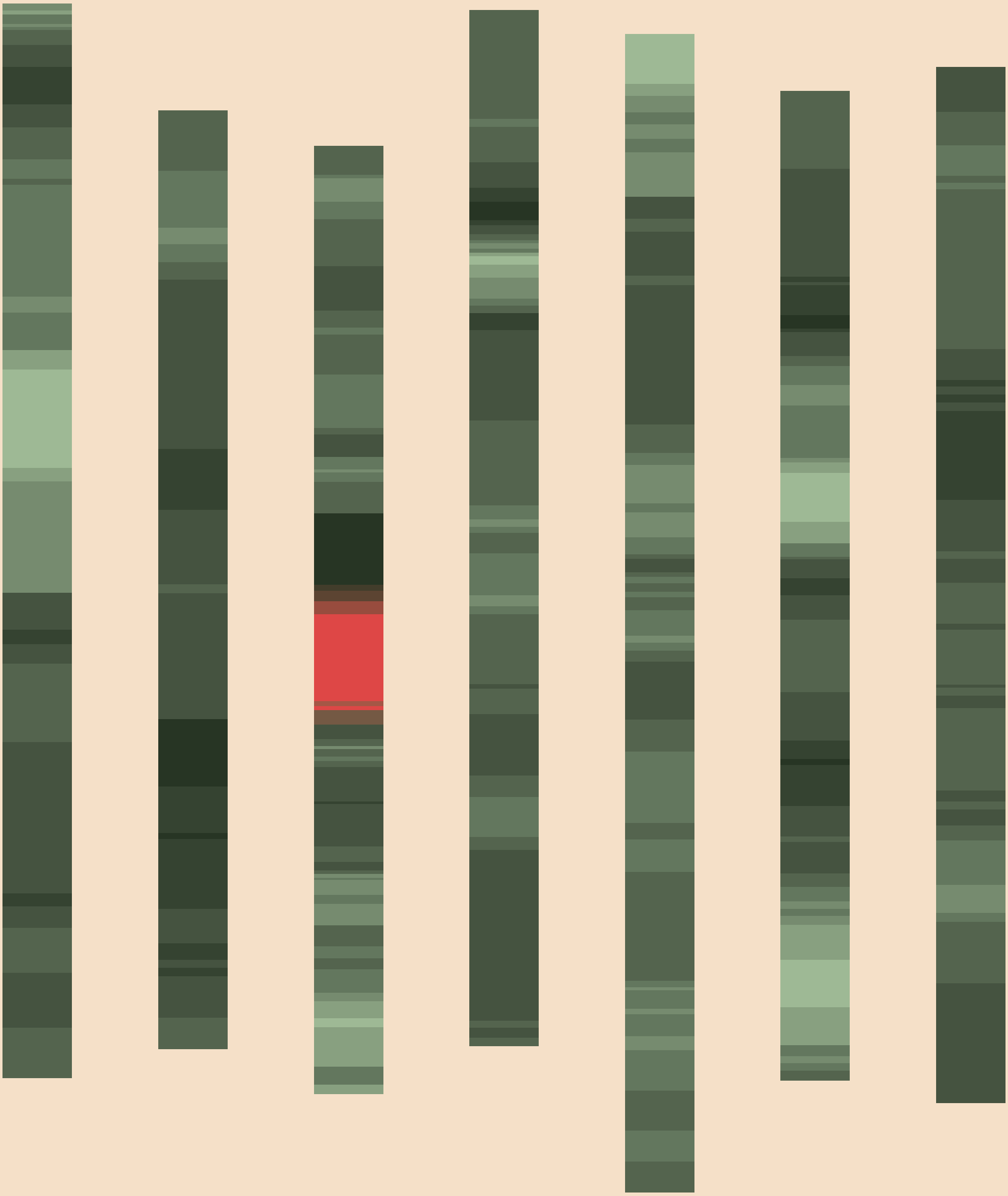
Harvesting data from polyploid plants

developing tools for genetic analysis in strawberry



ALEJANDRO THÉRÈSE NAVARRO

Harvesting data *from* polyploid plants

developing tools for genetic analysis in strawberry

Alejandro Thérèse Navarro

Thesis committee

Promotor

Prof. dr. R.G.F. Visser
Professor and chair of Plant Breeding
Wageningen University & Research

Co-promotors

Dr. C.A. Maliepaard
Associate professor, Plant Breeding
Wageningen University & Research

Dr. P.M. Bourke
Assistant professor, Plant Breeding
Wageningen University & Research

Other members

*Prof. dr. F.A. van Eeuwij*k, Wageningen University & Research
Prof. dr. B.J. Zwaan, Wageningen University & Research
Dr. S. Smit, Wageningen University & Research
Dr. A. Shahin, BASF Vegetables Seeds - Nunhems

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

Harvesting data *from* polyploid plants

developing tools for genetic analysis in strawberry

Alejandro Thérèse Navarro

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 Wednesday 28 June 2023
at 1.30 p.m. in the Omnia Auditorium.

Alejandro Thérèse Navarro
Harvesting data from polyploid plants: developing tools for genetic analysis in strawberry, 232 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English, Catalan and Spanish

ISBN: 978-94-6447-701-6
DOI: https://doi.org/10.18174/630278

Table of Contents

Summary

In the past 30 years there has been a quick progression of technological advances in DNA reading technologies, resulting in a vastly increased amount of genetic data. As a result, many techniques commonly applied in plant breeding have greatly advanced. Among such methods we find quantitative trait loci (QTL) studies, aimed at finding markers linked to a specific trait using genotype-phenotype associations. Also linkage mapping approaches, which leverage recombination frequencies between markers (linkage), to build genetic maps, have greatly advanced. Such developments have lagged in polyploids, organisms that contain more than two copies of each chromosome and need to be analysed using specific models. Many economically important crops are polyploid, such as potatoes, bananas, cotton, wheat, kiwifruit, or roses. The garden strawberry (*Fragaria* x *ananassa*) is also a polyploid, with eight copies of each of its seven chromosomes, organised as four diploid subgenomes. Due to its high ploidy and high similarity between subgenomes, genetic and genomic studies have been historically very challenging in strawberry. One of the most important and variable traits of strawberry is its aroma, a trait determined by a mixture of volatile compounds produced through a vast network of metabolic processes. Due to the complexity of aroma and of strawberry genetics, there is not much known about the underlying genetic mechanisms that control the wide diversity of strawberry aromas. This thesis elaborates on these topics, expanding current methodologies to scenarios that were previously not possible to be studied in polyploid organisms and shedding light on the effectivity of different approaches to handle polyploid genetic data.

In **Chapter 2** the concept of multiparental populations (MPPs) was reviewed in a polyploid context. MPPs are all those populations formed by a set of individuals whose ancestors can be traced back to a limited number of founder individuals. A breeding programme is the prime example of an MPP, since they are formed by concurrent crossings from a limited set of founders. QTL analyses usually hinge on the development of specific experimental populations, most commonly by crossing two parents. As such, many QTL analyses are limited to analysing the genetic diversity present in those parents, ignoring the wider diversity that can be present in the species under study. MPPs resolve this issue by including a wider germplasm, but without moving to the genome-wide association study (GWAS) setting, which carries statistical issues with rare alleles and the influence of genetic structure. At the moment of writing this review, there were no tools available to perform QTL analysis in multiparental populations of polyploids, however, many models had been proposed to perform such analysis in diploids. I reviewed these models and pointed towards the need of tracking identity-by-descent (IBD) in MPPs, especially in heterozygous polyploid organisms which may harbour many different alleles. Tracking such IBD is especially challenging in polyploids, but haplotyping methods show promise in being able to estimate IBD.

In **Chapter 3** a model to perform QTL analysis in MPP of polyploids was presented, evaluated and implemented in the R package mpQTL. To study the performance of this model a set of simulated autotetraploid multiparental populations was developed following the nested ancestral mapping (NAM) design. A NAM population consists of a single central parent that is crossed with multiple peripheral parents, thus producing a set of F1 populations that have one parent in common. Each NAM contained a varying degree of genetic diversity, which was expected to affect the number of segregating alleles in each NAM population and consequently impact the QTL analysis. The model was evaluated using biallelic markers, true IBD markers and haplotype markers that were expected to track IBD. Both IBD and haplotype markers were multiallelic, a key innovation of this model implementation that is not common in diploids or polyploids. The evaluation found that the scenario with lower genetic diversity yielded more statistical power to detect and locate QTLs, and that multiallelic models (IBD or haplotypes) were more powerful

than biallelic ones. Most crucial was the observation that multiallelic models produced more accurate QTL positions than biallelic ones. This chapter showed that MPP analysis is possible and powerful in polyploids, highlighting the impact of genetic diversity on QTL detection in MPP populations.

In **Chapter 4** an approach to perform linkage mapping using error-prone genotype data was developed and evaluated, with specific considerations for polyploid linkage mapping. Modern genotyping techniques based on sequencing technology can be used to genotype hundreds of thousands or even millions of polymorphisms cheaply. Typically, they produce less accurate data than older technologies like SNP arrays. This decrease in accuracy is particularly prominent using skim-sequencing, a technique that aims at sequencing using low read depth. Genotyping errors greatly affect linkage mapping due to inaccurate estimation of recombination frequencies. Thus, a methodology was developed which is able to utilize error-prone genotypes to iteratively construct linkage maps and this method was implemented in the R package Smooth Descent. Using a set of diploid and polyploid simulations, we showed that Smooth Descent can be used to increase genotype accuracy in diploids, while in polyploids it is mostly useful to improve linkage map construction. The tool was also applied to real datasets, where it was used to estimate error rates in a strawberry population and improve genotyping in tetraploid potato and hexaploid sweet potato. We also showed how our tool can correct genotypes at a similar rate than other tools in the field, with better performance in time and resulting in better genetic maps.

In **Chapter 5** linkage maps for a strawberry (*Fragaria* x *ananassa*) population were produced, using error-prone data generated through whole-genome skim sequencing. Initially, 10.24M polymorphic sites were discovered by comparing reads to the “Camarosa” genome assembly. An initial filtering based on read coverage and segregation reduced the number of markers to 4.04M. Since only 46 individuals were genotyped, we expected an oversaturated map with many completely co-segregating markers which were grouped into bins. Bin size, the number of markers within a bin, was very variable and could be used to remove markers with a high number of genotyping errors. Using the physical locations of markers within each bin, bins could be assigned to

chromosomes. An average of ~5% of markers within each bin presented conflicting information, being assigned to one chromosome but originating from a different chromosome of the genome assembly. This revealed assembly issues in the "Camarosa" genome. After removing conflict markers and those markers from bins that could not be clearly assigned to any chromosome, 1.85M markers remained, grouped in 6567 bins. These markers were used with Smooth Descent to produce linkage maps, with an estimated genotyping error ranging from 1 to 10% for samples with a read depth below 10x. Lastly, the maps produced using Smooth Descent were compared with the "Camarosa" assembly and an equivalent linkage map produced using more accurate SNP array genotypes. The results showed that both linkage maps are highly colinear (0.76 Spearman correlation). This research proved that, with adequate software adapted to error-prone situations, skim-sequencing genotypes can be successfully used to produce high quality linkage maps, even in the case of allopolyploid crops.

Lastly, in **Chapter 6,** metabolomic data from volatile organic compounds (VOC) of strawberry varieties were used to perform a multivariate QTL study. One of the most crucial traits of strawberry (*F.* x *ananassa*) is its aroma, which is greatly determined by the VOCs produced by the ripening fruit. Aroma is also one of the most variable traits in strawberry, with over 300 compounds having been reported. As a complex trait that is determined by the combined abundance of multiple compounds, we proposed that using a multivariate approach to summarise metabolites into a few multivariate traits, each representing a single group of correlated metabolites, would be helpful in uncovering main regulators of volatile production in strawberry. In a biparental cross together with a diverse population, 125 compounds were detected and 96 were confidently identified. Most compounds were esters, terpenoids and fatty acids or compounds derived from fatty acids. The multivariate analysis revealed that all terpenoids were tightly correlated with each other, as were esters within their own chemical class. We found a major regulator influencing the abundance of at least 8 and possibly up to 17 different terpenoid compounds in chromosome 3C. By estimating genomic positions of previously reported QTLs we were able to confirm this QTL in three previous studies. We did not find any major QTL for esters and we found little overlap of such

QTLs across the literature, confirming a lack of repeatability between studies that had previously been reported. In this Chapter we showed the usefulness of multivariate methods to coordinate the analysis of many related traits, simplifying the QTL study and helping interpret its results. Moreover, we discovered an important regulator of terpenoid biosynthesis that is an interesting target for further research and for applied breeding.

