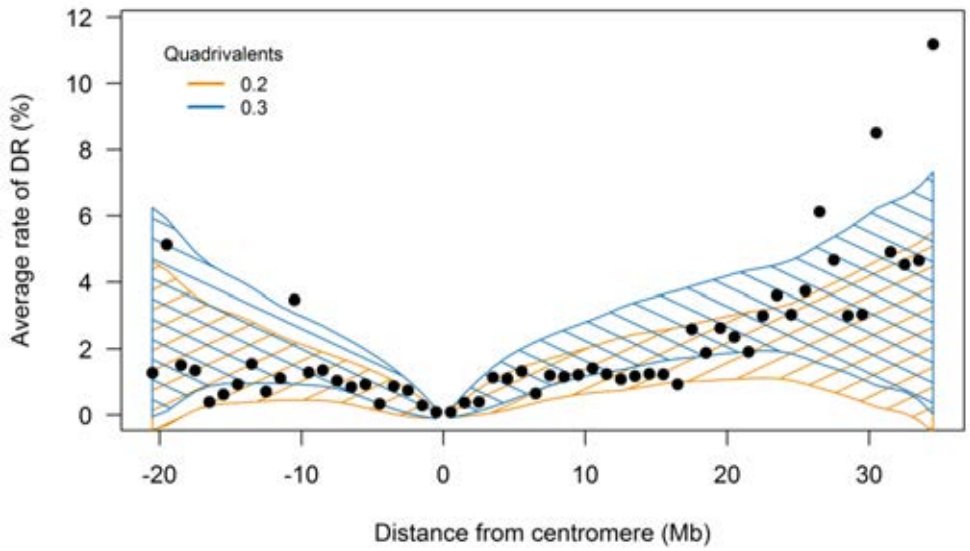



**Figure 4. Average rate of double reduction versus distance from the centromeres**. Shaded areas represent 95% confidence regions around the simulated mean rate of DR arising from a fraction quadrivalents of 0.2 and 0.3. The standard deviation of the simulated mean rate of DR increases towards the telomeres, coinciding with greater fluctuations in the true rate of DR in these regions.

We also used an algorithm which allowed for possible missing scores within a string of duplex values. Using this approach, we were able to reveal the relationship between double reduction and the average distance from the centromere (Figure 4) by pooling the estimates from all 96 homologue maps, giving the average rate of double reduction as a function of distance from the centromere. The rate of double reduction close to the centromeres approaches zero while towards the telomeres it increases substantially. Within the centromeres themselves there were twenty-two P1 markers and five P2

markers with duplex scores in the offspring. Of these, 18 were single occurrences which were probable errors (for example, the centromeric marker "PotVar0014900" which mapped to chromosome 1, homologue 4 in P1 gave five separate duplex scores. This marker was also found to have 16.2% missing values, suggesting a lower reliability. Other isolated cases would require a double recombination at both sides of the markers, which is highly unlikely to have occurred). There remained five cases of longer strings of duplex scores which partially entered the boundaries of the centromeric regions (Supplementary table 3), suggesting that in a very limited number of cases, recombination may occur within what is considered to be a non-recombining region.

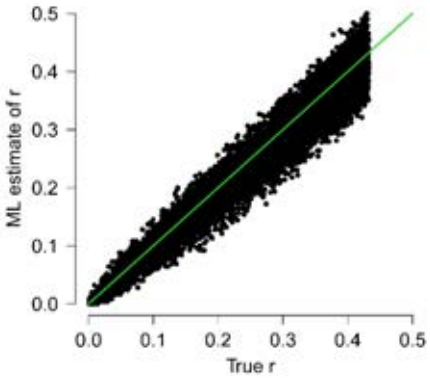



**Figure 5. True versus estimated *r* (using maximum likelihood) for coupling-phase simplex markers with fraction quadrivalents 0.2.** The straight line ($y = x$) shows the line of perfect correspondence between true and estimated values.

PedigreeSim has previously been used to determine the rate of double reduction in simulated populations and to visualise the relationship between (genetic) distance from the centromere and double reduction (Voorrips and Maliepaard, 2012). In this study we simulated phased marker data and a mapping population size of 235 in order to empirically fit a pairing model to the observed data. The observed rates of double reduction and those predicted by simulation overlap well when the fraction of quadrivalents was simulated in the range 0.2 – 0.3. Towards the telomeres the average rate of double reduction exceeded the expected rates (within a 95% confidence interval), although the confidence intervals were found to widen greatly in these regions. This may be due to the limited number of markers at these distances from the centromeres, causing greater uncertainty in the estimates.

## Evidence for preferential pairing

Using the repulsion-phase marker data, we investigated whether there was any evidence for preferential pairing in this population. We found almost no evidence for preferential pairing (correcting for multiple testing using the FDR correction). On chromosomes 5 and 8 in P1 there were four marker pairs (out of 18336 and 17205 pairs, respectively) which did show possible evidence of disomic pairing, but this was not considered strong enough evidence to support a hypothesis of preferential pairing. In P2, no markers

displayed disomic-like behavior. It was therefore concluded that potato follows tetrasomic inheritance as is generally assumed.

## Effect of quadrivalents on mapping of simplex markers

Our analysis of double reduction suggests that quadrivalents may account for between 20% and 30% of all meiotic pairing configurations in this population. Given that previous mapping studies in potato have assumed that the rate of quadrivalent formation is negligible, we wanted to examine what effect quadrivalents have on recombination frequency estimates (and hence on linkage mapping). We compared pairwise ML estimators for *r* to their true underlying value (Figure 5) for different rates of quadrivalents. Overall, the effect of quadrivalents on coupling-phase estimates for simplex marker pairs was relatively minor, as shown by the gradual decrease in the slope of the regression between the true and estimated values (Figure 6.b). Correct phasing in the coupling phase was also unaffected by quadrivalents (Figure 6.a). For a quadrivalent rate between 0.2 and 0.3, the effect on coupling-phase estimates can likely be ignored. For repulsion-phase marker pairs, a greater effect was found although remarkably, the assignment of marker phasing actually improves slightly with higher numbers of quadrivalents (Figure 6.a). Of the 2374 incorrect repulsion phase assignments in the purely bivalent situation, only 14 had an associated LOD score greater than one. This suggests that as a precaution against incorrect phase assignment within a linkage group, an "unknown" phase be assigned in cases where the LOD falls below a certain threshold (*e.g.* LOD of one).

## Effect of preferential pairing on mapping of simplex markers

Our study on the effect of preferential pairing on estimates of *r* revealed that preferential pairing has no effect on these estimates in coupling phase but has a dramatic impact in repulsion phase (Figure 7.b). This fact has already been reported (Qu and Hancock, 2001;Koning-Boucoiran et al., 2012) and forms the basis for a test of preferential pairing which we also exploit in this study. It is evident that preferential pairing can have a severe impact on the correct assignment of repulsion phase (Figure 7.a), regardless of whether MINR or MLL is used for phase-assignment (data not shown). Since we found no evidence to suggest that any systematic preferential pairing occurred we can be fairly confident that the estimates for recombination frequency and phase were accurately performed, as confirmed by the simulation study.

# Discussion

## Linkage maps

A recent publication describes the methods used to produce a high-density SNP linkage map of a well-studied tetraploid mapping population (Hackett et al., 2013) using the Infinium 8300 SolCap array (Felcher et al., 2012). Although we have not attempted to include all marker types in the current linkage maps, we have mapped a large number of markers (3273) in a tetraploid population which to the best of our knowledge is the highest-yet reported marker density of a tetraploid potato map.

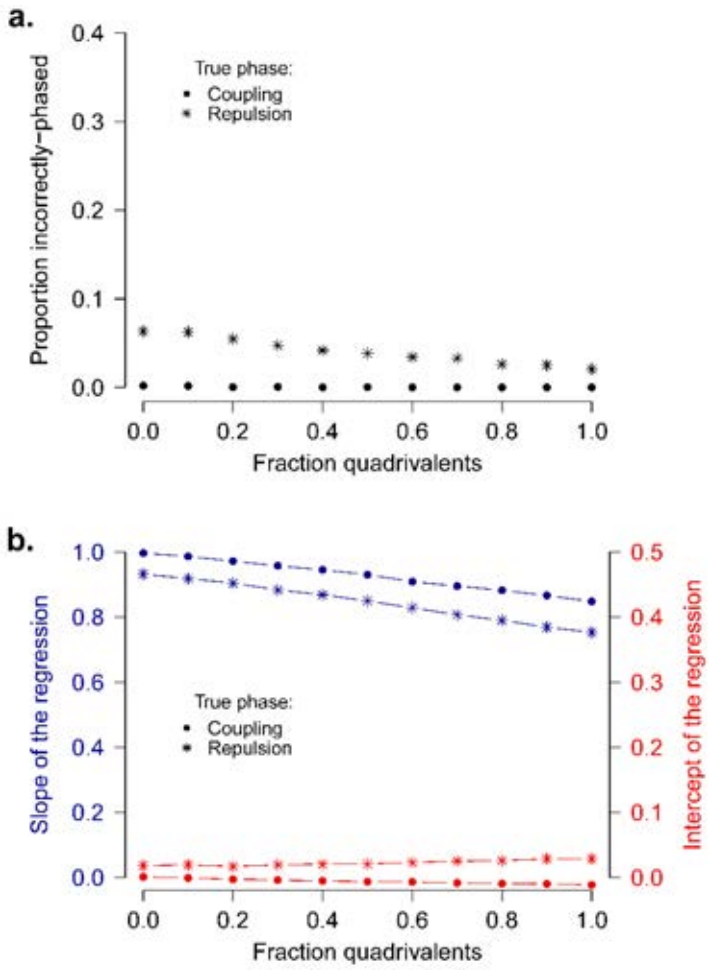



**Figure 6. Effect of quadrivalents on linkage analysis. a.** Proportion of incorrectly-phased markers pairs under different levels of quadrivalent formation**. b.** Effect of quadrivalents on accuracy of *r* ML estimates for coupling and repulsion-phase simplex marker pairs.

This has given us adequate coverage to recover all homologous chromosomes and develop an accurate picture of the double reduction landscape in this tetraploid species. We have presented separate homologue maps rather than a single consensus integrated map per chromosome as achieved by (Hackett et al., 2013). Separate homologue maps give one the ability to infer the phasing of markers directly from the map (long-range haplotyping) without recourse to hidden Markov models (Hackett et al., 2013), although ultimately integrated maps and genotype probabilities estimated using the integrated map will lead to greater power in subsequent QTL studies.

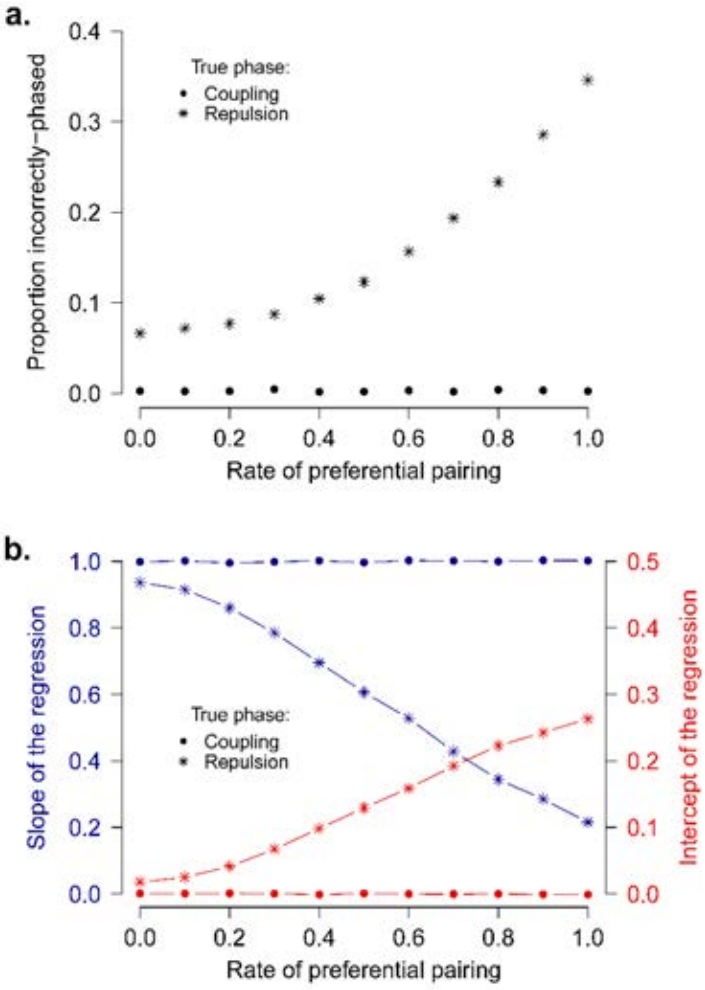



**Figure 7. Effect of preferential pairing on linkage analysis. a.** Proportion of incorrectly-phased marker pairs with different levels of preferential pairing. **b.** Effect of preferential pairing on accuracy of *r* ML estimates for coupling and repulsion-phase simplex marker pairs.

Our finding that the large-scale conversion rate between genetic and physical distance is essentially constant outside the centromeric regions (genome-wide recombination rate) has shown that the prospects for integrating maps across homologues and between parents are good and should not impose undue stress on the underlying homologue maps. We also found little evidence of recombination "hot-spots" or "cold-spots" outside the centromeres, as evidenced by the high $R^2$-values associated with our genetic-physical distance regressions (Supplementary table 4).

## Potato cytology

Information on the pairing behavior of polyploids has traditionally been generated from cytological studies. One of the more influential publications on potato cytology has been the 1953 review of Swaminathan and Howard who summarised the findings of previous researchers such as Cadman, Lamm and Bains for the mean number of multivalents per cell at diakinesis and first metaphase in tetraploid *S. tuberosum*, ranging from 1.70 to 5.24 (Swaminathan and Howard, 1953). This cytological evidence has been used to support the use of a simplified pairing model in potato mapping and QTL analysis since then (Hackett et al., 2001;Luo et al., 2001;Hackett et al., 2003;Bradshaw et al., 2008;Hackett et al., 2013). In our study we have used marker data to estimate the rate of double reduction and from this to extrapolate the likely frequency of multivalents (we only consider quadrivalents) involved. A fraction of 20-30% quadrivalents translates to between 2.4 and 3.6 quadrivalents per cell, consistent with the original cytological findings of Lamm performed on the cultivar 'Deodara' and the line '36/209' from the cross Greta x Fürstenkrone in 1945 (Lamm, 1945).

## General polyploid model

Attempts have previously been made to develop a general theory of linkage mapping in tetraploids which simultaneously considers the possibility of preferential pairing and multivalent formation (Wu et al., 2004). According to the authors, if the preferential pairing factor is set to 0 (for the case of random pairing) their model implies that the fraction of quadrivalents will equal 2/3 and that of bivalents 1/3. This is consistent with the random-end pairing model that assumes pairing initiation occurs at one set of telomeres, with probability of 1/3 that the pairing at the other telomeres will result in a separation into bivalents (John and Henderson, 1962). Our data shows that preferential pairing does not occur in potato yet we have not found a fraction of quadrivalents as high as 2/3. Our findings on quadrivalent pairing are in line with a previous review of autopolyploid meiosis which found a mean multivalent frequency (trivalents and quadrivalents) of 28.8% over 93 different studies (Ramsey and Schemske, 2002). It has also been shown that low numbers of multivalents does not necessarily suggest that preferential pairing behavior occurs (Sybenga, 1992;Sybenga, 1994).

## Identification of double reduction

We decided to take a more stringent approach than studies which consider two or even a single locus as sufficient evidence for double reduction (Luo et al., 2006;Hackett et al., 2013). This is likely to have led to an under-estimation of DR on our part. However, all quantification of DR using marker data are likely to under-estimate the true rate of double reduction to some extent. For instance, DR segments can be hidden (no segregating allele carried on the segment), or due to limited numbers of markers one might only recover part of a double reduction segment. Higher-density linkage maps (where all homologue parts are covered by segregating markers) will lead to more accurate estimates of the rate of double reduction, unless a strong bias exists in how markers are distributed or where DR occurs. In this study, with over 3000 well-distributed simplex x nulliplex markers, we feel we have sufficient marker coverage for a detailed understanding of the double-reduction landscape.

Simplex x nulliplex markers give the most unambiguous information about the presence of double reduction when compared with other marker segregation types. Other marker classes could have been used as well (for example simplex x simplex markers, which are expected to show triplex scores in 50% of the cases of double reduction involving one of the simplex alleles). However, no marker class other than simplex x nulliplex allow DR scores to be distinguished directly as a double-reduction product. Maximum likelihood approaches that estimate the rate of double reduction such as that described in (Luo et al., 2006) may be useful for the identification of double reduction in cases where it is not clear, although we feel that flanking simplex x nulliplex marker information which supports the duplex score should be used as we have done here.

## Double reduction increases towards the telomeres

It has been widely reported that the rate of double reduction is expected to increase towards the telomeres (Mather, 1936;Fisher, 1947;Butruille and Boiteux, 2000;Stift et al., 2008;Nemorin et al., 2012;Zielinski and Scheid, 2012), given that the probability of a cross-over occurring between the centromere and a locus should increase as that locus is situated further from the centromere. Nevertheless, it has only rarely been experimentally verified. The clearest evidence we found in the literature came from an analysis of tetraploid potato using isozyme markers, although the numbers of markers used were rather few, with less than 50 loci considered (Haynes and Douches, 1993). In our study we have clearly shown, using high-density marker data of over 3000 markers, that the rate of double reduction steadily increases with distance from the centromere. We have furthermore been able to visualise this phenomenon which has not previously been reported.

The fact that the frequency of double reduction increases towards the telomeres is perhaps cause for some concern as this could be considered a systematic source of error in the marker data. Nevertheless, with dense marker data it is now possible to accurately estimate the rate of double reduction in a mapping population. In cases where the rate of double reduction is low and marker number high, it is questionable whether highly complicated models with many parameters to be estimated are actually useful, particularly if they do not distinguish between singleton double reduction scores and genotyping errors. Our simulations have shown that even with fully quadrivalent pairing, pairwise estimators for recombination frequency between coupling-phase simplex x nulliplex markers under a bivalent pairing model are close to being exact (and as these are the most informative pairing scenario, they are the most important estimates for linkage map construction). We look forward to comparing our estimates for double reduction in tetraploid potato with other polyploid species and in gaining a deeper understanding of why these rates differ in what are otherwise classified collectively as autopolyploids.

## Double reduction in mapping

Some authors claim that double reduction should be included in map estimation and QTL analysis to increase the power and accuracy of the analysis (Li et al., 2010). Our findings show that quadrivalents have little effect on the mapping of simplex markers in the highly-informative coupling-phase. In potato at least, our data shows that the level of quadrivalent formation (and preferential pairing) is very low and is therefore not likely to be of serious worry for linkage mapping. However, confirmation of this finding for other marker types is still needed.

It is also worth pointing out that quadrivalent formation not only leads to double reduction but can also result in the formation of homologue combinations of more than two parental homologues (Sved, 1964) which can result from pairing-partner switches (Jones and Vincent, 1994) along the chromosome. The fact that this is already part of the simulation process in PedigreeSim (Voorrips and Maliepaard, 2012) increases the accuracy of our approach, not only in terms of modelling double reduction but also in our study of the effect of quadrivalents on map estimation.

## Double reduction in breeding

Double reduction has many implications for polyploid breeding. One consequence that has been described is its potential to lead to a higher inbreeding coefficient in dihaploids derived from tetraploid lines (Haynes and Douches, 1993). Given the efforts currently underway towards hybrid potato breeding (Lindhout et al., 2011), DR may have unwanted impacts on genetic diversity at the diploid level if future diploid founder

material is derived from tetraploid lines. On the other hand, hybrid breeding is dependent on the production of highly-homozygous inbred lines. Tetraploid potato breeding might welcome greater levels of homozygosity in a crop that is often complicated by high heterozygosity (Uitdewilligen et al., 2013), as well as the potential purging effect that DR can have by exposing deleterious alleles to selection (Butruille and Boiteux, 2000). DR could also speed up the accumulation of rare but favorable alleles through marker-assisted selection. Here we have developed the tools for the identification of DR in a segregating population which could be applied by breeders in the selection of founder parents for subsequent crossings or for confirmation studies of QTL positions.

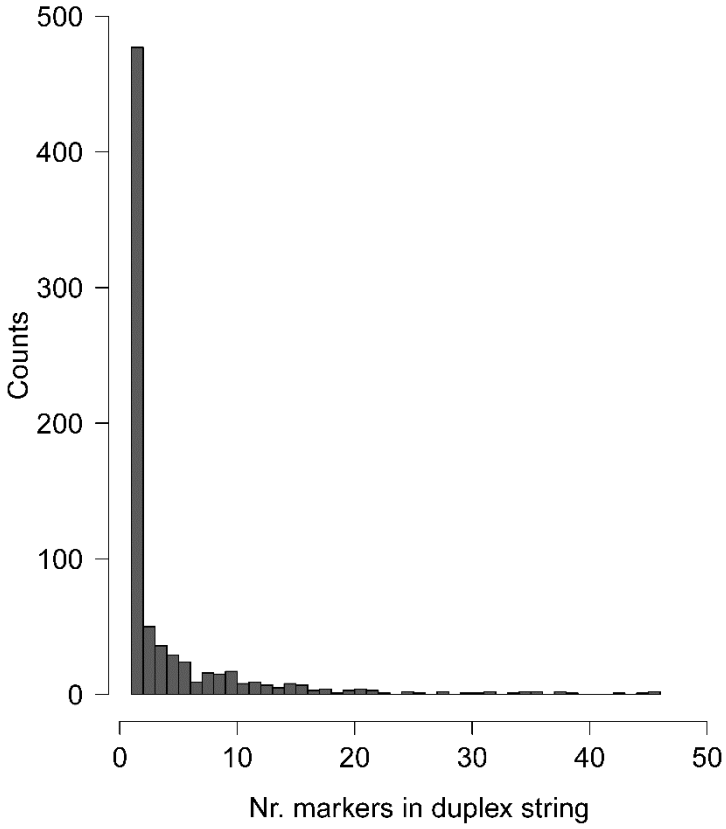
# Conclusions

In this study we constructed 96 separate homologue linkage maps of tetraploid potato using 1:1 segregating simplex markers. We estimated the approximate rate of double reduction (10% or more at the distal regions) and predicted by simulation that a fraction quadrivalents of 20 – 30% is required to account for this level of double reduction. We found no evidence of preferential pairing in our data, consistent with previous reports on the mode of inheritance in potato. Simulation studies using simplex x nulliplex markers revealed that marker phasing and recombination frequency estimation under a simplifying bivalent-pairing model are relatively robust, even when some level of multivalent pairing occurs.

# Acknowledgements

The authors would like to acknowledge Peter Vos for assistance with genotype calling and HZPC and Averis for providing potato varieties, as well as all partners involved in the TKI polyploids project “A genetic analysis pipeline for polyploid crops” (project number BO-26.03-002-001) which helped fund this research. The authors would also like to thank Jeffrey Endelman for helpful comments leading to a correction in the original manuscript. The development of the SolSTW SNP array was financially supported by a grant from the Dutch technology foundation STW (project WPB-7926).

***Supplementary data is available online at:***
http://www.genetics.org/content/201/3/853.supplemental




**Supplementary Figure 1.** Histogram showing distribution of lengths of strings of duplex scores in the dataset

# Chapter 4

# Integrating haplotype-specific linkage maps in tetraploid species using SNP markers

Peter M. Bourke[1], Roeland E. Voorrips[1], Twan Kranenburg[1], Johannes Jansen[2], Richard G. F. Visser[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[2] Biometris, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.

*Published as* Bourke, P.M., Voorrips, R.E., Kranenburg, T., Jansen, J., Visser, R.G., and Maliepaard, C. (2016). "Integrating haplotype-specific linkage maps in tetraploid species using SNP markers", **Theoretical and Applied Genetics 129 (11), 2211-2226**

# Abstract

High-density linkage mapping in autopolyploid species has become possible in recent years given the increasing number of molecular markers now available through modern genotyping platforms. Such maps along with larger experimental populations are needed before we can obtain sufficient accuracy to make marker-trait association studies useful in practice. Here, we describe a method to create genetic linkage maps for an autotetraploid species with large numbers of markers and apply it to an $F_1$ population of tetraploid potato (*Solanum tuberosum* L.) of 235 individuals genotyped using a 20K SNP array. SNP intensity values were converted to allele dosages after which we calculated pairwise maximum likelihood estimates of recombination frequencies between all marker segregation types under the assumption of random bivalent pairing. These estimates were used in the clustering of markers into linkage groups and their subsequent ordering into 96 homologue maps. The homologue maps were integrated per chromosome, resulting in a total map length of 1061 cM from 6910 markers covering all 12 potato chromosomes. We examined the questions of marker phasing and binning and propose optimal strategies for both. We also investigated the effect of quadrivalent formation and preferential pairing on recombination frequency estimation and marker phasing, which is of great relevance not only for potato but also for genetic studies in other tetraploid species for which the meiotic pairing behaviour is less well understood.



# Key words

High-density linkage mapping, autotetraploid, random bivalent pairing, haplotype map, marker phase, map integration.

# Introduction

Polyploid species, where the basic diploid number of chromosome copies is exceeded, are increasingly becoming the subject of studies that aim to determine the links between genetic polymorphisms and phenotypic traits. In order to do this, researchers have needed to create maps of these species through linkage studies or sequencing efforts (or both). Affordable, high-throughput genotyping technologies together with greater computing power and the software needed to assemble these maps are enhancing our ability to perform such studies.

There have been a relatively large number of published tetraploid linkage maps in economically-important allotetraploid species such as cotton (*Gossypium hirsutum* L.) and durum wheat (*Triticum durum* L.). In contrast, autotetraploid maps are far fewer, with the exception of alfalfa (*Medicago sativa* L.) (Brouwer and Osborn, 1999;Robins et al., 2008), potato (*Solanum tuberosum* L.) (Meyer et al., 1998;Hackett et al., 2013) and rose (*Rosa hybrida*) (Rajapakse et al., 2001;Koning-Boucoiran et al., 2012).

Methods for estimating marker dosage using (for example) SNP array data (*e.g.* fitTetra (Voorrips et al., 2011) or SuperMASSA (Serang et al., 2012)) have enabled researchers to exploit marker dosage information to generate polyploid linkage maps with a much higher marker density than before. Given the abundance of such marker sets, many polyploid maps continue to rely on 1:1 segregating markers, for which the coupling-phase recombination frequency estimates are identical to those for diploid species (Bertioli et al., 2014;Bourke et al., 2015;Yu et al., 2015;Vigna et al., 2016) (repulsion-phase estimates are not the same between species showing disomic inheritance, such as diploids, and those with polysomic inheritance, such as autotetraploids). However, there are many more marker segregation types that can be considered which may provide greater genome coverage as well as providing links between parental maps, important for subsequent analyses. In a tetraploid cross genotyped with bi-allelic markers for which dosage scores are available (assuming an absence of null alleles), there are nine fundamental marker segregation types: simplex x nulliplex (SxN), nulliplex x simplex (NxS), duplex x nulliplex (DxN), nulliplex x duplex (NxD), simplex x simplex (SxS), simplex x triplex (SxT), duplex x simplex (DxS), simplex x duplex (SxD) and duplex x duplex (DxD), according to the marker dosages carried by both parents. All other marker segregation types can be converted to one of these categories (Supplementary Table S1). These nine fundamental types have also been identified in previous studies, *e.g.* Hackett et al. (2013). Recently, methods to incorporate all marker segregation types from a tetraploid cross have been developed (Hackett et al., 2013). However, these methods do not automatically generate homologue maps, as these must be derived using chromosomal maps and phase information (Hackett et al., 2013;Massa et al., 2015).

Here, ‘phase’ or ‘phasing’ means determining whether linked markers are physically on the same homologous chromosome within a parent. In our approach, we first develop separate maps for every parental homologous chromosome (termed ‘homologue’ here) using all marker segregation types, integrating them afterwards into one chromosomal map for each set of eight homologues. Marker phasing is thus an essential aspect of our approach, which is of importance in the development of marker haplotypes consisting of more than a single SNP marker.

Although methods to include double reduction in a linkage analysis have already been developed (either using two-point estimation (Luo et al., 2006) or multi-point estimation (Leach et al., 2010)), linkage analysis can be considerably simplified in autopolyploid mapping populations if it is assumed that only random bivalent pairing occurs. A review of metaphase I of autopolyploid meiosis found that bivalents accounted for approximately 70% of the pairing structures observed (Ramsey and Schemske, 2002) with quadrivalents accounting for approximately 29% (there were relatively few univalents or trivalents observed; more complex multivalents were not recorded at higher ploidy levels). Comparable rates have also been reported for potato (Swaminathan and Howard, 1953;Bourke et al., 2015). In the computations of this study, we have assumed a complete absence of preferential pairing and multivalent formation. For example, in the case of SxN markers, a duplex score in the offspring would effectively be treated as a missing value, as it is not an expected score according to our model. However, we took care to examine what effect both preferential pairing and multivalents might have on the pairwise estimation of recombination frequency as well as the effects on the accuracy of marker phasing.

Although broadly similar, our mapping approach differs from that of Hackett et al. (2013) in the following respects:

- The initial clustering of all marker segregation types into linkage groups is defined by their linkage to SxN or NxS (1:1 segregating) markers, enabling automatic marker phasing during mapping.
- Homologue maps are first created (per parent) and then integrated into a single consensus map per chromosome using linear programming.
- We include the results of a comparison study between two different methods for deciding the most likely phase.
- Criteria are determined for binning markers together before map ordering.
- All marker segregation types are included in the mapping (in particular, we also include SxT and TxS marker types).

In this study, we describe a method to perform linkage mapping in an autotetraploid species under the assumption of random bivalent pairing, and apply it to a genotyped mapping population of tetraploid potato. We explore some of the potential complications involved in polyploid mapping and discuss the implication of these for future mapping efforts.

# Materials and methods

## Plant material and genotyping

An $F_1$ population of 237 individuals from the cross between two outbred tetraploid cultivars '*Altus*' (parent 1 or P1) and '*Columba*' (P2) was genotyped using the SolSTW Infinium SNP array which assays 17,987 SNPs (Vos et al., 2015). Markers were assigned dosages using the fitTetra package (Voorrips et al., 2011) as previously described (Bourke et al., 2015). Highly-skewed markers (using a $\chi^2$ test with $p < 0.001$) as well as markers with more than 10% missing values were removed from the dataset. In a previous study, a pair of duplicate offspring individuals were identified in this population as well as an individual which showed unrealistic numbers of recombinations (Bourke et al., 2015). The suspect individual was removed as well as the duplicate with most missing values, leaving a mapping population size of N = 235 individuals.

## Marker dosage conversion

A small number of markers for which one of the parental dosages was missing were examined and the likely parental dosage imputed using the observed offspring segregation (if possible), after which marker dosages were converted to their most fundamental form. In a tetraploid species genotyped using bi-allelic markers, the possible marker dosages classes are 0 (nulliplex), 1 (simplex), 2 (duplex), 3 (triplex) or 4 (quadruplex) depending on the number of copies of the 'reference' allele carried by that individual. Marker conversion simplifies the analysis by reducing the number of marker types that need to be considered. For example, simplex x nulliplex, triplex x quadruplex, triplex x nulliplex and simplex x quadruplex markers all segregate in a 1:1 fashion and all carry a segregating allele inherited from parent 1 (P1). They can therefore be re-coded as SxN markers using suitable score conversions in the offspring (Supplementary Table S1). Ultimately, this results in nine fundamental marker segregation types as previously mentioned. In determining linkage between SxT markers and other markers in P2, the set of SxT markers were re-coded by symmetry into TxS to facilitate the calculations. The physical distribution of the segregating markers was

visualised in MapChart 2.3 (Voorrips, 2002) using the previously-published centromere boundaries (Sharma et al., 2013).

## Linkage analysis

The maximum likelihood framework for determining pairwise estimators for recombination frequency (*r*) and their significance (LOD) scores under the assumption of random bivalent pairing has already been described in Hackett et al. (2013). We independently derived the likelihood functions for all possible marker pairs and phases (of which we counted 92 possible combinations between the nine fundamental marker types mentioned) using routines written in Mathematica 10.0 (Wolfram Research Inc., 2014). We describe the procedure through a worked example in Appendix 1 (Supplementary File S1). The maximum likelihood functions (for each of these 92 cases) were coded in R (R Core Team, 2016) for use in the linkage analysis.

## Marker phasing

To explain the concept of phasing by way of example, there are three possible phases between a DxN and a DxS marker: 'coupling', 'mixed' and 'repulsion', with 'mixed' implying only one pair of Duplex alleles is in coupling phase in P1 (there is no phase consideration in P2 since one of the markers is Nulliplex in that parent). For pairs of markers with segregating alleles from both parents, such phases are combined (*e.g.* 'coupling mixed' refers to coupling phase in P1 and mixed phase in P2). One criterion for selecting the correct phase (and hence the correct estimator for *r*) is to use the maximum of the log-likelihood function between phases for which $r \leq 0.5$ (Hackett et al., 2013) which we refer to as MLL. Another possibility is to choose the minimum estimate of *r* over all phases (which we term MINR). We performed a simulation study to determine which of these criteria was optimal across all possible marker pair combinations using the simulation software PedigreeSim (Voorrips and Maliepaard, 2012). One hundred separate populations were generated for each of three population sizes ($F_1$ = 100, 200 and 400). Each simulated individual carried a single chromosome with 100 marker positions spaced 1 cM apart. All possible marker types were assigned to each of these loci, with a random assignment of the segregating alleles across homologues. For each simulated population, phasing accuracy was determined by recording the proportions of correctly-phased pairs using both the MLL and MINR phasing strategies. In a few cases it was not possible to distinguish between phases (*e.g.* SxS with DxD 'coupling-repulsion' and 'repulsion-coupling' phases produce precisely the same *r* estimates and LOD scores). In diploid species, an analogous situation can arise in cross-pollinated species where certain marker type combinations cannot be phased (*e.g.* AB x AB with AB x BA), in which case other marker segregation types are needed to complete the phasing (Maliepaard et al., 1997). We dealt with such instances

by considering both phases to be equally correct (since we do not use these particular phase assignments themselves, only their *r* and LOD values).

## Linkage group identification and marker clustering

Preliminary identification of linkage groups was performed by clustering the SxN (and NxS) markers, based on the LOD of the recombination frequency estimate between marker pairs. A routine for marker clustering was written in R using a grouping algorithm analogous to that employed by the JoinMap software (Stam, 1993;Van Ooijen, 2006;Van Ooijen and Jansen, 2013). Of the 1497 NxS markers (Table 1), three did not have any strong linkage to other NxS markers and were therefore removed at this stage (we were later able to 'rescue' one of them when more markers were assigned to chromosomes).

**Table 1. Tetraploid marker segregation types after filtering (adapted from Chapter 3)**

| Parental dosage | Segregation | Amount [a] |
|---|---|---|
| Simplex x nulliplex | 1:1 | 1690 |
| Nulliplex x simplex | 1:1 | 1497 |
| Duplex x nulliplex | 1:4:1 | 409 |
| Nulliplex x duplex | 1:4:1 | 442 |
| Simplex x simplex | 1:2:1 | 924 |
| Simplex x triplex | 1:2:1 | 410 |
| Duplex x simplex | 1:5:5:1 | 596 |
| Simplex x duplex | 1:5:5:1 | 665 |
| Duplex x duplex | 1:8:18:8:1 | 279 |
| **Total** | | **6912** |

[a] Number of SNP markers after marker conversions have been made

At a clustering threshold of LOD 4, the NxS marker data divided into 12 clusters. We visualised how clusters split across different LOD values for these 12 putative chromosomes in R, allowing the identification of tightly-linked sub-clusters (putative homologues or fragments thereof). In the majority of cases, these large clusters split into four sub-clusters at higher LOD values (as expected for a tetrasomic species). However, one cluster (of 15 markers) could not be further subdivided at higher LOD values. We therefore assumed that this cluster represented (part of) one homologue and used the repulsion linkage information to assign it to one of the other clusters. Another cluster broke down into two clear sub-clusters at LOD 5 (*i.e.* it contained two chromosomal groups), which further sub-divided into 9 sub-clusters at LOD 6. Clustering in P1 was more straightforward, with 12 clear chromosomal clusters emerging at LOD 4.

Cluster numbers were replaced with chromosome numbering for consistency with the reference physical map using marker positions given by (Vos et al., 2015). In P1, chromosomes 4 and 10 contained five sub-clusters with the rest having four. In P2, chromosomes 3, 5 and 9 were found to contain five sub-clusters at this stage, the rest having four. In cases where more than four sub-clusters are identified, a visualisation of cross-cluster phase assignments allowed us to quickly identify which sub-clusters were (albeit distantly) linked in coupling phase, resolving the SxN and NxS marker data into 12 x 4 x 2 = 96 linkage groups, the expected number of homologues. Following this, the vast majority of markers within the complete dataset were unambiguously assigned to homologue clusters using coupling-phase linkage with SxN markers (a single linkage above a LOD threshold of 3 was used as evidence of linkage, although in practice there were often hundreds of such linkages identified). SxN markers are extremely useful for this step as they unambiguously tag a single homologue. Where multiple assignments were possible, assignments with the greatest number of significant coupling linkages (LOD > 3) were chosen as the most likely linkage groups. For those markers which could not be completely assigned to the expected number of homologues in both parents due to poor linkage with SxN markers, linkage analysis was performed between these markers and all other marker segregation types to identify their most likely chromosome and homologue assignment. Finally, the marker data was split into twelve subsets (one for each chromosome) and a complete pairwise linkage analysis was run between all markers within each chromosome.

## Map construction

***Homologue mapping***

Per chromosome, there are eight homologues that can be mapped separately in an autotetraploid. Markers which appear on these homologues are already identified through their linkage with SxN (or NxS) markers. In addition to the coupling-phase linkages considered, we also included some repulsion-phase linkages in our homologue maps. Markers with at least one Duplex allele are completely symmetrical between the ‘reference’ and ‘alternative’ alleles in the Duplex parent. For example, DxN markers are initially assigned to two homologues, these being the homologues on which the reference allele can be found. However, it is equally informative to consider the pair of ‘alternative’ alleles from the same parent, as these carry the same linkage information as the reference alleles. We therefore used DxN markers in the mapping of four homologues, SxD and DxS for the mapping of five homologues and DxD markers for

all eight. All of these marker types (as well as SxS and SxT) are extremely useful as "bridging" markers for the integration of homologue maps.

There remained some linkages that we did not exploit, *e.g.* SxN and SxN in repulsion. The variance of these repulsion-phase estimates is high (and hence, LOD values are low), and therefore the added computation time from including these estimates is not worth the marginal increase in linkage information that they yield (a similar conclusion was arrived at in previous studies. *e.g.* (Ripol et al., 1999)).

***Marker binning***

Linkage information per homologue was first assembled into two pairwise matrices (one for *r* estimates and one for LOD scores), after which the strength of linkage was tested to determine whether marker binning was possible – *i.e.* markers with a small recombination frequency estimate (*r*) of low variance (thus high LOD) were binned together. The minimum (non-zero) number of recombinations that can be observed in a mapping population of size N (and hence 2N gametes) is one, in which case the smallest non-zero *r* estimate should be 1/(2N) (ignoring the influence of errors). Given an average missing value rate of $\mu$, a population size adjusted for missing values is approximately $N_a = (1-\mu)N$ and therefore $r_{min} \approx 1/(2(1-\mu)N)$. Estimates of *r* that were smaller than this value were taken as being below the threshold of minimum resolution ($r_{min}$).

Not all estimates of the recombination frequency are equally accurate. We therefore determined criteria for binning markers together with a high degree of confidence. To achieve this, we ran simulations using PedigreeSim (Voorrips and Maliepaard, 2012) and recorded the LOD scores as well as the range of true recombination frequencies for *r* estimates below the threshold of minimum resolution over a wide range of population sizes ($F_1$ = 100 to 1000 in steps of 100) and rates of missing values (0% to 20% in steps of 5%). We chose a maximum allowable deviation between the true and estimated *r* as 0.01 (approximately 1 cM) and determined the corresponding LOD score to ensure this over all possible marker pair combinations (hence we took the most stringent LOD threshold to cover all cases). For each population size and rate of missing values, we examined the distribution of LOD scores for those recombination frequency estimates for which the deviation was less than 0.01 (Supplementary Figure 1.b). From this we could determine a suitable LOD threshold for marker binning as a function of mapping population size and rate of missing data.

Given a set of markers that have been binned together, we chose the SxN marker with the fewest number of missing values for mapping (SxN recombination frequency estimates with other marker types are exact as opposed to being numerically

approximated); in bins with no SxN markers, the marker with the fewest missing values was selected.

***Marker ordering***

The remaining marker data (after binning) were converted into pairwise-data file format (.pwd) for each homologue and imported into JoinMap 4.1 (Van Ooijen, 2006) for ordering. Three rounds of mapping using the weighted least squares algorithm were used (using the default settings with a "jump threshold" of 5), and with Haldane's mapping function used for distance conversion. Map files were subsequently exported and all binned markers were re-added to the maps.

***Map integration***

The homologue maps were first re-orientated (if necessary) before integrating. Map re-orientation was achieved by locating bridging markers between the maps and determining the correlation between the cM positions of these markers. A negative correlation suggests that maps are orientated in reverse order relative to one another. Since not all homologues necessarily share bridging markers, the R package igraph (Csardi and Nepusz, 2006) was used to find an order of comparison through the eight homologue maps per chromosome to allow stepwise correlations to be calculated (for example 5-3-7-6-4-2-8-1 might be one such order). An example of this is shown in Figure 1 for chromosome 7. In this example, three separate re-orientations were required to ensure consistency in orientation. When all eight maps were similarly orientated, the R package LPmerge (Endelman and Plomion, 2014) was used to integrate them. LPmerge eliminates the minimum number of constraints in the marker order of the underlying maps (conflicts in order between maps) in order to generate what is called a "feasible system", and then uses linear programming to find the solution with the minimum error between the underlying maps and the integrated map (Endelman and Plomion, 2014).

## Map quality checks

We checked the quality of our linkage maps using three different approaches:

- Consistency between the integrated maps and the underlying homologue maps
- Comparison with the reported physical position of the mapped markers
- Comparison to other published tetraploid potato maps

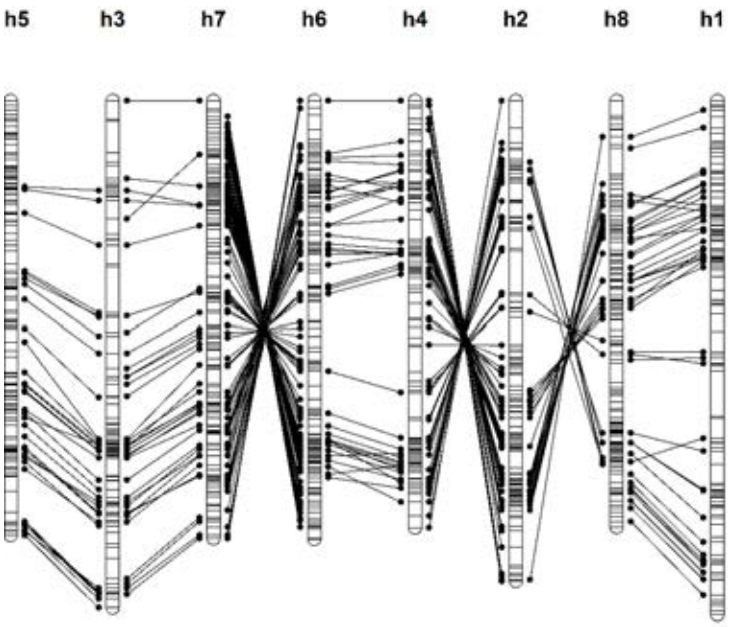



**Figure 1. Visualisation of map connections on chromosome 7.**
In this example, there were sufficient bridging markers to provide connections between all homologues. Three reversals were needed in order to ensure all homologues were consistently orientated before map integration.

***Consistency between the integrated maps and the underlying haplotype-specific homologue maps***

We compared the marker positions on the underlying homologue maps and the integrated maps in order to identify possible map distortion. Map distortion is partly reflected in the absolute error ($\delta$) between the underlying maps and the integrated maps. However, in cases where the telomeres of all homologues are not equally covered by markers, shifting the 0 cM position may occur to align the maps, contributing to an apparent increase in $\delta$ without implying any map distortion. We ran a simple linear regression between the integrated map positions (cM) and the underlying homologue positions (cM) and recorded the slope and adjusted $R^2$ values of the fit as well as visually inspecting each chromosome to identify potential distortion.

***Comparison with the physical position of the mapped markers***

The physical positions of the markers were taken as described in Vos et al. (2015). Plotting the genetic positions against the physical positions allowed us to identify whether our maps were correctly orientated (*i.e.* 0 cM corresponding to the lowest bp value) and we re-orientated our maps if necessary. In cases where a discrepancy was found between the reported chromosome and that found through our linkage analysis, we BLASTed the marker EST sequences (provided in the Supplementary Material of

Vos et al. (2015) and reproduced here) to the potato DM1-3 Pseudomolecules reference genome version 4.03 (Hirsch et al., 2014) to check the marker positions (website: http://solanaceae.plantbiology.msu.edu/blast.shtml, accessed 16.11.2015).

***Comparison to other published tetraploid potato maps***

The SolCAP 8303 Infinium array (Felcher et al., 2012) has been used to genotype at least two other published tetraploid potato mapping populations (Hackett et al., 2013;Massa et al., 2015). We compared the genetic map positions of the common markers as a further check on the validity of our maps.

## Simulation study to check mapping assumptions

Two crucial assumptions were made prior to mapping: that there is no preferential pairing behaviour between any homologues and that all pairing is between bivalents (as opposed to trivalents or quadrivalents), *i.e.* there is no double reduction. It had previously been established that the rate of quadrivalent pairing in this population was between 20-30% (Bourke et al., 2015). In contrast to previous studies which also rely on these assumptions, we wanted to test what effect deviations from these assumptions might have on our ability to produce unbiased and accurate estimates for *r* between marker pairs as well as the effects on phasing accuracy.

We simulated mapping populations using different degrees of quadrivalents and preferential pairing in PedigreeSim (Voorrips and Maliepaard, 2012). The simulation parameters we chose were: population sizes of 100, 200 and 400 $F_1$ offspring, levels of preferential pairing from 0 (fully random pairing or tetrasomic behaviour) to 1 (fully preferential pairing or disomic behaviour, associated with allopolyploidy) in steps of 0.1, or fraction quadrivalents from 0 to 1 in steps of 0.1. For each population size we generated 100 separate populations. Each simulated individual carried a single chromosome with 100 marker positions spaced 1 cM apart. All possible marker types were assigned to each of these loci, with a random assignment of the segregating alleles to homologues at all marker positions. In total, we simulated 6,600 populations to cover our chosen range of quadrivalents and preferential pairing for these 3 population sizes and number of repetitions ((11 + 11)*3*100).

After the populations were generated (*i.e.* their SNP dosage genotypes known), we ran our linkage analysis functions across all populations. Given that the datasets were simulated, the true recombination frequency and correct phasing between all marker positions was known, allowing us to test whether our estimation of recombination frequency and marker phasing was robust against these deviations from the random bivalent model. To generate average results from the 100 repeat populations per setting, we ran a simple linear regression on the estimated *r* versus true *r* values, recording the

slope, intercept, $R^2_{adj}$ and residual standard deviation of the regression. We also recorded the proportion of situations not estimated (*e.g.* due to undefined numbers in the likelihood equation) and the proportion of pairing situations that were correctly phased.

# Results

## Genotypes

The numbers of SNP markers that were available for mapping after marker filtering and quality checks was 6912, as outlined in Table 2. The breakdown of marker segregation types after marker conversions were performed is provided in Table 1.

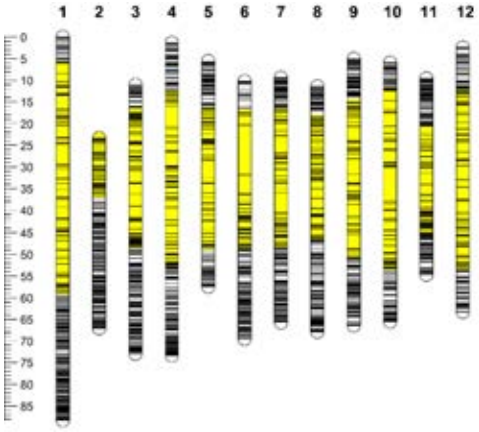



**Figure 2. Distribution of 6836 of the 6912 segregating markers used in this study for which a physical assignment was available.** Distances shown in Mbp. Shaded regions indicate centromeres, as previously defined (Sharma et al. 2013).

Approximately 46% of the markers segregate in a 1:1 fashion and these were mapped in a previous study (Bourke et al., 2015). The physical distribution of all marker types for which physical positions were available is given in Figure 2, highlighting the difference in marker distribution between telo- and centromeric regions.

**Table 2. Breakdown of SNP marker numbers after quality filtering (adapted from Chapter 3)**

| Steps in SNP filtering | Amount | % |
|---|---|---|
| SolSTW Infinium array total # SNPs | 17987 | 100.0 |
| Dosages assigned by fitTetra [a] | 15266 | 84.9 |
| $F_1$ pattern acceptable [b] | 13774 | 76.6 |
| - Monomorphic | 6558 | 36.5 |
| - Polymorphic | 7216 | 40.1 |
| Polymorphic and ≤ 10% NA values | **6912** | 38.4 |

[a] Markers not scored were monomorphic or not clearly resolved. Markers with a single missing parental dosage score which was imputed have been included.
[b] Criteria for lack of $F_1$ fit: presence of null alleles, > 3% invalid scores, highly-skewed segregation ($P < 0.001$)

# Linkage analysis and marker clustering

***Maximum likelihood pairwise estimates for r***

We counted 92 separate marker type and phase combinations for which we derived maximum likelihood functions, although this may contain scenarios that can be counted together (previously, 67 situations have been reported (Hackett et al., 2013)). For clarity we provide a table of all the possible marker type and phase combinations we considered (Supplementary Table S2). In many cases, the maximum likelihood equation cannot be solved analytically, in which case we used Brent's algorithm (Brent, 1973) to numerically estimate the recombination frequency with the highest likelihood, constrained to the interval [0,0.5]. It is also possible that a negative estimate for $r$ could be the most likely (this can occur in low-information situations, involving repulsion phases). We examined the true values of $r$ underlying such cases using simulated data and found a wide range of true $r$ values were possible. Whenever $r < 0$ was found, we artificially set $r = 0.499$, LOD = 0 and phase "unknown", thereby excluding these estimates from the map ordering step.

***Optimal phasing strategy***

Our simulations revealed the optimum phasing strategy to use for different marker combinations (Supplementary Figure 2). The maximum log likelihood strategy (MLL) as proposed by Hackett et al. (2013) proved in general to be a very good method of selecting the correct phase (and hence the correct estimate for $r$). In only one situation did we find MINR to outperform MLL, namely SxS with SxS markers. The improvement was, however, marginal: at a population size of 200 individuals for example, MINR gave 93.8% accuracy versus 90.9% accuracy using MLL. Whenever phase was incorrectly assigned, we found that the LOD score was also low (and the recombination frequency estimates tended to be high) – in other words, incorrect phasing only occurred in poorly-informative situations which would have little or no impact on the subsequent ordering step (especially since our mapping strategy favours coupling-phase estimates which tend to be more informative than those of repulsion-phase). In all situations involving at least one SxN marker there was no difference between the two methods. For the case of SxS paired with SxT markers, the accuracy of MINR appeared at first glance to be higher. However, this particular combination of markers contains an essentially un-estimable phase with extremely high variance (repulsion/coupling phase; see also Results section "Simplex x Triplex markers"). Removing this phase from the accuracy calculation, MLL was found to perform significantly better among the phases that actually *matter* in this combination. Finally, phasing accuracy was found to slightly increase as a function of population size, with 92% accuracy on average for a mapping population of size 400

(compared with 89% accuracy for a population of 200, and 84% for a population of 100). A breakdown of the phasing accuracy rates is provided in Supplementary File S2.

## Map construction and integration

***Simplex x Triplex markers***

SxT markers have previously been reported as problematic (where they are termed XSS markers (Hackett et al., 2013)). When we examined the issue, we found that SxT in combination with SxS produce highly variable estimates for *r* in repulsion/coupling phase, where the estimates for *r* are essentially random. They are the only marker combination (and phase) that exhibit such behaviour. We therefore artificially set LOD = 0 in this phase (it would be small but non-zero otherwise) which automatically excludes these estimates from exerting any influence on map ordering.

***Marker binning***

Once all linkages between markers had been estimated we were in a position to identify co-segregating markers. The *r* and LOD estimates give a convenient measure of linkage which can be applied across all marker segregation types in a binning procedure. An example of the relationship between *r* and LOD for pairs of DxD markers is visualised in Figure 3. As higher LOD values correspond to a lower standard error in *r*, we wanted to define thresholds for *r* and LOD which would identify markers which co-segregate with a high degree of confidence.

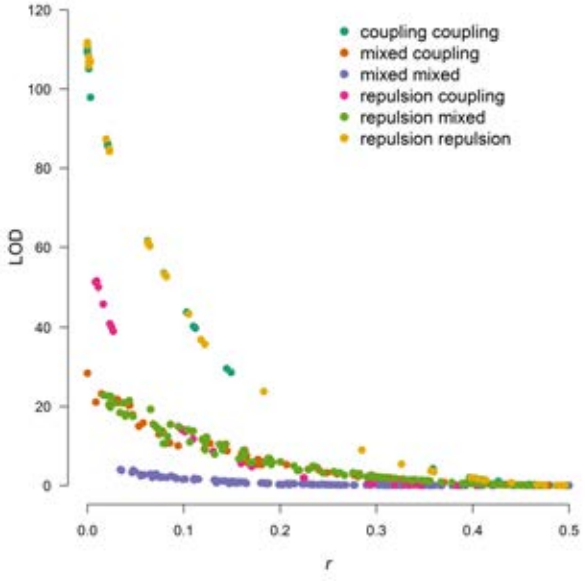



**Figure 3. Estimated recombination frequencies (*r*) versus associated LOD for duplex x duplex marker pairs on potato chromosome 1.** Only 6 of the 9 possible phases were identified for marker pairs on this chromosome.

We previously introduced the concept of the threshold of minimum resolution for recombination frequency, $r_{min}$. Given our mapping population size and missing error

rate (in the filtered dataset) we estimated $r_{min}$ to be approximately 0.0022, the smallest non-zero recombination frequency we should be able to observe. From our simulation of different population sizes and rates of missing data, we observed a clearly linear relationship between the mapping population size and the LOD threshold needed to ensure a margin of error of less than 0.01 in the estimation of *r* (Supplementary Figure 1.c). By performing a linear regression between the LOD thresholds and the adjusted mapping population sizes ($N_a$), we were able to derive an empirical relationship between a binning LOD threshold and the adjusted population size which ensures this margin of error in *r* is not exceeded:

$$LOD \approx 23.43 + 0.1158N_a$$

Given our dataset, we binned markers together if we found that the pairwise *r* estimate was less than 0.0022 and the LOD for that estimate exceeded 50.4. In total, 10,649 markers were used in the map ordering step across 96 separate homologue maps (note that some markers were present multiple times because they are present on multiple homologues), after which 7099 binned markers were re-assigned a position to give a total of 17,748 map positions (Supplementary Table S3). As binning was performed using a nearest-neighbour clustering, there is a danger that binned markers might have a non-negligible distance between them. When we examined this, we found that the maximum recombination frequency estimate between binned markers was 0.031, or approximately 3.3 cM using Haldane's mapping function (Supplementary File S5). However, the mean inter-marker distance within bins was only 0.12 cM, and almost 99% of binned markers were less than 1 cM from each other. In other words, our binning strategy rarely appears to have falsely binned markers together. These 17,748 map positions represented 6910 unique marker loci, *i.e.* only two of the 6912 markers available for mapping were not mapped. We suspect that the two unmapped markers which showed no linkage may have harboured abnormally-high numbers of errors, although we cannot verify this. A full list of all marker positions per chromosome is provided in Supplementary File S3.

***Map integration***

LPmerge (Endelman and Plomion, 2014) was run for all 12 linkage groups to determine the integrated maps with lowest absolute error $\delta$ between the homologue maps and the integrated map per linkage group (referred to as RMSE by the authors). Although the LPmerge algorithm is deterministic (J. Endelman, personal communiation), we found the resulting maps differed when the input maps were flipped. This suggests the current version of LPmerge should be run twice to identify the best integrated map. The maximum interval size (see Endelman and Plomion (2014) for a description) was set at 8 (default 4) and the map with the minimum error was saved for the final selection of the

"globally" optimal integrated maps. LPmerge currently reports the error associated with each possible solution but does not save this information automatically. We therefore altered part of the source code to create an output file of the errors and map lengths per maximum interval, allowing us to identify the best results. The altered source code of the LPmerge function is provided in Supplementary File S6.

# Map quality

***Consistency between the integrated maps and the underlying haplotype-specific homologue maps***

We visualised the relationship between the homologue maps and the integrated map for each chromosome (Figure 4). Apart from a small number of 'kinks', there was a very high level of linearity observed between the component maps and the integrated maps as well as an acceptable correspondence in map lengths (Supplementary Table S3), demonstrating that the integration step did not create undue distortion. This linearity was also reflected in the high $R^2_{adj}$ values associated with the regression analysis – with a minimum of 0.97 and a mean of 0.99 (*i.e.* essentially co-linear). The slopes and adjusted correlation coefficients of the different maps are provided in Supplementary Table S4.

***Comparison with the physical position of the mapped markers***

One of the advantages of developing mapping theory and software using data from potato is the availability of physical maps which provide a reference marker position (Potato Genome Sequencing Consortium, 2011;Felcher et al., 2012;Vos et al., 2015). Re-orienting the integrated genetic maps if necessary, we found the expected profiles for all chromosomes (Figure 5) and could also clearly identify markers for which the chromosome assignment on the physical map appears to be incorrect. The physical location of 68 other markers was previously unknown (recorded as 0 Mb on chromosome 0), for which we can now provide an approximate physical position based on these plots (Supplementary File S4). These plots also provided information on the location of the pericentromeric regions. We found differences between our identified pericentromeric boundaries and those previously reported for potato chromosomes 5, 6, 10 and 11 (Sharma et al., 2013). In all these cases we noted that the published regions were too large *i.e.* some stretches of the pericentromeric regions of these chromosomes show little or no suppression of recombination. One issue that did arise on chromosome 2 was the mapping of what we suspect is a centromeric marker to a non-centromeric position (Figure 5). All markers binned with this marker were therefore also at a non-centromeric position. When we re-ran the mapping without binning in this homologue, we saw essentially the same result – *i.e.* marker binning was not to blame. The integration of information from telocentric chromosomes may result in such minor ordering errors

given that multi-point estimates (from which the map is ultimately derived) are one-sided at the telomeres.

***Comparison to other published tetraploid potato maps***

A subset of the SolSTW 20K SNPs came directly from the SolCAP 8303 SNP array, with 3684 of these having an acceptable $F_1$ pattern after fitting, of which 2707 segregated. We mapped 2706 of these SolCAP markers, allowing a direct comparison of our genetic maps with previously-published tetraploid maps which use these markers (Hackett et al., 2013;Massa et al., 2015). Only a single marker from this set was found to have been assigned to a different linkage group between the three studies (solcap_snp_c1_15085), which was mapped on chromosome 6 by Hackett et al. and which both we and Massa et al. mapped on chromosome 4. We double-checked the physical position by BLASTing the marker sequence against the potato genome (Hirsch et al., 2014) and found it produced a single hit on chromosome 4.

A comparison of the maps is shown in Supplementary Figure 3. In general, our map positions correspond well with those of previous studies, apart from chromosome 9 where we found that a group of markers most likely to be centromeric (based on a comparison with the physical map) were mapped at 110 cM by Hackett et al. (2013) even though the pericentromeric region of chromosome 9 is positioned at approximately 50 cM in their map. There also appear to be several cases where markers were binned together by Hackett et al. (2013) which we assigned to different genetic positions (Supplementary Figure 3, chromosomes 2, 4, 8 and 11). However, the binning strategy employed by Hackett et al. (2013) differs considerably from our approach, in that markers are not automatically binned before map ordering, but only end up in a "bin" if they fail to map after two rounds of JoinMap's weighted regression ordering algorithm. Despite these discrepancies, the mapping performed in previous studies is broadly consistent with our results.

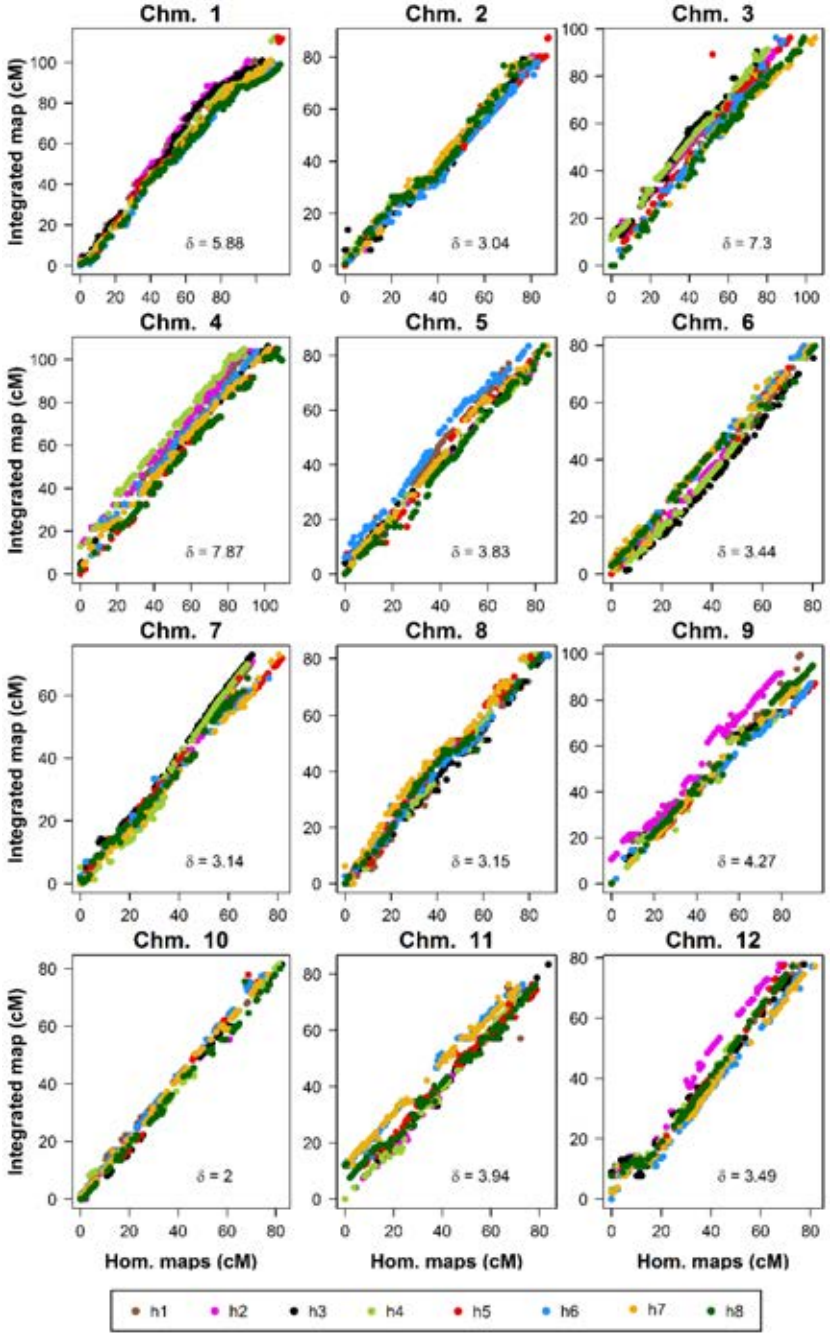



**Figure 4. Comparison between marker positions on underlying homologue maps and integrated map positions for potato chromosomes 1-12.** Different colours denote different homologues. $\delta$ denotes the absolute error between the eight homologue maps and the integrated map, as calculated by LPmerge.

# Effect of quadrivalents and preferential pairing on mapping

***Quadrivalents***

One notable effect of quadrivalent pairing in meiosis is the phenomenon known as double reduction, where two copies of the same parental homologue segment are transmitted to an offspring. It has previously been shown that quadrivalents have a relatively minor impact on recombination frequency estimates of pairs of SxN markers (Bourke et al., 2015). Here we extend the analysis to all possible marker segregation types of a tetraploid cross. In general, we can confirm our previous finding that quadrivalents have a minor impact on *r* estimates for most marker pairs and phases, but lead to an under-estimation of *r* when the proportion of quadrivalents approaches one (*e.g.* SxD and DxD in coupling/repulsion phase, Supplementary Figure 4). However, no observations of such high proportions of quadrivalents have been reported yet (as far as we know) in an autopolyploid species (Swaminathan and Howard, 1953;Ramsey and Schemske, 2002;Bourke et al., 2015). In Supplementary File S7.a, we provide full details of the results of this study.

***Preferential pairing***

Preferential pairing constitutes a much greater deviation from the assumption of random bivalent pairing, and this was reflected in the results of the simulation study. We again saw a downward bias in *r* with greater levels of disomic behaviour, which was accompanied by a drop in the correlation between the true and estimated values. A higher population size can help to mitigate these effects, but when the rate of preferential pairing ($p$) exceeds ~0.7 this makes little difference. Correct phase estimation was surprisingly robust against preferential pairing, although in a fully disomic situation ($p = 1$) it was not possible to estimate *r* in certain cases (specifically, the combination between a duplex and simplex allele in either parent when both alleles are present on the same bivalent, leads to on average 33% inestimable values). Nevertheless, recombination frequency estimates showed high levels of stability and robustness, even when significant levels of preferential pairing occur. For identifying linkage groups, preferential pairing has almost no impact on the accuracy of coupling linkage estimates with SxN markers (which we use for marker clustering), with the possible exception of DxD markers. An example of the results for SxD and DxS markers in coupling/coupling phase is given in Supplementary Figure 5. In Supplementary File S7.b, full details of the results of this study can be found.

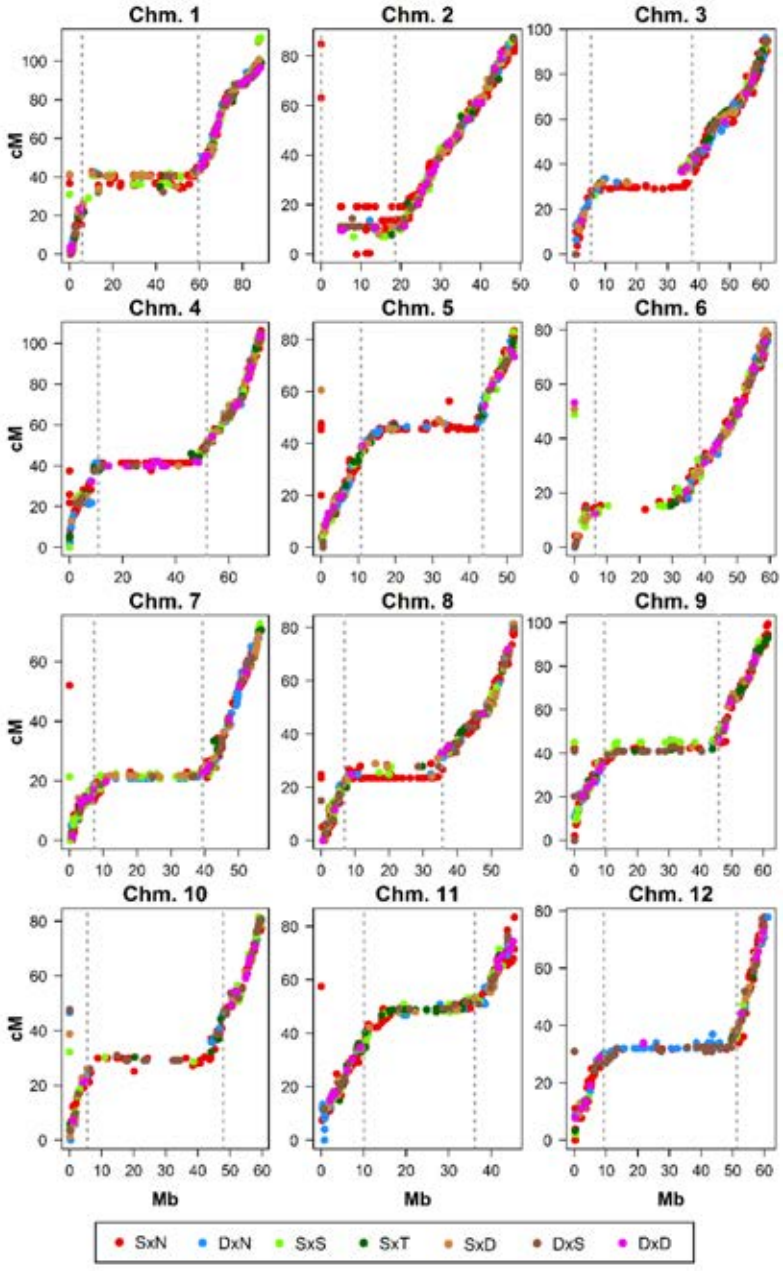



**Figure 5. Comparison between physical and integrated genetic maps for 6872 mapped markers on potato chromosomes 1-12.** Different colours denote different marker segregation types. Centromeres as defined in (Sharma et al. 2013) are shown with dashed lines. 38 markers for which the chromosome assignment differed were removed before plotting. Outlying markers positioned at 0 Mb had no physical position – for which we suggest approximate positions based on these plots (*c.f.* Supplementary File S4).

# Discussion

## Homologue mapping

In our approach, we identify all homologous linkage groups first and map them separately, combining them together in the final step using bridging markers (markers mapped on more than one homologue). There are a number of advantages to this, the first of which is the division of large computational tasks into parallel sub-tasks which results in a significant time-saving. Given that marker datasets are generally increasing in size, it is likely this approach will become increasingly necessary in future polyploid mapping studies. Marker phasing is performed automatically in the initial clustering step and does not have to be calculated afterwards. We also avoid potential map ordering issues by only using the most informative linkage information in map construction. In the case where we identified very high variance associated with $r$ (SxS and SxT in coupling/repulsion phase) we excluded these estimates from the map ordering step by artificially setting LOD = 0.

The use of haplotypes has been shown to have greater statistical power than single-marker approaches in diploid association studies, particularly those involving humans (de Bakker et al., 2005). Our mapping method focuses on creating chromosome-length SNP haplotypes (homologue maps) and therefore could facilitate multi-SNP marker QTL studies rather than those based on single marker positions. Having separate homologue maps will also enable the further exploration of QTL positions, allowing the identification of haplotypes responsible for the phenotypic variation observed. Integrating these homologue maps is a prerequisite for further QTL analyses that use inheritance probabilities instead of marker dosages as explanatory variables (Hackett et al., 2013;Hackett et al., 2014).

Finally, in mapping populations where the meiotic behaviour is not consistent between parents, it is preferable to map each parent separately, given a framework that can incorporate *e.g.* preferential pairing in the estimation of recombination frequency. Our work in other polyploids (particularly ornamental species) suggests that accommodating such meiotic differences is likely to become a regular feature of future mapping work. Parental mapping would also be needed in a tetraploid x diploid cross (for example) because of the different ploidy levels of the two parents.

## The necessity of simplex x nulliplex markers

One of the potential pitfalls of our mapping strategy is its reliance on SxN markers (both in terms of numbers and distribution). Without an abundance of this marker type, we would have to adapt our mapping approach. SxS markers can also be used to define homologous chromosomes, but with the added complication of dividing the marker data into 4x4=16 cross-parental groupings rather than 4+4=8. However, it is feasible to use additional phasing information to determine from which parental allele the coupling-phase linkage originates. A viable alternative would be to adopt the mapping strategy described in (Hackett et al., 2013). Nevertheless, we have yet to encounter situations where the number of SxN (or NxS) markers would cause such a restriction – indeed, they tend to be the most abundant marker segregation type that we have encountered across multiple populations.

## Marker binning

Marker binning has an enormous impact on the speed of marker ordering, particularly since the timing of the weighted linear regression map ordering algorithm is at least quadratic with the number of markers used. Almost half the markers were binned during the mapping, reducing the effective number of marker loci from 6910 to 3980. This was also reflected at the homologue level, reducing the mean number of markers from 185 to 111 markers per homologue map. An examination of where these binned markers came from revealed that they were relatively well-distributed, but were particularly abundant in the pericentromeric regions as one would expect (chromosome 7 is shown as an example in Figure 6).

Despite the efficacy of marker binning, the trend continues to be towards even larger marker datasets. In cases where the marker set size becomes unworkably large, there are a number of simple amendments to our method that could be considered, for example:

1. Binning more markers to create sparse framework maps initially; further saturation for fine-mapping can be confined to interesting regions after initial QTL analyses.

2. Sub-dividing homologue marker clusters into smaller groups and mapping these segments separately before merging.

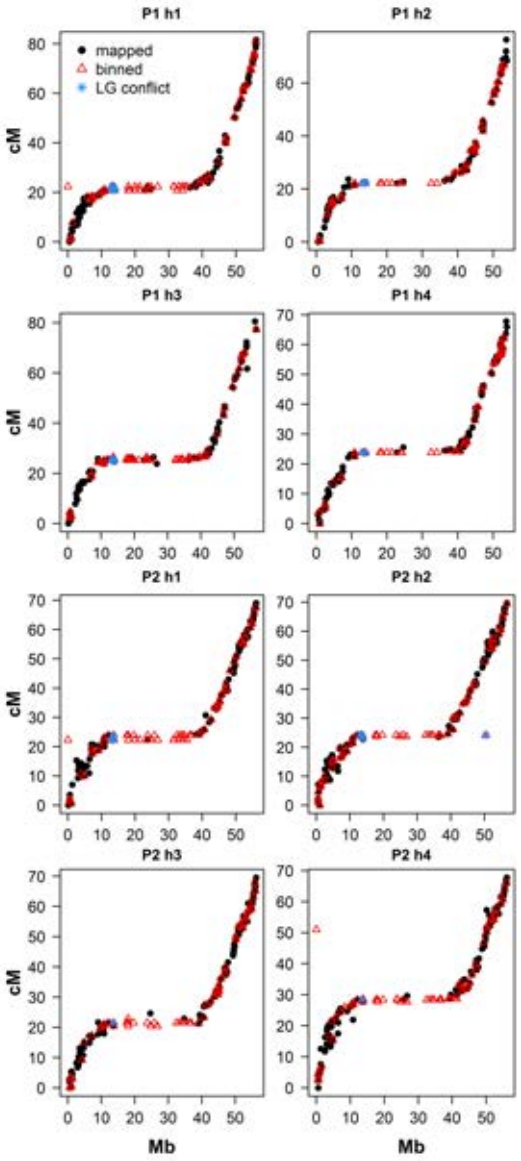



**Figure 6. Distribution of binned markers on eight homologue maps of potato chromosome 7.** Mapped markers (used in the marker ordering step) are shown as black dots, binned markers (removed prior to marker ordering) are shown as red open triangles and were re-added after mapping. LG conflict (blue stars) refers to markers for which the chromosome assignment on the physical and genetic maps differs.

Nevertheless, the development of faster algorithms for marker ordering is a likely prerequisite for future mapping studies in polyploid species involving large population sizes (and more markers). For now, it appears that the weighted linear regression criterion of JoinMap remains the best option to produce accurate maps given pairwise recombination frequency estimates with variable information content.

## Map integration

One behaviour which we did not expect was the variability of LPmerge depending on the relative orientation of the input maps, which has not been described by the authors (Endelman, 2011;Endelman and Plomion, 2014), or in any subsequent publication known to us which uses this package. Higher numbers of bridging markers will probably improve the stability of the integrated map solution found between successive runs, although we recommend that the integration step be repeated over a range of maximum interval sizes and using both forward and reverse orientations to ensure that the best integrated map has been found.

## Application to other tetraploid species

The methods developed here can be directly applied to other tetraploid species. Our results show that mapping under the assumption of random bivalents is a relatively robust simplification when there is a low amount of quadrivalent formation or preferential pairing. Of some concern are polyploid species for which the mode of inheritance is neither strictly polysomic nor disomic, but something in between. There have been various reports of "segmental allopolyploidy" (Stebbins, 1947;Sybenga, 1996), for example in rose (Koning-Boucoiran et al., 2012), garden dahlia (Schie et al., 2014) and peanut (Leal-Bertioli et al., 2015). One of the advantages of our approach is that it predominantly relies on coupling-phase estimates which have been shown to be more robust against preferential pairing than repulsion-phase estimates (the case of SxN markers is covered in detail in (Bourke et al., 2015)). We would caution against mapping in any polyploid species without first assessing the strength of preferential pairing, unless map construction is to be limited to a subset of marker segregation types (*e.g.* SxN and NxS with SxS or SxT markers, but not both). As mentioned, our mapping strategy can be tailored to accommodate differences in pairing behaviour between parents, chromosomes or even parts of a chromosome if necessary.