Resum

En els últims 30 anys la tecnologia de lectura del DNA ha progressat ràpidament, duent-nos com a conseqüència a un increment substancial de la quantitat de dades genètiques. Aquests desenvolupaments han propel·lit també les tècniques de millora genètica vegetal. Per exemple l'estudi de locus de caràcters quantitatius (QTL, acrònim de l'anglès *quantitative trait locus*), que empra marcadors genètics per trobar associació entre fenotip i genotip, localitzant així els gens reguladors del caràcter estudiat. També ha avançat significativament la cartografia genètica, una tècnica que permet la construcció de mapes genètics mitjançant l'estudi de freqüències de recombinació entre marcadors. Aquests desenvolupaments s'han produït més lentament en el camp dels poliploides, aquells organismes que contenen més de dues còpies de cada cromosoma, ja que requereixen models matemàtics especialitzats. No obstant això, molts cultius de gran importància econòmica són poliploides, entre d'altres les patates, els plàtans, el cotó, el blat, el kiwi o les roses. El maduixot (*Fragaria* x *ananassa*) també es poliploide, amb vuit copies de cadascun dels seus set cromosomes, que es poden dividir en quatre subgenomes diploides. L'alta ploidia i la semblança entre genomes han dificultat greument els estudis genètics i genòmics d'aquest fruit, el qual explica l'estat relativament endarrerit de la genètica del maduixot. Un dels caràcters més importants i variables d'aquest fruit és l'aroma, determinat pel perfil de compostos volàtils produïts per una gran xarxa de processos metabòlics. Degut a la complexitat d'aquest caràcter i de la genètica subjacent, encara no s'han esclarit els mecanismes que controlen la gran diversitat metabòlica de l'aroma dels maduixots. En aquesta tesi s'elaboren els temes presentats, expandint diverses tècniques a escenaris poliploides que prèviament no es po-

dien estudiar, esclarint així l'efectivitat de diferents mètodes d'estudi de la genètica poliploide.

Al **Capítol 2** es presenten les poblacions multiparentals (MPP, acrònim de l'anglés *multiparental population*) en una revisió orientada als poliploides. Les MPP són aquelles poblacions formades per un conjunt d'individus que comparteixen un grup limitat d'ancestres fundadors. Les poblacions produïdes en un programa de millora genètica son un exemple ideal de MPP, ja que es generen creuant un petit conjunt de fundadors. Els anàlisis de QTL usualment es duen a terme en poblacions descendents de dos progenitors. Conseqüentment, la diversitat genètica present en aquestes poblacions és limitada per la diversitat dels progenitors escollits, ignorant així doncs la diversitat genètica de la població general. Una MPP resol aquest problema incloent un germoplasma més ampli, però sense arribar al nivell d'un GWAS, que usualment implica problemes estadístics degut a la baixa freqüència d'al·lels rars i a la influencia de l'estructura genètica. Quan aquesta revisió es va escriure, no existia cap mètode per estudiar QTLs a MPP poliploides, tot i que s'havien proposat diversos models equivalents per diploides. En aquest capítol, s'han revisat els models diploides, senyalant la necessitat de rastrejar la identitat per descendència (IBD, acrònim de l'anglés *identity by descent*), especialment en organismes poliploides i heterozigots, que tenen una alta probabilitat de contenir múltiples al·lels. Rastrejar la IBD de organismes poliploides és especialment difícil, però nous mètodes d'haplotipat són esperançadors i prometen estimacions precises de la IBD.

Al **Capítol 3** es presenta i avalua un model d'anàlisi de QTLs en MPP poliploides, implementat en el paquet de R mpQTL. El rendiment d'aquest model s'ha estudiat fent servir un conjunt de simulacions de poblacions multiparentals autotetraploides, desenvolupades seguint l'esquema *nested ancestral mapping* (NAM, cartografia ancestral niuada). Una població NAM consisteix en un sol progenitor central creuat amb múltiples progenitors perifèrics, formant així una sèrie de poblacions F1 que comparteixen un sol progenitor. Cada NAM simulada conté una quantitat diferent de diversitat genètica, el qual s'esperava que afectés al número d'al·lels segregant en cada població, amb conseqüències directes per l'anàlisi de QTLs. El model s'ha avaluat fent servir marcadors bial·lelics, marcadors d'IBD real i marcadors d'haplotips, que s'espera que siguin capaços de rastrejar la IBD. Tant els marcadors d'IBD real com els haplotips eren multial·lelics, una innovació clau del model implementat en aquest capítol que no es comú a models diploides o poliploides. Durant l'avaluació s'ha trobat que l'escenari amb menor diversitat genètica és el més estadísticament poderós per detectar i localitzar QTLs, i que els models multial·lelics (tant d'IBD com d'haplotips) són més poderosos que els bial·lelics. L'observació més crucial ha estat que els models multial·lelics estimen la posició de QTLs amb major precisió. Així doncs, aquest capítol demostra que l'anàlisi de MPP poliploides es possible i poderós, destacant l'impacte de la diversitat genètica a l'hora de detectar QTLs en poblacions MPP.

Al **Capítol 4** es desenvolupa i avalua un mètode de mapeig de lligament genètic fent servir dades genotípiques propenses a error, amb consideracions específiques pel mapeig d'organismes poliploides. Els mètodes moderns de genotipat basats en tecnologies de seqüenciació produeixen genotips per centenars, milers o fins i tot milions de marcadors per un baix cost. Típicament, la precisió d'aquestes tecnologies es menor que la de tècniques més antigues com els *SNP array*. Aquesta disminució en precisió es particularment prominent emprant la tècnica *skim sequencing* (seqüenciació superficial), un mètode de seqüenciació a baixa profunditat de lectura. Els errors de genotipat afecten greument els mapes de lligament, ja que produeixen estimacions de lligament entre marcadors molt imprecises. Degut a això, al capítol 4 s'ha desenvolupat una metodologia que permet crear mapes genètics fent servir genotips propensos a errors a través del cartografiat iteratiu. Aquest mètode s'ha implementat en el paquet de R *Smooth Descent*. Fent servir simulacions diploides i poliploides he demostrat que *Smooth Descent* es pot fer servir per augmentar la precisió de genotipat en organismes diploides, mentre que en organismes poliploides aquesta eina millora considerablement els mapes genètics. *Smooth Descent* s'ha aplicat també a genotips de poblacions reals, on s'ha fet servir per estimar el percentatge d'errors de genotipat en maduixots, per millorar el genotipat de patates tetraploides i de moniatos hexaploides. S'ha demostrat també que aquesta eina es capaç de corregir genotips d'una manera similar o millor que altres eines del camp, fent servir menys temps i produint millors mapes genètics.

Al **Capítol 5** es van crear mapes genètics pel maduixot (*Fragaria* x *ananassa*) utilitzant genotips propensos a error generats mitjançant *skim sequencing* del genoma complet. Inicialment es van descobrir 10,24 milions de marcadors polimòrfics comparant les seqüències amb l'assemblatge del genoma "Camarosa". Després d'aplicar un filtre basat en la cobertura i segregació, es va reduir el conjunt de marcadors a 4,04 milions. Amb només 46 individus genotipats, s'esperava un mapa sobresaturat amb marcadors co-segregants, els quals es van agrupar en manats. El número de marcadors per manat es va utilitzar per detectar aquells manats amb un gran percentatge d'error, descartant-los de l'anàlisi. Fent servir les posicions físiques (al genoma) dels marcadors de cada manat, es van poder assignar cada manat a un cromosoma, el qual va revelar una mitjana de ~5% dels marcadors assignats a un cromosoma, però originaris d'un altre cromosoma del genoma "Camarosa". Després de depurar els marcadors conflictius i aquells en manats que no es podien assignar a cap cromosoma, es van obtenir 1,85 milions de marcadors agrupats en 6576 manats. L'eina *Smooth Descent* es va emprar per generar mapes genètics, estimant alhora un error de genotipat entre el 1% i el 10% per a mostres amb profunditats de lectura inferiors a 10x. Finalment, comparant aquest mapa amb l'assemblatge del genoma "Camarosa" i un mapa genètic equivalent produït utilitzant marcadors SNP més precisos, es va observar una alta correlació (0,76). Aquesta investigació demostra que, amb el software adequat per a dades propenses a errors, és possible utilitzar amb èxit genotips obtinguts mitjançant seqüenciació superficial per a la producció de mapes genètics d'alta qualitat, fins i tot en organismes al·lopoliploides.

Per últim, al **Capítol 6**, dades metabòliques de compostos volàtils orgànics (VOC, de l'anglès *volatile organic compound*) de maduixa es van fer servir per dur a terme un estudi de QTL multivariat. Un dels caràcters més importants del maduixot (*F.* x *ananassa*) és l'aroma, que en gran mesura és determinat pels VOCs produïts durant la maduració del fruit. L'aroma és també un dels caràcters més variables del maduixot: s'han descrit més de 300 compostos diferents. Degut a la complexitat d'aquest caràcter controlat per l'abundància de múltiples compostos, es va proposar fer servir un mètode multivariat per resumir els metabòlits en uns pocs caràcters multivariats, cadascun representant un grup de metabòlits correlacionats, ajudant així a descobrir regula-

dors principals de la producció de volàtils. Estudiant una població biparental i un panell de diversitat, es van descobrir 125 compostos dels quals 96 van ser identificats. La majoria de compostos eren èsters, terpenoides i lípids o derivats dels lípids. L'anàlisi multivariat va revelar que tots els terpenoides estaven altament correlacionats, així com els esters amb el seu grup químic. S'ha trobat un regulador principal influint l'abundància des de 8 fins a 17 terpenoides diferents al cromosoma 3C. També s'han estimat les posicions genòmiques de les QTLs descrites prèviament a la literatura, el qual ha permès confirmar aquest regulador principal. No s'ha trobat cap regulador principal de la producció d'èsters i els resultats prèviament descrits mostren poc solapament entre estudis, ressaltant així la poca repetibilitat d'aquest tipus de recerca. En aquest capítol s'ha demostrat la utilitat dels mètodes multivariats en la coordinació d'anàlisis complexos de caràcters interrelacionats, simplificant així l'estudi de QTLs i ajudant a la interpretació dels seus resultats. Addicionalment, s'ha descobert un regulador principal de la biosíntesi de terpenoides, el qual pot ser un caràcter interessant per futures investigacions i per millora genètica aplicada.

Resumen

Durante los últimos 30 años la tecnología de lectura del ADN ha progresado rápidamente, trayendo consigo un incremento substancial de la cantidad de datos genéticos disponibles. Estos desarrollos han propulsado también las técnicas de mejora genética vegetal. Entre ellas, el estudio de locus de caracteres cuantitativos (QTL, acrónimo del inglés *quantitative trait locus*), que usa marcadores genéticos para encontrar asociaciones entre genotipo y fenotipo, localizado así los genes reguladores de caracteres. También se han producido avances significativos en el mapeo genético, una técnica que permite la construcción de mapas genéticos mediante el estudio de las frecuencias de recombinación entre marcadores. Estos desarrollos se han producido con más lentitud en el campo de los poliploides, debido a que requieren models matemáticos especializados. Sin embargo, muchos cultivos de gran importancia son poliploides, entre otros las patatas, los plátanos, el algodón, el trigo, el kiwi o las rosas. El fresón (*Fragaria* x *ananassa*) también es poliploide, con cuatro copias de cada uno de sus siete cromosomas, organizados en cuatro subgenomas diploides. La alta ploidía i la similitud entre sus genomas han dificultado gravemente los estudios genéticos y genómicos de esta fruta, lo cual explica el estado relativamente atrasado de la genética del fresón. Uno de los caracteres más importantes y variables de esta fruta es su aroma, determinado por el perfil de compuestos volátiles que a su vez son producidos por una compleja red de procesos metabólicos. Debido a la complejidad de este carácter i a la genética subyacente, todavía no se han descubierto los mecanismos que controlan el aroma de los fresones. En esta tesis se elabora sobre los temas aquí presentados, expandiendo diversas técnicas a escenarios poliploides que previamente no podían estudiarse, aclarando así la

efectividad de diversos métodos para el estudio de la genética poliploide.

En el **Capítulo 2** se presentan las poblaciones multiparentales (MPP, acrónimo del inglés *multiparental population*) en una revisión orientada a los poliploides. Las MPP son aquellas poblaciones formadas por un conjunto de individuos que comparten un grupo limitado de ancestros fundadores. Las poblaciones producidas en un programa de mejora genética son un ejemplo ideal de MPP, ya que se generan cruzando un pequeño conjunto de fundadores. Los análisis de QTL usualmente se llevan a cabo en poblaciones descendientes de dos progenitores. Consecuentemente, la diversidad genética presente en estas poblaciones se ve limitada por la diversidad de los progenitores escogidos, ignorando así la diversidad genética de la población general. Una MPP resuelve este problema incluyendo un germoplasma más amplio, pero sin llegar al nivel de un GWAS (*genome-wide association study*), que usualmente implica problemas estadísticos debido a la baja frecuencia de los alelos raros y a la influencia de la estructura genética. Cuando esta revisión se escribió, no existía ningún método para estudiar QTLs en MPP poliploides, aunque se habían propuesto varios modelos equivalentes para diploides. En este capítulo, se revisan los modelos diploides, señalando la necesidad de rastrear la identidad por descendencia (IBD, acrónimo del inglés *identity by descent*), especialmente en organismos poliploides y heterocigotos, que tienen una alta probabilidad de contener múltiples alelos. Rastrear la IBD de organismos poliploides es especialmente difícil, pero nuevos métodos de haplotipado son esperanzadores y prometen estimaciones precisas de la IBD.

En el **Capítulo 3** se presenta y evalúa un modelo de análisis de QTLs en MPP poliploides, implementado en el paquete de R mpQTL. El rendimiento de este modelo se ha estudiado usando un conjunto de simulaciones de poblaciones multiparentales autotetraploides, desarrolladas con el esquema *nested ancestral mapping* (NAM, cartografía ancestral anidada). Una población NAM consiste en un solo progenitor central cruzado con múltiples progenitores periféricos, formando así una serie de poblaciones F1 que comparten un sólo progenitor. Cada NAM simulada contiene una cantidad diferente de diversidad genética, el cual se esperaba que afectara al número de alelos segregando en cada población, con consecuencias directas para el análisis de QTLs. El modelo se ha evaluado usando marcadores bialélicos, marcadores de IBD real y marcadores de haplotipos, que se esperaba que fueran capaces de rastrear la IBD. Tanto los marcadores de IBD real como los haplotipos eran multialélicos, una innovación clave del modelo implementado en este capítulo que no es común en modelos diploides o poliploides. Durante la evaluación se ha encontrado que el escenario con menor diversidad genética es el más estadísticamente poderoso para detectar y localizar QTLs, y que los modelos multialélicos (tanto de IBD como de haplotipos) son más poderosos que los bialélicos. La observación más crucial ha sido que los modelos multialélicos estiman la posición de QTLs con mayor precisión. Así pues, este capítulo demuestra que el análisis de MPP poliploides es posible, destacando el impacto de la diversidad genética en la detección de QTLs en poblaciones MPP.

En el **Capítulo 4** se desarrolla y evalúa un método de mapeo de ligamento genético usando datos genotípicos propensos a error, con consideraciones específicas para organismos poliploides. Los métodos modernos de genotipado basados en tecnologías de secuenciación producen genotipos para centenares, miles o incluso millones de marcadores por un bajo coste. Típicamente, la precisión de estas tecnologías es menor a la de técnicas más antiguas como los *SNP array*. Esta disminución en precisión es particularmente prominente empleando la técnica *skim sequencing* (secuenciación superficial), un método de secuenciación a baja profundidad de lectura. Los errores de genotipado afectan gravemente a los mapas de ligamento, ya que producen estimaciones de ligamento entre marcadores muy imprecisas. Debido a esto, en el capítulo 4 se ha desarrollado una metodología que permite crear mapas genéticos usando genotipos propensos a error a través del cartografiado iterativo. Este método se ha implementado en el paquete de R *Smooth Descent*. Usando simulaciones diploides y poliploides se ha demostrado que *Smooth Descent* se puede utilizar para aumentar la precisión de genotipado en organismos diploides, mientras que en organismos poliploides esta herramienta mejora considerablemente los mapas genéticos. *Smooth Descent* se ha aplicado también a genotipos de poblaciones reales, donde se ha usado para estimar el porcentaje de errores de genotipado en fresones, para mejorar el genotipado de patatas tetraploides y de moniatos hexaploides. Se ha demostrado también que esta herramienta es capaz de corregir genotipos de una manera similar o mejor que otras herramientas del campo, usando menos tiempo y produciendo mejores mapas genéticos.

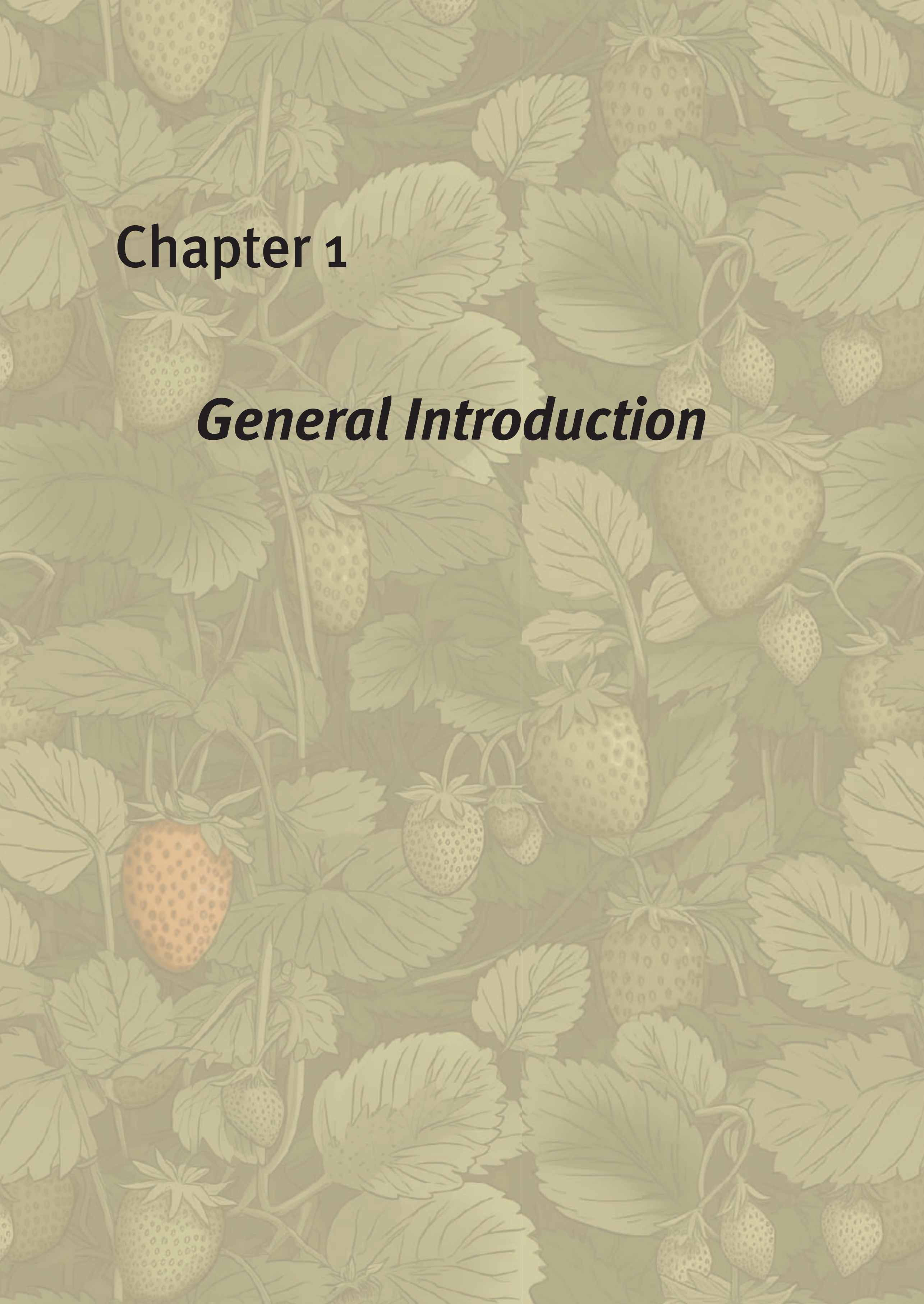
En el **Capítulo 5** se crearon mapas genéticos para el fresón (*Fragaria* x *ananassa*) utilizando genotipos propensos a error generados mediante *skim sequencing* del genoma completo. Inicialmente se descubrieron 10,24 millones de marcadores polimórficos comparando las secuencias con el ensamblaje del genoma "Camarosa". Tras aplicar un filtro basado en la cobertura y segregación, se redujo el conjunto de marcadores a 4,04 millones. Con sólo 46 individuos genotipados, se esperaba un mapa sobresaturado con marcadores co-segregantes, los cuales se agruparon en manojos. El número de marcadores por manojo se utilizó para detectar aquellos manojos con un gran porcentaje de error, descartándolos del análisis. Usando las posiciones físicas (en el genoma) de los marcadores de cada manojo, se pudo asignar cada manojo a un cromosoma, lo cual reveló un promedio de 5% de los marcadores asignados a un cromosoma, pero originarios de otro cromosoma del genoma "Camarosa". Tras depurar los marcadores conflictivos y aquellos en manojos que no se podían asignar a ningún cromosoma, se obtuvieron 1,85 millones de marcadores agrupados en 6576 manojos. La herramienta *Smooth Descent* se empleó para generar mapas genéticos, estimando a la vez un error de genotipado entre el 1% y el 10% para muestras con profundidades de lectura inferiores a 10x. Finalmente, comparando este mapa con el ensamblaje del genoma "Camarosa" y un mapa genético equivalente producido utilizando marcadores SNP más precisos, se observó una alta correlación entre ellos (0,76). Esta investigación demuestra que, con el software adecuado para datos propensas a errores, es posible utilizar con éxito genotipos obtenidos mediante secuenciación superficial para la producción de mapas genéticos de alta calidad, incluso en organismos alopoliploides.

Por último, en el **Capítulo 6**, datos metabólicos de compuestos volátiles orgánicos (VOC, del inglés *volatile organic compound*) de fersón se utilizaron para llevar a cabo un estudio de QTL multivariado. Uno de los caracteres más importantes del fresón (*F.* x *ananassa*) es el aroma, que en gran medida es determinado por los VOCs producidos durante la maduración del fruto. El aroma es también uno de los caracteres más variables del fresón: en él se

han descrito más de 300 compuestos diferentes. Debido a la complejidad de este carácter controlado por la abundancia de múltiples compuestos, se propuso utilizar un método multivariado para resumir los metabolitos en unos pocos caracteres multivariados, cada uno representando un grupo de metabolitos correlacionados, ayudando así a descubrir reguladores principales de la producción de volátiles. Estudiando una población biparental y un panel de diversidad, se descubrieron 125 compuestos de los cuales 96 fueron identificados. La mayoría de los compuestos eran ésteres, terpenoides y lípidos o derivados de los lípidos. El análisis multivariado reveló que todos los terpenoides estaban altamente correlacionados, así como los ésteres con su grupo químico. Se ha encontrado un regulador principal controlando la abundancia de desde 8 hasta 17 terpenoides diferentes en el cromosoma 3C. También se han estimado las posiciones genómicas de las QTLs descritas previamente en la literatura, lo cual ha permitido confirmar este regulador principal. No se ha encontrado ningún regulador principal de la producción de ésteres y los resultados previamente descritos muestran poco solapamiento entre estudios, resaltando así la poca repetibilidad de este tipo de investigación. En este capítulo se ha demostrado la utilidad de los métodos multivariados en la coordinación de análisis complejos de caracteres interrelacionados, simplificando así el estudio de QTLs y ayudando a la interpretación de sus resultados. Adicionalmente, se ha descubierto un regulador principal de la biosíntesis de terpenoides, el cual puede ser un carácter interesante para futuras investigaciones y para la mejora genética aplicada.

Chapter 1

General Introduction



General Introduction

The study of plants has fascinated humanity since ancient times. Such can be seen by Theophrastus' *Historia Plantarum* and Dioscorides' *De Materia Medica*, which provide us a glimpse into the study of plants in antiquity. These texts already contained some of the questions that still intrigue us today and that contributed to the development of this thesis. Most noteworthy for this book is the question of *heredity*: why and how traits are inherited from parents to descendants. Nowadays we know this scientific field as genetics, a blooming area of the biological sciences that has grown enormously in the last century. At its origin, many wrong theories were proposed to explain heredity, among them *preformationism*, *spermism* or *ovism.* Only the theory of blending inheritance, which proposed that offspring were the result of a mixture of both parents, was somewhat on the right track, although still largely inaccurate.

Modern genetics cannot start with anyone other than Gregor Mendel, a monk from Moravia, in the modern Czech Republic. During the mid-19th century Mendel cultivated an experimental garden of 2 hectares where he grew about 29.000 pea plants. His studies on the shape and colour of seeds and flowers showed that many of his traits were independently inherited. He theorised what are now known as the *Mendelian laws of inheritance* (Corcos and Monaghan 2008; Gayon 2016; Abbott and Fairbanks 2016). However, the major contribution of his research was the proposition of factors -now called *genes*-, discrete units that would be independently assorted and somehow carry information to the next generation. The exact nature of these factors, however, would remain a mystery for another century.

Several discoveries corroborated and expanded Mendel's theory. First came Nettie Stevens, a biologist that studied inheritance by analysing the cell nucleus of mealworm sperm. By observing their chromosomes, she was able to show that sperm cells carrying a large (X) chromosome would produce females, while those carrying a small (Y) chromosome would produce males (Stevens 1905; Carey et al. 2022). She simultaneously discovered sex chromosomes and proved that trait inheritance was somehow related to chromosomal inheritance. Her discoveries led Thomas Hunt Morgan and his students to the analysis of sex-linked traits in *Drosophila melanogaster,* the fruit fly. By studying these traits, he developed the idea of genetic linkage: some genes are linked to others and often, but not always, inherited together. By observing the number of times that two traits were co-inherited, he hypothesized the phenomenon of cross-over, or recombination. These findings led Sturtevant, his student, to the development of the first ever genetic map (Sturtevant 1913), which finally convinced their colleagues that chromosomes contained genes organised in a linear fashion. His cross-over hypothesis was later confirmed when Barbara McClintock was able to show recombination in the meiosis of maize chromosomes with her advanced chromosome staining techniques (Creighton and McClintock 1931). However, the exact nature of the gene and the molecule that was carrying information were still controversial topics.

Only twenty years later this long open question was finally concluded. Nikolai Koltsov proposed in 1927 that traits would be inherited by a "giant hereditary molecule" made up of "two mirror strands" (Koltsov 1927; Soyfer 2001), an idea later popularised by Erwin Schrödinger as an "aperiodic crystal" (Schrödinger 1945; Varn and Crutchfield 2016). They would be proven correct with the discovery of the double helix structure of deoxyribonucleic acid (DNA) by James D. Watson and Francis Crick using X-ray diffraction images obtained by Rosalind Franklin (Watson and Crick 1953). However, it was the Hershey-Chase experiments what would convince the scientific community that DNA was the molecule transmitting information (Hershey and Chase 1952). They were able to show that the DNA injected by bacteriophages was the one transmitting genes, and not proteins as many had previously thought. Together, these findings painted a clear picture of how DNA was replicated, inherited and pointed towards answers on how it transmitted information.