# Conclusions

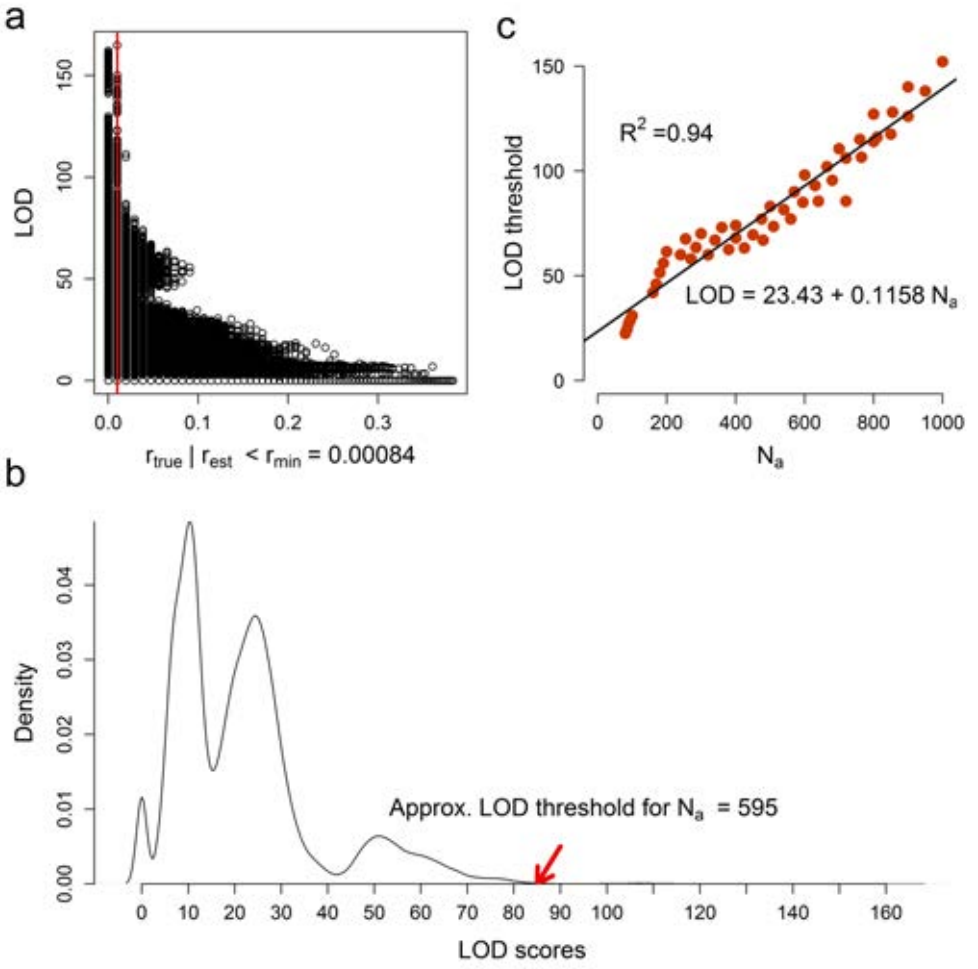
In this study we have demonstrated that high-quality, high-density linkage maps can be efficiently produced in tetraploid species, which we have applied to a dataset from a bi-parental cross in the economically-important crop species potato. These maps will facilitate down-stream applications such as QTL analysis and marker-assisted selection in polyploids. Our mapping approach results in the relatively fast creation of linkage maps in tetraploid species for which the assumption of random bivalent pairing holds to a reasonable extent. Extension to higher ploidy levels is theoretically straightforward, but remains to be realised in practice. Homologue mapping facilitates the parallelisation

of map computation as well as providing long-range haplotype information, with marker phase being automatically assigned prior to mapping without the need for manual intervention. The time-limiting step remains marker ordering, but we have found that our binning approach offers a substantial speed-up in computational time without adversely affecting map quality.

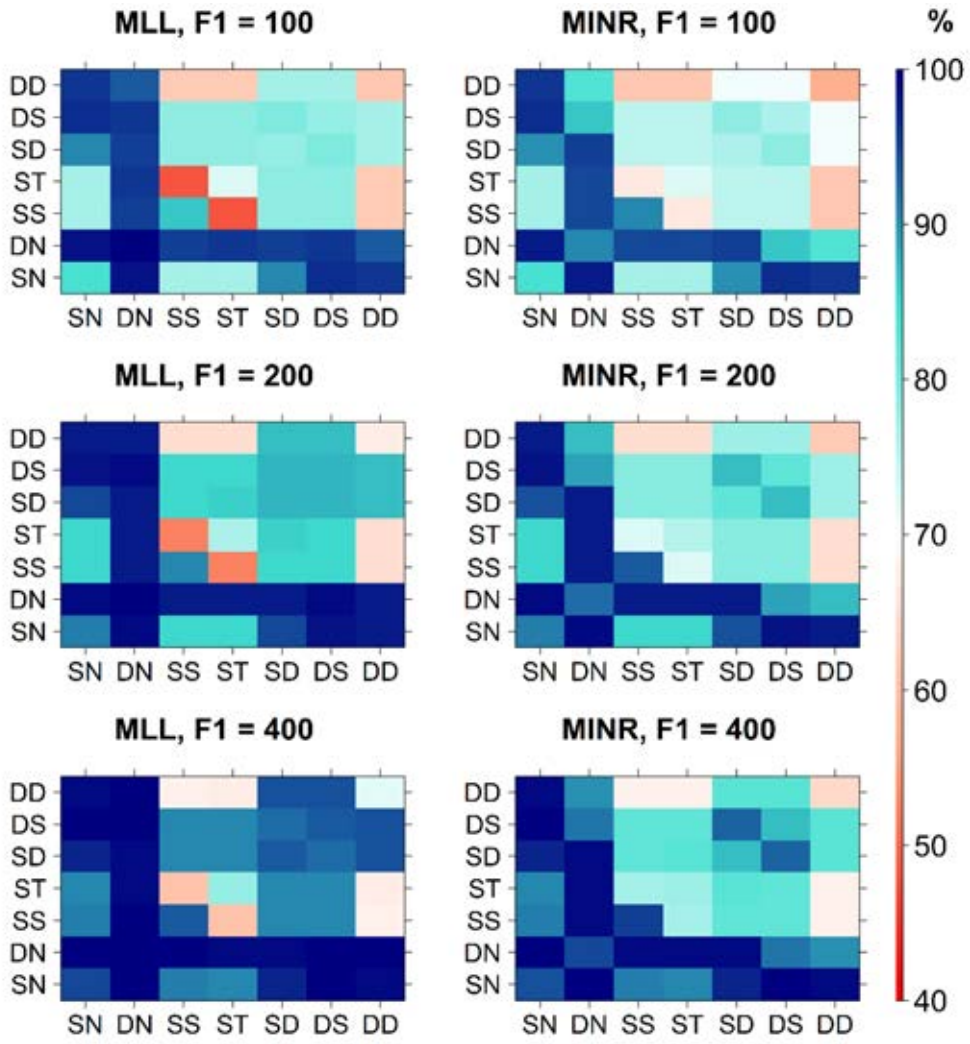
# Acknowledgements

The authors would like to acknowledge Peter Vos and Herman van Eck for sharing their expertise, Paul Arens for critically reading the manuscript and Johan van Ooijen (Kyazma) for helpful suggestions. The authors also wish to acknowledge the editor and two anonymous reviewers, whose comments helped improve the manuscript. Funding for this research was provided through the TKI polyploids project "A genetic analysis pipeline for polyploid crops", project number BO-26.03-002-001.

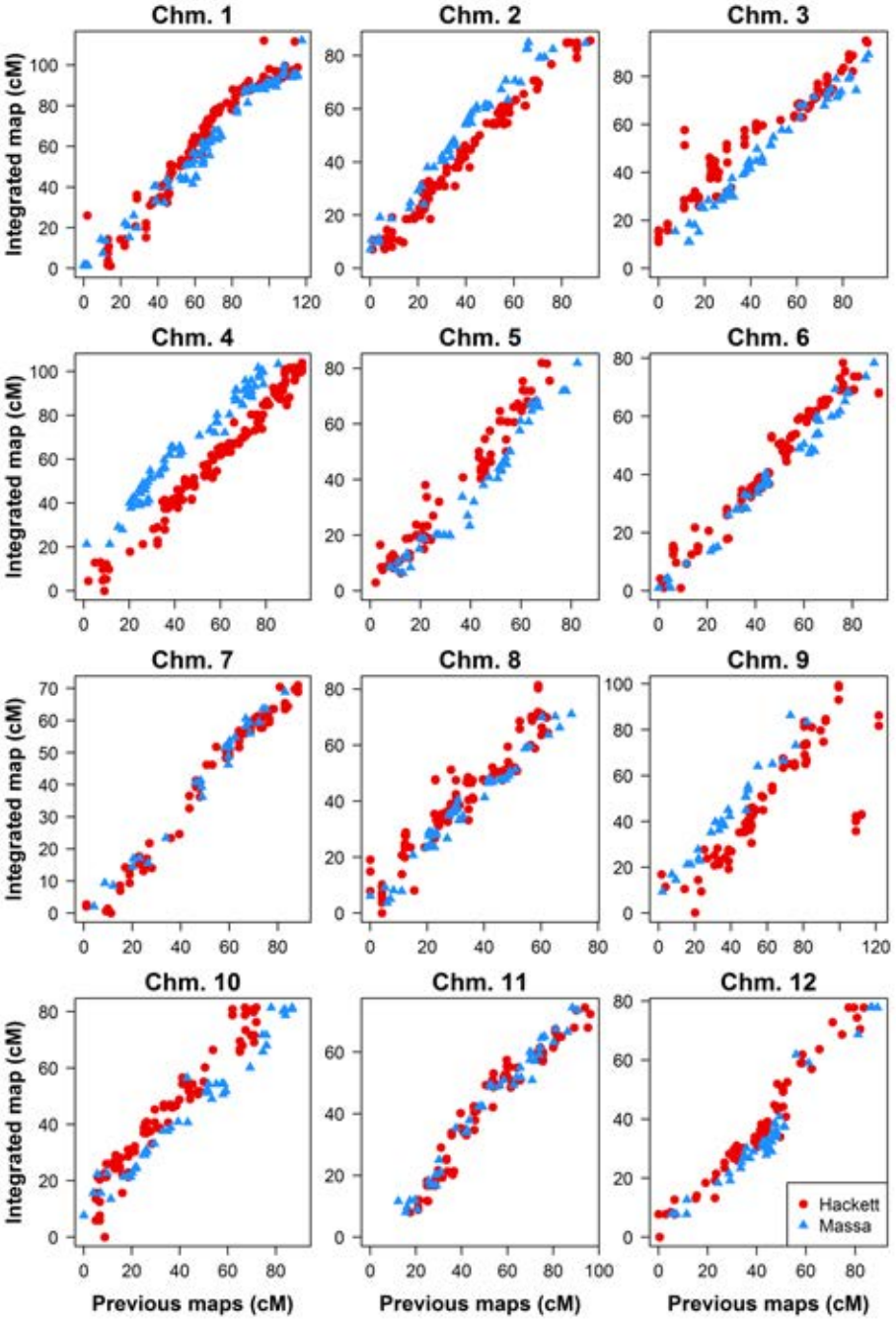
***Supplementary data is available online at:***
https://link.springer.com/article/10.1007/s00122-016-2768-1#SupplementaryMaterial



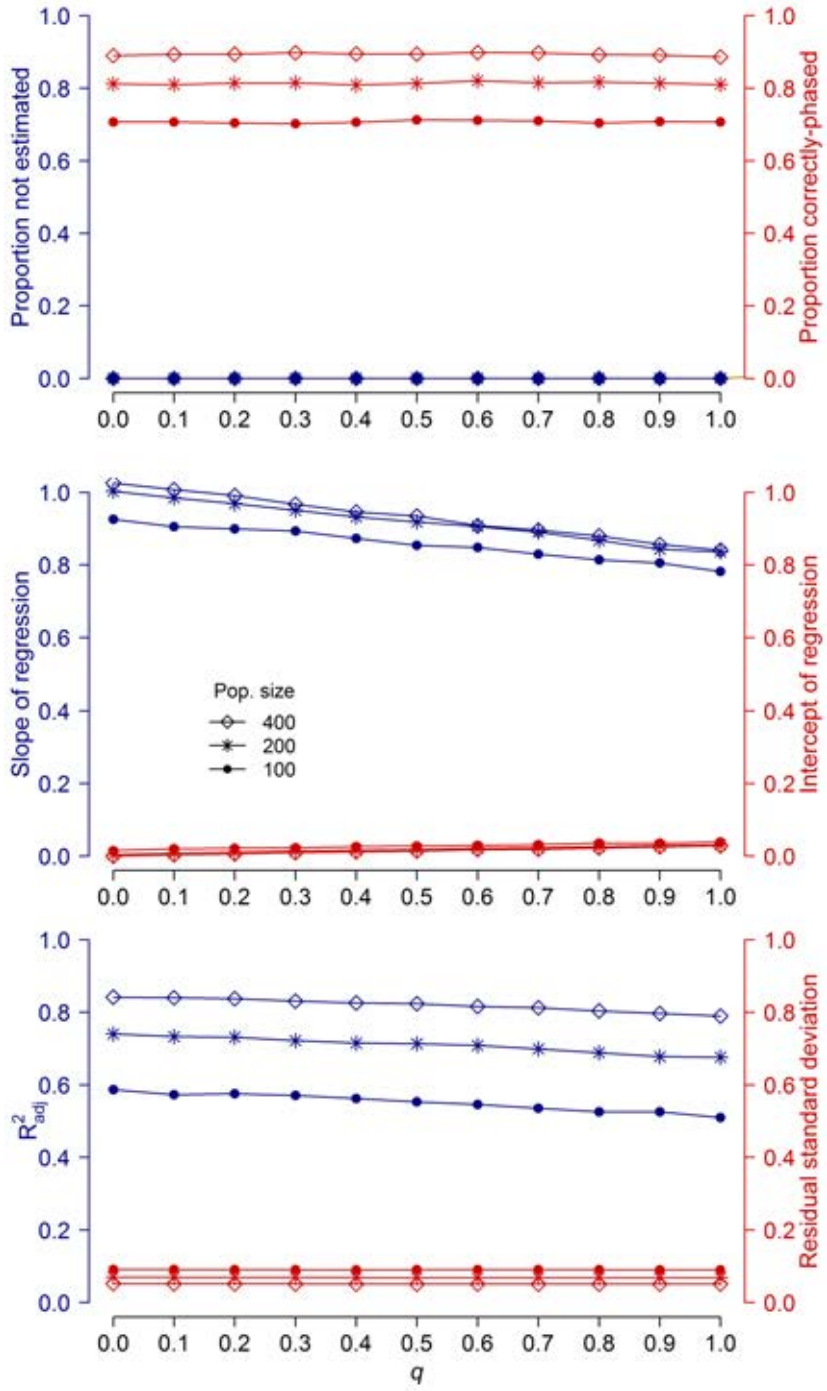

**Supplementary Figure 1. a.** Plot of true recombination frequencies versus their LOD scores when the estimated *r* value fell below the threshold of minimum resolution $r_{min}$. In this example, the simulated population size was 700 and there were 15% missing values in the data (thus the adjusted population size ($N_a$) was 595). The red line shows the "acceptable" deviation of 0.01 adopted in this study. **b.** Distribution of LOD values for recombination frequency estimates that fall below the threshold of minimum resolution $r_{min}$ and have a deviation of less than 0.01 from the true value. The approximate LOD threshold for this population size is shown by the arrow. **c.** Relationship between LOD threshold and $N_a$ to guarantee a maximum deviation of 0.01 between an estimate of zero recombination frequency and its true value for all possible tetraploid marker pair combinations (derived using simulated data). The fitted line was used to determine an appropriate binning threshold given the population size and average rate of missing data used in this study.



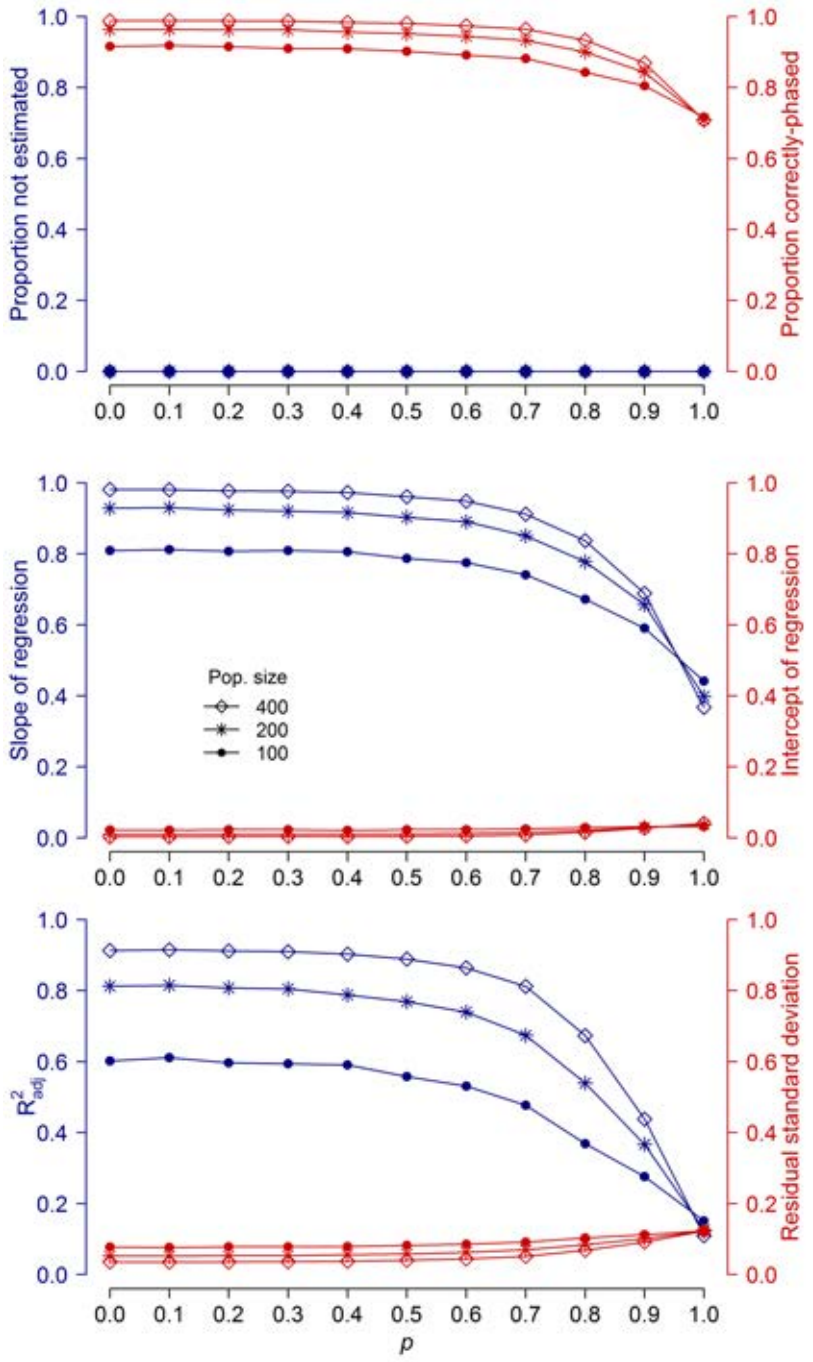

**Supplementary Figure 2.** Heatplots showing results of the simulation to determine the overall phasing accuracy in all pairs of marker types using maximum log likelihood (MLL) or minimum *r* (MINR) for various mapping population sizes.




**Supplementary Figure 3.** Comparison between previously-published tetraploid potato linkage maps (Hackett et al., 2013;Massa et al., 2015) and the integrated map described here.




**Supplementary Figure 4.** Visualisation of the effect of quadrivalents on pairwise *r* estimates for SxD and DxD markers in coupling/repulsion phase. *q* denotes rate of quadrivalent pairing.




**Supplementary Figure 5.** Visualisation of the effect of preferential pairing on pairwise *r* estimates for SxD and DxS markers in coupling/coupling phase. *p* denotes strength of preferential pairing, from 0 (polysomic) to 1 (disomic).

# Chapter 5

# Partial preferential chromosome pairing is genotype dependent in tetraploid rose

Peter M. Bourke[1], Paul Arens[1], Roeland E. Voorrips[1], G. Danny Esselink[1], Carole F. S. Koning-Boucoiran[1], Wendy P. C. van 't Westende[1], Tiago Santos Leonardo[1], Patrick Wissink[1], Chaozhi Zheng[2], Geert van Geest[1,3], Richard G. F. Visser[1], Frans. A. Krens[1], Marinus J. M. Smulders[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[2] Biometris, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[3] Horticulture and Product Physiology, Dept. of Plant Sciences, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

*Published as* Bourke, P.M., Arens, P., Voorrips, R.E., Esselink, G.D., Koning-Boucoiran, C.F.S., Van 'T Westende, W.P.C., *et al.* (2017), "Partial preferential chromosome pairing is genotype dependent in tetraploid rose", **The Plant Journal 90 (2), 330-343**

# Abstract

It has long been recognised that polyploid species do not always neatly fall into the categories of auto- or allopolyploid, leading to the term "segmental allopolyploid" to describe everything in between. The meiotic behaviour of such intermediate species is not fully understood, nor is there consensus as to how to model their inheritance patterns. In this study, we used a tetraploid cut rose (*Rosa hybrida*) population, genotyped using the 68K WagRhSNP array, to construct an ultra-high density linkage map of all homologous chromosomes, using methods previously developed for autotetraploids. Using the predicted bivalent configurations in this population, we quantified differences in pairing behaviour among and along homologous chromosomes, leading us to correct our recombination frequency estimates to account for this behaviour. This resulted in the re-mapping of 25,695 SNP markers across all homologues of the seven rose chromosomes, tailored to the pairing behaviour of each chromosome in each parent. We confirmed the inferred differences in pairing behaviour among chromosomes by examining repulsion-phase linkage estimates, which also carry information about preferential pairing and recombination. Currently, the closest-sequenced relative to rose is *Fragaria vesca*. Aligning the integrated ultra-dense rose map with the strawberry genome sequence provided a detailed picture of the synteny, confirming overall co-linearity but also revealing new genomic rearrangements. Our results suggest that pairing affinities may vary along chromosome arms, which broadens our current understanding of segmental allopolyploidy.



# Key words

High-density integrated map, segmental allopolyploid, polyploid genetic linkage map, *Rosa hybrida*, meiotic chromosomal pairing behaviour.

# Introduction

Polyploids are generally divided into two types – autopolyploids and allopolyploids. There continues to be debate about the definition of these categories – namely, whether they should be defined by their (presumed or known) mode of origin or their (observed) mode of inheritance (Ramsey and Schemske, 1998), otherwise described as the taxonomic or genetic definitions (Doyle and Egan, 2010). According to the taxonomic definition, polyploids are distinguished by their number of founder species (one in the case of autopolyploids, two or more in the case of allopolyploids) whereas the genetic definition distinguishes polysomic inheritance resulting from random pairing of chromosomes during meiosis (autopolyploid), from disomic inheritance resulting from non-random or preferential pairing (allopolyploid) (Doyle and Sherman-Broyles, 2016).

Theoretically at least, it has long been recognised that there may also be intermediate forms of polyploidy, variously termed segmental allopolyploidy (Stebbins, 1947), partial preferential pairing (Wu et al., 1992), incomplete polysomy (Guimarães et al., 1997), heterosomy (Roux and Pannell, 2015) or mixosomy (Soltis et al., 2016). In these intermediate categories, the genetic definition (*i.e.* pairing behaviour during meiosis) takes precedence (Figure 1.a). This pairing behaviour is primarily important as it determines the extent of recombination between homologues or homoeologues, providing a diagnostic of the type of polyploidy (Parisod et al., 2010;Doyle and Sherman-Broyles, 2016). Furthermore, it may influence our ability to produce linkage maps and subsequently perform accurate quantitative trait locus (QTL) analysis, of importance for both fundamental and applied plant research. Understanding the pairing behaviour of polyploid species is also relevant for the study of genome evolution upon genome duplication.

Various ornamental crops, *Rosa* species in particular, have a long history of hybridisation and polyploid formation between and within species to generate the diversity of cultivars we have today (Smulders et al., 2011). This breeding and selection is likely to have contributed to differences in homology between various chromosomes, thought to play a role in meiotic pairing behaviour (Bingham and Gillies, 1971;Lentz et al., 1983). Pairing in other species has been shown to be under genetic control (for example, the *Ph1* locus of hexaploid wheat (Okamoto, 1957;Riley and Chapman, 1958) or the *PrBn* locus in *Brassica napus* (Jenczewski et al., 2003;Nicolas et al., 2009)), but may also be influenced by environmental factors (Bomblies et al., 2015). Despite some advances in

our understanding of polyploid meiotic regulation, there remain many unanswered questions, particularly in species with mixed inheritance types.

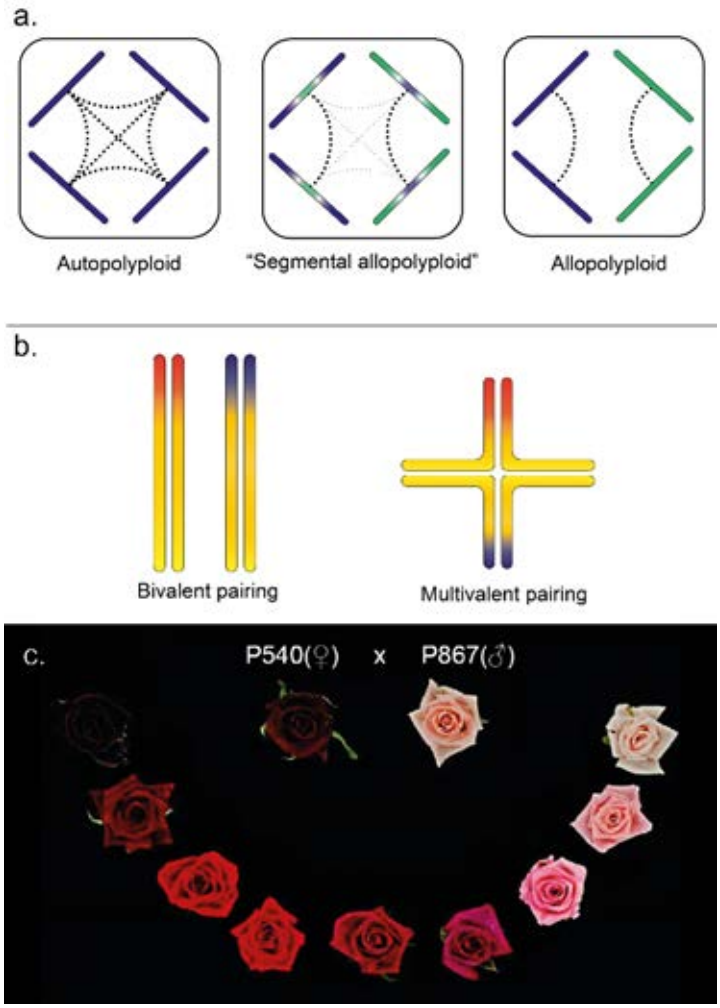



**Figure 1. The meiotic pairing behaviour of tetraploid rose is not fully understood, yet exhibits many features of a segmental allopolyploid; mapping populations and high-resolution maps can shed light on such questions. a.** Schematic representation of auto- and allopolyploid pairing (in this case, for a tetraploid), with intermediate pairing behaviour of a segmental allopolyploid shown in between. **b.** Hypothesised model to account for variable rates of preferential pairing along a chromosome. If preferential pairing initiation (telomeric bouquet formation) is confined to one end of the chromosome, random pairing between the other telomeres can lead to both bivalent and multivalent pairing as shown. Some degree of preferential pairing is still expected given that pairing between one set of telomeres is highly preferential; this agrees with the observations of integrated consensus map (ICM) chromosome 1 of P1. **c.** Parents P540 and P867 of the K5 rose population used in this study, with examples showing the range of segregation for flower colour in the $F_1$ progeny.

## Genetic mapping in rose

The genus *Rosa* consists of a highly complex and much-debated phylogeny, divided into four subgenera, of which subgenus *Rosa* comprises over 95% of all species (Wissemann and Ritz, 2005). *Rosa* is generally divided into ten sections, four of which (*Synstylae*, *Gallicanae*, *Indicae* and *Pimpinellifoliae*) have contributed to the domesticated rose genepool (Smulders et al., 2011). *Rosa hybrida* or the hybrid tea rose is nowadays the most well-known and commercially-important representative of the genus and has a complex mixture of hybrid perpetuals derived from China rose, Noisettes (*Rosa chinensis*), Bourbons as well as *R. gallica* and *R. alba*, and tea roses (*R. x odorata*) in its pedigree (Koning-Boucoiran et al., 2012;Liorzou et al., 2016). Given this complexity of origin, it is not surprising that it remains a poorly-understood species genetically (Debener and Linde, 2009). Studies of the mode of inheritance of wild tetraploid rose populations have found evidence for both disomic and tetrasomic behaviour (Wissemann and Ritz, 2005;Joly et al., 2006). Tetrasomic inheritance is generally assumed for *Rosa hybrida* (Gar et al., 2011) with the possibility of some preferential pairing (Koning-Boucoiran et al., 2012) although this has never been quantified. A large number of publications have studied linkage mapping in *Rosa*, including an integrated map at the diploid level (Spiller et al., 2011) and maps at the tetraploid level (Gar et al., 2011;Koning-Boucoiran et al., 2012;Vukosavljev et al., 2016), but none have taken full account of its pairing behaviour. An initiative is underway to sequence the genome of the diploid species *Rosa chinensis* (Bendahmane et al., 2016), for which ultra-high density genetic linkage maps will likely provide useful information for connecting and orientating scaffolds (Bartholomé et al., 2015).

## Identifying and quantifying preferential pairing

Identifying pairing preferences to help formulate a model of meiosis has been a long-standing challenge for researchers. The earliest methods to determine pairing behaviour were cytological and relied on counting the frequency of bivalents and multivalents during diakinesis or metaphase I of the meiosis (Lentz et al., 1983;Sybenga, 1994), under the assumption that allopolyploids should exhibit more bivalent pairing than autopolyploids. However, certain autopolyploid species such as potato predominantly pair as bivalents (Swaminathan and Howard, 1953) and yet show no evidence for preferential pairing. Bivalent to multivalent ratios are no longer seen as an accurate method to determine pairing type.

More recent methods have used molecular markers, which carry a signature of the parental meioses. Comparing the observed segregation ratios of different marker types to those expected under an assumed pairing model is a simple test for pairing behaviour, but can be influenced by marker skewness, which may be caused by selection or other,

unknown factors. If preferential pairing is incomplete, marker segregation ratios cannot be tested against a fixed set of expected ratios (Allendorf and Danzmann, 1997). Repulsion-phase linkages (*i.e.* linkage between markers which target the same genomic region but which tag alternative homologues of the chromosome) are also sensitive to deviations from random pairing. Initially, identifying significant repulsion linkages was taken as evidence of disomic behaviour (Da Silva et al., 1993;Al-Janabi et al., 1994). Later studies attempted to estimate the degree of preferential pairing from repulsion-phase recombination frequency estimates, although they relied on prior knowledge or assumptions about repulsion-phase inter-marker distances (Qu and Hancock, 2001;Cao et al., 2004).

Recently, methods for creating integrated genetic maps in autotetraploid species using SNP marker data have been developed (Hackett et al., 2013;Bourke et al., 2016). These methods offer the possibility of identifying repulsion-linkages that can help reveal pairing behaviour. Furthermore, methods to reconstruct offspring haplotypes in mapping populations of tetraploids have also become available (Hackett et al., 2013;Zheng et al., 2016). In particular, TetraOrigin (Zheng et al., 2016) provides the most likely predicted pairing structures per offspring (either in bivalents or quadrivalents), enabling the estimation of population-wide meiotic pairing behaviour as well as the strength of preferential pairing ($\rho$) should it exist.

In this study, we used a genotyped tetraploid mapping population of rose to investigate its meiotic behaviour. The 68K WagRhSNP array (Koning-Boucoiran et al., 2015;Schulz et al., 2016;Vukosavljev et al., 2016) was used to create an ultra-high density tetraploid linkage map, enabling the quantification of preferential pairing, quadrivalent formation and double reduction during parental meiosis. In a previous study using simulated data, it was hypothesised that preferential pairing could effectively be ignored in polyploid linkage mapping (up to a 70% deviation from random pairing) (Bourke et al., 2016). In this study we had the opportunity to test this prediction by correcting each linkage map for the observed pairing behaviour, and comparing them to those created under a random model.

# Experimental Procedures

## Plant material, DNA isolation and genotyping

The tetraploid “K5” cut rose mapping population, consisting of 172 individuals of the cross between “P540” (mother) and “P867” (father) was used in this study (Figure 1.c).

This population has previously been used in studies on powdery mildew (*Podosphaera pannosa*) resistance (Yan et al., 2006), a range of morphological traits (Gitonga et al., 2014;Gitonga et al., 2016), stomatal functioning (Carvalho et al., 2015), and linkage map construction using AFLP and SSR markers (Koning-Boucoiran et al., 2012). Genomic DNA was extracted from freeze-dried young leaves, using the DNeasy Plant Mini Kit (Qiagen, http://www.qiagen.com/) following the protocol of (Esselink et al., 2003). Samples were sent to Affymetrix for genotyping using the WagRhSNP 68k Axiom SNP array (Koning-Boucoiran et al., 2015). This array targets 68,893 SNPs with every SNP targeted with two probes (which we refer to as the P and Q probes). Both parents were genotyped in triplicate (two biological replicates and one technical replicate).

## Genotype calling and data preparation

The SNP array data was converted into dosage scores using the fitTetra package (Voorrips et al., 2011). Quality checks were subsequently performed to ensure consistency between the parental scores and those of the offspring, and to check for the possibility of "shifts" in dosage assignments which can occur when fewer than the five possible dosage classes occur. As each SNP was targeted with two probes (P and Q), the genotype calls for these probe pairs had to be compared and merged if they were found to be consistent (< 10% conflicts, where missing values were not considered conflicting), with conflicting scores made missing. Probes with more than 10% conflicting scores (including parental scores) were kept as separate markers, by appending the letters "P" and "Q" to the marker names. Markers with more than 10% missing values were removed from the dataset, as well as those markers showing highly-skewed segregation patterns under either a tetrasomic or disomic model ($P < 0.001$).

In a tetraploid genotyped using bi-allelic SNP markers, the possible dosage classes range from nulliplex (0 copies of the alternative allele, coded 'N') to quadruplex (4 copies, coded 'Q'), with marker segregation types defined by their parental scores. Marker dosage scores were re-coded as described in (Bourke et al., 2016), resulting in 9 marker classes (SxN, NxS, DxN, NxD, SxS, SxT, SxD, DxS, DxD). Duplicate individuals were identified by genotype pairs with an unusually-high correlation (> 95% similar scores); these individuals were merged (if both scores were non-missing and conflicting, the merged score was made missing). Individuals with more than 10% missing values were also removed.

## Marker clustering and linkage group assignment

Initially, the simplex x nulliplex (SxN) and nulliplex x simplex (NxS) markers were clustered using the LOD for linkage. A LOD value of 4 was found to split the data evenly, with clusters of fifteen or more markers selected as candidate homologue groups. In P1

this resulted in 29 clusters, one more than the 28 expected. Each cluster was then tested over a range of thresholds (from LOD 4 to LOD 10) to ensure the markers remained clustered together. In P2 at a LOD threshold of 4, 28 clusters were identified. Two of these clusters split at LOD 5, resulting in 30 P2 clusters. These clusters were subsequently assigned to linkage groups using their linkage to DxN (or NxD) markers, which provide cross-homologue linkage. Where more than four putative homologues were present, the predicted phasing across clusters was used to merge these into single homologues. In total, 28 P1 homologues across 7 chromosomes were identified, and 27 P2 homologues across 7 chromosomes in P2 (*i.e.* we missed one P2 homologue). Chromosomes were re-numbered according to the ICM numbering previously introduced (Spiller et al., 2011), through linkage with SSR markers from a previous study (Koning-Boucoiran et al., 2012). All other marker segregation types were subsequently assigned to both chromosomes and homologues based on their linkage to SxN markers (LOD > 3).

## Linkage analysis and map construction under a tetrasomic model

Pairwise linkage analysis between all marker types was performed per chromosome and per parent. The maximum likelihood framework used to estimate the (phased) recombination frequency and LOD score under the assumption of random bivalent pairing (*i.e.* no double reduction and no preferential pairing) is already described elsewhere (Hackett et al., 2013;Bourke et al., 2016). Pairwise marker phase was primarily based on likelihood maximisation (Hackett et al., 2013), although minimum recombination frequency was used to phase SxS marker pairs (as described in Bourke et al. (2016)).

A prototype version of the MDSMap software was kindly made available by its authors for map ordering (Preedy and Hackett, 2016). Maps were produced using unconstrained weighted metric multi-dimensional scaling (with $LOD^2$ as weights and using Haldane's mapping function) followed by principal curve fitting in two dimensions. Poorly-mapping markers were identified either as outliers in the PCO plots (judged by eye), or if their nearest-neighbour fit exceeded 5. Such markers were removed, and up to three rounds of MDSMap were performed until stable maps free of outlying markers were produced. The final map positions of the 705 pairs of (unmerged) P and Q probes were compared. In 235 cases, one of the probes was lost (either during the clustering stage or later when outliers were removed from the linkage maps) and the remaining probe was taken as the consensus marker. In cases where both probes were mapped, only 148 mapped less than 1 cM apart and had the same parental dosages; these probes could be

merged while the rest were discarded. The mapping procedure was repeated using the amended dataset.

## Estimation of a preferential pairing parameter

We modelled preferential pairing in the context of bivalent pairing by considering deviations $\rho$ from the expected probabilities under a random model, where the probability of pairing between homologues 1 and 2, and between 3 and 4 (denoted 12/34) is $\frac{1}{3} + \rho$, and that between 13/24 or 14/23 is $\frac{1}{3} - \frac{\rho}{2}$. We avoided simultaneously estimating a preferential pairing factor and a recombination frequency (which was found to lead to an overestimation of the level of preferential pairing (Wu et al., 2002)), in favour of a two-step procedure that first estimates a map using two-point tetrasomic linkage analysis and then revises those estimates when a preferential pairing factor has been determined using a multi-point hidden Markov model (HMM) approach (Zheng et al., 2016).

The marker data was simplified by rounding marker positions to the nearest centiMorgan (cM) after which one of each marker segregation type was selected at each cM position. Where more than one marker of a particular type was present at a locus, the marker with the fewest missing values was chosen. TetraOrigin provides the most likely bivalent pairing in each individual (classes 12/34, 13/24 or 14/23). We re-numbered homologues such that the pairing configuration with the highest predicted count in TetraOrigin was 12/34 *etc*. We then applied a $\chi^2$ test on the counts of each class to test for deviations from $\frac{1}{3}$ ($P < 0.001$). If the number of structures predicted in each of the three pairing classes are $n_1$, $n_2$ and $n_3$ and assuming $n_1 \geq n_2$ and $n_1 \geq n_3$, the likelihood function given the observed counts is:

$$\mathcal{L}(\rho) \propto \left(\frac{1}{3} + \rho\right)^{n_1} \left(\frac{1}{3} - \frac{\rho}{2}\right)^{n_2+n_3}$$

Solving the likelihood equation leads to the maximum likelihood estimate for $\rho$:

$$\hat{\rho} = \frac{\frac{2}{3}n_1 - \frac{1}{3}(n_2 + n_3)}{n_1 + n_2 + n_3}$$

TetraOrigin was re-run, allowing for the possibility of quadrivalent pairing in the offspring, to investigate the level of double reduction and to see whether preferential pairing was still predicted under a model that included quadrivalents. The rate of double reduction was determined per locus as the average of the haplotype probabilities that exceeded 1 across the population.

## Preferential pairing estimated from repulsion-phase linkages

We estimated the error rate in the genotype data from pairs of duplicated individuals, of which nineteen pairs were identified. The approximate error rate was taken as the average conflict rate over all duplicate pairs. We followed Brzustowicz et al. (1993) in their definition of the correspondence between the true ($\theta_0$) and apparent ($\varphi$) rates of recombination given an error rate *s* (Brzustowicz et al., 1993;Hackett and Broadfoot, 2003)

$$\varphi = \theta_0(1-s) + (1-\theta_0)s$$

The minimum resolution of recombination frequency in a mapping population of size *N* and missing value rate of $\mu$ is $r_{min} \approx 1/(2(1-\mu)N)$ (Bourke et al., 2016), assuming error-free data. In the presence of errors, the above equation then implies:

$$r_{min} \approx \frac{1-2\mu Ns}{2(1-\mu)N}$$

We identified all pairs of SxN or NxS markers that mapped within this distance ($r \sim 0.0117$) on different homologues (repulsion pairs). Under a purely disomic model, the repulsion-phase maximum likelihood estimate for the recombination frequency is given by $r_{disom} = \frac{n_{00}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$, where $n_{01}$ is the number of offspring with a dosage of 0 at marker A and a dosage of 1 at marker B *etc*. If inheritance is tetrasomic, this estimate never falls below $\frac{1}{3}$ (Qu and Hancock, 2001;Bourke et al., 2015). This forms the basis of a Binomial test ($H_0 : r_{disom} \geq 1/3$), corrected for multiple testing using FDR with $\alpha = 0.05$ (Benjamini and Hochberg, 1995). This identified cross-homologue pairs that showed significant deviations from a tetrasomic model. We then estimated a preferential pairing parameter for the homologue combinations that showed significant evidence of preferential pairing. For two repulsion-phase SxN markers A and B located at the same genetic position, the recombination frequency estimate between them is a function of independent assortment with no contribution from cross-overs (Qu and Hancock, 2001). Given a preferential pairing parameter $\rho$, the probabilities of observing each of the classes $n_{00}, n_{01}, n_{10}, n_{11}$ in the offspring are $\frac{1}{6}-\frac{\rho}{4}, \frac{1}{3}+\frac{\rho}{4}, \frac{1}{3}+\frac{\rho}{4}, \frac{1}{6}-\frac{\rho}{4}$ respectively. This leads to the following equation:

$$\frac{\partial}{\partial\rho}(\ln \mathcal{L}(\rho)) \propto \frac{-(n_{00}+n_{11})}{4\left(\frac{1}{6}-\frac{\rho}{4}\right)} + \frac{n_{01}+n_{01}}{4\left(\frac{1}{3}+\frac{\rho}{4}\right)} = 0$$

Solving for $\rho$ yields the maximum likelihood estimate for the preferential pairing parameter:

$$\hat{\rho} = \frac{2(n_{00} + n_{11}) - 4(n_{01} + n_{01})}{3(n_{00} + n_{01} + n_{10} + n_{11})}$$

We recorded the mean (and standard deviation) of the non-negative estimates to generate pairing-specific estimates of $\rho$.

## Linkage analysis of a segmental allopolyploid

Having estimated the strength of preferential pairing for all affected linkage groups, we re-estimated the pairwise recombination frequencies for these cases. The maximum likelihood estimates for recombination frequency were derived in Mathematica (Wolfram Research Inc., 2014) and exported for use in R (R Core Team, 2016). A description of the likelihood model including preferential pairing is provided in Supplementary Methods S1. Markers were subsequently re-assigned to chromosomes and homologues given the updated information, and the maps were re-calculated using MDSMap (Preedy and Hackett, 2016).

## Visualisation and quantification of haplotype diversity

A subset of 20,431 markers for which there were at least 10 significant linkages (LOD > 3) to each of the expected number of homologue clusters was used to define haplotypes. These allowed us to examine pairwise diversity between homologues, to investigate the relationship between homologue similarity and pairing behaviour. We defined a simple dissimilarity measure between homologues A and B within a window centred at marker position *j* as:

$$\mathcal{D}_{A-B,j} = \frac{1}{N_j} \sum_{i=1}^{N_j} \left| d_{A,i} - d_{B,i} \right|$$

where $d_{A,i}$ is the SNP allele score of homologue A at marker position *i* (either '0' or '1') and $N_j$ is the number of markers within a 5 cM sliding window around each marker position *j*.

## Integration with SSR and AFLP map

Previously, a genetic linkage map of the same population using SSR and AFLP markers was published (Koning-Boucoiran et al., 2012). We initially attempted to map all markers together, but found that the SSR / AFLP markers tended to cluster together at the ends of a linkage group, suggesting high tension between the two marker datasets. Linkage between SSR / AFLP markers and the set of SxN (or NxS) markers was evaluated, with the three most significant linkages recorded, giving these markers approximate positions on the SNP map.

## Synteny analysis with *Fragaria vesca*

The expressed sequence tags (ESTs) from which the WagRhSNP 68k Axiom SNP array markers were derived were used in a BLASTn search against the woodland strawberry genome assembly (*Fragaria vesca* v2.0.a1 pseudomolecules (downloaded on 19/09/2016 from http://www.rosaceae.org)) with 'N' used at the SNP position, as provided in Supplementary Table 2 of Koning-Boucoiran et al. (2015). An E-value threshold of $1x10^{-20}$ was used to retain only highly-homologous hits, with multiple hits filtered out to avoid targeting multi-gene families. The rose linkage maps were orientated according to the order on the *Fragaria* genome sequence (as in Vukosavljev et al. (2016)), and the resulting synteny was visualised using the Circos software (Krzywinski et al., 2009). The most likely map positions of a set of non-segregating DxN markers were also determined based on the position of their hit in the strawberry genome, providing additional indications for disomic segregation.

# Results

## Marker filtering and dosage assignment

All 68,893 SNPs on the WagRhSNP array were assayed by two probes, designed to anneal to the flanking sequence at both strands (Koning-Boucoiran et al., 2015). Probes were independently scored in fitTetra (Voorrips et al., 2011) and were filtered according to a number of quality criteria, such as containing at most 25% missing values (in a later filtering step after merging probes, this was reduced to 10%), being non-skewed ($P < 0.001$) and containing fewer than 5% invalid or unexpected scores (Table S1). Filtered probes were merged where possible, with conflicting probes being kept separate, using the mapping as a quality check. Although parents were genotyped in triplicate, two of the Parent 1 (P1) replicates showed high numbers of missing values and low concordance with the offspring (we suspect some genotypes were incorrectly-labelled during previous multiplication steps). Therefore, the replicate of P1 with the best concordance with the offspring scores was selected, whereas for P2 two of the three replicates were found to have few conflicts (0.4%) and could be merged. There were 28,109 segregating markers to begin mapping, with a good spread over the nine marker classes (Table S2). Nineteen pairs of duplicate individuals were identified (pairwise genotype correlation > 0.95). Two individuals with more than 10% missing values were also removed, resulting in a mapping population size of 151 individuals.

## Linkage map construction under a tetrasomic model

We followed the mapping procedure as described for potato in Bourke et al. (2016), although it was not possible to associate homologue (simplex x nulliplex, or SxN)

clusters into chromosomal groups on the basis of repulsion-phase linkages due to more noise in the data. Instead, coupling linkages with duplex x nulliplex (DxN) markers were used to provide these associations. We used new marker ordering software MDSMap (Preedy and Hackett, 2016) which greatly facilitated the creation of integrated chromosomal linkage maps in this study. Rose is a predominantly tetrasomic species and thus we can identify four homologous copies of each chromosome from each parent, assuming sufficient SxN and NxS markers exist. Over both parents and seven linkage groups there are therefore 4 x 7 x 2 = 56 homologue groups expected. Fifty-five of the expected 56 homologues were identified, following which we assigned every other marker type to both its most probable linkage group (chromosomes, termed integrated consensus map (ICM) groups 1 to 7, *c.f.* Spiller et al. (2011)) and homologue(s) using a linkage logarithm of odds (LOD) threshold of 3.

# Estimation of preferential pairing parameters

Given an integrated linkage map and a population with dosage scores, TetraOrigin (Zheng et al., 2016) can be used to infer the identity-by-descent (IBD) probabilities for all offspring, useful for subsequent QTL analyses (Hackett et al., 2014). Here, we used it to estimate preferential pairing between chromosomes. Using the assumption of bivalent pairing, we applied a $\chi^2$ test to the predicted bivalent counts under the hypothesis of random pairing (Table 1). At a significance level of 0.001 three chromosomes exhibited unusual behaviour (ICM1 of P1 and ICM3 and ICM4 of P2), with an over-

**Table 1. Predicted pairing behaviour according to TetraOrigin, showing significance of the deviation from random pairing and the estimated preferential pairing parameter, $\rho$**

| P1 [a] | ICM[b] | 12_34 [c] | 13_24 | 14_23 | $\chi^2$ | P-value [d] | $\rho$ |
|---|---|---|---|---|---|---|---|
| | 1 | 96 | 35 | 20 | 64.4 | **1.00E-14** | 0.302 |
| | 2 | 55 | 50 | 46 | 0.8 | 0.7 | - |
| | 3 | 56 | 48 | 47 | 1.0 | 0.6 | - |
| | 4 | 60 | 56 | 35 | 7.2 | 0.03 | - |
| | 5 | 68 | 50 | 33 | 12.2 | 0.002 | - |
| | 6 | 61 | 56 | 34 | 8.2 | 0.02 | - |
| | 7 | 66 | 47 | 38 | 8.1 | 0.02 | - |
| **P2** | **ICM** | **56_78** | **57_68** | **58_67** | **$\chi^2$** | **P value** | **$\rho$** |
| | 1 | 63 | 51 | 37 | 6.7 | 0.04 | - |
| | 2 | 58 | 55 | 38 | 4.6 | 0.1 | - |
| | 3 | 77 | 43 | 31 | 22.6 | **0.00001** | 0.177 |
| | 4 | 80 | 40 | 31 | 27.0 | **1.3E-06** | 0.196 |
| | 5 | 67 | 48 | 36 | 9.7 | 0.01 | - |
| | 6 | 52 | 52 | 47 | 0.3 | 0.9 | - |
| | 7 | 67 | 42 | 42 | 8.3 | 0.02 | - |

[a] P1 = Parent 1, P2 = Parent 2; [b] ICM = integrated consensus map; [c] The predicted number of offspring with bivalent pairing between homologues 1 and 2, and the second bivalent pairing between 3 and 4.
[d] P-value of the $\chi^2$ test of the hypothesis of random pairing, significant values ($P < 0.001$) in bold

-representation of one bivalent configuration and an under-representation of the other two, which was most extreme on chromosome ICM1 of P1 (96 out of 151 counts were of the same pairing configuration). For ICM5 in P1, there was a near-significant departure from random pairing ($P = 0.002$) but this arose from an under-representation of one configuration, not two, which we have not attempted to model.

## Preferential pairing estimated from repulsion-phase linkages

Preferential pairing has previously been deduced from repulsion-phase estimates between "nearby" SxN marker pairs in other studies. We estimated the minimum resolution of recombination frequency ($r_{min}$) as approximately 0.0117 and found 73,386 repulsion SxN marker pairs in P1 and 103,496 pairs in P2 mapped to within this distance on the integrated map (Table S3). Of these, 2896 had significant evidence for non-tetrasomic behaviour (using a False Discovery Rate (FDR)-corrected threshold of 0.00081, where $\alpha = 0.05$). Strong preferential pairing was identified on ICM1 in P1, confirming our previous findings. The negative $\log_{10}(P)$ values of these tests in visualised in Figure 2. The strength of significance was not constant across the chromosome, partly due to uneven marker distribution. However, in the case of ICM1 of P1 this appears to also be due to differences in pairing behaviour between chromosome arms. A possible model explaining this behaviour is represented in Figure 1.b in which variable pairing affinities are realised through a mixture of bivalent and non-random quadrivalent formation. Our estimated preferential pairing parameters corresponded well with those from TetraOrigin (Table 2), although ICM7 of P1 also shows a region of preferential pairing, which was not predicted by TetraOrigin. The strongest evidence for preferential pairing from repulsion-phase linkages remains that of ICM1 in P1 and ICM3 in P2.

**Table 2. Significant preferential pairing identified using closely-mapping SxN marker pairs in repulsion**

| Parent | ICM [a] | h_a | h_b | $\rho$ | sd [b] | N [c] | $N_s$ [d] |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 3 | 0.35 | 0.14 | 612 | 360 |
| 1 | 1 | 2 | 4 | 0.23 | 0.02 | 88 | 16 |
| 2 | 3 | 1 | 2 | 0.22 | 0.07 | 4676 | 2388 |
| 2 | 4 | 2 | 3 | 0.17 | 0.05 | 106 | 4 |
| 2 | 6 | 1 | 2 | 0.17 | 0.02 | 6760 | 4 |
| 1 | 7 | 2 | 3 | 0.16 | 0.07 | 2414 | 152 |
| 2 | 7 | 1 | 2 | 0.06 | 0.05 | 7538 | 2 |

[a] ICM = integrated consensus map; [b] The standard deviation of the estimate for $\rho$, the preferential pairing parameter between homologues h_a and h_b; [c] The number of non-negative estimates on which the estimate of $\rho$ was based; [d] The number of repulsion-pairs that showed significant evidence for non-tetrasomic behaviour

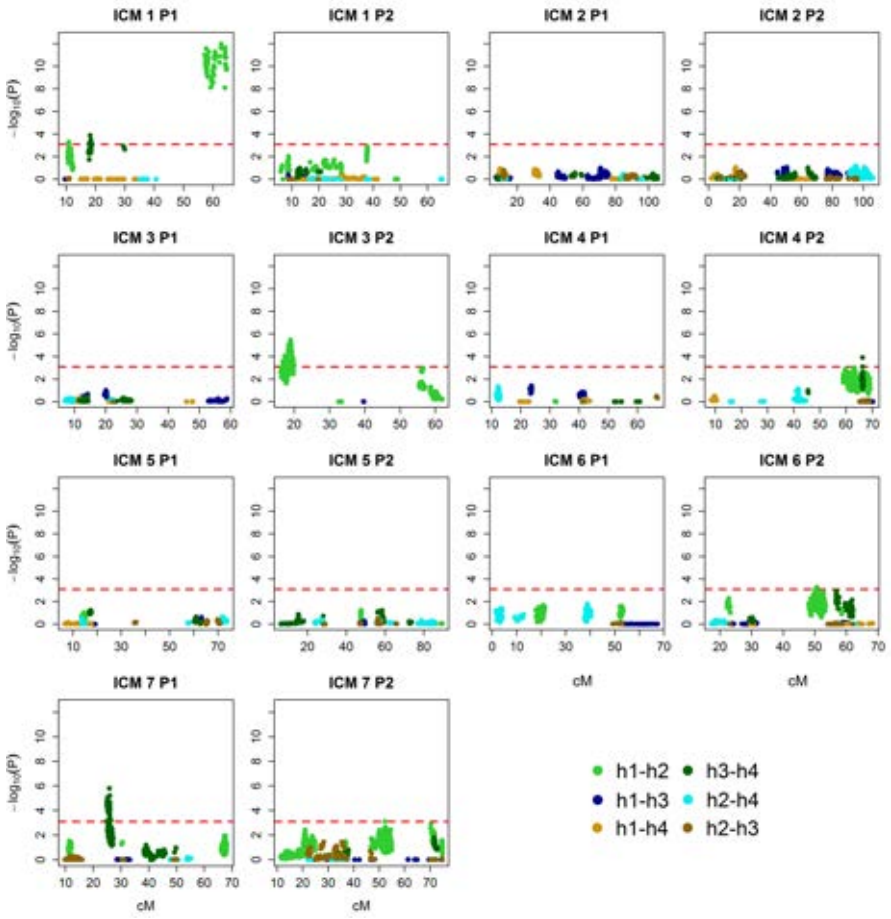



**Figure 2. Evidence for both disomic and tetrasomic behaviour from pairs of closely-mapped repulsion-phase SxN markers.** The $-\log_{10}(P)$ values of a Binomial test against the hypothesis $r_{r\_d} \geq \frac{1}{3}$ for SxN marker pairs in repulsion are plotted against their approximate genetic position. Regions of disomic behaviour (high values of $-\log_{10}(P)$) are not uniformly distributed, suggesting differences in pairing affinities along chromosomes. The significance threshold is shown as a dashed red line. h1-h2 refers to marker pairs from homologues 1 and 2. For simplicity, homologues 5-8 in parent 2 have been labelled 1-4.

## Ultra-high density *Rosa hybrida* linkage maps, tailored for segmental allopolyploidy

We modified our pairwise recombination frequency framework to include preferential pairing and re-calculated the recombination frequency, LOD score and phase for all affected linkage groups and then re-mapped these chromosomes using the amended estimates. For the vast majority of markers in the dataset, there was no change in the assignment to chromosomes or homologues (*i.e.* marker phasing was unaffected). We

found almost no difference in map order (Figure 3) but there were some differences in map lengths (between -0.6 and 5.2 cM longer). We expected maps to be on average longer from previous work with simulated data, which showed that ignoring preferential pairing when it occurs will lead to an under-estimation of recombination frequency (Bourke et al., 2016). In this population, the most extreme example of preferential pairing was found on ICM 1 of P1 ($\rho$ = 0.302, or about 45%), for which the corrected map was 5.21 cM longer.

All linkage groups were densely covered with markers, with the largest gap being 4.32 cM on ICM1 (Table 3). The final integrated map positions of 25,695 markers, using the corrections for ICM 1, 3 and 4 are provided in Table S4. The closest-linked SxN marker positions on the SNP map to a set of previously-mapped 1:1 segregating SSR and AFLP markers (Koning-Boucoiran et al., 2012) are provided in Table S5, which facilitated numbering the linkage groups according to the ICM numbering (Spiller et al., 2011).

**Table 3. Numbers of markers mapped per linkage group on the ultra-high density *Rosa hybrida* map.**
These maps represent the highest-density linkage maps in the genus *Rosa* published to date, helping to provide linkage information for current genome assembly efforts.

| ICM [a] | length / cM | N markers | Max. gap (cM) |
|---|---|---|---|
| 1 | 79.19 | 1865 | 4.32 |
| 2 | 108.67 | 6154 | 1.00 |
| 3 | 72.16 | 2912 | 1.18 |
| 4 | 77.30 | 2866 | 3.48 |
| 5 | 89.76 | 3799 | 0.69 |
| 6 | 71.82 | 4193 | 2.02 |
| 7 | 74.76 | 3906 | 0.53 |

[a] ICM = integrated consensus map (chromosome numbering)

# Prediction of quadrivalents and double reduction

When TetraOrigin was re-run with the possibility of quadrivalent pairing, we found a relatively high proportion of quadrivalents predicted across all chromosomes in both parents, ranging from 38% to 64% in P1 and 44% to 67% in P2 (Table S6). At a significance threshold of 0.001, we still found significant deviations from random pairing on ICM1, ICM3 and ICM4 (in the same parents, between the same homologues), but also on ICM6 of P1 and ICM7 of P2. In ICM6 of P1, the significant score stemmed from an underrepresentation of one bivalent pairing configuration. There was also some evidence for preferential pairing on ICM 7 of P2, similar to the prediction made using the repulsion-phase SxN markers (which was between the same two homologues). In other words, the results are consistent with those from the bivalent pairing model as well

as from the analysis of repulsion-phase linkages, with a similar set of chromosomes and parents implicated each time. The double reduction plots (Figure S1) were more jagged than those from a previous analysis of potato (Bourke et al., 2015), suggesting some difficulty in predicting double reduction in this dataset, although we did observe an increase in the rate of double reduction away from the centromeres, reaching a maximum of between 5 - 10% at the telomeres (Figure S1).

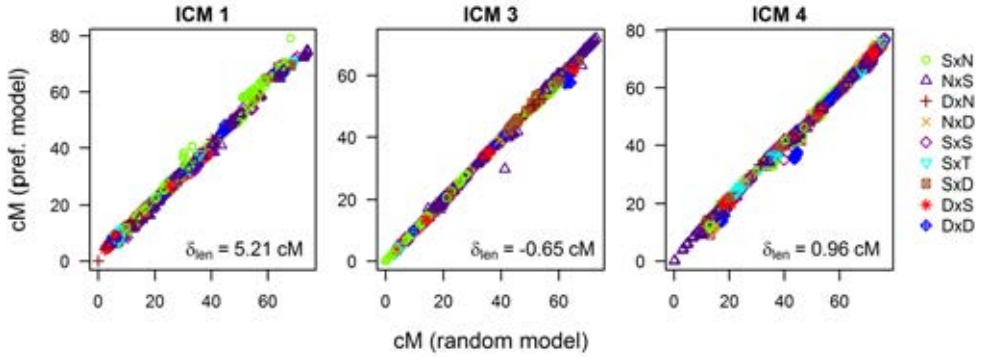



**Figure 3. Comparison between the integrated genetic map (in centiMorgans) under a random bivalent model (x-axis) and a bivalent model that takes account of preferential pairing (y-axis) for rose chromosomes 1, 3 and 4.** $\delta_{len}$ refers to the difference between the tetrasomic and mixosomic map lengths. Symbols are given for each of the nine marker segregation types encountered in a tetraploid. Correcting for mild preferential pairing in an otherwise tetrasomic species has little impact on the map order or length.

## Haplotype diversity and pairing behaviour

Of the 25,695 mapped markers, we could confidently assign full phase information to 20,090 of them (at least 10 linkages with LOD > 3 to a SxN homologue cluster), provided in Table S7. This set of markers was used to visualise the distribution of markers over the parental homologues (Figure S2). Using this marker subset, we compared the level of homology between homologues. There are 28 possible pairwise homologue combinations in a tetraploid cross, with 12 of these comparing within-parent haplotype diversity and 16 comparing between-parent diversity. There appeared to be no difference between the average within-parent homology versus between-parent homology (Figure S3) suggesting both parents carry haplotypes from the same breeding pool. The dissimilarity between haplotypes per parent varied greatly per chromosome, but could not be used as a predictor of pairing behaviour (Figure S4). That said, for ICM1 of P1 we observed a high level of homology between h1 and h2 from 20 cM to 50 cM, and a high level of homology between h3 and h4 from 50 cM to 80 cM, consistent with the predicted preferential pairing conformation.

## Synteny analysis with *Fragaria vesca*

Synteny has been reported between *Rosa* and woodland strawberry before (Gar et al., 2011;Kirov et al., 2014;Vukosavljev et al., 2016). In this study, we provide the most detailed picture of this synteny to date (Figure 4.a), which helps confirm the marker order of our maps. *Fragaria* chromosomes Fv1 and Fv6 are closely related to rose ICM2 and ICM3, with what appears to have been a reciprocal translocation occurring between them. We identified a major telomeric inversion between the gene orders on strawberry chromosome 5 and rose ICM7, as well as a second possible inversion on rose ICM6 or strawberry Fv2 (*c.f.* outer track of Figure 4.a). There is also a small fragment of the strawberry genome assembly that is currently unassigned (Fv0) with synteny to rose ICM5 from 34.8 – 48.1 cM. This suggests that this unassembled fragment should form part of chromosome Fv3, between 10.5 Mb and 14.9 Mb. We were unable to locate the pericentromeric regions of rose from this comparison; this may be due to the method of marker development for the WagRhSNP array, which used transcriptome data (targeting gene-rich euchromatin rather than heterochromatin) to identify marker sequences.

A set of 176 non-segregating DxN markers and 7 non-segregating NxD markers, which were originally filtered from the dataset, were double-checked against the BLAST results. Although these markers cannot be genetically mapped, they provide an indication of disomic behaviour (since they are present on exclusively-pairing homologues). Eighty-one of these markers produced a BLAST hit (72 of which had a BLAST E-value < $1 \times 10^{-20}$), and all 81 mapped to *Fragaria* chromosome 7, for which we have unambiguous evidence for synteny with rose ICM1 (Figure 4.a). Their putative positions ranged from 12.14 Mb – 19.87 Mb on the strawberry physical map (Table S8), which corresponds to the genetic region 47.7 cM – 65.9 cM on rose ICM 1. Interestingly, this is precisely the region for which the most highly-disomic repulsion-phase estimates were previously identified (Figure 4.b) and they therefore provide further evidence for highly preferential pairing on this chromosome in this parent, which may be confined to a specific chromosomal region. Regarding the non-segregating NxD markers, only two produced BLAST hits, one to Fv4 and the other to Fv5 (Table S8), which show synteny to rose ICM4 and ICM7 respectively, consistent with our previous findings.

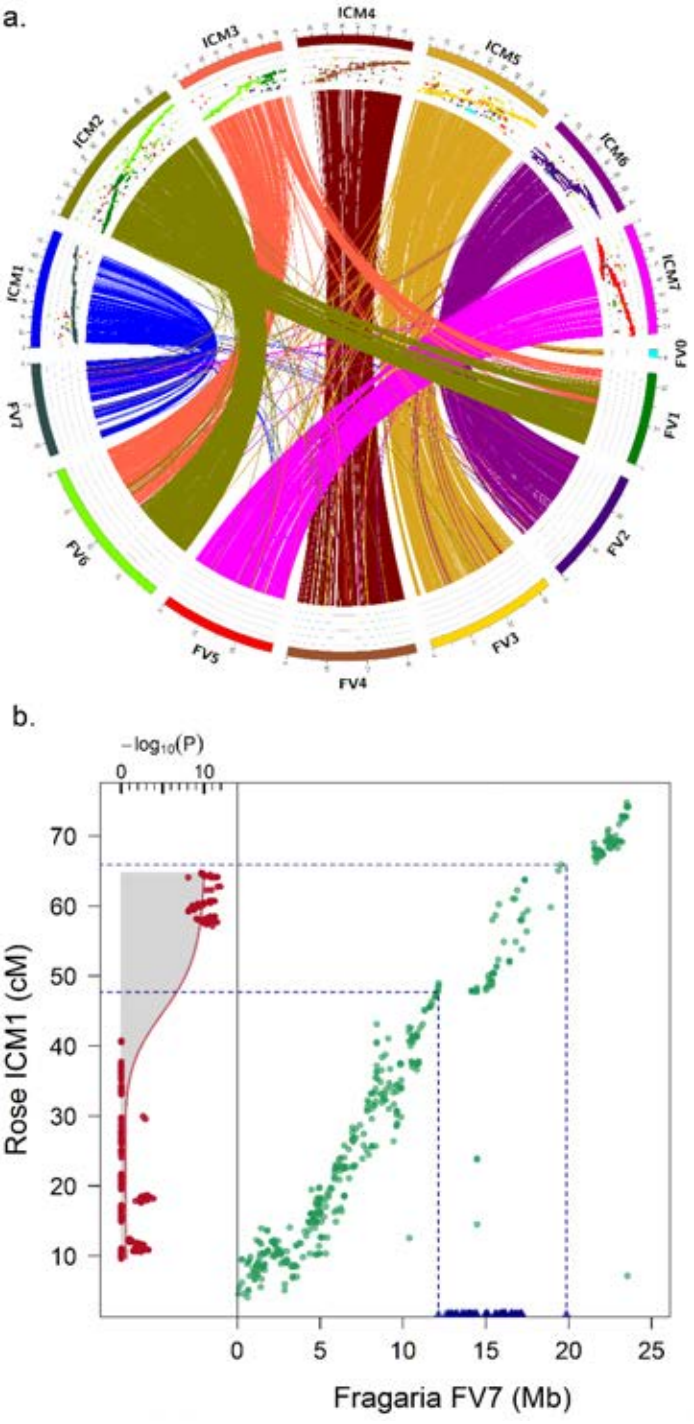



**Figure 4. Overview of the synteny between *Rosa hybrida* and *Fragaria vesca*. a.** Synteny between integrated genetic linkage maps of cut rose and *Fragaria vesca* pseudomolecules v2.0.a1, with precise correspondence of position displayed in the outer track**.** FV0 indicates unassigned assembly of *Fragaria vesca* (available from www.rosaceae.org). Rose genetic linkage group numbering follows the integrated consensus map (ICM) numbering of rose. Gene order is highly conserved between these species, and a number of interesting features such as inversions (ICM 6 and ICM7) and reciprocal translocations (ICM2 and ICM3) are clearly visible. **b.** Detailed comparison of the syntenic relationship between *Fragaria* chromosome 7 and rose ICM1, with non-segregating DxN markers shown as blue triangles**.** Non-segregating DxN markers suggest complete disomic behaviour. These set of markers would have mapped to the same region of rose ICM1 that contains strong evidence for disomic behaviour from repulsion-phase SxN marker pairs, confirming this area differs in its pairing affinity from the rest of the chromosome.

# Discussion

## Non-uniform distribution of preferential pairing

The observations in this study suggest that preferential pairing is not uniformly distributed in the rose genome. We detected differences in pairing behaviour between parents, among chromosomes, and even along a single chromosome where pairing behaviour can be preferential at one chromosome arm but tetrasomic at the other. The original description of segmental allopolyploidy does not preclude this sort of behaviour (Stebbins, 1947). Indeed, in keeping with Stebbins' description, *Rosa* displays many characteristics of an autopolyploid, with evidence for quadrivalent pairing, double reduction as well as the majority of our marker data fitting a tetrasomic segregation model. We have established with this work that rose may be categorised as a segmental allotetraploid, although it is impossible to predict whether this is ultimately a stable conformation (Sybenga, 1996) and whether it is genus-wide rather than population-specific. We have uncovered evidence that the strength of pairing can vary along a chromosome, potentially complicating genetic studies and methodologies, which invariably start with some assumptions about the uniformity of pairing behaviour. This is not the first reported instance of intra-chromosomal "mixosomy"; in rainbow trout it has been proposed that there may be "residual tetrasomy" (Allendorf and Danzmann, 1997) leading to variable pairing behaviour along a chromosome, albeit with disomic regions confined to the more central regions. A similar phenomenon has also recently been reported in peanut (Leal-Bertioli et al., 2015;Nguepjop et al., 2016). It is possible that the pairing behaviour in cultivated tetraploid *Rosa* is a consequence of its rather exotic pedigree, similar to an observation on the increased occurrence of mixed segregation patterns in trout populations derived from interspecific hybridisation (Allendorf et al., 2015). On the other hand, if rose is the result of hybridisation among 7-14 species, it is surprising that most chromosomes of this cross exhibit random pairing. Species hybridisation may exhibit tetrasomic pairing if the underlying species are young, as suggested by a previous study of the low levels of variation in ribosomal and chloroplast markers in rose (Wissemann and Ritz, 2005). Cultivated roses have an unusual recent pedigree (which ornamental species in particular, but also many other domesticated crop species have experienced), but the same may be true for *e.g.* naturally occurring polyploid roses (including the unbalanced meiosis in pentaploid dogroses (Herklotz and Ritz, 2016)). The genus *Rosa* is more than 30 million years old, but hybridsation between species has been extensive (Koopman et al., 2008;Fougère-Danezan et al., 2015), which may have led to widespread mixing of germplasm within hybrid rose varieties. Indeed, (Zhang et al., 2013b) found the same microsatellite haplotypes with flanking SNP markers in diploid and polyploid rose species, while

(Bruneau et al., 2007) found the same chloroplast gene haplotypes in multiple species. A more precise answer may therefore come from genome sequences of various diploid and polyploid accessions.

It is also worth noting here that preferential pairing and marker skewness are different phenomena that may be confused. Marker skewness can be caused by selection and can cause problems in linkage map construction (Van Ooijen and Jansen, 2013). For example, lethality or reduced fitness is unlikely to be caused by a combination of alleles present in one of the parents (since that parent is alive) and therefore, cross-parental lethal combinations are more probable, which will not result in a preferential pairing signal in one of the parents. Similarly, if selection is positive (*e.g.* at a resistance locus) then the skewness is not in favour of a particular *combination* of parental homologues but rather of specific allele(s); in cases where there are two functional copies in a single parent the effect would appear as selection *against* a particular homologue combination (*i.e.* "unpreferential" pairing). In any case, in this study we only mapped non-skewed markers, so that all our evidence for preferential pairing comes from markers with no significant skewness. This is also visualised in Figure S5, showing that skewness levels among mapped markers were essentially uniform across the rose genome.

## Predicting pairing behaviour through homology

We were interested to see whether pairing behaviour might be predicted from the homology as defined by our haplotype-specific linkage maps. We chose a simple dissimilarity measure to analyse the diversity between homologues in this study, similar to the most recent common ancestor (MRCA) measure used in population genetic studies (Joly et al., 2015). We did not observe any clear relationship between homology as estimated from the mapped markers and pairing behaviour. However, as in all mapping studies, our data was biased towards higher diversities since only segregating markers can be included on a genetic map. We should therefore treat these results with caution as they do not give a complete picture of homology, which only sequencing can fully reveal. Telomeric homology, where pairing initiation is thought to occur (Sved, 1966;Sybenga, 1975;Cifuentes et al., 2010), might have more influence than chromosome-length homology and could be a target of future sequencing efforts to clarify this.

## Tailored genetic linkage maps

In this study we generated genetic linkage maps tailored to the particular pairing behaviour of each chromosome. We could have further refined our maps by allowing the estimate for the rate of preferential pairing to vary along the chromosome arms if this was deemed necessary. However, in this population the rate of preferential pairing on

the affected chromosomes was low enough that we could have ignored it in map construction. We can confirm our prediction that linkage mapping can be safely performed under the assumption of random pairing up to a level of preferential pairing as high as 70% ($\rho \sim 0.467$) after which estimates become seriously biased (Bourke et al., 2016). This finding will likely be welcomed by researchers in *Rosa*, for which the assumption of tetrasomic inheritance has generally been used. In this study, we used DxN markers to provide links between putative homologue clusters of SxN markers (helping to identify chromosomal linkage groups). However, these may not always be informative in cases of extreme preferential pairing, where simplex x simplex (SxS) markers may need to be used to provide cross-homologue linkages instead.

## Synteny between *Rosa* and *Fragaria*

In this study we used the close relationship between *Rosa* and *Fragaria* to confirm the order and organisation of the linkage maps produced, as well as allowing us to position a set of non-segregating DxN markers which were indicative of extreme preferential pairing in a distal region of maternal chromosome ICM1. It also yielded a detailed picture of the high level of genomic conservation between these species (Figure 4a,b). Both *Rosa* and *Fragaria* belong to the *Rosoideae* clade (in the *Potentilleae* and *Roseae* tribes, respectively), whose divergence is estimated to have occurred approximately 60 mya (million years ago) (Xiang et al., 2016). High levels of gene conservation have also been reported in the grass family, for example (Bennetzen, 2007). On the other hand, despite an estimated divergence time of 27 mya, there is a much more striking difference in the genomic configuration of *Brassica rapa* and *Arabidopsis thaliana* (Murat et al., 2015), although this could be due to the fact that Brassicaceae possess an exceptionally high rate of biological radiation and diversification (Franzke et al., 2011). It is tempting to speculate why rates of genomic divergence differ between plant families (for example, the level of activity of transposable elements within the genome (Bennetzen and Wang, 2014)), although such a meta-analysis using data across multiple plant families has yet to appear in the literature. What our study does make clear is that the published reference sequence of *Fragaria vesca* continues to be a very useful genomic resource for the Rose research community, at least until such time as a Rose sequence becomes publically available.