Once DNA was found as the key culprit of inheritance, it was time to characterize it. In that regard, the study of bacteriophages and bacteria proved instrumental. A cornerstone of modern biotechnological research was hidden within their cells: sequence-specific restriction enzymes (Smith 1979). By selectively cleaving DNA at precise locations, these enzymes offer researchers a highly accurate means of identifying specific DNA sequences. Soon after, scientists realized that these cleavage sites could serve as markers akin to the traits used by T.H. Morgan and his students, but with the added advantage of observing DNA directly rather than inferring genotypes from Mendelian phenotypes. This discovery of the first DNA markers, known as restriction length fragment polymorphisms (RFLP), paved the way for the further development of genetic mapping. While RFLPs had been previously used to produce recombination maps in adenoviruses, Botstein ground-breaking paper was the first to use RFLP markers to create a chromosome map (Botstein et al. 1980). The publication of this paper opened a fertile field of research that continues to thrive nowadays. Crucially, genetic mapping was greatly propelled by the discovery of RFLP markers, a dynamic that has been repeated over the years as new genetic marker technologies have been developed.

These genetic maps showed much promise, due to the simplicity of their analysis and the possibility to include many more markers than was previously possible. However, Lander and Botstein realized that genetic maps could be used beyond the chromosomal organisation of restriction sites. In a landmark paper they described a method that would allow locating, within a map, the locus of a gene of interest for a particular trait (Lander and Botstein 1989). The principle would be to cross two parents with a contrasting phenotype, analyse the offspring's genotype and phenotype, and associate the marker inheritances with the observed phenotype, thus obtaining a probability profile along each chromosome defining the most likely locus for the causal gene of a trait. Each region of high probability would be a quantitative trait locus (QTL). Their method was designed to use a few dozen markers per chromosome and could analyse biparental and family populations. Although limited, geneticists quickly realised the potential of such studies to underpin causal genes in humans, animals and crop plants, with clear impacts in medicine and agriculture. Eventually, many variations of such models would be developed, adapted to all kinds of situations. Most notably, genome-wide association studies (GWAS), which would allow to perform a similar analysis in natural, not experimental populations, including humans (Ozaki et al. 2002; Klein et al. 2005).

With this, we have reviewed the foundational work that led to the modern field of quantitative genetics. We now know the mechanism by which genes are inherited between generations, how they are organised in the genome and, to a large extent, how they function. DNA markers have become an essential tool in genetics and have developed greatly in the recent years, remaining a vital instrument in genetic analysis. Genetic linkage maps have become the backbone of many important discoveries and their possibilities are being expanded every year, in many cases goaded by innovations in marker technologies. Finally, QTL and association studies have become the magnifier by which to find these needles in the large haystack of DNA. In recent years, all these crucial topics have received major attention, so let us dive into each in order to contextualize this thesis.

Plant breeding in the 21st century

The past 40 years have brought several advancements into the field of plant breeding. From RFLP markers we have moved to high-throughput genotyping techniques that yield thousands of genotypes at a low cost. Sequencing entire genomes has become easily achievable and the computational capacity required to analyse this data is being developed. As a result, plant breeding has moved forward together with the genetics field. Although many aspects of this revolution present interesting challenges, in this thesis I have dealt mostly with the study of DNA, with particular applications in strawberry.

Reading DNA

Since its discovery, DNA has become a key interest in molecular biology. Its organisation, the distribution of genes, the presence or absence of polymorphisms, all have direct implications in the biology of organisms. Novel techniques to study genomes are of special interest to plant breeding, since they allow us to understand the crops we work with more deeply. For instance, an evolutionary study on the rose genome revealed the loci responsible for aroma production and continuous flowering, as well as their historical origin in the Middle East and China (Raymond et al. 2018). Nowadays, two complementary methods are commonly used to read DNA: SNP genotyping and high-throughput sequencing.

If you have read any molecular genetics article in the past 20 years you will likely be familiar with the acronym SNP, the most common marker type. It stands for single nucleotide polymorphism, one base-pair differences, usually biallelic, that are abundant in the whole genome of virtually all species. The first step to genotyping SNPs is discovery, locating SNPs in the genome. After discovery, the actual genotyping step is performed. At each SNP position, the abundance of the SNP alleles are measured, and a genotype score is given. Two main methods exist for high-throughput SNP genotyping, the first based on probe-based SNP arrays (Ganal et al. 2012), the second using exclusively sequencing technologies (Zargar et al. 2016; Torkamaneh et al. 2016). Importantly, sequence-based genotyping provides a larger number of markers, but at a lower accuracy than SNP arrays. For this reason, linkage mapping has required major adaptations in the past years. Firstly, algorithms for linkage mapping can become quite slow as the number of markers increase, prompting the development of efficient methods for sequence-based genotypes (Wu et al. 2008; Liu et al. 2014; Preedy and Hackett 2016; Rastas 2017). Additionally, since accuracy of linkage maps is greatly affected by genotyping error rates (Lincoln and Lander 1992; Cartwright et al. 2007) it is questionable how useful linkage mapping can be with error prone data. The major driver of errors in sequence-based genotyping is read depth (Deschamps et al. 2012; Sims et al. 2014), and since higher depth is equivalent to higher costs, one can expect most researchers to avoid high depth experiments. In fact, some have argued in favour of skim-sequencing approaches, purposefully lowering sequencing depth to reduce costs (Kumar et al. 2021; Adhikari et al. 2022). As we will see, this will have a clear effect on linkage mapping approaches, meaning that new adaptations will be required.

Sequencing, the reading of DNA strands, is slowly becoming the prevalent technology for anything genetics. Its development seems non-stopping (Heather and Chain 2016; Hu et al. 2021b). With increased read length, accuracy and volume one might think that sequences will take over markers. Indeed, assembling entire genomes with a phased sequence per chromosome is becoming a standard, available even at high levels of genome complexity (Colle et al. 2019; Piet et al. 2022). Nevertheless, comparing sequences across individuals remains challenging, with the *pangenomics* approach of comparing entire genomes still under development (Bayer et al. 2020). The main unit of study will likely be the *haplotype*, the exact combination of alleles of a DNA sequence (Garrison and Marth 2012; Clevenger et al. 2018; He et al. 2018). Although a haplotype can span an entire chromosome, as is the case in haplotype-phased assemblies, they can also span much shorter regions, thus becoming multiallelic markers. Importantly, multiallelic markers seem to provide more accurate results in QTL studies, genome association and genomic selection (Wang et al. 2016; Sallam et al. 2020; Bajgain and Anderson 2021; Li et al. 2021). Further development of these type of sequence-based markers seems a promising step forward in genetic analysis.

Understanding DNA

Beyond finding DNA sequences, a major interest of plant breeding is understanding what those sequences do in a biological context. Several approaches exist to do this. Functional analyses and mutation experiments, often called reverse genetics approaches, study specific genes and characterize their action upon phenotypes (Gilchrist and Haughn 2010; Malzahn et al. 2017). Gene prediction approaches can use sequence features, RNA data and protein databases in order to annotate entire genomes with relatively high accuracy (Brůna et al. 2021). However, in the field of quantitative genetics we use the QTL study as the major approach. A QTL study is always based on finding associations between phenotype data and genotype data, and as such it has greatly evolved as both types of datasets have changed.

Originally, QTL studies were only performed in biparental populations. These populations were mostly crossings between diploid, inbred parents, meaning only two alleles segregated in the population. Obviously, this represented an extremely narrow range of the genetic diversity within a species. To alleviate this issue, the GWAS approach was developed, similar in principle to a QTL study, but using evolutionary linkage disequilibrium (LD) among markers to find associations between markers and causal loci. It quickly became evident that the non-random distribution of LD, known as genetic structure, was a major hinderance for QTL analyses and prompted the development of several approaches (Yu et al. 2006; Huang et al. 2019). Nowadays, there is an increased interest in multiparental populations, which can harbour larger diversity than biparental crosses while reducing the effect of genetic structure (Würschum 2012; Garin et al. 2015; Mangandi et al. 2017; Li et al. 2022b). However, they pose significant challenges, especially when these population types deal with more complex polyploid crops. In a biparental QTL analysis we attempt to estimate which parental alleles each individual has inherited, in other words, which markers are *identical by descent* (IBD). In multiparental populations, particularly polyploid ones, estimating IBD is more challenging. Besides the interconnectedness of populations, unknown relatedness between the founder parents of the population complicate IBD estimates. Although several approaches are being designed to solve this issue, they remain exclusive to diploids (Jacquin et al. 2014; Broman et al. 2019; Li et al. 2021; Zheng et al. 2021). It seems likely that using haplotype-based, multiallelic markers in polyploids can help bridge that gap, and although challenging, many tools have been developed to obtain polyploid haplotypes (Motazedi et al. 2017; Clevenger et al. 2018; He et al. 2018; Majidian et al. 2020).

Models must also adapt to changing phenotyping data. Large and complex phenotypic datasets are becoming available (Yang et al. 2020; Hall et al. 2022). One common approach, which involves performing a separate QTL analysis for each new available phenotype, is not practical for large imaging or metabolomic datasets that may contain thousands of variables with intercorrelated features. In this context, it is crucial to develop models that can handle high-dimensional data and account for the usual correlation among variables (e.g. Mitteroecker et al. 2016). More importantly, such methods should increase model interpretability when using such large datasets, a feat that is not easy to achieve.

Application to strawberry

Strawberry is a perfect example of the evolution of Plant Breeding in the past decades due to innovations in the field of genetics. Genetic research has advanced at large strides, quickly moving from molecular marker development to genetic mapping, to genome assembly and candidate gene research. Let's consider some of the challenges of these developments.

The complex puzzle of strawberry DNA

The word "strawberry" can refer to more than 20 different species, and even more interspecific hybrids and cultivars. The most well-known are the woodland strawberry (*F. vesca*) and the large-fruited, commercially cultivated garden strawberry (*F.* x *ananassa*). Due to its horticultural popularity and extensive breeding history, strawberry has been deeply researched since the beginning of the 20th century (Darrow 1966; Folta and Davis 2007). However, the octoploid nature of its genome, with 28 pairs of chromosomes, have made genetic studies on strawberry especially difficult. Nowadays it is believed that strawberry is an allopolyploid due to its disomic segregation and preferential chromosome pairing (Sargent et al. 2009; Whitaker 2011). Allopolyploidy often means that the resulting polyploid is a hybrid between closely related species, thus resulting in two or more *subgenomes* within a nucleus that, although very similar sequence-wise, segregate disomically and have diploid-like meiosis (Tate et al. 2005; Birchler 2012; Soltis et al. 2016). While segregation studies are clear in garden strawberry, its polyploidization history is still a mystery (Fig. 1A).

The origin of *F.* x *ananassa* is well documented as a cross between *F. virginiana* and *F. chiloensis* (Darrow 1966). These two North American octoploid species, found in the west and east coast, share a common octoploid ancestor. However, the origin of this octoploid and therefore the origin of its four diploid subgenomes is a heavily debated topic (Tennessen et al. 2014; Sargent et al. 2015; Edger et al. 2020; Liston et al. 2020; Feng et al. 2021). The development of genetic markers that could help clarify subgenome ancestry was realised early on, yet it proved to be particularly difficult (Galletta and Maas 1990; Hokanson 2001; Davis et al. 2006; Bassil et al. 2015). Such markers are still interesting nowadays, since they could greatly help in the identification of wild germplasm that could provide valuable genetic diversity to breeding programmes (Galletta and Maas 1990; Hancock and Luby 1993; Marta et al. 2004; Davis et al. 2006). And naturally, modern breeding techniques based on markers could not be applied without reasonably large genetic maps and marker sets (Folta and Davis 2007).

Obtaining DNA markers in an allopolyploid is no easy task. Allopolyploids contain multiple diploid subgenomes, so for each chromosome there is one homolog and *n* pairs of homeologs. In the case of allo-ocotoploid strawberry, there are 7 chromosome pairs in each subgenome, thus each of the seven chromosome pairs has 3 other homoeologous pairs (Fig. 1B). This is particularly problematic when designing markers, since any probe must characterize within-subgenome variation without including noise from between-subgenome variation, a feat that can only be achieved when there is sufficient subgenome differentiation (Bassil et al. 2015; Edger et al. 2018; Cheng et al. 2018). Consequently, development of genetic maps was greatly delayed in strawberry in comparison to other crops (Folta and Davis 2007). Once maps started to be generated, the lack of clear methods to distinguish linkage groups added to the problem: each map published its own chromosome naming system without a clear way to integrate genetic maps. Subgenome similarity also complicated genome assembly, which explains why the diploid *F. vesca* genome was produced almost a decade before the allo-octoploid genome of *F.* x *ananassa* (Shulaev et al. 2010; Edger et al. 2019; Hardigan et al. 2021a). The octoploid sequence helped resolve the naming ambiguity of genetic maps, finally providing a translation table between them and a naming convention from A to D (Hardigan et al. 2021b). Although this problem is resolved, the fact that it existed in the first place highlights the complexity of subgenome genetics. With the rise in popularity of sequence-based genotyping methods, one must wonder if the complexities of allopolyploid genomes will not become a major roadblock. Read mapping in particular seems likely to be problematic, since reads might easily map across subgenomes, potentially adding great sources of noise and error to genotyping methods (Fig. 1C). It would be a shame if such were the case, as these approaches enable the discovery of millions of markers, a veritable cornucopia of abundance for a crop in which marker development has been so hindered.

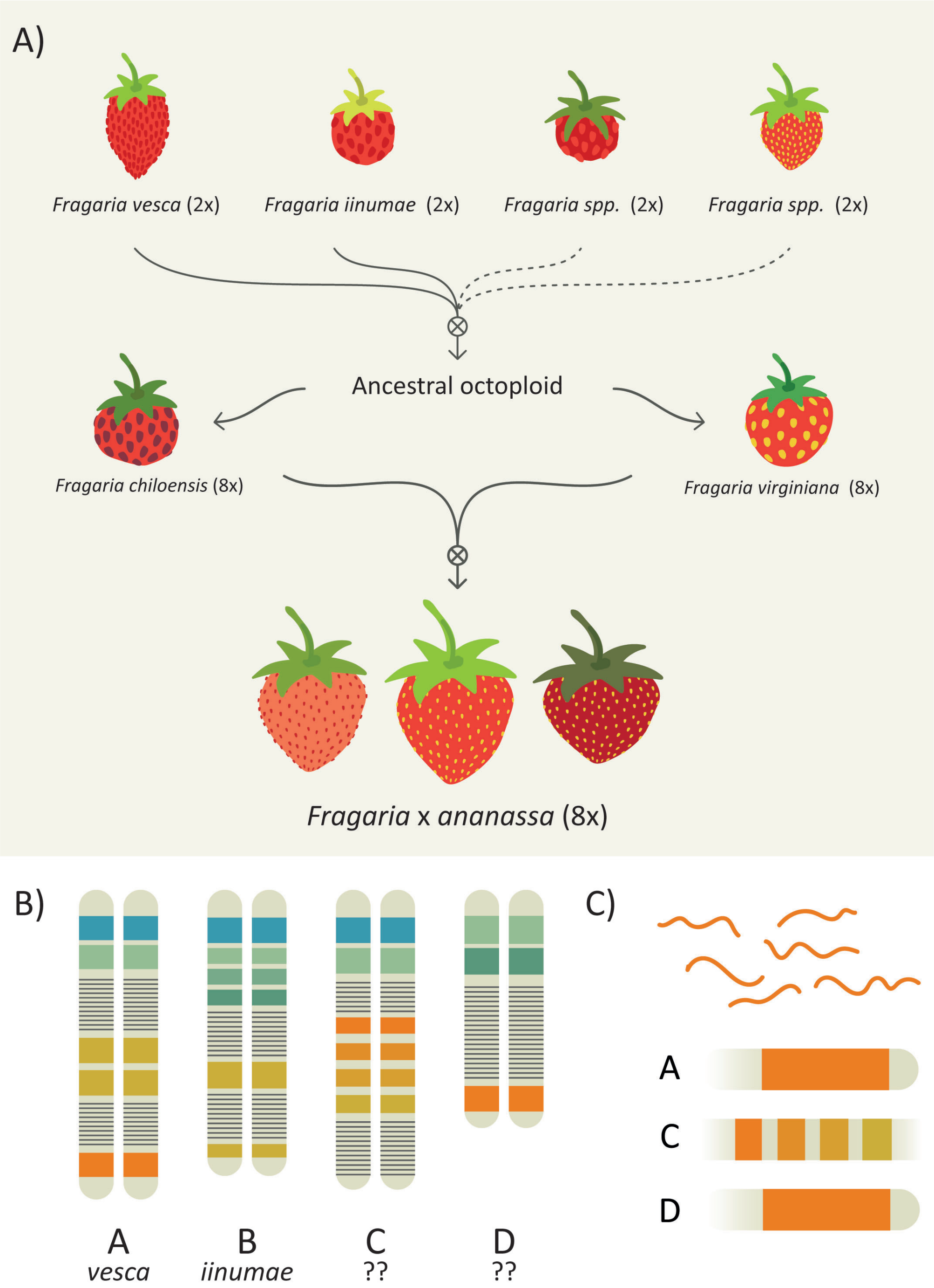
Breeding and research on strawberry aroma

As a horticultural crop, many breeding-related traits have received great attention in strawberry research. Adaptation to climate, flowering time, modes of reproduction and susceptibility to diseases have all been crucial to strawberry breeding (Darrow 1966; Galletta and Maas 1990; Folta and Davis 2007; van Dijk et al. 2014; Dominique et al. 2018; Anciro et al. 2018; Labadie et al. 2020; Castillejo et al. 2020; Tapia et al. 2021). However, flavour is a long-recognised key trait. While sugar, water content and acidity are key components of flavour, *aroma* is what ultimately makes strawberry unique (Ulrich and Olbricht 2016; Yan et al. 2018; Fan et al. 2021).

Aroma is determined by the composition and abundance of volatile organic compounds (VOC). These low-weight molecules are characterized by their odorous activity and wide range of functions. Plant VOCs have a great impact beyond their hedonic value, including as regulators of plant-plant and plant-insect interactions (de Boer et al. 2004; Bruce et al. 2005; Clavijo McCormick et al. 2012; Effah et al. 2019). However, in strawberries the major

Figure 1: Garden strawberry evolution and genome structure.

A) The commercial garden strawberry Fragaria x ananassa is the result of hybridization between two octoploid North American Fragaria species. Their common allo-octoploid ancestor was likely the result of a cross between diploids F. vesca and F. iinumae. Whether two other diploid species contributed to this polyploidization, and in which order these hybridizations took place is still under debate. B) The genome of F. x ananassa is composed of four subgenomes, A to D. The ancestors of A and B are clearly established, while C and D remain under debate. C) Schematic representation of read mapping on an allopolyploid. Due to high similarity between the orange regions in subgenomes A, C and D, mapping orange reads is an ambiguous process.


A)
Fragaria vesca (2x)
Fragaria iinumae (2x)
Fragaria spp. (2x)
Fragaria spp. (2x)
Ancestral octoploid
Fragaria chiloensis (8x)
Fragaria virginiana (8x)
Fragaria x ananassa (8x)
B)
A
vesca
B
iinumae
C
??
D
??
C)
A
C
D

VOC compounds are those expressed in fruits as part of the ripening process. There are a few major classes of compounds relevant to strawberry: esters, terpenes, lactones, aldehydes, benzenes and sulphur-related compounds are some of them (Ulrich et al. 2018; Yan et al. 2018). Strawberries are quite unique among fruits due to their unexpectedly high variability of aroma profiles. A review of 2018 found that during more than 30 years of research, over 900 different compounds have been reported in *Fragaria spp.* and over 300 different ones in *F.* x *ananassa* (Ulrich et al. 2018). They also highlight the seeming lack of overlap in the literature on identified compounds, with only a few being reported more than once. Clearly, experimental differences between studies are partly to blame for this discrepancy, but it is more than likely that true biological diversity contributes to this lack of consensus. Such can be seen by several studies on biparental populations that highlight the absence of certain compounds in part of the germplasm (Zorrilla-Fontanesi et al. 2012; Barbey et al. 2021). Despite a few successes, it is evident that very little is understood of volatile synthesis in strawberry. After years of experimental research, only a few relatively recent studies have uncovered functional genes (Aharoni et al. 2000, 2004; Zorrilla-Fontanesi et al. 2012; Oh et al. 2021; Barbey et al. 2021; Rey-Serra et al. 2022). However, no overarching paradigm has been sketched that allows to effectively control or predict the desired aroma when breeding strawberries. Moreover, there is a lack of integrated results across studies, an issue that the recent genome assembly could likely help to address.

Focusing on single aroma compounds and their causal genes is in stark contrast with the multivariate nature of aroma. Smell and taste perception is clearly the combined result of *all compounds*, not single ones (Jetti et al. 2007). Similarly, metabolites are not produced alone, but as parts of large metabolic networks. Therefore, it seems desirable to study aroma genetics through the lens of overarching regulators of aroma production, rather than individual metabolites. To do this, information across compounds should be somehow integrated. Moreover, studying a broader set of samples could help more accurately dissect the chemical diversity of strawberry aroma.

While many questions remain regarding the physiology, evolution and origin of strawberry, some answers become clear that highlight the complex history of this crop and suggest future directions (Whitaker et al. 2020). This thesis will hopefully contribute to the growing wealth of strawberry research, adding new tools and approaches with which to study this complex plant.

This thesis

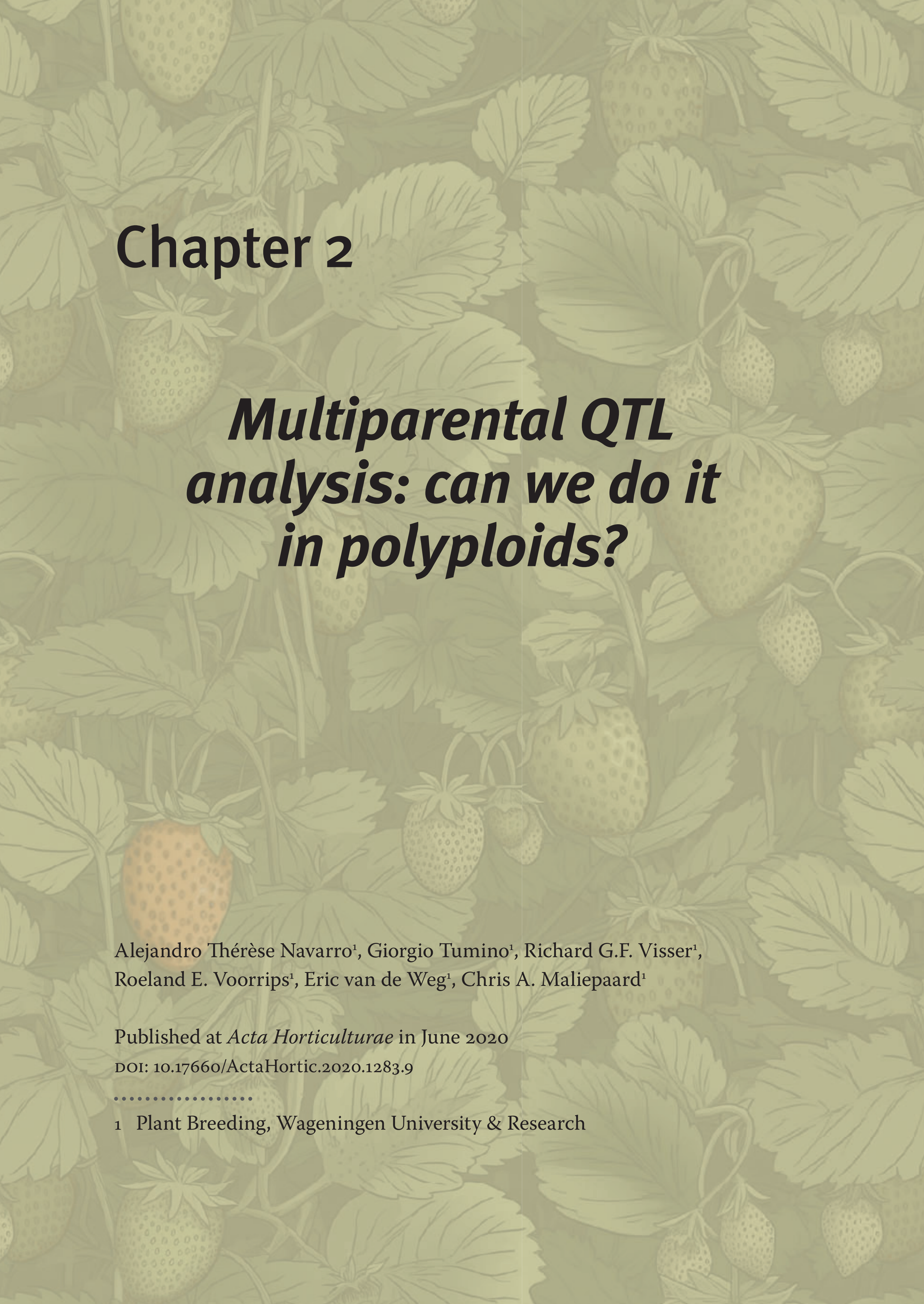
After this introduction, five chapters are contained in this thesis, followed by a general discussion. Their overarching theme revolves around methodological developments related to novel technologies and experimental settings. However, the biological nature of this research should not be overlooked. All models and methodologies here described mean little if they do not accurately represent a natural reality. In this sense, I have focused on polyploid organisms, which are usually underserved in the methodological field and more particularly on allo-octoploid strawberry.

Chapter 2 and Chapter 3 deal with the topic of multiparental populations in QTL mapping of polyploids. As outlined above, multiallelic markers are likely to be particularly useful in multiparental populations since they can closely track IBD. At the time of writing these chapters, however, no model nor tool was available that could readily perform QTL mapping in multiparental polyploid populations. It was also unclear whether such models would be actually better. While theoretically multiallelic models are more powerful, in practice they require many more parameters than biallelic models, and thus if the accuracy improvement were too small, they would not be more statistically powerful than their biallelic counterparts. In Chapter 2 I review this issue. I provide some ideas on how to model QTLs in these types of populations and point towards tools that allow to address the issue of haplotyping in polyploid organisms. In Chapter 3 I go a step further and propose my own model, implemented in the tool mpQTL. Using a complex set of simulations that account for different levels of genetic diversity and genetic architectures underlying the simulated traits, I was able to clearly test if there was any statistical improvement using such models.