## Concluding remarks

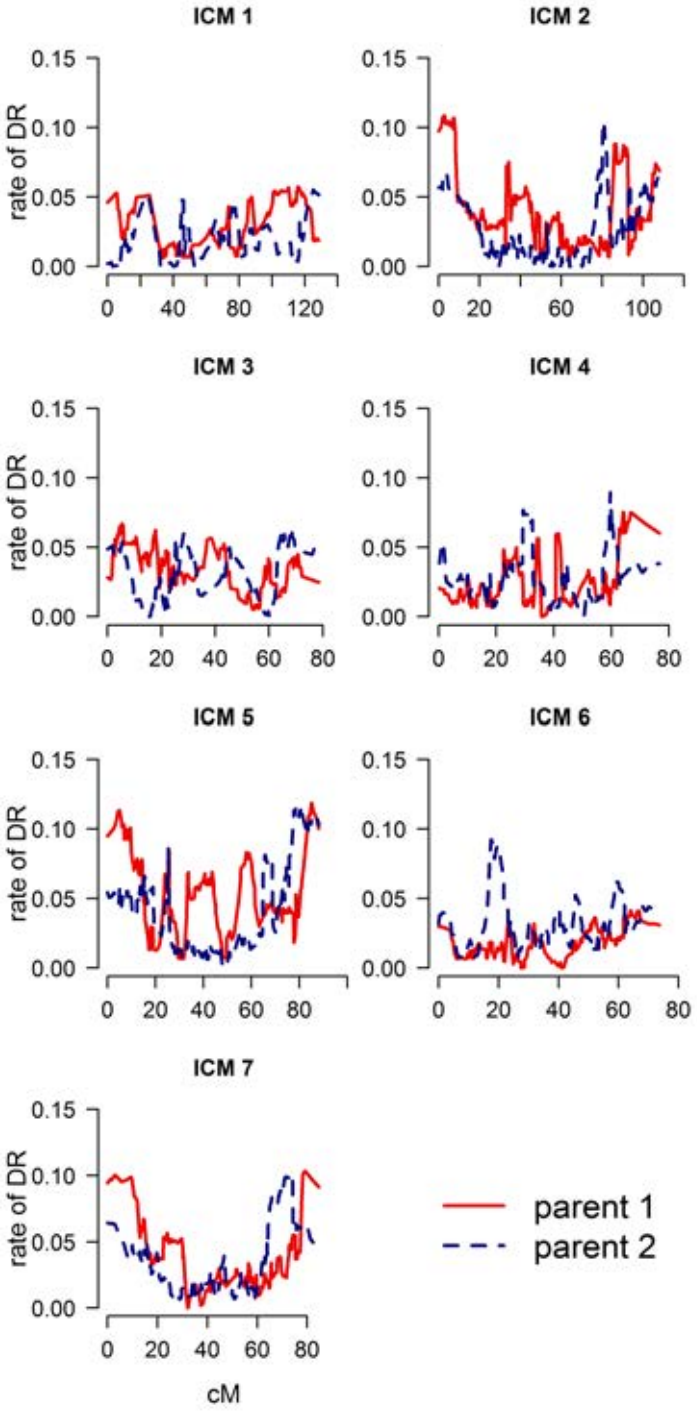
In this study we have applied a novel approach for detecting and quantifying the level of preferential pairing in a tetraploid biparental population of cut rose, and used this information to refine our estimates of recombination to produce a tailored genetic map of this economically important species. We found conclusive evidence for parent-specific preferential pairing behaviour on three chromosomes whereas the other

chromosomes showed tetrasomic behaviour. The picture that emerges is that the meiotic behaviour in polyploids may be difficult to predict without detailed analysis. We found that meiotic pairing behaviour can vary from chromosome to chromosome and from genotype to genotype, and perhaps even along chromosome arms. Tools which utilise chromosome-wide inheritance information across all homologues are very powerful in revealing underlying meiotic mechanisms and confirm our findings using repulsion-phase linkage information as well as information from syntenic mapping of non-segregating duplex markers to infer inheritance type. We uncovered a detailed picture of conserved synteny with the closely-related woodland strawberry, highlighting both telomeric inversions and reciprocal translocations which occurred somewhere in the lineages of one or both of these species. Our ultra-high-density linkage map will facilitate down-stream applications such as current efforts at genome assembly as well as providing a basis for detailed QTL analysis in the future.

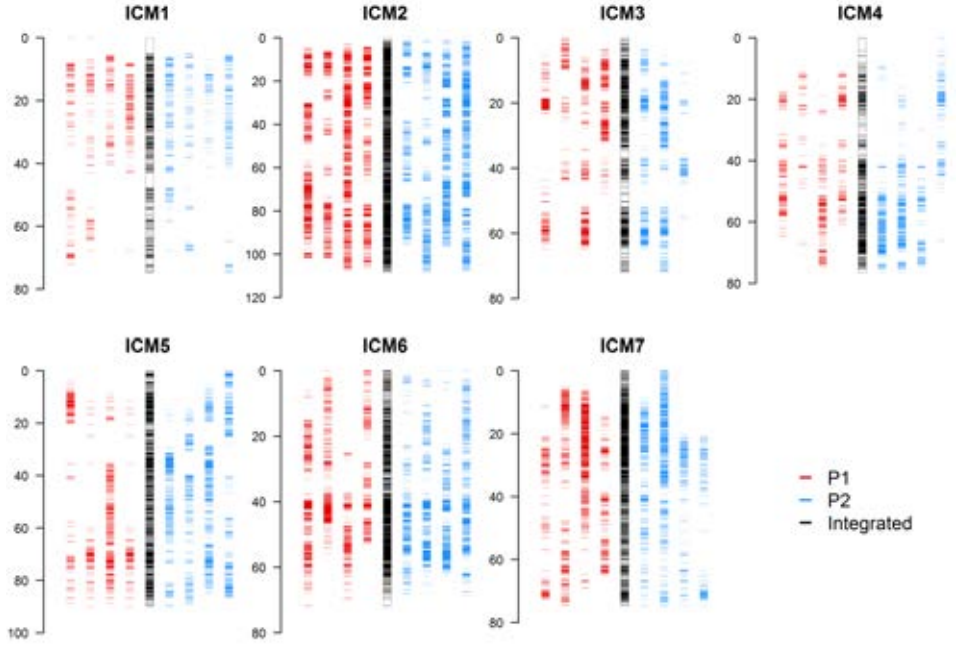
# Acknowledgements

The authors would like to thank Dr. Christine Hackett and Dr. Katherine Preedy (BioSS) for sharing a developmental version of MDSMap, which was used in this study. The authors also wish to acknowledge the two anonymous reviewers, whose comments helped improve the manuscript. Funding for this research was provided through the TKI polyploids project “A genetic analysis pipeline for polyploid crops”, project number BO-26.03-002-001.

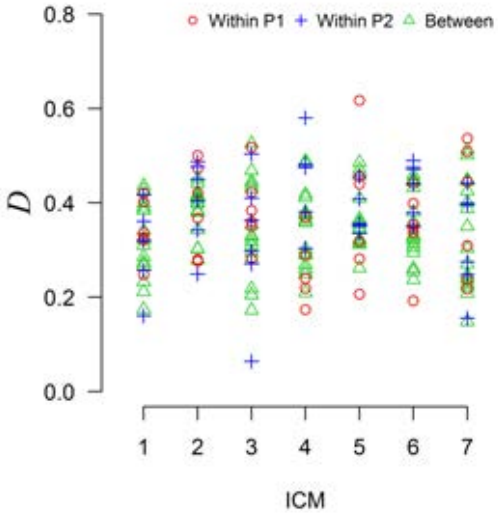
***Supplementary data is available online at:***
http://onlinelibrary.wiley.com/doi/10.1111/tpj.13496/full



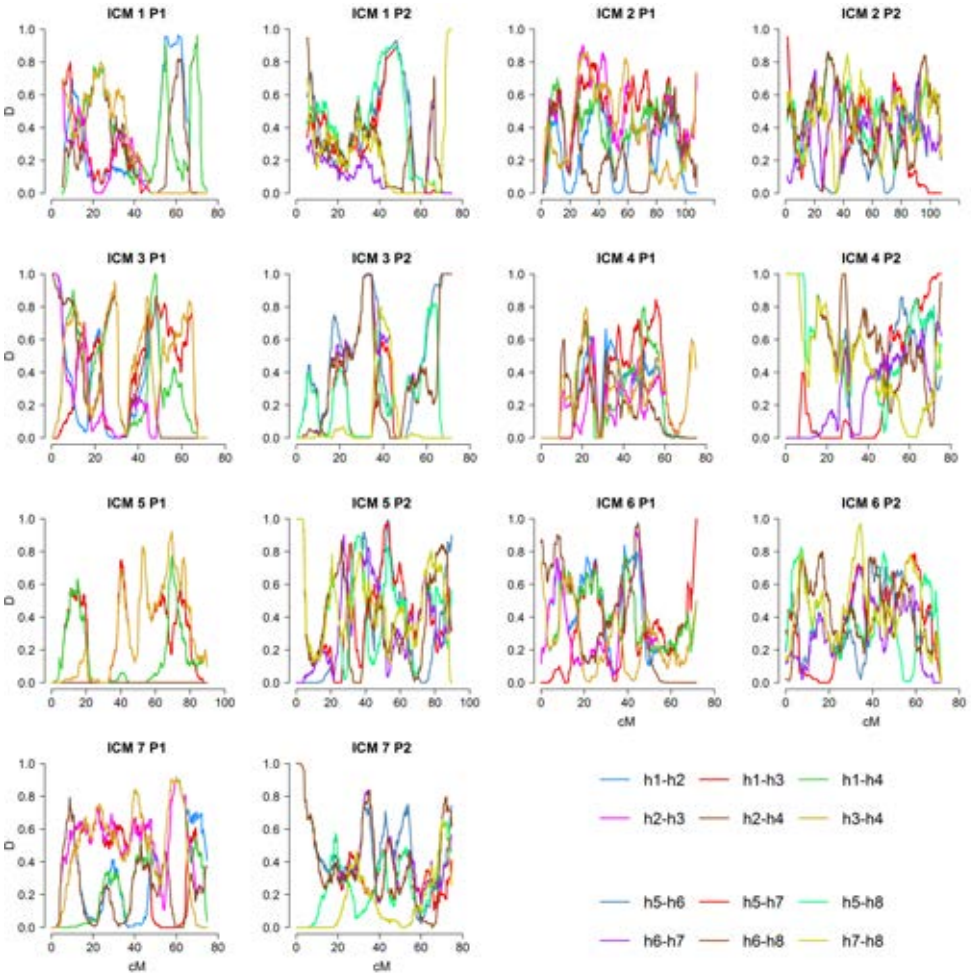

**Supplementary Figure S1.** Average estimated probability of double reduction (DR) per parental meiosis over the seven rose chromosomes.



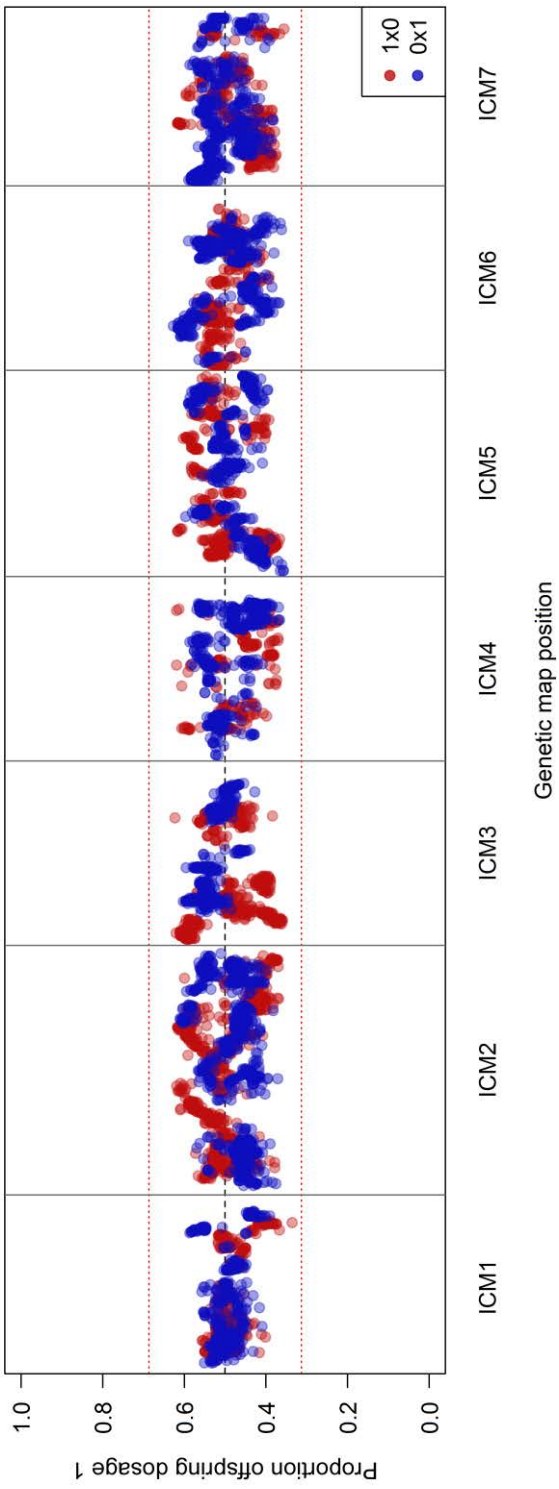

**Supplementary Figure S2.** Visualisation of meiotically-tailored high-density linkage maps, showing the distribution of segregating marker alleles across all eight parental homologues for each of the seven rose linkage groups (ICM1 – 7), also listed in Supplementary Table S7 using original marker dosage coding.




**Supplementary Figure S3.** Average within-parent and between-parent dissimilarity (*D*) for each of the seven rose chromosomes. Multiple points correspond to each set of homologue pairs within these categories.




**Supplementary Figure S4.** Average pairwise dissimilarity (*D*) between parental haplotypes, plotted against genetic position (cM).




**Supplementary Figure S5.** Proportion of offspring with simplex scores (dosage 1) at 10,651 SxN and NxS marker positions across the rose linkage maps. Parent 1 markers (1x0) are shown in red and parent 2 markers (0x1) are shown in blue (deeper colours correspond to overlapping markers). Boundaries (red dotted lines) correspond to an FDR-adjusted p-value of 0.05 (*i.e.* corrected for multiple testing using FDR with $\alpha = 0.05$) assuming a $\chi^2$ distribution. At this level of significance there were no highly-skewed simplex markers.

# Chapter 6

# polymapR – linkage analysis and genetic map construction from $F_1$ populations of outcrossing polyploids

Peter M. Bourke[1], Geert van Geest[1,2], Roeland E. Voorrips[1*], Johannes Jansen[3], Twan Kranenburg[1], Arwa Shahin[1,4], Richard G. F. Visser[1], Paul Arens[1], Marinus J. M. Smulders[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[2] Deliflor Chrysanten B.V., Korte Kruisweg 163, 2676 BS Maasdijk, The Netherlands.

[3] Biometris, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[4] Van Zanten Breeding B. V., Lavendelweg 15, 1435 EW, Rijsenhout, The Netherlands.

*Published (with modifications) as* Bourke, P.M., van Geest, G., Voorrips, R.E., Jansen, J., Kranenburg, T., Shahin, A. *et al.* (2018). "polymapR – linkage analysis and genetic map construction from $F_1$ populations of outcrossing polyploids", **Bioinformatics, doi: 10.1093/bioinformatics/bty371**

# Abstract

## Motivation

Polyploid species carry more than two copies of each chromosome, a condition found in many of the world's most important crops. Genetic mapping in polyploids is more complex than in diploid species, resulting in a lack of available software tools. These are needed if we are to realise all the opportunities offered by modern genotyping platforms for genetic research and breeding in polyploid crops.

## Results

polymapR is an R package for genetic linkage analysis and integrated genetic map construction from bi-parental populations of outcrossing autopolyploids. It can currently analyse triploid, tetraploid and hexaploid marker datasets and is applicable to various crops including potato, leek, alfalfa, blueberry, chrysanthemum, sweet potato or kiwifruit. It can detect, estimate and correct for preferential chromosome pairing, and has been tested on high-density marker datasets from potato, rose and chrysanthemum, generating high-density integrated linkage maps in all of these crops.

## Availability and Implementation

polymapR is freely available under the general public license from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/package=polymapR.



# Key words

autopolyploid genetic map, triploid, tetraploid, hexaploid, partial preferential chromosomal pairing, segmental allopolyploid

# Introduction

In recent years there has been an acceleration of progress in the understanding of the genetics underlying important traits in autopolyploid species. This has been to a large extent due to developments in high-density genotyping platforms for single nucleotide polymorphism (SNP) markers, which have found increasing application in polyploids. For example, high-density SNP arrays have been developed in potato (Felcher et al., 2012;Vos et al., 2015), rose (Koning-Boucoiran et al., 2015), alfalfa (Li et al., 2014b) and chrysanthemum (van Geest et al., 2017c), enhancing the scope for genetic studies in these species.

In polyploid species, as opposed to diploids, co-dominantly scored markers can possess multiple classes in the heterozygous condition, usually termed marker "dosage". In a tetraploid there are five possible dosage classes of a bi-allelic SNP marker, namely nulliplex with a dosage 0 for one of the alleles, simplex with dosage 1, duplex with dosage 2, triplex with dosage 3, and quadruplex with dosage 4. In a hexaploid, the number of dosage classes at a bi-allelic locus rises to seven. Various software have been developed to convert the signal from *e.g.* SNP arrays into these discrete dosage calls for polyploids, such as fitTetra (Voorrips et al., 2011) or ClusterCall (Schmitz Carley et al., 2017).

Genetic linkage maps have traditionally been used for both exploratory trait mapping (often termed QTL analysis) and the subsequent fine mapping of traits, as well as for assisting genome assembly efforts by guiding the integration and orientation of contigs. High-density linkage maps may also improve our understanding of the chromosomal composition and genetics of polyploid species, uncovering such phenomena as double reduction or partially-preferential chromosome pairing. In many polyploid species which lack reference genome sequences, linkage maps are also a (vital) first genomic description of that species.

Despite the importance of both linkage maps and polyploid species, there are still relatively few software tools available for polyploid linkage map construction. Allopolyploid species showing disomic inheritance can be treated (genetically-speaking) as diploids, with a wide range of software options available. In the case of polysomic polyploids (autopolyploids and segmental allopolyploids), the options available to the research community are limited. Probably the most well-known autopolyploid mapping software is TetraploidMap (Hackett and Luo, 2003;Hackett et al., 2007), which has been used in studies of various autotetraploid species such as potato, alfalfa, rose and blueberry (*e.g.* (Bradshaw et al., 2008;Robins et al., 2008;Gar et al., 2011;McCallum et al., 2016)). Recently, its successor TetraploidSNPMap (TSNPM) has been released to

accommodate high-density marker data from SNP arrays (Hackett et al., 2017). However, it can only handle autotetraploid datasets and provides a graphical user interface for the Windows platform only. Linkage studies in species exhibiting strong preferential chromosomal pairing or other ploidy levels are not currently possible using this software. An alternative polyploid mapping software is the PERGOLA package in R (Grandke et al., 2017). However, this software has been developed for use with $F_2$ or backcross populations from homozygous parents only. In many cases, either due to inbreeding depression or the difficulties imposed by polysomic inheritance, $F_1$ populations from two heterozygous parents are typically used instead.

In short, there is currently no software which can perform linkage mapping at various ploidy levels under a variety of inheritance models for outcrossing species using dosage-scored marker data. Here we present polymapR, an R package (R Core Team, 2016) for linkage mapping in outcrossing polyploid species which can generate linkage maps for polysomic triploids, tetraploids and hexaploids, accommodating either fully tetrasomic or mixed meiotic pairing behaviour (segmental allopolyploidy) at the tetraploid level. Its modularity will facilitate its adaption to other marker genotyping technologies or ploidy levels in the future.

# System and methods

The polymapR pipeline consists of four parts – data inspection, linkage analysis, linkage group assignment and marker ordering, which are detailed below. A description of the functions within polymapR is described in the vignette which accompanies the package, going through all the steps in a typical mapping project. For consistency and simplicity, all examples mentioned here describe a tetraploid cross.

## 1. Data inspection, filtering and preparation for linkage analysis

The input data for polymapR is dosage-scored marker data, available from a number of packages such as fitTetra (Voorrips et al., 2011) or ClusterCall (Schmitz Carley et al., 2017). Both fitTetra and ClusterCall are limited to tetraploid data; extensions to fitTetra to accommodate multiple ploidy levels are underway. Regardless of how it is generated, the input dosage-scored marker data should consist of a column of marker dosages for the mother, one for the father followed by a column for each of the offspring of the $F_1$ cross. Checks for marker skewness and shifted markers (when dosage scores are shifted by a fixed amount) are currently provided in polymapR from a suite of tools developed for the fitTetra package (Voorrips et al., 2011).

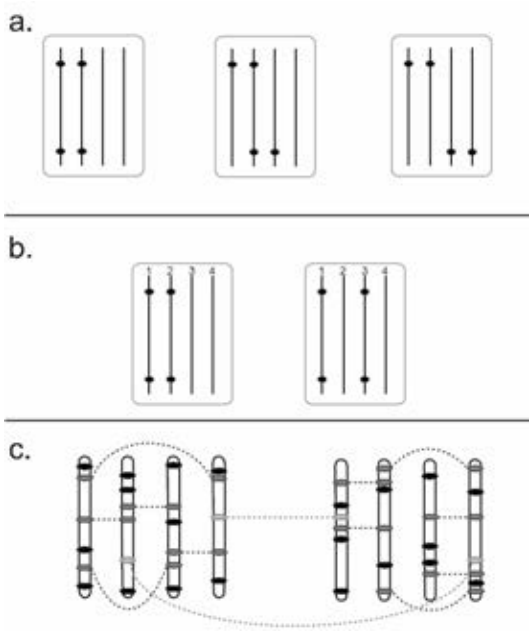



**Figure 1. Phase considerations and clustering strategy in a tetraploid. a.** The three phases considered for a pair of 2x0 markers, from left to right, "coupling", "mixed" and "repulsion". **b.** In the case of preferential pairing between homologues 1-2 and 3-4, we must consider two separate types of coupling phase, either coupling within preferential bivalents (left) or coupling between preferential bivalents (right). In the extreme case of an allotetraploid, this distinction could also be termed "subgenome-specific" versus "subgenome-straddling".
**c.** Simplex x nulliplex (1x0) markers (solid black dots) uniquely define homologous chromosomes and are initially clustered together. Higher-dosage marker types such as duplex x nulliplex (2x0) markers (dark grey) provide linkage associations between simplex x nulliplex homologues, helping to identify chromosomal linkage groups. Cross-parental markers such as simplex x simplex (1x1, light grey) can also link these groups together, leading to consistent linkage group numbering across parents.

The next step in data preparation is the conversion of marker dosages to their simplest form, such that the sum of the parental dosage scores is minimised. There are two possible conversions – a relabelling of the reference and alternative allele in both parents, or a single-parent relabelling if the other parent is homozygous. Marker conversions are performed to reduce the number of marker segregation classes for the linkage analysis (which is directed according to the parental dosages), but have no effect on the pairwise results. In a tetraploid there are nine fundamental segregation types, rising to nineteen for a hexaploid. Identifiable double reduction scores are preserved during conversion (*e.g.* a dosage of 0 from a triplex x nulliplex (3x0) marker becomes a dosage of 2 in its converted form as a simplex x nulliplex (1x0) marker), allowing an investigation of double reduction post-mapping. Any impossible scores (like a dosage of 3 or 4 from a 1x0 marker) are made missing.

High-quality data facilitates the generation of high-quality maps. One indication of poor data quality is a high proportion of missing values. The user may choose to screen out markers or individuals with more than a desired rate of missing values (by default up to 10% is tolerated), or duplicate individuals. Identical markers, which often occur in high-density marker datasets with limited population sizes and hence a limited number of recombination events, can be identified and reduced to one representative marker for the mapping steps, and reintegrated later. A principal component analysis (PCA) can also be performed and visualised, which may highlight some unwanted structure in the population (for example due to pollination from an unknown external pollen parent or from self-pollination) or outlying individuals (for example because of admixture).

# 2. Linkage analysis

### *2.1 Linkage analysis under a polysomic model*

In autopolyploid species with polysomic inheritance, it is possible to model meiotic pairing structures as random bivalents or multivalents. In practice, both pairing structures tend to occur, with a relatively low frequency of multivalents in stable autopolyploids (Santos et al., 2003;Bomblies et al., 2016). The main consequence of multivalent formation from a genetic perspective is the phenomenon of double reduction, where two segments of a particular homologue can end up in the same gamete and become transmitted together to $F_1$ offspring. It has been demonstrated that double reduction introduces some bias in recombination frequency estimates under a random bivalent model. This can be safely ignored if the rate of quadrivalent pairing is low (Bourke et al., 2015;Bourke et al., 2016).

Under a random bivalent model, there are three possible bivalent pairing conformations in a tetraploid. In general, for any even ploidy $p = 2n$ there are $c = \frac{(2n)!}{(2^n).n!}$ possible bivalent pairing conformations to be considered. Given any pair of marker loci with unknown recombination frequency $r$, we consider the contribution of recombinant homologues with a within-bivalent probability of $\frac{1}{2}r$ and non-recombinant homologues with a within-bivalent probability of $\frac{1}{2}(1-r)$. In cases where recombinants and non-recombinations cannot be distinguished, both are assigned a probability of $\frac{1}{2}$. Assuming random pairing, the probability of any particular pairing configuration is $\frac{1}{c}$ (in the case of preferential pairing, we introduce a preferential pairing factor to model deviations from randomness here).

The expected frequency of each offspring class $n_{ij}$ $(0 \leq i, j \leq 2n)$ is first summed over all $c$ bivalent conformations:

$$E(n_{ij}) = \sum_{k=1}^{c} \frac{1}{c} f_k(r, 1-r)$$

where $f_k(r, 1-r)$ denotes a function of $r$ and $1-r$, dependant on the marker combination considered. Given these expected frequencies, we relate them to the observed counts of individuals in each class $O(n_{ij})$ to yield the likelihood function $\mathcal{L}(r)$:

$$\mathcal{L}(r) \propto \prod_{i,j=0}^{2n} E(n_{ij})^{O(n_{ij})}$$

The likelihood equation results from equating the first derivative of the log of the likelihood function with zero:

$$\sum_{i,j=0}^{2n} \left( O(n_{ij}) \,.\, \sum_{k=1}^{c} \frac{1}{c} \frac{d}{dr} \ln(f(r, 1-r)) \right) = 0$$

In cases where no analytical solution exists, we use Brent's algorithm (Brent, 1973) to numerically maximise the log likelihood function in the bounded interval $0 \leq r < 0.5$. For any pair of markers there are a number of possible phases between these markers to consider, which describe the physical linkage between marker alleles. In the case of a pair of duplex x nulliplex (2x0) markers, these phases are termed "coupling", "mixed" and "repulsion" (Figure 1.a). As the phase between markers is initially unknown, we must compute expressions for each of the possible phases, and select the most likely as the phase for which $0 \leq r < 0.5$ and which maximises the log of the likelihood (Hackett et al., 2013).

Finally, we also compute the logarithm of odds (LOD) score, which provides a useful measure of the confidence in the estimate and is used for both marker clustering and marker ordering:

$$LOD = log_{10} \left( \frac{\mathcal{L}(r = \hat{r})}{\mathcal{L}(r = 0.5)} \right)$$

where $\hat{r}$ is the maximum likelihood estimate of $r$.

### *2.2 Linkage analysis in the presence of preferential chromosomal pairing*

In certain polyploid species the meiotic pairing is neither fully random nor fully partitioned into exclusively-pairing subgenomes, a situation described as segmental allopolyploidy (Stebbins, 1947). Regardless of the underlying mechanism, the result of

preferential pairing is that both the segregation ratios and the co-inheritance of marker alleles are affected. In the example of a 2x0 marker introduced earlier, the expected segregation ratio in a polysomic autotetraploid is 1:4:1. With increasing preferential pairing, this ratio will approach 1:2:1 in the case of subgenome-straddling markers (Figure 1.b right), or approach non-segregation in the case of subgenome-specific markers (Figure 1.b left).

In order to model this behaviour, we introduce a preferential pairing parameter $\rho$, such that (in the case of a tetraploid) the probability of the chromosome pairing configuration 1-2 / 3-4 is $\frac{1}{3} + \rho$ and the probability of pairing configurations 1-3 / 2-4 and 1-4 / 2-4 is $\frac{1}{3} - \frac{\rho}{2}$. Attempting to model preferential pairing at higher ploidy levels introduces further complications; Zhu *et al.* (2016) have proposed a solution for hexaploids by introducing three preferential pairing parameters $\theta_1$, $\theta_2$ and $\theta_3$ to model deviations in bivalent configurations 1-2, 3-4 and 5-6 respectively, with all other configurations having a probability of $\frac{1}{15} - \frac{1}{12}(\theta_1 + \theta_2 + \theta_3)$. In our software, we have not yet attempted to model segmental allohexaploidy, and confine our attention to the tetraploid level for now.

We do not simultaneously estimate $\rho$ and $r$, which can lead to an over-estimation of the preferential pairing parameter (Wu et al., 2002). Instead, we estimate the chromosome-wide strength of preferential pairing after map construction and thereafter correct the pairwise recombination frequency estimates to revise the maps. A robust method of preferential pairing detection and estimation is to use inheritance probability estimates such as those provided by TetraOrigin (Zheng et al., 2016); in polymapR we offer a simpler likelihood-based approach which uses closely-linked repulsion marker pairs to test for deviations from random pairing and simultaneously estimate the strength of this deviation:

$$\rho = \frac{2(n_{00} + n_{11}) - 4(n_{01} + n_{01})}{3(n_{00} + n_{01} + n_{10} + n_{11})}$$

where $n_{01}$ is the number of offspring with a dosage 0 at marker A and 1 at marker B *etc.*

Given a parent- and chromosome-specific estimate for the preferential pairing factor $\rho$, we modify the expression for the expected frequency of individuals in marker class $n_{ij}$ of a tetraploid as follows:

$$E(n_{ij}) = \left(\frac{1}{3} + \rho\right) f_1(r, 1-r) + \left(\frac{1}{3} - \frac{\rho}{2}\right) f_2(r, 1-r) + \left(\frac{1}{3} - \frac{\rho}{2}\right) f_3(r, 1-r)$$

Due to the lack of symmetry, we must consider all possible conformations *within* each phase, an example of which is shown in Figure 1.b. The procedure for estimating *r* and LOD remain otherwise the same. The inclusion of preferential pairing imposes an extra computational burden as each phase can have up to four sub-phase conformations, all of which are calculated prior to selection of the most likely phase and its associated *r* and LOD score.

Finally, in both the case of random and preferential pairing, linkage calculations can be run in parallel (using doParallel (Microsoft Corporation and Weston, 2017)) on any Windows or Unix-like multi-core desktop computer resulting in significant time-savings. High-density marker datasets with tens of thousands of markers can be processed in a few hours.

# 3. Linkage group assignment

In diploid studies, the term linkage group is loosely synonymous with the term chromosome. In autopolyploids two levels of linkage group exist – homologue groups and integrated chromosomal groups. The first step in linkage group assignment is to cluster the 1x0 linkage data into homologue groups, for which we currently use the R package igraph (Csardi and Nepusz, 2006). Clustering is performed using the pairwise linkage LOD scores, although the LOD for independence can be used if desired, which may be more robust with skewed marker data (Van Ooijen and Jansen, 2013).

A number of visual aids are provided to assist in clustering (Figure 2). In general, clustering should be performed over a suitable range of LOD thresholds (*e.g.* from LOD 3 to 10) in order to inform the choice of LOD score to partition the data into both homologues and chromosomes (Figure 2.a, b). If chromosome and homologue clusters cannot be readily identified using 1x0 markers alone, coupling-phase homologue clusters are first identified at a high LOD and later re-connected into chromosomal clusters using a higher-dose marker type (Figure 1.c). Visualisations help display the strength of associations between homologues (Figure 2.c). Occasionally homologues may split apart; various possibilities to merge these fragments are provided (Figure 2. d, e, f).

In the case of triploid populations, the phasing approach differs between the diploid and tetraploid parents: for the diploid parent, phasing can be achieved directly using the phase assignment from the linkage analysis. Following the definition of the chromosome and homologue structure using the 1x0 markers, all other marker segregation classes are assigned to both homologues and chromosomes using their linkage to these markers, generating the final phase assignment of all marker types.

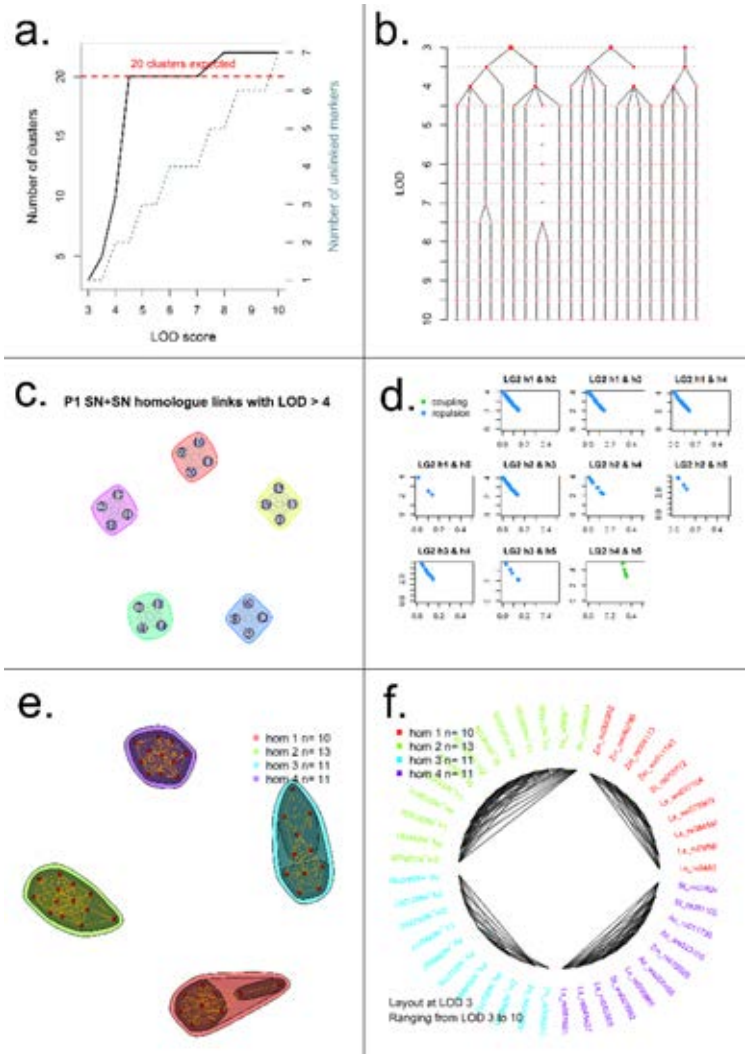



**Figure 2. Example visualisations produced by polymapR to facilitate linkage group identification and marker clustering. a.** As LOD score is increased, the number of 1x0 clusters increases, as does the number of single-marker clusters (unlinked markers). For a given ploidy and chromosome base number, the expected number of (homologue) clusters is also shown. **b.** Alternative representation of (a) which shows the splitting of each cluster as the LOD score is increased. In this example, five chromosomal clusters are identified at LOD 3.5, each of which split into four homologue clusters between LOD 4.5 and 7. **c.** Using linkage to other marker segregation types such as 2x0 markers, homologue clusters can be associated into chromosomal clusters, if this was not achieved using 1x0 data alone. Here, five chromosomes are represented. **d.** If homologues fragment, cross-homologue phase information can help determine which fragments to merge. Here, homologues 4 and 5 show only coupling-phase linkage and should therefore be joined as a single homologue.
**e.** Alternative approach to merge fragments showing network of linkages over a range of LOD scores. Here, four homologues were successfully identified and merged directly.
**f.** Alternative representation of (e) showing these connections in a circular format instead.

## 4. Marker ordering

One of the challenges of marker ordering and map construction in autopolyploid species using marker dosages is the variable accuracy of recombination frequency estimates which must be integrated somehow. Ordering algorithms which only use unweighted recombination frequency estimates are unlikely to find an optimal map order, as there is no distinction between equal estimates of *r* from situations with vastly different information contents and variances. A thorough description of this issue is provided in Preedy and Hackett (2016). Within the polymapR package, marker ordering can be achieved in two ways – either using the weighted regression algorithm as originally developed by Piet Stam (Stam, 1993) and implemented in JoinMap (Van Ooijen, 2006) and now also in polymapR, or to use the multi-dimensional scaling algorithm as implemented in the MDSmap package (Preedy and Hackett, 2016). Given the computation efficiency of the MDS algorithm, in almost all circumstances this will be the preferred choice. Identical markers that were originally set aside can be added back to the final maps after marker ordering is complete.

# Implementation

## Software output – final linkage maps

The final output of the polymapR package is a phased integrated map. Maps can either be generated per homologue or per chromosome, facilitating the definition of haplotypes within a population. A record is kept in a log file of any markers that were removed at any stage during the procedure, as well as logging the function calls that generated each step, improving project reproducibility and later reporting. Visualisations are provided throughout the mapping procedure, facilitating the diagnosis of issues as well as summarising the results. An example of an integrated map with five chromosomes, generated using the sample data provided with the package, is shown in Figure 3.a. Phased linkage maps, giving the position of the SNP alleles on each parental homologue are also generated, as visualised in Figure 3.b for a triploid species. polymapR also generates input files for TetraOrigin (Zheng et al., 2016) which can calculate IBD probabilities for the population, useful for QTL analysis.

## Application of polymapR to real data

Various developmental versions of the polymapR package have been used for linkage map construction in potato, rose and chrysanthemum (Bourke et al., 2016;Vukosavljev et al., 2016;Bourke et al., 2017 ;van Geest et al., 2017a). The current version brings together all the capacities developed previously, while extending the algorithm to triploid populations as well (produced in a tetraploid x diploid cross). Cross-ploidy hybrids are

commonly produced in ornamental breeding, as well as in certain fruit species such as watermelon (*Citrullus lanatus* var. *lanatus*) or grape (*Vitis vinifera*) to generate seedless fruit (Acquaah, 2012). polymapR is applicable to a wide range of commercially-important crop species such as potato, leek, alfalfa, blueberry, chrysanthemum, sweet potato and kiwifruit, as well as the myriad of cross-ploidy populations developed in ornamental and fruit breeding programmes.

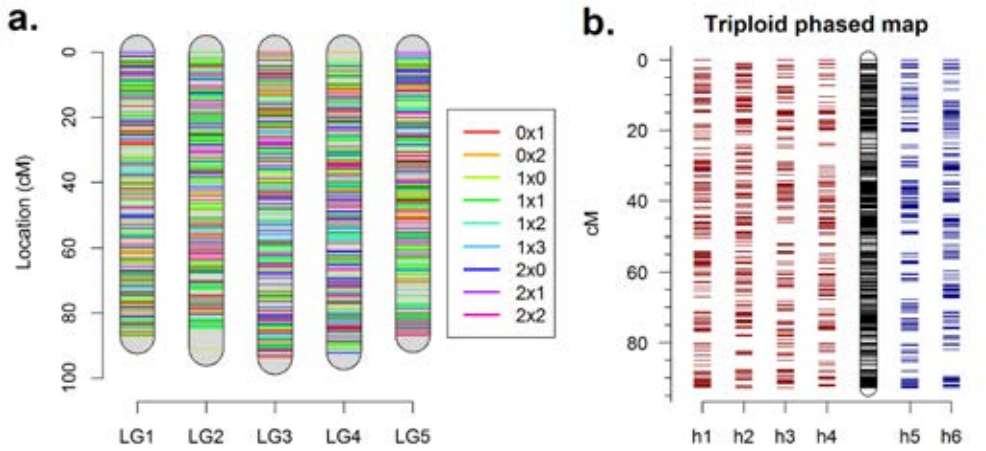



**Figure 3. Linkage map visualisations of polymapR. a.** Integrated chromosomal linkage maps generated using the sample tetraploid dataset provided with the package, with each marker segregation type highlighted. **b.** Phased homologue-specific maps of a single chromosomal linkage group from a triploid dataset (simulated with PedigreeSim (Voorrips and Maliepaard, 2012)). Maternal homologue maps (h1 – h4) from the tetraploid parent are shown on the left, and paternal homologue maps (h5 – h6) from the diploid parent are shown on the right, with the integrated chromosomal map in the middle.

# Discussion

## Comparison with other polyploid mapping software

The range of options for linkage mapping in autopolyploid species is quite limited. We compared the performance and applicability of polymapR with two alternative software, TetraploidSNPMap and PERGOLA.

### *TetraploidSNPMap (TSNPM)*

TSNPM possesses a graphical user interface for Windows, uses optimised routines for marker clustering and offers interactive cluster plots for linkage group assignment. It goes beyond linkage map construction to compute IBD probabilities and perform QTL interval mapping as well. Given that polymapR uses the same random bivalent pairing assumption and the same ordering algorithm (MDSmap (Preedy and Hackett, 2016)), we did not expect much difference in output. Using the sample tetraploid dataset provided

with polymapR (with 3000 markers over 5 chromosomes and 207 $F_1$ individuals, including 7 pairs of duplicate individuals), polymapR produced phased maps within 24 minutes on an Intel i7 desktop with 16 Gb RAM; TSNPM took 5 minutes, but took another 10 minutes to phase (so a total of 15 minutes were needed). However, the phased output of TSNPM is more difficult to interpret than that of polymapR and would likely require extra time for curation. The maps themselves were remarkably similar in terms of numbers of mapped markers, map length and marker order (Supplementary Figure 1).

Marker phasing in polymapR is automatic, by selecting phase based on the counts of significant linkages to 1x0 homologue clusters and ignoring any spurious linkages that go against the general trend. On the other hand, phase assignment seems to (generally) require manual intervention in the TSNPM pipeline. Despite its computational efficiency, TSNPM has also set an upper limit of 8000 SNP markers, and the maximum mapping population size is currently 300 $F_1$ individuals. polymapR sets no limits on marker numbers or population sizes, employing parallel processing to help speed up calculations for large datasets. Duplicated markers are initially binned (also possible in TSNPM) and identical individuals are merged (this feature was missing from TSNPM) to avoid needless calculations. Overall, the main difference between TSNPM and polymapR appears to be in applicability: polymapR can analyse autotriploid, autotetraploid, autohexaploid as well as segmental allotetraploid data, whereas TSNPM is currently confined to autotetraploid data. polymapR is also cross-platform given that it is written in R (R Core Team, 2016).

***PERGOLA***

The PERGOLA package in R has been developed for $F_2$ or backcross populations from an initial cross between homozygous parents. Such a situation is highly unusual for most polysomic polyploids, since inbreeding requires many more generations before homozygosity is reached compared to a diploid or disomic polyploid. In a polysomic hexaploid for example, it would take 25 generations of selfing an $F_1$ individual before 90% homozygosity is reached (ignoring the effects of double reduction (Haldane, 1930)). The applicability of the PERGOLA software to real populations in polysomic polyploids is therefore limited.

Despite the highly unusual type of population, we simulated a small $F_2$ dataset of selfed $F_1$ individuals randomly chosen from a cross between two inbred parental lines using PedigreeSim (Voorrips and Maliepaard, 2012), leading to a marker dataset of 500 duplex x duplex markers over 5 chromosomes. The calculation of recombination frequencies took a mere 3.54 seconds in PERGOLA, in comparison to 28 minutes using polymapR (on a single core; using 6 cores this step took 8 minutes). However, for polymapR this particular marker combination is complex, with nine possible phase combinations in the

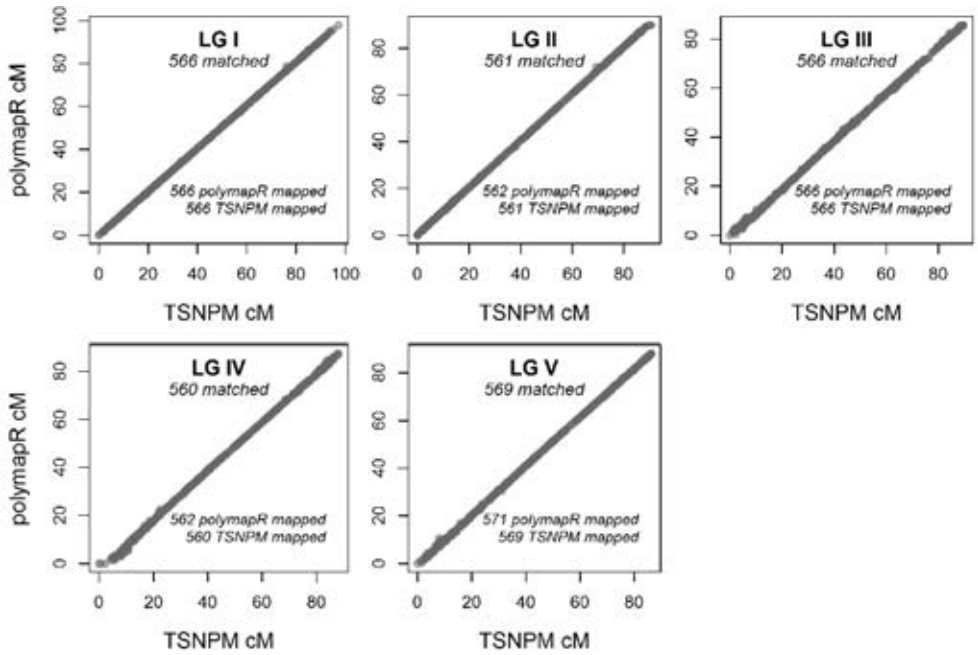
parents to be separately calculated per marker pair, and with extremely complicated likelihood functions for each phase (all 25 dosage combinations are possible in a tetraploid, from $n_{00}$ to $n_{44}$). It is therefore a somewhat unfair comparison, as PERGOLA labours under no such "generalist" difficulties. Phase considerations are trivial and therefore ignored by PERGOLA because of their simplistic population assumptions. If such populations could be generated, PERGOLA would produce excellent maps. In our test, PERGOLA identified all five chromosomes, with near perfect marker order in each, although the map lengths were inflated – from 200 cM using the Kosambi mapping function to 400 cM using Haldane's (when 100 cM was expected). polymapR also produced near-perfect maps with map-lengths of approximately 90 cM using Haldane's mapping function. The polymapR package can handle data from both cross-pollinating *and* inbred populations whereas PERGOLA cannot, but given the performance difference, PERGOLA would appear to be the software of choice for inbred polyploid populations, should they be developed.

## Concluding remarks

The development and release of polymapR comes at a time when there is increasing need for tools to perform genetic analysis in polyploids. Understanding the genetic control of important biological traits in polyploid species will have a large impact on plant breeding (or in the case of certain salmonid fish, animal breeding as well), facilitating the adoption of genomics-driven breeding decisions such as marker-assisted selection or genomic prediction into breeding programs. For these advances to take place, high-density and accurate maps showing the relative position of markers on chromosomal groups are needed – which is precisely what polymapR delivers.

# Acknowledgements

The authors wish to thank Dr. Katherine Preedy and Dr. Christine A. Hackett (Biomathematics Scotland, Dundee, Scotland) for providing a developmental version of the MDSMap scripts before their package became publically available, and to Dr. Johan van Ooijen (Kyazma B.V. Wageningen, The Netherlands) for helpful comments. This work was supported through the TKI projects "A genetic analysis pipeline for polyploid crops" (project number BO-26.03-002-001) and "Novel genetic and genomic tools for polyploid crops" (project number BO-26.03-009-004). Polyploid project partners are gratefully acknowledged for their feedback and contributions in the development of this software.




**Supplementary Figure 1.** Comparison of final linkage maps produced by polymapR (y-axis) and TetraploidSNPMap (x-axis) using the sample dataset provided with the polymapR package. For all linkage groups, the marker order, map length and the number of mapped markers were almost identical between both software.

# Chapter 7

# An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis

Geert van Geest[1,2,3], Peter M. Bourke[1], Roeland E. Voorrips[1], Agnieszka Marasek-Ciolakowska[4], Yanlin Liao[1], Aike Post[2], Uulke van Meeteren[3], Richard G.F. Visser[1], Chris Maliepaard[1], Paul Arens[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[2] Deliflor Chrysanten B.V., Korte Kruisweg 163, 2676 BS, Maasdijk, the Netherlands

[3] Horticulture & Product Physiology, Department of Plant Sciences, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

[4] Research Institute of Horticulture, Konstytucji 3 Maja 1/3, 96-100 Skierniewice, Poland

*Published (with modifications) as* Van Geest, G., Bourke, P.M., Voorrips, R.E., Marasek-Ciolakowska, A., Liao, Y., Post, A., *et al.* (2017), "An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis", **Theoretical and Applied Genetics 130, 2527-2541**

# Abstract

Construction and use of linkage maps is challenging in hexaploids with polysomic inheritance. Full map integration requires calculations of recombination frequency between markers with complex segregation types. In addition, detection of QTL in hexaploids requires information on all six alleles at one locus for each individual. We describe a method that we used to construct a fully integrated linkage map for chrysanthemum (*Chrysanthemum x morifolium*, 2n = 6x = 54). A bi-parental $F_1$ population of 406 individuals was genotyped with an 183,000 SNP genotyping array. The resulting linkage map consisted of 30,312 segregating SNP markers of all possible marker dosage types, representing nine chromosomal linkage groups and 107 out of 108 expected homologues. Synteny with lettuce (*Lactuca sativa*) showed local colinearity. Overall, it was high enough to number the chrysanthemum chromosomal linkage groups according to those in lettuce. We used the integrated and phased linkage map to reconstruct inheritance of parental haplotypes in the $F_1$ population. Estimated probabilities for the parental haplotypes were used for multi-allelic QTL analyses on four traits with different underlying genetic architectures. This resulted in the identification of major QTL that were affected by multiple alleles having a differential effect on the phenotype. The presented linkage map sets a standard for future genetic mapping analyses in chrysanthemum and closely related species. Moreover, the described methods are a major step forward for linkage mapping and QTL analysis in hexaploids.



# Keywords

autohexaploid, marker dosage, identity-by-descent, haplotype, polysomic polyploid

# Introduction

A linkage map is a starting point for localisation of genomic regions that are associated with agriculturally important traits. This makes it an important tool for DNA-informed breeding (Peace, 2017). For polyploids, DNA-informed breeding has lagged behind compared to diploids, because genotyping co-dominant markers and linkage map construction in polyploids requires specialised methods. Such methods need to be able to handle higher-dose markers. As opposed to diploids, polyploids have multiple conformations of heterozygous genotypes; on a locus with two alleles, a hexaploid can have five different heterozygous genotypes ranging from a dosage of one to a dosage of five. Together with the two homozygous conformations, this adds up to seven different dosage scores.

Many linkage maps of polyploids are constructed with single-dose (present/absent) markers using methods developed for diploids. Those kinds of maps are limited to representing only individual homologues. Integration of separate maps of homologous chromosomes is needed for transferability of results between mapping studies and mapping of traits with a complex genetic architecture. In an integrated map, all markers are located relative to each other, resulting in one representation of the positions of all mapped loci, irrespective of the phase of their alleles. This enables comparisons of linkage maps based on different populations. Map integration requires estimation of linkage between single-dose markers in repulsion or linkage between higher dose markers. Estimation of linkage of markers in repulsion is different from diploids and can only be done with very low confidence, especially in a hexaploid (Wu et al., 1992). Segregation ratios of higher dose markers are fairly complex, and calculation of recombination frequency needs specific statistical methods (Hackett et al., 1998). Because of the complicated nature of recombination frequency estimation between higher dose markers, dedicated software is required.

In an outcrossing species, the number of alleles that can affect a trait in a single individual is the same as the ploidy level (Figure 1.a). For QTL detection in an outcrossing full-sib population without any prior knowledge on the involved alleles, all twelve possible alleles that can be inherited from the parents should therefore be taken into account (Figure 1.b). With use of the positions of markers on a non-integrated linkage map of homologues, only information on the presence or absence of one out of twelve parental alleles is available (Figure 1.c). If the other eleven alleles are ignored, any QTL that does

not have underlying alleles with major effect on the trait will be missed. Multi-allelic QTL mapping therefore needs information of all alleles per locus.

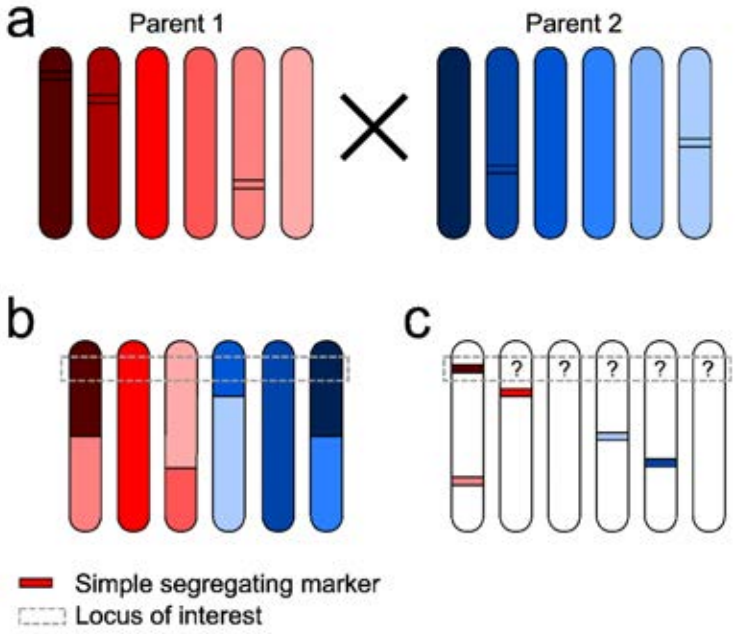



**Figure 1. Marker allele considerations in an autohexaploid. a.** Example of a cross of two parents with each six different alleles. **b.** An $F_1$ progeny is composed of two gametes with recombinations from both parents. **c.** By considering a single 1:1 segregating marker, we can only characterise absence or presence of one allele in the $F_1$ offspring (coloured band indicates observed presence of a marker allele). However, the actual situation is much more complex. On that same locus there are five other alleles that all could have specific effects on the phenotype. Genotype probability models try to reconstruct the actual situation by estimating absence and presence for each allele per locus.

Most progress in linkage mapping in an outcrossing hexaploid with polysomic inheritance has been reported in sweet potato (*Ipomoea batatas*). In this species, several non-integrated maps have been published (Ukoskit and Thompson, 1997;Kriegner et al., 2003;Cervantes-Flores et al., 2008;Chang et al., 2009;Monden et al., 2015;Shirasawa et al., 2017), with three publications reporting information on homologous chromosomes without actually integrating the maps. In two publications, this information was based on markers that had a dosage of two (duplex) in one parent and zero (nulliplex) in the other (2x0 markers), and markers with a dosage three (triplex) in one parent and zero in the other (3x0 markers;(Ukoskit and Thompson, 1997;Cervantes-Flores et al., 2008)). Others have identified homologous chromosomes based on alignment to a reference genome (Shirasawa et al., 2017). Similar to sweet potato, chrysanthemum is an outcrossing hexaploid with polysomic inheritance (van Geest et al., 2017c). Pairing at meiosis is primarily through bivalents, but multivalents do occur (Roxas et al.,

1995;Chen et al., 2009). Reported methods for linkage map construction (Zhang et al., 2010a;Zhang et al., 2011b) and QTL analysis (Zhang et al., 2012a;Zhang et al., 2012b;Zhang et al., 2013a) for chrysanthemum have been limited to methods developed for diploids, and constructed maps are therefore not integrated.

In a hexaploid, an integrated linkage map is most preferably constructed by estimation of linkage with higher dose markers. Those multi-dose markers can connect homologous chromosomes within parents and between parents and can therefore be used to integrate them. For tetraploids, methods to estimate linkage between higher dose dominant markers have been developed (Hackett et al., 1998), and applied to construct integrated linkage maps (Meyer et al., 1998;Luo et al., 2001;McCallum et al., 2016). Later, these methods have been extended and applied to bi-allelic SNP markers (Hackett et al., 2013;Bourke et al., 2016;Bourke et al., 2017). Such methods would need to be extended to hexaploids in order to generate integrated linkage maps with use of higher dose markers.

An integrated linkage map can be used to reconstruct inheritance of parental haplotypes to approach a representation as in the example in Figure 1.b. The two alleles of bi-allelic SNPs can be in linkage disequilibrium with multiple haplotypes, each having a different effect on the phenotype. Such haplotypes can be identified based on the configuration of neighbouring alleles. Methods for reconstruction of haplotype inheritance by estimating probabilities of identity-by-descent (IBD) in tetraploid bi-parental populations have been developed (Hackett et al., 2013;Bourke, 2014;Zheng et al., 2016). Although all methods are theoretically extendible to hexaploids, the method developed by (Bourke, 2014) is ploidy-level independent and is therefore directly applicable to hexaploids.

In this paper, we describe the construction of an integrated linkage map from all possible marker dosage types in hexaploid chrysanthemum. We are setting the standard for transferability of results by chromosomal linkage group numbering based on synteny with lettuce (*Lactuca sativa*) and by generating a core set of SNP markers that can be used to anchor future maps. With the integrated linkage map, we reconstruct haplotypes based on parental origins using a relatively simple procedure. We demonstrate the usefulness for QTL mapping for four traits for which information of all twelve segregating alleles was taken into account.

# Materials and methods

## Plant material and phenotyping

We analysed the segregation of SNP markers in a bi-parental population that consisted of 406 individuals originating from a cross between DB36451 (P1) and DB39287 (P2), two daisy-type, white chrysanthemum cultivars. Phenotyping took place in the same experiment as described by (van Geest et al., 2017b). In short, the offspring and parents were grown in three randomised blocks in each of three seasons: summer (May to July 2015), late summer (August to October 2015), and autumn (September to November 2015). A replicate consisted of a field containing 10 to 50 plants. Plants were grown in 12, 12 and 14 days of 18, 21 and 21 h photoperiods for the summer, late summer and autumn, respectively. To induce flowering, they were subsequently grown in 12 h photoperiods for the plants grown in summer and late summer and in 11 h photoperiods for the plants grown in autumn. Flower colour was recorded based on visual observation. If flowers were completely white, they were scored as 0, if they were slightly pink as a 1, and pink flowers were scored as 2. Flowering time was recorded as the number of days (at short photoperiod) needed to reach commercial maturity for at least 50% of the plants grown in a single field. The number of ray florets was counted from the third flower head from the top of one flower stem for each replicate. The phenotypic scores obtained for disk floret degreening are described by (van Geest et al., 2017b). Heritability was calculated by dividing the estimated genotypic variance by phenotypic variance. Variances were estimated using an analysis of variance (ANOVA) with trial and genotype as fixed effects.

## Mitotic chromosome counting

For mitotic metaphase chromosome analysis, ± 1 cm long roots were collected from DB36451 (P1) and DB39287 (P2) and incubated in eppendorf tubes in ice water for 24 hr and then fixed in ethanol–acetic acid (3:1) solution for 12–24 hr. Roots were stored in fixative at -20°C until use. For chromosome preparations, the root tips were washed 4 times 5 min in enzyme buffer (0.01 M citric acid-sodium citrate, pH 4.8) and incubated in an enzyme mixture containing 1% (w/v) pectolyase Y23, 1 % (w/v) cellulase RS at 37°C for about 1.5 h. Squash preparations were made in a drop of 45% acetic acid and frozen in liquid nitrogen. The cover slips were removed by using a razor blade. The slides were then dehydrated in absolute ethanol, air dried and stained with 1 µg/ml 4,6-diamidino-2-phenylindole (DAPI, Sigma) in Vectashield (Vector Laboratories). Images of fluorescently stained chromosomes were acquired using a Canon digital camera attached to an Axiophot microscope with an appropriate filter and then processed using

software (Axio Vision 4.2). For each genotype, the total number of chromosomes was determined for 5-10 metaphases.

## Genotyping and marker quality filtering

Genotyping was performed with a 183k Affymetrix SNP array, as described by (van Geest et al., 2017c). In short, the array was designed based on RNA-seq data of 13 cultivars, including both parents of the population. A reference transcriptome was assembled based on the reads originating from DB36451 (the female parent of the population), and reads of all 13 other cultivars were aligned against this assembly. From these alignment files SNPs were called, while retaining information from which transcript contig they originated.

Dosage scoring from array output was mainly performed as described by (van Geest et al., 2017c). Because genomic dosage was highly correlated with the number of reads per allele from our sequence data (van Geest et al., 2017c), we estimated dosage per SNP of the parents a-priori based on the sequence data, and used this information for SNP calling. This resulted in 67,916 SNP markers with expected segregation in the population based on parental dosages. Similar to the description provided by (van Geest et al., 2017c) we removed non-segregating markers, markers with >5% missing values, and skewed markers ($p < 0.001$ based on a $\chi^2$-test assuming polysomic inheritance). Of the individuals we removed selfings and indviduals with >10% missing values, resulting in 400 out of 406 individuals. We grouped identical markers together if markers had identical non-missing dosage scores for each individual. These groups of non-unique markers were represented by a single marker from that group that had the least missing values. This representative marker was used in further mapping steps with all other unique markers. After ordering, the other markers in the represented group were assigned to the same position as the representative marker.

## Linkage map construction

To calculate recombination frequency ($r$) and LOD scores of marker pairs, the method as described by (Bourke et al., 2016) was modified for hexaploids, *i.e.* using the assumption of completely random bivalent pairing. Initially, marker dosages were converted to their most fundamental form as previously described (Bourke et al., 2016), resulting in nineteen separate marker segregation types. For all possible marker combinations, functions for pairwise estimation of $r$ were then derived. In a hexaploid species, fifteen bivalent pairing scenarios are possible, in comparison to three for a tetraploid. For each combination of marker types, there are multiple phases possible depending on the conformation of the markers within one or both parents. All possible phase combinations were calculated for each marker pair (*i.e.* all phases having a distinct

likelihood function), since the phasing of marker pairs is unknown before mapping. The recombination frequency (and associated LOD) was selected among those estimates of $r$ in the range $0 \leq r < 0.5$ which maximised the log of the likelihood function (Hackett et al., 2013). The accuracy of recombination frequency estimation and phase assignment was checked using a small simulated hexaploid dataset generated in PedigreeSim (Voorrips and Maliepaard, 2012), which showed a high degree of concordance between the true and expected results for most marker combinations. In cases where the accuracy of the estimate was lower, the LOD score reflected this (being loosely related to the inverse of the variance of the estimate). Overall, for each marker type combination a total of 104 linkage functions were derived in Mathematica 10.0 (Wolfram Research Inc., 2014) and converted to R language (R Core Team, 2016) for the linkage analysis.

To construct backbone clusters that would represent homologues, simplex x nulliplex (1x0) markers were clustered at a LOD score of 10. To identify chromosomal linkage groups (CLG), multi-dose markers can be used to provide bridge linkages between pairs of 1x0 markers, therefore associating clusters into CLG. Abundant multi-dose markers provide the most information, among those are uniparental duplex x nulliplex (2x0) markers (Bourke et al., 2017) or bi-parental markers, like simplex x simplex (1x1) markers. In our case, the use of 1x1 markers showed the clearest associations between 1x0 clusters, and these marker types were therefore used to identify CLG. Markers in clusters smaller than five markers were not used in further mapping steps. Linkage information of the bi-parental 1x1 markers were used to assign consensus numbering to the linkage groups between parents. After construction of this backbone clustering, all other marker types were assigned and phased to a CLG and homologue based on linkages with 1x0 markers with a LOD score greater than five. To complete information on all pairwise linkages, for each marker combination within a linkage group, recombination frequency and LOD were calculated with the derived functions. The markers were ordered using MDSMap (Preedy and Hackett, 2016), with parameter settings as suggested by the authors: we used Haldane's mapping function, two dimensions for the principal curves, and $LOD^2$ as weights. We did not observe any notable change on the map ordering between two and three principal curve dimensions, and we therefore chose to use the simplest setting of two dimensions. After the first round, problematic markers were removed based on visual inspection of the principal curves and the difference in distance between nearest neighbouring markers as estimated from recombination frequency and the distance on the map. This difference is represented by the nearest neighbour fit (Preedy and Hackett, 2016) and markers exceeding a value of four were considered problematic and thus removed. This was repeated if the next round resulted in a reduction of the total nearest neighbour fit. From the integrated map, all marker alleles were assigned to a homologue. This assignment was based on coupling linkages

with 1x0 markers that formed the backbone clustering. If there were at least five coupling linkages with 1x0 markers at LOD greater than 5, alleles were assigned to a homologue. If the number of marker alleles was not equal to the number of assigned homologues, the marker was not included in the phased map.

As SNP markers were discovered from an RNA-seq derived transcriptome assembly, each marker is associated with a transcript contig sequence. This information was used to investigate the quality of the map. Markers of the type 1x0 that originated from the same transcript contig should have a distance approaching 0 cM on the integrated map (assuming the contig was assembled correctly). For each homologue combination and for each linkage group, an overall deviation was quantified by calculating the root mean square error (RMSE) of these differences on the integrated map for all mapped 1x0 markers originating from the same transcript contig.

To enable alignment of any future linkage maps in chrysanthemum by gene sequences, markers were identified that originated from contigs representing characterised genes. For this, mapped markers were aligned to all proteins from the UniProt database from *Chrysanthemum x morifolium* (taxonomy ID 41568) using BLASTX (Altschul et al., 1997). Hits were filtered for alignment lengths greater than 100 and more than 95% identity. A subset of markers originating from these filtered transcript contigs spread over all linkage groups was selected to form a reference linkage map.

## Synteny with lettuce

To investigate the synteny of the integrated linkage map with lettuce (*Lactuca sativa*), mapped transcript contigs were aligned to the mapped unigenes of lettuce as available from the Lettuce SFP Chip Project website (Truco et al., 2013) using BLAST (Altschul et al., 1997). Unique hits with an e-value smaller than 1E-100 were used to assess synteny. Chrysanthemum CLG were renumbered based on the number of alignment hits with the lettuce linkage groups.

## IBD probabilities

In order to estimate presence of parental haplotypes in the offspring, we calculated IBD probabilities as described by (Bourke, 2014). A schematic overview of the method is shown in Figure 2. Based on the phased map and linkage information, IBD probabilities per marker locus were calculated for each member of the $F_1$ population in two steps. The information was stored in a three-dimensional array for each chromosomal linkage group, with marker, offspring individuals and homologue on the x, y, and z dimensions. In the first step, only fully informative dosage scores were used to fill the IBD probability array. This means that if a progeny has inherited all alleles of a marker on specific homologues, this progeny will be assigned an IBD probability of 1 for these homologues

at that marker locus. If none of the alleles are inherited, the IBD probabilities for these homologues at the locus would be 0. In general, any scores in the progeny that were larger than zero and smaller than the sum of the parental dosage scores were considered as non-informative (*e.g.* for a 1x1 marker, progeny with a dosage of 0 or a dosage of 2 were considered informative, and with a dosage of 1 non-informative). Probabilities of loci at homologues of progeny that had non-informative marker scores were given a starting probability of 0.5. In the second step, inter-marker distance was used to estimate the IBD probabilities of homologue loci with non-informative marker scores. For each marker locus at each homologue in each progeny the closest informative marker was located, and IBD-probabilities were calculated based on this closest informative marker, where:

$$P_i = 1 - r_{ij}$$

if $P_j = 1$, and

$$P_i = r_{ij}$$

if $P_j = 0$.

Here, $P$ represents the IBD probability, $i$ indicates a marker without full IBD information and $j$ indicates a fully informative marker, and $r$ represents recombination frequency between an informative and non-informative marker as calculated using Haldane's mapping function from estimated distance on the integrated map. After assignment of IBD probabilities, the sum of IBD probabilities per parent was normalised to three (as there are three homologous chromosomes in a gamete). For each homologue in each $F_1$ individual, a cubic spline was fitted over IBD probabilities versus position to calculate IBD probability interpolations over 1 cM intervals. Genotype information content (GIC) for interval $k$ at homologue $h$ for $n$ individuals was calculated with the following formula:

$$GIC_{hk} = 1 - \frac{2}{n}\sum_{i=1}^{n} | P_i - \lfloor P_i \rceil |$$

where

$$\lfloor P_i \rceil = 0, \quad 0 \leq P_i \leq 0.5,$$
$$\lfloor P_i \rceil = 1, \quad 0.5 < P_i \leq 1$$

This results in a score for GIC ranging from 0 to 1, where 0 represents a locus with little information, and 1 with complete information.

# QTL mapping

QTL analysis was performed on block-corrected mean phenotypic values using an IBD probability model, as described before for tetraploids (Bourke, 2014). An additive model modified from (Kempthorne, 1957) as suggested by (Hackett et al., 2013;Hackett et al., 2014) was modified to the hexaploid level:

$$Y = \mu + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_8 X_8 + \alpha_9 X_9 + \alpha_{10} X_{10} + \alpha_{11} X_{11} + \alpha_{12} X_{12}$$

where $\alpha_i$ and $X_i$ are the main effects and indicator variables for allele *i*, respectively. The parameters representing homologue 1 and homologue 7 were taken as the reference classes and were therefore omitted from the model as in all cases three alleles are inherited per parent. To calculate the significance threshold for detecting significant QTL, a thousand permutations were run with randomly permuted phenotypes (Churchill and Doerge, 1994), taking the 5$^{th}$ percentile of the (ordered) minimum p-values from each genome-scan analysis as an approximate significance threshold. To identify homologues affecting the trait, a simple linear model was run for each of the twelve alleles separately.

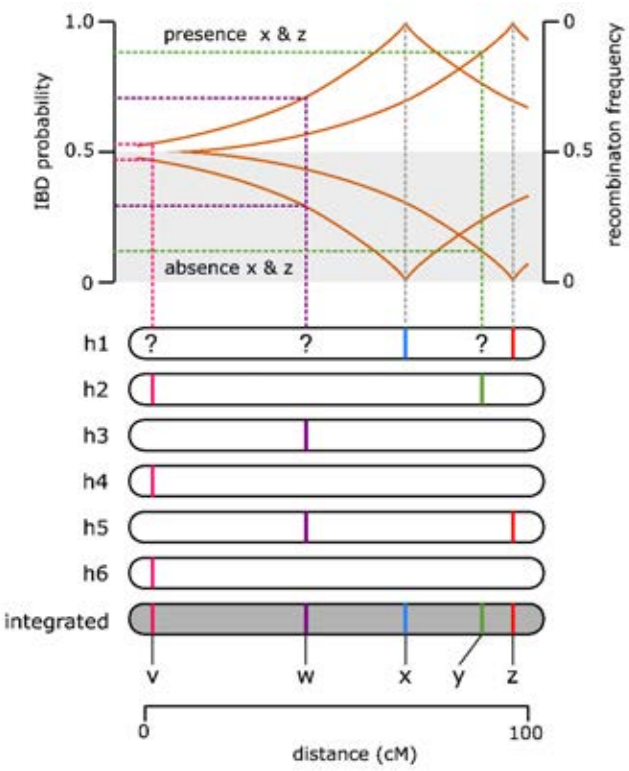



**Figure 2. Visualisation of the estimation of IBD-probabilities.** A hypothetical integrated linkage map and the separate linkage maps of the six homologues of one parent are shown in dark grey and white respectively. In the upper panel of the line graph (IBD probability > 0.5), the calculation of IBD-probabilities for homologue 1 (h1) are shown for marker loci *v* (pink; triplex), *w* (purple; duplex) and *y* (green; simplex) in a situation in which all alleles of marker *x* (blue; simplex) and *z* (red; duplex) are inherited. Since all alleles of loci *x* and *z* are inherited, these loci get an IBD-probability of 1 for inheritance of homologue 1. For marker loci *v*, *w* and *y* none of the marker alleles are present on homologue 1. It is therefore not known whether h1 is inherited at these loci. The orange lines depict the relationship between genetic distance and recombination frequency (*r*), as a function of map distance. Because distance between all marker combinations is known from the integrated map, we estimate the IBD-probabilities of loci *v*, *w* and *y* as 1-*r* (in case of inheritance of all alleles of *x* and *z*), where *r* is the recombination frequency between the locus and the closest informative marker (marker *x* in the case of *w* and *v*, and *z* in the case of *y*). The lower panel of the line graph (shaded in grey; IBD probability < 0.5) depicts the situation where none of the alleles of loci *x* and *z* are inherited. Here, IBD-probabilities for *v*, *w* and *y* are estimated as *r*

# Results

## Linkage map

After removal of markers that were non-segregating, had distorted segregation or had more than 5% missing values, 30,532 markers remained in the dataset. Of those, 21,345 had unique dosage scores across the progeny (Figure 3). Because markers with identical dosage scores in each individual (non-unique markers) will map to the exact same position, they were reduced to a single unique marker for calculation of linkage and ordering. The others were added to the linkage map after map construction with only unique markers.

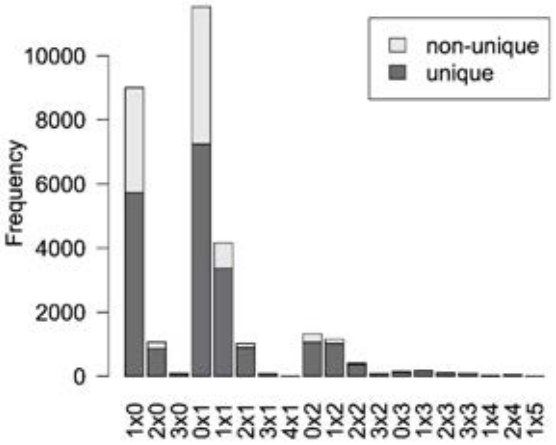



**Figure 3. Distribution of 19 different marker types segregating in the bi-parental population.** Total number of markers: 30,532, of which 21,345 were unique. The non-unique markers had duplicate dosage scores across the population. The labels on the x-axis represent marker segregation types such as simples x nulliplex (1x0) etc. ("dosage parent 1" x "dosage parent 2")

Simplex x nulliplex and nulliplex x simplex (1x0 and 0x1) markers were used to construct backbone clusters that represent homologues. This resulted in 54 clusters for P1 and 53 clusters for P2 each containing five or more markers. Chromosome counting showed that both parents had 2n=54 chromosomes, the expected euploid chromosome number. Our dataset was therefore lacking 0x1 markers identifying one out of the 54 homologous chromosomes of P2. Identification of CLG (chromosomal linkage groups) with simplex x simplex (1x1) markers resulted in a network of nine CLG representing all homologue clusters of both parents. All other marker types were subsequently assigned to a CLG based on linkage with 1x0 and 0x1 markers. In total, 21,159 unique markers (99.1%) could be assigned. Markers were ordered per CLG based on recombination frequency with $LOD^2$ as weights, resulting in CLG map lengths ranging

from 64.5 to 95.0 cM. After ordering, the groups of non-unique markers were added to the linkage map based on the position of their unique representing marker, resulting in a linkage map containing 30,312 markers (99.3% of initial; Table 1). Of the ordered markers, the alleles of 28,638 (93.8% of initial) could be phased to an expected number of homologues based on parental dosages with at least five significant linkages to 1x0 markers (Table 1), resulting in a fully phased linkage map.

**Table 1. Summary statistics of integrated linkage map**

| CLG [a] | Length (cM) | N [b] | Phased markers | Contigs[a] | Rounds[b] |
|---|---|---|---|---|---|
| **1** | 82 | 2,595 | 2,528 | 1,199 | 2 |
| **2** | 77.3 | 3,184 | 3,110 | 1,411 | 3 |
| **3** | 64.5 | 2,970 | 2,786 | 1,269 | 3 |
| **4** | 84.2 | 3,601 | 3,215 | 1,557 | 3 |
| **5** | 90.3 | 3,498 | 3,427 | 1,508 | 3 |
| **6** | 91.1 | 3,619 | 3,533 | 1,585 | 2 |
| **7** | 81.6 | 3,936 | 3,464 | 1,621 | 2 |
| **8** | 95 | 3,805 | 3,499 | 1,604 | 2 |
| **9** | 86.1 | 3,104 | 3,076 | 1,338 | 3 |
| **sum** | 752.1 | 30,312 | 28,638 | 13,092 | - |
| **mean** | 83.6 | 3,368 | 3,182 | 1,455 | - |

[a] CLG, chromosomal linkage group
[b] N, number of mapped markers
[c] Number of transcript contigs associated with mapped markers.
[b] Number of rounds of problematic marker removal and re-ordering after the first ordering.

## Synteny with lettuce and reference map

We aligned mapped transcript contigs of chrysanthemum with mapped lettuce unigenes (Figure 4). We aligned the 13,092 mapped chrysanthemum transcript contigs to 12,841 mapped lettuce unigenes, and obtained 4,757 unique hits with an e-value smaller than 1E-100. This resulted in the identification of syntenic linkage groups between lettuce and chrysanthemum. All combinations of linkage groups of chrysanthemum and lettuce with maximum number of hits were unique, except for CLG9 and CLG4. These two CLG had both most hits with lettuce LG4. The chrysanthemum CLG9, with least hits to lettuce LG4 was renumbered based on LG9, the non-assigned linkage group from lettuce. This combination still had 126 hits, indicating partial similarity. Syntenic analysis per LG resulted in identification of large regions with linear correspondence between locations of genes, so the genomes appear to be partly co-linear at local scale. This was not clear at a larger scale, as syntenic regions were scattered across the linkage map of lettuce, which can be

interpreted as that each chromosome carries major inversions and translocations. With use of this data, we based the numbering of chrysanthemum CLG on the number of significant alignments of mapped transcripts. To mark these nine chrysanthemum linkage groups, we present 92 CLG-defining SNP markers. These are evenly spread over all nine CLG and originate from 85 contigs representing genes coding for protein entries of the UniProt database (Figure 5). This should be a useful tool for future studies in chrysanthemum.