The following research was focused on genetic linkage mapping. As discussed, genetic maps are an old but very useful technique to study chromosomes and genomes, yet modern DNA marker technologies challenge their classical algorithms. That is more so the case with skim-sequencing approaches, those in which depth is purposefully lowered, which have even a larger error rate in polyploid crops. Applying this technique to allopolyploids seemed even more challenging. With reduced genotyping accuracy it seemed plausible that subgenome differentiation could not be achieved. Testing this method would require an application that would enable assessing the genotyping quality. Genetic mapping came to mind. If skim-sequencing genotyping was possible, it would allow to produce maps with much larger numbers of markers at a similar cost than modern linkage maps. With high hopes, I aimed to answer this question using a dataset from strawberry population, and while doing so I updated an algorithm named SMOOTH (van Os et al. 2005). The results are described in Chapter 4 and Chapter 5. In the first, I describe Smooth Descent, the upgraded SMOOTH that implements IBD ideas in order to correct genotyping errors and improve genetic maps (in diploids and polyploids). In Chapter 5 I show the results of applying Smooth Descent to a skim-sequencing population of strawberry and compare the resulting maps with maps produced using SNP array genotypes, with surprisingly positive results.

Lastly, in Chapter 6 I return to association mapping with a multivariate QTL study of strawberry volatile organic compounds. While previously I worked on the issues caused by genetic structure and large, low-quality genetic datasets, now I focused on large *phenotypic* datasets. As hinted above, metabolomics studies often struggle to integrate the results of all metabolites at once, an issue echoed in imaging datasets and other large phenotypic data. To address this problem, I opted to perform a multivariate QTL study, an idea that is far from new but is rarely used and for which there are no current standards. A multivariate QTL study aims to understand the relationship between all phenotype variables (metabolites in our case) as well as QTLs for those relationships. As a result, one can study groups of metabolites instead of individual ones, finding QTLs not only for the metabolites, but also for the overall pathway. While the theory is sound, the application of this approach was somewhat unclear. In Chapter 6 I test two possible approaches, one more readily obvious from metabolomics literature, and another less common but more statistically sound.

Together, these chapters form the thesis that you are about to read. They also reflect a current trend in the quantitative genetics field, where research is aimed at understanding challenging datasets with the appropriate tools. Often, that has required adapting existing tools to new contexts, updating and expanding already existing models. The computational aspect of this work cannot be overstated, most of my work has meant programming these models into usable and shareable tools and creating reproducible research. Although I studied plants, the experiments of this thesis have mostly been carried out on the silica boards of our department's supercomputer.



Chapter 2

Multiparental QTL analysis: can we do it in polyploids?

Alejandro Thérèse Navarro[1], Giorgio Tumino[1], Richard G.F. Visser[1], Roeland E. Voorrips[1], Eric van de Weg[1], Chris A. Maliepaard[1]

Published at *Acta Horticulturae* in June 2020
DOI: 10.17660/ActaHortic.2020.1283.9

[1] Plant Breeding, Wageningen University & Research

Abstract

Many ornamental crops are polyploid or even exist at different ploidy levels. Polyploid QTL analysis tools have been developed in recent years, yet they are limited in the population types they accept. Biparental populations are nowadays being regarded as a limited tool for QTL discovery, as only a limited number of QTLs occurs in an experimental cross and their effects might not be stable across genetic backgrounds. Genome-Wide Association Studies include more genetic diversity but suffer from (hidden) genetic structure and low frequency of QTL alleles. Both factors influence QTL detection and effect estimation, decreasing the sensitivity of QTL analysis. Alternatively, multiparental populations (MPP) can be used, potentially combining multiple QTLs and QTL alleles with known population structure and balanced allele frequencies. Breeding populations of interconnected crosses also constitute a form of MPP and QTLs identified in them might be more applicable to commercial cultivars. To perform QTL analysis in polyploids, mixed models or Bayesian approaches that consider pedigree information are recommended. During the analysis, QTL effects are ideally estimated using IBD information, which can be obtained through haplotype estimation. Although MPPs could thus be a powerful set-up to estimate polyploid haplotypes, a software gap was identified as no current polyploid haplotyping tools are able to utilize MPP pedigree information to obtain haplotypes across an MPP. In order to utilize MPPs to their full extent and expand polyploid QTL analyses to encompass typical breeding populations, new haplotyping tools must be developed.

Keywords

Breeding populations, IBD, GWAS, family-based analysis, pedigree-based analysis.

Multiparental QTL analysis: can we do it in polyploids?

Introduction

Polyploidy, the multiplicity of genome copies within a cell, is an important evolutionary phenomenon that has played a crucial role in plant evolution (Comai 2005; Soltis and Soltis 2012). This genetic condition has also been utilized in breeding, particularly in the ornamental field, due to its direct effects on organ size and morphology, and its ability to restore fertility in interspecific hybrids. In fact, in the recent book *Ornamental Crops* (van Huylenbroeck 2018), in which molecular breeding techniques in ornamentals are reviewed, virtually all crops mentioned deal with polyploidy in one or more of these cases: i) in natural polyploid or mixed ploidy populations, ii) in cultivars that had been unconsciously selected for polyploidy, iii) in plants with induced polyploidy to alter morphology or bridge interspecific fertility barriers. Given the interest of moving from classical to molecular breeding approaches, it is essential to develop and expand methodologies that allow breeders and researchers to analyse organisms that differ from the diploid standard.

One of those techniques is Quantitative Trait Loci (QTL) mapping. The term QTL arose almost accidentally in a mathematical article by (Geldermann 1975), in which he described a marker-based method to associate variation in a *quantitative trait* with genetic *loci* (in a population of segregating individuals). Although Geldermann did not pioneer the idea, his acronym was rapidly adopted and has nowadays become an essential tool in breeding and research.

Polyploid QTL models were proposed early on (Kempthorne 1957), but their application has lagged in comparison to diploids until genotyping technologies and computational resources were good enough to handle polyploid genetic complexity (Doerge and Craig 2000; Xie and Xu 2000).

Interpretation of a QTL analysis and its results depends directly on the population type, the genotyping platform and the statistical method used to detect QTLs. For instance, a QTL found in an F2 biparental population identifies those genomic regions *for which the parents are polymorphic* and whose variation in the F2 can be associated with phenotypic variation, suggesting a link between the genes in that area and the trait in question. Although the usefulness of this method has allowed a great variety of functional genes to be uncovered, the limitations of this approach are well known: QTLs detected in a biparental population might not be functional in other genetic backgrounds and not all causative loci can be detected due to the limited genetic diversity of the population's parents.

Alternatively, Genome-Wide Association Studies (GWAS) can be used, where a group of genetically diverse individuals (generally with unknown relationship between them) are used. In these populations, linkage between marker and QTL allele is due to evolutionary Linkage Disequilibrium (LD) rather than the LD caused by recent relatedness in biparental populations. Despite their great usefulness, GWAS designs remain controversial (Tam et al. 2019). Although they allow to study a wider range of genetic diversity, effects from rare or weak QTL alleles cannot be adequately estimated, and thus are generally missed. Additionally, population structure and allelic diversity act as confounding factors that must be taken into account in order to avoid statistical artifacts (Yu et al. 2006; Korte and Farlow 2013).

Nevertheless, both approaches represent the two extremes of a gradient of population diversity (Würschum 2012). Alternatively, multiple biparental populations that share parents can be used. We will refer to them as multiparental populations (MPP), although in literature they are also known as pedigreed populations, connected populations or families. MPPs harbour a higher level of genetic diversity than a single biparental population, allow testing parental genes in multiple genetic backgrounds, and can be expected to have a more balanced genetic structure compared to GWAS panels composed of a random sample of diverse individuals. Moreover, MPPs are particularly suited for the type of exploratory crossing that is common in breeding efforts: a few interesting parents are selected, intercrossed and a small population is raised from each cross. While traditionally each cross is analysed separately, the MPP approach proposes joining all crosses in a single analysis.

In this article, we consider the existing MPP populations in plant breeding and discuss the statistical implications of using MPPs in QTL analysis, with special attention to the analytical issues that arise with polyploid genetic analysis.

Multiparental population types

Experimental Populations

In plants, experimental MPPs have been developed for a long time. Diallel crosses, populations where a set of parents are crossed in all possible combinations (full diallel) or omitting reciprocal crosses (half diallel), were and are still used in breeding since the definition of general and specific combining abilities were laid down by (Sprague and Tatum 1942). However, they were not developed for QTL analysis, but as a form of evaluating parental contributions to hybrids, obtaining an evaluation of the quality of a parent as a source for breeding (Griffing 1956).

More recently, complex MPP schemes have been developed. Multi-parent Advanced Generation Intercross (MAGIC) populations (Cavanagh et al. 2008) have already been developed in a variety of crops, both diploid (e.g. maize, rice, tomato) and (allo)polyploid (e.g. wheat, peanut) (Huang et al. 2015). The principle of MAGIC is to combine alleles from different founders in a single genome, and thus evaluate each QTL allele in a variety of different genetic backgrounds. Another MPP scheme is termed Nested Association Mapping (NAM), in which a central parent is crossed with a set of peripheral parents, followed by a series of back-crosses and selfings. This population design has been adopted in fewer crops, but nevertheless with great success. The maize NAM population (McMullen et al. 2009) is undoubtedly its most famous example, but other NAMs have also been developed in sorghum (Bouchet et al. 2017), soybean (Song et al. 2017), barley (Maurer et al. 2015) and a single polyploid: wheat (Bajgain et al. 2016).

Breeding populations

Although the experimental MPP schemes mentioned above are useful innovations, they represent only a small fraction of existing MPPs. In breeding programmes, it is common practice to select phenotypically interesting parents and cross them together or with other cultivars, in order to explore new trait combinations. Thus, each parent contributes to many biparental populations, and these sets of connected crosses can be regarded as an MPP.

Breeding populations are regularly screened for interesting phenotypes, and genotyping these populations is becoming an increasingly common practice. Thus, both genotypic and phenotypic data are likely to already be available for these *ad-hoc* MPPs. Additionally, it has been suggested that utilizing such populations for QTL detection provides results that are more readily translatable into breeding application due to their direct connection to the final cultivars (Jansen et al. 2003; Verhoeven et al. 2006; Würschum 2012; Bink et al. 2012; Bardol et al. 2013; Han et al. 2016).

For these reasons, development of analytical methods that allow QTL analysis in polyploid MPPs is especially relevant for breeding efforts, as it will allow breeders to describe and explore more deeply the breeding material present in a program.

Statistical considerations

Genetic Structure

QTL analyses rely on association between marker and QTL alleles (also known as allelic disequilibrium, linkage disequilibrium, gametic phase disequilibrium or gametic disequilibrium), which occurs when alleles of two loci are found together more often than by chance (Flint-Garcia et al. 2003). Physical linkage is not the only possible cause of linkage disequilibrium (LD): bottlenecks, genetic drift, natural selection, domestication, breeding history or recent relatedness can generate long-range LD across a population, even on different chromosomes, a property generally called *population structure* (Yu et al. 2006).

The presence of population structure is one of the major differences between biparental populations and MPPs. In biparental populations all individuals

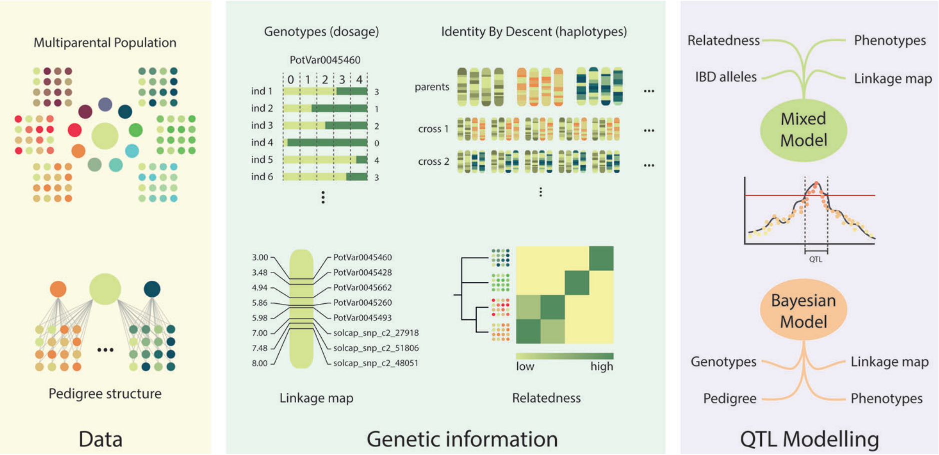


Figure 1: Required elements for polyploid multiparental population (MPP) analysis.

Left panel: a multiparental population is obtained. Its pedigree structure can be used in later analysis. Middle panel: raw genetic data can be used to estimate: i) genotypes (dosages) to determine heterozygote classes, ii) haplotypes, to determine which parents and offspring are identical by descent along chromosomes, iii) linkage map, to determine marker distance and order, and iv) genetic structure, as a relatedness matrix, from pedigree and or genotype information. Right panel: phenotypes and linkage maps are used to predict QTL positions. With mixed models, genetic structure is accounted for by the relatedness matrix and IBD alleles, while in Bayesian models that is achieved by incorporating pedigree information in the QTL models.

are “equally” related (i.e. that any two individuals have the same probability of having a level of relatedness *k*), but in MPPs, those individuals that originate from the same cross (full siblings) are more highly related than those that originate from different crosses (half-siblings or unrelated). Importantly, unlike in GWAS panels, these patterns of relatedness are predictable and can be incorporated in the MPP design process in order to obtain a balanced genetic structure.

A direct consequence of genetic structure is a non-random distribution of alleles across the population, which may lead to an increased rate of type I errors when genetic structure correlates to the phenotype. To resolve this issue, genetic structure must be considered. Two statistical frameworks have been used so far for this purpose: mixed models, where a relatedness matrix is used (Yu et al. 2006); and Bayesian QTL models in which pedigree information is included in their likelihoods (Bink et al. 2014). Strictly speaking, only mixed models have been applied using a polyploid model (Ferrão et al. 2018) through the R package GWASpoly (Rosyara et al. 2016). However, the diploid models of FlexQTL™ (Bink et al. 2014) have been used in allo-octoploid strawberry by treating each subgenome separately (Mangandi et al. 2017; Verma et al. 2017; Anciro et al. 2018). Expansion of this Bayesian framework to polyploidy would thus be an interesting development for autopolyploids and allopolyploids for which genotyping tools cannot obtain subgenome-specific genotypes.