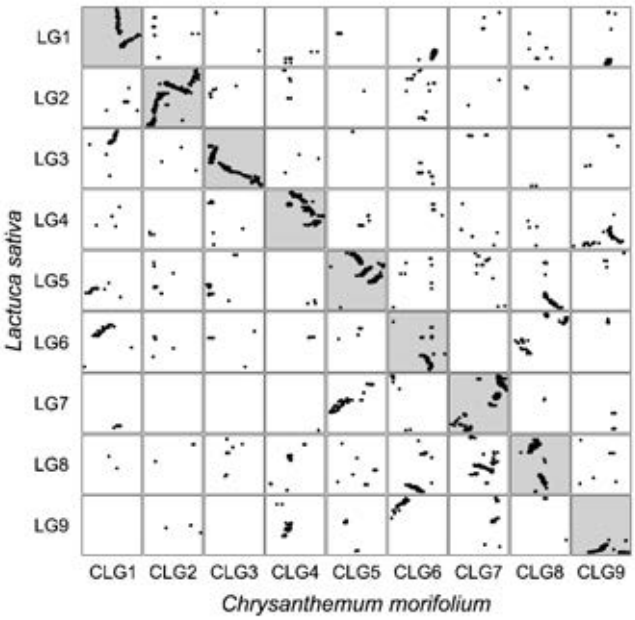



**Figure 4. Synteny between the lettuce ultra-high density map (Truco et al. 2013) and chrysanthemum.** Each dot represents a significant alignment between lettuce unigenes and chrysanthemum transcript contigs

# Linkage map quality

We used two analyses to evaluate the quality of the linkage map. First, to investigate the concordance between estimated pairwise *r* ($r_{pairwise}$) and *r* based on map distance ($r_{map}$), these two estimators of *r* were plotted against each other. With high LOD scores, these two estimators were in concordance with each other over a wide range of *r* (from 0 to 0.3). Second, to evaluate the position of nearby 1x0 markers in coupling and repulsion, we aligned the position on the integrated map of 1x0 markers that originated from the same transcript contig from the RNA-seq assembly. The positions of markers that originated from the same contig and had the same phase (6,937 markers in total) aligned nearly perfectly (Figure 6), indicating low error-rates. The position of markers phased

on different homologues (8,352 markers in total) was more spread. The residual mean squared error (RMSE) was calculated for each linkage group (Figure 6) and each combination of homologues from the same linkage group. RMSE was generally below 5 cM, with some outliers, on CLG 2, 5 and 8. These outliers were caused by one, two, and two markers respectively.

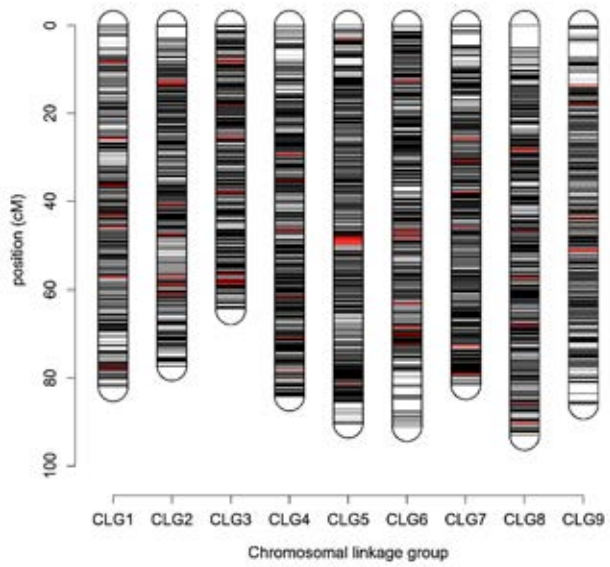



**Figure 5. Integrated linkage map of phased markers with 1x0 markers (black), other marker types (grey) and CLG defining markers (red)**

## IBD probabilities

The presence of each of the twelve segregating haplotypes per locus was estimated in all progeny individuals at 1 cM map intervals, which was expressed in IBD probabilities. In the middle of the CLG, the IBD probabilities could be estimated with high confidence. If there were no markers in large parts of one homologue, the IBD probabilities could still be close to 0 or 1, because information from the five other homologues can complement the missing information. Even if no markers were mapped on the entire homologue, e.g. homologue 12 from P2 on linkage group 4, IBD probabilities were complemented with information from the other five homologues. Genotype information content was lower towards telomeres, because in those regions markers were often missing in a large range in at least two homologues and informative markers were present on only one side.

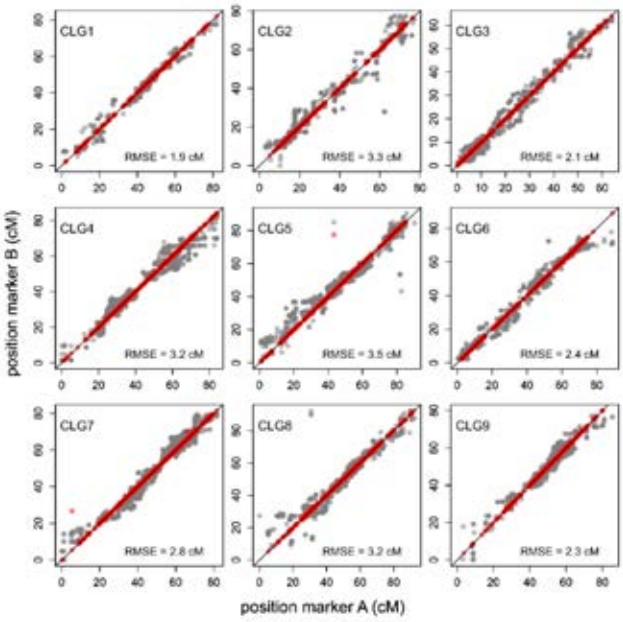



**Figure 6. Scatterplot of marker positions of 1x0 markers on the integrated map that originated from the same transcript contig**. Each dot represents a combination of markers that originated from the same transcript contig. The red dots indicate markers phased on the same homologue, grey dots on different homologues. The black line represents $y = x$

# QTL mapping

The population was phenotyped for four different traits: flower colour, flowering time, disk floret degreening and number of ray florets. All four traits had a moderately high heritability ranging from 0.68 to 0.72. The phenotypes were fitted against the IBD probabilities at 1 cM intervals with a main effects model.

Two regions were highly-significantly associated with flower colour, at CLG5 and 7, and one region at CLG9 was slightly associated (Figure 7). The highly significant loci were both simplex QTL (Figure 8.a). Analysis of variance of the interaction between the associated alleles showed a highly significant ($p < 1E\text{-}16$) interaction, indicating that both alleles need to be present to get a pink flower colour. Together, the two 1x0 markers that were most closely linked to each of the QTL explained 47.8% of the variation, indicating that the trait is mainly inherited by two major alleles segregating from two loci from each of the two parents. There is a minor QTL on CLG9, but some genotypic variation is still to be explained by undetected QTL.

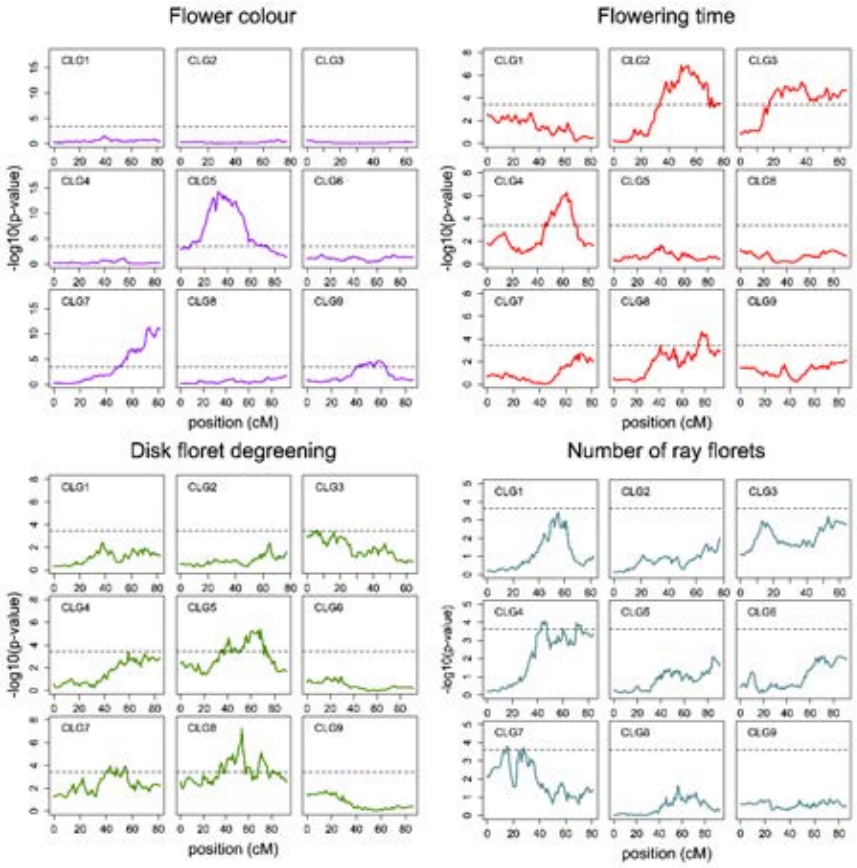



**Figure 7. QTL analysis of flower colour (purple), flowering time (red), disk floret degreening (green) and number of ray florets (blue).** Significance thresholds based on 1000 permutations are indicated by a dashed line

For flowering time, we found three clear QTLs on CLG2, 3 and 4 and one minor QTL at CLG8 (Figure 7). For the simplex QTL on CLG4, presence or absence of the allele at homologue 11 had a major effect on the trait. In both loci on CLG2 and 4 presence of different alleles could have a positive effect, a negative effect or no significant effect on the phenotype (Figure 8.b). Therefore, at least three alleles underlie the QTLs. F or disk floret degreening, three QTL located on CLG5, 7 and 8 were detected (Figure 7). For all three QTL, multiple alleles played a role (Figure 8.c). The QTL on CLG8 explained most phenotypic variation, and presence of the allele on homologue 5 had the strongest effect on the mean value of disk floret degreening. For number of ray florets, two minor QTL were found on CLG4 and 7. The QTL on CLG7 was affected by one allele from the maternal parent. The QTL on CLG4 was affected by alleles that originated from only the maternal parent with opposite effects from different homologues (Figure 8.d).

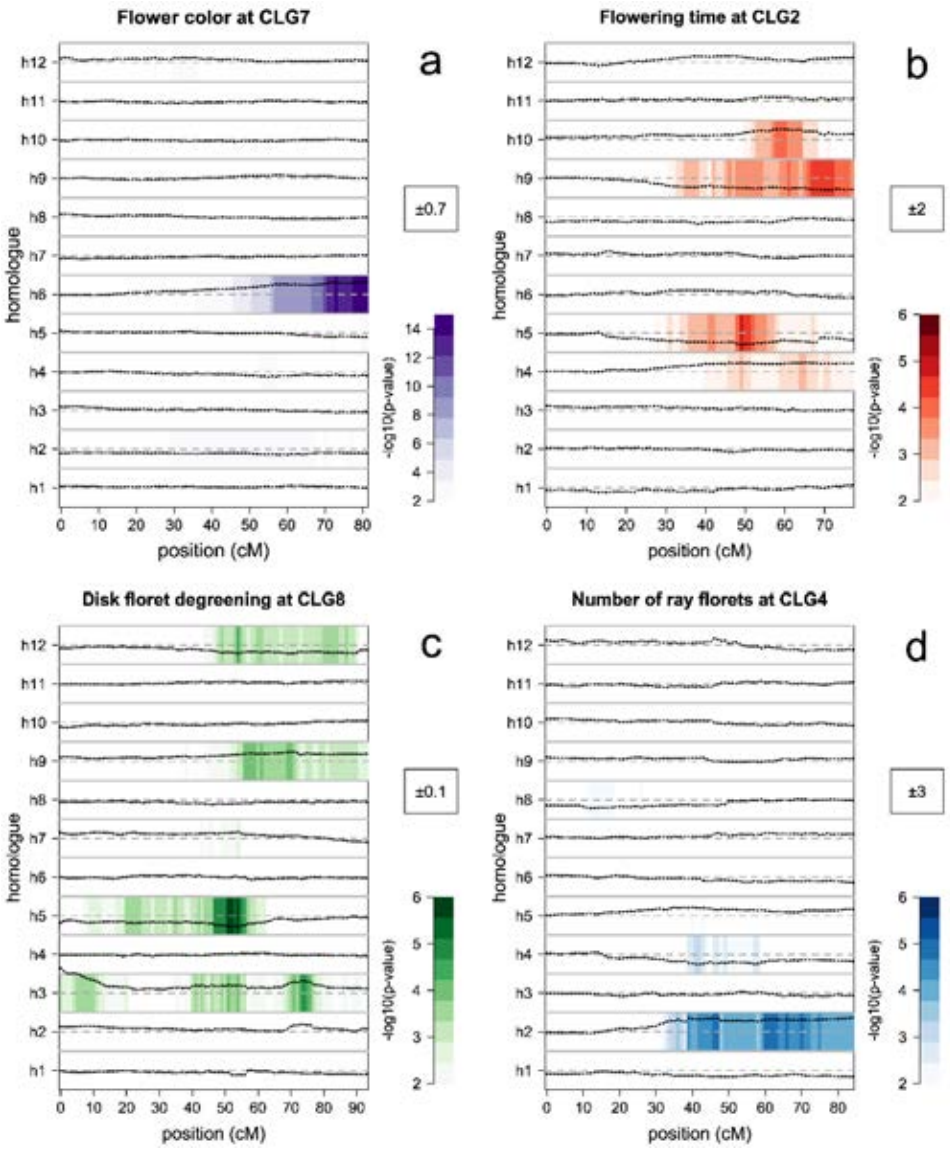



**Figure 8. Analysis per homologue for four QTL**. Traits studied were **a.** flower colour at CLG7; **b.** flowering time at CLG2; **c.** disk floret degreening at CLG8; **d.** number of ray florets at CLG4. The p-value for testing the significance of the explained variation of IBD probabilities of a single allele versus phenotype ($Y_G = \mu_C + \alpha_i X_i$) is shown as a heatmap. The estimated effect of full absence or presence of an allele on the phenotypic value is shown in the *black points*. The plot limits of the effect is shown in the *black box*, meaning that the *black points* can range within the negative and positive value between the boundaries between homologous indicated by *grey lines*. The *dashed grey lines* represent an effect of zero

# Discussion

In this paper, we report the first integrated linkage map in a hexaploid species with polysomic inheritance. We were able to assign multi-dose markers to their parental homologues. With this phasing information, we could reconstruct inheritance of haplotype alleles in the bi-parental population and perform QTL analyses. We provide major steps to overcome a number of limitations to linkage map construction based on SNP markers in hexaploids, including full map integration and phasing.

## An integrated and phased linkage map

The ultra-dense integrated map contained markers of all 19 possible types in a hexaploid. With our approach, we first defined backbone marker clusters that represented homologues based on linkages between simplex x nulliplex (1x0) markers. With 1x1 markers that contained information about homologous chromosomes, we created networks of 1x0 linkage groups that represented CLG (chromosomal linkage groups). Subsequently the other marker types were assigned to these backbone clusters. Because we first defined backbone marker clusters based on 1x0 markers, definition of homologues relied on presence of 1x0 markers. Lack of 1x0 markers on a homologue is possible if there is a high degree of inbreeding, the population is from a selfing, or if there is selection for a phenotype for which alleles have an additive effect. However, our experience in polyploid mapping to date has shown that low-dosage markers tend to be the most abundant marker type (Bourke et al., 2016;Bourke et al., 2017;van Geest et al., 2017c). Nevertheless, we were not able to define one homologue on linkage group 4, even though all others each had 175 SNP markers or more. Too few markers therefore seems unlikely. A reason could be that some combinations with alleles on this homologue might have been lethal, and 1x0 markers on this homologue might therefore have had highly distorted segregation. These markers would have been filtered out prior to linkage mapping.

The MDSMap algorithm (Preedy and Hackett, 2016) has proved particularly useful for the weighted ordering of our large number of diverse marker types. As the LOD for linkage varies for the same values of *r* for different marker type combinations and phases, a weighted ordering algorithm was required. The most frequently used algorithm for weighted ordering is based on a weighted linear regression (WLR) algorithm as deployed in JoinMap (Van Ooijen, 2006). However, current processor speed of desktop computers limits the use of this algorithm to approximately a hundred markers per linkage group. Because of this restriction, (Bourke et al., 2016) used the WLR algorithm to construct homologue maps separately, which were later integrated. In a subsequent mapping study in tetraploid rose, the MDSMap algorithm was used to construct an integrated map containing over 25k SNPs, without the need for binning or the separation of homologue

maps before integration (Bourke et al., 2017). The MDSMap algorithm also forms the core of the map-ordering module within TetraploidSNPMap software (Hackett et al., 2017), although its release as a separate R package opens up the possibility of high-density mapping at any conceivable ploidy level given pairwise recombination frequency information. With the MDSMap algorithm, we were able to order all markers from a CLG in one run (ranging from 1,721 to 2,404 markers per CLG), resulting directly in an integrated map. This number of markers would previously have been completely intractable using a WLR. With the new algorithm, such maps can be produced on an average desktop computer within hours. As the ordering step is time and resource efficient, running multiple rounds of mapping with removal of problematic markers is much more feasible.

Linkage map quality was assessed based on two analyses: on the concordance between $r_{pairwise}$ and $r_{map}$, and on the relative position between 1x0 markers originating from the same contig from our transcriptome assembly. According to the comparison of $r_{pairwise}$ and $r_{map}$, the two estimators were in concordance if $r_{pairwise}$ could be estimated with high confidence (high LOD). Therefore, there is little discrepancy between the distance and ordering of different combinations of markers on the linkage map and their initial estimation of $r$. The second analysis resulted in information on the quality of local integration of homologous chromosomes. This is based on the assumption that recombinations are essentially absent within a transcript contig. Therefore, we would expect that markers originating from the same contig have a distance very close to 0 cM. From the position of markers originating from the same contigs, a difference from zero can be calculated, and with that a measure for error; we used the RMSE. The RMSE of 1x0 markers mapped on the same homologue was generally very low indicating both a high-quality assembly of transcripts containing mapped markers, and high-quality local ordering at the level of the homologue. The RMSE of 1x0 markers phased on different homologues was higher. Of all marker-type combinations, the estimation of genetic distance between 1x0 markers in coupling is most accurate. Estimation of distance between 1x0 markers in repulsion relies on higher dose markers, because high-confidence estimation of recombination frequency in repulsion of 1x0 markers was not possible with our population size of 406. The positions of 1x0 markers that are on different homologues relative to each other are therefore estimated with lower certainty than if they were in coupling phase. However, errors in estimating distance between these 1x0 markers in repulsion were in general lower than 5 cM. If serious ordering issues occurred, a much higher value would be anticipated. Nevertheless, there were four homologue combinations with RMSE values higher than 10 cM. Only one or two markers per CLG caused these high values. As these markers were not associated with any notable stress on the map, it is likely that these markers were actually from different

loci in the genome, and the contigs they originated may be the result of a chimeric contig assembly from two very similar transcripts originating from the same chromosome.

Earlier linkage maps of chrysanthemum are based on RAPD, ISSR, AFLP (Zhang et al., 2010a) and SRAP markers (Zhang et al., 2011a). A disadvantage of these types of molecular markers is that they are difficult to transfer, and different linkage maps therefore cannot be integrated. SNP markers are sequence based, and executing single SNP assays like KASP™ or TaqMan™ are commonly applied laboratory procedures. They can therefore be flawlessly transferred between laboratories. To set a standard for chrysanthemum, we present the sequences of a set of 92 well-distributed SNP markers originating from conserved coding sequences that can be used as a core set to align future linkage maps to each of the chromosomal linkage groups presented here.

## Estimating IBD probabilities

We used a relatively simple approach to estimate IBD probabilities for absence or presence of parental haplotypes in our segregating population (Bourke, 2014). The method only uses information of dosage scores if they are fully informative. This means that in case of a 1x1 marker for example, a dosage of 0 and a dosage of 2 in the progeny is fully informative (while assuming absence of double reduction), because it represents inheritance of respectively none of the associated homologues or both. A dosage of 1 is not fully informative as it is not known from which parental homologue the allele originated. Therefore, higher dose markers carry relatively few informative dosage scores. A consequence of our method is that it is only accurate if markers with a large fraction of informative dosage scores are equally distributed over the homologues. In our data, parts of homologues were sometimes poorly endowed with informative markers. This did not turn out to be problematic if at that position all other five homologues for that parent carried enough information. More sophisticated methods have shown that higher dose markers add more information to the estimation of IBD probabilities (Hackett et al., 2013;Zheng et al., 2016). Such methods could result in more accurate IBD estimates, but an adequate marker distribution over all homologues is key to all methods.

The accuracy of genetic analysis based on IBD probabilities relies on the quality of the integrated map. If the estimation of distance between markers with alleles on different homologues is poor, estimation of IBD probabilities of alleles on the presumed same locus will be wrong, and will therefore provide a poor representation. However, the RMSE of the marker positions on the integrated map was generally well below 5 cM. This would not have a large effect on the estimation of IBD probabilities, because according to Haldane's mapping function a distance of 5 cM corresponds to a

recombination frequency of 0.047, resulting in a relatively low error of 4.7% on the estimation of IBD probabilities.

## QTL mapping

With the integrated map and IBD probabilities, we were able to perform a multi-allelic QTL analysis. In a polyploid, this type of analysis has large advantages over the use of methods that are developed for diploids, because QTL that are regulated by multiple different alleles can be detected and their genetic architecture investigated (Hackett et al., 2014). In a polyploid, more than two alleles can underlie a QTL. This means that the QTL genotype does not only have a dosage, but can also be multi-allelic (*i.e.* not only different conformations of the alleles A and B, but also combinations of *e.g.* A,B,C,D,E and F are possible within a locus). To investigate the genetic architecture and with that the occurrence of multi-allelic QTL, we performed a QTL analysis that makes use of using IBD probabilities for four traits with different underlying genetic architecture.

The major loci associated with flower colour were bi-allelic. Together, they explained a large part (47.8%) of the phenotypic variation and were affected by one allele for each of the two loci. The two loci clearly showed an interaction, suggesting that presence of both alleles is needed for pink colouration. In chrysanthemum, pink colouration is caused by anthocyanin accumulation (Stickland, 1972). The interaction between alleles could be caused by the requirement of two enzyme variants needed for the production or regulation of production of anthocyanin, or two gene copies that are required for the same limiting step, needing the additive effect of both to become visible.

Several QTLs associated with flowering time, disk floret degreening and number of ray florets were multi-allelic. These QTLs had underlying alleles with a positive effect, a negative effect and no significant effect on the phenotype, indicating presence of at least three alleles. The exact number of unique alleles that affect the phenotype is difficult to determine. Two haplotypes that have the same effect on the phenotype could have the same underlying polymorphism affecting the phenotype, which would make them the same alleles. On the other hand, they could contain different causative polymorphisms that have a similar effect on the phenotype. Based on our data, it is not possible to uniquely identify such alleles, because our analysis is based on genetic linkage, and the causative alleles cannot be identified.

Compared to flower colour, the genetic architecture for flowering time was more complex. The QTL at CLG4 was bi-allelic, meaning that presence of one allele affected the trait, whereas the other eleven alleles did not significantly affect the phenotype. However, in two other major QTL on CLG2 and 3 multiple alleles were involved. Other studies on the inheritance of flowering time in chrysanthemum also suggested

involvement of multiple loci (Zhang et al., 2011b;Zhang et al., 2013a). Flowering time in short day plants is mainly the result of an interaction between growth rate and signal transduction of environmental cues like day-length and temperature. As these cues are strictly controlled in a greenhouse, the role of the environment would be expected to be relatively small. This is supported by the relatively high heritability (0.70), which was also found earlier (De Jong, 1984). However, genetic regulation of signal transduction and growth rate is likely complex and it is therefore not surprising that multiple loci are involved.

Disk floret degreening is an important determinant of postharvest performance of chrysanthemum after long storage (van Geest et al., 2017b). Three multi-allelic QTL were identified. These QTL explained a relatively small fraction of the phenotypic variation. Disk floret degreening is a physiologically complex trait; in the investigated population it is related to carbohydrate content of the disk floret at harvest (van Geest et al., 2017b). Many sub-traits could affect carbohydrate content, including genotypic variation related to photosynthetic rate and source-sink relationships. Furthermore, it was shown that carbohydrate content is not the only factor affecting degreening (van Geest et al., 2016). It is therefore not surprising that we did not find major QTL for disk floret degreening. Dissecting the trait further by phenotyping for sub-traits such as carbohydrate content, or by backcrossing progeny harbouring specific trait characteristics might help to further identify specific loci underlying this complex trait.

The number of ray florets had the highest heritability of the investigated traits (0.72), but least variation could be explained by detected QTL. Asteraceae plants carry composite flower heads that are comprised of multiple florets. Those florets can be categorised into disk florets and ray florets. The number of ray florets is affected by the number of florets on a capitulum and organ identity of those florets. Regulation of floret identity is generally inherited through one or two major loci in Asteraceae (Gillies et al., 2002). It is therefore quite unexpected we did not find any major QTL associated with the trait. As both parents were of the single flower type, it is possible that both lacked allelic variation in the major genes, and we only found variation in more complexly regulated minor allelic effects.

In the QTL analyses, possible interactions between alleles were not taken into account. An alternative model as described by (Hackett et al., 2014) that uses all possible genotype classes as parameters would enable detection of interactions. However, the method we used to estimate IBD probabilities is not able to estimate probabilities for these genotype classes directly. More importantly, in a tetraploid, there are 36 possible genotype classes ($\binom{4}{2} \times \binom{4}{2}$), leading to a model with 36 parameters that is already prone

to over-fitting. In a hexaploid this would be 400 genotype classes ($\binom{6}{3} \times \binom{6}{3}$), leading to 400 parameters; over-fitting would definitely become an issue.

Our results show that hexaploidy in chrysanthemum complicates QTL analysis because multiple alleles with a differential effect can underlie an associated locus. With the integrated map and IBD probabilities we were able to identify inheritance of parental haplotypes in the progeny, enabling us to identify effects of specific alleles that affected the phenotype. We indeed found clear examples in which different alleles from the same locus and parent affected the trait negatively or positively. With these findings we show that polyploids with polysomic inheritance can harbour much more diversity on a single locus compared to a diploid, and this is very important to take into account during QTL detection and breeding.

# Conclusions

The methods described in this paper enable construction of integrated linkage maps in hexaploids with polysomic inheritance. Our presented methods can be used for future projects that aim to construct integrated linkage maps and perform multi-allelic QTL analyses in hexaploids. Success of such projects depends on several features of the investigated organism and the obtained dataset. First, it depends on the predominance of random bivalent pairing at meiosis. Second, sufficient and evenly distributed 1x0 markers are required that can define each homologous chromosome. Last, higher-dose co-dominant markers (both uni-parental and bi-parental *e.g.* 1x1, 2x0 and 3x0) with alleles on each homologous linkage group are needed that provide information to integrate homologous linkage groups into chromosomal linkage groups. With the resulting integrated linkage maps, it is possible to perform QTL analysis that takes all possible alleles into account at the same locus. This has major impact on the possibilities for localisation of genomic loci and their genetic architecture associated with traits in chrysanthemum, but also for other agriculturally-important hexaploid species such as sweet potato, kiwi and persimmon.

# Acknowledgements

The authors would like to thank Katherine Preedy and Christine Hackett of BioSS, the James Hutton Institute in Dundee for making a developmental version of the MDSMap software available. The authors would also like to thank René Smulders for suggesting revisions to the manuscript.

***Supplementary data is available online at:***
https://link.springer.com/article/10.1007/s00122-017-2974-5#SupplementaryMaterial

# Chapter 8

# Quantifying the power and precision of QTL analysis in autopolyploids under bivalent and multivalent genetic models

Peter M. Bourke[1], Christine A. Hackett[2], Roeland E. Voorrips[1], Richard G. F. Visser[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[2] Biomathematics and Statistics Scotland, Invergowrie, Dundee DD2 5DA, UK.

*Submitted*

# Abstract

New genotyping technologies, offering the possibility of high genetic resolution at low cost, have helped fuel a surge in interest in the genetic analysis of polyploid species. Nevertheless, autopolyploid species present extra challenges not encountered in diploids and allopolyploids, such as polysomic inheritance or double reduction. Here we investigate the power and precision of quantitative trait locus (QTL) analysis in outcrossing autotetraploids, comparing the results of a model that assumes random bivalent chromosomal pairing during meiosis to one that also allows for multivalents and double reduction. Through a series of simulation studies we found that marginal gains in QTL detection power are achieved using the double reduction model but at the cost of an impaired ability to determine the most likely QTL segregation type and mode of action. We also explored the effect of variable genotypic information across parental homologues and found that both QTL detection power and precision require high and uniform information contents. This suggests linkage analysis results for autopolyploids should be accompanied by marker coverage information across all parental homologues along with the per-homologue genotypic information coefficients (GIC). Visualising the GIC landscape of the homologues of interest around QTL peaks will help elucidate the limitations of QTL power and precision in further studies. Application of these methods to an autotetraploid potato (*Solanum tuberosum* L.) mapping population confirmed our ability to locate and dissect QTL in highly heterozygous outcrossing autotetraploid populations.

# Key words

Quantitative Trait Locus (QTL) analysis, autopolyploid, double reduction, QTL power, Bayesian Information Criterion (BIC), genotypic information coefficient (GIC).

# Introduction

Autopolyploid species, characterised by having more than two homologous copies of each chromosome, present a number of challenges to genetic research not present in diploids or allopolyploids (which are essentially already diploidised, genetically-speaking). They have therefore been somewhat left behind when it comes to tools and methods for their genetic analysis. Among these challenges are the complexity of modelling polysomic inheritance and the occurrence of double reduction (the phenomenon whereby a particular segment of a parental chromatid and its recombinant "sister" copy migrate to the same gamete, which can only occur if multivalent pairing structures are formed and maintained through meiosis 1 (Haldane, 1930;Mather, 1935)). However, we are now at a stage where the availability and low cost of genotyping tools (based on single nucleotide polymorphisms, or SNP markers) are making the analysis of autopolyploids not just feasible, but of practical importance to crop breeders, increasing the need for both the methods and the tools to conduct these analyses, as well as knowledge on how best to apply these methods.

Breeders and researchers are often interested in knowing the genetic architecture of important traits, for example: 1. whether they are oligo- or polygenic; 2. where the quantitative trait loci (QTL) influencing the trait lie on the genome; 3. from which specific parental homologous chromosome (which we term "homologue") the favourable alleles originate; 4. whether these alleles exhibit additive or dominant gene action; 5. whether they interact with alleles at other loci, or with the environment. Up to now, approaches such as QTL mapping in bi-parental populations or genome-wide association studies have been proposed, although not always addressing all of the above points. More recently, genomic selection has been advocated as a powerful method to increase genetic gain, without necessarily needing an understanding of the genetic architecture (Meuwissen et al., 2001;Slater et al., 2016). For quantitative traits with hundreds or thousands of causative genes, this is likely to be more appropriate in breeding programs. However, for traits with a few major causative loci, QTL mapping remains a viable option that additionally offers the promise of both understanding the genetics underlying the trait (including the possibility of finding the underlying genes) while also facilitating selection for it through marker-assisted selection.

QTL mapping in autopolyploids has evolved in the last 20 years to keep pace with changes in genotyping technologies. Approaches have been developed for both co-dominant and dominant marker systems, and range from simplified models that only consider bivalent pairing to more complex models that also include double reduction. However, there has been almost no investigation into the applicability or advantages of different models. Despite this, it is often asserted that models that ignore double

reduction are *a priori* inferior to those that include it (Luo et al., 2004). More complete models of polysomic inheritance that include double reduction are often developed under the assumption of completely-informative marker systems (*e.g.* (Xie and Xu, 2000;Li et al., 2010;Xu et al., 2013)), thereby avoiding the statistical complexities imposed by partially-informative markers (such as dosage-scored SNP markers). In fact, it is quite telling that the only publically-available tools for QTL analysis in autopolyploids have adopted the simplifying assumption of random bivalent pairing (*e.g.* TetraploidMap (Hackett and Luo, 2003;Hackett et al., 2007) and TetraploidSNPMap (Hackett et al., 2017)), whereas more complete QTL model descriptions remain unimplemented.

Recently, a method to reconstruct inheritance probabilities (or identity-by-descent (IBD) probabilities) under both bivalent and multivalent pairing models has been developed into a software package TetraOrigin (Zheng et al., 2016). We used TetraOrigin and the simulation software PedigreeSim (Voorrips and Maliepaard, 2012) to investigate QTL mapping in autopolyploids, estimating QTL detection power and precision, the effect of double reduction and multivalent pairing (while comparing models that both ignore and include it), the impact of population size, trait heritability and marker distribution, and the differences in QTL detection and diagnostic power (*i.e.* correctly predicting the QTL position as well as the composition of the QTL alleles) between simple or more complex QTL segregation types and different modes of action (additive or dominant). We also examined the well-studied traits of plant maturity (earliness) and flesh colour in a bi-parental tetraploid potato population to further illustrate our findings, comparing the QTL locations to the physical positions of candidate genes underlying these loci.

One important aspect of QTL mapping that remains conspicuously absent from most published QTL studies in both diploid and polyploid species is the topic of information content. Originally it was noted that "marker information content" could adversely affect the estimated position of a QTL if markers of variable informativeness were located near a QTL (Knott and Haley, 1992). Increasing information content was found to lead to an increased test statistic, which could bias the location of a QTL peak (Knott and Haley, 1992;Knott et al., 1997). Other authors have proposed alternative measures of information content than that of Knott and Haley, such as one based on Shannon's information content (Reyes-Valdes and Williams, 2005). In autopolyploid species the issue of information content is arguably even more important than in diploids, as information content generally varies between homologues. We prefer to use the term "genotypic information coefficient" (Van Ooijen, 1992;Van Ooijen, 2009) as it avoids confusion when dealing with IBD probabilities, which are multi-point estimates of homologue transmission probabilities and not the marker genotypes themselves. In this article we extend the definition of the genotypic information coefficient (GIC) to

autopolyploids and explore its usefulness in predicting QTL detection power and precision. Our work will help to guide better-informed QTL analyses in autopolyploids as well as deepening our understanding of the genetic architecture controlling important traits in these species.

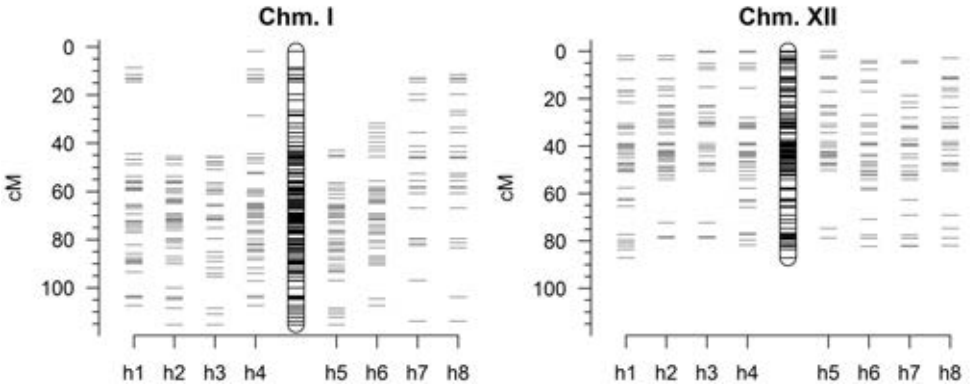



**Figure 1. Distribution of markers on potato chromosomes 1 and 12 from the integrated genetic maps of Hackett et al. (2013).** Chromosome 1 shows poor marker coverage across most homologues in the region 0 – 40 cM and was therefore chosen to investigate the genotypic information coefficient (GIC). Chromosome 12 has better overall marker coverage. Parental homologue (h) numbers are shown as h1 – h4 for parent 1 and h5 – h8 for parent 2.

# Materials and Methods

Simulated tetraploid populations were the basis for much of the work presented here. We based our simulated populations on previously-published genetic maps of tetraploid potato developed from dosage-scored SNP markers (Hackett et al., 2013). In polyploids, marker dosages are generally understood to correspond to the allele counts of the "alternative" allele (as opposed to the "reference" allele of a bi-allelic SNP marker). In heterozygous autotetraploids the possible dosages are 0, 1, 2, 3 and 4, with a marker being defined by its maternal and paternal dosages (*e.g.* 1x0 means a dosage of 1 in parent 1 (the mother) and a dosage of 0 in parent 2 (the father)). There are nine "fundamental" marker segregation types to consider in an autotetraploid cross, namely 1x0, 0x1, 2x0, 0x2, 1x1, 1x3, 1x2, 2x1 and 2x2, to which all other marker types can be converted without loss or distortion of linkage information (Bourke et al., 2016). For convenience, we often refer to simplex x nulliplex or "SxN" markers to indicate both 1x0 or 0x1 markers (similarly, duplex x nulliplex markers (DxN) imply either 2x0 or 0x2). Simplex x simplex (SxS) refers to markers whose fundamental form is 1x1.

**Table 1. Details of parameter levels used in the simulation study.**

| | q | N | $h^2$ | QTL seg. | QTL action | QTL pos. | QTL model |
|---|---|---|---|---|---|---|---|
| **Chm. 1** | 0 | 200 | 0.1 | SxN | additive | random | noDR |
| | 0.2 | 400 | 0.2 | DxN | dominant | | DR |
| | 0.5 | | | SxS | | | |
| **Chm. 12** | 0 | 200 | 0.1 | SxN | additive | 14 cM | noDR |
| | 0.1 | 400 | 0.2 | DxN | dominant | 49 cM | DR |
| | 0.2 | | | SxS | | | |
| | 0.3 | | | | | | |
| | 0.4 | | | | | | |
| | 0.5 | | | | | | |

The phased genetic linkage maps of potato chromosomes 1 and 12 from Hackett et al. (2013) were used in the simulation of genotyped tetraploid populations with PedigreeSim in this study. *Chm.* = chromosome; $q$ = rate of quadrivalent formation (specified in PedigreeSim .chrom file); $N$ = population size of $F_1$ population; $h^2$ = (broad-sense) trait heritability; *QTL seg.* = QTL segregation type, given as the maternal x paternal dosage of the (alternative) marker allele. The codes SxN refers to 1x0 and 0x1 markers, DxN refers to 2x0 and 0x2 markers and SxS refers to 1x1 markers; *QTL pos.* = genetic position of the QTL. This was either random (chm. 1) or confined to a telomeric (14 cM) versus centromeric (49 cM) position on chm. 12; *QTL model* = model used in QTL analysis, either with double reduction (DR) or not (noDR).

We limited our attention to potato chromosomes 1 and 12 as these displayed contrasting levels of marker coverage (Figure 1). The range of parameters used in both simulation sets are outlined in Table 1. We used potato chromosome 1, with its uneven marker distribution, for the investigations into the genotypic information coefficient (GIC), while chromosome 12 was used for the power analysis. Unless otherwise stated, all statistical analyses and visualisations were performed using the R statistical computing environment version 3.3.2 (R Core Team, 2016).

# GIC study – potato chromosome 1

For each set of population parameters (all possible combinations of population size (Pop.) and rate of quadrivalent formation ($q$)) we simulated 10 separate populations using PedigreeSim and the phased linkage map of chromosome 1 from Hackett et al. (2013) for the phased parental marker positions and dosages (visualised in Figure 1, left-hand side). Each simulated individual carried a single chromosome. For each population, we generated 100 phenotype sets for all possible combinations of the factors heritability ($h^2$), QTL segregation type (QTL seg.) and QTL action (Table 1). The phenotype of the $i^{\text{th}}$ individual ($P_i$) with QTL dosage $d_i$ was randomly sampled from a Normal distribution according to:

$$P_i \sim \mathcal{N}(\mu + Q * d_i, \sigma_e^2)$$

where $\mu = 10$, $Q = 1$. The environmental variance $\sigma_e^2 = \left(\frac{1-h^2}{h^2}\right)\sigma_g^2$ was determined by first calculating the genotypic variance $\sigma_g^2$ across the whole population given the individual QTL dosages (in the case of a dominant QTL these were taken as a dosage of 0 and 1 only). Offspring QTL genotypes were derived from the .hsa and .hsb output files of PedigreeSim (which provide the exact location and origin of recombination points along offspring homologues). Both the position and the configuration of the QTL were randomised for each phenotype set (to 0.01 cM accuracy).

TetraOrigin (Zheng et al., 2016) was run on Mathematica version 10 (Wolfram Research Inc., 2014) with input files derived from the integrated linkage maps and dosage output of PedigreeSim, using both bivalent_decoding options (False / True) to generate IBD probabilities under both a model that allowed for double reduction (DR) and one that did not (noDR). The other parameter settings used were parental dosage error probability (epsF) = 0, offspring dosage error probability (eps) = 0.001, and parental bivalentPhasing = True (*i.e.* assuming purely bivalent pairing predominates to determine parental marker phase, for computational efficiency (Zheng et al., 2016)). The IBD probabilities at the marker positions were used to fit splines (using the `smooth.spline` function in R (R Core Team, 2016)) from which re-normalised probabilities were interpolated at a 1 cM grid of positions (using the `predict` function in R) for subsequent QTL analysis.

QTL analysis was performed using a weighted regression of the homologue effects, weighted by the IBD genotype probabilities (Hackett et al., 2014). The QTL model described by Hackett et al. (2013, 2014), derived from the earlier work of Kempthorne (Kempthorne, 1957), can be written as:

$$Y = \mu' + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 + \varepsilon$$

having taken the constraints $X_1 + X_2 + X_3 + X_4 = 2$ and $X_5 + X_6 + X_7 + X_8 = 2$ into account. Here, $Y$ corresponds to the trait values, $X_i$ the indicator variables for the presence / absence of a particular parental homologue (1-4 for parent 1, 5-8 for parent 2) and $\varepsilon$ the residual term. Hackett et al. (2014) describe this as the "additive" model, and used the probabilities of the 36 possible genotypes as weights in a regression using the above model form. We firstly applied this approach (which we term the "no double reduction" or "noDR" model, where all $X_i = 0$ or 1), subsequently extending it to use the set of 100 genotype probabilities as weights (termed the "DR" model which includes the additional genotypes resulting from double reduction, *i.e.* $X_i$ are no longer constrained to equal 0 and 1, but rather $X_i = 0$, 1 or 2). The "logarithm of odds ratio" (LOD) score for the regression was calculated using the formula

$$LOD = \frac{N}{2} \log_{10}\left(\frac{RSS_0}{RSS_1}\right)$$

where $N$ is the population size, $RSS_0$ is the residual sum of squares under the null hypothesis of no QTL ($RSS_0 = \sum_i (y_i - \bar{y})^2$ for trait values $y_i$ and overall trait mean $\bar{y}$), and $RSS_1$ is the residual sum of squares from the regression model (Broman et al., 2003). A chromosome-wide scan was performed at 1 cM intervals and the LOD score recorded at each position.

Significance thresholds were determined through permutation tests (Churchill and Doerge, 1994), with each of the 1000 simulated phenotype sets per parameter set (10 populations x 100 phenotypes) permuted once before recording the maximum LOD score from the chromosome-wide scan (*i.e.* recording 1000 maxima). This generated approximate experiment-wise LOD thresholds by taking the 0.95 quantile of the sorted LOD values. A QTL was declared detected if the significance at the QTL position exceeded the significance threshold. Because the true positions of most QTL were not at the grid of 1 cM positions tested, splines were fitted to the LOD profile to enable interpolation of the approximate LOD score at the position of the QTL itself (which was used to derive QTL detection rates).

The GIC values for homologue $j$ at a particular locus were determined as follows:

$$GIC_j = 1 - \frac{4}{N} \sum_{n=1}^{N} \pi_r (1 - \pi_r)$$

using the noDR IBD probabilities, where $\pi_r$ is the probability of presence of homologue $j$ in individual $n$ at this locus (Appendix 1). GIC values were calculated at all 1 cM splined positions used in the QTL scan. We considered the extension of the GIC to include the case of double reduction, but found a homologue-specific GIC was no longer easily defined when an offspring can inherit more than one copy of part of a particular homologue.

To better understand the relative importance of GIC on the power of QTL detection, we re-coded QTL detection / non-detection as 1 / 0 and ran both a simple ANOVA as well as a GLM (using a Binomial model with logit link) using the following model:

$$D = q + Pop + h^2 + QTLseg + QTLact + GIC$$

where $D$ is the QTL detection indicator variable, and the explanatory variables are $q$ (rate of multivalent formation), $Pop$ (Population size), $h^2$ (heritability), $QTLseg$ (QTL

segregation type), *QTLact* (mode of QTL action, either additive or dominant) and *GIC* (the product of per-homologue GIC values underlying the QTL alleles).

To understand the influence of GIC on the detection of more complex QTL segregation types, we categorised the per-homologue GIC as either high (H) or low (L) using a threshold of both 0.9 and 0.95 for high GIC (so for example in the former, Low GIC < 0.9 and High GIC ≥ 0.9). A DxN or SxS QTL could then be categorised as either LL, LH, HL or HH, depending on the underlying GIC at each of the alleles with positive effect. For each parameter set we compared the power of detection of LL, LH / HL (since both have one low and one high-GIC allele they were grouped together) and HH QTLs.

## QTL power analysis – potato chromosome 12

The simulations using chromosome 12 were similar to those of chromosome 1 with some differences (Table 1). Six different rates of quadrivalent formation were tested ($q$ = 0, 0.1, ..., 0.5) and for each set of population-wise parameters, 50 separate populations were simulated. For each simulated population, 50 sets of phenotypes were generated as described above, except that the position of the QTL was confined to two positions, namely 14 cM (telomeric) and 49 cM (centromeric). The choice of these positions was not arbitrary: they were chosen to minimise the effect that differences in GIC might have on the results, whilst noting that QTL positioned at the telomere itself (0 cM) would be likely to suffer from lower detection rates due to lower information contents typically observed at the telomeric extremes. The centromeric position (49 cM) was selected for study as it is known that the rate of double reduction typically falls to zero at the centromeres (Bourke et al., 2015).

The QTL analysis and setting of significance thresholds was performed as described above (although permutation tests were now based on 2500 permutations, one of each phenotype set, with $\alpha = 0.05$ as before). We also wished to investigate the rate at which the QTL segregation type and mode of action was correctly reconstructed. Hackett et al. (2014) describe a QTL model-selection method by fitting the 36 phenotype means at the QTL peak and comparing the Bayesian Information Criterion (BIC) (Schwarz, 1978) for SxN, DxN and SxS models. This is given by the formula:

$$BIC = -2\log(L) + n * p$$

where $L$ is the likelihood of the QTL model being tested, $n$ is the number of observations (either 36 or 100 for the noDR and DR models, respectively) and $p$ is the number of parameters in the QTL model.

The (log) likelihood is a sum over the different genotype classes in the population, defined by their dosages in the case of an additive QTL, or presence / absence of the QTL in the case of a dominant QTL:

$$\log(L) = -\frac{n}{2}\left(\log(2\pi) + 1 + \log\left(\frac{1}{n}\sum_j \sigma_j^2 (n_j - 1)\right)\right)$$

where $\sigma_j^2$ is the variance associated with QTL genotype class *j*, and $n_j$ is the number of observations within genotype class *j* (so that $\sum_j n_j = n$).

As well as the BIC, the Akaike Information Criterion (AIC) (Akaike, 1974) was also recorded for comparison, where

$$AIC = -2\log(L) + 2 * p$$

Whereas Hackett et al. (2014) restricted their model search to only three QTL segregation types, we expanded the search to all possible bi-allelic QTL, comprising in total 240 different QTL models (listed in Supplementary Table 1). We compared the performance of the search of the full bi-allelic model space to that of a restricted model space (the first 58 models of Supplementary Table 1, containing only SxN, DxN and SxS QTL).

# Application to real data

The *Altus* x *Colomba* (AxC) tetraploid potato mapping population was used to explore both QTL models and test the methods described earlier for simulated data. The genetic positions of 6910 SNP markers from the SolSTW 20K SNP array (Vos et al., 2015) were taken from a previously-published high-density linkage map developed using this population (Bourke et al., 2016), distinct from the maps of Hackett et al. (2013) used in the simulation study which were based on a different population. A subset of these markers were selected as input data for TetraOrigin (Zheng et al., 2016). Markers were selected so that each consecutive 0.5 cM window had (if possible) one marker of every segregation type (in total there are nine; see beginning of Methods section), selecting markers randomly among those with fewest missing values. TetraOrigin IBD probabilities computed under the assumption of no double reduction (noDR) or allowing for the possibility of double reduction (DR) were saved for later QTL analysis (after confirming that homologue numbering between the noDR and DR datasets was consistent).

The two traits investigated were plant maturity and tuber flesh colour. Both were scored on an ordinal scale, with maturity scored from 1 (very late) to 9 (very early) in increments of 1, and flesh colour scored from 4 to 8 through varying shades (4 = white, 5 = cream, 6 = light yellow, 7 = yellow, 8 = dark yellow). Maturity was scored visually in the field during the growing seasons 2012, 2013 and 2014, with flesh colour scored post-harvest for each of these years (three replicates). QTL analysis was performed as described in the previous sections, using splined IBD probabilities as weights in both a noDR and DR model for comparison purposes. Individual analyses per year were performed, as well as a general analysis using best linear unbiased estimates (BLUEs) generated using the `lme` function from the nlme R package (Pinheiro et al., 2017) with Year as random effect and genotype as fixed effect. QTLs were re-mapped by saturating the LOD-5 support intervals around the QTL peaks (no marker binning performed) and re-estimating IBD probabilities in TetraOrigin. QTL analysis was subsequently performed at the marker positions themselves (rather than at splined positions) to better estimate the peak positions. The location of the CYCLING DOF FACTOR 1 (*StCDF1*) locus on chromosome 5 (Kloosterman et al., 2013) with gene annotation PGSC0003DMG400018408, and the BETA-CAROTENE HYDROXYLASE 2 (*StChy2*) locus on chromosome 3 (Wolters et al., 2010) with gene annotation PGSC0003DMT400026363 from the potato genome sequence version 4.03 (SpudDB Genome Browser *S. tuberosum* group Phureja DM1-3 (Potato Genome Sequencing Consortium, 2011;Hirsch et al., 2014)) were used to compare the physical and genetic positions of the QTL peaks. The most likely QTL model at the peak positions was explored by calculating the BIC for all 240 bi-allelic QTL models (see previous section for details) and selecting the minimum as most likely. Models within 10 BIC of the minima were also deemed plausible and recorded.

# Results

## Effect of GIC on QTL analyses

As expected, there was a clear relationship found between the GIC per homologue and the marker coverage of that particular homologue (an example is given in Figure 2.a). Differences between coupling and repulsion marker information are detectable, for example where a single 1x0 (SxN) marker tagging homologue 4 in parent 1 at 28.5 cM gave a large boost to the otherwise low GIC values in that region on homologue 4, but also slightly increased the GIC values on homologues 1, 2 and 3 (there is no information about the meiosis of parent 2 from such a marker). The results of both the ANOVA and GLM analyses showed that the GIC explains a large proportion of the variance for QTL

detection power, although not as much as the population size or trait heritability (Supplementary File 1).

Apart from the influence of GIC on detection power, we were also interested in understanding the influence of GIC on the accuracy of QTL analysis. For this we examined more closely the position of QTL peaks in relation to their true position, for SxN QTL only (since these originate from a single homologue and are simpler to track). We noted a dramatic influence of GIC on the QTL peak position in regions of variable GIC, even in situations with 100% detection power (Figure 2.b). Local maxima in GIC such as that observed in parent 1 homologue 4 at 28.5 cM serve as local "attractors" for QTL peaks, an effect seen across all homologues. Generally-speaking, there is a tapering of GIC profiles at the telomeres, a consequence of poorer marker information (coming from one side only).

Where GIC is high, the true position and detected QTL peak closely corresponded (Figure 2.b, 40 – 100 cM region). A visualisation of the homologue-specific variation in QTL detection power is given in Figure 2.c. For more complex QTL types such as DxN or SxS QTL, we were curious to know whether the presence of a single QTL allele on a homologue with high GIC would be enough to detect that QTL, or whether high GIC was needed on both homologues. As described in the Methods section, we categorised QTLs as either LL, LH, HL or HH depending on the per-homologue GIC underlying the QTL alleles with positive effect. As can be seen in Table 2, the detection power of LL-type QTL tended to be lower than that of LH- or HL-type QTL, which themselves tended to be detected less often than HH-type QTL. In fact, the intermediate class (having only one positive QTL allele residing on a high-GIC homologue) were detected at approximately the midpoint of the LL-type and HH-type detection rates (Supplementary Figure 1).

## Power to detect QTL

As noted in the previous section, population size and trait heritability were found to have the most impact on QTL detection power, followed by GIC. On chromosome 12 we deliberately chose two QTL locations to minimise the impact of GIC and allow a comparison of centromeric versus telomeric effects (with average cross-homologue GICs of 0.95 and 0.98 for 14 and 49 cM respectively). The four most important factors in determining QTL detection power (excluding GIC) were population size, trait heritability, QTL segregation type and QTL mode of action (Figure 3). When we ran an ANOVA using all the available explanatory variables we found that neither the rate of multivalent formation ($q$) nor the form of the model used (DR or noDR) had any real impact on the QTL detection power overall (Supplementary File 2). However, there were

some instances where the DR model could improve detection power. For example, when the population size or trait heritability is low and rate of multivalent pairing is high, the DR model does offer a slight advantage (Figure 4.a), helping to maintain the same level of power as that achieved when there is strictly bivalent pairing ($q = 0$). The average distance from the QTL peak to the true QTL position is also adversely affected by double reduction (Figure 4.b). However in this instance, no matter what model is used, the QTL analysis will become slightly less accurate at higher values of $q$ (although there is some mitigation of the loss of accuracy when the DR model is used). Here we used distance as an absolute measure – the direction of this distance appeared to be biased towards the side of the QTL with greater genetic length (Supplementary Figure 2).

**Table 2. Power of detection of DxN or SxS QTL, categorised by the GIC on the homologues carrying the positive QTL alleles.**

| | | | | *High GIC threshold = 0.9* | | | *High GIC threshold = 0.95* | | |
|---|---|---|---|---|---|---|---|---|---|
| **N** | **$h^2$** | **Seg.** | **Act.** | **LL** | **LH/HL** | **HH** | **LL** | **LH/HL** | **HH** |
| 200 | 0.1 | SxS | *A* | **0.38** (643) | **0.58** (1179) | **0.75** (114) | **0.48** (1204) | **0.59** (716) | **0.81** (16) |
| 200 | 0.2 | SxS | *A* | **0.83** (632) | **0.97** (1285) | **0.98** (133) | **0.90** (1260) | **0.98** (774) | **1.00** (16) |
| 200 | 0.1 | SxS | *D* | **0.12** (641) | **0.31** (1210) | **0.49** (129) | **0.21** (1247) | **0.33** (710) | **0.70** (23) |
| 200 | 0.2 | SxS | *D* | **0.52** (574) | **0.78** (1228) | **0.93** (150) | **0.63** (1169) | **0.83** (761) | **0.82** (22) |
| 200 | 0.1 | DxN | *A* | **0.43** (947) | **0.48** (710) | **0.74** (415) | **0.46** (1396) | **0.54** (531) | **0.81** (145) |
| 200 | 0.2 | DxN | *A* | **0.89** (940) | **0.95** (626) | **1.00** (436) | **0.92** (1361) | **0.95** (507) | **1.00** (134) |
| 200 | 0.1 | DxN | *D* | **0.16** (944) | **0.20** (646) | **0.35** (372) | **0.18** (1378) | **0.25** (471) | **0.38** (113) |
| 200 | 0.2 | DxN | *D* | **0.53** (985) | **0.60** (678) | **0.87** (373) | **0.56** (1418) | **0.70** (483) | **0.90** (135) |
| 400 | 0.1 | SxS | *A* | **0.86** (654) | **0.96** (1255) | **0.98** (127) | **0.91** (1261) | **0.96** (756) | **1.00** (19) |
| 400 | 0.2 | SxS | *A* | **1.00** (670) | **1.00** (1230) | **1.00** (134) | **1.00** (1246) | **1.00** (780) | **1.00** (8) |
| 400 | 0.1 | SxS | *D* | **0.55** (681) | **0.79** (1194) | **0.91** (125) | **0.64** (1275) | **0.84** (708) | **0.71** (17) |
| 400 | 0.2 | SxS | *D* | **0.96** (684) | **1.00** (1168) | **1.00** (132) | **0.98** (1312) | **1.00** (652) | **1.00** (20) |
| 400 | 0.1 | DxN | *A* | **0.86** (946) | **0.93** (683) | **0.99** (415) | **0.89** (1392) | **0.95** (492) | **0.99** (160) |
| 400 | 0.2 | DxN | *A* | **1.00** (956) | **1.00** (643) | **1.00** (421) | **1.00** (1369) | **1.00** (502) | **1.00** (149) |
| 400 | 0.1 | DxN | *D* | **0.57** (935) | **0.57** (604) | **0.84** (465) | **0.59** (1380) | **0.70** (488) | **0.85** (136) |
| 400 | 0.2 | DxN | *D* | **0.95** (900) | **0.96** (611) | **1.00** (393) | **0.96** (1316) | **0.97** (440) | **1.00** (148) |

QTL detection power was determined for two different definitions of "high GIC" – either exceeding 0.9, or exceeding 0.95. Results for different levels of multivalent pairing and QTL model used (DR and noDR) were combined. Numbers in brackets refer to the numbers of separate QTL on which the estimates of power are based.

*N* = Population size; $h^2$ = heritability; *Seg.* = QTL segregation type, where DxN denotes a duplex QTL, so either 2x0 or 0x2, and SxS implies a 1x1 marker; *Act.* = mode of QTL action, either *A* additive or *D* dominant. LL, LH and HH refer to QTL with alleles on homologues of both Low GIC, of Low and High GIC and of both High GIC, respectively. Results for LH and HL were considered equivalent and were combined.

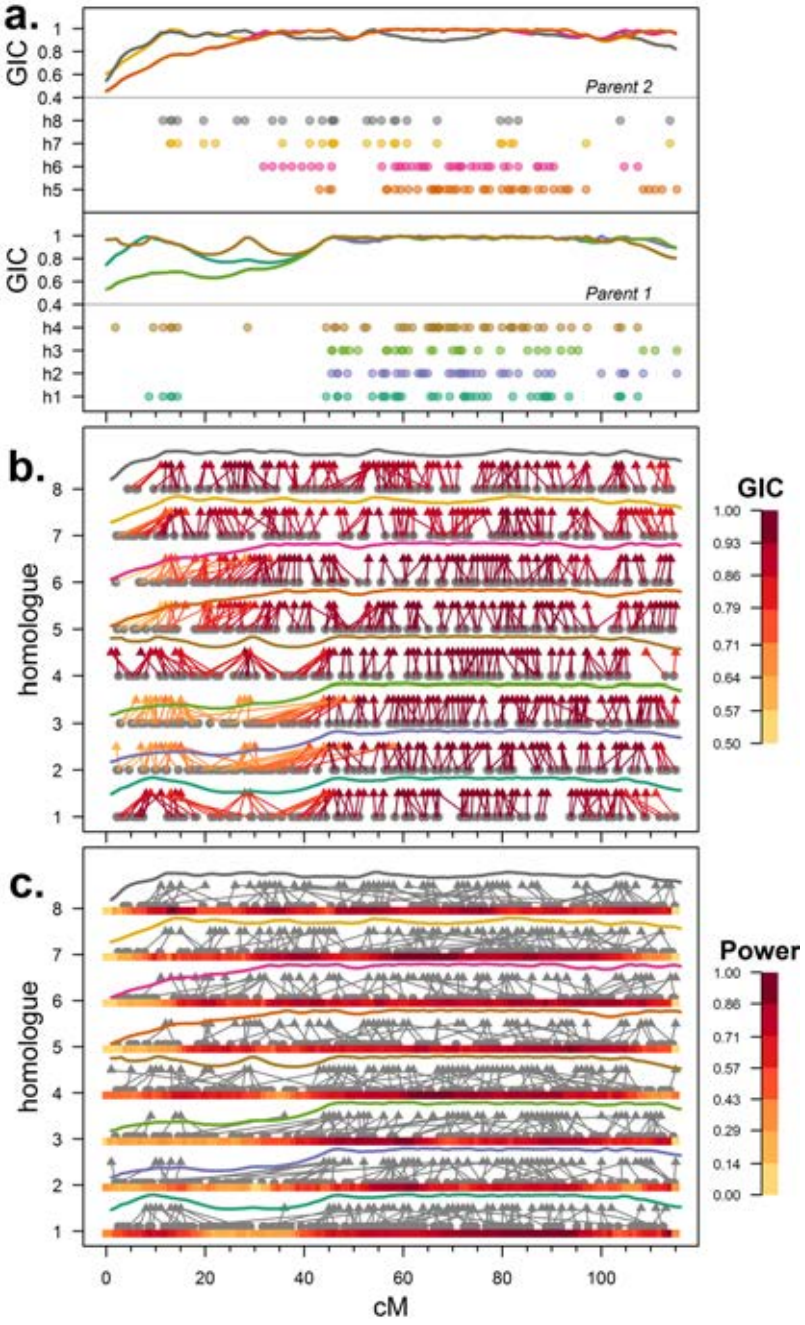



**Figure 2. The effect of variable Genotypic Information Coefficient (GIC) explored.**

**a.** Influence of marker distribution on GIC values (calculated using noDR IBD probabilities) on potato chromosome 1 (Hackett et al, 2013), with parent 1 values in the lower panel and parent 2 values in the upper panel. Average GIC values over the 10 simulated populations are shown as lines above the marker distribution (and carry the same colour). On the y-axis, h1 – h4 = parent 1

homologues 1 to 4, and h5 – h8 = parent 2 homologues 5 to 8. **b.** Effect of variable GIC on the precision of QTL detection. True QTL positions per homologue are represented by grey dots (each from a separate analysis – we did not simulate multi-QTL scenarios), with arrows indicating the position of the discovered QTL peak. Arrows are coloured by the GIC content at the QTL position itself, with average GIC lines from (a) shown above the arrows. The example shown corresponds to SxN additive QTLs with a population size of 400, heritability of 0.2 and multivalent rate $q = 0$ analysed using the noDR model. In this figure, all simulated QTL were detected (full power). **c.** Effect of variable GIC on the power of QTL detection, visualised on a per-homologue basis. Here, the power in a 10 cM sliding window is shown by a heat-map track below each homologue. QTL positions are shown as grey dots, with arrows indicating the position of the discovered QTL peaks. In contrast to (b), there was not full detection power (population size 200 and heritability = 0.1) – hence variation in QTL detection power along each homologue is apparent, and corresponds quite well with variations in the estimated GIC per homologue, shown above the arrows.

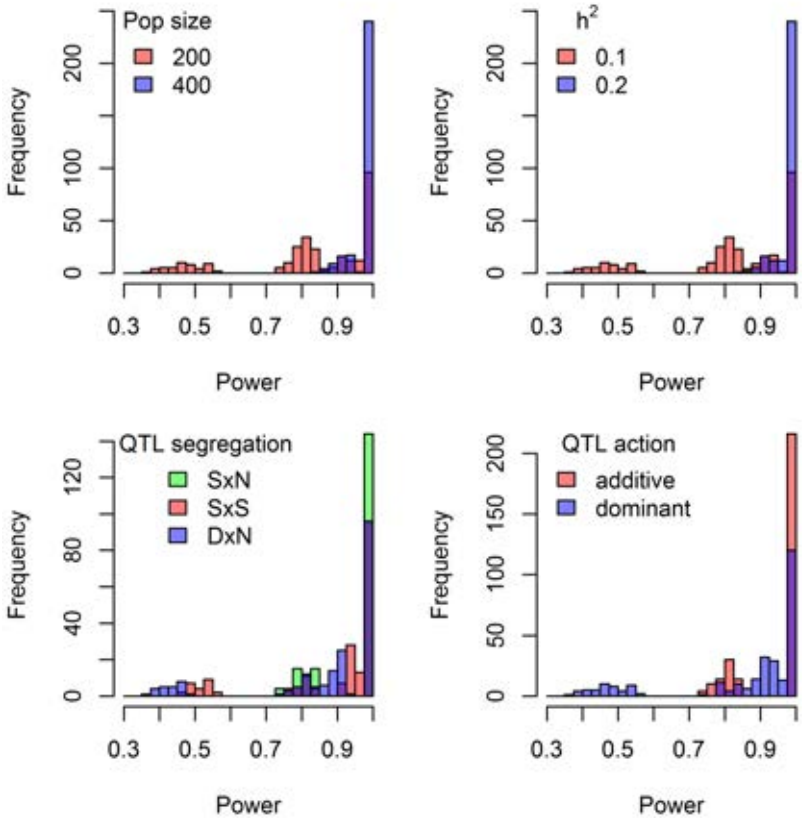



**Figure 3. Distribution of QTL detection power, divided by main explanatory effects.** Most scenarios had a very high detection power (> 0.95), although powers as low as 0.35 were also observed. Additive SxN QTL with a mapping population of 400 and heritability of 0.2 are likely to always be detected, whereas a dominant DxN QTL with a mapping population of 200 and heritability of 0.1 is unlikely to be detected more than 50% of the time.

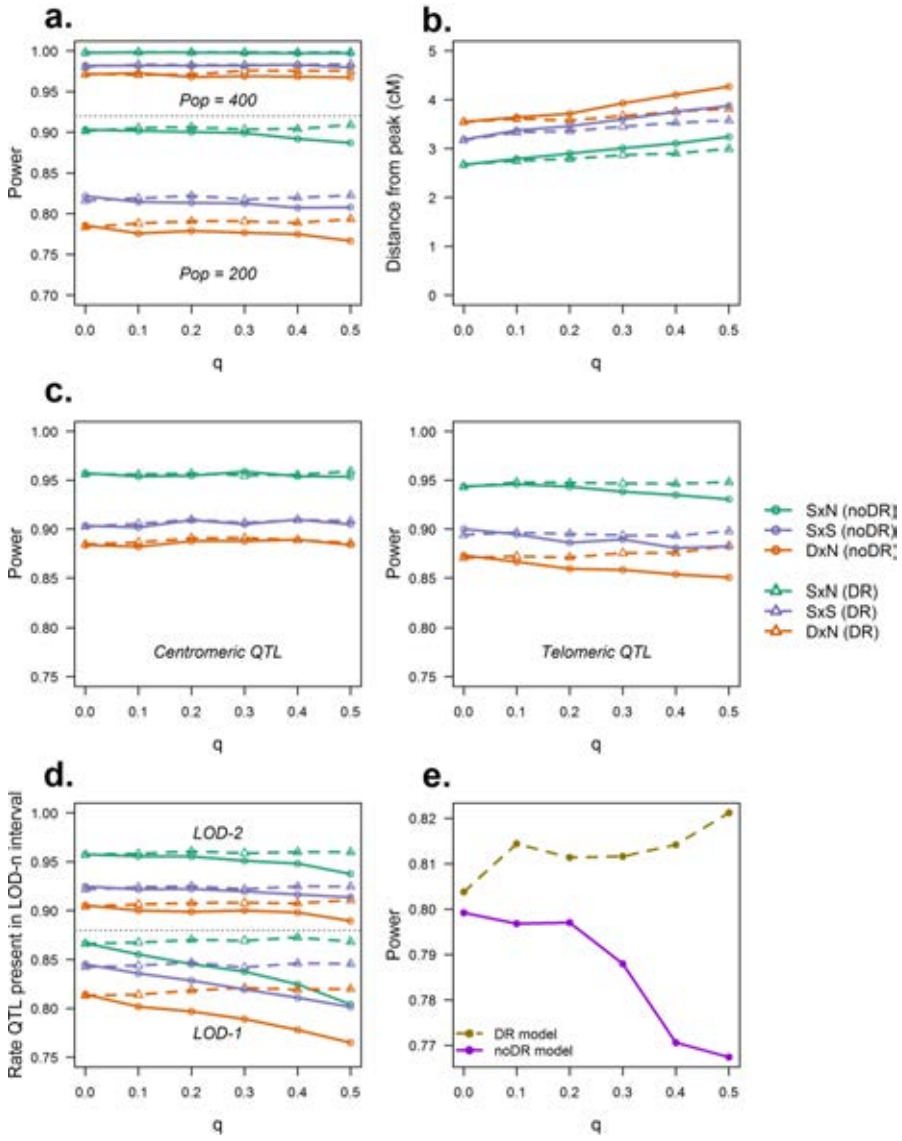



**Figure 4. Effect of various experimental parameters on QTL power and precision.**
**a.** QTL detection power for population size 200 and 400 compared over a range of different rates of multivalent formation (*q*). Solid lines indicate the results using the random bivalent noDR model, with dashed lines indicating the results using the DR model (including double reduction). Results over both trait heritabilities are combined here - the comparison between lower and higher trait heritabilities showed the same trend. **b.** Average distance between QTL peak and true QTL position across all experimental scenarios. **c.** Comparison between detection power for a centromeric QTL (left) and a telomeric QTL (right). **d.** Rate at which QTL fall within LOD-2 and LOD-1 intervals. **e.** Specific example of the difference in power between the DR and noDR models for the case of an additive SxN QTL with population size 200 and trait heritability of 0.1, up to 5% higher power on average when the rate of quadrivalent pairing was high.

As expected, there is essentially no difference between the two models for centromeric QTL, but differences do appear for telomeric QTL (Figure 4.c). If we consider the rate at which QTL were present in the LOD-1 and LOD-2 intervals instead, we see a very sharp decline in the performance of the LOD-1 interval at high levels of multivalent formation if the noDR model is used (Figure 4.d). However, it is questionable whether the LOD-1 interval should be used at all – even in the case of purely bivalent pairing, on average 16% of these support intervals contain no QTL – a value which increases to almost 32% at the lower rates of population size and heritability. The width of the support intervals around QTL peaks was also found to increase as the levels of multivalent formation increased, an undesirable effect that cannot be mitigated by using the DR model (Supplementary Figure 3). If we remove the major sources of variation from the data by considering only an additive SxN QTL with population size 200 and heritability of 0.1, we find that the DR model has the potential to increase detection power by up to 5% when the rate of multivalent formation is high (Figure 4.e). We also noted that the significance thresholds from permutation tests for both models were almost identical across all parameter settings.

## Accuracy in predicting QTL configuration and mode of action

Apart from the ability to detect QTL, we were also interested in investigating methods to correctly predict the QTL configuration (*i.e.* predicting on which parental homologues the favourable QTL alleles reside (QTL segregation), and what the most likely mode of action is (additive or dominant)). We followed the procedure described in Hackett et al. (2014) for this, using the Bayesian information criterion (BIC) to compare different bi-allelic QTL models (either a "restricted" search over 58 models (SxN, DxN or SxS only) or a "full" over 240 models (all possible bi-allelic QTL models) were tested, as listed in Supplementary Table 1). The main results are summarised in Figure 5. The BIC correctly predicted the QTL configuration and mode of action in most cases when the noDR model was used. On the other hand, the DR model was found to produce quite unreliable predictions (Figures 5.a, b, c). For example, SxS QTL were incorrectly identified for more than 50% of all simulated QTL. Increasing the breadth of the model space (from 58 (SxN, DxN and SxS) to 240 (all bi-allelic) QTL models) came at a cost – although the cost was not equal across QTL segregation types: SxN QTL suffered a smaller cost than the more complex SxS or DxN QTL. An indication of how inaccurate the DR model was can be seen in Figure 5.c – on average, the simulated SxS or DxN QTL were between 2 and 3 BIC from the most likely model. If we follow the rule of thumb that a BIC difference of between 2 – 6 constitutes positive evidence that one model is better than another (Neath and Cavanaugh, 2012), then on average we will predict that a competing (wrong) model is significantly more likely than the true QTL model if we use the DR framework.

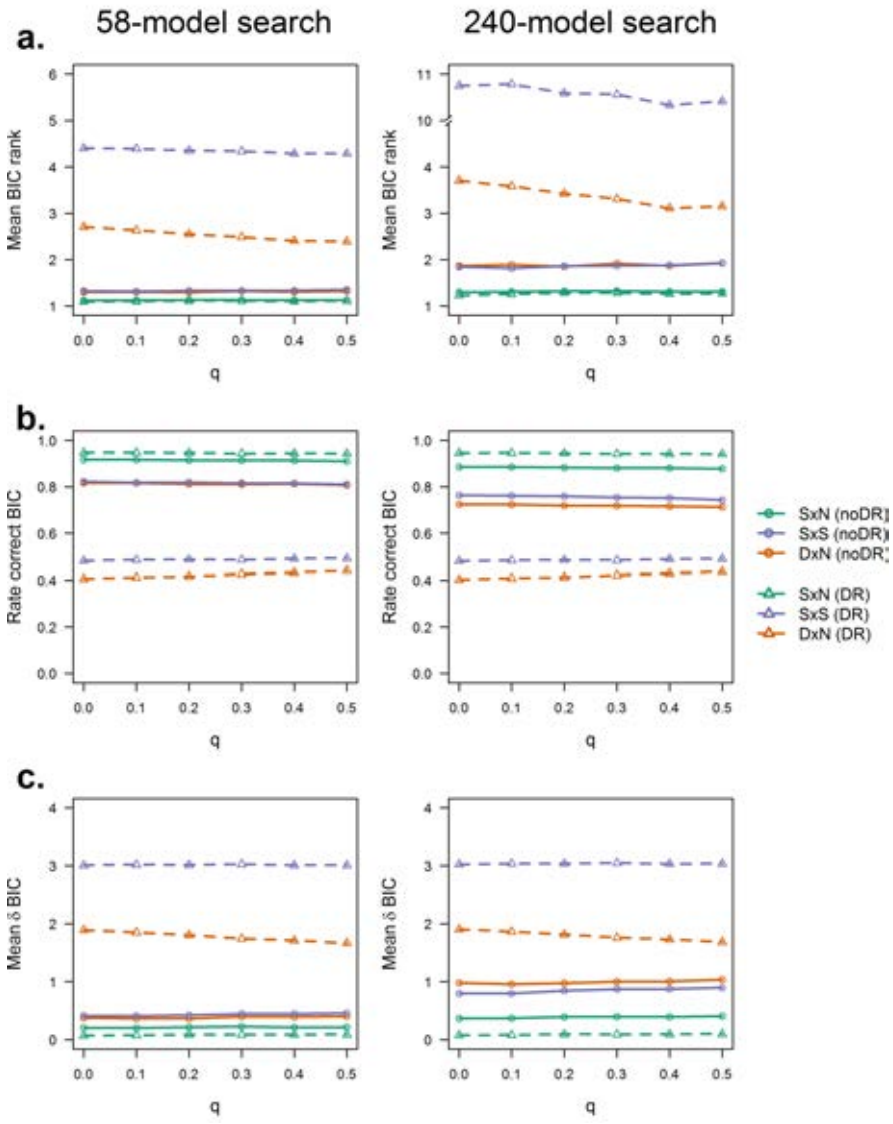



**Figure 5. Performance of the Bayesian Information Criterion (BIC) in correctly identifying QTL configuration under different experimental conditions. a.** The mean rank of the correct QTL model among those tested over a range of different rates of multivalent formation ($q$). The left panel depicts results for the restricted model search (only SxN, DxN and SxS QTL considered, in total 58 models) whereas the right panel depicts the results when the full bi-allelic QTL model space was searched (240 models). Solid lines with circles indicate the noDR results, while dashed lines with triangles indicate the DR results (including double reduction). A mean rank of 1 implies that on average the correct QTL model was always ranked as most likely (ignoring ties). **b.** The average rates at which the BIC correctly identified the QTL configuration (*i.e.* the fraction of cases where the BIC of the correct model was ranked as most likely). **c.** The average difference (in units of BIC) between the BIC-selected model and the true QTL model.

If we examine the cause of this poor performance, we find that the BIC has difficulties correctly predicting *additive* SxS or DxN QTL, whereas dominant QTL of these types cause no such issue (Supplementary Figure 4). The Akaike Information Criterion (AIC) was found to perform marginally better than the BIC if the analysis was conducted using the DR model (data not shown), although it would still be unwise to use it in conjunction with the DR model.

## Application to real data

To help illustrate our findings we looked at two well-studied traits for which phenotypic data was available from the AxC tetraploid $F_1$ potato population. AxC is the result of a wide cross between the late, white/cream fleshed starch cultivar *Altus* and the early, yellow fleshed ware cultivar *Colomba*. On average the rate of quadrivalent pairing was 0.24 (and was similar between parents: parent 1 = 0.23 ± 0.07 and parent 2 = 0.25 ± 0.05) (Supplementary Figure 5), consistent with a previous estimate of 0.2 – 0.3 from this population using only SxN marker information (Bourke et al., 2015). The phenotypic traits themselves (plant maturity and flesh colour) are already genetically well-characterised, offering the opportunity to compare QTL peak positions with the physical location of the underlying candidate genes as well as an exploration of the most likely QTL models. For both traits a single major QTL was found with both the noDR and DR models (Figure 6). As can be seen from the bottom panel of Figure 6, the GIC for some homologues was quite variable (*e.g.* chromosomes 3, 4, 11 or 12) but was overall relatively high across both parental maps.

There was a single QTL peak for plant maturity on chromosome 5 around 18 – 20 cM (Table 3). Using the DR model, slightly more of the variance was explained than with the noDR model (44% versus 41%), and the width of the LOD-2 intervals was narrower (1 cM versus 3 cM). When we saturated the QTL region with marker information and re-ran the IBD calculations and QTL analysis, we were able to increase the proportion of explained variance at the peak QTL position to 47%, with the peak occurring at approximately 20 cM using both models (Table 4). The most likely QTL model was an additive oooo x qqqQ with the early allele ('Q') coming from *Colomba* using both the noDR and DR models (Table 4). Here 'o' denotes an allele about which we have no information, which could be 'q', 'Q' or any other non-segregating allele. Note that for a simplex QTL, an additive and a dominant model produce the same expected 1:1 segregation and therefore cannot be distinguished within an $F_1$ population. The mean maturity of offspring with the early 'Q' allele from *Colomba* was 6.9 ± 0.8, and without was 5.8 ± 0.9. The next most likely QTL model was the additive model QQqQ x qqqQ, although this was far less likely (δ BIC = 18.1 or 16.8 for the noDR or DR models, respectively). It is possible that a multi-allelic model (with different earliness alleles of

different effect sizes) may have fitted the data better (Hackett et al., 2014), although we did not see the need to model that here.