Modelling QTLs in multiparental populations

QTL detection requires estimation of QTL effects. The number and properties of these effects are determined by the type of situation we expect to encounter. In general, we can summarize MPP QTL modelling in four categories (Han et al. 2016):

1) Each effect is specific for each cross and parent genotype (Jannink and Jansen 2001; Blanc et al. 2006). It is assumed that there is no shared information between crosses. We know that protein and gene functions might vary depending on the context in which they are expressed (i.e. genetic background, environ-mental factors), but completely disregarding shared information between crosses is an over-conservative approach that ignores the potential to increase statistical power by using MPPs.

2) Parental alleles are unique and effects are shared between crosses (Jannink and Jansen 2001; Blanc et al. 2006; Garin et al. 2017). QTLs are estimated across families, but we assume that each parent contributes different alleles to the populations. While that might be realistic when all parents originate from completely divergent genepools, it does not reflect most breeding populations, where parents have some degree of relatedness and thus might share alleles at certain loci.

3) Identity-by-descent (IBD) segments among parents harbour the same alleles, with identical effects between crosses (Jansen et al. 2003; Bardol et al. 2013; Garin et al. 2017). Two alleles are said to be identical by descent if they have originated from a common ancestor. This method requires the identification of parental alleles that are IBD and estimation of the effects for each unique allele. If IBD can be estimated (see below) and QTL effects are stable between crosses, this is the most powerful method in MPPs (Jansen et al. 2003).

4) Effects are estimated per marker and identical between crosses (Würschum 2012; Garin et al. 2017). In this case it is assumed that a marker allele indicates QTL allele, i.e. that *each* QTL allele is in linkage with a different marker allele.

Since we expect multiple QTL alleles, this would require a multiallelic marker system densely spread across the genome. These requirements are not yet met by any modern marker system: high-density SNPs are mostly biallelic, and multiallelic markers lack the high-density (and cost effectiveness) of SNPs. Whole genome sequencing could meet these criteria, but fully reconstructing all chromosomes (haplotyping of the whole genome) is still impossible in polyploids (see below).

Consequences of polyploidy in MPP analysis

Polyploid and diploid MPPs do not differ qualitatively. Although some population parameters such as genetic drift, heterozygosity and allele frequencies are different in polyploids, these do not impact significantly the design or properties of MPPs. Complications arise on the analytical side during genotyping, linkage mapping and haplotyping.

The polyploid genotyping problem

Unlike diploids, polyploids present multiple heterozygote classes which must be distinguished. When two alleles (e.g. A and C) are detected, they must be quantified in order to estimate their *dosage* (ACCC, AACC or AAAC). Various tools have been developed for this using both fluorescence intensities of SNP arrays (Voorrips et al. 2011; Serang et al. 2012; Schmitz Carley et al. 2017; Zych et al. 2019) and read data from genotyping by sequencing (Gerard et al. 2018).These tools have helped not only in genotyping polyploids, but also in understanding the types of uncertainties that arise with each measurement technique (e.g. background fluorescence or allele bias in SNP arrays, sequencing error or overdispersion in sequence counts of GBS). In order to improve genotype estimation, allele frequency expectations in a population are usually included. There are two common frequency assumptions: a biparental F1 population, where frequencies depend directly on parental genotypes; and Hardy-Weinberg equilibrium, which is useful in randomly sampled natural populations (Voorrips et al. 2011; Gerard et al. 2018). To accommodate MPPs, frequency expectations must be adapted to reflect the structure of an MPP, i.e. to model multiple F1s. Such work has already been performed for fitTetra (Zych et al. 2019) and there exist programs initially developed with MPPs in mind (Schmitz Carley et al. 2017). These are positive improvements, yet more complex MPPs (e.g. combining pedigreed individuals and F1s) might still require further developments.

The polyploid mapping problem

Linkage mapping is an important tool as it allows to characterize the recombination behaviour along the chromosomes of a species. More importantly, in order to detect QTLs, marker order must be known. Allopolyploids, that segregate as diploids, can use diploid mapping software to obtain a linkage map. However, autopolyploids, that generally segregate polysomically require dedicated models. Few programs are available for mapping in autopolyploids (Hackett et al. 2014; Bourke et al. 2018a), the most flexible being polymapR (Bourke et al. 2018a), as it can estimate integrated maps under different ploidies and segregation models (bivalent, preferential or multivalent pairing of chromosomes). Nevertheless, the package can only estimate maps in F1 populations. In order to generate a map for an MPP either multiple F1 maps should be generated and integrated, or polymapR should be adapted to accept other population structures. In any case, current methodologies do not allow to generate linkage maps for (auto)polyploid MPPs.

The polyploid haplotyping problem

More specifically relevant to MPPs is **estimation of IBD.** One can speak of two “kinds” of IBD: on the one hand family-IBD, e.g. regions of chromosomes from two offspring individuals that originate from the same *parental chromosome*; and on the other, ancestral-IBD, those chromosomal regions that originate from the same *common ancestor* and that are broken down by recombination events (Browning and Browning 2011a). The latter might even span across closely related species.

Generally, IBD is estimated using haplotypes (Meuwissen et al. 2001), the concatenation of adjacent polymorphisms, most commonly SNPs. Finding the haplotypes is simple when the number of possible combinations is low, (with high homozygosity and low ploidy). In heterozygous polyploids the issue of finding the underlying haplotypes becomes increasingly complex due to the great number of possible haplotypes. Combining polymorphisms to form haplotypes is a process known as phasing or haplotyping and can be performed in different ways depending on the type of data one uses.

Firstly, genotypes can be obtained as independent polymorphism scores (e.g. from a SNP array), or as sequence reads, where each read might contain information about multiple polymorphic sites, thus providing short-range SNP phase information. Secondly, we might wish to resolve the haplotypes of a single individual, or of a group of individuals (related or unrelated).

Haplotypes of a single individual can currently only be resolved using sequence reads, as independent SNP scores do not allow us to decide between the multiple haplotype possibilities. There already exist multiple tools that can phase polyploid haplotypes using next-generation sequencing (NGS) reads (Aguiar and Istrail 2012; Berger et al. 2014; Das and Vikalo 2015; He et al. 2018), and for short reads these will provide accurate results if sequencing is performed at adequate depth (Motazedi et al. 2017). In the future, haplotype library methods based on inferring the most likely haplotype given a set of previously identified haplotypes (Pook et al. 2019), might provide haplotyping solutions for single individuals if SNP-arrays are used. These methods, however, are still in development even for diploids and thus might take some time to reach polyploids.

Phasing using populations has, comparatively, received less attention. (Browning and Browning 2011b) reviewed existing methods for diploids, and divided them in two main groups: i) phasing methods for unrelated individuals, which use either coalescent theory to haplotype likelihood via Hidden Markov Models (Scheet and Stephens 2006; Li et al. 2010; Browning and Browning 2011a) or a parsimony principle (Neigenfind et al. 2008); and ii) phasing models for related individuals, in which pedigree information and Mendelian constraints allow to determine likely haplotypes (Abecasis et al. 2002). Since 2011, other population haplotyping tools have been released for polyploids for independent SNPs in F1s (Zheng et al. 2016), for long reads in pedigreed individuals (Garg et al. 2016) and for short reads in parent-offspring trios (Motazedi et al. 2018).

None of the above-mentioned methods can currently exploit information across MPPs. Thus, there is no available methodology that is able to transform unphased SNP genotypes in polyploid MPPs into the multiallelic markers that are required to apply the modelling strategy 3 described above. This gap does not allow to fully utilize multiparental population QTL detection methods in polyploids and represents a lag of polyploid methodology with respect to diploids.

Conclusion

Multiparental populations (MPPs) are an interesting prospect that could allow to identify and utilize QTLs with more relevance for breeding applications. In that regard, we must consider also MPPs beyond experimental populations and realize that the breeding populations of interconnected crosses that are regularly generated as a form of exploratory cultivar evaluation also constitute useful MPPs.

QTL modelling of MPPs is more challenging than in biparental crosses due to genetic structure and higher allelic diversity, but mixed models have shown their usefulness in analysing polyploid MPPs and Bayesian models, if adapted to polyploid organisms, could also prove a useful tool. Regarding the models of QTL effects, IBD-based (haplotype) estimates are the most theoretically consistent method to perform QTL analysis in MPPs, as they account both for family-based linkage and possible sharing of alleles between parents. However, estimation of IBD in polyploids is a challenging task and no method has yet been developed that can adequately obtain haplotypes across multiparental populations fully capitalizing on the shared information between crosses, either from sequencing reads or from unphased SNPs.

Polyploid genetic tools are usually developed as extensions from less general diploid models. Similarly, MPP polyploid tools must be developed as generalizations of methodologies that were developed for application in different population types. To that end it will be useful to look both at polyploid tools for biparental and GWAS populations and at diploid tools for MPP analysis, harvesting developments from both fields.

Declarations

Author's contributions

Literature research and manuscript writing was performed by ATN. Main points were outlined by GT, REV, EvdW and CAM. Revision was performed by all co-authors. CAM obtained funding.

Acknowledgements

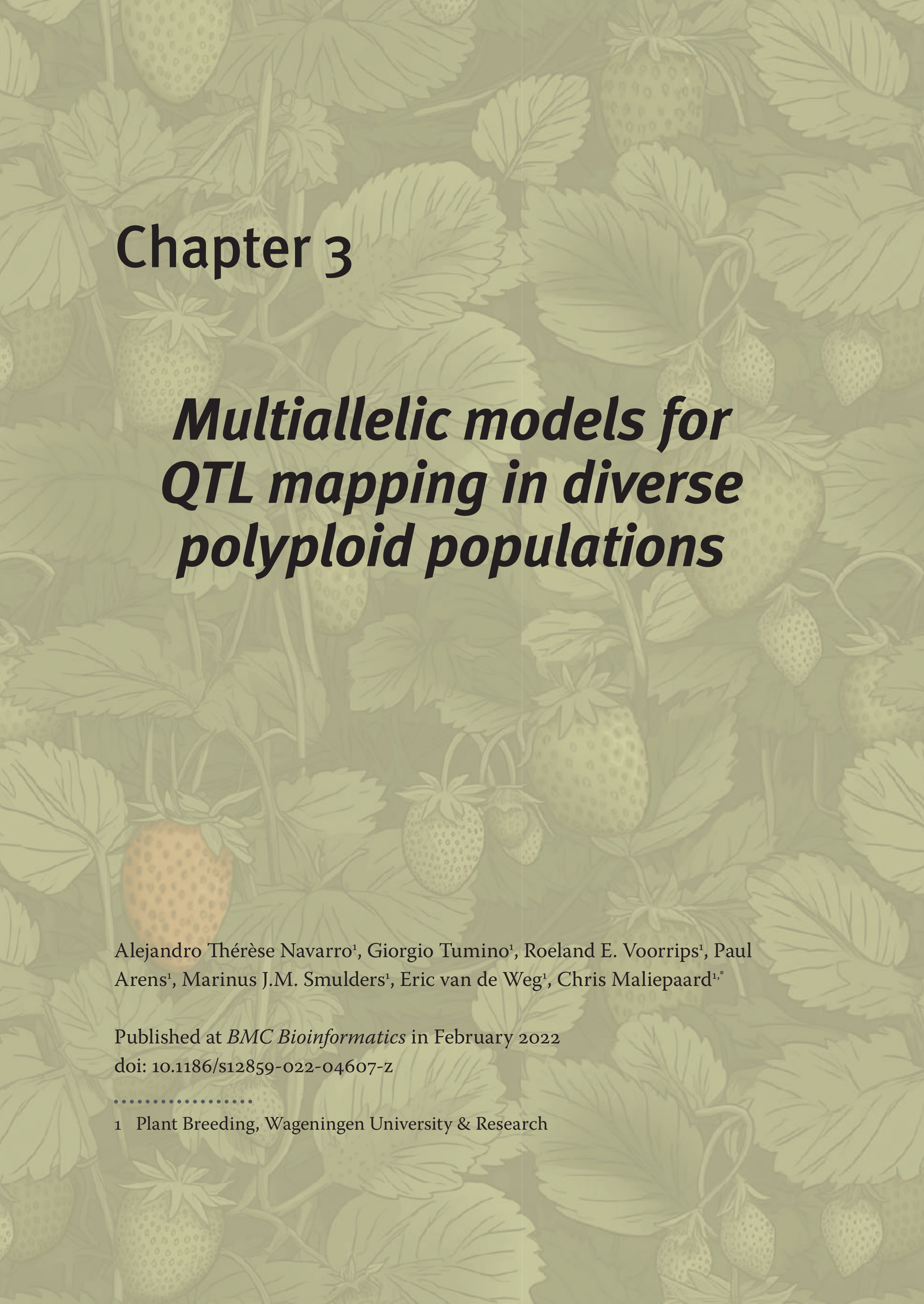
The author would like to personally acknowledge the contributions of Eric van de Weg and Roeland Voorrips, with whom fruitful discussions on this topic allowed to define how to perform analysis in multiparental populations, and Paul Arens, Peter Bourke and René Smulders for their useful guidance on writing this manuscript.