**Table 3. Major QTL peaks for potato maturity (chromosome 5) and flesh colour (chromosome 3) detected in AxC, phenotyped over 3 seasons (2012 – 2014).**

| Trait | Chm. | Year | Model | Peak (cM) | N | LOD | LOD-2 (cM) | \|LOD-2\| | Var. |
|---|---|---|---|---|---|---|---|---|---|
| Maturity | 5 | 2012 | noDR | 18 | 222 | 22.3 | 17-20 | 3 | 0.37 |
| | | 2012 | DR | 20 | 222 | 25.1 | 20 | 0 | 0.41 |
| | | 2013 | noDR | 18 | 221 | 19.0 | 15-20 | 5 | 0.33 |
| | | 2013 | DR | 20 | 221 | 20.9 | 19-20 | 1 | 0.35 |
| | | 2014 | noDR | 19 | 219 | 26.6 | 17-20 | 3 | 0.43 |
| | | 2014 | DR | 19 | 219 | 28.3 | 17-20 | 3 | 0.45 |
| | | BLUEs | noDR | 18 | 222 | 25.6 | 17-20 | 3 | 0.41 |
| | | BLUEs | DR | 20 | 222 | 28.1 | 19-20 | 1 | 0.44 |
| Flesh colour | 3 | 2012 | noDR | 60 | 222 | 29.3 | 57-61 | 4 | 0.46 |
| | | 2012 | DR | 60 | 222 | 29.1 | 57-63 | 6 | 0.45 |
| | | 2013 | noDR | 60 | 221 | 30.0 | 57-60 | 3 | 0.47 |
| | | 2013 | DR | 60 | 221 | 29.0 | 57-63 | 6 | 0.45 |
| | | 2014 | noDR | 60 | 219 | 35.4 | 57-61 | 4 | 0.52 |
| | | 2014 | DR | 60 | 219 | 36.3 | 57-63 | 6 | 0.53 |
| | | BLUEs | noDR | 60 | 222 | 36.1 | 57-60 | 3 | 0.53 |
| | | BLUEs | DR | 60 | 222 | 36.0 | 57-63 | 6 | 0.53 |

*Chm.* = chromosome number; *Year* = year of phenotypic measurement, including best linear unbiased estimates (BLUEs) over the 3 years; *Model* = QTL model used, either random bivalents (noDR) or also allowing for double reduction (DR); *Peak (cM)* = position of QTL peak in centiMorgans; *N* = number of individuals with matching phenotypic and genotypic data; *LOD* = LOD score at the peak; *LOD-2* (cM) = range of QTL support interval cM positions (loci within 2 LOD of maximum LOD); |*LOD-2*| = width of the LOD-2 support interval in centiMorgans; *Var.* = proportion of phenotypic variance explained by QTL peak.

For flesh colour, the noDR and DR models resulted in the same rate of explained variance at the chromosome 3 peak (53%), with the LOD-2 interval for the noDR narrower than that for the DR model (3 cM versus 6 cM), in contrast to the results for plant maturity. When we searched for the most likely QTL model we found in both cases (noDR and DR) that a dominant model fit the data best, with segregation type oooo x QQqq (Table 4), while the additive version was not far behind (δ BIC = 3.4 in both cases). Individuals who inherited either the allele from h5 or h6 had an average flesh colour of 6.4 ± 0.2, whereas those without them had an average flesh colour of 4.7 ± 0.2. Interestingly, the next most likely model was 26.9 BIC away using the noDR model, while only 8 BIC away using the DR model.

We were interested in comparing the position of the QTL peaks with the physical location of the candidate genes *StCDF1* (for maturity) and *StChy2* (for tuber flesh colour). As described in the Methods section, we saturated the LOD-5 support interval

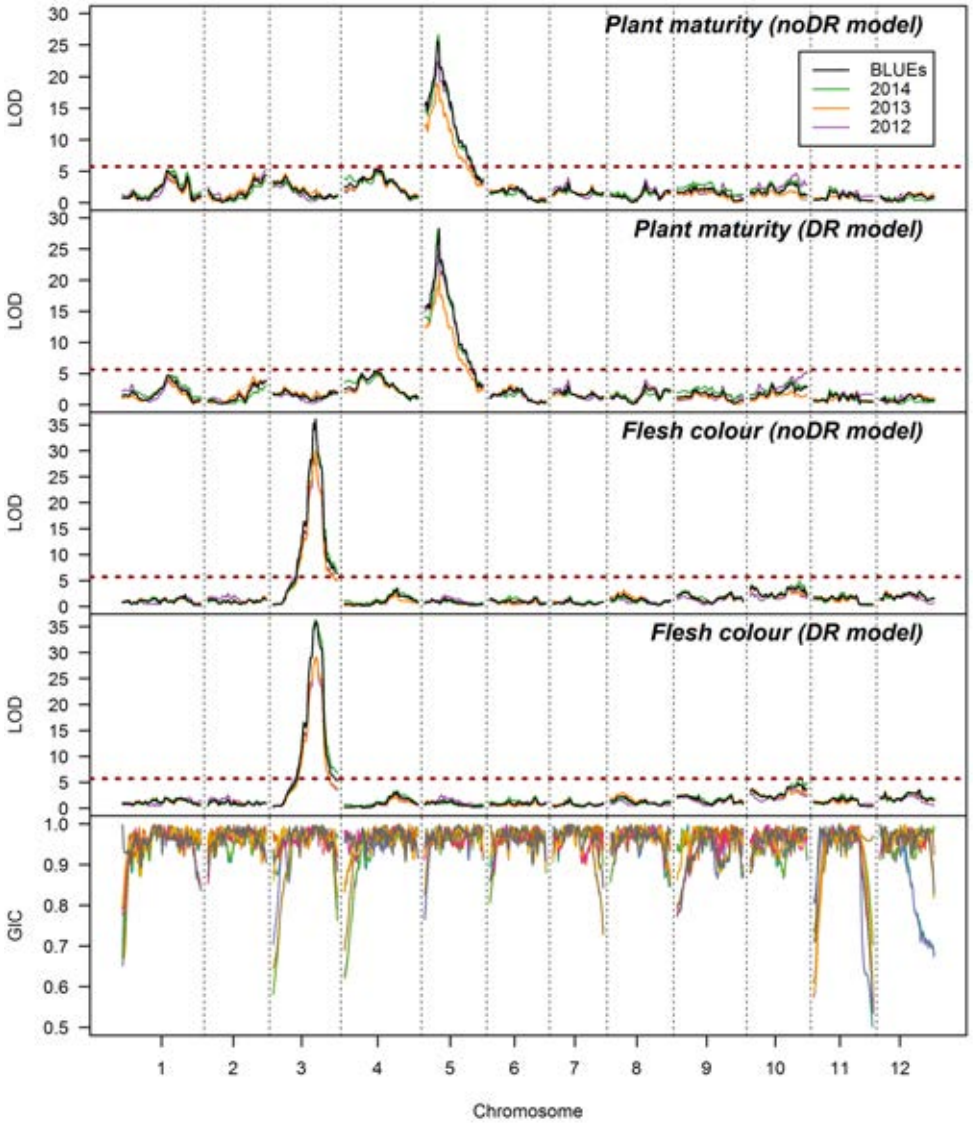



**Figure 6. Results of the QTL scans for traits plant maturity and tuber flesh colour in the tetraploid AxC potato population (N = 222 for most analyses, *c.f.* Table 3).** Results using the noDR model (random bivalent) are shown above those of the DR model (allowing double reduction) for both traits. LOD significance thresholds are shown as dashed red lines. The lowest panel shows the GIC per homologue for the eight parental homologues (using the noDR IBD probabilities), using the same colour scheme as Figure 2.a.

around the QTL peaks with markers and re-ran both the IBD calculations and QTL analysis, in a "re-mapping" of both traits. For both chromosomes three and five, the expected sigmoid correspondence between the genetic and physical position of markers

was observed, with an approximately linear correspondence in the regions of interest (Figure 7.a and 7.c). The peak LOD positions were comparable using both the noDR and DR models (Figure 7.b and 7.d and Table 4), which we compared to the GIC profiles for the homologues in question (homologue 8 for *StCDF1*, and homologues 5 and 6 for *StChy2*). In both cases, the gene position appeared to fall within the LOD-2 intervals of the QTL peaks, although there was no difference in these intervals for plant maturity; for flesh colour, the LOD-2 interval was narrower using the noDR model, and appeared to give a better indication of the QTL position that the DR model (Figure 7.c). In both cases, there appeared to be sufficient marker coverage on the important homologues within the QTL support intervals, as reflected by the relatively high GIC values. For plant maturity, the *StCDF1* region had far more mapped markers than elsewhere, suggesting that this locus was specifically targeted in the development of the SolSTW SNP array (Uitdewilligen et al., 2013;Vos et al., 2015). As can be seen from Figure 7.a, we were unable to separate these markers genetically due to the limited population size used for linkage map construction (N = 235), highlighting the inadequacy of this population size for fine-mapping work.

**Table 4. Exploration of major QTL peaks**

| *Trait* | *Chm.* | *Model* | *Peak (cM)* | *LOD* | *Var.* | *QTL seg.* | *BIC* |
|---|---|---|---|---|---|---|---|
| Maturity | 5 | noDR | 20.03 | 30.3 | 0.47 | *A/D*: oooo x qqqQ | 331.3 |
| | 5 | DR | 19.75 | 30.4 | 0.47 | *A/D*: oooo x qqqQ | 896.5 |
| Flesh colour | 3 | noDR | 56.72 | 36.3 | 0.53 | *D*: oooo x QQqq | 166.1 |
| | | | | | | *A*: oooo x QQqq | 169.5 |
| | 3 | DR | 59.66 | 35.6 | 0.52 | *D*: oooo x QQqq | 459.2 |
| | | | | | | *A*: oooo x QQqq | 462.6 |

Exploration of the major QTL peaks using all available marker information in the LOD-5 support regions. *Chm.* = chromosome number; *Peak (cM)* = position of QTL peak in centiMorgans; *LOD* = LOD score at the peak; *Var.* = variance explained by QTL peak; *QTL seg.* = QTL segregation type and mode of action, either *A* (additive) or *D* (dominant). *A/D* is used to indicate that either an additive or dominant model is possible, as these cannot be distinguished in the case of simple QTL. QTL alleles with a positive effect are denoted 'Q', alleles with a negative effect are denoted 'q' and those with unknown effect 'o'; *BIC* = Bayesian Information Criterion.

# Discussion

## The effect of a variable Genotypic Information Coefficient

Although reported as early as 1992 (Knott and Haley, 1992), the influence of a variable GIC in the vicinity of QTL has essentially been ignored in many subsequent QTL studies both at the diploid and polyploid level. In this study we hope to re-emphasise its importance by demonstrating the effect of low GIC on QTL detection power, as well as the effect of variable GIC on QTL precision.

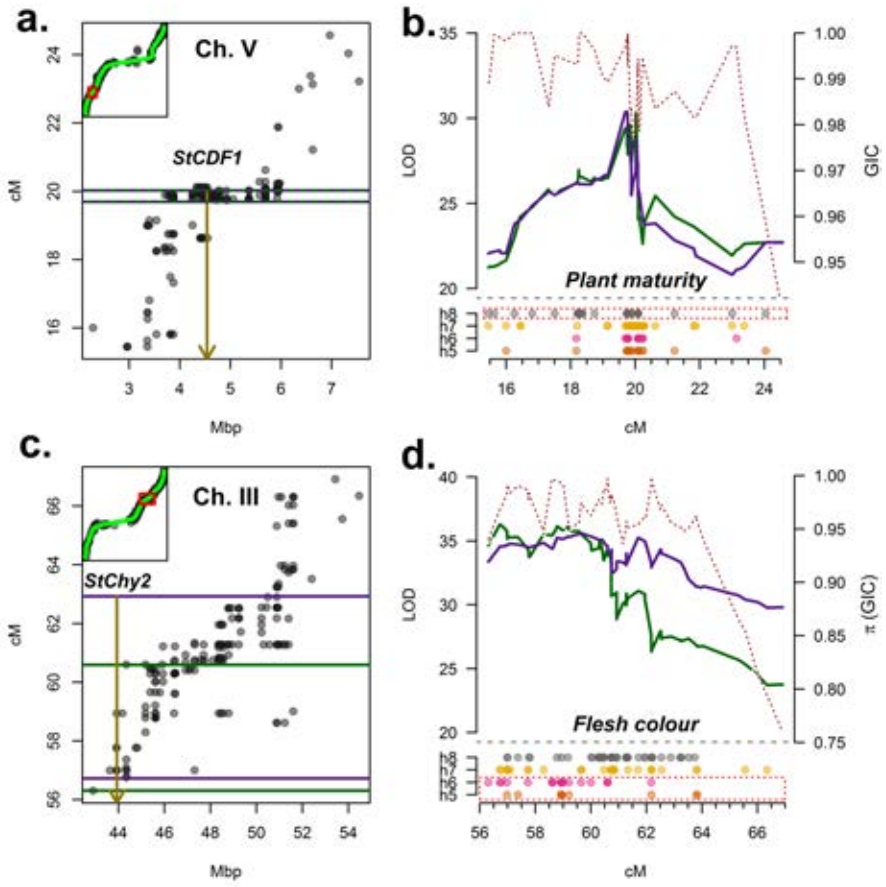



**Figure 7. Re-mapping of the plant maturity QTL on potato chromosome 5 and the flesh colour QTL on chromosome 3. a.** Genetic versus physical position of subset of markers within the re-mapped QTL region of chromosome 5. LOD-2 support intervals for the noDR and DR models overlapped, shown here as horizontal green / purple lines. The *StCDF1* locus is highlighted by a vertical arrow at ~4.54 Mbp. *Inset*: Genetic v's physical position of markers used in initial TetraOrigin analysis on chromosome 5, with re-mapped region highlighted in red. **b.** LOD profiles of the re-mapping of the chromosome 5 QTL for plant maturity. noDR model results are shown as a green solid line, with DR model a purple solid line. The GICs of homologue 8 (on which the QTL early allele was predicted to lie, highlighted underneath) for each marker position used in the QTL scan are also shown (dotted red line). y-axis labels "h5" – "h8" signify parent 2 homologues 5 through 8.
**c.** Genetic versus physical position of subset of markers within the re-mapped QTL region of chromosome 3. LOD-2 support intervals for the noDR and DR models are shown here as horizontal green / purple lines. The *StChy2* locus is highlighted by a vertical arrow at ~43.94 Mbp. *Inset*: Genetic v's physical position of markers used in initial TetraOrigin analysis on chromosome 3, with LOD-5 support re-mapped region highlighted in red. **d.** LOD profiles of the re-mapping of the chromosome 3 QTL for flesh colour noDR model results are shown as a green solid line, with DR model a purple solid line. The product of GICs ($\pi(GIC)$) for homologue 5 & 6 (on which the QTL alleles are predicted to lie (highlighted underneath) for each marker position used in the QTL scan are also shown (dotted red line).

According to our analysis, the GIC is the third-most important consideration in a QTL study (after population size and trait heritability), suggesting that dense marker coverage across all homologous chromosomes is important for successful QTL mapping. GIC values were found to drop at the telomeres, a consequence of the one-sided information available at these regions in the multi-point IBD estimation. This has the unwanted effect of biasing the QTL detection positions inwards, making it unlikely for a telomeric QTL to be found at the correct position. This could be cause for some concern, given that telomeric regions tend to be more gene-rich than more centromeric positions. However, as telomeric regions are also known to undergo more recombination (Gaut et al., 2007;Li et al., 2015), the extent of this effect is likely to be diminished by the genetic extension of telomeric regions.

Particularly in the case of autopolyploids, knowledge of homologue-specific GIC values is crucial in predicting whether a QTL is likely to lie beneath a QTL peak, since variable GIC profiles can lead to a significant bias in the estimated QTL position. We were unable to model homologue-specific GIC in the context of double reduction, as it was not obvious how GIC should behave when an offspring can inherit more than one copy of part of a particular homologue. This could also be seen as an advantage of using the noDR model, where a homologue-specific GIC is a clearly-defined concept (Appendix 1).

The GIC cannot be further increased by increasing mapping population sizes (which is often thought to be the only way to increase power), above the limitations imposed by marker density, distribution and informativeness. If GIC values are found to be low on certain homologues, it could be worthwhile to develop more markers within that region on the affected homologues. In scenarios where this is impossible (*e.g.* due to long stretches of homozygosity across homologues), the investigator remains blind to any potential QTL within that region, although it could be argued that such regions are unlikely to harbour segregating QTL either. For complex QTL types with more than one positive allele contributing to the trait, our results show that it is preferable to tag all QTL alleles through nearby informative markers rather than just one, or none.

## Double reduction or double trouble?

We initiated this study to determine whether it is worthwhile to include double reduction in a QTL model. Through a large simulation study we have demonstrated that the improvements to QTL analysis are at best marginal and sometimes counter-productive. We saw at most a 5% increase in detection power for “low-power” situations (population size 200 and trait heritability of 0.1) when a significant proportion of multivalents are formed ($q = 0.5$). In practice however, one is unlikely to encounter rates of multivalent

formation this high (Bourke et al., 2015). The main disadvantage to using the double reduction (DR) model is that it can befuddle later exploration of QTL peaks, introducing a large degree of uncertainty into the most likely QTL segregation type and mode of action. However, there is no huge computational burden to running both analyses and comparing results. The time-limiting step in the current pipeline was the calculation of parental marker phase using TetraOrigin (which is arguably a redundant step given current linkage mapping methodologies which also determine parental marker phase (Hackett et al., 2013;Bourke et al., 2016)). The subsequent estimation of both noDR and DR IBD probabilities can be very quickly generated by TetraOrigin, following which both QTL models can be fitted. Permutation test results for both models were found to be extremely comparable and do not need to be run twice per trait analysed. Although hardly surprising, our results suggest that QTL studies in autopolyploid species should focus more on experimental set-up and marker distribution than on whether double reduction should be included in the QTL model or not.

Previous studies have attempted to incorporate double reduction in their QTL models without providing sufficient motivation for such models apart from the completeness of their model (Xie and Xu, 2000;Li et al., 2010;Xu et al., 2013). As demonstrated here and elsewhere (Bourke et al., 2015;Zheng et al., 2016), estimation of the rates of double reduction per parental chromosome as well as the frequency of multivalent pairing are now relatively straightforward given modern high-density marker datasets. With this study, we hope to have provided a sufficiently balanced analysis of both the noDR and DR models to allow an informed decision about which model is most appropriate, given knowledge of the specific meiotic behaviour of the population in question.

## Full model space search versus restricted space search

We compared a search of the full bi-allelic QTL model space (under a purely additive or dominant model only) with one restricted to just SxN, SxS and DxN bi-allelic QTL and found that a penalty was incurred by including more models in the search space. Considering all possible multi-allelic QTL ($Q_1Q_2Q_3Q_4$ x $Q_5Q_6Q_7Q_8$) and a more general formulation of inter-allelic interactions would increase the computation time of this step without necessarily increasing predictive accuracy. Nevertheless for higher ploidy levels, the likelihood of having more than two distinct functional alleles at a single locus increases, which also increases the possibility of novel interactions between these alleles (which cannot simply be termed “dominance effects” anymore). Extension of our current approach to include tri-allelic QTL, or novel allelic interaction effects would be relatively straightforward. It is also interesting to speculate whether machine learning algorithms might be applied to solve this problem rather than providing a fixed number of input models as potential candidates, using some form of cross-validation to check the

results. We found that a naive application of the BIC was not suitable for the more complex DR model. This was because by using the DR model, we were generally introducing extra QTL classes, unnecessarily so as it turned out. Developing a model-comparison framework that weights offspring genotype classes by their probability (since double reduction is a position-dependent phenomenon, occurring with greater frequency towards the telomeres) appears to be necessary, although investigations in this direction fell outside the scope of the current study.

## Concluding remarks

We have shown with this work that slight gains in QTL detection power in autopolyploids can be achieved using a model that incorporates double reduction, although this comes at the cost of a diminished ability to correctly predict the QTL segregation type and mode of action using current methods such as the Bayesian Information Coefficient. This study also re-instates the importance of high genotypic information content, particularly when framed within the context of autopolyploids. We now live in the age of "high density" marker sets and linkage maps, even in polyploid species. However, high marker densities on an integrated map do not necessarily translate to high marker densities per homologue. Therefore, such terminology may offer a misleading impression and fail to provide an accurate description of marker density and distribution unless haplotype-specific maps have also been examined and provided (*e.g.* Figure 1). We have shown here that GIC has an impact on QTL detection power and accuracy. In extending the definition of GIC to autopolyploids, we hope to encourage future QTL studies in polyploid species to include this important genetic description alongside any reported QTL results.

# Acknowledgements

Funding for this research was provided through the TKI polyploids projects "A genetic analysis pipeline for polyploid crops" (project number BO-26.03-002-001) and "Novel genetic and genomic tools for polyploid crops" (project number BO-26.03-009-004). PMB received an EMBO short term fellowship to work in the group of CAH at BioSS, Dundee, Scotland (ASTF number 228 – 2016), during which this research was initiated. The work of CAH and the TetraploidSNPMap software development were funded by the Rural & Environment Science & Analytical Services Division of the Scottish Government. The authors wish to acknowledge the potato breeding companies Averis Seeds B.V. and HZPC B.V. for providing the phenotypic data on the AxC population, Michiel Klaassen for helpful comments and discussion and Glenn Marion for suggestions on the final manuscript.

*Supplementary data will accompany the online version of the published manuscript.*

# Appendix 1

## Derivation of expression for Genotypic Information Coefficient, GIC

Van Ooijen (2009) described a procedure to decompose the variance associated with a QTL ($V_Q$) into that which is explained by the markers ($V_M$) and a residual variance for which uncertainty remains ($V_R$). The GIC is then defined as (Van Ooijen, 2009):

$$GIC = {V_M}/{V_Q} = 1 - {V_R}/{V_Q}$$

We consider the case of the GIC for one homologue of a tetraploid, and limit our attention to the assumption of random bivalent pairing (noDR model). The derivation of the GIC per homologue when double reduction is admissible is less intuitive and has been omitted here.

At a given locus, we assume we have already calculated the IBD probability of inheritance of homologue $j$ ($1 \leq j \leq 8$) using TetraOrigin (Zheng et al., 2016) or some alternative method (*e.g.* (Hackett et al., 2013;Bourke, 2014)). We can consider the IBD probability of inheritance of a particular homologue to be the inheritance probability of the so-called alternative allele ($\pi_a$), with the corresponding probability of inheritance of the reference allele $\pi_r = 1 - \pi_a$.

Using the definition of variance as $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$, and using a mean of +1 for the reference allele and -1 for the alternative allele, the residual variance ($V_R$) is given by:

$$V_R = \frac{1}{N}\sum_{n=1}^{N}(\pi_r(1)^2 + \pi_a(-1)^2 - (\pi_r(1) + \pi_a(-1))^2)$$

$$\Rightarrow V_R = \frac{1}{N}\sum_{n=1}^{N}((\pi_r + \pi_a) - (\pi_r - \pi_a)^2)$$

and since $\pi_r + \pi_a = 1$,

$$V_R = \frac{1}{N}\sum_{n=1}^{N}(1 - (2\pi_r - 1)^2)$$

$$\Rightarrow V_R = \frac{1}{N}\sum_{n=1}^{N} 2\pi_r(2 - 2\pi_r)$$

$$\Rightarrow V_R = \frac{4}{N}\sum_{n=1}^{N} \pi_r(1 - \pi_r)$$

where the sum is over $N$ individuals in the $F_1$ population. If we assume that the QTL is not under selection so that the expected proportions of the reference and alternative allele are both 0.5, then the total variance is simply $0.5(1)^2 + 0.5(-1)^2 = 1$, and therefore the genotypic information content for homologue $j$ is given by

$$GIC_j = 1 - \frac{4}{N}\sum_{n=1}^{N} \pi_r(1 - \pi_r)$$

This is a quadratic function of the IBD probability π, taking a minimum when the IBD probability equals 0.5 (Figure 8).

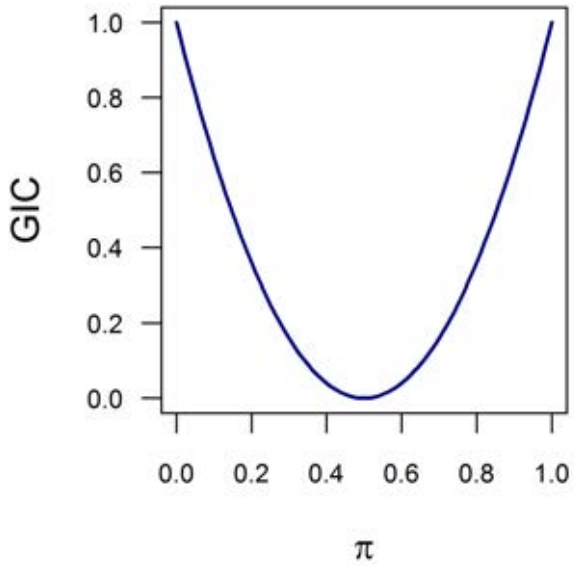
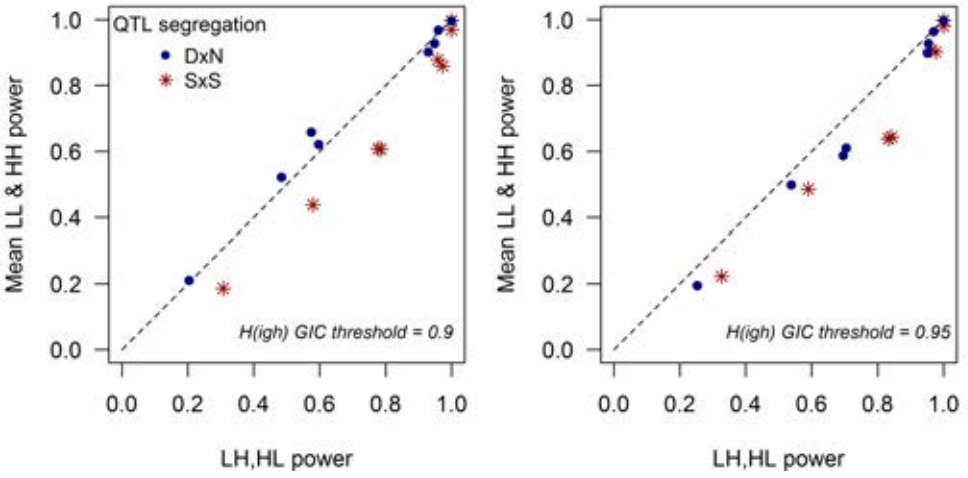
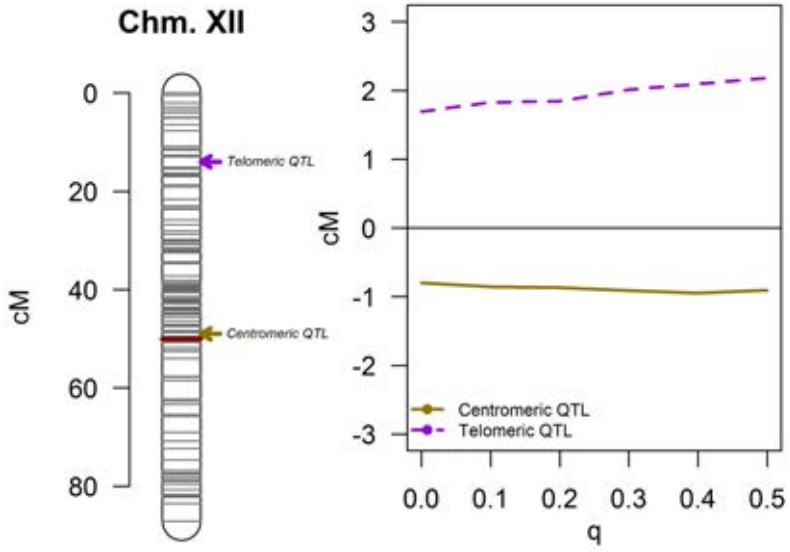


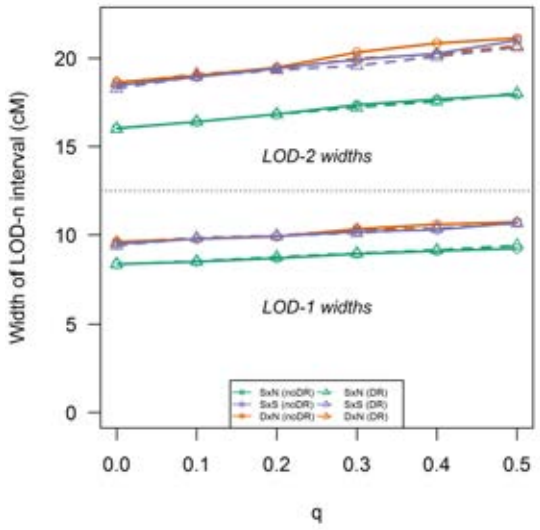

**Figure 8. Relationship between genotypic information coefficient (GIC) and probability of presence of the reference allele (π)**



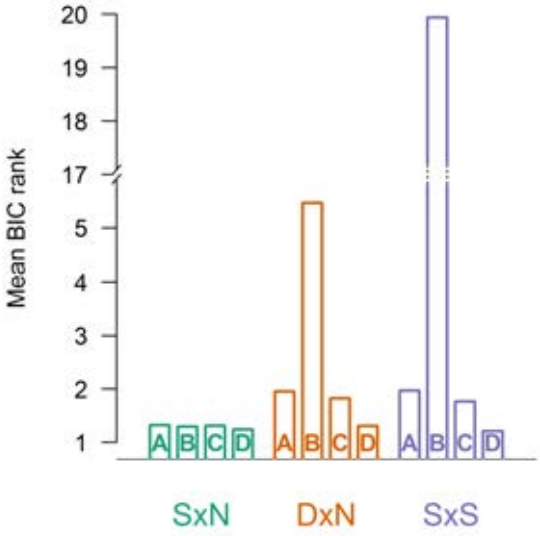

**Supplementary Figure 1.** Prediction of LH or HL detection power from LL and HH detection powers. In general, the power of detection of QTL with a single allele residing on a homologue of high GIC (LH, HL) is approximately mid-way between the detection powers of either LL-type or HH-type QTL.



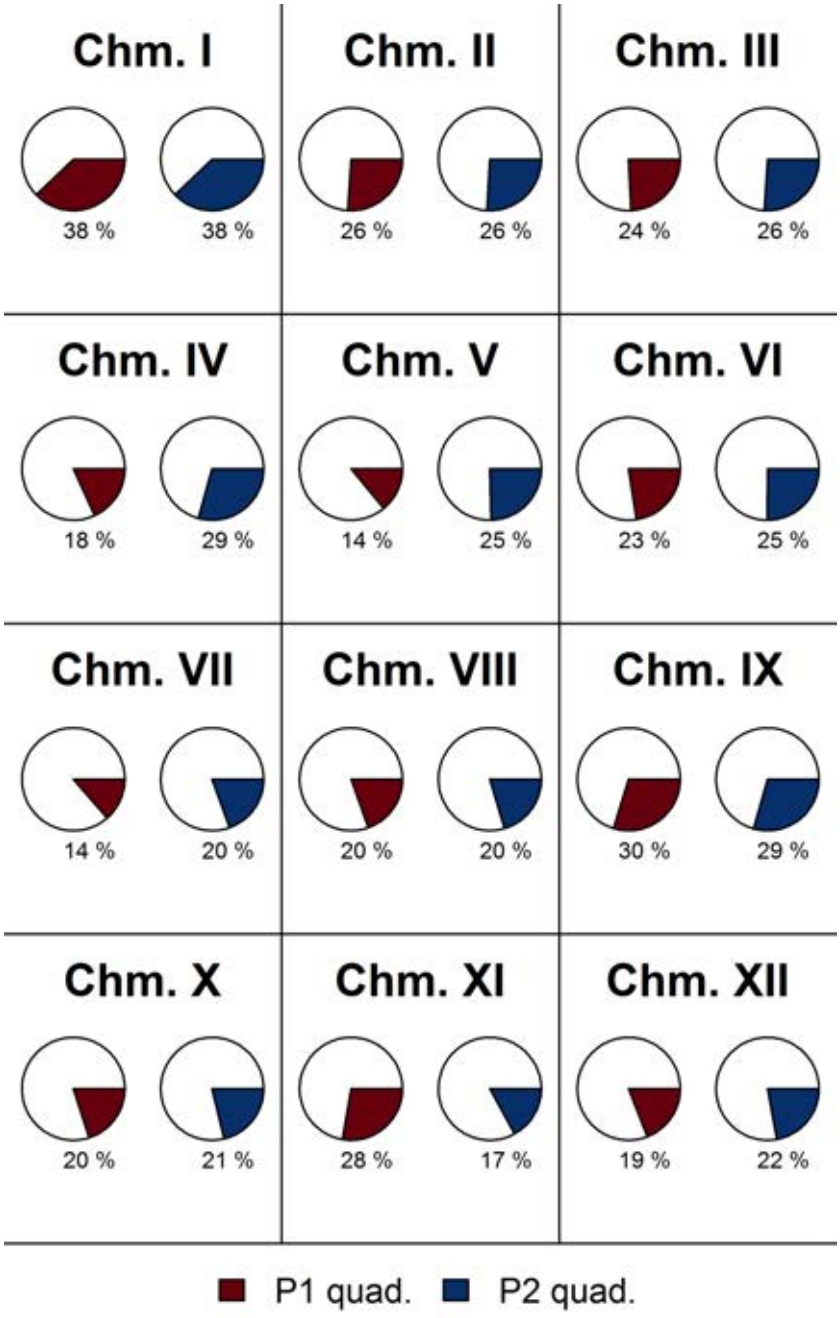

**Supplementary Figure 2.** Bias in the average position of the LOD peak. For both telomeric and centromeric QTL the average position of the QTL peak was biased, although in different directions. Greater levels of quadrivalent pairing ($q \rightarrow 0.5$) introduce extra bias at the telomeres, which is not (or hardly) seen at the centromere.




**Supplementary Figure 3.** Mean width of LOD-1 and LOD-2 confidence intervals (in centiMorgans) around QTL peaks. The size of these confidence intervals was found to increase with higher levels of multivalent formation (*q*), an effect that cannot be mitigated by use of the DR model (dashed lines).




**Supplementary Figure 4.** Split bar-plot showing the performance of the BIC (mean rank) for additive and dominant QTL across different experimental conditions (using full 240 model search). *Codes: A* = Additive noDR; *B* = Additive DR; *C* = Dominant noDR; *D* = Dominant DR. As already seen in Figure 5.a, the mean rank varies for different QTL segregation types. Here we see the cause of the poor performance of the DR model – it occurs only for *additive* 1x1 or 2x0 QTL.




**Supplementary Figure 5**: Predicted rate of quadrivalent pairing from TetraOrigin. Results for Chromosomes 1 to 12 (Chm. I – XII) are shown separately, with parent 1 rates shown in red and parent 2 rates in blue. The rounded percentages are shown beneath each pie chart for clarity.

# Chapter 9

# Multi-environment QTL analysis of plant and flower morphological traits in tetraploid rose

Peter M. Bourke[1*], Virginia W. Gitonga[1,2*], Roeland E. Voorrips[1], Richard G. F. Visser[1], Frans A. Krens[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[2] *Current address*: Selecta Kenya GmbH & Co. KG, P. O. Box 64132 – 00620, Nairobi, Kenya

[*] These authors contributed equally to this work.

*Submitted*

# Abstract

Rose, one of the world's most-loved and commercially-important ornamental plants, is predominantly tetraploid, possessing four rather than two copies of each chromosome. This condition complicates genetic analysis and so the majority of previous genetic studies in rose have been performed at the diploid level instead. However, there may be advantages to performing such analyses at the tetraploid level, not least because this is the ploidy level of most breeding germplasm. Here, we apply recently-developed methods for quantitative trait loci (QTL) detection in a segregating tetraploid rose population ($F_1$ = 151) to unravel the genetic control of a number of key morphological traits. These traits were measured both in the Netherlands and Kenya. Since ornamental plant breeding and selection is increasingly being performed at locations other than the production sites, environment-neutral QTL are required to maximise the effectiveness of breeding programs. We detected a number of robust, multi-environment QTL for such traits as stem and petiole prickles, petal number and stem width and length, that were localised on the recently-developed high-density SNP linkage map for rose. Our work explores the complex genetic architecture of these important morphological traits at the tetraploid level, while helping to advance the methods for marker-trait exploration in polyploid species.



# Key words

Quantitative Trait Locus (QTL) analysis, *Rosa x hybrida* L., autotetraploid, morphological traits, multi-environment trials

# Introduction

Rose (*Rosa x hybrida* L.) is widely considered to be one of the most important ornamental plant species currently cultivated. The genus *Rosa* contains both diploid and tetraploid species with a base chromosome number of seven (2n = 4x = 28). A consistent chromosomal numbering scheme has been proposed, based on an integrated consensus linkage map (ICM) which incorporated markers mapped in a number of different rose mapping populations (Spiller et al., 2011) and which we follow here also. This consistency in chromosomal numbering has facilitated the comparison of results between studies, helping to confirm the position of important sources of genetic variation for a large number of traits of interest.

Among these traits, understanding the genetic basis of plant and flower morphology in rose has been a major research aim in the rose community for many years. Morphological traits (such as plant vigour, presence of prickles on stem or petioles, the number of flower petals *etc.*) have predominantly been investigated at the diploid level (Debener, 1999;Crespel et al., 2002;Shupert et al., 2005;Yan et al., 2005;Hibrand-Saint Oyant et al., 2008;Roman et al., 2015;Li-Marchetti et al., 2017) with far fewer studies conducted at the tetraploid level (Rajapakse et al., 2001;Koning-Boucoiran et al., 2012). This is no doubt due to the complications of genotype calling, linkage mapping and quantitative trait loci (QTL) analysis in autotetraploids, for which methods and software options remain much more limited than for diploids. However, there is an increasing interest in the genetic analysis of autopolyploid species (Voorrips et al., 2011;Hackett et al., 2013;Bourke et al., 2016;Bourke et al., 2017;Hackett et al., 2017;Schmitz Carley et al., 2017), spurred on by ever-decreasing genotyping costs.

The majority of rose cultivars are tetraploid (Smulders et al., 2011) and most breeding work is performed at the tetraploid level (Gar et al., 2011). Indeed, the development of large segregating diploid mapping populations is arguably of little or no commercial importance (Debener and Linde, 2009) which may impose constraints on the size of such populations. Genetic insights gained at the diploid level are often directly applied to related polyploids, but studies are increasingly identifying genetic and epigenetic control mechanisms at the polyploid level that cannot be directly predicted by progenitor diploid species. Newly-formed polyploids may experience homeologue loss and genome restructuring, altered patterns of gene expression as well as transcriptional and epigenetic changes (Soltis et al., 2015). For example in autotetraploid potato, about 10% of a set of 9000 genes were found to be differentially expressed in an experimental autopolyploid series (Stupar et al., 2007). More recently, preferential allele expression for a large set of genes (~3000) was detected among a panel of six autotetraploid potato cultivars (Pham et al., 2017). As modern cultivated cut rose is probably best classified as a segmental

allopolyploid with mainly tetrasomic inheritance (Bourke et al., 2017), it is likely to have derived from the hybridisation of distinct but closely-related progenitor species. As such, insights into both allo- and auto-polyploidisation are potentially relevant. In the well-studied allotetraploid cotton (*Gossypium hirsutum* L.), tetraploids were found to have higher yields and produced a higher-quality fibre than their diploid progenitors grown in the same environment (Jiang et al., 1998). Allohexaploid wheat (*Triticum aestivum*) was found to be significantly more salt-tolerant than either its diploid (*Aegiolops tauschii*) or tetraploid (*T. turgidum*) progenitors. This appears to have been a consequence of combining favourable traits from both parents, but also from polyploid-specific phenomena such as the salt-induced expression of a $Na^+$ transporter HKT1;5 which was transcriptionally reprogrammed following polyploidisation (Yang et al., 2014). Studies like these emphasise that polyploids may possess emergent properties and traits not found at the diploid level. It therefore appears preferable to investigate important breeding traits at the tetraploid level, both to validate previous studies in diploid rose as well as to understand the genetic control of these traits at the ploidy level at which they are usually selected for.

In this study we examined a number of morphological traits such as the number of petals, the presence of prickles on stems and petioles, the stem width and length, the chlorophyll content of leaves and the presence of side shoots. These traits were assessed in different locations, enabling an investigation of possible genotype x environment interactions (Gitonga et al., 2014). This is of particular relevance in modern-day ornamental breeding, where selection and production are often performed in different locations (for details see Gitonga et al. (2014)). Stability of QTL expression over both selection and target environments is needed if marker-assisted selection is to be effectively applied. Here, we re-analyse the data of Gitonga et al. (2014) to perform a QTL analysis using the recently-published high-density tetraploid rose linkage map (containing over 25K single-nucleotide polymorphism (SNP) markers (Bourke et al., 2017)). Our current study helps to increase our understanding of the genetic control of these traits in tetraploid rose, as well as testing and exploring the effectiveness of recently-developed methods (Chapter 8) for the genetic analysis of autopolyploids.

# Materials and Methods

## Plant material and genotyping

The tetraploid "K5" rose population, the result of a cross between contrasting lines "P540" and "P867" was used in this study. P540 (the maternal parent) possesses dark red flowers, has prickles on both stem and petiole and is susceptible to powdery mildew, whereas P867 (paternal) has pale salmon-coloured flowers, few to no prickles and is more resistant to powdery mildew (Koning-Boucoiran et al., 2012;Gitonga et al., 2014). The $F_1$ population resulting from this cross originally comprised of 181 individuals (Yan et al., 2006) but subsequently was found to consist of only 151 unique individuals (Bourke et al., 2017) after genotyping with the 68K WagRhSNP Axiom SNP array (Koning-Boucoiran et al., 2015). The population segregates for quite a number of important traits (including flower colour, for which a QTL analysis has already been performed (Gitonga et al., 2016)), making it suitable for the current study into morphological traits (an example of the range of phenotypes observed for the number of flower petals is shown in Figure 1). Discrete dosage calls (ranging from nulliplex condition (dosage = 0) to quadruplex condition (dosage = 4)) were assigned using the fitTetra package in R (Voorrips et al., 2011) as previously described (Bourke et al., 2017).

**Figure 1. Example of the phenotypic diversity for petal number (and flower colour) in the tetraploid rose K5 population**

## Phenotyping

Plant trials were performed at three locations: Wageningen (WAG) in The Netherlands and at Winchester farm, Nairobi (NAI) and Agriflora farm, Njoro (NJO), two production sites in Kenya. In the Netherlands, observations were made during both the summer of 2007 (WAG_S) and the winter of 2007 / 2008 (WAG_W) whereas in Kenya the observations were made during the period January – July 2009. Full details of plant propagation and growth conditions are described in Gitonga et al. (2014). A description of the traits measured is provided in Supplementary Table 1. Three juvenile traits (date of bending, plant height and plant vigour) were only recorded in the Wageningen summer trial (WAG_S). All traits were measured with at least two replicates, with 4 individual plants per genotype constituting a replicate. There were two completely randomised blocks for both Wageningen trials, and three completely randomised blocks for both Kenyan trials. Best linear unbiased estimates (BLUEs) for traits across environments were calculated using the nlme package (Pinheiro et al., 2017) in R (R Core Team, 2016). Pearson correlations between single-environment traits and multi-environment BLUEs were calculated in R, as well as the frequency distributions of the traits in each of the environments (using the `density` function in R).

## Linkage map construction and QTL analysis

The linkage map used in the current study has already been published, and was created using R scripts developed using previously described methods (Bourke et al., 2016;Preedy and Hackett, 2016;Bourke et al., 2017). Full details of map construction are described in Bourke et al. (2017). The final integrated linkage maps had 25,695 SNP markers (not all unique positions) and covered 55 of the 56 expected parental homologues (the base chromosome number in *Rosa* is 7; each tetraploid parent is expected therefore to have 28 “homologue” maps, resulting in 56 maps across both parents).

A subset of these markers was chosen for the estimation of inheritance probabilities in the population using the TetraOrigin software (Zheng et al., 2016). In an outcrossing autotetraploid there are nine distinct marker segregation types, namely 1x0, 0x1, 2x0, 0x2, 1x1, 1x3, 1x2, 2x1 and 2x2, where the numbers represent the dosage of the marker in the mother and father, respectively. All other marker types (*e.g.* 4x1) can be converted to one of these 9 types without loss or distortion of their linkage information (Bourke et al., 2016). For each 1 centiMorgan (cM) interval, a single marker from each marker segregation type was selected (if possible) which had the lowest number of missing observations across the population. TetraOrigin (Zheng et al., 2016) was run on Mathematica version 10 (Wolfram Research Inc., 2014), allowing both bivalent_decoding options (False / True) in the ancestral inference stage. This generated

identity-by-descent (IBD) probabilities for the population under a double reduction model (which we subsequently refer to by "DR") that allowed for both bivalents and multivalents in the parental meiosis (*i.e.* bivalent_decoding = False), as well as a purely bivalent model ("noDR") for which double reduction is ignored (*i.e.* bivalent_decoding = True). In the latter case, unexpected scores are treated as genotyping errors by the software. The following settings were used: parental dosage error probability (epsF) = 0; offspring dosage error probability (eps) = 0.001; parental bivalentPhasing = True (which assumes bivalent pairing predominates across the population in the determination of parental marker phase). The IBD probabilities at the marker positions were interpolated at 1 cM intervals using the default settings of the `smooth.spline` function in R (R Core Team, 2016)) and saved for subsequent QTL analysis.

A QTL scan was performed using a weighted regression on the trait phenotypes with IBD probabilities from the DR and noDR models as weights. The form of the model used has been described in detail elsewhere (Kempthorne, 1957;Hackett et al., 2013;Hackett et al., 2014), namely:

$$Y = \mu' + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 + \varepsilon$$

where each $X_i$ is an indicator variable for one of the eight parental homologues, having taken the inheritance constraints $\sum_{i=1}^{4} X_i = 2$ and $\sum_{i=5}^{8} X_i = 2$ into account, and weighting by the IBD probabilities as calculated by TetraOrigin.

To simplify the analysis (due to unequal numbers of observations both within and across blocks), we used the mean values per environment in the analysis. The model fit environmental and QTL effects (but not their interaction) by first fitting environment effects and saving the residuals for the QTL scan. For the traits bending time, height and vigour that were measured in the Wageningen summer season alone (WAG_S), a single-environment analysis was performed, using the phenotype values rather than residuals as the dependent variable. Genome-wide significance thresholds per trait were determined using permutation tests by recording the maximum LOD score from each of 1000 genome-wide QTL scans using permuted genotypes, with the 95-quantile of the sorted LOD scores taken as the threshold. Single-environment analyses were also performed to assess the stability of QTL across environments, with significance thresholds per environment and per trait determined using permutation tests as described. To facilitate visualisation, threshold-corrected LOD profiles were determined, assigning a value of 0 to any LOD score below the threshold, and a value of (LOD – threshold) for any score above.

Regions for which the LOD profile of the multi-environment QTL analysis exceeded the significance threshold were re-mapped by saturating the LOD-2 intervals of the QTL peaks with extra markers before re-running TetraOrigin to generate more precise IBD probabilities in the vicinity of QTL. These extra markers were selected as previously described but with a binning window of 0.1 cM in the QTL interval, added to the already selected marker set from the initial QTL scan. We followed the same approach as before for QTL detection (weighted regression with the IBD probabilities as weights), albeit limited to the linkage groups where QTL were originally detected.

QTL peaks from the (marker-saturated) multi-environment analysis were also explored to determine the most likely QTL segregation type and mode of action using the Bayesian Information Criterion (BIC) (Schwarz, 1978) as described by Hackett et al. (2013, 2014). All possible bi-allelic QTL models (assuming additivity or dominance) were tested at each peak, which meant comparing 240 models for each QTL detected. For a polyploid, various dominance models are possible; here we only tested for the simplest case (*i.e.* Aaaa = AAaa = AAAa = AAAA versus aaaa). Any model within 6 BIC of the most likely model (*i.e.* minimising the BIC) was recorded as a potential candidate model. In cases where there were more than five possible QTL models, the QTL segregation type and mode of action were classified as "unclear".

A single-marker analysis of variance (ANOVA) was also conducted for each trait on the marker dosage classes for all mapped markers, with the $-\log_{10}$(p-value) of the model fit used as a proxy for the LOD score. Significance thresholds were determined using permutation tests with N = 1000 and $\alpha = 0.05$ as described above. ANOVA and IBD-based results were visualised together to enable a comparison of the two approaches.

The genotypic information coefficient (GIC) per homologue was calculated using the following formula:

$$GIC_j = 1 - \frac{4}{N}\sum_{n=1}^{N} \pi(1-\pi)$$

where $\pi$ is the inheritance probability of homologue *j* in individual *n* at a particular locus, estimated using the noDR model (Chapter 8).

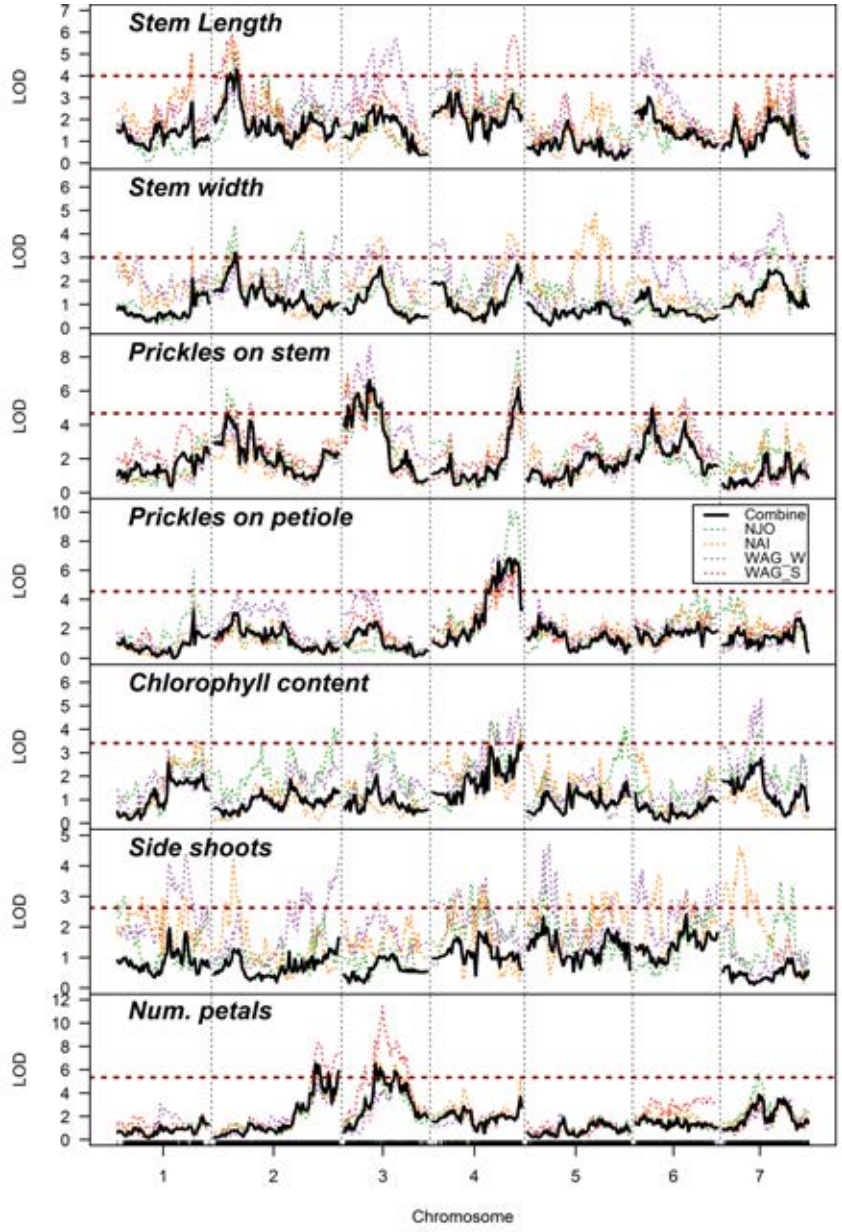



**Figure 2. LOD profiles of QTL scans for morphological traits across different growing environments in the K5 tetraploid rose population.** The results from single-environment analyses are shown as coloured dashed lines, with the multi-environment analysis results as a solid black line ("Combine"). Environments tested were Njoro ("NJO"), Nairobi ("NAI"), Wageningen winter ("WAG_W") and Wageningen summer ("WAG_S"). Significance thresholds for the multi-environment analysis, as determined by permutation tests, are shown as red dashed lines. Note that single-environment LOD profiles (dashed QTL lines) exceeding the multi-environment LOD thresholds are not necessarily significant.

# Results

## Trait correlations across environments

Trait values were in general relatively consistent across environments (Supplementary Figure 1). However for the traits chlorophyll content and side shoots, we observed a poor correlation across environments, with correlation coefficients as low as $R^2 = 0.57$ in the case of the number of side shoots recorded in Nairobi when compared to the multi-environment BLUEs. We also visualised the frequency distributions of the phenotypes across each environment, showing that some traits (like the number of petals) had a very similar distribution of values across environments, whereas other (like chlorophyll content) did not (Supplementary Figure 2).

### Detection of QTL using IBD probabilities

We discovered a number of QTL in this study, although not all traits produced significant peaks (Figure 2). LOD significance thresholds for the multi-environment analyses were in the region of 3.0 – 5.3 as determined by permutation tests. Table 1 summarises the main QTL findings from the initial QTL scan as well as those from the subsequent exploration of the QTL peaks. For most traits, the QTL position, LOD score and explained variances were relatively consistent between the initial QTL scan and the subsequent re-mapping. We were unable to detect QTL for the traits chlorophyll content or number of side shoots, either in the single- or multi-environment analyses (Figure 3). Furthermore, some QTL that were detected in single-environment analyses were not confirmed in the multi-environment analysis (*e.g.* QTLs for stem length on linkage groups 3 and 4 in Wageningen winter and summer, respectively (Figure 3)).

The prediction of QTL segregation type using the Bayesian Information Criterion (BIC) was not particularly clear, since many QTL could be assigned multiple segregation types and modes of action (Table 1). Only the QTL peak for stem prickles on linkage group ICM 6 had unequivocal support for a single model (qqQq x QqQQ with additive action), with as many as 14 potential QTL models identified in the case of the ICM 2 peak for stem width (Table 1). We compared the results of the multi-environment QTL analysis between a model incorporating double reduction (DR model) with one that ignored it (noDR model). In general the results were very comparable, although the model that included double reduction had slightly more detection power in that the QTL peaks rose slightly higher above the threshold (Supplementary Figure 3). Previous simulation work has shown that the deciphering of QTL configuration and mode of action is less accurate when the DR model is used (details in Chapter 8). Therefore, given that the DR model did not alter the QTL results significantly, we opted to only report results from the noDR model for simplicity (whilst still providing a visual comparison in Supplementary Figure

3). For the juvenile traits bending time, plant height and vigour we found two QTL for height on ICM 2 and 6, with no QTL for the other two traits (Supplementary Figure 4). However, since we only had phenotypic data for these traits in a single environment (WAG_S), we were unable to make any prediction about the robustness of these QTL across environments.

**Table 1. Summary of multi-environment QTL detected in this study using noDR IBD probabilities**

| Trait | N | Thr. | ICM | cM | LOD | Expl. Var | Model | Act. | BIC |
|---|---|---|---|---|---|---|---|---|---|
| **Stem Length** | 121 | 4.0 | 2 | 19 (19.8)* | 4.3 (4.6)* | 0.15 (0.16)* | QqQQ x QQqq | *A* | 303.3 |
| | | | | | | | *oooo x QQqq* | *A* | *307.8* |
| | | | | | | | *qqQq x QQqq* | *D* | *308.1* |
| | | | | | | | *QQqQ x QQQq* | *D* | *309.0* |
| | | | | | | | *QQqQ x QQqq* | *A* | *309.1* |
| **Stem width** | 115 | 3.0 | 2 | 18 (18.0) | 3.2 (3.6) | 0.12 (0.14) | *Unclear: 14 models* | | |
| **Prickles on stem** | 121 | 4.7 | 3 | 22 (26.3) | 6.7 (7.8) | 0.22 (0.26) | qqQQ x Qqqq | *D* | 245.9 |
| | | | | | | | *qqQQ x Qqqq* | *A* | *248.9* |
| | | | 4 | 74 (74.5) | 6.2 (6.3) | 0.21 (0.21) | *Unclear: 6 models* | | |
| | | | 6 | 14 (14.4) | 5.0 (4.5) | 0.17 (0.16) | qqQq x QqQQ | *A* | 228.7 |
| **Prickles on petiole** | 121 | 4.5 | 4 | 66 (64.4) | 6.8 (7.2) | 0.23 (0.24) | QQqq x Qqqq | *A* | 83.2 |
| | | | | | | | *qQqq x Qqqq* | *A* | *85.1* |
| | | | | | | | *qQqq x Qqqq* | *D* | *86.5* |
| | | | | | | | *oooo x Qqqq* | *A* | *88.9* |
| **Number of petals** | 120 | 5.3 | 2 | 88 (92.7) | 6.5 (7.3) | 0.22 (0.24) | qqQQ x qqQq | *A* | 342.8 |
| | | | | | | | *qqQQ x QqQq* | *A* | *343.9* |
| | | | | | | | *oooo x qqQq* | *A* | *345.7* |
| | | | | | | | *qqQQ x QQQq* | *A* | *348.2* |
| | | | 3 | 27 (33.7) | 6.5 (6.8) | 0.22 (0.23) | QQQQ x QQqQ | *A* | 345.0 |
| | | | | | | | *QQqQ x QQqQ* | *A* | *346.6* |
| | | | | | | | *QQqq x QQqQ* | *A* | *346.8* |
| | | | | | | | *QQqQ x QQqQ* | *D* | *350.9* |

* Numbers shown are results from initial QTL scan, with those in parenthesis the results following re-saturation of QTL region with additional markers; *N* = number of offspring for which genotype and phenotype data were available for each trait; *Thr.* = experiment-wide LOD significance threshold, determined by permutation test with $N = 1000$ and $\alpha = 0.05$; *ICM* = chromosomal linkage group, using the integrated consensus map (ICM) numbering of Spiller et al. (2011) and Bourke et al. (2017); *cM* = centiMorgan position of QTL peak; *LOD* = logarithm of odds at QTL peak; *Expl. Var.* = fraction of phenotypic variance explained by the QTL model at the peak position; *Model* = QTL models that best fit the data at the QTL position (only those within 6 BIC of the minimum BIC model are shown). "*Unclear*" refers to situations where there were > 5 competing models within 6 BIC of the minimum BIC value (*i.e.* there was no clearly best model(s) suggested). "*Q*" signifies a QTL allele with positive effect effect, "*q*" with a negative effect, and "*o*" a neutral effect (or more precisely, no segregation of alleles originating from that parent); *Act.* = mode of action of QTL model, either additive or dominant; *BIC* = Bayesian Information Criterion (Schwarz, 1978) of the most likely QTL models.

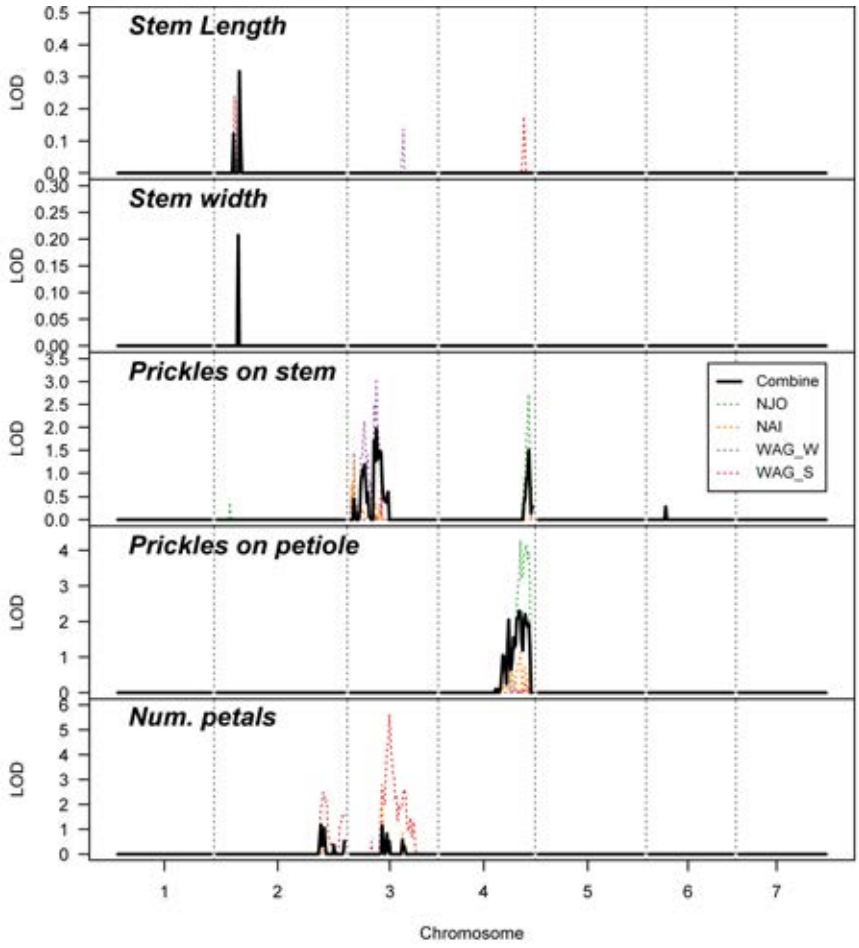



**Figure 3. Threshold-corrected LOD profiles for QTL scans of morphological traits in the K5 tetraploid rose population.** For each trait and each environment (including the multi-environment analysis), significance thresholds were determined using permutation tests, and subtracted from the LOD profiles shown in Figure 2 to show only significant results. The y-axes therefore correspond to the LOD-excess above the significance thresholds. For traits Chlorophyll content and Side shoots, no significant regions of the genome were identified and were therefore not plotted. The single-environment analyses were conducted for trait data from Njoro (NJO), Nairobi (NAI), Wageningen winter (WAG_W) and Wageningen summer (WAG_S).

## Single-marker ANOVA

The single-marker ANOVA analysis on the seven multi-environment morphological traits in this study produced slightly different results to that using the IBD probabilities (Table 2; Supplementary Figure 5). In some cases, the power of QTL detection was slightly higher using the ANOVA approach, for example a single marker "K5520_777" on linkage group ICM 4 was found to be significantly associated with chlorophyll content whereas no QTL were detected for this trait using the IBD model. A single QTL for stem width was identified on ICM 7 with the ANOVA model, as opposed to ICM 2

using the IBD model (which co-localised with the QTL detected for stem length on ICM 2). For the other traits (stem length, prickles on stem or petiole and number of petals) the same set of QTL were detected, albeit with somewhat different peak positions (Tables 1 and 2).

**Table 2. Summary of multi-environment QTL detected in this study using the single-marker ANOVA additive effects model**

| Trait | N | Thr. | ICM | cM | -log10(p) | Expl. Var | Peak marker | Marker phase |
|---|---|---|---|---|---|---|---|---|
| Stem Length | 120 | 4.8 | 2 | 14.6 | 5.0 | 0.15 | M36499_924 | qqqq x QQqq |
| Stem width | 120 | 4.7 | 7 | 47.6 | 5.0 | 0.15 | K14737_275 | D x N |
| Prickles on stem | 120 | 4.8 | 3 | 21.4 | 7.3 | 0.22 | K7826_576 | qqQq x qqQq |
| | | | 4 | 72.2 | 6.7 | 0.20 | K978_98 | S x S |
| | | | 6 | 41.6 | 4.9 | 0.15 | K10843_324 | qQQq x Qqqq |
| Prickles on petiole | 120 | 4.8 | 4 | 66.9 | 8.4 | 0.26 | G38418_730 | QqqQ x qQQq |
| Chlorophyll content | 120 | 4.8 | 4 | 45.9 | 5.1 | 0.16 | K5520_777 | S x S |
| Num. petals | 120 | 4.9 | 2 | 98.3 | 5.8 | 0.18 | G66895_409 | qqqq x qqQq |
| | | | 3 | 31.5 | 6.4 | 0.19 | K599_2377 | qqQq x qqQq |

*N* = number of offspring for which genotype and phenotype data were available for each trait; *Thr.* = experiment-wide significance threshold, determined by permutation test with N = 1000 and $\alpha = 0.05$; *ICM* = chromosomal linkage group, using the integrated consensus map (ICM) numbering of Spiller et al. (2011) and Bourke et al. (2017); *cM* = centiMorgan position of QTL peak; $-\log_{10}(p)$= significance at QTL peak, using the p-value from the model fit; *Expl. Var.* = fraction of phenotypic variance explained by the QTL model at the peak position (adjusted $R^2$ from the regression); *Peak marker* = marker with the highest trait association above the threshold on that linkage group; *Marker phase* = parental marker phase (consistent with homologue numbering from Table 1). In cases where the marker was not phased due to insufficient linkage to 1x0 markers, the segregation type is given instead.

## Comparison with previously-reported QTL

A comparison of our results with those of previous studies showed that some QTL had already been identified in other populations on the same linkage groups as we identified, while some QTL differed (Table 3). For example, an earlier study in the diploid 94/1 population found a QTL for both stem length and width on ICM 2 (Yan et al., 2007), although they also identified QTL on ICM 1 and 5 for these traits which we did not. For the trait stem prickles we identified QTL on ICM 3, 4 and 6, whereas previously stem prickle QTL have been detected on ICM 2, 3 and 4 (Crespel et al., 2002;Linde et al., 2006;Koning-Boucoiran et al., 2012). The QTL on ICM 2 as reported by Koning-Boucoiran et al. (2012) was found using the same K5 mapping population (and same phenotype data) as here (although with a different type of analysis and different genotypes). There, the trait was found to be associated with linkage groups A2-2 and A2-3, corresponding to our ICM 2 (as well as A3-1, corresponding to ICM 3) in parent 1. In our study, there was a rise in LOD significance values on ICM 2 at the same position for this trait, although falling just below the significance threshold (Figure 2).

**Table 3. Overview of previous QTL detected for the morphological traits in this study**

| Trait | Pop. | N | Ploidy | LG | Reference |
|---|---|---|---|---|---|
| Stem length | 94/1 | 88 | 2x | 1 | |
| | | | | 2 | |
| | | | | 5 | (Yan et al., 2007) |
| Stem width | 94/1 | 88 | 2x | 1 | |
| | | | | 2 | |
| | | | | 5 | (Yan et al., 2007) |
| Stem prickles | K5 | 184 | 4x | 2 | |
| | | | | 3 | (Koning-Boucoiran et al., 2012) |
| | HW | 91 | 2x | 4 | (Crespel et al., 2002) |
| | 97/7 | 270 | 2x | 3 | (Linde et al., 2006) |
| Petiole prickles | 90-69 $F_2$ | 52 | 4x | 6* | (Rajapakse et al., 2001) |
| | 90-69 $F_2$ | 52 | 4x | 6* | (Zhang et al., 2006) |
| Chlorophyll content | 94/1 | 88 | 2x | 2 | |
| | | | | 3 | |
| | | | | 6 | |
| | | | | 7 | (Yan et al., 2007) |
| Num. petals | 94/1 | 60 | 2x | 3 | (Debener, 1999) |
| | HW | 91 | 2x | 6 | (Crespel et al., 2002) |
| | 97/7 | 270 | 2x | 3 | (Linde et al., 2006) |
| | HW | 91 | 2x | 4 | (Hibrand-Saint Oyant et al., 2008) |
| | K5 | 184 | 4x | 3 | (Koning-Boucoiran et al., 2012) |

*N* = mapping population size used; *Ploidy* of mapping population used, either diploid (2x) or tetraploid (4x); *LG* = linkage group numbering according to the integrated consensus map (ICM) numbering of Spiller et al (2010).

* In the case of the Rajapakse et al. (2001) and Zhang et al. (2006) studies, the numbering was imputed through linkage with microsatellite markers (see main text for details).

## Genotypic Information Coefficient (GIC)

The genotypic information coefficient gives a measure of the level of information provided by marker data on a scale from 0 to 1, with 0 corresponding to no information, and 1 to full information (and is homologue-specific in our formulation). When based on IBD probabilities, GIC can be thought of a measure of how certain we are about the inheritance of a parental homologue at a particular position, averaged across the population. Most GIC values exceeded 0.9, reflecting the dense marker coverage across all parental homologues. An example of the GIC profile for linkage group ICM 3 is shown in Figure 4. The telomeric region of homologue h7 of parent 2 had additional marker coverage when compared to the other parental homologues (due to a 10 cM stretch of simplex x nulliplex markers which mapped there). The GIC content on all other seven homologues steadily decreases towards the telomere in this region, while for homologue h7 it remains close to 1. Dips in the GIC can also be seen to correspond to

areas of lower marker density along the chromosome (Figure 4). The GIC profiles for all seven rose linkage groups are provided in Supplementary Figure 6.

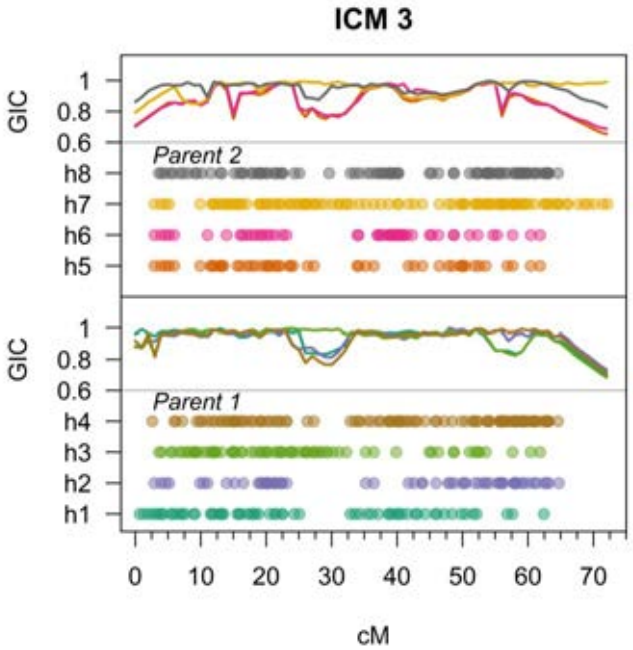



**Figure 4. Example of the distribution of marker alleles across the eight parental homologues on linkage group ICM 3, with the per-homologue genotypic information coefficient (GIC) plotted above.** The lower section shows the four homologues of Parent 1, the upper section those of Parent 2. GIC values were found to be in the range 0.6 – 1 (approximately) for both parents, with higher GIC values indicating greater amounts of genetic information for that homologue.

# Discussion

## Multi-environment QTL

QTL studies often report QTL without necessarily considering whether these QTL will be useful in practice. One of the more important considerations is whether the reported QTL are robust across environments or are only expressed in a specific environment. This is particularly relevant for modern rose breeding, for which the selection and production environments often differ (Gitonga et al., 2014). In this study we analysed phenotypic data from three locations, two in Kenya and one in the Netherlands, with an extra set of phenotypes evaluated during both a summer and winter season in the Netherlands. We were most interested in identifying QTL that are effective across both the selection and target environment. We found that a number of QTL were environment-specific, such as the QTL for stem length present on linkage groups ICM 3 and 4, but most QTL were identified in both the Kenyan and Dutch trial data (Figure 3). The implication is that for the traits studied here, selection in a non-target environment should

also produce rose cultivars that express the desired traits in the production environment. The previous analysis of Gitonga et al. (2014) reported high heritabilities for these traits across environments, suggesting either that there was very little environmental noise in the data, or that the growing environment had little impact on the expression of the phenotype. Among the traits with the lowest reported heritabilities were chlorophyll content and number of side shoots (Gitonga et al., 2014), both traits for which no QTL were detected in the current analysis. These traits both showed a lower consistency across environments (Supplementary Figure 1), as well as large genotype x environment (GxE) interaction effects (Gitonga et al., 2014), which we did not attempt to model in our QTL analysis. Selection of traits with lower heritabilities is by definition less effective than that of highly-heritable traits (according to the "breeder's equation"); those that also show significant GxE interactions should be selected for at the target environment where the crop will ultimately be produced.

## Consistency with previous studies

Many of the traits examined in the current study have previously been mapped in other populations (mainly, but not exclusively at the diploid level). Table 3 summarises the main QTL identified by other groups for the morphological traits examined here. Attempting to compare results from previous studies relies on consistent linkage group numbering; the work of Spiller et al. (2010) has done much to improve this situation, but not all previously published maps were included in that study.

Somewhat unexpectedly, there were differences in our detection of QTL for stem prickles (on ICM 3, 4 and 6) and those of a previous study (on linkage groups A2 and A3, corresponding to ICM 2 and 3) that used the same population and phenotypic data but different genotype data (Koning-Boucoiran et al., 2012). The main difference in these studies was the amount of marker data, which went from 469 mapped AFLP and SSR markers to 25,695 mapped SNP markers. This greater coverage probably enabled the detection of QTL on (parts of) linkage groups that were not covered by markers in the initial study (like the QTL peaks on ICM 4 and 6 that we detected). The QTL we missed on ICM 2 could be localised to the region 14 – 23 cM by assigning approximate map positions to markers Rh91-135-1 and P11M55-237-1 (linked to the QTL in the study of Koning-Boucoiran et al. (2012)). These approximate map positions were provided in the supplementary material Table S5 of Bourke et al. (2017). A definite rise in significance was detected in this region (Figure 2), although not exceeding the significance thresholds. In order to ensure the robustness of the reported QTL we used the permutation test procedure with relatively stringent genome-wide thresholds (Churchill and Doerge, 1994), which may have resulted in some minor QTL being missed (Type II errors).

Both Rajapakse et al. (2001) and Zhang et al. (2006) reported finding a QTL for petiole prickles in a tetraploid $F_2$ mapping population on their linkage group "B7" (Table 3). Zhang (2006) provided the primer sequences of microsatellite markers Rw55D22 and Rw5D11 which were reportedly linked to this trait (Zhang et al., 2006). A BLASTn of these primer sequences on the *Fragaria vesca* genome produced hits on chromosome *Fv*2, which shares synteny with rose ICM 6 (Bourke et al., 2017). Intriguingly, the marker Rw5D11 was previously used in genotyping the K5 mapping population (Koning-Boucoiran et al., 2012) and was mapped on LG 4, which would be consistent with our result of a single major QTL for petiole prickles on ICM 4 (Figure 2). However, this marker (Rw5D11) showed greatest linkage to simplex markers on ICM 6 (Bourke et al. (2017), Table S5: recombination frequency $r \leq 0.09$ and LOD $\geq 20.4$), consistent with the BLASTn results. Indeed, a more detailed look at the AFLP and SSR maps of Koning-Boucoiran et al. (2012) with the SNP maps of Bourke et al. (2017) shows that linkage groups A4-1, A4-2 and A4-3 should probably have been assigned to ICM 6, and A4-4 to ICM 4 (in other words, the homologues of linkage group A4 do not appear to form a single chromosomal linkage group). It is therefore unclear where the QTL for petiole prickles of Rajapakse et al. (2001) and Zhang et al. (2006) should be placed, on ICM 4 or ICM 6. These sorts of examples demonstrate the pressing need for a reliable physical sequence in *Rosa*, so that reported QTL positions can be traced and quickly checked against the reference physical map to help compare and integrate results from different studies.

Despite these hurdles, we were able to confirm many of the previously-reported QTL in the present study, as well as providing robust evidence for some previously-unreported QTL for these important morphological traits (*e.g.* the QTL for stem prickles on ICM 6 and the QTL for petal number on ICM 2). In general, the results from diploid studies tally well with our results from a tetraploid analysis (*e.g.* Linde et al. (2006) or Debener and Mattiesch (1999)). Conversely, many of the QTL reported from the diploid 94/1 population by Yan et al. (2007) were not detected in this study (*e.g.* the QTL for stem length and width on linkage groups ICM 1 and 5, or the four separate QTL they detected for chlorophyll content (Table 3)). This may suggest that either large differences in the genetic control of some morphological traits could exist between ploidy levels, or (more likely) that QTL for these traits segregated in their population whilst not in ours. Another possible cause is differences in the methodologies of QTL detection and particularly significance threshold setting, which may have led to either inflated Type I or Type II errors between their study and this one, respectively.

As long as rose breeding and production continues to be performed at the tetraploid level, QTL studies at the tetraploid level continue to have clear advantages over diploid studies.

For example, marker assays that have been identified at the diploid level may not necessarily perform well at the tetraploid level (*e.g.* due to unknown flanking SNPs, off-target hybridisation *etc.*). Tetraploid offspring with desirable combinations of traits can be directly used in further breeding work (or may represent a finished variety) whereas there is less use for offspring of a diploid cross. Furthermore, although there are more tools available for diploid genetic analysis, those for tetraploid studies are becoming increasingly sophisticated (*e.g.* TetraploidSNPMap (Hackett et al., 2017) or the methods described here). We therefore believe that genetic analyses at the polyploid level will become increasingly commonplace in the future.

## Phasing QTL

One of the principal advantages of QTL analyses using IBD probabilities in polyploids is that QTL detection is performed across all homologues simultaneously. Single-marker approaches rely on coupling linkage between marker alleles and QTL alleles, and may be less powerful due to "genotypic noise" from other homologues if the marker is not in complete coupling phase with the QTL alleles. This becomes even more important at higher ploidy levels where there are even more QTL configurations possible, decreasing the likelihood of full coupling linkage between a QTL and a nearby marker. In this study we performed both an IBD-based as well as single-marker ANOVA QTL analysis. The results from both approaches were quite consistent (as a comparison of Tables 1 and 2 shows). However, we did find a single significant marker ("K5520_777" on linkage group ICM 4) which was associated with chlorophyll content through the ANOVA analysis, with no corresponding peak using the IBD-based approach. It is possible that noise in the IBD probabilities near the QTL adversely affected the detection power (the GIC at 46 cM (marker "K5520_777" is located at 45.9 cM) for homologues h2, h4 and h5 were ≈ 0.858 , 0.843 and 0.895 respectively, below the chromosome-scale averages of 0.933, 0.908 and 0.935, although the GIC of the other homologues were all above 0.9). On the other hand, we detected a QTL for stem width on ICM 2 using the IBD model which was not picked up in the single-marker analysis. This ICM 2 peak co-localised with the QTL detected for stem length on ICM 2, increasing the confidence that an important growth-determining factor is located in that region.

Apart from detection power, IBD-based analyses also offer the possibility to determine the most likely QTL segregation type and mode of action, which can only be guessed at using single-marker approaches (*i.e.* by looking at the segregation type of the most significant marker, Table 2). The Bayesian Information Criterion (Schwarz, 1978) has previously been shown to correctly identify the QTL model in polyploid QTL studies, with prediction accuracies of up to 90% for simplex QTL (details in Chapter 8). We were therefore somewhat disappointed with the results obtained in this study for QTL

segregation type and mode of action (Table 1). In only one case (the QTL on ICM 6 for stem prickles) did we have a clear indication of the most likely QTL phase, namely a simplex x triplex additive QTL segregating as qqQq x QqQQ. Phased QTL information could help improve the accuracy of marker-assisted selection, where breeders would no longer select individuals based on single marker alleles, but on desirable combinations of parental haplotypes (through a combination of specific SNP alleles). Haplotype-assisted selection in polyploids has yet to be applied (to our knowledge) but would require much clearer QTL phasing than we were able to achieve in this study to be able to reliably construct a potential QTL-tagging haplotype-marker. The possible causes for the uncertainty are many – conflicts in genotype information, inaccurate linkage maps, more complex (or multi-allelic rather than bi-allelic) QTL types, poorly-scored or confused phenotype data *etc.* Despite this, our approach of saturating QTL regions with extra markers tended in general to increase the significance and proportion of explained variances at QTL positions (Table 1), with a single exception that may have been due to a sub-optimal phasing solution returned by TetraOrigin (which is a stochastic rather than deterministic procedure (Zheng et al., 2016)). High genotypic information coefficients around QTL positions on the relevant homologues have been shown to be necessary for the accurate and clear detection of QTL as well as the precise estimation of their location and phase (Chapter 8) and need to be taken into account if QTL studies in polyploids are to deliver applicable results.

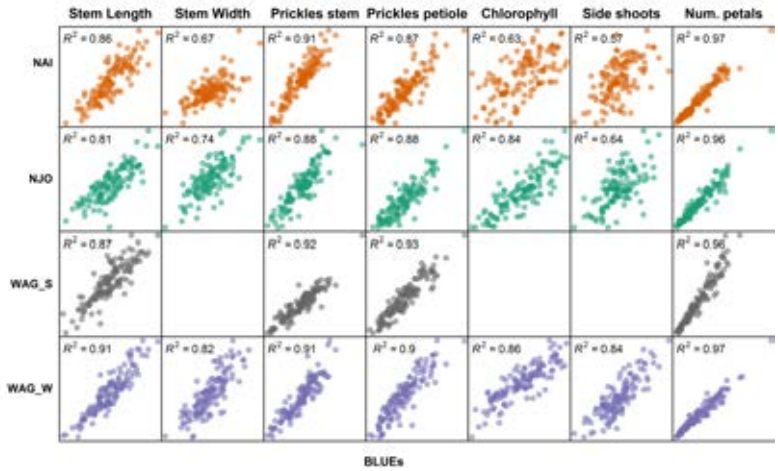
## Concluding remarks

In this study we applied recently developed methods to help unravel the genetic architecture underlying a number of important morphological traits in tetraploid rose. Most of the detected QTL were found to be robust across environments, suggesting that selection for these traits can successfully be performed in locations other than the production site. However, traits displaying significant genotype x environment interaction effects should probably not be selected for in another than the target location. Apart from identifying and confirming an important set of QTL at the tetraploid level, our work helps pave the way towards haplotype-assisted selection methods in the future.

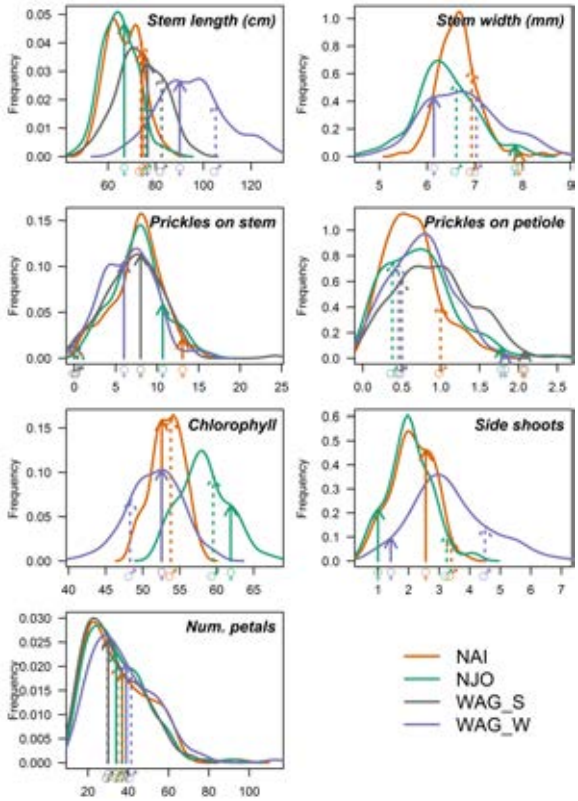
# Acknowledgements

Funding for this research was provided through the TKI polyploids project “Novel genetic and genomic tools for polyploid crops” (project number BO-26.03-009-004). The support of the companies participating in the polyploids projects is gratefully acknowledged. Terra Nigra B.V. is acknowledged for generating the K5 mapping population and making it available.

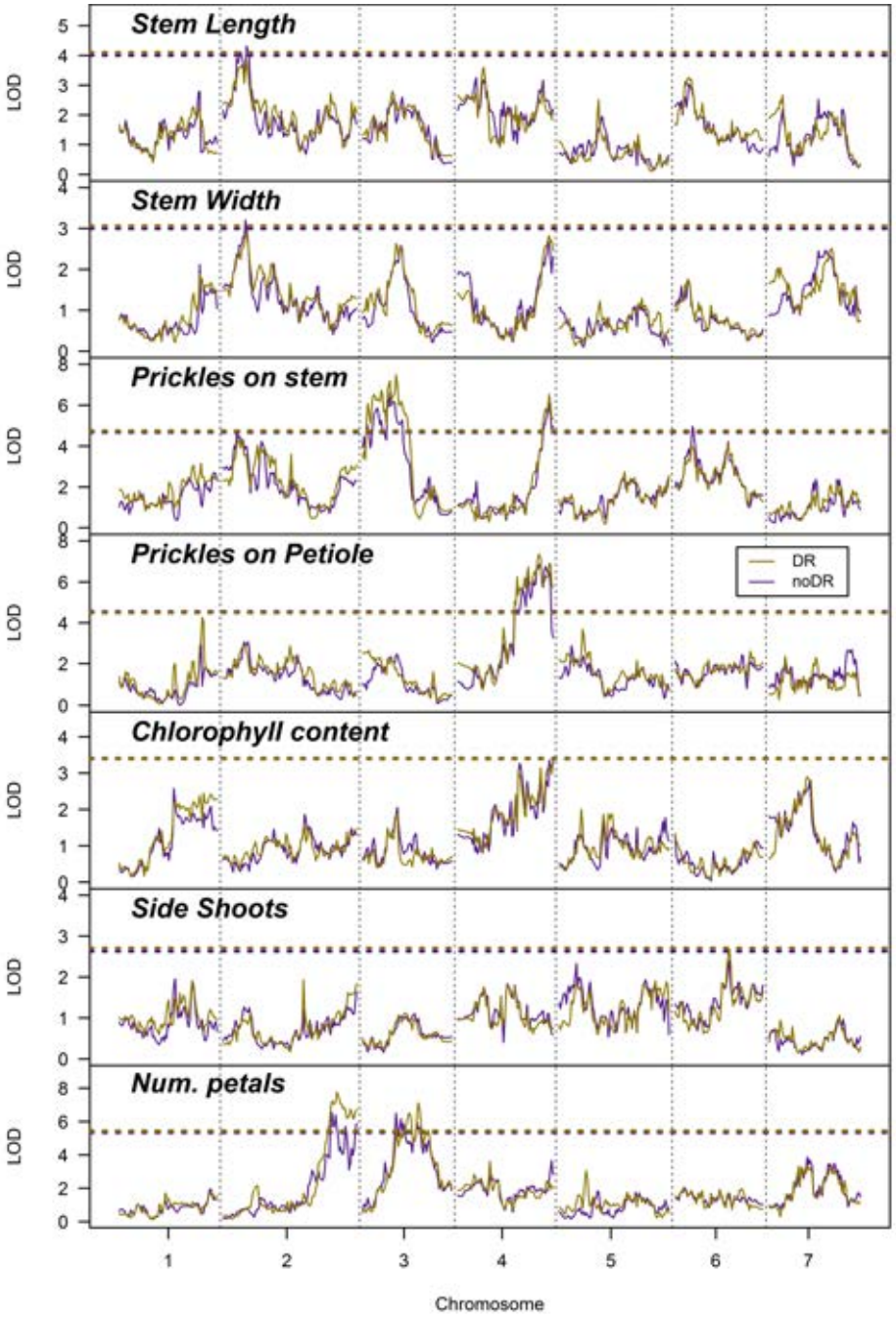
*Supplementary data will accompany the online version of the published article.*



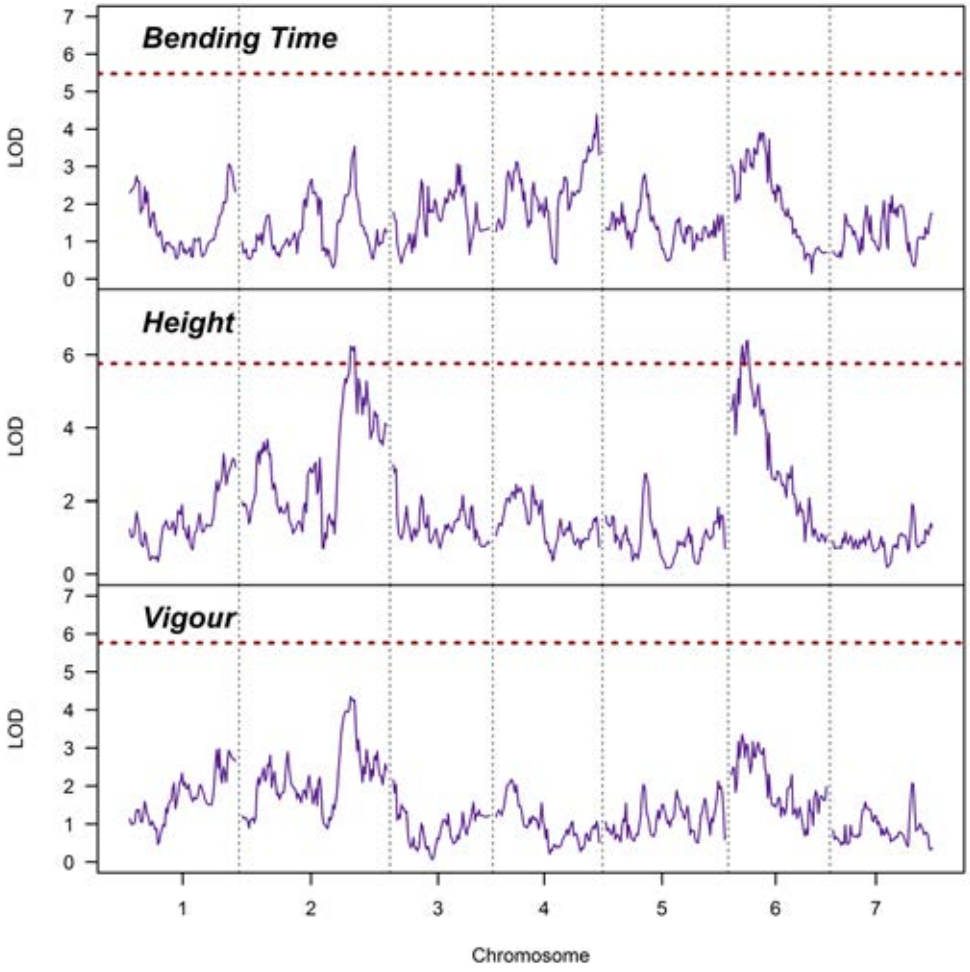

**Supplementary Figure 1.** Comparison between single environment trait values and Best Linear Unbiased Estimates (BLUEs) across environments for each of the seven morphological traits recorded in this study. The four environments are shown on the left-hand margin, namely Nairobi (NAI), Njoro (NJO), Wageningen summer (WAG_S) and Wageningen winter (WAG_W). The squared Pearson correlation coefficients ($R^2$) are printed for each comparison. BLUEs values (x-axis) are equal across the four environmental comparisons.



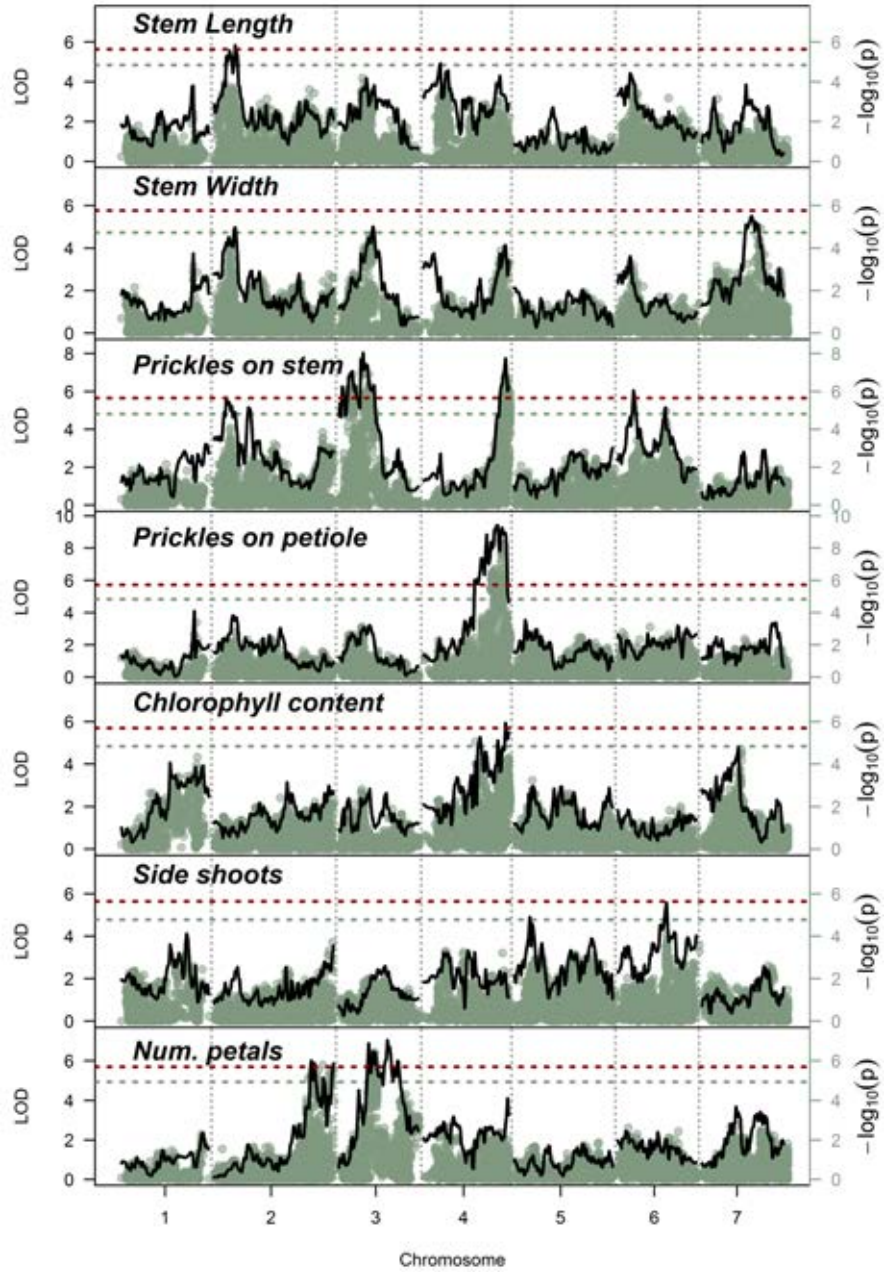

**Supplementary Figure 2.** Distribution of rose morphological trait values over four different environments (Nairobi (NAI), Njoro (NJO), Wageningen summer (WAG_S) and Wageningen winter (WAG_W)). Mean parental values are shown as either solid vertical arrows (maternal mean, ♀) or dashed vertical arrows (paternal mean, ♂).



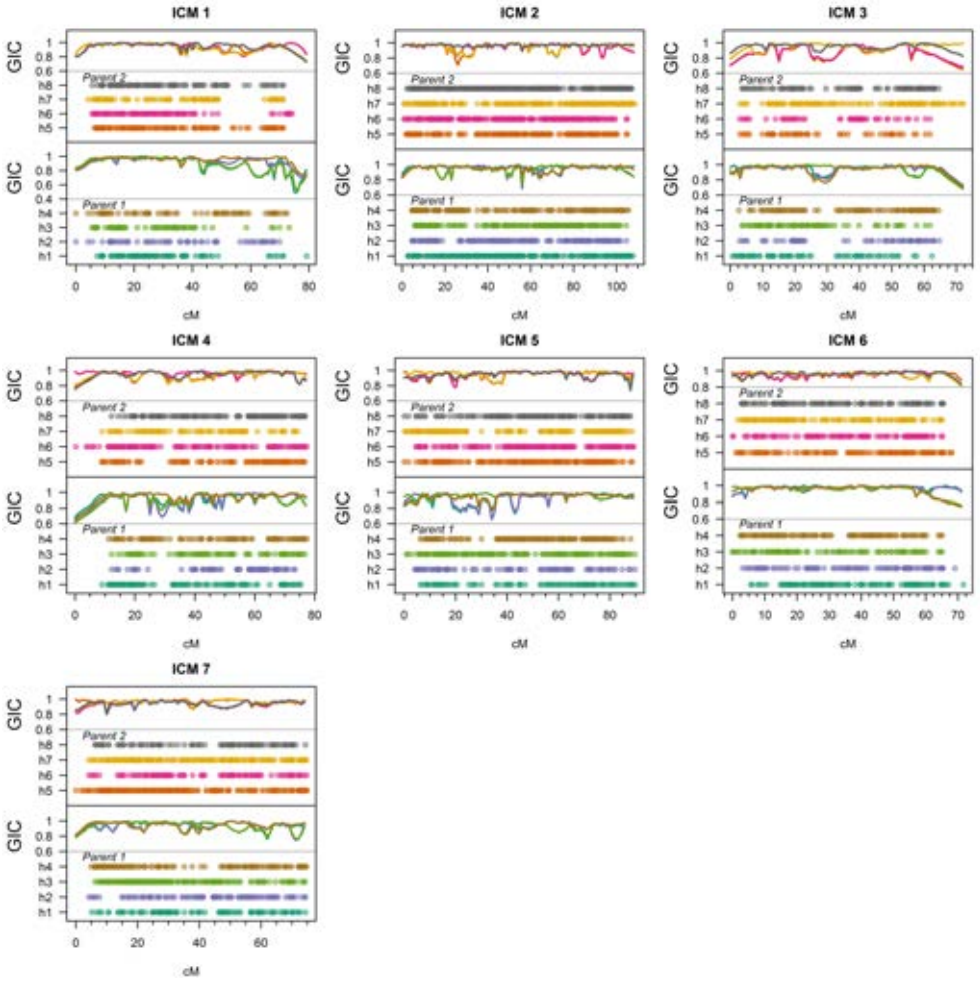

**Supplementary Figure 3.** Comparison between multi-environment QTL analysis for the seven multi-environment morphological traits studied, using a model that ignores double reduction (noDR) versus one that includes it (DR). Significance thresholds (as determined by permutation tests) are shown as dashed lines, with little difference found between thresholds for the noDR and DR models.




**Supplementary Figure 4.** Single-environment QTL analysis results for the traits bending time, plant height and plant vigour, assessed in Wageningen (WAG_S) only. Significance thresholds (as determined by permutation tests) are shown as dashed lines; the analysis shown used the noDR model which assumes random bivalent pairing during meiosis only.




**Supplementary Figure 5.** Comparison between IBD-based QTL analysis results (black line) with a single-marker ANOVA results (green dots) for the seven multi-environment morphological traits studied. Significance thresholds (as determined by permutation tests) are shown as dashed lines, with the red line corresponding to the IBD-based thresholds, and light green line the ANOVA-based thresholds. Significance values for the IBD-based analysis are given on the left-hand axes (LOD scores) with the ANOVA significance values on the right-hand axes ($-\log_{10}(p)$ values).




**Supplementary Figure 6.** Genotypic Information Coefficient (GIC) plots for rose linkage groups 1 – 7, visualising the information content per homologue of the IBD probabilities used in the initial QTL scan. Marker allele distributions for Parent 1 on homologues h1 – h4 are shown in the lower section, with marker allele distributions for Parent 2 (homologues h5 – h8) shown in the upper section. GIC values were found to be in the range 0.6 – 1 (approximately) for both parents, with higher GIC values indicating greater amounts of genetic information for that homologue. The colouring scheme of the GIC per homologue and the marker distribution on that homologue are identical.

# Chapter 10

# polyqtlR – an R package to analyse quantitative trait loci in autopolyploid populations

Peter M. Bourke[1], Roeland E. Voorrips[1], Christine Hackett[2], Geert van Geest[1,3], Richard G. F. Visser[1], Chris Maliepaard[1]

[1] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.

[2] Biomathematics and Statistics Scotland, Invergowrie, Dundee DD2 5DA, UK.

[3] Deliflor Chrysanten B.V., Korte Kruisweg 163, 2676 BS Maasdijk, The Netherlands.

# Abstract

The investigation of quantitative trait loci (QTL) is an essential component in our understanding of how organisms develop and respond to their environment through the identification of genes and their alleles that contribute to trait variation. This knowledge can help guide and accelerate breeding programs aimed at developing superior genotypes through genomics-assisted breeding. Here we present polyqtlR, a new software tool to facilitate QTL analysis in autopolyploids that performs QTL interval mapping in $F_1$ populations of outcrossing autopolyploids and segmental allopolyploids using identity-by-descent (IBD) probabilities. The allelic composition of discovered QTL can be explored, enabling favourable alleles to be identified and tracked in the population. Visualisation tools within the package facilitate this process, and options to include genetic co-factors (QTL positions or markers), covariates, randomised blocks or different environments are included.



# Key words

Quantitative Trait Locus (QTL) analysis, polyploidy, identity-by-descent (IBD) probability, interval QTL mapping

# Introduction

Polyploids, which carry more than the usual two copies of each chromosome, are an important group of organisms that occur widely among plants, particularly domesticated ones (Salman-Minkov et al., 2016). Many theories to explain their prevalence among crop species have been proposed, identifying features which may have appealed to early farmers in their domestication of wild species. Such features include their larger organs such as tubers, fruits or flowers (the so-called "gigas" effect) (Sattler et al., 2016), increased heterosis (Comai, 2005), their genomic plasticity (te Beest et al., 2011), phenotypic novelty (Udall and Wendel, 2006), their ability to be clonally propagated (Herben et al., 2017), increased seedling and juvenile vigour (Levin, 1983), the masking of deleterious alleles (Renny-Byfield and Wendel, 2014) or the possibility of seedlessness which accompanies aneuploidy (Bradshaw, 2016). It is currently believed that all flowering plants have experienced at least one whole genome duplication (WGD) during the course of their evolution, with many lineages undergoing multiple rounds of WGD followed by re-diploidisation (Vanneste et al., 2014). Polyploidy may also be induced deliberately (usually through the use of some chemical cell division inhibitor such as colchicine (Blakeslee and Avery, 1937)), often to combine properties of parents that could not otherwise be crossed (Van Tuyl and Lim, 2003), or to benefit from some of the other advantages listed above.

However, not all features of polyploids are necessarily advantageous. Perhaps one of the greatest difficulties in polyploid cultivation and breeding is the constant re-shuffling of alleles in each generation, a consequence of polysomic inheritance. We are currently witnessing a genomics revolution which could hold the key to understanding and tracking polyploid inheritance in detail. From a breeding perspective, we would like to be able to identify genomic regions that contribute favourable alleles to a particular trait of interest (quantitative trait loci (QTL)), and predict which offspring in a population carry favourable combinations of parental alleles. The software tools to perform such analysis at the polyploid level are still relatively rare – here, we introduce a novel R package, polyqtlR, for performing these analyses in outbreeding populations of autopolyploid and segmental allopolyploid species.

The analysis of QTL within polyqtlR can either be single marker-based, or identity-by-descent (IBD) probability-based. For single marker analyses, the dosage of each bi-allelic marker is used as the explanatory variable in a genome-wide scan. In polyploid species, marker dosage refers to the count of marker alleles (by convention the count of the "alternative" allele at a bi-allelic locus, as opposed to the "reference" allele). In an autotetraploid for example, the possible dosages range from nulliplex (0 copies of the alternative allele), simplex (1 copy), duplex (2 copies), triplex (3 copies) to quadruplex

(4 copies). The assignment of marker dosage in polyploids is a non-trivial problem in itself, but there are a number possibilities for achieving this using dedicated software (Gidskehaug et al., 2010;Voorrips et al., 2011;Serang et al., 2012;Schmitz Carley et al., 2017).

On the other hand, identity-by-descent (IBD) probabilities are the inheritance probabilities of parental alleles in a mapping population. An algorithm employing Hidden Markov Models was recently proposed for tetraploids and implemented in the software program TetraOrigin (Zheng et al., 2016). polyqtlR can utilise the output of TetraOrigin for the analysis of quantitative traits. Alternatively, a more simplistic algorithm to approximate IBD probabilities was developed within our lab (Bourke, 2014) and has also been implemented in polyqtlR. This algorithm has the advantage of being applicable to all ploidy levels as well as being computationally very efficient, and was recently used in the analysis of several traits in hexaploid chrysanthemum (van Geest et al., 2017a).

polyqtlR offers the possibility to explore the possible allelic composition at QTL positions and to track the transmission of such alleles within the mapping population. Significance thresholds can be determined using permutation tests, with parallelised calculations to improve performance. The inclusion of genetic co-factors, phenotypic co-variates or co-factors to specify experimental design (*e.g.* blocking factors) can increase the power of detection, as well as allowing for possible interactions between QTL to be detected. As with most new tools, polyqtlR will continue to be developed to deal with advances in genotyping technologies as well as methodologies for QTL analysis in polyploid populations. That said, the current implementation of polyqtlR should already have a significant positive impact on the breeding of polyploid crops.

# Implementation

polyqtlR requires dosage-scored marker information with an accompanying integrated linkage map from an $F_1$ population. Linkage maps can be generated for tetraploid populations using software such as TetraploidSNPMap (Hackett et al., 2017) or polymapR (Chapter 6). For triploid or hexaploid populations polymapR is recommended as TetraploidSNPMap does not currently handle these ploidy levels. In the case of tetraploids, linkage map phase and IBD probabilities can be estimated externally using the TetraOrigin software (Zheng et al., 2016). Currently, the only public implementation of TetraOrigin is written in the proprietary software Mathematica (Wolfram Research Inc., 2014), although an R version (R Core Team, 2016) may also become available in

future. Otherwise, IBD probabilities can be estimated within the package using a computationally efficient internal algorithm that uses the phased map output of polymapR.

## Estimation of IBD probabilities

Details of the methodology behind the calculation of IBD probabilities in TetraOrigin can be found in Zheng et al. (2016). The internal algorithm for estimating IBD probabilities in polyqtlR currently relies on a method originally described in Bourke et al. (2014) and re-implemented by van Geest et al. (2017). One pre-requisite to calculating IBD probabilities within polyqtlR is a fully-phased linkage map such as those produced by polymapR (Chapter 6). In a hexaploid species, for example, such a map would be of the form shown in Table 1.

**Table 1. Example of a phased linkage map in a hexaploid species.**

| Marker | LG | position | h1 | h2 | h3 | h4 | h5 | h6 | h7 | h8 | h9 | h10 | h11 | h12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ps_I1.2 | 1 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Ac_I11.7 | 1 | 6.57 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| Zm_I3.25 | 1 | 9.73 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Ap_I21.3 | 1 | 14.41 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Zm_I19.8 | 1 | 18.30 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| Zm_I24.9 | 1 | 23.31 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Zm_I33.55 | 1 | 30.26 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ap_I33 | 1 | 33.60 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Ps_I42.65 | 1 | 37.86 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| St_I47.05 | 1 | 42.24 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |

As well as the marker position on an integrated map (columns "LG" for linkage group and "position" in centiMorgans), the configuration of marker alleles across the twelve parental homologues is also shown (h1 – h6 for parent 1, h7 – h12 for parent 2). Generally a "1" corresponds to the presence of the segregating allele, although if the original marker coding is used, "1" corresponds to the alternative and "0" to the reference allele.

IBD probabilities of 0.5 are initially assigned to the population for all parental homologues at all marker positions (completely uninformative priors). Fully-informative marker scores are then used to assign probabilities of 1 in the case of inheritance, or 0 in the case of no inheritance. For example, in a tetraploid population at a 1x0 marker (single segregating allele originating from parent 1) which was phased 0001 | 0000, offspring with marker dosage 1 would initially be assigned probabilities of (0.5,0.5,0.5,**1** | 0.5,0.5,0.5,0.5), and offspring with marker dosage 0 would initially be assigned probabilities of (0.5,0.5,0.5,**0** | 0.5,0.5,0.5,0.5) (normalisation of probabilities occurs later). Partially-informative dosages in the offspring are not used (*e.g.* an offspring dosage of 1 from a 1x1 marker – the origin of this allele could be from either parent). Once all such probabilities have been assigned, probabilities at all other marker positions are approximated by first locating the nearest linked marker with an informative

probability (0 or 1). The probability is then estimated as *r* in the case of a starting probability of 0, or 1 – *r* in the case of a starting probability of 1, where *r* is the recombination frequency between the two markers, as derived from the map positions (for example, if Haldane's mapping function (Haldane, 1919) was used to generate the map, the inverse of this function is used to re-calculate *r*). The IBD probabilities $P_i$ are then normalised to ensure $\sum_i P_i = {ploidy}/{2}$ for each parent. Finally, IBD probabilities are imputed at a grid of positions using the `smooth.spline` function in R (by default, at 1 cM spacings).

# Form of the QTL model

***Single marker analysis***

For the single marker analysis, a genome-wide scan is performed by fitting the following additive model at each marker position:

$$Y = \bar{y} + \alpha D + \varepsilon$$

where $Y$ is the vector of phenotypes, $D$ is the vector of marker dosage scores, $\bar{y}$ is the overall mean and $\varepsilon$ the residuals. This model assumes additivity of QTL effects depending on the dosage, which may or may not hold. From the ANOVA table, the sum of squared residuals ($RSS_1$) is recorded and used to calculate the logarithm of odds (LOD) score as follows (Broman et al., 2003):

$$LOD = \frac{N}{2}\log_{10}\left(\frac{RSS_0}{RSS_1}\right)$$

where *N* is the population size, and $RSS_0$ is the residual sum of squares under the Null Hypothesis of no QTL, *i.e.* $RSS_0 = \sum_i(\bar{y} - y_i)^2$. The amount of explained variance at any tested position is also calculated using:

$$expl.var. = 1 - \frac{MS_R}{MS_T}$$

where $MS_R$ is the mean squared residuals from the ANOVA table, and $MS_T$ is the total mean squares, $MS_T = \frac{\sum SS}{N*nr.blocks-1}$ (where $SS$ are the total sum of squares terms from the ANOVA table, *N* the population size and $nr.blocks$ the number of blocks, taken to be one in the case of no experimental blocking). This is the same formula used to generate the adjusted $R^2$ from the linear model fit (as returned by the `summary.lm` function in R).

***IBD-based analysis***

The IBD-based analysis has two different forms, depending on the origin of the IBD probabilities. If the probabilities were estimated in TetraOrigin, then the probabilities of all possible tetraploid genotype combinations are known, either 36 possibilities in the

case of purely bivalent pairing, or 100 possibilities in the case where quadrivalent pairing is also allowed. These genotype probabilities are then used as weights in a regression model originally proposed by Kempthorne (Kempthorne, 1957), but more recently adopted in the TetraploidSNPMap pipeline (Hackett et al., 2013;Hackett et al., 2014;Hackett et al., 2017):

$$Y = \mu' + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 + \varepsilon$$

where the constraints $X_1 + X_2 + X_3 + X_4 = 2$ and $X_5 + X_6 + X_7 + X_8 = 2$ have already been applied. Here, the $X_i$ are indicator variables for each parental homologue (1-4 for parent 1, 5-8 for parent 2 in a tetraploid) and $\varepsilon$ is the residual term.

For the IBD probabilities estimated internally, the probabilities of *combinations* of parental alleles are unknown, but the inheritance probabilities of each *individual* parental homologue is known. The form of the model appears identical, *i.e.*

$$Y = \mu' + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 + \varepsilon$$

except now the regression is unweighted, and the $X_i$ are vectors of inheritance probabilities of a specific parental homologue at a locus.

For a hexaploid, the model is extended to include ten of the twelve parental homologues, *i.e.*

$$Y = \mu' + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_8 X_8 + \alpha_9 X_9 + \alpha_{10} X_{10} + \alpha_{11} X_{11} + \alpha_{12} X_{12} + \varepsilon$$

*etc.*

## Significance thresholds

Approximate LOD significance thresholds are determined using Permutation Tests (Churchill and Doerge, 1994). By default, 1000 permutations of the genotypes are generated, following which the maximum genome-wide LOD scores are recorded from each of the 1000 analyses, and the 95$^{th}$ quantile of the ordered LOD score is taken as an approximate 95% significance threshold. However, these number are merely guidelines, and the user has full control over the number of permutations and the selection of the approximate Type I error rate α.

## Inclusion of co-factors, covariates and experimental blocks

In cases where a large-effect QTL is present and segregating in a population, it can be advantageous to reduce the level of background noise at other loci by accounting first for the major QTL and running an analysis on the QTL-corrected phenotypes. Such an approach has previously been termed multiple QTL mapping (Jansen, 1993) or composite interval mapping (Zeng, 1993). In polyqtlR, we offer a simple approach to correct for genetic co-factors, either by supplying the name of a marker closely linked to

the major QTL peak, or the QTL peak position from the genome-wide scan (which is usually performed at a grid of splined positions). There is no limit to the number of co-factors that can be added, but a parsimonious analysis with only significant, unlinked QTL as genetic co-factors is recommended (as well as analyses including these co-factors singly). The QTL model described above for IBD probabilities is initially fitted at the supplied position(s) and the residuals are saved to replace the vector of phenotype values in the QTL scan.

Phenotypic covariates that may represent a source of confounding variation can also be supplied, and these are also initially fitted ($Y \sim Covariates$) and the residuals saved as a replacement for the phenotypes under investigation. In cases where incomplete covariates are supplied (*i.e.* containing missing values) the missing values are replaced with the mean covariate value, whilst warning the user of the extent of this problem. If experimental design factors are included in the analysis, they are also fitted ($Y \sim Blocks$) after which the residuals are used to perform the genome-wide QTL scan. Note that if blocks are included, the apparent dimension of the model increase. The genotype matrix $Y$ is vertically concatenated according to the number of blocks, which also motivated our decision to permute the *genotypes* rather than the *phenotypes* in the Permutation Test (the association is broken in both cases, but permuting genotypes and then concatenating them ensures the correct level of nesting of permutations is achieved). Both blocks, covariates and co-factors can be included (or any combination of these three), in which case blocks are first fit, followed by covariates and then co-factors. Because the analysis relies on simple linear models (using the `lm` function in R (R Core Team, 2016)), missing phenotypic values can be problematic. If blocks are used, there is the possibility to impute missing phenotypes using the fitted model for block effects and the non-missing phenotype scores for that individual in the other blocks. By default, at least 50% observations are required for imputation (*e.g.* minimum 2 out of 3 phenotypes non-missing for that individual in a 3-block situation).

## Exploration of QTL configuration and mode of action

One of the main advantages of an IBD-based analysis over single-marker methods is the ability to explore QTL peak positions to determine the most likely QTL configuration and mode of action (additive / dominant). These can be compared using a model-selection procedure such as the Bayesian Information Criterion (BIC) (Schwarz, 1978) as previously proposed (Hackett et al., 2014). In addition, we can run a per-homologue QTL analysis on linkage groups known to possess segregating QTL. To achieve this, we re-run the analysis while restricting the genotypic matrix $Y$ to a single column of haplotype-specific IBD probabilities. This may assist in the interpretation of the most likely origin of favourable (or unfavourable) QTL alleles (although only reliably so in

cases of a simplex QTL – for more complex, multi-allelic QTL the output can be misleading). In cases where a blocking factor(s) is included in the analysis, best linear unbiased estimates (BLUEs) should first be calculated using the BLUE function provided (which uses the `lme` function from the R package nlme (Pinheiro et al., 2017)).

## Genotypic Information Coefficient (GIC)

The Genotypic Information Coefficient (GIC) is a convenient measure of the precision of our knowledge on the composition of parental alleles at a particular position in each offspring, averaged across the mapping population. The GIC of homologue $j$ ($1 \leq j \leq \sum q$) at each position is calculated from the IBD probabilities using the formula:

$$GIC_j = 1 - \frac{4}{N}\sum_{n=1}^{N} P_{n,j}(1 - P_{n,j})$$

where $\sum q$ is the sum of parental ploidy levels, $N$ is the population size and $P_{n,j}$ is the probability of inheriting homologue $j$ in individual $n$ at that position (this is a generalisation of the GIC measure used in MapQTL (Van Ooijen, 1992;Van Ooijen, 2009)).

## Visualisations

The visualisation of results is an important aspect of any QTL analysis and is not neglected in polyqtlR. LOD profile plots across all chromosomes can be simply generated, and can be overlaid with previous analyses if direct comparisons are required (such as the impact of including a cofactor (or multiple co-factors) in the analysis). The GIC profiles of each parental homologue can also be plotted in a similar fashion, facilitating a comparison between the position of QTL peaks and the GIC landscape in that region. The results of homologue-specific analyses can also be plotted to help interpret the possible QTL configuration, and the results of the QTL model selection procedure (BIC) are also plotted to give an indication of the likelihood of competing models. The haplotype conformation of each individual for any specified genomic region can be plotted, either for selected individuals or for those individuals with extreme phenotypes. It is also possible to automatically screen for and identify recombinant individuals within a specified region, or individuals carrying double reduction products.

**Table 2. Simulated traits for a tetraploid population ($F_1$ = 200) and their underlying genetic architecture.**

| Phenotype | LG[a] | cM[b] | Q[c] | Action[d] | Configuration[e] |
|---|---|---|---|---|---|
| Trait_1 | 1 | 80 | 10 | additive | 1,5,8 |
| ($h^2$ = 0.2[†]) | 2 | 60 | 10 | additive | 1 |
| | | | - | | |
| | 2 | 60 | 10 | additive | 6 |
| | 5 | 10 | 10 | additive | 1,4 |
| Trait_2 ($h^2$ = 0.6) | 3 | 20 | 15 | additive | 1,2 |
| Trait_3 | 1 | 70 | 7 | dominant | 3,7,8 |
| ($h^2$ = 0.6) | 3 | 60 | 7 | additive | 3,7,8 |
| | 2 | 20 | 10 | epistatic additive | 2,5 |
| | 4 | 65 | 10 | epistatic additive | 1,6 |
| | 5 | 50 | 10 | epistatic additive | 6,7 |

[a] Linkage group of the simulated QTL; [b] Position of the simulated QTL in centiMorgans; [c] Effect size of segregating QTL allele (QTL alleles not mentioned were assigned an effect size of 0); [d] Simulated gene action, either "additive", "dominant" or "epistatic additive" (as described in the main text);
[e] Simulated QTL configuration with numbers 1-4 corresponding to parent 1 (for a tetraploid) and 5-8 for parent 2; [†] "Broad-sense" heritability of the simulated trait, as defined by $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$

# Sample dataset

The polyqtlR package contains a sample simulated dataset of marker data for a tetraploid species with five linkage groups (and no preferential chromosome pairing) simulated using PedigreeSim (Voorrips and Maliepaard, 2012). Phenotypes for three traits were generated for three years, harbouring various QTL as outlined in Table 2. The phenotype of the $i^{th}$ individual ($P_i$) with QTL dosage $d_i$ for each year was generated by sampling from a N-distribution: $P_i \sim \mathcal{N}(Q * d_i, \sigma_e^2) + \mathcal{N}(Y, s)$, where the second term corresponds to a random year effect (no genotype x environment interaction effects were included). The environmental variance $\sigma_e^2 = \left(\frac{1-h^2}{h^2}\right)\sigma_g^2$ was derived from the genotypic variance $\sigma_g^2$, which was estimated by calculating the variance of the QTL dosage scores at the QTL position (for dominant QTL these were taken to be 0 and 1 only), where $h^2$ refers to the "broad-sense" heritability (Acquaah, 2012). Offspring QTL genotypes were known at any chosen position from the .hsa and .hsb output files of PedigreeSim. Trait_1 was simulated to be affected by three unlinked QTL, one of which was multi-allelic, with an allele of effect size +10 on homologue 1 (originating from the mother), an allele of effect size -10 on homologue 6 (from the father) while all other alleles were of effect size 0 (Table 2). Trait_2 was simulated to be monogenic, with a single major-effect locus on linkage group 3 (Table 2). For Trait_3, we modelled epistasis through complementary (additive) gene action between QTL on three linkage groups. It was assumed that at least 1 copy of each plus-allele was required to have a positive effect on the phenotype (as might occur in a biosynthetic pathway for instance); where multiple copies were

inherited, dosage-effects were also included. For traits 1 and 3 where multiple QTL contributed, single-locus phenotypes were added together to produce a "multi-locus" phenotype.

# Results

Using the sample tetraploid dataset described, we estimated IBD probabilities using TetraOrigin (Zheng et al., 2016) as well as using our own IBD estimation algorithm for comparison. The TetraOrigin IBDs were used to perform a QTL analysis for the three simulated traits described in Table 2, with the results visualised in Figure 1. For Trait_1, three significant peaks were detected (as listed in Table 3), corresponding to the positions of the three simulated QTL. For Trait_2, a single very sharp QTL peak was observed at

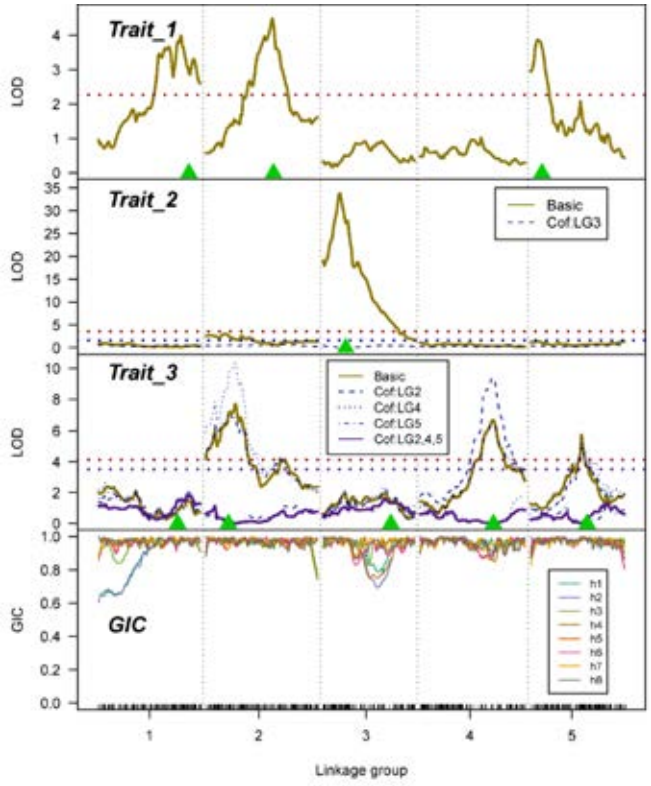



**Figure 1. Visualised QTL results of polyqtlR for the three simulated traits, with the genotypic information coefficients (GIC) per homologue shown in the lowest panel.** Significance thresholds are shown as horizontal dotted lines, as determined by permutation tests. For Trait_2, a co-factor analysis using the major peak on linkage group 3 yielded no extra QTL (although the significance threshold reduced slightly, as shown by the lower (blue) dotted line). For Trait_3, the addition of the three QTL peaks as co-factors reduced the significance threshold also (lower purple dotted horizontal line), although no new QTL were detected (purple LOD profile line). The distribution of markers on the integrated linkage map are shown at the x-axis as vertical dashes. True QTL positions are shown by green triangles.

15 cM, although the LOD-2 support interval did not include the true QTL position (which was at 20 cM). A co-factor analysis using the peak position on linkage group 3 did not result in any extra QTL being detected (Figure 1, second panel). For Trait_3, three significant peaks were detected on linkage groups 2, 4 and 5, with the remaining smaller QTL on linkage groups 1 and 3 not detected. Including either of the peaks on linkage groups 2 and 4 as a co-factor increased the significance at the other peak, but had little effect on the LOD profile on linkage group 5 (Figure 1). Including all three peaks simultaneously as co-factors in the analysis reduced the significance threshold slightly, but did not result in the detection of any new QTL.

The GIC values per homologue were calculated and found to often be close to 1 (full information). However, some major dips in the GIC profile can be seen, for example on linkage group 1 for homologues 1 and 2, corresponding to the region on the genetic linkage maps which harbours no segregating markers (Supplementary Figure 1).

**Table 3. Results of the QTL analysis for the three simulated traits.**

| Phenotype | LG[a] | Peak (cM)[b] | LOD-2 (cM)[c] | LOD | Thr.[d] | $R^2_{adj}$[e] | Configuration[f] | Act.[g] | BIC[h] |
|---|---|---|---|---|---|---|---|---|---|
| Trait_1 | 1 | 73 | 51 - 90 | 3.99 | 2.27 | 0.079 | Qqqq x QqqQ | *A* | 386 |
| | | | | | | | Qqqq x qqqQ | *D* | 391 |
| | 2 | 59 | 38 - 71 | 4.49 | 2.27 | 0.089 | Qqqq x QqQQ | *A* | 382 |
| | 5 | 6 | 0 - 15 | 3.87 | 2.27 | 0.076 | QqqQ x qqqq | *A* | 379 |
| | | | | | | | QqqQ x qqqq | *D* | 383 |
| Trait_2 | 3 | 15 | 14-16 | 33.79 | 3.58 | 0.536 | QQqq x qqqq | *A* | 332.0 |
| Trait_3 | 2 | 26 | 8 - 35 | 7.69 | 4.13 | 0.154 | QqqQ x QQqq | *D* | 411 |
| | | | | | | | Qqqq x qQqq | *A* | 413 |
| | | | | | | | QQqQ x QQqq | *D* | 417 |
| | 4 | 64 | 57 - 68 | 6.70 | 4.13 | 0.134 | qQqq x qqqq | *A* | 381.3 |
| | 5 | 45 | 45 - 45 | 5.74 | 4.13 | 0.115 | *Unclear : 10 models* | | |

[a] Linkage Group (LG) of QTL position; [b] Position of QTL peak in centiMorgans; [c] QTL 2-LOD support interval in centiMorgans; [d] LOD significance threshold as estimated from permutation testing with $N = 1000$ and $\alpha = 0.05$; [e] Adjusted $R^2$ from the model fit at the QTL peak, the proportion of variance explained by the QTL peak; [f] Predicted QTL configuration using model selection procedure based on BIC. "*Unclear : 10 models*" refers to situation when no clear minimum BIC occurred, here with 10 competing QTL models within 6 BIC units of the minimum; [g] Mode of QTL action, either additive (*A*) or dominant (*D*); [h] Bayesian Information Criterion (BIC) values for models listed in column 'Configuration'

The prediction of QTL segregation type and mode of action by selecting the model which minimised the BIC was accurate for Trait_1 and Trait_2, but prediction of the configuration of QTL underlying Trait_3 proved more difficult (Table 3). At each QTL peak, both additive and dominant bi-allelic QTL models are tested (of which there are 240 in total – *c.f.* Chapter 8) and any QTL models within 6 BIC of the minimum are

returned as possible QTL models (Figure 2). For Trait_1, a bi-allelic model for the LG1 was not entirely appropriate, and the predicted QTL type (*Qqqq* x *qQqq*, Table 3) was not exactly that which was simulated (*Qooo* x *oqoo*, where *o* is an allele which does not contribute to trait variation, while *Q* and *q* do). In some cases, more than one model was suggested (Figure 2.a) while at times one clearly optimal model was found (Figure 2.b). The output of a single-homologue analysis was also visualised, which may help in the interpretation of results (an example of which is shown in Figure 2.c). The predicted QTL model can be verified by examining the haplotype structure of offspring with extreme phenotypes, which in theory should carry some or all of the "positive alleles" for the QTL (Figure 3).

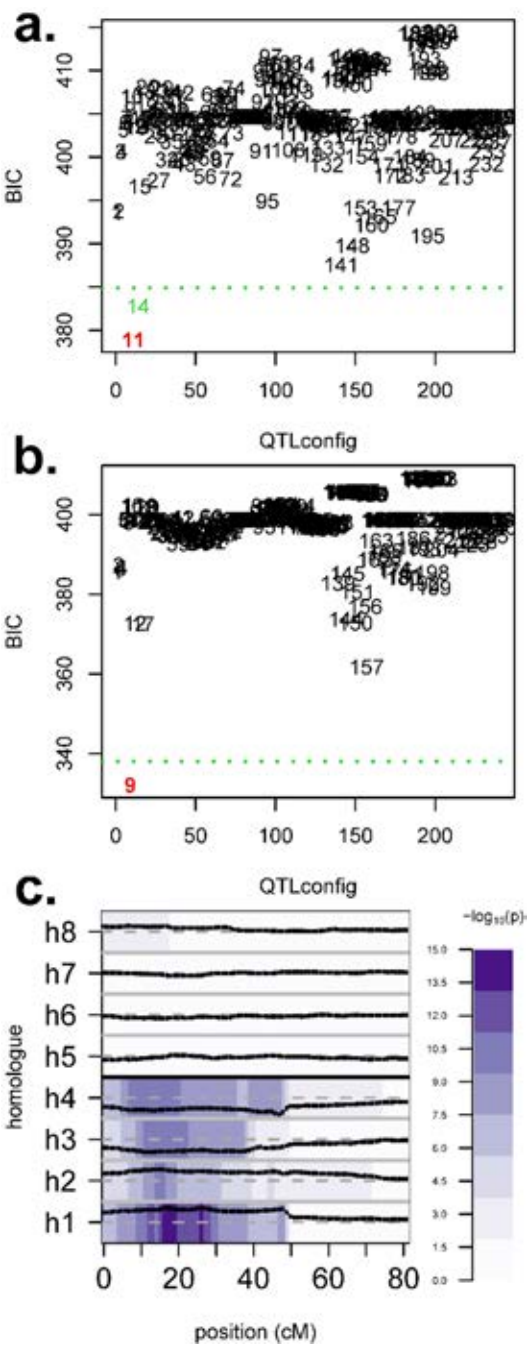



**Figure 2. Exploration of QTL peaks using the polyqtlR package.**
**a.** The Bayesian Information Criterion (BIC) results for the linkage group 5 (LG 5) peak of Trait_1. Each number corresponds to a different configuration, for which there are 240 in total (Supplementary Table 1). The additive model QqqQ x qqqq (model 11) minimises the BIC for this trait (and model 14, the dominant version of this model, was second most likely).
**b.** BIC results for the LG 3 peak of Trait_2, with the additive model QQqq x qqqq suggested (model 9). Here, there are no other models within 6 BIC of the minimum.
**c.** Per-homologue analyses can also complement the analysis; here, the LG 3 results of Trait_2 are shown. Colour scales are used to indicate significance of the single-homologue model fit, with the direction of the homologue effect shown by black lines (above axis = positive, below axis = negative). Here we see that positive allele effects originate from homologues 1 and 2.

In the simulated dataset provided with the package we did not include quadrivalent pairing structures or preferential chromosomal pairing during the population simulation in PedigreeSim. However, a certain proportion of multivalent pairing structures are expected to form during autopolyploid meiosis which can lead to the phenomenon of double reduction. An analysis of the potato AxC $F_1$ population genotyped using the SolSTW single nucleotide polymorphism (SNP) array (Vos et al., 2015) as described in Bourke et al. (2015, 2016) revealed the presence of double reduction products in a subset of individuals in the population (a sample for potato linkage group 1 is shown in Supplementary Figure 2).

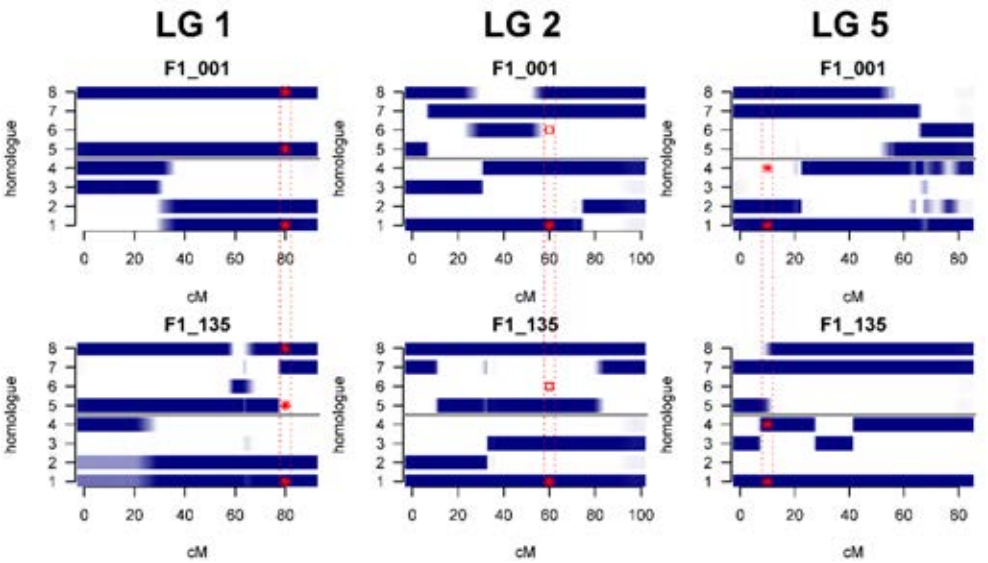



**Figure 3. Predicted haplotype conformation of two individuals (F1_001 and F1_135) with extreme phenotypes for Trait_1.** Three QTL were detected for Trait_1 (Figure 1) on linkage groups LG 1, 2 and 5. The QTL positions are depicted by dotted red boxed, with plus-alleles of positive effect shown by red stars and alleles of negative effect by an open square. Individuals F1_001 and F1_135 had the highest trait values (BLUE = 145 and 142 resp.; population mean = 96 ± 18). In general, these individuals carried the predicted favourable alleles at these positions. Such visualisations can help confirm QTL analysis results as well as facilitate haplotype-assisted selection.

Double reduction in general exerts only a minor influence on linkage analyses (Bourke et al., 2015;Bourke et al., 2016) or QTL analyses (Chapter 8) and can safely be ignored in most cases. In polyqtlR, QTL analysis using a model that includes double reduction is straightforward, provided that IBD probabilities from a model that allowed for double reduction are available (for example using TetraOrigin (Zheng et al., 2016)). One additional feature of polyqtlR that may be useful for gene mapping and $F_1$ selection is its ability to identify recombinant $F_1$ individuals within a specified region of interest, an

example of which is shown in Supplementary Figure 3 where we identified eight individuals with a recombination between homologues 1 and 2 in the interval 15 – 25 cM on linkage group 3.

Finally, we also performed a single-marker analysis using the marker dosages as an additive predictor of phenotypes, highlighting some differences between IBD-based and single marker-based analyses. The IBD-based analysis was found to be slightly more powerful, with the LG 5 peak for Trait_3 being detected using IBD probabilities (Figure 1), but missed entirely using only marker dosages (Supplementary Figure 4). Significance thresholds varied between approaches, but were not consistently higher or lower in one set or the other. A single extremely-significant marker was found to be associated with Trait_2 on LG 3 (as can be seen in Supplementary Figure 4). This marker ("DN_III35.6") mapped to 19.3 cM (just 0.7 cM from the QTL) and was phased exactly as the QTL itself (1100 | 0000). In other words, it was a near-perfect marker for this particular QTL.

# Discussion

In this paper we have described the possibilities for performing QTL analysis in bi-parental $F_1$ populations of autopolyploid species using the polyqtlR package in R. We did this using simulated tetraploid SNP marker dosage data and three associated quantitative traits with contrasting genetic architecture recorded over three years. We confined our attention to a tetraploid dataset and have not demonstrated trait analysis in triploid or hexaploid populations, the other most commonly encountered ploidy levels for experimental plant populations and breeding programs. However, all functions within polyqtlR have been developed to be able to analyse these (and potentially any other) ploidy levels and therefore the results would be substantially similar (see for example van Geest et al. (2017) for a description of a QTL analysis in the autohexaploid ornamental species *Chrysanthemum* × *morifolium*). In this discussion we consider a number of points – the relative advantages or disadvantages of IBD-based analyses over single-marker approaches, the importance of GIC, a comparison with alternative software, and the scope for future developments.

## IBD versus single marker analyses

polyqtlR offers the possibility to perform QTL analyses either using identity-by-descent (IBD) probabilities or using the marker dosage scores directly. Although we showed that the IBD-based analysis had somewhat more power by detecting a QTL that was missed by the single-marker analysis, we have not performed a comprehensive power analysis

to verify this. IBD-based analyses allow us to predict the most probable QTL segregation type and mode of gene action, as well as tracking the transmission of favourable or unfavourable alleles in the population. They therefore offer a more comprehensive approach to QTL detection than single-marker studies. However, there may also be situations in which single marker analyses may be preferable to IBD-based approaches. For example, in cases where it is not possible to reliably assign discrete dosage scores, continuous marker genotypes might be the only available genotyping data. These have already proven to be useful in QTL association mapping studies in various polyploid species (Grandke et al., 2016;Tumino et al., 2016). The methods to determine IBD probabilities using continuous genotypes rather than discrete dosage scores are not yet developed (nor indeed are there methods to construct linkage maps from such genotypes to the best of our knowledge). In such circumstances, a single marker analysis may be the only available option. In the unlikely event that an individual marker co-segregates nearly perfectly with a QTL allele (as was the case here for Trait_2) then single-marker approaches will also be at least as powerful than IBD-based analyses, if not more so. In cases where the estimation of IBD probabilities is difficult due to serious errors in the genotypes, linkage map and/or marker phasing, then (accurate) single markers may prove more suitable to detect QTL effects than an IBD-based analysis. However, on balance both approaches have their merits, and it is generally a good idea to perform both analyses and combine results for a more comprehensive picture.

## Genotypic Information Coefficient (GIC)

It has previously been shown that a high and uniform GIC is needed for accurate QTL analyses (Chapter 8). Here, we describe the calculation of GIC and applied it to a simulated tetraploid dataset from the polyqtlR package. However, the usefulness of the GIC measure depends to a large extent on the quality of the algorithm to generate IBD probabilities upon which it is based. In cases where the IBD probabilities are close to 0 or 1 but are also incorrect, the GIC may be misleading: the GIC can only indicate regions where there is *low* information content, not *incorrect* information. In what we present here, we enjoyed unrealistically fortunate circumstances, with error-free simulated data containing no missing values. This was reflected in consistently-high GICs for most homologues and linkage groups (Figure 1, bottom panel). In reality, we are unlikely to encounter such high-quality datasets. Although we have studied the GIC in detail elsewhere (Chapter 8), we have not yet attempted to quantify what constitutes a "sufficiently-high" GIC. However, we found that regions of variable GIC will tend to bias the detection of QTL to neighbouring loci with high GIC, and that GIC is a strong predictor of QTL detection power (Chapter 8). Therefore, it is prudent to examine the GIC profiles around QTL peaks, particularly on the homologues which are indicated to carry the QTL alleles.

## Software alternatives

The only other software that is currently available to perform similar analyses is TetraploidSNPMap (TSNPM) (Hackett et al., 2017), available as free downloadable software with an advanced user-friendly GUI that can produce integrated and phased linkage maps as well as perform interval QTL mapping in tetraploid $F_1$ populations. We transformed our simulated marker dosages, phased linkage map and phenotypes into TSNPM-compatible input and re-ran the analysis (for the phenotypes we used the BLUEs as phenotypes as there is no facility for including blocks in TSNPM). An overview of the results is given in Supplementary Table 2. On the whole the results are very comparable. TSNPM detected the same QTL peaks in (almost precisely) the same positions, although the peak LOD scores and estimated explained variances were not exactly the same (TSNPM tended to have slightly higher LOD scores and greater explained variances at the QTL peaks, although both software employ similar methods). The major difference in the results was in the predicted QTL segregation type and mode of action. Apart from additive and dominant QTL models, TSNPM also tests "codominant factors", where the QTL genotype classes are allowed to have independent means. Despite this, the QTL model search space is currently limited to 1x0 (0x1), 1x1 and 2x0 (0x2) QTL types, whereas we currently consider all possible QTL segregation types (there are 240 bi-allelic models to consider in total (Chapter 8, Supplementary Table 1)), albeit without the codominant model currently included. If we compare the predictions of polyqtlR (Table 3) and TSNPM (Supplementary Table 2) with the true QTL configurations (Table 2), both performed well, but there are some small differences. In general, TSNPM returned more possible models (which we take as being within 6 BIC units of the minimum BIC, which corresponds to between positive and strong evidence of a meaningful difference between models (Neath and Cavanaugh, 2012)). For the linkage group 1 QTL of Trait_1 with configuration Qqqq x QqqQ, TSNPM found that Qqqq x qqqQ was the best fitting, which is as far as it is able to go. polyqtlR also tests all possible 1x2 QTL types (what we term "simplex x duplex" QTL), and therefore found that the model Qqqq x QqqQ was the most likely. For Trait_2 (the "monogenic" trait) both software correctly guessed that the QTL segregated as QQqq x qqqq with additive effect. However, TSNPM also returned a second competing model, namely QQqq x qqqq with codominant effect, with a ΔBIC of 1.9 (Supplementary Table 2). For polyqtlR there was no competing model within at least 10 BIC, *i.e.* it was a very clear front-runner. Interestingly, both software had some difficulties with Trait_3 – neither detected the minor QTLs on LG 1 or 3, and polyqtlR in particular had problems in correctly assigning the QTL segregation type for the epistatic QTLs of LG 2, 4 and 5. TSNPM did better, correctly predicting the configuration of all three QTL, but not always the mode of action (which was difficult to define – an allele count of at least 1 was needed at all three loci to produce a subsequent (additive) effect). Both software use

the Bayesian Information Criterion (BIC) for model comparisons, and therefore discrepancies are most likely due to slight differences in the IBD probabilities (which we were unable to compare as they are not part of the TSNPM output).

# Future directions of polyqtlR

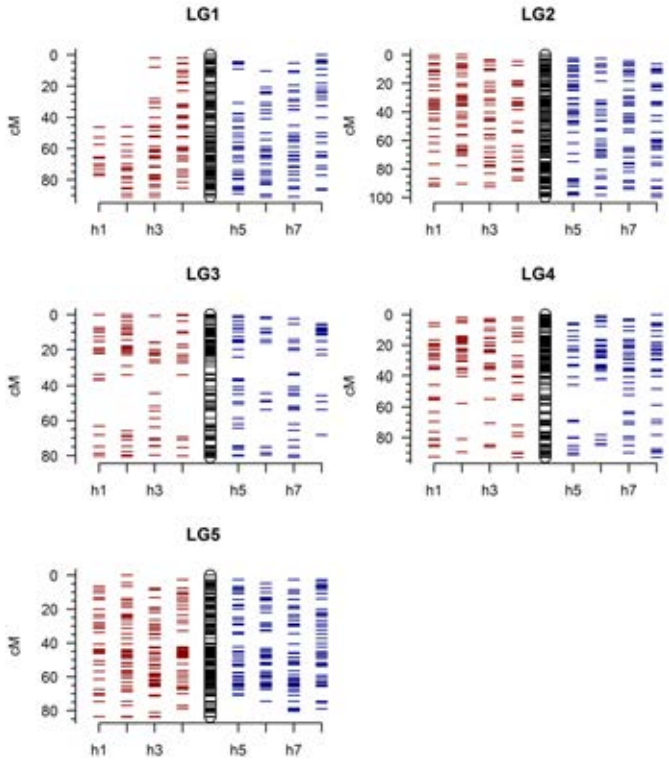
The statistical models behind the QTL analysis in polyqtlR are simple yet powerful, with most of the complications of polyploid inheritance already dealt with in the earlier stages of phased linkage map construction and in the calculation of IBD probabilities. There is however still scope to expand upon the developments already implemented. For example, we have not yet implemented any formal procedure for detecting epistasis. As we saw with Trait_3, polyqtlR appears to have difficulties in correctly identifying the QTL configuration in such circumstances. In our treatment of genetic co-factors we do not currently compare between models, but a model comparison framework such as the Akaike or Bayesian Information Criterion could be quite instructive here. There is also currently no provision for investigating genotype x environment (GxE) interactions, despite these interactions being relatively common (van Eeuwijk et al., 2010). One reason we have not (yet) implemented GxE in polyqtlR is the difficulty of including an interaction term for a polyploid population. Should this interaction effect be investigated homologue-by-homologue? Perhaps two at a time, or simultaneously? In other words, capturing an interaction effect in a polyploid QTL model would likely require some simplifying assumptions which may or may not be appropriate. In time these will no doubt be investigated and implemented.

Another direction polyqtlR could take in future is to integrate information across multiple populations in a single analysis either using pedigree or genomic relationship matrices, such as is done for example in FlexQTL (Bink et al., 2008). The current dependence on bi-parental $F_1$ populations means that each analysis must be done separately, with perhaps some *ad hoc* integration of QTL results afterwards. In future, merging such analyses into a single study will improve not only detection power but also the robustness of results, but will require the development of novel approaches to estimate IBD probabilities across populations and pedigrees.

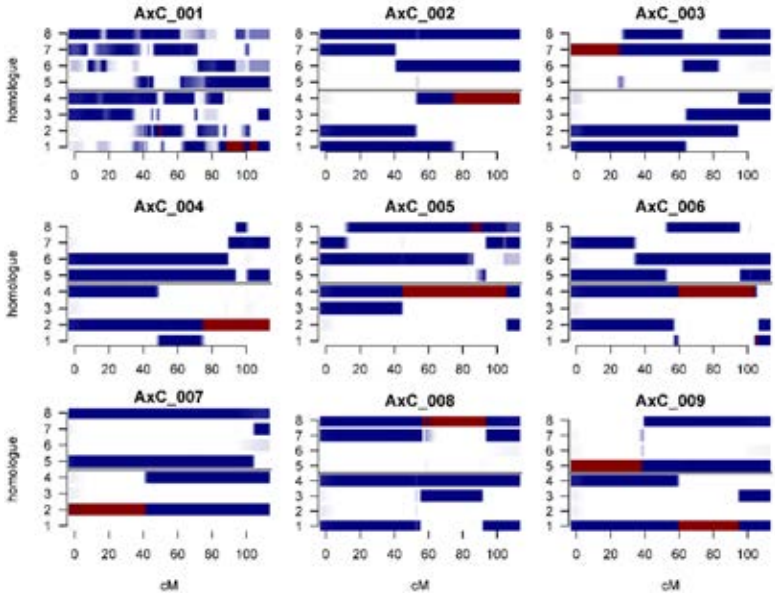
Finally, we also expect that polyploid genotyping may soon differ from the single-SNP dosages that are currently the basis for much of the analysis in polyqtlR. Time will tell how successful these new genotyping technologies will be at accurately capturing the full breadth of allelic diversity in polyploids and exploiting it for research and breeding purposes. polyqtlR is flexible enough to allow adaption to deal with such multi-allelic haplotype-based markers, offering greater choice to users in future experiments and analyses.

# Acknowledgements

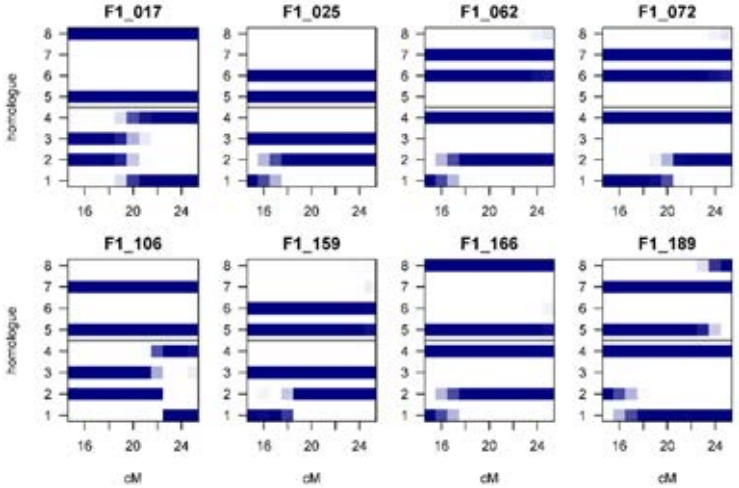
This work was supported through the TKI projects “A genetic analysis pipeline for polyploid crops” (project number BO-26.03-002-001) and “Novel genetic and genomic tools for polyploid crops” (project number BO-26.03-009-004). The authors would like to thank Eric van de Weg, Herman van Eck, René Smulders, Michiel Klaassen (WUR Plant Breeding) and Camillo Berenos (Dümmen Orange B.V.) for their constructive feedback and discussions.



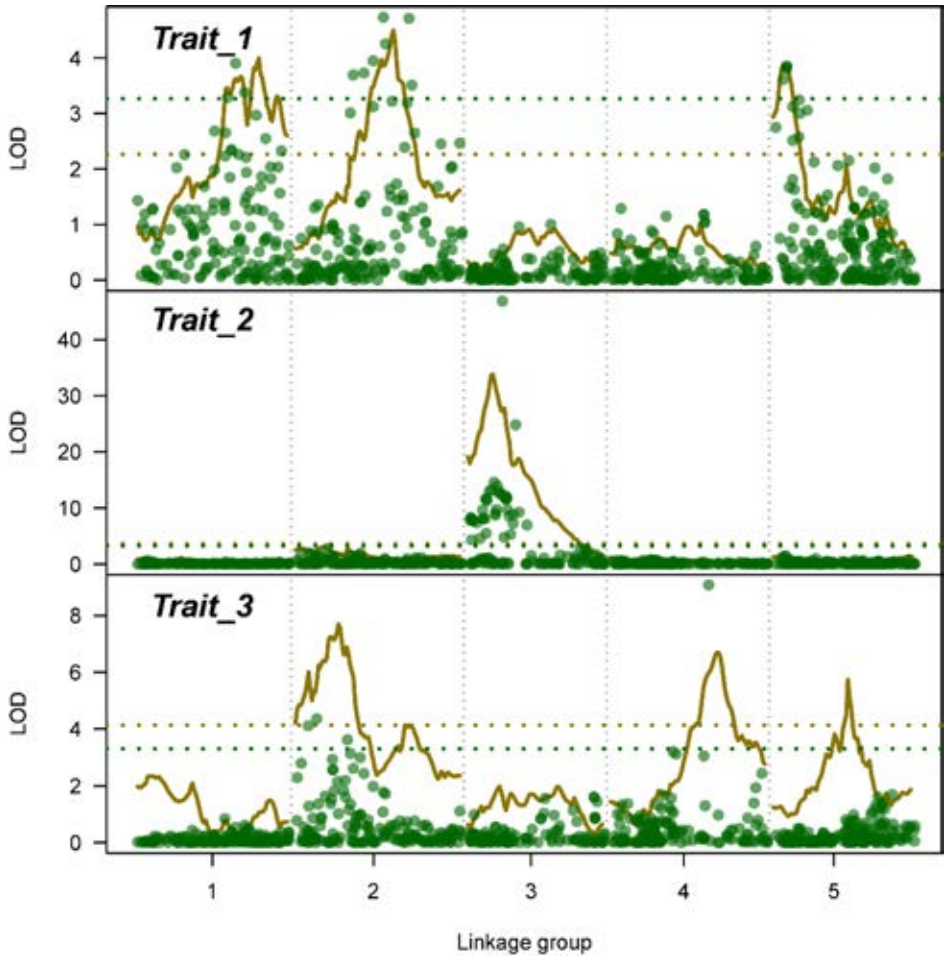

**Supplementary Figure 1.** Visualisation of the marker coverage of the tetraploid linkage maps provided with the *polyqtlR* package. Parent 1 homologue maps are numbered h1 – h4, with parent 2 homologue maps numbered h5 – h8. The integrated maps are shown in the centre for each linkage group (LG1 – LG5 corresponding to the five linkage groups). Maps were generated and plotted using the *polymapR* package (Chapter 6). Certain regions *e.g.* 0 – 40 cM for linkage group 1 homologues 1 and 2 contain no segregating marker information, which is reflected in low GIC values for the affected homologues (*c.f.* Figure 1).




**Supplementary Figure 2.** Predicted haplotype conformation showing possible double reduction segments in red, from selected individuals of the potato AxC $F_1$ population, chromosome 1. Lighter shades of blue correspond to intermediate probabilities. Certain individuals *e.g* AxC_001 or AxC_005 proved difficult to reconstruct with TetraOrigin and harbour double reduction segments of improbably-short length.




**Supplementary Figure 3.** Visualisation of the haplotypes of eight $F_1$ individuals from the simulated dataset identified as having a recombination between homologues 1 and 2 on linkage group 3 between 15 cM and 25 cM. The user has the ability to select any set of homologues involved in recombination and any desired stretch of chromosome, as well as any threshold probability with which a homologue (or segment thereof) is considered to be inherited (by default 0.95).




**Supplementary Figure 4.** Single-marker analysis results (dark green points) using additive linear model on marker dosages to detect QTL. Green horizontal dotted lines correspond to LOD significance thresholds as determined by permutation tests ($N = 1000$, $\alpha = 0.05$). Also shown are the results of interval-QTL mapping using IBD probabilities (reproduced from Figure 1), with significance thresholds as horizontal gold dotted lines. The QTL peak on LG 5 for Trait_3 was not detected using the single-marker approach but was found using the IBD-based approach.

# Chapter 11

# General Discussion

# Foreword

Working under the auspices of the department of plant breeding, one experiences a sense of responsibility that research findings should ultimately have a practical application in the plant breeding sector. This is particularly true of a project embedded within a public-private partnership, with a consortium of a dozen breeding companies providing matching funding for this work. That said, there was never the feeling that those enticing yet often time-consuming theoretical side-alleys should not be explored. In this discussion therefore, I address both aspects – how the findings of this thesis might be used to help in the breeding of polyploid crops, but also how this thesis has deepened our understanding of polyploids. Ultimately, both can help in the pursuit of better polyploid crops. As the saying goes, sometimes "there's nothing more practical than a good theory" (Lewin, 1951). Or taken from a plant-breeding perspective, "the full potential and effectiveness of a plant-breeding program can be realised only when the reproductive, genetic and breeding behaviour of a species is known and understood" (Dewey, 1966).

# Can experimental populations yield unbiased information on meiosis?

In this thesis I have tried to investigate real marker datasets beyond the traditional boundaries of genetic mapping, with a particular interest in polyploid meiotic phenomena. This began in Chapter 3 with a description of the "double-reduction landscape" in a tetraploid potato population. In Chapter 5 we reconstructed the pairing behaviour in a tetraploid rose population and from that inferred that homologue pairing and recombination was not random across all chromosomes. We took the further step of trying to relate homology at the level of SNP markers to the pairing behaviour, testing the theory that chromosomal homology is an important factor in pairing recognition and synapsis formation during meiosis (Naranjo and Corredor, 2008). Later in this discussion I also present some results on meiotic cross-over frequencies in polyploid meiosis using a tetraploid potato mapping population. In all of these cases, I am assuming that an experimental mapping population represents an unbiased record of parental meiosis. This is of course an unrealistic assumption, and like many experiments using living populations, suffers from what could be termed the *survivors' paradox*. We only use plants that have survived in a mapping study, which often means that an initial selection has already occurred before the experiment has even started (*e.g.* during gamete formation, post-fertilisation death of embryos, non-germination, pre- or post-emergence diseases *etc.*). One of the well-known consequences of such unknown selection is marker skewness (also called segregation distortion (Mangelsdorf and Jones, 1926)), where

unexpected or unbalanced proportions of marker alleles are observed. These markers are often removed prior to mapping to avoid the possible complication of false linkages, although in so doing, parts of entire linkage groups may unwittingly be removed (Van Ooijen and Jansen, 2013).

One of the questions that arose during our work on rose meiosis (Chapter 5) was whether segregation distortion and preferential homologue pairing could leave identical signatures in a mapping population and hence be confounded effects. We argued that the over-abundance of certain pairing combinations was unlikely to have been caused by segregation distortion, unless for example a combination of alleles proved to be debilitating (but non-lethal) to the spore in four of the six homologue pairings possible (which is quite unlikely, but not impossible). In potato we estimated the average rate of double reduction at the centromeres and telomeres (Chapter 3), which could have been influenced by systematic bias if for example gametes resulting from multivalents were more frequently infertile than those from bivalents (Lloyd and Bomblies, 2016). With only genotyping data from a mapping population we cannot definitively resolve these ambiguities.

There are some complementary experimental approaches that could help provide more conclusive answers to the precise nature of the biological phenomena involved. Many of the earliest papers on polyploid meiosis were cytological – *i.e.* physically observing the pairing behaviour of chromosomes during meiosis and (for example) counting the number of bivalents and multivalents that persisted into metaphase I (Westergaard, 1940;Cadman, 1943;Lamm, 1945;Swaminathan and Howard, 1953). FISH (Fluorescence *in situ* hybridisation) or GISH (Genomic *in situ* hybridisation) techniques could be used to identify homologous (and homoeologous) pairing during meiosis, although till now these techniques have mainly been used to establish base chromosomal number or identify sub-genomes in polyploids (D'Hont, 2005;Eberhard et al., 2010;Chester et al., 2012;Liu et al., 2016). Another approach to investigate meiosis is tetrad analysis (Preuss et al., 1994). In *Arabidopsis thaliana*, mutant lines carrying mutations in the *QUARTET* genes *QRT1* or *QRT2* produce conjoined pollen tetrad cells which can help reveal the exact pairing and recombination behaviour at meiosis (Wijnker et al., 2013). However, being a recessive mutation (Preuss et al., 1994) it is uncertain how useful *QRT* mutants would be in genetic studies of polyploids, assuming of course that *QRT* homologues or genes of similar function exist. Tetrad analyses also suffer from the survivors' paradox, given that experimental results are based on surviving pollen cells (the original study found that only half of all tetrads possessed four viable pollen grains (Preuss et al., 1994)). Therefore it seems that almost all experimental approaches, no matter how elegant, rely on surviving germplasm with which to make inferences on

meiosis. The original cytological studies were arguably the most unbiased regarding meiosis itself, but the viability of the observed structures was inevitably unknown. Perhaps an exhaustive analysis that tracks every single spore and egg cell could be imagined, but in practice it seems hardly feasible. For a breeder the issue of population-induced bias is perhaps moot, as breeders only work with surviving germplasm. For a fundamental researcher however, such biases need to be carefully considered and accounted for. Our work sits somewhere between theoretical and applied genetics, and it is therefore hoped that highlighting such issues here and in previous discussions is sufficient for now. The results we have obtained in this thesis regarding polyploid meiosis using experimental populations are certainly relevant for breeders, although corroborating data, ideally cytogenetic, would be advantageous in future studies to fully address the more fundamental meiotic questions we have attempted to answer.

# Cross-over rates in bivalent and multivalent meiosis

One of the key factors that ensures the orderly division of homologues during meiosis is the formation of chiasmata which helps generate mechanical tension at the centromeres and kinetochores during spindle-fibre formation. This tension is sensed by the cell, signalling that spindle fibres are correctly attached and that homologues should segregate correctly to opposite poles (Nicklas, 1997;Lampson and Cheeseman, 2011). Bivalent pairing in a polyploid poses no additional challenges to the mechanisms that have evolved to ensure balanced chromosomal segregation in diploids (since all pairing in diploids occurs in bivalents). A multivalent on the other hand is a "novel" meiotic structure that potentially disrupts these mechanisms. If homologous pairing were left unchecked it could lead to meiotic chaos, with many unresolved entanglements between homologues leading to improper segregation. It has long been hypothesised that a reduction in cross-over frequency occurs in autopolyploid meiosis which could have a stabilising effect on meiosis (Kostoff, 1940;Myers, 1943;McCollum, 1957;Hazarika and Rees, 1967). This hypothesis follows a certain line of logic: fewer cross-overs would result in fewer multivalents and hence a more orderly meiosis, since at least three separate cross-overs are required to generate an effective multivalent structure that can persist through to metaphase I (Chapter 1). Note that an "effective" multivalent in this context generally means a quadrivalent; a trivalent + univalent combination (in a tetraploid) is not tolerated by the cell as it would lead to unbalanced segregation (Bomblies et al., 2016). This has been borne out by a meta-analysis of autopolyploid meiosis, where very few univalent or trivalent structures were recorded (Ramsey and Schemske, 2002). In theory hexavalent structures are also possible at higher ploidy levels, but these have only rarely been reported in the literature (Ramsey and Schemske,

2002). Assuming that fewer cross-overs do indeed occur in autopolyploids than would be expected from recombination rates in progenitor diploid species (Hazarika and Rees, 1967), the question still remains how this might be achieved by the cell. An interesting possible mechanism that was recently proposed is that an increase in the effective distance of cross-over interference could play a significant role (Bomblies et al., 2016). As breeding very often relies on recombinations to sort and shuffle favourable alleles into a single genotype, any limitation on the number of cross-overs could have implications on the efficacy of traditional breeding approaches.

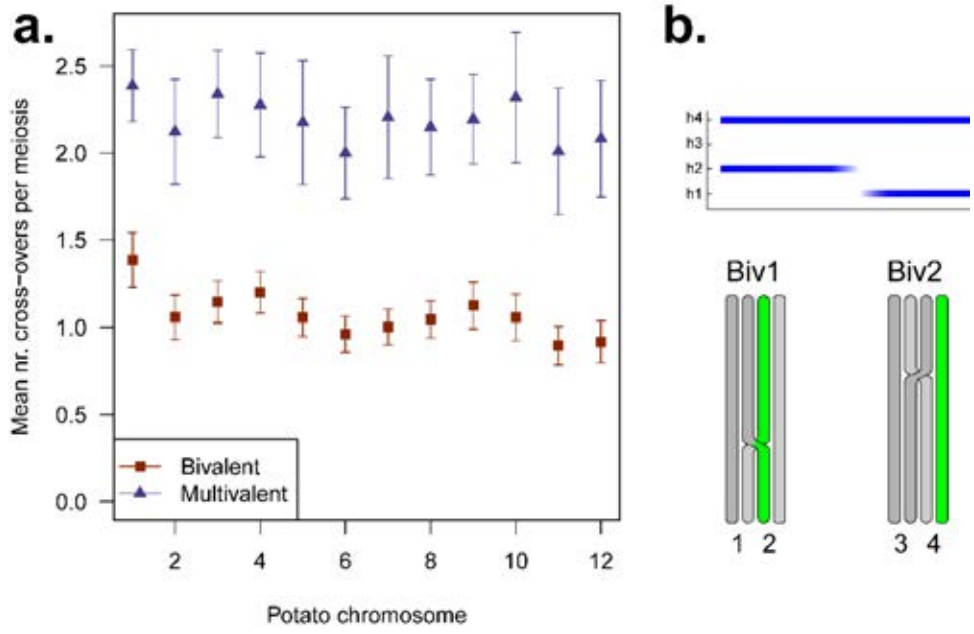



**Figure 1. Cross-over rates in bivalent and multivalent pairing structures. a.** Average rate of *observed* cross-overs per meiosis in the AxC tetraploid potato population, either resulting from bivalent pairing (red squares) or quadrivalent pairing (blue triangles). Error bars show 99% confidence intervals around the means for each chromosome. Two-sided t-tests on the difference of means were all highly-significant ($P < 1 \times 10^{-11}$). **b.** Example of possible discrepancy between the observed number of cross-overs from IBD probabilities (above) and the actual cross-overs that occurred during meiosis (below). Inherited chromosomal segments are highlighted in green, and correspond to the IBD predictions. However, the recombinant products from Bivalent 2 are not transmitted to the gamete and therefore the cross-over is not observed. The data used to generate figure (a) only consider observed recombinations. Cross-overs were identified from the output of TetraOrigin (Zheng et al., 2016), using a presence probability threshold of 0.9 (*i.e.* any inheritance probabilities < 0.9 were discounted) and a minimum recombinant segment length of 5cM.

However, there remains considerable uncertainty as to whether cross-over rates (and hence quadrivalents) are actively suppressed in autopolyploids. In this thesis we have provided clear evidence for the existence of quadrivalents in the meiosis of autopolyploids like potato and rose (Chapter 3 and Chapter 5, respectively). Therefore, the assertion that quadrivalents are aberrant structures that cause a failure in the proper division of homologues (Darlington, 1937;Kostoff, 1940;Carvalho et al., 2010;Lloyd and Bomblies, 2016) does not hold across all species (and particularly not among established autopolyploids, many of which are domesticated crop species). Here we are once more plagued by the *survivors' paradox*: it could be that the 20 – 30% quadrivalent pairings that we estimated to have contributed to the potato AxC population (Chapters 3, 8) might only represent a fraction of the multivalent meioses that occurred, the rest having perished. However, even were this to be true, it still represents a large proportion of successful and balanced meioses that involved quadrivalents, a fact that some of the literature mentioned above would have you doubt is even possible.

In fact, we can actually test the hypothesis that cross-over rates influence the pairing behaviour using our population datasets. To demonstrate this point, I re-analysed the identity-by-descent (IBD) probabilities of the AxC potato population, reconstructing the cross-over events that were predicted for each of the 235 $F_1$ individuals (Figure 1). I chose a presence threshold probability of 0.9 (which meant that if the inheritance probability at a marker for a homologue was less than 0.9 I did not consider that marker to have been inherited) and ignored any recombinant sections less than 5 cM in length (to prevent over-estimation of the number of recombinations in the case of noisy results). After this initial filtering, I counted the number of cross-overs from bivalent or quadrivalent meioses (I treated double reduction products separately and added them afterwards). As can be seen in Figure 1.a, the rate of observed cross-overs from bivalent pairing was significantly lower than the number observed due to multivalent pairing across all 12 potato chromosomes ($P < 1 \times 10^{-11}$). The rates for bivalents comprised of recombinations observed from two bivalents; the actual rates of cross-over formation *per bivalent* would be expected to be approximately the same, if we consider that potentially 50% of all cross-overs are unobserved (Figure 1.b). The findings here are consistent with the previously-reported rates of cross-overs per bivalent of approximately 1.1 in the autotetraploids *Physaria vitulifera* L., *Lotus corniculantus* L. and *Arabidopsis arenosa* L. (Mulligan, 1967;Davies et al., 1990;Yant et al., 2013). There is also the possibility of unobserved recombinations when multivalents occur, although an estimation of what proportion this represents is less obvious than for bivalents. It will be somewhat less than 50%, depending on the total number of cross-overs and how they are arranged between chromatids. Cross-overs from a quadrivalent structure can be hidden in the gamete if the

recombining chromatids involved sort to the same gamete (*e.g.* Chapter 1, Figure 3, gamete 4 the cross-over between the blue and orange homologues in that figure would have been invisible). Therefore, the estimated rates of recombination for the multivalent situation presented in Figure 1 are probably an underestimation, meaning that we have quite strong evidence here to suggest that cross-over rates in multivalents are higher than in bivalents. Once more this demonstrates the potential of experimental populations to help answer fundamental questions about polyploid meiosis.

It was also fascinating to note the occasional presence of offspring carrying chromosomal mosaics originating from three parental homologues (Figure 2). In the case of F1_090, the pattern we observe could also reflect a combination of mosaics of only two chromatids (6-8 and 8-7-8), although this would require three rather than two recombination break points, with independent cross-over positions precisely coinciding (which seems a less parsimonious explanation). This phenomenon ("tri-mosaicism") is curiously ignored in the current literature on polyploid meiosis. This is perhaps because it is considered of little practical relevance, but also because we didn't have the tools to recognise it until recently. Like double reduction, it is one of the clearest diagnostics for quadrivalent pairing using marker data.

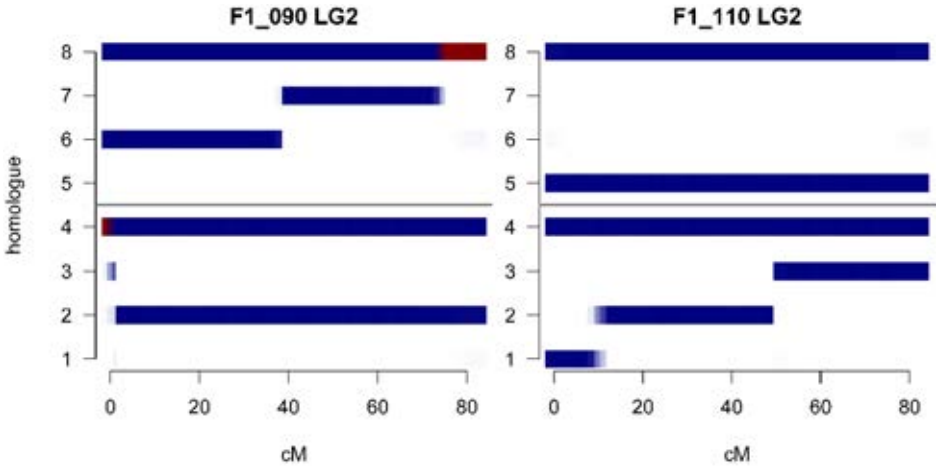



**Figure 2. Examples of offspring harbouring chromosomal mosaics of three parental homologues from potato chromosome 2 (AxC population).** Offspring F1_090 (left) carries such a mosaic inherited from parent 2 (involving homologues 6, 7 and 8), part of which is a double reduction product (coloured red). A similar mosaic was observed in offspring F1_110 (right), this time without involving double reduction.

# Opening the polysomic "black box"

There are many challenges in breeding polysomic polyploid crops. One of the major difficulties encountered is that every offspring of a cross contains an unknown combination of parental homologue mosaics. In inbreeding diploid populations, parental lines are often fully homozygous, in which case the transmission of parental homologues is essentially known. In outbreeding diploids this is not the case, but at any locus each individual carries one of $\binom{2}{1} \times \binom{2}{1} = 4$ possible allelic combinations. For tetraploids this becomes $\binom{4}{2} \times \binom{4}{2} = 36$ combinations (assuming no double reduction – it is 100 otherwise), while for hexaploids this rises to $\binom{6}{3} \times \binom{6}{3} = 400$ combinations (or 2500 if we include double reduction). When we start to consider multiple loci together (for example in the case of a trait influenced by a number of different genes) the number of possible allelic combinations explodes. Polysomic inheritance, coupled with high heterozygosity, leads to far more complex inheritance patterns than anything encountered at the diploid level.

Developing homozygous lines in polyploids would simplify matters considerably, both from a breeding and a genetics perspective. However, Haldane demonstrated the infeasibility of reaching homozygosity through repeated selfing (Haldane, 1930), a technique that is regularly used in diploid breeding and research (*c.f.* Figure 5 of Chapter 2). Apart from the impracticality of inbreeding, many polyploids are also outbreeding species that poorly tolerate inbreeding (tetraploid leek (*Allium ampeloprasum* L.) is a good example (De Clercq and Van Bockstaele, 2002)). This is generally thought to be caused by large numbers of deleterious recessive alleles (also called the *genetic load*, a point to which I later return). Because of these and related difficulties, the use of polyploids in breeding programs is sometimes viewed as an unnecessary complication. Efforts to convert potato from a highly-heterozygous outbreeding polyploid to an inbreeding, homozygous diploid are currently underway (Lindhout et al., 2011; Jansky et al., 2016). In sugar beet (*Beta vulgaris* L.) breeding, the use of tetraploid parental lines has been largely abandoned (Dewey, 1980;Draycott, 2008) because of difficulties in autopolyploid breeding. Previously, triploids and anisoploids (mixture of diploids, triploids and tetraploids from an open pollination) were developed for their superior yields, but these have largely been replaced by diploids (McGrath and Jung, 2016).

Nevertheless, it is unlikely that polysomic polyploids will completely disappear from breeding programs. In ornamental breeding, it is routine to induce polyploidisation while developing new varieties which has led to a preponderance of polyploidy in many ornamental breeding programs (Van Tuyl and Lim, 2003;Marasek-Ciolakowska et al.,

2016). Triploid fruit-bearing crops like banana (*Musa* spp.), watermelon (*Citrullus lanatus*) or grape (*Vitis* spp.) are attractive to breeders and consumers alike due to their seedlessness and are usually derived from a tetraploid × diploid cross. In forage species like alfalfa (*Medicago sativa* L.) and red clover (*Trifolium pratense* L.), tetraploids are generally considered to have favourable agronomic characteristics and are widely employed (Taylor, 2008). Therefore, instead of retreating from the challenges of polyploidy, we could also embrace the developments in technology and bioinformatics which are helping to turn the key to open the polysomic "black box".

One may wonder how we might reconcile the advantages of polyploid crops on the one hand with the difficulties in following their complex inheritance patterns on the other. For example, would it be possible to develop an economical marker set to track the inheritance of alleles in an autopolyploid breeding program, allowing a breeder to select offspring on genomic composition alone (and track inheritance over multiple generations)? Or is polysomic inheritance just too complex to decipher without the use of high-density marker datasets? To give an idea of whether the polysomic "black box" can actually be opened in an economically-feasible manner, consider autotetraploid potato, a crop species carrying twelve sets of chromosomes (2n = 4x = 48). Assuming bivalent pairing, if each bivalent had a single cross-over (the minimum requirement for stable meiosis) then 50% of all homologues carried in the gametes would be recombinant, with 50% non-recombinant. If we could develop markers at both the telomeres and centromeres of all chromosomes (Figure 3.a), we would already have quite a good, albeit low-resolution, picture of meiosis. The number of cross-overs per bivalent is rarely more than three (Mercier et al., 2015), which means that we are unlikely to find more than two recombination break-points along the length of a single chromatid. Of course, for fine-mapping or QTL analyses it is advantageous to have a higher resolution and that is why most QTL studies now employ vast numbers of markers (far more than are needed when the size of the mapping populations are considered).

But assuming our objective were to design a minimum marker set to track inheritance in an autotetraploid, we might employ DxD markers (Figure 3.a) since the have been shown to carry the most inheritance information in an autotetraploid (Zheng et al., 2016). Measures such as the root mean squared error (RMSE) (calculated using the deviations between the predicted and true genotypes) or the genotypic information content (GIC, described in Chapter 8) demonstrate that it is theoretically possible to have a reasonable picture of polysomic inheritance in autotetraploid potato with as few as 180 markers (*e.g.* using five DxD markers per chromosome, Figure 3.b). A mixture of marker types might even be better, and/or a staggering of marker positions – these variables could easily be

investigated by simulation. Given that the costs of genotyping are continually falling, the economics of such an exercise may start to make sense very soon.

Apart from understanding inheritance, this thesis has been primarily motivated by the development of tools to find *specific* markers to tag important traits. A breeder might rightly wonder why 180 markers should be used if one or five might be used instead. Indeed, specific markers that tag important loci (*e.g.* resistance genes or introgressed segments) would represent a more economical first use of marker technology. However, the argument that polysomic inheritance is an insurmountable hurdle to breeding programs needs to be put into its proper context given modern techniques. A breeder may not necessarily fully understand the mode of inheritance of the breeding material, which such a marker set could help reveal. As well as that, knowledge of genome-wide inheritance could help facilitate so-called "genomic selection", which has been advocated for polyploid breeding (*e.g.* (Slater et al., 2016)) but has yet to be shown to produce substantial gains in practice. Genomic selection theory as currently implemented assumes that all markers are potentially in linkage disequilibrium with genes that affect quantitative traits of interest (Meuwissen et al., 2001;Meuwissen et al., 2016), and therefore very large marker datasets are often used. More recently, haplotype-based approaches have been proposed which may offer greater accuracy and power than single-SNP approaches (Hess et al., 2017). Therefore, the selection of a set of highly-informative and well-spaced markers for a range of applications is likely to become a regular feature of polyploid breeding programs in the future.

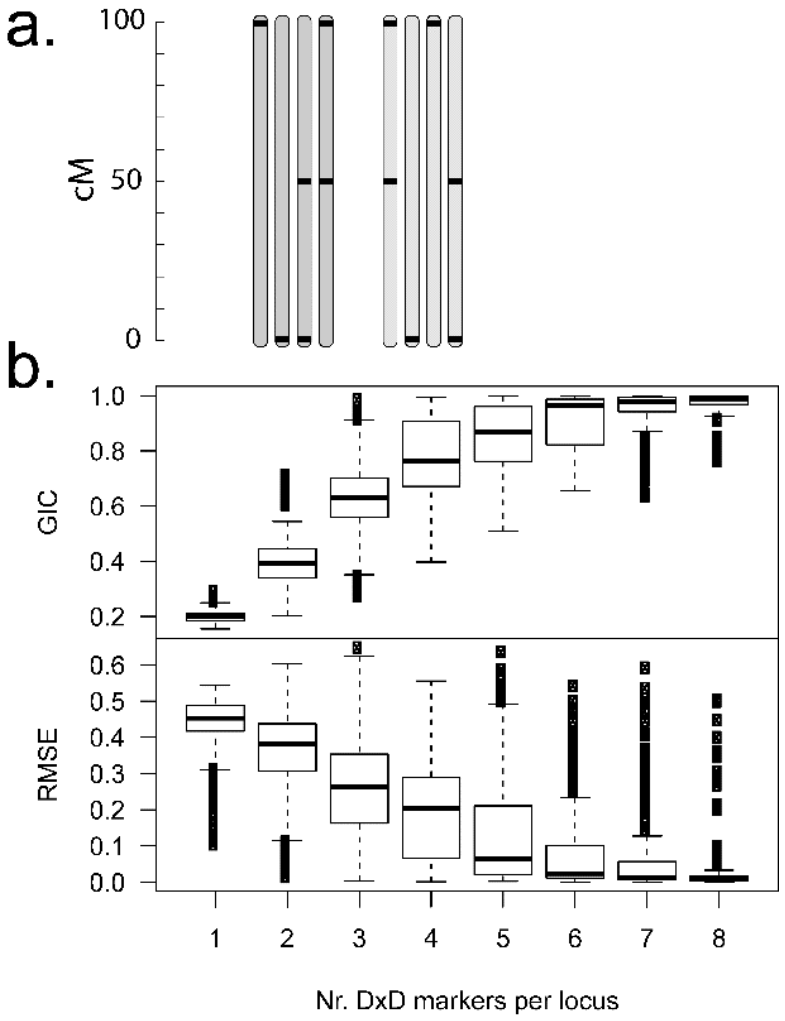



**Figure 3. Conceptual strategy for minimal marker set needed to track meiosis in an autotetraploid. a.** At 3 chosen positions (0, 50 and 100 cM) a single DxD marker is deployed. DxD markers are considered to provide the most inheritance information in an autotetraploid population (Zheng et al., 2016). **b.** Effectiveness of using a minimal marker set to track meiosis in an autotetraploid. The RMSE and GIC of genotype predictions associated with different marker numbers are compared, starting with 1 DxD marker at each of the three positions and progressing through to 8 per locus. Both the RMSE and the genotypic information coefficient (GIC) are affected by marker density. Populations were simulated in PedigreeSim (Voorrips and Maliepaard, 2012) and IBD probabilities were estimated in TetraOrigin (Zheng et al., 2016).

# Does double reduction add anything to the analysis?

One of the phenomena I looked at in some detail during this thesis was double reduction (DR), which it is claimed is the principal source of *systematic* marker segregation distortion in polyploid genetic studies (Luo et al., 2004). The non-randomness of this distortion arises from the fact that the rate of DR increases towards the telomeres (Mather, 1936;Fisher, 1947). In that sense the bias is positional, but unlike true segregation distortion that may be associated with alleles or allelic combinations of a particular locus, DR is combinatorically a random process which occurs with equal frequency on all homologues. As mentioned in Chapter 1, some authors like to emphasise the importance of including DR in their models for completeness (often castigating those who do not), but so far all have failed to substantiate these claims with actual data (Luo et al., 2004;Wu et al., 2004;Wu and Ma, 2005;Luo et al., 2006;Li et al., 2010;Lu et al., 2012;Xu et al., 2013).

In this thesis I began by looking at the bias that DR introduces in two-point linkage analyses using a random bivalent pairing model and simplex x nulliplex marker data (Chapter 2). I subsequently expanded this to include all possible autotetraploid marker segregation types (Chapter 3). I also compared the power and precision of a QTL model that included DR versus one that did not (Chapter 8). In the case of linkage mapping, DR introduces a small bias in the estimation of recombination frequencies, although it's negligible if the rate of quadivalent pairing is low. I argued that this bias could safely be ignored, particularly when we consider that linkage maps are primarily built from highly-informative linkages between nearby markers that are unlikely to be severely affected by a small number of unexpected genotypes such as DR introduces. Crucially, DR generates a novel offspring genotype class for simplex x nulliplex markers, namely duplex offspring, which are not expected in a random bivalent setting. In a mapping approach that ignores DR they effectively become missing values in the initial steps, which is quite a different proposition than segregation distortion or genotyping errors. Randomly-missing values will reduce the effective population size upon which recombination frequency is estimated, thereby increasing the variance and decreasing the LOD score. But this noise is essentially unbiased apart from positional considerations (it is certainly not skewness). Simplex x nulliplex markers form the centrepiece of the theoretical framework we have developed for polyploid linkage mapping (Chapters 4, 6) and are also the marker type that appear to occur in the greatest abundance in $F_1$ populations (Chapters 4, 5, 7). Therefore, no systematic deviations are introduced by DR to the principal structural elements of our linkage maps.

Regarding the prevalence of DR, it has been reported in a number of separate studies that DR is a rare phenomenon that occurs at low frequencies (Bradshaw et al., 2008;Stift et al., 2008;Hackett et al., 2013). Theoreticians like to argue over the maximum rate of DR, with suggestions ranging from $^1/_4$ (Luo et al., 2006;Bradshaw, 2007;Zheng et al., 2016), $^1/_6$ (Mather, 1935;Stift et al., 2010), $^1/_7$ (Voorrips and Maliepaard, 2012) and even $^1/_8$ (Sybenga, 1972). Bradshaw notes that the rate of DR depends on whether a ring or chain quadrivalent is involved, which complicates matters further (Bradshaw, 2007). In our work with tetraploid potato we found that the rate of DR reached a maximum of about 10 – 12% at the telomeres (Chapter 3), which is one of the few studies that actually quantifies the rate of DR in a population as opposed to merely mentioning it. A theoretical maximum rate of DR represents the expected rate at an infinite genetic distance from the centromere. Often, such rates are used to motivate more advanced descriptions of polyploid meiosis, although it is questionable whether such rates would ever occur in practice. Furthermore, if bivalent pairing is indeed somehow promoted by the polyploid cell it is even more unlikely that high rates of DR will occur (a point that is all but ignored by theoreticians). The consistent reports of low rates of DR can therefore be seen as further justification for the use of random bivalents to model polysomic inheritance.

In exploring the effects of ignoring double reduction (DR) in QTL mapping we found that the gains in power or precision were dwarfed by other considerations such as the trait heritability, the population size or the marker distribution near QTL regions (Chapter 8). Its inclusion was also found to adversely affect our ability to correctly predict the QTL segregation type (using either the Akaike Information Criterion (Akaike, 1974) or the Bayesian Information Criterion (Schwarz, 1978)). Once again we found that DR only played a minor role and could safely be ignored. However, a flexible model of inheritance that allows for the possibility of multivalents and DR can at times achieve much more accurate reconstructions of IBD probabilities over random bivalent models. This is apparent when one examines some of the output of TetraOrigin (Zheng et al., 2016) when applied to real datasets. For example, we found that individual F1_097 from the potato AxC population carried an extensive DR segment on chomosome 1. When we ran the analysis using a bivalent model only, the predicted haplotypes reveal a patchwork of segments across both sets of parental homologues (Figure 4.a). It is unlikely that such a high number of recombinations actually occurred. The individual F1_110 highlighted earlier as carrying a mosaic of three parental homologues (Figure 2) was also poorly reconstructed when only bivalent pairing structures were allowed (Figure 4.b). Given that the accuracy and power of our QTL analysis largely rests on the accuracy of the IBD probabilities, it is worrying to see this sort of behaviour. Here at last we have clear evidence that allowing for multivalents (including but not restricted to the phenomenon

of DR) yields significantly better results than assuming only bivalents occur. This does not contradict the conclusions of Chapter 8, where we recommended running both QTL models and comparing results. In summary, it appears that more complete models of polysomic meiosis and inheritance can have a big impact in specific instances, but when taken over a population of two hundred individuals or more the positive effects are generally negligible.

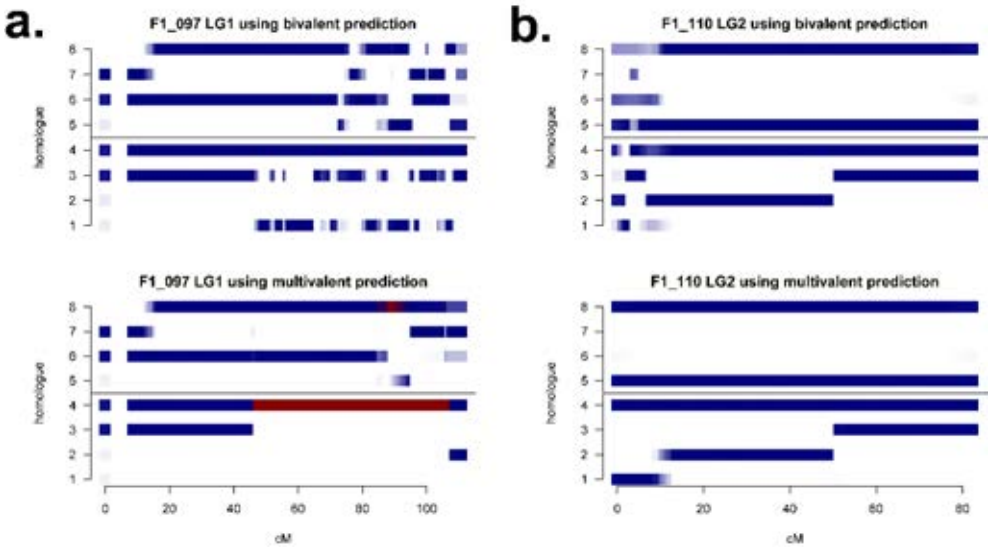



**Figure 4. Examples of where a multivalent model can out-perform a bivalent model in predicting offspring chromosome composition. a.** Offspring F1_097 carries a double reduction segment inherited from parent 1 (shown in red). If a bivalent model is used instead, TetraOrigin appears unable to cope. **b.** Double reduction is not the only unusual outcome of multivalent pairing. Here, offspring F1_110 carries a homologue inherited from segments of three parental homologues. Without the possibility of multivalents, TetraOrigin is again incapable of assigning reliable IBD probabilities in the affected region.

# Understanding the mode of inheritance

While taxonomists may argue over the classification of a polyploid species by speculating about the relatedness of diploid progenitors, geneticists and breeders are usually more interested in how a polyploid species reproduces, sometimes referred to as its *mode of inheritance*. In Chapter 1 the three known polyploid modes of inheritance were introduced – disomic, polysomic (also termed "tetrasomic" for tetraploids and "hexasomic" for hexaploids) and mixosomic. Mixosomy can be interpreted as either a mixture of disomic and polysomic inheritance at separate genomic regions, or as an intermediate between the two extreme types resulting from partial preferential

chromosome pairing, and is the mode of inheritance associated with segmental allopolyploidy (Stebbins, 1947;Soltis et al., 2016). In many important polyploid crop species the mode of inheritance is fully agreed upon (examples include wheat, oilseed rape (disomic) and potato (tetrasomic)), while in other species it is not (examples include sugarcane, leek and the ornamental species *Alstroemeria*). It may be that there is no fixed mode of inheritance for a species that, once diagnosed, will predict all future meiotic behaviour of members of the same species. We have already encountered an example of this in rose: the mode of inheritance varied between the parents of the K5 mapping population, with both disomic and polysomic inheritance detected for different chromosomes in both parents (Chapter 5).

One may wonder whether the mode of inheritance is merely of academic interest – do we need to worry about it as long as crosses can be made and viable seeds produced? If we would like to employ markers in our breeding program the answer is clearly yes, understanding the mode of inheritance is vital. Genotyping polyploids is in general more complex and less reliable than genotyping diploids (Mason, 2015). This could be caused by off-target amplification due to duplication of marker sequences at other loci (Limborg et al., 2016), but is also due in large part to the difficulty in correctly estimating the number of copies of a marker allele, also known as marker dosage assignment (Chapter 2). One of the standard practices when assigning dosage is to compare marker scores between parents and offspring. This acts as both a quality-check, by comparing the expected and observed allele frequencies, but also acts as a means of determining the correct range of dosage values (Voorrips et al., 2011;Hackett et al., 2013;Schmitz Carley et al., 2017). In order to compute the expected allele frequencies in a polyploid population given the parental dosage scores, one needs to know the mode of inheritance *in advance*. This is potentially one of the greatest hurdles in the development of genomic resources for novel polyploid research projects. When additional issues such as aneuploidy occur (including the possibility of compensated aneuploidy (Ising, 1966;Chester et al., 2012)), it may be very hard to interpret marker data, with no solid ground upon which to build a hypothesis. However, all such issues can potentially be resolved if they are understood, but require pioneering genotyping (and possibly cytological) studies to pave the way for future breeding and research applications.

Understanding the mode of inheritance is also required for the construction of phased, integrated linkage maps. We demonstrated this through simulation studies in Chapters 3 and 4, developing a flexible approach to linkage analysis in Chapter 5 which can accommodate fluctuations in the mode of inheritance. Novel mapping algorithms that do not require estimates of recombination frequency ("model-free" approaches) might be more robust. However, so far none have demonstrated the ability to generate *phased*

linkage maps, of fundamental importance in polyploid genetics (Chapter 2). Therefore, in order to use current mapping tools such as TetraploidSNPMap (Hackett et al., 2017) or polymapR (Chapter 6) the mode of inheritance of the population must be established. TetraploidSNPMap is only suitable for tetrasomic species, whereas polymapR has been developed to accommodate all modes of inheritance from disomy through to tetrasomy (but at the tetraploid level only). Regarding mixosomic linkage mapping in non-tetraploids, it should be noted that extending the linkage mapping framework within polymapR to include hexasomic inheritance was itself a major undertaking, necessitating 105 linkage functions that required over 1.9 million characters of R code (R Core Team, 2016). In contrast, tetrasomy required only 26 linkage functions and just over 250,000 R characters. Including the possibility of partial preferential pairing for tetraploids required almost 1.2 million characters, representing a 5-fold increase over the tetrasomic functions. For hexaploidy, this fold-increase can be expected to be much greater than five given the increased complexity. Therefore as currently implemented, allowing for mixosomy in hexaploids would be a gargantuan undertaking. One solution would be to implement a fully-general likelihood framework rather than implementing each likelihood function separately, similar to the approach used in TetraploidSNPMap (C.A. Hackett, personal communication). Currently however, implementing dynamic algebra in R is cumbersome and that is why pre-programmed solutions, generated in Mathematica, were used instead.

We demonstrated how important the mode of inheritance was for unbiased linkage analyses in Chapter 4. How the mode of inheritance affects QTL detection models was not investigated in this thesis, although IBD-based approaches could be expected to be quite robust against deviations from a polysomic model. However this was not verified. There therefore remains more work to do in properly accounting for the varied modes of inheritance in our polyploid genetic mapping pipeline.

# Improving power and precision

One of the main preoccupations of diploid bio-statisticians is to improve their models to glean more predictive power from the data they are presented with. This hunger for bigger and better has yielded many fine developments over the years. For linkage mapping it is typical to consider two-point analyses as an initial step towards more accurate three-point or multi-point analyses (Ridout et al., 1998;Leach et al., 2010;Tong et al., 2010). For QTL mapping, more powerful approaches might include mixed model methodologies, Bayesian mapping or machine-learning algorithms (Wang et al., 2014a;Xavier et al., 2016), or employing more complex or pedigreed populations (Bink et al., 2008;Cavanagh et al., 2008;Huang et al., 2015). In this thesis we did not venture beyond two-point linkage analysis, simple linear models or $F_1$ bi-parental mapping

populations. There appears to be plenty of opportunities to explore more complex (and potentially more powerful) approaches at detecting and deciphering QTL effects in polyploid populations.

However, we are now living in a time when the developments in genotyping technologies are dictating the development of statistical methodologies in genetic mapping and not the other way around. Perhaps this has always been the case, but I feel the pace of change has increased in recent decades. No sooner have we established an effective pipeline to perform genetic mapping in autopolyploids using SNP data than it becomes potentially redundant due to a shift towards genotyping-by-sequencing (GBS) data (Elshire et al., 2011), optical mapping data (Goodwin et al., 2016), Hi-C data (Lieberman-Aiden et al., 2009) *etc*. We are all guilty of technology-chasing, as nobody wants to be seen to be stuck working with yesteryears' tools. On the other hand, by putting all our hope in a technological fix rather than a methodological fix to improve power or precision issues, we risk becoming locked into a cycle of using sub-optimal methods to keep pace.

Particularly in polysomic polyploids where the complexity is already substantial for "basic" methods and models, it seems implausible to have sufficient time to replicate the whole suite of approaches that have been developed for diploid species in the time afforded by inter-technological cool-off periods. Each new genotyping technology carries with it its own set of challenges and characteristics, from different ways to scoring marker dosage (Voorrips et al., 2011;van Dijk et al., 2012;Blischak et al., 2016) to differences in the amount and type of missing data or genotyping errors (Elshire et al., 2011;Bajgain et al., 2016). The later steps of linkage mapping, QTL analysis, genome-wide association mapping *etc.* cannot be performed without the initial pre-processing of genotype data, the tools for which can often appear many years after the first introduction of the technology (diploid tools tend to be developed sooner, often by the genotyping provider but also by the much larger research community engaged in diploid genotyping experiments). Unless a stable genotyping technology for polyploids comes along, the polyploid genetics community is unlikely to ascend to the upper echelons of diploid biostatistics before building a new ladder from scratch again. Currently, improvements in power and precision of genetic mapping experiments are being driven by improvements in genotyping approaches (although the production of robust and accurate phenotypic data remains a critical component as well).

# Where next for polysomic polyploid breeding?

To the early agriculturalists, the polyploid nature of their crops probably wasn't a major hindrance given the relatively simple techniques that were employed in the selection and maintenance of germplasm. Modern plant breeding as a science and art did not evolve until relatively recently, and therefore our ancestors were hardly troubled by issues such as polysomic inheritance, partial preferential chromosome pairing or double reduction (indeed they were most likely unaware of these issues in the first place). Nowadays, society is demanding the production of more with less, given the projected increase in the world's population coupled with dwindling fossil fuel resources, increased climatic instability and the degradation of agricultural land (Bradshaw, 2016;Jez et al., 2016). These appear to be insurmountable challenges, yet we have at our disposal a much wider array of intellectual resources than ever before. The question is whether these can be translated into effective solutions quickly enough, particularly given that many of our most important agricultural crops are polyploid.

Perhaps the single most important trait within all breeding programs across all crops is yield, a trait that is notoriously complex even within diploids. It has been suggested that despite 150 years of breeding for increased yield, phenotypic selection methods have done nothing to raise average potato yields in that time (Jansky, 2009;Slater et al., 2016), in contrast to other major diploid crops like maize or rice. On the other hand, some authors remain convinced about the merit of polyploids and their potential to deliver higher yields through heterosis (Renny-Byfield and Wendel, 2014). I don't believe that polyploids lack the necessary qualities to rise to the challenges we are facing. However, I do believe we need to carefully consider how polyploids are currently bred and what the implications are for future breeding material given our present selection decisions.

One of the main difficulties in developing elite breeding material in polysomic polyploids is their *genetic load*. Here, genetic load refers to lethal or debilitating recessive mutations that are masked in the heterozygous condition but which carry a high fitness cost when present in homozygous condition. Although it is claimed that neo-autopolyploids have a lower genetic load than their progenitor diploids (Otto and Whitton, 2000), in time it builds up to greater levels in established autopolyploids, leading to inbreeding depression (Ronfort, 1999;Otto, 2007). This is not to be confused with self-incompatibility which, if present, is usually under the control of one or more specific self-incompatibility loci (Ridout et al., 2005;Jansky et al., 2016). Polyploids tolerate higher genetic loads than diploids, as for example elegantly demonstrated in a study of diploid and tetraploid pollen vigour (Husband, 2016). Possibilities to select parents with a lower genetic load include evaluating derived haploids (which may express the deleterious alleles) or by self-pollinating parental lines and observing the

proportion of non-vigorous offspring (Jansky, 2009). However, it is unlikely that stringent selection against genetic load is routinely performed in polyploid breeding programs. Without the benefits of purging selection, polysomic polyploids can be expected to slowly accumulate deleterious mutations which may reduce the quality of breeding lines for many future generations.

One solution would be to simply avoid the effects of genetic load by increasing the diversity of breeding material, thereby ensuring heterozygosity at as many loci as possible. The development of different heterotic pools is one possibility to achieve this in. For example the AxC tetraploid potato population (Chapters 3, 4, 8), a cross between a starch and a table variety, represents a population that normally would not be generated in a breeding program (since these breeding pools are usually kept distinct (Vos et al., 2015)). In AxC we observed transgressive segregation for a number of important traits, demonstrating that wide crosses may generate unexpected and favourable results. Alternatively, genome-wide marker data could be used to predict heterotic effects, by crossing parents known to carry different alleles across multiple loci.

An alternative to maximising heterozygosity is to try to reduce the genetic load. We have seen that inbreeding is not a practical proposition at the polyploid level (Haldane, 1930); it therefore seems we might have to "clean" polyploids at the haploid or diploid level by exposing deleterious mutations to selection. However if we were to go that route, returning to the polyploid level might not be needed (Lindhout et al., 2011;Renny-Byfield and Wendel, 2014;Jansky et al., 2016). Reducing genetic load at the polyploid level might be aided by genomic tools but it would take a significant (and perhaps unrealistic) effort to pinpoint all such loci, over multiple traits, and exclude them from future breeding germplasm using marker data. Nevertheless, the identification of some major lethality loci using the tools developed in this thesis could be a very useful research aim of breeding programs (for example by identifying regions of skewed segregation in selfed progenies and associating these with reconstructed parental haplotypes in these regions). By tracking and purging such alleles from breeding material, breeders could enjoy slightly more freedom to mildly-inbreed their parental lines, as well as improving the overall fitness of their breeding stock. This may have direct implications to improve complex traits like yield, but could more generally be seen as improving the elite germplasm within a crop, with positive knock-on effects for a whole range of traits.

In considering the future of polysomic plant breeding it is important to distinguish between two categories, namely vegetatively-propagated and seed-propagated crops. Examples of the former include potato and chrysanthemum, while examples of the latter include alfalfa, red clover and leek. The potato cultivar Russett's Burbank (introduced in 1902 as "Netted Gem" (Bethke et al., 2014)) is an example of a single genotype that

possessed an above-average combination of alleles for multiple traits that has led to it become North America's most widely-cultivated potato variety, despite all the competition from cultivars that have been bred in the intervening century (Hardigan et al., 2017). In the U.K, Maris Piper (introduced in 1966) continues to be the most widely-grown variety, while in the Netherlands Bintje (introduced in 1910) still holds this honour. Vegetative or clonal propagation immortalises favourable allelic combinations and heterotic effects which can lead to a single variety having a very long production life. However, clonal propagation also comes with its own inherent disadvantages such as a slow rate of seed stock multiplication, increased risk of spreading soil-borne diseases or endemic viral pathogens, more costly distribution and storage of propagation material and difficulties in maintaining germplasm in genebanks. Despite these drawbacks, clonal propagation continues to dominate in crops like potato, chrysanthemum and rose.

The breeding of seed-propagated polysomic crops is arguably much more challenging given the varietal registration requirements of distinctness, uniformity and stability (DUS). Obtaining uniform progeny through seed can only be reliably predicted if parental lines are homozygous at all loci affecting descriptor traits (these usually concern the morphology of the adult plant). For example, uniformity continues to be one of the main breeding goals for autotetraploid leek (*Allium ampeloprasum*) (De Clercq and Van Bockstaele, 2002). In recent years there has been a move from open-pollinated varieties to $F_1$ hybrids in leek breeding. Clonally-propagated male-sterile plants as maternal lines restricts the amount of selfing that occurs, thought to account for up to 20-30% of open pollinations (De Clercq and Van Bockstaele, 2002). Given the extremely high genetic load in leek, selfed progeny are likely to be the biggest contributor to lack of uniformity, with substantial loss of yield after only one or two rounds of self-pollination (De Clercq and Van Bockstaele, 2002). Therefore, the use of specialised $F_1$ hybrid breeding is another example of how breeders have devised strategies to overcome the high genetic load in polysomic polyploids.

Another example of a seed-producing autopolyploid is the autotetraploid red clover (*Trifolium pratense* L.). Tetraploid clovers are often favoured over diploids due to their higher yields, greater persistence and higher levels of resistance to biotic and abiotic stresses (Taylor, 2008;Vleugels et al., 2016). However, the main drawback to tetraploid cultivars is their reduced seed yield (Vleugels et al., 2016). The precise cause(s) of this reduction remain unknown despite many years of investigation. The mechanism could be related to problems during meiosis or abnormal pollen tube growth (Büyükkartal, 2002;2008), but it has also been proposed that flower morphology (particularly the length of the corolla) may result in lower pollination levels in tetraploid clovers (Bender, 1999;Furuya, 2001). This trait represents one of the most important breeding aims in

tetraploid clover breeding which has until now proved extremely difficult to understand or improve. With the deployment of molecular markers across tetraploid populations, it could be possible to determine the genetic control (if any) of seed number in tetraploid clover, which could lead to more profitable and reliable seed yields.

One of the more interesting possibilities for polysomic polyploid breeding that has been suggested is the use of *apomictic* reproduction. Apomixis can be defined as the production of seeds from maternal tissues, bypassing normal sexual reproduction (Comai, 2005;Bicknell and Catanach, 2015). There are numerous reports of the association of gametophytic apomixis with polyploidy (Wet and Harlan, 1970;Whitton et al., 2008), with a number of possible theories proposed (both mechanistic and evolutionary) but no consensus on the precise reason why this should be (Quarin et al., 2001;Comai, 2005;Zielinski and Scheid, 2012). Whatever the connection, the potential attractiveness of programmable apomixis for breeding has widely been acknowledged (Spillane et al., 2004;Abdi et al., 2016;Bicknell et al., 2016;Bradshaw, 2016), and not just at the polyploid level. Just like clonal propagation, apomixis would allow the fixation of heterosis and has been suggested as a superior method to produce $F_1$ hybrids by immortalising the $F_1$ rather than recreating it each season through repeated crossing of inbred parental lines. This would combine the advantages of true seed production while avoiding the numerous disadvantages of vegetative clones as already discussed. Apomixis has been shown to be under genetic control in multiple species (reviewed in (Bicknell and Catanach, 2015)), with a number of recent genetic mapping studies in tropical polyploid forage grasses such as the tetraploid *Brachiaria decumbens* or the hexaploid *Urochloa humidicola* revealing major QTL underlying the trait (Vigna et al., 2016;Worthington et al., 2016). A nice example of where apomixis is actively used in polyploid breeding is in Kentucky bluegrass (*Poa pratensis*) (Huff, 2010). In this species, apomixis is the norm whereas true outcrossings are termed “aberrants”. Techniques to increase the production of aberrants during the breeding cycle are used (Huff, 2010;Bradshaw, 2016), following which superior lines are selected. Most progeny are found to be apomictic once more, resulting in immortal maternal lines which can subsequently be marketed as stable $F_1$ hybrids. Although it might seem like an ideal scenario, the breeding of apomicts presents its own set of challenges, particularly if the number of aberrants generated is too low to allow crossing. Aneuploidy may also accompany apomixis as it is not selected against during meiosis (Huff, 2010), which may lead to problems in future crosses if the aneuploid is anything other than the apomictic parent. Finding sources of apomixis genes in related and crossable species is a further challenge, if transgenic approaches to introducing the trait are not feasible or acceptable. In this search, the use of molecular markers and genomic resources (through candidate gene approaches) will be vital. Although apomixis may offer many advantages, there has

yet to be an example of it being successfully bred into a polyploid crop and commercially exploited. Time will tell whether this form of breeding will be adopted by the polyploid community or not.

Returning finally to the humble potato, the apparent lack of progress in yield over the past 150 years can perhaps be best explained in terms of polysomic inheritance, with each new variety carrying a reshuffled set of alleles (as well as perhaps some specific disease resistance loci introgressed from wild species). I have (hopefully) demonstrated that polysomic inheritance need no longer be considered a black box. We are entering a time where autopolyploid breeders will have the choice to no longer just accept the shuffled hand they are dealt, but to choose their own hand based on (among other factors) marker data. Only time will tell whether this will become an intrinsic part of polyploid breeding. Economics, and the ease at which markers can be incorporated into the infrastructure of existing breeding programs, will ultimately dictate the future direction of polyploid breeding.

# Concluding remarks

In this thesis I have tried to strike a balance between the development of tools to help unravel the complexities of polyploid genetics, and what these tools can provide to breeders and researchers. I do not consider these tools as ends in themselves, but rather see them as a means to achieve certain goals, be it the identification of markers linked to important traits, the creation of linkage maps representing the molecular karyotype of a population, or the understanding of polyploid meiosis including how pairing and recombination proceed. We have seen that such tools are extremely powerful at opening up a window into polyploid genetics, which I believe will in time allow for many more genomics-informed breeding decisions to be made. We are just at the beginning of a new era in polyploid breeding, with many exciting opportunities ahead.

# References


Abdi, S., Dwivedi, A., and Bhat, V. (2016). "Harnessing apomixis for heterosis breeding in crop improvement," in *Molecular Breeding for Sustainable Crop Improvement*. (Springer), 79-99.

Acquaah, G. (2012). *Principles of Plant Genetics and Breeding*. Wiley-Blackwell.

Aguiar, D., and Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* 29**,** i352-i360.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control* 19**,** 716-723.

Akhunov, E., Nicolet, C., and Dvorak, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theoretical and Applied Genetics* 119**,** 507-517.

Al-Janabi, S.M., Honeycutt, R.J., and Sobral, B.W.S. (1994). Chromosome assortment in Saccharum. *Theoretical and Applied Genetics* 89**,** 959-963.

Albrechtsen, A., Nielsen, F.C., and Nielsen, R. (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution* 27**,** 2534-2547.

Allendorf, F.W., Bassham, S., Cresko, W.A., Limborg, M.T., Seeb, L.W., and Seeb, J.E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity* 106**,** 217-227.

Allendorf, F.W., and Danzmann, R.G. (1997). Secondary tetrasomic segregation of *MDH-B* and preferential pairing of homeologues in rainbow trout. *Genetics* 145**,** 1083-1092.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25,3389-3402.

Andrade, J., and Estévez-Pérez, M. (2014). Statistical comparison of the slopes of two regression lines: A tutorial. *Analytica chimica acta* 838**,** 1-12.

Bajgain, P., Rouse, M.N., and Anderson, J.A. (2016). Comparing genotyping-by-sequencing and single nucleotide polymorphism chip genotyping for quantitative trait loci mapping in wheat. *Crop science* 56**,** 232-248.

Baker, P. (2014). polySegratio: Simulate and test marker dosage for dominant markers in autopolyploids. R package version 0.2-4.

Baker, P., Jackson, P., and Aitken, K. (2010). Bayesian estimation of marker dosage in sugarcane and other autopolyploids. *Theoretical and applied genetics* 120**,** 1653-1672.

Balsalobre, T.W.A., Da Silva Pereira, G., Margarido, G.R.A., Gazaffi, R., Barreto, F.Z., Anoni, C.O., et al. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC genomics* 18**,** 72.

Barker, M.S., Arrigo, N., Baniaga, A.E., Li, Z., and Levin, D.A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 210**,** 391-398.

Barringer, B.C. (2007). Polyploidy and self-fertilization in flowering plants. *American Journal of Botany* 94**,** 1527-1533.

Bartholomé, J., Mandrou, E., Mabiala, A., Jenkins, J., Nabihoudine, I., Klopp, C., et al. (2015). High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytologist* 206**,** 1283-1296.

Bassil, N.V., Davis, T.M., Zhang, H., Ficklin, S., Mittmann, M., Webster, T., et al. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *BMC Genomics* 16**,** 155.

Beçak, M.L., Beçak, W., and Rabello, M.N. (1966). Cytological evidence of constant tetraploidy in the bisexual South American frog *Odontophrynus americanus*. *Chromosoma* 19**,** 188-193.

Behrouzi, P., and Wit, E.C. (2017a). De novo construction of q-ploid linkage maps using discrete graphical models. *arXiv preprint arXiv:1710.01063*.

Behrouzi, P., and Wit, E.C. (2017b). netgwas: An R Package for Network-Based Genome-Wide Association Studies. *arXiv preprint arXiv:1710.01236*.

Benabdelmouna, A., and Ledu, C. (2015). Autotetraploid Pacific oysters (*Crassostrea gigas*) obtained using normal diploid eggs: induction and impact on cytogenetic stability. *Genome* 58**,** 333-348.

Bendahmane, M., Just, J., Vergne, P., Raymond, O., Dubois, A., Szcesi, J., et al. (Year). "The Rose Genome Sequencing Initiative, Prospects and Perspectives. Abstract W659", in: *Plant and Animal Genome XXIV Conference*: https://pag.confex.com/pag/xxiv/webprogram/Paper19873.html).

Bender, A. (1999). An impact of morphological and physiological transformations of red clover flowers accompanying polyploidization on the pollinators' working speed and value as a guarantee for cross pollination. *Agraarteadus*, 4, 9-23.

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

Bennetzen, J.L. (2007). Patterns in grass genome evolution. *Current Opinion in Plant Biology* 10, 176-181.

Bennetzen, J.L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology* 65, 505-530.

Berger, E., Yorukoglu, D., Peng, J., and Berger, B. (2014). Haptree: A novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS computational biology* 10, e1003502.

Bernardo, R. (2008). Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. *Crop Science* 48, 1649-1664.

Bernardo, R. (2016). Bandwagons I, too, have known. *Theoretical and Applied Genetics* 129, 2323-2332.

Bertioli, D.J., Cannon, S.B., Froenicke, L., Huang, G., Farmer, A.D., Cannon, E.K., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics* 48, 438-446.

Bertioli, D.J., Ozias-Akins, P., Chu, Y., Dantas, K.M., Santos, S.P., Gouvea, E., et al. (2014). The use of SNP markers for linkage mapping in diploid and tetraploid peanuts. *G3: Genes| Genomes| Genetics* 4, 89-96.

Bethke, P.C., Nassar, A.M., Kubow, S., Leclerc, Y.N., Li, X.-Q., Haroon, M., et al. (2014). History and origin of Russet Burbank (Netted Gem) a sport of Burbank. *American journal of potato research* 91, 594-609.

Bicknell, R., and Catanach, A. (2015). "Apomixis: the asexual formation of seed," in *Somatic Genome Manipulation*. (Springer), 147-167.

Bicknell, R., Catanach, A., Hand, M., and Koltunow, A. (2016). Seeds of doubt: Mendel's choice of *Hieracium* to study inheritance, a case of right plant, wrong trait. *Theoretical and Applied Genetics* 129, 2253-2266.

Bingham, E.T., and Gillies, C.B. (1971). Chromosome pairing, fertility, and crossing behavior of haploids of tetraploid alfalfa, *Medicago sativa* L. *Canadian Journal of Genetics and Cytology* 13, 195-202.

Bink, M., Boer, M., Ter Braak, C., Jansen, J., Voorrips, R., and Van De Weg, W. (2008). Bayesian analysis of complex traits in pedigreed plant populations. *Euphytica* 161, 85-96.

Blakeslee, A.F., and Avery, A.G. (1937). Methods of inducing doubling of chromosomes in plants: By treatment with colchicine. *Journal of Heredity* 28, 393-411.

Blischak, P.D., Kubatko, L.S., and Wolfe, A.D. (2016). Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Molecular ecology resources* 16, 742-754.

Bomblies, K., Higgins, J.D., and Yant, L. (2015). Meiosis evolves: adaptation to external and internal environments. *New Phytologist* 208, 306-323.

Bomblies, K., Jones, G., Franklin, C., Zickler, D., and Kleckner, N. (2016). The challenge of evolving stable polyploidy: could an increase in "crossover interference distance" play a central role? *Chromosoma* 125, 287-300.

Bourke, P.M. (2014). QTL analysis in polyploids. *MSc thesis, Wageningen University, Wageningen*.

Bourke, P.M., Arens, P., Voorrips, R.E., Esselink, G.D., Koning-Boucoiran, C.F.S., Van 'T Westende, W.P.C., et al. (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal* 90, 330-343.

Bourke, P.M., Voorrips, R.E., Kranenburg, T., Jansen, J., Visser, R.G., and Maliepaard, C. (2016). Integrating haplotype-specific linkage maps in tetraploid species using SNP markers. *Theoretical and Applied Genetics* 129, 2211-2226.

Bourke, P.M., Voorrips, R.E., Visser, R.G.F., and Maliepaard, C. (2015). The Double Reduction Landscape in Tetraploid Potato as Revealed by a High-Density Linkage Map. *Genetics* 201, 853-863.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633-2635.

Bradshaw, J. (2007). The canon of potato science: 4. Tetrasomic inheritance. *Potato Research* 50, 219-222.

Bradshaw, J.E. (2016). *Plant Breeding: Past, Present and Future.* New York: Springer.

Bradshaw, J.E., Hackett, C.A., Pande, B., Waugh, R., and Bryan, G.J. (2008). QTL mapping of yield, agronomic and quality traits in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*). *Theoretical and Applied Genetics* 116, 193-211.

Bradshaw, J.E., Pande, B., Bryan, G.J., Hackett, C.A., Mclean, K., Stewart, H.E., et al. (2004). Interval mapping of quantitative trait loci for resistance to late blight [*Phytophthora infestans* (Mont.) de Bary], height and maturity in a tetraploid population of potato (*Solanum tuberosum* subsp. *tuberosum*). *Genetics* 168**,** 983-995.

Brent, R. (1973). "Algorithms for minimizing without derivatives". Prentice-Hall, Englewood Cliffs (NJ.).

Broman, K.W., Wu, H., Sen, Ś., and Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19**,** 889-890.

Brouwer, D., and Osborn, T. (1999). A molecular marker linkage map of tetraploid alfalfa (*Medicago sativa* L.). *Theoretical and Applied Genetics* 99**,** 1194-1200.

Bruneau, A., Starr, J.R., and Joly, S. (2007). Phylogenetic relationships in the genus *Rosa*: new evidence from chloroplast DNA sequences and an appraisal of current knowledge. *Systematic Botany* 32**,** 366-378.

Brzustowicz, L., Merette, C., Xie, X., Townsend, L., Gilliam, T., and Ott, J. (1993). Molecular and statistical approaches to the detection and correction of errors in genotype databases. *American Journal of human Genetics* 53**,** 1137-1145.

Buitendijk, J.H., Boon, E.J., and Ramanna, M.S. (1997). Nuclear DNA Content in Twelve Species of *Alstroemeria* L. and Some of their Hybrids. *Annals of Botany* 79**,** 343-353.

Bushman, B., Robbins, M., Larson, S., and Staub, J. (2016). "Genotyping by Sequencing in Autotetraploid Cocksfoot (*Dactylis glomerata*) without a Reference Genome," in *Breeding in a World of Scarcity*. (Springer), 133-137.

Butruille, D., and Boiteux, L. (2000). Selection–mutation balance in polysomic tetraploids: impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proceedings of the National Academy of Sciences* 97**,** 6608-6613.

Büyükkartal, H.N.B. (2002). In Vitro Pollen Germination and Pollen Tube Characteristics in Tetraploid Red Clover (*Trifolium pratense* L.). *Turkish Journal of Botany* 27**,** 57-61.

Büyükkartal, H.N.B. (2008). Causes of low seed set in the natural tetraploid *Trifolium pratense* L. (Fabaceae). *African Journal of Biotechnology* 7**,** 1240.

Cadman, C. (1943). Nature of tetraploidy in cultivated European potatoes. *Nature* 152**,** 103-104.

Cao, D., Osborn, T.C., and Doerge, R.W. (2004). Correct estimation of preferential chromosome pairing in autotetraploids. *Genome Research* 14**,** 459-462.

Carvalho, A., Delgado, M., Barão, A., Frescatada, M., Ribeiro, E., Pikaard, C.S., et al. (2010). Chromosome and DNA methylation dynamics during meiosis in the autotetraploid *Arabidopsis arenosa*. *Sexual plant reproduction* 23**,** 29-37.

Carvalho, D.R., Koning-Boucoiran, C.F., Fanourakis, D., Vasconcelos, M.W., Carvalho, S.M., Heuvelink, E., et al. (2015). QTL analysis for stomatal functioning in tetraploid *Rosa× hybrida* grown at high relative air humidity and its implications on postharvest longevity. *Molecular Breeding* 35**,** 1-11.

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Current opinion in plant biology* 11**,** 215-221.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences* 110**,** 8057-8062.

Cervantes-Flores, J.C., Yencho, G.C., Kriegner, A., Pecota, K.V., Faulk, M.A., Mwanga, R.O., et al. (2008). Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose AFLP markers. *Molecular Breeding* 21**,** 511-532.

Chaffin, A.S., Huang, Y.-F., Smith, S., Bekele, W.A., Babiker, E., Gnanesh, B.N., et al. (2016). A consensus map in cultivated hexaploid oat reveals conserved grass synteny with substantial subgenome rearrangement. *The plant genome* 9.

Chakravarti, A. (1991). A graphical representation of genetic and physical maps: the Marey map. *Genomics* 11**,** 219-222.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345**,** 950-953.

Chang, K.Y., Lo, H.F., Lai, Y.C., Yao, P.J., Lin, K.H., and Hwang, S.-Y. (2009). Identification of quantitative trait loci associated with yield-related traits in sweet potato (*Ipomoea batatas*). *Botanical studies* 50**,** 43-55.

Cheema, J., and Dicks, J. (2009). Computational approaches and software tools for genetic linkage map estimation in plants. *Briefings in bioinformatics* 10**,** 595-608.

Chen, F.-D., Li, F.-T., Chen, S.-M., Guan, Z.-Y., and Fang, W.-M. (2009). Meiosis and pollen germinability in small-flowered anemone type chrysanthemum cultivars. *Plant systematics and evolution* 280**,** 143.

Chester, M., Gallagher, J.P., Symonds, V.V., Da Silva, A.V.C., Mavrodiev, E.V., Leitch, A.R., et al. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proceedings of the National Academy of Sciences* 109**,** 1176-1181.

Churchill, G.A., and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138**,** 963-971.

Cifuentes, M., Grandont, L., Moore, G., Chèvre, A.M., and Jenczewski, E. (2010). Genetic regulation of meiosis in polyploid species: new insights into an old question. *New Phytologist* 186**,** 29-36.

Clarke, W.E., Higgins, E.E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., et al. (2016). A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theoretical and Applied Genetics* 129**,** 1887-1899.

Collard, B.C., and Mackill, D.J. (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**,** 557-572.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6**,** 836-846.

Cornille, A., Salcedo, A., Kryvokhyzha, D., Glémin, S., Holm, K., Wright, S.I., et al. (2016). Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). *Molecular ecology* 25**,** 616-629.

Crawford, J., Brown, P.J., Voigt, T., and Lee, D. (2016). Linkage mapping in prairie cordgrass (*Spartina pectinata* Link) using genotyping-by-sequencing. *Molecular breeding* 36**:**62.

Crespel, L., Chirollet, M., Durel, C., Zhang, D., Meynet, J., and Gudin, S. (2002). Mapping of qualitative and quantitative phenotypic traits in *Rosa* using AFLP markers. *Theoretical and Applied Genetics* 105**,** 1207-1214.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* 1695**,** 1-9.

D'hont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenetic and genome research* 109**,** 27-33.

D'hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488**,** 213-217.

Da Ines, O., Gallego, M.E., and White, C.I. (2014). Recombination-independent mechanisms and pairing of homologous chromosomes during meiosis in plants. *Molecular plant* 7**,** 492-501.

Da Silva, J.a.G., Sorrells, M.E., Burnquist, W.L., and Tanksley, S.D. (1993). RFLP linkage map and genome analysis of *Saccharum spontaneum*. *Genome* 36**,** 782-791.

Dalton-Morgan, J., Hayward, A., Alamery, S., Tollenaere, R., Mason, A.S., Campbell, E., et al. (2014). A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Functional & Integrative Genomics* 14**,** 643-655.

Darlington, C.D. (1937). *Recent advances in cytology.* J And A Churchill; London.

Das, S., and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC genomics* 16**:**260.

Davies, A., Jenkins, G., and Rees, H. (1990). Diploidisation of *Lotus corniculatus* L. (Fabaceae) by elimination of multivalents. *Chromosoma* 99**,** 289-295.

De Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature genetics* 37**,** 1217-1223.

De Clercq, H., and Van Bockstaele, E. (2002). "Leek: Advances in agronomy and breeding," in *Allium crop science: recent advances,* (eds. H.D. Rabinowitch & L. Currah.), 431-458.

De Jong, J. (1984). Genetic analysis in *Chrysanthemum morifolium*. I. Flowering time and flower number at low and optimum temperature. *Euphytica* 33**,** 455-463.

De Storme, N., and Geelen, D. (2014). The impact of environmental stress on male reproductive development in plants: biological processes and molecular mechanisms. *Plant, cell & environment* 37**,** 1-18.

Debener, T. (1999). Genetic analysis of horticulturally important morphological and physiological characters in diploid roses. *Gartenbauwissenschaft* 64**,** 14-19.

Debener, T., and Linde, M. (2009). Exploring complex ornamental genomes: the rose as a model plant. *Critical reviews in plant sciences* 28**,** 267-280.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345**,** 1181-1184.

Dewey, D.R. (1966). Inbreeding depression in diploid, tetraploid, and hexaploid crested wheatgrass. *Crop Science* 6**,** 144-147.

Dewey, D.R. (1980). "Some Applications and Misapplications of Induced Polyploidy to Plant Breeding," in *Polyploidy: Biological Relevance,* ed. W.H. Lewis. (Boston, MA: Springer US), 445-470.

Didion, J.P., Yang, H., Sheppard, K., Fu, C.-P., Mcmillan, L., De Villena, F.P.-M., et al. (2012). Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC genomics* 13**:**34.

Doyle, J.J., and Egan, A.N. (2010). Dating the origins of polyploidy events. *New Phytologist* 186**,** 73-85.

Doyle, J.J., and Sherman-Broyles, S. (2016). Double trouble: taxonomy and definitions of polyploidy. *New Phytologist,* 213, 487-493.

Draycott, A.P. (2008). *Sugar beet.* John Wiley & Sons.

Dresselhaus, T., and Johnson, M.A. (2018). Reproduction: Plant Parentage à Trois. *Current Biology* 28**,** R28-R30.

Eberhard, F.S., Zhang, P., Lehmensiek, A., Hare, R.A., Simpfendorfer, S., and Sutherland, M.W. (2010). Chromosome composition of an F2 *Triticum aestivum* × *T. turgidum* spp. durum cross analysed by DArT markers and MCFISH. *Crop and Pasture Science* 61**,** 619-624.

Edae, E.A., Bowden, R.L., and Poland, J. (2015). Application of population sequencing (POPSEQ) for ordering and imputing genotyping-by-sequencing markers in hexaploid wheat. *G3: Genes, Genomes, Genetics* 5**,** 2547-2553.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* 6**,** e19379.

Endelman, J.B. (2011). New algorithm improves fine structure of the barley consensus SNP map. *BMC Genomics* 12**:**407.

Endelman, J.B., and Plomion, C. (2014). LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* 30**,** 1623-1624.

Esselink, G., Smulders, M., and Vosman, B. (2003). Identification of cut rose (*Rosa hybrida*) and rootstock varieties using robust sequence tagged microsatellite site markers. *Theoretical and Applied Genetics* 106**,** 277-286.

Evans, B.J., Upham, N.S., Golding, G.B., Ojeda, R.A., and Ojeda, A.A. (2017). Evolution of the Largest Mammalian Genome. *Genome Biology and Evolution* 9**,** 1711-1724.

Felcher, K.J., Coombs, J.J., Massa, A.N., Hansey, C.N., Hamilton, J.P., Veilleux, R.E., et al. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One* 7**,** e36347.

Fierst, J.L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in Genetics* 6:220.

Fisher, R.A. (1947). The theory of linkage in polysomic inheritance. *Philosophical Transactions of the Royal Society B: Biological Sciences* 233**,** 55-87.

Fjellstrom, R., Beuselinck, P., and Steiner, J. (2001). RFLP marker analysis supports tetrasonic inheritance in *Lotus corniculatus* L. *Theoretical and Applied Genetics* 102**,** 718-725.

Flint-Garcia, S.A., Thornsberry, J.M., and Buckler Iv, E.S. (2003). Structure of linkage disequilibrium in plants. *Annual review of plant biology* 54**,** 357-374.

Fougère-Danezan, M., Joly, S., Bruneau, A., Gao, X.-F., and Zhang, L.-B. (2015). Phylogeny and biogeography of wild roses with specific attention to polyploids. *Annals of Botany* 115**,** 275-291.

Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A., and Mummenhoff, K. (2011). Cabbage family affairs: the evolutionary history of *Brassicaceae*. *Trends in plant science* 16**,** 108-116.

Furuya, H. (2001). *Comparisons of seed weight and seedling characteristics of diploid and autotetraploid red clover.* Division of Colleges & Universities, Florida Board of Education.

Gajardo, H.A., Wittkop, B., Soto-Cerda, B., Higgins, E.E., Parkin, I.a.P., Snowdon, R.J., et al. (2015). Association mapping of seed quality traits in *Brassica napus* L. using GWAS and candidate QTL approaches. *Molecular Breeding* 35**:**143.

Gallardo, M.H., Bickham, J., Honeycutt, R., Ojeda, R., and Köhler, N. (1999). Discovery of tetraploidy in a mammal. *Nature* 401**,** 341-341.

Galloway, L.F., and Etterson, J.R. (2007). Inbreeding depression in an autotetraploid herb: a three cohort field study. *New Phytologist* 173**,** 383-392.

Galloway, L.F., Etterson, J.R., and Hamrick, J.L. (2003). Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, *Campanula americana*. *Heredity*, 90**,** 308-315.

Gar, O., Sargent, D.J., Tsai, C.-J., Pleban, T., Shalev, G., Byrne, D.H., et al. (2011). An autotetraploid linkage map of rose (*Rosa hybrida*) validated using the strawberry (*Fragaria vesca*) genome sequence. *PLoS One* 6**,** e20463.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J., and Anderson, L.K. (2007). Recombination: an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics* 8**,** 77-84.

Geldermann, H. (1975). Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. *Theoretical and Applied Genetics* 46**,** 319-330.

Gidskehaug, L., Kent, M., Hayes, B.J., and Lien, S. (2010). Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics* 27**,** 303-310.

Gillies, A.C., Cubas, P., Coen, E.S., and Abbott, R.J. (2002). "Making rays in the Asteraceae: genetics and evolution of radiate versus discoid flower heads," in *Developmental genetics and plant evolution,* (eds. Q. Cronk, R. Bateman & J. Hawkins. Taylor and Francis, London), 233-246.

Gitonga, V.W., Koning-Boucoiran, C.F., Verlinden, K., Dolstra, O., Visser, R.G., Maliepaard, C., et al. (2014). Genetic variation, heritability and genotype by environment interaction of morphological traits in a tetraploid rose population. *BMC Genetics* 15**:**1.

Gitonga, V.W., Stolker, R., Koning-Boucoiran, C.F.S., Aelaei, M., Visser, R.G.F., Maliepaard, C., et al. (2016). Inheritance and QTL analysis of the determinants of flower color in tetraploid cut roses. *Molecular Breeding* 36**:**143.

Glover, N.M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: what are they and how do we infer them? *Trends in plant science* 21**,** 609-621.

Goldschmidt, R. (1933). Some aspects of evolution. *Science* 78**,** 539-547.

Goodwin, S., Mcpherson, J.D., and Mccombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17**,** 333-351.

Grandke, F., Ranganathan, S., Van Bers, N., De Haan, J.R., and Metzler, D. (2017). PERGOLA: fast and deterministic linkage mapping of polyploids. *BMC Bioinformatics* 18**:**12.

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

Grandont, L., Jenczewski, E., and Lloyd, A. (2013). Meiosis and its deviations in polyploid plants. *Cytogenetic and genome research* 140**,** 171-184.

Griffiths, A.J., Wessler, S.R., Carroll, S.B., and Doebley, J. (2012). *An introduction to genetic analysis (10th edition).* Macmillan.

Grivet, L., and Arruda, P. (2002). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Current Opinion in Plant Biology* 5**,** 122-127.

Grivet, L., D'hont, A., Roques, D., Feldmann, P., Lanaud, C., and Glaszmann, J.C. (1996). RFLP Mapping in Cultivated Sugarcane (*Saccharum* spp.): Genome Organization in a Highly Polyploid and Aneuploid Interspecific Hybrid. *Genetics* 142**,** 987-1000.

Guimarães, C.T., Sills, G.R., and Sobral, B.W.S. (1997). Comparative mapping of *Andropogoneae*: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. *Proceedings of the National Academy of Sciences* 94**,** 14261-14266.

Hackett, C.A., Bradshaw, J., and McNicol, J. (2001). Interval mapping of quantitative trait loci in autotetraploid species. *Genetics* 159**,** 1819-1832.

Hackett, C.A., Bradshaw, J., Meyer, R., McNicol, J., Milbourne, D., and Waugh, R. (1998). Linkage analysis in tetraploid species: a simulation study. *Genetics Research* 71**,** 143-153.

Hackett, C.A., and Luo, Z. (2003). TetraploidMap: construction of a linkage map in autotetraploid species. *Journal of Heredity* 94**,** 358-359.

Hackett, C.A., Pande, B., and Bryan, G. (2003). Constructing linkage maps in autotetraploid species using simulated annealing. *Theoretical and Applied Genetics* 106**,** 1107-1115.

Hackett, C.A., Boskamp, B., Vogogias, A., Preedy, K.F., and Milne, I. (2017). TetraploidSNPMap: Software for Linkage Analysis and QTL Mapping in Autotetraploid Populations Using SNP Dosage Data. *Journal of Heredity* 108**,** 438-442.

Hackett, C.A., Bradshaw, J.E., and Bryan, G.J. (2014). QTL mapping in autotetraploids using SNP dosage information. *Theoretical and Applied Genetics* 127**,** 1885-1904.

Hackett, C.A., and Broadfoot, L.B. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* 90**,** 33-38.

Hackett, C.A., Mclean, K., and Bryan, G.J. (2013). Linkage Analysis and QTL Mapping Using SNP Dosage Data in a Tetraploid Potato Mapping Population. *PLoS One* 8**,** e63939.

Hackett, C.A., Milne, I., Bradshaw, J.E., and Luo, Z. (2007). TetraploidMap for Windows: linkage map construction and QTL mapping in autotetraploid species. *Journal of Heredity* 98**,** 727-729.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* 8**,** 299-309.

Haldane, J.B. (1930). Theoretical genetics of autopolyploids. *Journal of Genetics* 22**,** 359-372.

Hamilton, J.P., Hansey, C.N., Whitty, B.R., Stoffel, K., Massa, A.N., Van Deynze, A., et al. (2011). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics* 12**,** 302.

Hamilton, M.G., and Kerr, R.J. (2018). Computation of the inverse additive relationship matrix for autopolyploid and multiple-ploidy populations. *Theoretical and Applied Genetics* 131, 851-860.

Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B., et al. (2017). Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences* 114:46.

Harlan, J.R., and De Wet, J.M.J. (1975). On Ö. Winge and a prayer: the origins of polyploidy. *The botanical review* 41**,** 361-390.

Harrison, C.J., Alvey, E., and Henderson, I.R. (2010). Meiosis in flowering plants and other green organisms. *Journal of experimental botany* 61**,** 2863-2875.

Haynes, K., and Douches, D. (1993). Estimation of the coefficient of double reduction in the cultivated tetraploid potato. *Theoretical and applied genetics* 85**,** 857-862.

Hazarika, M., and Rees, H. (1967). Genotypic control of chromosome behaviour in rye. X. Chromosome pairing and fertility in autotetraploids. *Heredity* 22**,** 317-332.

He, Y., Xu, X., Tobutt, K.R., and Ridout, M.S. (2001). Polylink: to support two-point linkage analysis in autotetraploids. *Bioinformatics* 17**,** 740-741.

Henderson, K.A., and Keeney, S. (2004). Tying synaptonemal complex initiation to the formation and programmed repair of DNA double-strand breaks. *Proceedings of the National Academy of Sciences of the United States of America* 101**,** 4519-4524.

Herben, T., Suda, J., and Klimešová, J. (2017). Polyploid species rely on vegetative reproduction more than diploids: a re-examination of the old hypothesis. *Annals of Botany* 120**,** 341-349.

Herklotz, V., and Ritz, C. (2016). Multiple and asymmetrical origin of polyploid dog rose hybrids (*Rosa* L. sect. *Caninae* (DC.) Ser.) involving unreduced gametes. *Annals of Botany* 120**,** 209-220.

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genetics Selection Evolution* 49**:**54.

Hibrand-Saint Oyant, L., Crespel, L., Rajapakse, S., Zhang, L., and Foucher, F. (2008). Genetic linkage maps of rose constructed with new microsatellite markers and locating QTL controlling flowering traits. *Tree Genetics & Genomes* 4**,** 11-23.

Hirsch, C.D., Hamilton, J.P., Childs, K.L., Cepela, J., Crisovan, E., Vaillancourt, B., et al. (2014). Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *The Plant Genome* 7.

Hospital, F. (2009). Challenges for effective marker-assisted selection in plants. *Genetica* 136**,** 303-310.

Huang, B.E., Verbyla, K.L., Verbyla, A.P., Raghavan, C., Singh, V.K., Gaur, P., et al. (2015). MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics* 128**,** 999-1017.

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., et al. (2013). Draft genome of the kiwifruit *Actinidia chinensis*. *Nature communications* 4**:**2640.

Huff, D.R. (2010). "Bluegrasses," in *Fodder crops and amenity grasses,* (eds. B. Boller, U.K. Posselt & F. Veronesi. Springer, New York), 345-379.

Hulse-Kemp, A.M., Lemm, J., Plieske, J., Ashrafi, H., Buyyarapu, R., Fang, D.D., et al. (2015). Development of a 63K SNP Array for Cotton and High-Density Mapping of Intra- and Inter-Specific Populations of *Gossypium* spp. *G3: Genes|Genomes|Genetics* 5, 1187-1209.

Hunter, N. (2015). Meiotic recombination: the essence of heredity. *Cold Spring Harbor perspectives in biology* 7**,** a016618.

Husband, B.C. (2000). Constraints on polyploid evolution: a test of the minority cytotype exclusion principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267**,** 217-223.

Husband, B.C. (2016). Effect of inbreeding on pollen tube growth in diploid and tetraploid *Chamerion angustifolium*: Do polyploids mask mutational load in pollen? *American journal of botany* 103**,** 532-540.

International Wheat Genome Sequencing Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:6194.

Ising, G. (1966). Cytogenetic studies in *Cyrtanthus* I. Segregation in an allotetraploid. *Hereditas* 56**,** 27-53.

Islam, M.S., Thyssen, G.N., Jenkins, J.N., and Fang, D.D. (2015). Detection, validation, and application of genotyping-by-sequencing based single nucleotide polymorphisms in Upland cotton. *The Plant Genome* 8.

Jannoo, N., Grivet, L., David, J., D'hont, A., and Glaszmann, J. (2004). Differential chromosome pairing affinities at meiosis in polyploid sugarcane revealed by molecular markers. *Heredity* 93**,** 460-467.

Jannoo, N., Grivet, L., Dookun, A., D'hont, A., and Glaszmann, J.C. (1999). Linkage disequilibrium among modern sugarcane cultivars. *Theoretical and Applied Genetics* 99**,** 1053-1060.

Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* 135**,** 205-211.

Jansky, S. (2009). "Chapter 2 - Breeding, Genetics, and Cultivar Development," in *Advances in Potato Chemistry and Technology*. (San Diego: Academic Press), 27-62.

Jansky, S.H., Charkowski, A.O., Douches, D.S., Gusmini, G., Richael, C., Bethke, P.C., et al. (2016). Reinventing potato as a diploid inbred line–based crop. *Crop Science* 56**,** 1412-1422.

Jarvis, D.E., Ho, Y.S., Lightfoot, D.J., Schmöckel, S.M., Li, B., Borm, T.J.A., et al. (2017). The genome of *Chenopodium quinoa*. *Nature* 542**:**307.

Jenczewski, E., Eber, F., Grimaud, A., Huet, S., Lucas, M.O., Monod, H., et al. (2003). *PrBn*, a major gene controlling homeologous pairing in oilseed rape (*Brassica napus*) haploids. *Genetics* 164**,** 645-653.

Jez, J.M., Lee, S.G., and Sherp, A.M. (2016). The next green movement: plant biology for the environment and sustainability. *Science* 353**,** 1241-1244.

Jiang, C.-X., Wright, R.J., El-Zik, K.M., and Paterson, A.H. (1998). Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). *Proceedings of the National Academy of Sciences* 95**,** 4419-4424.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473**,** 97-100.

John, B., and Henderson, S. (1962). Asynapsis and polyploidy in *Schistocerca paranensis*. *Chromosoma* 13**,** 111-147.

Joly, S., Bryant, D., and Lockhart, P.J. (2015). Flexible methods for estimating genetic distances from single nucleotide polymorphisms. *Methods in Ecology and Evolution* 6**,** 938-948.

Joly, S., Starr, J.R., Lewis, W.H., and Bruneau, A. (2006). Polyploid and hybrid evolution in roses east of the Rocky Mountains. *American Journal of Botany* 93**,** 412-425.

Jones, D.B., Jerry, D.R., Khatkar, M.S., Raadsma, H.W., Van Der Steen, H., Prochaska, J., et al. (2017). A comparative integrated gene-based linkage and locus ordering by linkage disequilibrium map for the Pacific white shrimp, *Litopenaeus vannamei*. *Scientific Reports* 7.

Jones, G., and Vincent, J. (1994). Meiosis in autopolyploid *Crepis capillaris*. II. Autotetraploids. *Genome* 37**,** 497-505.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* 24**,** 1384-1395.

Kantarski, T., Larson, S., Zhang, X., Dehaan, L., Borevitz, J., Anderson, J., et al. (2017). Development of the first consensus genetic map of intermediate wheatgrass (*Thinopyrum intermedium*) using genotyping-by-sequencing. *Theoretical and Applied Genetics* 130**,** 137-150.

Kaur, S., Francki, M.G., and Forster, J.W. (2012). Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant biotechnology journal* 10**,** 125-138.

Kempthorne, O. (1957). *An introduction to genetic statistics.* John Wiley & Sons, New York.

Kerr, R.J., Li, L., Tier, B., Dutkowski, G.W., and Mcrae, T.A. (2012). Use of the numerator relationship matrix in genetic analysis of autopolyploid species. *Theoretical and applied genetics* 124**,** 1271-1282.

Khawaja, H.I.T., Sybenga, J., and Ellis, J.R. (1997). Chromosome pairing and chiasma formation in autopolyploids of different *Lathyrus* species. *Genome* 40**,** 937-944.

Kihara, H., and Ono, T. (1926). Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. *Zeitschrift für Zellforschung und mikroskopische Anatomie* 4**,** 475-481.

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., and Paterson, A.H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science* 242**,** 14-22.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., et al. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature genetics* 39**,** 1151-1155.

Kirov, I., Van Laere, K., De Riek, J., De Keyser, E., Van Roy, N., and Khrustaleva, L. (2014). Anchoring linkage groups of the *Rosa* genetic map to physical chromosomes with Tyramide-FISH and EST-SNP markers. *PloS one* 9**,** e95793.

Kloosterman, B., Abelenda, J.A., Gomez, M.D.M.C., Oortwijn, M., De Boer, J.M., Kowitwanich, K., et al. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature* 495**:**246.

Knott, S., Neale, D., Sewell, M., and Haley, C. (1997). Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. *Theoretical and Applied Genetics* 94**,** 810-820.

Knott, S.A., and Haley, C.S. (1992). Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* 132**,** 1211-1222.

Koning-Boucoiran, C.F.S., Esselink, G.D., Vukosavljev, M., Van't Westende, W.P.C., Gitonga, V.W., Krens, F.A., et al. (2015). Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa* L.). *Frontiers in Plant Science* 6**:**249.

Koning-Boucoiran, C.F.S., Gitonga, V.W., Yan, Z., Dolstra, O., Van Der Linden, C.G., Van Der Schoot, J., et al. (2012). The mode of inheritance in tetraploid cut roses. *Theoretical and Applied Genetics* 125**,** 591-607.

Koopman, W.J., Vosman, B., Sabatino, G.J., Visser, D., Van Huylenbroeck, J., De Riek, J., et al. (2008). AFLP markers as a tool to reconstruct complex relationships in the genus *Rosa* (*Rosaceae*). *American Journal of Botany* 95**,** 353-366.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* 9**:**29.

Kostoff, D. (1940). Fertility and chromosome length. Correlations between chromosome length and viability of gametes in autopolyploid plants. *Journal of Heredity* 31**,** 33-34.

Krebs, S.L., and Hancock, J.F. (1990). Early-acting inbreeding depression and reproductive success in the highbush blueberry, *Vaccinium corymbosum* L. *Theoretical and Applied Genetics* 79**,** 825-832.

Kriegner, A., Cervantes, J.C., Burg, K., Mwanga, R.O., and Zhang, D. (2003). A genetic linkage map of sweetpotato [*Ipomoea batatas* (L.) Lam.] based on AFLP markers. *Molecular Breeding* 11**,** 169-185.

Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research* 19**,** 1639-1653.

Lachance, J., and Tishkoff, S.A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35**,** 780-786.

Lamm, R. (1945). Cytogenetic studies in *Solanum*, sect. Tuberarium. *Hereditas* 31**,** 1-128.

Lampson, M.A., and Cheeseman, I.M. (2011). Sensing centromere tension: Aurora B and the regulation of kinetochore function. *Trends in cell biology* 21**,** 133-140.

Le Comber, S., Ainouche, M., Kovarik, A., and Leitch, A. (2010). Making a functional diploid: from polysomic to disomic inheritance. *New Phytologist* 186**,** 113-122.

Leach, L.J., Wang, L., Kearsey, M.J., and Luo, Z. (2010). Multilocus tetrasomic linkage analysis using hidden Markov chain model. *Proceedings of the National Academy of Sciences* 107**,** 4270-4274.

Leal-Bertioli, S., Shirasawa, K., Abernathy, B., Moretzsohn, M., Chavarro, C., Clevenger, J., et al. (2015). Tetrasomic recombination is surprisingly frequent in allotetraploid *Arachis*. *Genetics* 199**,** 1093-1105.

Lentz, E., Crane, C., Sleper, D., and Loegering, W. (1983). An assessment of preferential chromosome pairing at meiosis in *Dactylis glomerata*. *Canadian Journal of Genetics and Cytology* 25**,** 222-232.

Levin, D.A. (1975). Minority cytotype exclusion in local plant populations. *Taxon***,** 35-43.

Levin, D.A. (1983). Polyploidy and Novelty in Flowering Plants. *The American Naturalist* 122**,** 1-25.

Lewin, K. (1951). *Field theory in social science: selected theoretical papers.* London: Tavistock.

Li-Marchetti, C., Le Bras, C., Chastellier, A., Relion, D., Morel, P., Sakr, S., et al. (2017). 3D phenotyping and QTL analysis of a complex character: rose bush architecture. *Tree Genetics & Genomes* 13**:**112.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., et al. (2014a). Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nature genetics* 46**,** 567-572.

Li, J., Das, K., Fu, G., Tong, C., Li, Y., Tobias, C., et al. (2010). EM algorithm for mapping quantitative trait loci in multivalent tetraploids. *International Journal of Plant Genomics* 2010, 1-10.

Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A.D., Ho, J., et al. (2014b). Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS One* 9**,** e84329.

Li, X., Li, L., and Yan, J. (2015). Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nature Communications* 6**:**6648.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326**,** 289-293.

Limborg, M.T., Seeb, L.W., and Seeb, J.E. (2016). Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Molecular ecology* 25**,** 2117-2129.

Linde, M., Hattendorf, A., Kaufmann, H., and Debener, T. (2006). Powdery mildew resistance in roses: QTL mapping in different environments using selective genotyping. *Theoretical and applied genetics* 113**,** 1081-1092.

Lindhout, P., Meijer, D., Schotte, T., Hutten, R.C., Visser, R.G., and Van Eck, H.J. (2011). Towards F1 hybrid seed potato breeding. *Potato Research* 54**,** 301-312.

Liorzou, M., Pernet, A., Li, S., Chastellier, A., Thouroude, T., Michel, G., et al. (2016). Nineteenth century French rose (*Rosa* sp.) germplasm shows a shift over time from a European to an Asian genetic background. *Journal of experimental botany* 67**,** 4711-4725.

Liu, B., Poulsen, E.G., and Davis, T.M. (2016). Insight into octoploid strawberry (*Fragaria*) subgenome composition revealed by GISH analysis of pentaploid hybrids. *Genome* 59**,** 79-86.

Lloyd, A., and Bomblies, K. (2016). Meiosis in autopolyploid and allopolyploid *Arabidopsis*. *Current opinion in plant biology* 30**,** 116-122.

Lu, Y., Yang, X., Tong, C., Li, X., Feng, S., Wang, Z., et al. (2012). A multivalent three-point linkage analysis model of autotetraploids. *Briefings in bioinformatics* 14**,** 460-468.

Luo, Z., Hackett, C., Bradshaw, J., Mcnicol, J., and Milbourne, D. (2001). Construction of a genetic linkage map in tetraploid species using molecular markers. *Genetics* 157**,** 1369-1385.

Luo, Z., Zhang, R., and Kearsey, M. (2004). Theoretical basis for genetic linkage analysis in autotetraploid species. *Proceedings of the National Academy of Sciences of the United States of America* 101**,** 7040-7045.

Luo, Z.W., Zhang, Z., Leach, L., Zhang, R.M., Bradshaw, J.E., and Kearsey, M.J. (2006). Constructing Genetic Linkage Maps Under a Tetrasomic Model. *Genetics* 172**,** 2635-2645.

Mable, B., Alexandrou, M., and Taylor, M. (2011). Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology* 284**,** 151-182.

Maenhout, S. (2018). "Progeno". (Ghent University, Belgium).

Maliepaard, C., Jansen, J., and Van Ooijen, J. (1997). Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. *Genetical Research* 70**,** 237-250.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., et al. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7**,** 111-118.

Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., and Lysak, M.A. (2010). Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *The Plant Cell* 22**,** 2277-2290.

Mangelsdorf, P.C., and Jones, D.F. (1926). The expression of Mendelian factors in the gametophyte of maize. *Genetics* 11**:**423.

Marasek-Ciolakowska, A., Arens, P., and Van Tuyl, J. (2016). "The Role of Polyploidization and Interspecific Hybridization in the Breeding of Ornamental Crops," in *Polyploidy and Hybridization for Crop Improvement*. CRC Press), 159-181.

Mason, A.S. (2015). "Challenges of Genotyping Polyploid Species," in *Plant Genotyping: Methods and Protocols,* ed. J. Batley. (New York, NY: Springer New York), 161-168.

Mason, A.S., Higgins, E.E., Snowdon, R.J., Batley, J., Stein, A., Werner, C., et al. (2017). A user guide to the Brassica 60K Illumina Infinium™ SNP genotyping array. *Theoretical and Applied Genetics* 130**,** 621-633.

Massa, A.N., Manrique-Carpintero, N.C., Coombs, J.J., Zarka, D.G., Boone, A.E., Kirk, W.W., et al. (2015). Genetic linkage mapping of economically important traits in cultivated tetraploid potato (*Solanum tuberosum* L.). *G3: Genes| Genomes| Genetics* 5**,** 2357-2364.

Mather, K. (1935). Reductional and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics* 30**,** 53-78.

Mather, K. (1936). Segregation and linkage in autotetraploids. *Journal of Genetics* 32**,** 287-314.

McAllister, C.A., and Miller, A.J. (2016). Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. *American journal of botany* 103**,** 1314-1325.

McCallum, S., Graham, J., Jorgensen, L., Rowland, L.J., Bassil, N.V., Hancock, J.F., et al. (2016). Construction of a SNP and SSR linkage map in autotetraploid blueberry using genotyping by sequencing. *Molecular Breeding* 36**:**41.

McCamy, P., Holloway, H., Yu, X., Dunne, J.C., Schwartz, B.M., Patton, A.J., et al. (2018). A SNP-based high-density linkage map of zoysiagrass (*Zoysia japonica* Steud.) and its use for the identification of QTL associated with winter hardiness. *Molecular Breeding* 38:10.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226**,** 792-801.

McCollum, G.D. (1957). Comparative studies of chromosome pairing in natural and induced tetraploid *Dactylis*. *Chromosoma* 9**,** 571-605.

McDonald, M.J., Rice, D.P., and Desai, M.M. (2016). Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* 531**,** 233-236.

McGrath, J.M., and Jung, C. (2016). "Use of polyploids. interspecific, and intergeneric wide hybrids in sugar beet improvement," in *Polyploidy and Hybridization for Crop Improvement*. (CRC Press), 408-420.

Mendel, J.G. (1866). Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn* Bd. IV Abhandlungen**,** 3–47.

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N., and Grelon, M. (2015). The molecular biology of meiosis in plants. *Annual review of plant biology* 66**,** 297-327.

Meuwissen, T., Hayes, B., and Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal frontiers* 6**,** 6-14.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157**,** 1819-1829.

Meyer, R., Milbourne, D., Hackett, C., Bradshaw, J., Mcnichol, J., and Waugh, R. (1998). Linkage analysis in tetraploid potato and association of markers with quantitative resistance to late blight (*Phytophthora infestans*). *Molecular and General Genetics* 259**,** 150-160.

Microsoft Corporation, and Weston, S. (2017). doParallel: Foreach parallel adaptor for the 'parallel' package. *R package version* 1.0.11.

Milbourne, D., Bradshaw, J.E., and Hackett, C.A. (2008). "Molecular Mapping and Breeding in Polyploid Crop Plants," in *Principles and Practices of Plant Genomics (Vol. 2: Molecular Breeding),* (eds. C. Kole & A.G. Abbott. Science Publishers), 355 - 394.

Moncada, M.D.P., Tovar, E., Montoya, J.C., González, A., Spindel, J., and Mccouch, S. (2016). A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. *Tree genetics & genomes* 12**:**5.

Monden, Y., Hara, T., Okada, Y., Jahana, O., Kobayashi, A., Tabuchi, H., et al. (2015). Construction of a linkage map based on retrotransposon insertion polymorphisms in sweetpotato via high-throughput sequencing. *Breeding science* 65**,** 145-153.

Monroe, J.G., Allen, Z.A., Tanger, P., Mullen, J.L., Lovell, J.T., Moyers, B.T., et al. (2017). TSPmap, a tool making use of traveling salesperson problem solvers in the efficient and accurate construction of high-density genetic linkage maps. *BioData mining* 10**:**38.

Moore, G. (2013). "Meiosis in polyploids," in *Polyploid and hybrid genomics,* (eds. Z.J. Chen & J.A. Birchler. JohnWiley & Sons, Inc.), 241-255.

Moragues, M., Comadran, J., Waugh, R., Milne, I., Flavell, A., and Russell, J.R. (2010). Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theoretical and applied genetics* 120**,** 1525-1534.

Morgan, T.H. (1911). Random segregation versus coupling in Mendelian inheritance. *Science* 34**,** 384-384.

Morrison, J., and Rajhathy, T. (1960). Frequency of quadrivalents in autotetraploid plants. *Nature* 187**:**528.

Motazedi, E., De Ridder, D., Finkers, R., and Maliepaard, C. (2017a). TriPoly: a haplotype estimation approach for polyploids using sequencing data of related individuals. *bioRxiv, doi 10.1101/163162*.

Motazedi, E., Finkers, R., Maliepaard, C., and De Ridder, D. (2017b). Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in bioinformatics***,** 19, 387-403.

Muller, H. (1925). Why polyploidy is rarer in animals than in plants. *The American Naturalist* 59**,** 346-353.

Mulligan, G.A. (1967). Diploid and autotetraploid *Physaria vitulifera* (Cruciferae). *Canadian Journal of Botany* 45**,** 183-188.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., et al. (2015). Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome biology* 16**:**262.

Myers, W. (1943). Analysis of variance and covariance of chromosomal association and behavior during meiosis in clones of *Dactylis glomerata*. *Botanical Gazette* 104**,** 541-552.

Naranjo, T., and Corredor, E. (2008). Nuclear architecture and chromosome dynamics in the search of the pairing partner in meiosis in plants. *Cytogenetic and genome research* 120**,** 320-330.

Neath, A.A., and Cavanaugh, J.E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics* 4**,** 199-203.

Neigenfind, J., Gyetvai, G., Basekow, R., Diehl, S., Achenbach, U., Gebhardt, C., et al. (2008). Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC genomics* 9**,** 356.

Nemorin, A., Abraham, K., David, J., and Arnau, G. (2012). Inheritance pattern of tetraploid *Dioscorea alata* and evidence of double reduction using microsatellite marker segregation analysis. *Molecular Breeding* 30**,** 1657-1667.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461**,** 272-276.

Nguepjop, J.R., Tossim, H.-A., Bell, J.M., Rami, J.-F., Sharma, S., Courtois, B., et al. (2016). Evidence of genomic exchanges between homeologous chromosomes in a cross of peanut with newly synthetized allotetraploid hybrids. *Frontiers in Plant Science* 7:1635.

Nicklas, R.B. (1997). How cells get the right chromosomes. *Science* 275**,** 632-637.

Nicolas, S.D., Leflon, M., Monod, H., Eber, F., Coriton, O., Huteau, V., et al. (2009). Genetic regulation of meiotic cross-overs between related genomes in *Brassica napus* haploids and hybrids. *Plant Cell* 21**,** 373-385.

Ohno, S. (1970). *Evolution by gene duplication.* Springer-Verlag.

Okamoto, M. (1957). Asynaptic effect of chromosome V. *Wheat Inf Serv* 5.

Orr, H.A. (1990). " Why polyploidy is rarer in animals than in plants" revisited. *The American Naturalist* 136**,** 759-770.

Ott, J., Wang, J., and Leal, S.M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics* 16**,** 275-284.

Otto, S.P. (2007). The evolutionary consequences of polyploidy. *Cell* 131**,** 452-462.

Otto, S.P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annual review of genetics* 34**,** 401-437.

Pandey, M.K., Agarwal, G., Kale, S.M., Clevenger, J., Nayak, S.N., Sriswathi, M., et al. (2017). Development and Evaluation of a High Density Genotyping 'Axiom_Arachis' Array with 58 K SNPs for Accelerating Genetics and Breeding in Groundnut. *Scientific Reports* 7:40577.

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytologist* 186**,** 5-17.

Peace, C., Bassil, N., Main, D., Ficklin, S., Rosyara, U.R., Stegmeir, T., et al. (2012). Development and evaluation of a genome-wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. *PLoS One* 7**,** e48305.

Peace, C.P. (2017). DNA-informed breeding of rosaceous crops: promises, progress and prospects. *Horticulture research* 4**:**17006.

Pécrix, Y., Rallo, G., Folzer, H., Cigna, M., Gudin, S., and Le Bris, M. (2011). Polyploidization mechanisms: temperature environment can induce diploid gamete formation in *Rosa* sp. *Journal of experimental botany* 62**,** 3587-3597.

Pham, G.M., Newton, L., Wiegert-Rininger, K., Vaillancourt, B., Douches, D.S., and Buell, C.R. (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *The Plant Journal* 92, 624-637.

Piepho, H.-P., and Koch, G. (2000). Codominant Analysis of Banding Data From a Dominant Marker System by Normal Mixtures. *Genetics* 155**,** 1459-1468.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., and R Core Team (2017). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131. *https://CRAN.R-project.org/package=nlme*.

Plummer, M. (Year). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling", in: *Proceedings of the 3rd international workshop on distributed statistical computing*: Vienna, Austria), 125.

Poland, J.A., Brown, P.J., Sorrells, M.E., and Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS one* 7**,** e32253.

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Genetic Reviews* 6**:**847.

Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475**,** 189-195.

Preedy, K.F., and Hackett, C.A. (2016). A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theoretical and Applied Genetics* 129**,** 2117-2132.

Preuss, D., Rhee, S.Y., and Davis, R.W. (1994). Tetrad analysis possible in *Arabidopsis* with mutation of the *QUARTET* (*QRT*) genes. *Science* 264**,** 1458-1459.

Ptacek, M.B., Gerhardt, H.C., and Sage, R.D. (1994). Speciation by polyploidy in treefrogs: multiple origins of the tetraploid, *Hyla versicolor*. *Evolution* 48**,** 898-908.

Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453**,** 1064-1071.

Qu, L., and Hancock, J. (2001). Detecting and mapping repulsion-phase linkage in polyploids with polysomic inheritance. *Theoretical and Applied Genetics* 103**,** 136-143.

Quarin, C.L., Espinoza, F., Martinez, E.J., Pessino, S.C., and Bovo, O.A. (2001). A rise of ploidy level induces the expression of apomixis in *Paspalum notatum*. *Sexual Plant Reproduction* 13**,** 243-249.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R version 3.3.2.

Rajapakse, S., Byrne, D., Zhang, L., Anderson, N., Arumuganathan, K., and Ballard, R. (2001). Two genetic linkage maps of tetraploid roses. *Theoretical and Applied Genetics* 103**,** 575-583.

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

Ramsey, J., and Schemske, D.W. (2002). Neopolyploidy in flowering plants. *Annual review of ecology and systematics***,** 589-639.

Rasmussen, S.W., and Holm, P.B. (1979). Chromosome pairing in autotetraploid *Bombyx* females. *Carlsberg Research Communications* 44**,** 101-125.

Reddy, U.K., Nimmakayala, P., Abburi, V.L., Reddy, C., Saminathan, T., Percy, R.G., et al. (2017). Genome-wide divergence, haplotype distribution and population demographic histories for *Gossypium hirsutum* and *Gossypium barbadense* as revealed by genome-anchored SNPs. *Scientific Reports* 7:41285.

Renne, P.R., Deino, A.L., Hilgen, F.J., Kuiper, K.F., Mark, D.F., Mitchell, W.S., et al. (2013). Time scales of critical events around the Cretaceous-Paleogene boundary. *Science* 339**,** 684-687.

Renny-Byfield, S., and Wendel, J.F. (2014). Doubling down on genomes: polyploidy and crop plants. *American journal of botany* 101**,** 1711-1725.

Reyes-Valdes, M.H., and Williams, C.G. (2005). An entropy-based measure of founder informativeness. *Genetics Research* 85**,** 81-88.

Ridout, M., Tong, S., Vowden, C., and Tobutt, K. (1998). Three-point linkage analysis in crosses of allogamous plant species. *Genetics Research* 72**,** 111-121.

Ridout, M.S., Xu, X.-M., and Tobutt, K.R. (2005). Single-locus gametophytic incompatibility in autotetraploids. *Journal of heredity* 96**,** 430-433.

Riley, R., and Chapman, V. (1958). Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature* 182**,** 713-715.

Ripol, M., Churchill, G., Da Silva, J., and Sorrells, M. (1999). Statistical aspects of genetic mapping in autopolyploids. *Gene* 235**,** 31-41.

Robins, J.G., Hansen, J.L., Viands, D.R., and Brummer, E.C. (2008). Genetic mapping of persistence in tetraploid alfalfa. *Crop Science* 48**,** 1780-1786.

Roman, H., Rapicault, M., Miclot, A., Larenaudie, M., Kawamura, K., Thouroude, T., et al. (2015). Genetic analysis of the flowering date and number of petals in rose. *Tree genetics & genomes* 11**:**85.

Ronfort, J. (1999). The mutation load under tetrasomic inheritance and its consequences for the evolution of the selfing rate in autotetraploid species. *Genetics Research* 74**,** 31-42.

Rosyara, U.R., De Jong, W.S., Douches, D.S., and Endelman, J.B. (2016). Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome* 9.

Rothfels, C.J., Pryer, K.M., and Li, F.W. (2017). Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist* 213**,** 413-429.

Roux, C., and Pannell, J.R. (2015). Inferring the mode of origin of polyploid species from next-generation sequence data. *Molecular Ecology* 24**,** 1047-1059.

Roxas, N.J., Tashiro, Y., Miyazaki, S., Isshiki, S., and Takeshita, A. (1995). Meiosis and pollen fertility in *Higo chrysanthemum* (*Dendranthema* × *grandiflorum* (Ramat.) Kitam.). *Journal of the Japanese Society for Horticultural Science* 64**,** 161-168.

Salman-Minkov, A., Sabath, N., and Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nature plants* 2**:**16115.

Santos, J., Alfaro, D., Sanchez-Moran, E., Armstrong, S., Franklin, F., and Jones, G. (2003). Partial diploidization of meiosis in autotetraploid *Arabidopsis thaliana*. *Genetics* 165**,** 1533-1540.

Sattler, M.C., Carvalho, C.R., and Clarindo, W.R. (2016). The polyploidy and its key role in plant breeding. *Planta* 243**,** 281-296.

Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8**:**552.

Schie, S., Chaudhary, R., and Debener, T. (2014). Analysis of a Complex Polyploid Plant Genome using Molecular Markers: Strong Evidence for Segmental Allooctoploidy in Garden Dahlias. *The Plant Genome* 7.

Schinkel, C.C.F., Kirchheimer, B., Dullinger, S., Geelen, D., De Storme, N., and Hörandl, E. (2017). Pathways to polyploidy: indications of a female triploid bridge in the alpine species *Ranunculus kuepferi* (*Ranunculaceae*). *Plant Systematics and Evolution* 303**,** 1093-1108.

Schmitz Carley, C.A., Coombs, J.J., Douches, D.S., Bethke, P.C., Palta, J.P., Novy, R.G., et al. (2017). Automated tetraploid genotype calling by hierarchical clustering. *Theoretical and Applied Genetics***,** 130, 717-726 .

Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current opinion in plant biology* 15**,** 147-153.

Schultz, L., Cogan, N.O.I., Mclean, K., Dale, M.F.B., Bryan, G.J., Forster, J.W., et al. (2012). Evaluation and implementation of a potential diagnostic molecular marker for *H1*-conferred potato cyst nematode resistance in potato (*Solanum tuberosum* L.). *Plant Breeding* 131**,** 315-321.

Schulz, D.F., Schott, R.T., Voorrips, R.E., Smulders, M.J.M., Linde, M., and Debener, T. (2016). Genome-wide association analysis of the anthocyanin and carotenoid contents of rose petals. *Frontiers in Plant Science* 7:1798.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics* 6**,** 461-464.

Seeb, J., Carvalho, G., Hauser, L., Naish, K., Roberts, S., and Seeb, L. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources* 11**,** 1-8.

Serang, O., Mollinari, M., and Garcia, A.a.F. (2012). Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PLoS One* 7**,** e30906.

Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., et al. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* 538**:**336.

Sharma, S.K., Bolser, D., De Boer, J., Sønderkær, M., Amoros, W., Carboni, M.F., et al. (2013). Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3: Genes| Genomes| Genetics* 3**,** 2031-2047.

Shen, J., Li, Z., Chen, J., Song, Z., Zhou, Z., and Shi, Y. (2016). SHEsisPlus, a toolset for genetic studies on polyploid species. *Scientific Reports* 6:24095.

Shirasawa, K., Tanaka, M., Takahata, Y., Ma, D., Cao, Q., Liu, Q., et al. (2017). A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (*Ipomoea batatas*). *Scientific Reports* 7:44207.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nature genetics* 43**,** 109-116.

Shupert, D.A., Byrne, D.H., and Brent Pemberton, H. (2005). Inheritance of flower traits, leaflet number and prickles in roses. *Acta Hort* 751**,** 331-335.

Slater, A.T., Cogan, N.O., and Forster, J.W. (2013). Cost analysis of the application of marker-assisted selection in potato breeding. *Molecular breeding* 32**,** 299-310.

Slater, A.T., Cogan, N.O., Forster, J.W., Hayes, B.J., and Daetwyler, H.D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *The Plant Genome* 9.

Slater, A.T., Wilson, G.M., Cogan, N.O., Forster, J.W., and Hayes, B.J. (2014). Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. *Theoretical and applied genetics* 127**,** 809-820.

Smulders, M.J.M., Arens, P., Koning-Boucoiran, C.F.S., Gitonga, V.W., Krens, F.A., Atanassov, A., et al. (2011). "*Rosa*," in *Wild Crop Relatives: Genomic and Breeding Resources Plantation and Ornamental Crops*. (Springer-Verlag Berlin Heidelberg), 243-274.

Soltis, D.E., Visger, C.J., Marchant, D.B., and Soltis, P.S. (2016). Polyploidy: Pitfalls and paths to a paradigm. *American Journal of Botany* 103**,** 1146-1166.

Soltis, P.S., Marchant, D.B., Van De Peer, Y., and Soltis, D.E. (2015). Polyploidy and genome evolution in plants. *Current opinion in genetics & development* 35**,** 119-125.

Soltis, P.S., and Soltis, D.E. (2000). The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences* 97**,** 7051-7057.

Soltis, P.S., and Soltis, D.E. (2009). The role of hybridization in plant speciation. *Annual review of plant biology* 60**,** 561-588.

Spillane, C., Curtis, M.D., and Grossniklaus, U. (2004). Apomixis technology development—virgin births in farmers' fields? *Nature biotechnology* 22**,** 687-691.

Spiller, M., Linde, M., Hibrand-Saint Oyant, L., Tsai, C.-J., Byrne, D.H., Smulders, M.J.M., et al. (2011). Towards a unified genetic map for diploid roses. *Theoretical and Applied Genetics* 122**,** 489-500.

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., et al. (2013). Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theoretical and Applied Genetics* 126**,** 2699-2716.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *The Plant Journal* 3**,** 739-744.

Stebbins, G.L. (1947). Types of polyploids: their classification and significance. *Advances in Genetics* 1**,** 403-429.

Stickland, R. (1972). Changes in anthocyanin, carotenoid, chlorophyll, and protein in developing florets of the chrysanthemum. *Annals of Botany* 36**,** 459-469.

Stift, M., Berenos, C., Kuperus, P., and Van Tienderen, P.H. (2008). Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to Rorippa (yellow cress) microsatellite data. *Genetics* 179**,** 2113-2123.

Stift, M., Reeve, R., and Van Tienderen, P. (2010). Inheritance in tetraploid yeast revisited: segregation patterns and statistical power under different inheritance models. *Journal of evolutionary biology* 23**,** 1570-1578.

Strasburger, E. (1910). Sexuelle und apogame Fortpflanzung bei *Urticaceen*. *Jahrb Wiss Bot* 47**,** 491-502.

Stringham, H.M., and Boehnke, M. (1996). Identifying marker typing incompatibilities in linkage analysis. *American Journal of Human Genetics* 59**,** 946-950.

Stupar, R.M., Bhaskar, P.B., Yandell, B.S., Rensink, W.A., Hart, A.L., Ouyang, S., et al. (2007). Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics* 176**,** 2055-2067.

Sturtevant, A.H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology* 14**,** 43-59.

Su, S.-Y., White, J., Balding, D.J., and Coin, L.J. (2008). Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC bioinformatics* 9**:**513.

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P., and Reynolds, M.P. (2015). Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theoretical and Applied Genetics* 128**,** 353-363.

Svartman, M., Stone, G., and Stanyon, R. (2005). Molecular cytogenetics discards polyploidy in mammals. *Genomics* 85**,** 425-430.

Sved, J. (1966). Telomere attachment of chromosomes. Some genetical and cytological consequences. *Genetics* 53**:**747.

Sved, J.A. (1964). The relationship between diploid and tetraploid recombination frequencies. *Heredity* 19**,** 585-596.

Sverrisdóttir, E., Byrne, S., Sundmark, E.H.R., Johnsen, H.Ø., Kirk, H.G., Asp, T., et al. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theoretical and Applied Genetics* 130**,** 2091-2108.

Swaminathan, M.S., and Howard, H. (1953). Cytology and genetics of the potato (*Solanum tuberosum*) and related species. *Bibliographia Genetica* 16**,** 1–192.

Sybenga, J. (1972). *General Cytogenetics*. North-Holland Publishing Company, Amsterdam.

Sybenga, J. (1975). The quantitative analysis of chromosome pairing and chiasma formation based on the relative frequencies of MI configurations. *Chromosoma* 50**,** 211-222.

Sybenga, J. (1992). "Cytogenetics in plant breeding". (Springer-Verlag).

Sybenga, J. (1994). Preferential pairing estimates from multivalent frequencies in tetraploids. *Genome* 37**,** 1045-1055.

Sybenga, J. (1996). Chromosome pairing affinity and quadrivalent formation in polyploids: do segmental allopolyploids exist? *Genome* 39**,** 1176-1184.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., et al. (2014). An improved genome release (version Mt4. 0) for the model legume *Medicago truncatula*. *BMC genomics* 15**:**312.

Taylor, N.L. (2008). A century of clover breeding developments in the United States. *Crop Science* 48**,** 1-13.

Te Beest, M., Le Roux, J.J., Richardson, D.M., Brysting, A.K., Suda, J., Kubešová, M., et al. (2011). The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany* 109**,** 19-45.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* 102**,** 13950-13955.

Tinker, N.A., Bekele, W.A., and Hattori, J. (2016). Haplotag: software for haplotype-based genotyping-by-sequencing analysis. *G3: Genes, Genomes, Genetics* 6**,** 857-863.

Tinker, N.A., Chao, S., Lazo, G.R., Oliver, R.E., Huang, Y.-F., Poland, J.A., et al. (2014). A SNP genotyping array for hexaploid oat. *The Plant Genome* 7.

Tong, C., Zhang, B., and Shi, J. (2010). A hidden Markov model approach to multilocus linkage analysis in a full-sib family. *Tree genetics & genomes* 6**,** 651-662.

Truco, M.J., Ashrafi, H., Kozik, A., Van Leeuwen, H., Bowers, J., Wo, S.R.C., et al. (2013). An ultra-high-density, transcript-based, genetic map of lettuce. *G3: Genes, Genomes, Genetics* 3**,** 617-631.

Tumino, G., Voorrips, R.E., Morcia, C., Ghizzoni, R., Germeier, C.U., Paulo, M.-J., et al. (2017). Genome-wide association analysis for lodging tolerance and plant height in a diverse European hexaploid oat collection. *Euphytica* 213**:**163.

Tumino, G., Voorrips, R.E., Rizza, F., Badeck, F.W., Morcia, C., Ghizzoni, R., et al. (2016). Population structure and genome-wide association analysis for frost tolerance in oat using continuous SNP array signal intensity ratios. *Theoretical and Applied Genetics* 129**,** 1711-1724.

Udall, J.A., and Wendel, J.F. (2006). Polyploidy and crop improvement. *Crop Science* 46**,** S-3-S-14.

Uitdewilligen, J.G., Wolters, A.-M.A., D'hoop, B.B., Borm, T.J., Visser, R.G., and Van Eck, H.J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One* 8(5)**,** e62355.

Ukoskit, K., and Thompson, P.G. (1997). Autopolyploidy versus allopolyploidy and low-density randomly amplified polymorphic DNA linkage maps of sweetpotato. *Journal of the American Society for Horticultural Science* 122**,** 822-828.

Uzzell, T.M.J. (1964). Relations of the diploid and triploid species of the *Ambystoma jeffersonianum* complex (Amphibia, Caudata). *Copeia* 2**,** 257-300.

V.S.N. International (2018). "ASreml".

Van De Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18**,** 411-424.

Van Dijk, T., Noordijk, Y., Dubos, T., Bink, M.C., Meulenbroek, B.J., Visser, R.G., et al. (2012). Microsatellite allele dose and configuration establishment (MADCE): an integrated approach for genetic studies in allopolyploids. *BMC Plant Biology* 12**:**25.

Van Eeuwijk, F.A., Bink, M.C., Chenu, K., and Chapman, S.C. (2010). Detection and use of QTL for complex traits in multiple environments. *Current opinion in plant biology* 13**,** 193-205.

Van Geest, G., Bourke, P.M., Voorrips, R.E., Marasek-Ciolakowska, A., Liao, Y., Post, A., et al. (2017a). An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. *Theoretical and Applied Genetics* 130**,** 2527-2541.

Van Geest, G., Choi, Y.H., Arens, P., Post, A., Liu, Y., and Van Meeteren, U. (2016). Genotypic differences in metabolomic changes during storage induced-degreening of chrysanthemum disk florets. *Postharvest Biology and Technology* 115**,** 48-59.

Van Geest, G., Post, A., Arens, P., Visser, R.G., and Van Meeteren, U. (2017b). Breeding for postharvest performance in chrysanthemum by selection against storage-induced degreening of disk florets. *Postharvest Biology and Technology* 124**,** 45-53.

Van Geest, G., Voorrips, R.E., Esselink, D., Post, A., Visser, R.G., and Arens, P. (2017c). Conclusive evidence for hexasomic inheritance in chrysanthemum based on analysis of a 183 k SNP array. *BMC genomics* 18**:**585.

Van Ooijen, J. (2009). MapQTL ® 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma B.V., Wageningen, Netherlands.

Van Ooijen, J.W. (1992). Accuracy of mapping quantitative trait loci in autogamous species. *Theoretical and Applied Genetics* 84**,** 803-811.

Van Ooijen, J.W. (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. *Kyazma B.V., Wageningen, The Netherlands.*

Van Ooijen, J.W., and Jansen, J. (2013). *Genetic mapping in experimental populations.* Cambridge University Press.

Van Os, H., Andrzejewski, S., Bakker, E., Barrena, I., Bryan, G.J., Caromel, B., et al. (2006). Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* 173, 1075-1087.

Van Tuyl, J.M., and Lim, K.B. (2003). Interspecific hybridisation and polyploidisation as tools in ornamental plant breeding. *Acta Horticulturae* 612**,** 13-22.

Vanneste, K., Baele, G., Maere, S., and Van De Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome research* 24**,** 1334-1347.

Vanraden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91**,** 4414-4423.

Vigna, B.B., Santos, J.C., Jungmann, L., Do Valle, C.B., Mollinari, M., Pastina, M.M., et al. (2016). Evidence of allopolyploidy in *Urochloa humidicola* based on cytological analysis and genetic linkage mapping. *PloS one* 11**,** e0153764.

Vleugels, T., Ceuppens, B., Cnops, G., Lootens, P., Van Parijs, F.R., Smagghe, G., et al. (2016). Models with only two predictor variables can accurately predict seed yield in diploid and tetraploid red clover. *Euphytica* 209**,** 507-523.

Voorrips, R. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *Journal of Heredity* 93**,** 77-78.

Voorrips, R.E., Gort, G., and Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12:172.

Voorrips, R.E., and Maliepaard, C.A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics* 13**:**248.

Vos, P.G., Paulo, M.J., Voorrips, R.E., Visser, R.G., Van Eck, H.J., and Van Eeuwijk, F.A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics* 130**,** 123-135.

Vos, P.G., Uitdewilligen, J.G.a.M.L., Voorrips, R.E., Visser, R.G.F., and Van Eck, H.J. (2015). Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *Theoretical and Applied Genetics* 128**,** 2387-2401.

Vukosavljev, M., Arens, P., Voorrips, R., Van 'T Westende, W., Esselink, G.D., Bourke, P.M., et al. (2016). High-density SNP-based genetic maps for the parents of an outcrossed and a selfed tetraploid garden rose cross, inferred from admixed progeny using the 68k rose SNP array. *Horticulture Research* 3:16052.

Wang, H., Zhang, Z., Rose, S., and Van Der Laan, M. (2014a). A Novel Targeted Learning Method for Quantitative Trait Loci Mapping. *Genetics* 198**,** 1369-1376.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., et al. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. *Nature genetics* 44**,** 1098-1103.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E., et al. (2014b). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant biotechnology journal* 12**,** 787-796.

Westergaard, M. (1940). Studies on cytology and sex determination in polyploid forms of *Melandrium album*. *Dansk Botanisk Arkiv* 10**,** 1-131.

Wet, J.D., and Harlan, J. (1970). Apomixis, polyploidy, and speciation in Dichanthium. *Evolution* 24**,** 270-277.

Whitton, J., Sears, C.J., Baack, E.J., and Otto, S.P. (2008). The dynamic nature of apomixis in the angiosperms. *International Journal of Plant Sciences* 169**,** 169-182.

Wijnker, E., James, G.V., Ding, J., Becker, F., Klasen, J.R., Rawat, V., et al. (2013). The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife* 2**,** e01426.

Winfield, M.O., Allen, A.M., Burridge, A.J., Barker, G.L., Benbow, H.R., Wilkinson, P.A., et al. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant biotechnology journal* 14**,** 1195-1206.

Wissemann, V., and Ritz, C.M. (2005). The genus *Rosa* (Rosoideae, Rosaceae) revisited: molecular analysis of nrITS-1 and atpB-rbcL intergenic spacer (IGS) versus conventional taxonomy. *Botanical Journal of the Linnean Society* 147**,** 275-290.

Wolfram Research Inc. (2014). "Mathematica Version 10.0". (Champaign, Illinois).

Wolters, A.-M.A., Uitdewilligen, J.G., Kloosterman, B.A., Hutten, R.C., Visser, R.G., and Van Eck, H.J. (2010). Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. *Plant Molecular Biology* 73**,** 659-671.

Worthington, M., Heffelfinger, C., Bernal, D., Quintero, C., Zapata, Y.P., Perez, J.G., et al. (2016). A parthenogenesis gene candidate and evidence for segmental allopolyploidy in apomictic *Brachiaria decumbens*. *Genetics* 203**,** 1117-1132.

Wu, K., Burnquist, W., Sorrells, M., Tew, T., Moore, P., and Tanksley, S. (1992). The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theoretical and Applied Genetics* 83**,** 294-300.

Wu, R., and Ma, C.-X. (2005). A general framework for statistical linkage analysis in multivalent tetraploids. *Genetics* 170**,** 899-907.

Wu, R., Ma, C.-X., and Casella, G. (2002). A bivalent polyploid model for linkage analysis in outcrossing tetraploids. *Theoretical Population Biology* 62**,** 129-151.

Wu, R., Ma, C.-X., and Casella, G. (2004). A mixed polyploid model for linkage analysis in outcrossing tetraploids using a pseudo-test backcross design. *Journal of Computational Biology* 11**,** 562-580.

Xavier, A., Muir, W.M., Craig, B., and Rainey, K.M. (2016). Walking through the statistical black boxes of plant breeding. *Theoretical and applied genetics* 129**,** 1933-1949.

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., et al. (2016). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution* 34, 262-281.

Xie, C., and Xu, S. (2000). Mapping quantitative trait loci in tetraploid populations. *Genetical Research* 76**,** 105-115.

Xu, F., Lyu, Y., Tong, C., Wu, W., Zhu, X., Yin, D., et al. (2013). A statistical model for QTL mapping in polysomic autotetraploids underlying double reduction. *Briefings in Bioinformatics* 15**,** 1044-1056.

Yan, Z., Dolstra, O., Hendriks, T., Prins, T., Stam, P., and Visser, P. (2005). Vigour evaluation for genetics and breeding in rose. *Euphytica* 145**,** 339-347.

Yan, Z., Dolstra, O., Prins, T.W., Stam, P., and Visser, P.B. (2006). Assessment of partial resistance to powdery mildew (*Podosphaera pannosa*) in a tetraploid rose population using a spore-suspension inoculation method. *European journal of plant pathology* 114**,** 301-308.

Yan, Z., Visser, P., Hendriks, T., Prins, T., Stam, P., and Dolstra, O. (2007). QTL analysis of variation for vigour in rose. *Euphytica* 154**:**53.

Yang, C., Zhao, L., Zhang, H., Yang, Z., Wang, H., Wen, S., et al. (2014). Evolution of physiological responses to salt stress in hexaploid wheat. *Proceedings of the National Academy of Sciences* 111**,** 11882-11887.

Yang, J., Moeinzadeh, M.-H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., et al. (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature Plants* 3**:**696.

Yang, X., Sood, S., Glynn, N., Islam, M.S., Comstock, J., and Wang, J. (2017). Constructing high-density genetic maps for polyploid sugarcane (*Saccharum* spp.) and identifying quantitative trait loci controlling brown rust resistance. *Molecular Breeding* 37**:**116.

Yant, L., Hollister, J.D., Wright, K.M., Arnold, B.J., Higgins, J.D., Franklin, F.C.H., et al. (2013). Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Current Biology* 23**,** 2151-2156.

Young, N.D., Debellé, F., Oldroyd, G.E., Geurts, R., Cannon, S.B., Udvardi, M.K., et al. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* 480**,** 520-524.

Yu, C., Luo, L., Pan, H., Guo, X., Wan, H., and Zhang, Q. (2015). Filling gaps with construction of a genetic linkage map in tetraploid roses. *Frontiers in Plant Science* 5:796.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. 38**,** 203-208.

Yu, L.X., Zheng, P., Zhang, T., Rodringuez, J., and Main, D. (2017). Genotyping-by-sequencing-based genome-wide association studies on Verticillium wilt resistance in autotetraploid alfalfa (*Medicago sativa* L.). *Molecular plant pathology* 18**,** 187-194.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences* 90**,** 10972-10976.

Zhang, F., Chen, S., Chen, F., Fang, W., Chen, Y., and Li, F. (2011a). SRAP-based mapping and QTL detection for inflorescence-related traits in chrysanthemum (*Dendranthema morifolium*). *Molecular breeding* 27**,** 11-23.

Zhang, F., Chen, S., Chen, F., Fang, W., Deng, Y., Chang, Q., et al. (2011b). Genetic analysis and associated SRAP markers for flowering traits of chrysanthemum (*Chrysanthemum morifolium*). *Euphytica* 177**,** 15-24.

Zhang, F., Chen, S., Chen, F., Fang, W., and Li, F. (2010a). A preliminary genetic linkage map of chrysanthemum (*Chrysanthemum morifolium*) cultivars using RAPD, ISSR and AFLP markers. *Scientia Horticulturae* 125**,** 422-428.

Zhang, F., Chen, S., Jiang, J., Guan, Z., Fang, W., and Chen, F. (2013a). Genetic mapping of quantitative trait loci underlying flowering time in chrysanthemum (*Chrysanthemum morifolium*). *PLoS one* 8**,** e83023.

Zhang, F., Jiang, J., Chen, S., Chen, F., and Fang, W. (2012a). Detection of quantitative trait loci for leaf traits in chrysanthemum. *The Journal of Horticultural Science and Biotechnology* 87**,** 613-618.

Zhang, F., Jiang, J., Chen, S., Chen, F., and Fang, W. (2012b). Mapping single-locus and epistatic quantitative trait loci for plant architectural traits in chrysanthemum. *Molecular breeding* 30**,** 1027-1036.

Zhang, J., Esselink, G., Che, D., Fougère-Danezan, M., Arens, P., and Smulders, M. (2013b). The diploid origins of allopolyploid rose species studied using single nucleotide polymorphism haplotypes flanking a microsatellite repeat. *The Journal of Horticultural Science and Biotechnology* 88**,** 85-92.

Zhang, K., Calabrese, P., Nordborg, M., and Sun, F. (2002). Haplotype block structure and its applications to association studies: power and study designs. *The American Journal of Human Genetics* 71**,** 1386-1394.

Zhang, L., Byrne, D., Ballard, R., and Rajapakse, S. (2006). Microsatellite marker development in rose and its application in tetraploid mapping. *Journal of the American Society for Horticultural Science* 131**,** 380-387.

Zhang, T., Yu, L.-X., Zheng, P., Li, Y., Rivera, M., Main, D., et al. (2015). Identification of loci associated with drought resistance traits in heterozygous autotetraploid alfalfa (*Medicago sativa* L.) using genome-wide association studies with genotyping by sequencing. *PloS one* 10**,** e0138931.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., et al. (2010b). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42**,** 355-360.

Zheng, C., Voorrips, R.E., Jansen, J., Hackett, C.A., Ho, J., and Bink, M.C. (2016). Probabilistic Multilocus Haplotype Reconstruction in Outcrossing Tetraploids. *Genetics* 203**,** 119-131.

Zhou, C., Olukolu, B., Gemenet, D., Wu, S., Gruneberg, W., Cao, M.D., et al. (2017). Assembly Of Whole-Chromosome Pseudomolecules For Polyploid Plant Genomes Using Outcrossed Mapping Populations. *bioRxiv, doi 10.1101/119271*.

Zickler, D., and Kleckner, N. (2015). Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harbor perspectives in biology* 7**,** a016626.

Zielinski, M.-L., and Scheid, O.M. (2012). "Meiosis in polyploid plants," in *Polyploidy and genome evolution*. Springer, 33-55.

# Summary

Almost all higher organisms are diploid *i.e.* they carry two copies of each chromosome, one inherited from each parent. Although diploidy appears to be the norm in higher organisms, it is certainly not the rule. A small but significant proportion of higher organisms possess more than two copies of each chromosome. These are collectively termed "polyploids" and make up the central subject matter for this thesis.

We are now living in a time of unprecedented development in scientific tools, particularly for the life sciences. For many years diploid species have been the subject of genetic investigations, often with the aim of understanding the inheritance of specific traits. For polyploid species, many of which are important agricultural crops, this has often remained infeasible due to the increased complexity of polyploid genomics. This thesis represents part of the ongoing development of methods and software tools required to make polyploid genomics a reality, enabling genomics-assisted breeding decisions to be made for polyploid crops.

We begin in **Chapter 2** with a review of the tools currently available to analyse polyploid populations. These tools encompass activities such as genotype assignment and marker dosage calling, haplotype assembly, linkage analysis and genetic map construction, genome-wide association analysis, genomic selection, the creation of reference sequence assemblies and the simulation of polyploid populations. Although introductory to the thesis, this review represents the current state of the art in polyploid software tools (and was compiled *after* the tools described in this thesis were developed).

Linkage maps represent a molecular karyotype of a species and its recombination landscape and have numerous applications such as in quantitative trait loci (QTL) analysis, synteny analysis, reference sequence assembly, fine mapping studies *etc.* They can also provide insights into meiosis and for polyploids, can be used to uncover some of the more curious features of polyploid meiosis. In **Chapter 3** we generated a high-density linkage map from single nucleotide polymorphism (SNP) genotype data of a tetraploid potato population. This map allowed us to study the phenomenon of double reduction, where both copies of (part of) a chromatid end up being transmitted to an offspring. This was achieved using simplex x nulliplex markers only, which give unambiguous information about double reduction. We broadened our mapping approach to include all other SNP marker types in **Chapter 4**, while also investigating the effects of double reduction and non-random chromosomal pairing on the estimation of recombination frequency in autotetraploids. The topic of non-random, or preferential, chromosome pairing returns in **Chapter 5**, where we again used high-density linkage mapping to understand the patterns of meiosis in a tetraploid rose population. We

discovered that pairing affinities varied between parents, chromosomes and even chromosomal position in our population, with evidence for strongly preferential pairing localised to certain parts of a number of parental chromosomes. As well as diagnosing this behaviour, we corrected for it in our linkage analysis, leading to the creation of a tailored high-density genetic map for this important ornamental species. This map allowed us to uncover interesting genomic rearrangements between rose and its closest sequenced relative, the woodland strawberry (*Fragaria vesca* L.), and has since been used in efforts to create a reference sequence assembly in rose. In **Chapter 6** we describe a new software tool *polymapR*, which is freely available as an R package distributed through CRAN. *polymapR* is currently able to generate integrated and phased linkage maps using dosage-scored marker data in triploid, tetraploid and hexaploid populations that exhibit random chromosomal pairing (polysomic inheritance), as well as segmental allotetraploids showing preferential chromosome pairing, and represents a timely addition to the suite of tools available to polyploid geneticists.

In **Chapter 7**, we used *polymapR* to analyse a SNP marker dataset from a large population of hexaploid *Chrysanthemum* × *morifolium*, generating the first high-density integrated linkage map in this important ornamental species. We used the phased map and marker dosage scores to generate identity-by-descent (IBD) probabilities which formed the basis of a QTL analysis for a number of important ornamental traits. We return to the topic of double reduction in **Chapter 8**, quantifying the effect of double reduction on the power and precision of QTL studies in autotetraploids. As well as gaining insight into the mechanics of IBD-based QTL analysis at the polyploid level, we explored the effect of variable marker densities on genetic mapping using the genotypic information coefficient (GIC) and found that this measure has important implications for the results of polyploid QTL studies. We applied these methods in **Chapter 9** to perform a multi-environment analysis for some important morphological traits in rose. This allowed us not only to describe the genetic architecture of these traits, but to test whether the environment has a strong influence on the phenotypic expression, of great relevance to breeders and researchers performing QTL studies away from the target environment. In **Chapter 10** these developments form the basis of *polyqtlR*, an R package to perform QTL interval mapping in autopolyploid populations. Finally, in **Chapter 11** I take a step back from the technical innovations presented in this thesis, considering the impact that these tools can have on our understanding of, and our ability to breed, polyploid crops.

It is hoped that the methods and tools described in this thesis will simplify genetic mapping in polyploids and facilitate genomics-assisted breeding decisions in the future for these fascinating species.

# Acknowledgements

Being one of the most-read sections of a thesis, I am aware of the responsibility to properly thank all the people who in some way or another contributed to the work described within these covers. Given that the list is long and space is short, I will do my best, but please accept my apologies if you came here looking for your name and failed to find it! I had planned a special dedication to you further down in white ink, but the printers were unable to render it legibly...

But back to the serious topic of thanking people, first of all to my daily supervisors **Chris Maliepaard** and **Roeland Voorrips**, I wish to thank you both for your time, enthusiasm and expertise, and for inviting me to do this PhD in the first place. Our fortnightly supervision meetings were both friendly and focused, and we always seemed to end up laughing at some stage. Both of you contributed much to the content of the thesis, but also to the quality of the written output. Roeland, I hadn't fully appreciated what the term "scrupulous" meant until I received feedback from you. Your attention to detail and command of the English language are impressive, indeed I am often left with the feeling that your English is more English than mine. Thank you both for enduring the editorial demands of this somewhat frenetic writer. Chris, thanks also for giving me the room to follow my own research goals during this work. Together with Roeland you helped define the work to be done, but you allowed me the space to develop many of the details. I think this is one of the strengths of a good supervisor and I'm very aware of how lucky I've been with the supervision team I've had here.

To my promoter **Richard Visser**, you provided not only the opportunity to do the PhD within the department of plant breeding, but also very valuable feedback on my project, particularly the written output. After 2.5 years you told me I could finish up whenever I liked (you were probably getting worried that you would have no free weekends left if I kept writing) – this really spurred me on to "opschiet een beetje" and encouraged me to finish writing the thesis in 3.5 years instead of the usual 4 (or 5!). You have always been approachable and friendly, and your dedication to the department is second to none. Thanks to my external supervisor **Duur Aanen** for making time for our annual meetings which I enjoyed. Thankfully you weren't required to mediate at any stage! Within the department there are also a large number of individuals that have helped in many ways. I would like to thank **René Smulders** for the many interesting conversations and collaborations we've had, particularly on rose genetics but also more generally on quantitative genetics in polyploids. To **Herman van Eck**, like René you are also one of the people that helped me think more deeply and consider alternative ways to look at things. You have also given me a fuller appreciation of the power of a good question,

and I must admit I am slightly relieved that you are not eligible to sit on the opponents' bench for my defence! Thanks to **Paul Arens** for the help and feedback during my project, as well as being such a friendly colleague with whom I could have a chat about "gewoon dinges" as well. You have helped me feel more at home in the Netherlands. Thanks also to **Eric van de Weg**, you possess both a deep knowledge on genetic mapping as well as a meticulousness that has helped separate the signal from the noise in a way no computer could. Your ideas and questions helped broaden the boundaries of the work in this thesis. I also appreciated your enthusiasm for my presenting style, particularly when it led to my presenting Chapter 3 of this thesis to the polyploids session of the Plant and Animal Genome conference in 2016. Thanks to **Gerard van der Linden** and **Christian Bachem**, my MSc thesis supervisors. Christian, you managed to convince me to stay and do a PhD in plant breeding despite my advancing years. I probably wouldn't still be here otherwise! There are also many other people that deserve a mention who have been involved in my project (be it inspirational or otherwise), such as **Arwa Shahin**, **Carole Koning-Boucoiran**, **Chaozhi Zheng**, **Danny Esselink**, **Ehsan Motazedi**, **Frans Krens**, **Giorgio Tumino**, **Herma Koehorst-van Putten** (who, along with **Michela Appiano** helped fly the PBR flag at my wedding!), **Johan Willemsen**, **Katherine Preedy**, **Konrad Zych**, **Linda Kodde, Marco Bink**, **Peter Vos**, **Richard Finkers**, **Robert van Loo**, **Ronald Hutten** and **Thijs van Dijk.** Thanks also to everyone who kept the journal club going over the years, hopefully it will continue into the future!

A special thanks to **Christine Hackett** for accepting my request to join BioSS for a month in 2016. Chris, you have been one of the leading lights in polyploid genetic mapping for many years, and it was a great experience to work together with you. Our collaboration led directly to Chapter 8 of this thesis, but also helped greatly in developing Chapters 9 and 10. I would like to thank **Hans Jansen** for helping to define the goals in polyploid linkage mapping early on in the project, and for your infectious enthusiasm for mapping and genetics. I also want to thank **Johan van Ooijen** for being so generous with your time and advice in helping us understand some of the mapping steps in JoinMap and trying to translate these ideas into *polymapR*. Thank-you to **Geert van Geest** for all the work in improving *polymapR*, you took a set of functions and helped turn them into a great package. I learned a lot about programming from you, and our collaboration really helped push forward the progress in both of our projects.

I also want to acknowledge the contribution of the MSc students I co-supervised **- Twan Kranenburg**, **Patrick Wissink**, **Yanlin Liao**, and **Jitpanu Yamjabok**. All of you helped my project by exploring new ideas and I learnt a lot from the experience of supervision. Thanks to my paranymphs **Jordi Petit-Pedro** and **Michiel Klaassen** for your meditative pose and perfect enunciation of any and all propositions during my

defence, as well as being all-round great guys. Thanks to my present and former office-mates **Anne Giesbers**, **Charlie Chen**, **Deniz Gol**, **Ehsan Motazedi**, **Mas Muniroh**, **Michela Appiano**, **Myluska Caro-Rios**, **Narges Sedaghat Telgerd**, **Yi Wu** and **Ying Liu** (and everyone from my short stay in E2.163) for putting up with a colleague that never goes away to the lab or the greenhouse – the office was my greenhouse, the computer was my lab! To the many other friends and acquaintances I've made along the way, thanks for making my time at plant breeding so enjoyable. I won't list you all by name as that would involve literally listing everyone at PBR and quite a few people at the other side of Radix as well, so you know who you are! I should like to particularly mention **Anne Giesbers**, **Jeroen Berg** and **Tim van der Weijde** with whom I travelled to Nanjing, China for the international graduate conference in 2016. Apart from being a great bunch of adventurous spirits, you also put up with my "wat vervelend Nederlands" and I think the week and a half I spent in China really paid dividends for my ability to communicate using your "ingewikkeld taal". To the secretaries (past and present) in the department of plant breeding: **Nicole Trefflich**, **Letty Dijker**, **Daniëlle van der Wee** and **Janneke van Deursen**, thank you all for being such a friendly and efficient team, keeping the wheels in motion within the department. You have always been very helpful with any problems or questions that I had, and always did it with a smile. A special thanks to Nicole for booking my thesis defence first thing on a Friday morning on your day off! I also want to thank all the polyploid project partners who were there throughout my project during presentations and workshops, as well as the occasional off-site visit which I really enjoyed. Your feedback was essential and I think the involvement of so many partners with so many diverse polyploid crops has really helped broaden the horizons of polyploid genetics here in Wageningen. To my fellow members of the EPS PhD council – **Amalia**, **Francesca**, **Jarst**, **Jesse**, **Jordi**, **Kiki**, **Kim**, **Malaika**, **Manos**, **Martina**, **Sandeep**, **Sara**, **Shanice**, **Tieme** and **Tom**, I think we achieved a lot together on the council and had some fun along the way! Thanks also to **Douwe Zuidema**, **Ingrid Vleghels** and **Ria Fonteyn** from EPS for all the help. I also want to thank the reading committee (my "opponents") for agreeing to read this thesis – without foreknowledge about its length. They say brevity is the soul of wit – apologies for not writing a wittier thesis then!

Finally, I would like to acknowledge my family and in particular my parents who have always encouraged and supported me despite the many twists and turns I have taken along the way. Thank you for trusting that it will all work out one day, as indeed I hope someday it shall. A final word of thanks to my lovely wife Alena who has been a constant source of support and encouragement throughout. By the time these words are in print we will already be knee-deep in nappies, and I look forward to continue sharing all the happiness in the world together with you.

# About the author

Peter Michael Bourke was born on 28th July 1982 in Cork, Ireland. In 2001 he represented his country twice – at the World Schools' Debating Championship in Johannesburg, South Africa and the International Mathematical Olympiad in Washington DC, USA. He continued speaking and doing maths during his bachelors studies, graduating with joint first class honours in mathematics and physics from University College Cork. Having completed his studies, he bought a bicycle and headed overland and sea for China, with an English-language teaching qualification in his bag. Unfortunately, Russian immigration regulations and its winter intervened, forcing him to about-turn and return to western Europe, eventually pedalling 2300m above sea level to look for work in a ski resort in the French Alps. Having survived the winter (and learnt how to ski), he enrolled the following spring in a course on practical sustainability in Kinsale F.E.C., Cork. This led to a job with the Irish Seed Savers Association, an organisation dedicated to conserving open-pollinated landrace varieties of vegetables, grains and fruit trees in Ireland. After four years working as the garden co-ordinator he again took to his bicycle, visiting and working with organic breeding companies and seedbanks across Europe on a 10,000 km cycling tour. A love of cycling and a desire to learn more about plant breeding brought him to the Netherlands in 2012 where he enrolled in the MSc plant sciences program at Wageningen University (specialisation plant breeding and genetic resources). During his masters he was fortunate to do a minor thesis with Chris Maliepaard and Roeland Voorrips in the department of plant breeding, who offered him the possibility to continue this research as part of a PhD. In 2016 he was awarded an EMBO short-term fellowship to spend one month working in the group of Dr. Christine Hackett at BioSS, Dundee, Scotland. The results of his PhD project, started in September 2014, are described in this thesis.

# List of publications

**Bourke, P.M.**, van Geest, G., Voorrips, R.E., Jansen, J., Kranenburg, T., Shahin, A. *et al.* (2018). polymapR – linkage analysis and genetic map construction from $F_1$ populations of outcrossing polyploids. *Bioinformatics (in press)* doi: 10.1093/bioinformatics/bty371

Hibrand Saint-Oyant, L., Ruttink, T., ... **Bourke, P.M.** *et al.* (2018). A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nature Plants (accepted)*

**Bourke, P.M.**, Voorrips, R.E., Visser, R.G.F., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Frontiers in Plant Science* 9:513. doi: 10.3389/fpls.2018.00513

van Geest, G., **Bourke, P.M.** *et al.* (2017). An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. *Theoretical & Applied Genetics* 130:12, 2527-2541

**Bourke, P.M.**, Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F. S. *et al.* (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal* 90:2, 330 - 343

Vukosavljev, M., Arens, P., ... **Bourke, P.M.** *et al.* (2016). High-density SNP-based genetic maps for the parents of an outcrossed and a selfed tetraploid garden rose cross, inferred from admixed progeny using the 68k rose SNP array. *Horticulture Research* 3:16052

**Bourke, P.M.**, Voorrips, R.E., Kranenburg, T., Jansen, J., Visser, R.G.F., Maliepaard, C. (2016). Integrating haplotype-specific linkage maps in a tetraploid species using SNP markers. *Theoretical & Applied Genetics* 129:11, 2211-2226

**Bourke, P.M.**, Voorrips, R.E., Visser, R.G.F., and Maliepaard, C. (2015). The double reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics* 201, 853-863

# Education Statement of the Graduate School Experimental Plant Sciences

**Issued to:** **Peter M. Bourke**
**Date:** **15 June 2018**
**Group:** **Laboratory of Plant Breeding**
**University:** **Wageningen University & Research, Wageningen, The Netherlands**

| **1) Start-up phase** | *date* |
|---|---|
| ▶ **First presentation of your project** | |
| *Title:* QTL analysis in polyploids - models and methods | 23 Jan 2015 |
| ▶ **Writing or rewriting a project proposal** | |
| *Title:* Models and Methods for QTL analysis in polyploid crops | 09 Dec 2014 |
| ▶ **Writing a review or book chapter** | |
| "Tools for genetic studies in experimental populations of polyploids", Front. Plant Sci (2018), 9:513. doi: 10.3389/fpls.2018.00513 | 28 Dec 2017 |
| ▶ **MSc courses** | |
| Programming in Python (INF-22306) | Oct-Dec 2014 |
| ▶ **Laboratory use of isotopes** | |
| *Subtotal Start-up Phase* | *13.5 credits** |

| **2) Scientific Exposure** | *date* |
|---|---|
| ▶ **EPS PhD student days** | |
| EPS PhD student day 'Get2Gether', Soest, NL | 29-30 Jan 2015 |
| EPS PhD student day 'Get2Gether', Soest, NL | 28-29 Jan 2016 |
| EPS PhD student day 'Get2Gether', Soest, NL | 09-10 Feb 2017 |
| EPS PhD student day 'Get2Gether', Soest, NL | 15-16 Feb 2018 |
| ▶ **EPS theme symposia** | |
| EPS Theme 4 symposium 'Genome Biology', Amsterdam, NL | 15 Dec 2015 |
| EPS Theme 4 symposium 'Genome Biology', Wageningen, NL | 16 Dec 2016 |
| ▶ **National meetings (e.g. Lunteren days) and other National Platforms** | |
| Annual meeting 'Experimental Plant Sciences', Lunteren, NL | 14-15 Apr 2014 |
| Annual meeting 'Experimental Plant Sciences', Lunteren, NL | 13-14 Apr 2015 |
| Annual meeting 'Experimental Plant Sciences', Lunteren, NL | 10-11 Apr 2017 |
| ▶ **Seminars (series), workshops and symposia** | |
| EPS symposium: Omics advances for academia and industry | 11 Dec 2014 |
| EPS flying seminar: Whole-genome duplications, Yves van de Peer | 02 Feb 2015 |
| EPS flying seminar: Comparative & functional genomics of polyploidy, Jeff J. Doyle | 12 May 2015 |
| Biometris Symposium: From big data to biological solutions | 18 Jun 2015 |
| EPS flying seminar: Perianth evolution in Ranunculaceae, Sophie Nadot | 20 May 2016 |
| EPS seminar: Alternative splicing of a heat stress transcription factor, Sotirios Fragkostefanakis | 02 Nov 2016 |
| Seminar: Automated tetraploid genotype calling and its application to pedigree reconstruction in potato, Jeffrey Endelman | 16 Nov 2016 |
| Seminar: From QTLs to routine DNA-informed breeding: prospects, advances and needs, Cameron Peace | 16 Nov 2016 |
| Seminar: Genome-wide association analysis and prediction in tetraploid potato, Jeffrey Endelman | 18 Nov 2016 |
| EPS symposium 'WURomics-Technology-driven innovations for plant breeding', Wageningen, NL | 15 Dec 2016 |
| Seminar: Impact of ploidy level & genome evolution on the control of the frequency & distribution of recombination events in Brassicas, Alexander Pele | 04 Jul 2017 |
| Seminar: Breeding differently - The digital revolution: high-throughput phenotyping and genotyping, Tony Slater | 17 Jul 2017 |
| WEES seminar: From sex chromosomes to sex determination in Ledidoptera, Frantisek Marec | 25 Oct 2017 |
| ▶ **Seminar plus** | |
| ▶ **International symposia and congresses** | |
| EUCARPIA section Biometrics in Plant Breeding, Wageningen NL | 09-11 Sep 2015 |
| 2nd International graduate conference Nanjing, China | 27-30 Oct 2015 |
| Plant and Animal Genome XXIV, San Diego CA. | 09-13 Jan 2016 |
| ▶ **Presentations** | |

| | |
|---|---|
| *Talks:* | |
| Polyploids consortium meeting, Wageningen, NL | 22 Jan 2015 |
| B-WISE bioinformatics seminar, Wageningen, NL | 07 Apr 2015 |
| Polyploids consortium meeting, Wageningen, NL | 25 Jun 2015 |
| EUCARPIA Biometrics section in Plant Breeding 2015, Wageningen, NL | 11 Sep 2015 |
| 2nd International graduate conference Nanjing, China (invited speaker) | 27 Oct 2015 |
| Plant and Animal Genome XXIV, San Diego CA, USA | 10 Jan 2016 |
| Polyploids consortium meeting, Wageningen, NL | 04 Feb 2016 |
| BioSS General Meeting, James Hutton Institute, Dundee, Scotland | 26 Apr 2016 |
| Plant meets Animal, Wageningen, NL | 22 Jun 2016 |
| EPS Theme 4 symposium, Wageningen, NL | 16 Dec 2016 |
| AWL Lunteren 2017 | 11 Apr 2017 |
| Ernst van der Ende visits Plant Breeding, Wageningen, NL (flash presentation) | 08 May 2017 |
| Dümmen Orange B.V., De Lier, Netherlands (invited speaker) | 01 Jun 2017 |
| Polyploids consortium meeting, Wageningen, NL | 27 Jun 2017 |
| Polyploids consortium meeting, Wageningen, NL | 29 Nov 2017 |
| *Poster:* | |
| AWL Lunteren 2015 | 13-14 Apr 2015 |
| ▶ **IAB interview** | |
| ▶ **Excursions** | |
| Visit to Enza Zaden (EPS trip) | 12 Jun 2015 |
| *Subtotal Scientific Exposure* | *24.4 credits** |

| **3) In-Depth Studies** | *date* |
|---|---|
| ▶ **EPS courses or other PhD courses** | |
| Introduction to theory and implementation of genomic selection | 13-17 Oct 2014 |
| Bayesian statistics | 20-21 Oct 2014 |
| ▶ **Journal club** | |
| Member of literature discussion group Quantitative Genetics | 2015-2018 |
| ▶ **Individual research training** | |
| BioSS, James Hutton Institute, Invergowrie, Dundee, Scotland. | 04 Apr-04 May 2016 |
| *Subtotal In-Depth Studies* | *6.5 credits** |

| **4) Personal development** | *date* |
|---|---|
| ▶ **Skill training courses** | |
| Competence Assessment (CA) | 04 Nov 2014 |
| EPS Introduction course | 20 Jan 2015 |
| Scientific Publishing | 15 Oct 2015 |
| Brain Training | 20 Sep 2016 |
| Reviewing a scientific paper (RSP) | 22 Sep 2016 |
| Teaching Outside of Academia | Sep-Dec 2017 |
| ▶ **Organisation of PhD students day, course or conference** | |
| EPS Get2Gether 2017 (Sponsorship, Abstract book, Company stalls) | 2016-2017 |
| EPS Get2Gether 2018 (Speakers, Poster) | 2017-2018 |
| ▶ **Membership of Board, Committee or PhD council** | |
| EPS Council member | Mar 2016-Mar 2018 |
| *Subtotal Personal Development* | *9.6 credits** |

| **TOTAL NUMBER OF CREDIT POINTS*** | **54.0** |
|---|---|

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

This research was funded through the Topconsortium voor Kennis en Innovatie (TKI) Tuinbouw & Uitgangsmaterialen, project numbers BO-26.03-002-001, BO-26.03-009-004 and BO-50-002-022. The author received an EMBO short-term fellowship (ASTF number 228 – 2016) to work at Biomathematics & Statistics Scotland (BioSS), Dundee, Scotland.

Financial support from Wageningen University & Research for the printing of this thesis is gratefully acknowledged.

Cover design: Peter M. Bourke (using Circos | circos.ca)
Printed by: GVO drukkers & vormgevers, Ede | www.gvo.nl