Competing interests

The authors have no conflicts of interest regarding this article.

Funding

This paper was written as part of the Polyploids project "Novel genetic and genomic tools for polyploid crops" (project numbers BO-26.03-009-004 and BO-50-002-022).



Chapter 3

Multiallelic models for QTL mapping in diverse polyploid populations

Alejandro Thérèse Navarro[1], Giorgio Tumino[1], Roeland E. Voorrips[1], Paul Arens[1], Marinus J.M. Smulders[1], Eric van de Weg[1], Chris Maliepaard[1,*]

Published at *BMC Bioinformatics* in February 2022
doi: 10.1186/s12859-022-04607-z

1 Plant Breeding, Wageningen University & Research

Abstract

Quantitative Trait Locus (QTL) analysis allows to identify regions responsible for a trait and to associate alleles with their effect on phenotypes. When using biallelic markers to find these QTL regions, two alleles per QTL are modelled. This assumption might be close to reality in specific biparental crosses but is unrealistic in situations where broader genetic diversity is studied. Diversity panels used in genome-wide association studies or multi-parental populations can easily harbour multiple QTL alleles at each locus, more so in the case of polyploids that carry more than two alleles per individual. In such situations a multiallelic model would be closer to reality, allowing for different genetic effects for each potential allele in the population. To obtain such multiallelic markers we propose the usage of haplotypes, concatenations of nearby SNPs. We developed "mpQTL" an R package that can perform a QTL analysis at any ploidy level under biallelic and multiallelic models, depending on the marker type given. We tested the effect of genetic diversity on the power and accuracy difference between bi-allelic and multiallelic models using a set of simulated multiparental autotetraploid, outbreeding populations. Multiallelic models had higher detection power and were more precise than biallelic, SNP-based models, particularly when genetic diversity was higher. This confirms that moving to multi-allelic QTL models can lead to improved detection and characterization of QTLs.

Keywords

Polyploidy, multiparental, QTL, multiallelic, genetic diversity

Multiallelic models for QTL mapping in diverse polyploid populations

Introduction

Quantitative Trait Locus (QTL) analyses are those experiments in which a population is genotyped with many markers that cover the whole genome, and phenotyped for traits of interest. Once that is done, positions along the genome are tested for association, either defined by the markers or by some clever estimate such as those used in interval mapping (Lander and Botstein 1989; Akond et al. 2019). QTL studies have been extremely useful in unravelling genomic regions that control or contribute to important plant traits such as disease resistance, yield, crop quality or tolerance to abiotic stresses. The precision of these studies has been improved by the advent of high-throughput technologies, that facilitated genotyping of thousands to millions of Single Nucleotide Polymorphisms (SNPs) in a single analysis. This is nowadays also possible in polyploid organisms, thanks to statistical and computational developments in the areas of genotyping, linkage map construction and QTL analysis (Rosyara et al. 2016; Bourke et al. 2018b).

When trying to find QTLs two aspects will define the outcome obtained: the type of population studied, and the QTL modelling approach chosen.

Population types

A classical population type is the biparental cross, a population of siblings obtained by crossing two parents, usually contrasting in the trait of interest.

If both parents are homozygous, as is the case in many self-fertilizing species, QTLs found in this type of population will reflect the allelic differences between the two parents. If the parents are diploids, there will likely be only two alleles per QTL segregating in that population. Since the cross contains only a small fraction of the genetic diversity of the species, QTL results from these populations may not be applicable to other populations and markers linked to QTLs cannot easily be used in other crosses.

Another possibility is to use a genome-wide association study (GWAS), in which a large set of diverse individuals are studied, and thus a large number of QTL alleles is expected to segregate. Unlike in biparental crosses, an association between markers and QTLs is expected due to Linkage Disequilibrium (LD) rather than direct family linkage. These studies produce more widely applicable QTL results, but introduce some drawbacks: i) rare allele variants, which will be present at low frequency in a GWAS panel, will easily be missed even if they affect the phenotype, and ii) linkage disequilibrium (LD) is not spread homogeneously across the population or the genome, an effect known as "genetic structure", and this may generate false positives if not taken into account (Yu et al. 2006; Korte and Farlow 2013).

Nevertheless, as described in (Würschum 2012), mapping in biparental populations or GWAS panels represent two extremes of a genetic diversity gradient. An intermediate form can be found in multi-parental populations (MPP). An MPP is formed by individuals that share a limited number of known ancestors, for instance, a set of connected biparental crosses, or multiple lines originating from a small set of founders. As such, the number of QTL alleles will be at most of *ploidy* × *founders*. Additionally, as the genetic structure in an MPP originates from mostly known pedigree relationships, it will be less complex than that of GWAS populations, and the allele frequencies will often be more balanced.

The MPP concept fits well the type of populations usually available in breeding programmes, where multiple crosses are made with some interesting parents. Breeding populations become then *ad-hoc* MPPs and instead of analysing each cross separately, the whole breeding program could be analysed at once, increasing statistical power. The idea that utilizing breeding populations for QTL analysis might be a better option than creating specific experimental populations has been studied previously (Jansen et al. 2003; Würschum 2012; Bardol et al. 2013; Bink et al. 2014), although in diploid species under biallelic models.

Modelling approaches

The type of mathematical model used for QTL analysis will heavily depend on the population under study. In a classical biparental population an analysis of variance (ANOVA) will easily provide accurate QTL estimates. In contrast, in a GWAS panel, genetic structure must be taken into account, usually in the form of a mixed model (Yu et al. 2006). In the case of a MPP, a similar mixed model could be used, although if the genetic structure is simple enough, a fixed factor accounting for subpopulations may perform well also (Yu et al. 2006).

The number of modelled QTL alleles is also relevant. Typically, since biallelic markers are used, two alleles per QTL are modelled. Assuming the presence of only two alleles, however, is sensible under very few scenarios. As ploidy, heterozygosity or the number of founders of a population increase, the number of expected QTL alleles rises. The larger the number of alleles, the less realistic the biallelic model becomes for describing the observed variance. Nevertheless, as SNP markers have become the standard polymorphism in modern genotyping, using them directly implicitly tests a biallelic scenario. However, SNP information can be used differently. By combining adjacent SNPs, biallelic SNPs can be turned into multiallelic haplotype markers (Leroux et al. 2014).

Due to the increased genetic diversity present in GWAS and MPP populations, it is foreseeable that moving to multiallelic QTL models will provide a gain in statistical power. Nevertheless, biallelic models are simpler and thus more powerful, and they have a long trajectory of success. There is currently no software available that can perform multiallelic QTL analyses in polyploid populations in the presence of genetic structure, but such software is being developed. Under which circumstances, if any, will a genetically diverse population benefit from a multiallelic QTL modelling approach?

To answer this question, we have simulated a series of autotetraploid MPPs with different levels of genetic diversity. Populations were designed following the Nested Association Mapping (NAM) structure, where one central parent is crossed with many peripheral parents (McMullen et al. 2009). We adapted the QTL modelling approach presented in (Garin et al. 2017) for diploid MPPs with inbred founders, expanding it to a polyploid and heterozygous case. We present this approach as an R package (R Core Team 2016) named “mpQTL” to perform QTL analysis. This package together with the simulated MPPs allowed us to assess the effect of biallelic or multiallelic markers on QTL detection and QTL precision under different genetic diversity scenarios.

Materials and Methods

Statistical Models

Mixed models allow to correct for dependence between observations due to genetic structure. Yu et al. (2006) defined a “unified mixed model”, also known as the $Q+K$ model (Rosyara et al. 2016), that can accommodate both a population structure matrix (Q) and a kinship matrix (K):

$$y = X\boldsymbol{\beta} + Q\boldsymbol{v} + \underline{Z\boldsymbol{u}} + \underline{\boldsymbol{\varepsilon}} \qquad Var(\boldsymbol{u}) = K\sigma_G^2 \quad Var(\boldsymbol{\varepsilon}) = R\sigma_\varepsilon^2 \qquad [1]$$

Where y is the vector of phenotypic trait values, $X\boldsymbol{\beta}$ represents the incidence matrix and marker effects (SNP effect in (Yu et al. 2006)); $Q\boldsymbol{v}$ are the population structure matrix and vector, respectively; $\underline{Z\boldsymbol{u}}$ are design matrix and vector of genetic background effects (polygene component in (Yu et al. 2006)); and $\underline{\boldsymbol{\varepsilon}}$ is the residuals vector. The variances of the random effects, $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are also defined: K is the kinship matrix and σ_G^2, the genetic variance; R is a matrix with off-diagonal numbers being 0 and the diagonal is the reciprocal of the number of observations underlying each genotype estimation, and σ_ε^2 is the residual variance.

Fixed term: allele parametrization

Definition of *X* requires a genetic model, that is, a method to transform genetic data into an incidence matrix *X*. Polyploid genetic models have existed for a long time (Kempthorne 1957) and have inspired more recent versions applied to SNP data (Hackett et al. 2001; Luo et al. 2005). The simplest of them is the ***biallelic model*** (model B in (Würschum 2012), association mapping in (Liu et al. 2012)), which considers SNP alleles as equal to QTL alleles. In a biallelic model, the SNP dosages are used to predict genetic effects, giving the $X\beta$ term the following form:

$$X_b\beta = \begin{bmatrix} 1 & \delta_1 \\ 1 & \delta_2 \\ \vdots & \vdots \\ 1 & \delta_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} \qquad [2]$$

Where δ_i are the dosages (a value from 0 to *ploidy*) of one of the SNP alleles, μ is the intercept and β the genetic effect of that SNP allele. We denote the incidence matrix as X_b for this modelling strategy. Note that this represents an additive model without intra or inter-locus interaction, i.e. no dominance or epistasis between alleles is modelled.

Alternatively, Identity-By-Descent (IBD) information can be used to generate an ancestral model (Garin et al. 2017), or a PBA model (Bink et al. 2014) or an LDLA model (Bardol et al. 2013; Giraud et al. 2014). Under the ancestral model, the dosage of each ancestral allele or haplotype in the NAM population is used to estimate genetic effects. The shape of the $X\beta$ term then takes the form:

$$X_a\beta = \begin{bmatrix} 1 & \delta_{11} & \delta_{12} & & \delta_{1k} \\ 1 & \delta_{21} & \delta_{22} & & \delta_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \delta_{n1} & \delta_{n2} & & \delta_{nk} \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \qquad [3]$$

In this case, the dosages of all alleles *except one* (the reference allele) are specified. Therefore, *k* is the number of alleles -1. Each β represents the additive genetic effect of each ancestral allele, relative to the effect of the reference ancestral.

Random term: kinship matrix calculation

In this model, a kinship matrix *K* is calculated using the *realized relationship* (Rosyara et al. 2016):

$$K = \frac{DD^t}{\Delta} \qquad \Delta = \overline{diag(DD^t)}$$

Where *D* is a dosage matrix with markers on columns and individuals on rows, and the mean of each column is zero (column means have been subtracted for each column); and Δ is the mean of the diagonal of the DD^t matrix. If haplotypes are used instead of biallelic SNPs, *D* can consist of concatenated matrices similar to X_a (without the intercept column), so that the number of columns is equal to the total number of alleles present across all markers used. To mitigate the bias due to differences in marker density across the genome, kinship estimates are calculated on a subset of evenly distributed SNPs (one marker per cM).

Haplotyping

Haploblocks were arbitrarily defined using a sliding window of 6 consecutive SNPs with an overlap of 4 SNPs (first haplotype is SNP1-SNP2-…-SNP6, second is SNP3-SNP4…-SNP8). A haploblock of length 6 can tag a maximum of 2^6 = 64 alleles if all combinations are present, although in our simulations the number of observed alleles was much lower, with the average number of observed alleles ranging from 11.23 in NAM1 to 21.8 in NAM10. To obtain a haploblock position, the average position of the 6 SNP markers was taken. Haplotypes were obtained from the simulated phased SNP genotypes generated by PedigreeSim.

Power Study

Definition of QTL interval

Single marker QTL methods do not provide an estimate for the QTL interval, yet with a defined threshold and a genetic map one can interpret the p-value distribution to obtain them. Since adjacent markers are not independent, and the closer to a true QTL position, the more significant the p-value becomes, one expects a chain of increasingly significant markers, pointing towards a true QTL position. Based on this, we define a QTL interval as a set of ordered markers above the significance threshold such that:

$$QTL = \{m_1, \dots, m_n\} \text{ where } d_{ij} < l$$

where d_{ij} is the distance between adjacent significant markers i and j, and l represents a *linking distance*. As a result, a QTL interval is defined by a chain of significant markers, where adjacent significant markers are at a distance smaller than l. Therefore, for each value of l we can define a set of detected QTL intervals. Since the choice of l is arbitrary, we performed power calculations with l from 0 to 10 cM in steps of 0.5 cM.

Significance threshold

To adjust for multiple testing, an empirical permutation threshold was calculated for each QTL analysis (Churchill and Doerge 1994). Thresholds were obtained with 100 permutations on a single population for each model, as threshold values did not change substantially between populations.

Power estimates

To evaluate the QTL models here presented we will use 1) QTL detection power, the probability of detecting a QTL position when present; 2) false positive rate, the probability of having a significant marker outside a QTL region; 3) QTL accuracy, the closeness of a QTL peak (position of maximum probability within an interval) to the true position and 4) QTL and marker precision, the probability that a significant QTL interval or marker is a true positive.

QTL detection power can be calculated as the proportion of true QTLs that are found by the model. While this is informative, one can easily increase detection power by increasing the number of false positives. To estimate the false positive rate, we must define the true negative markers (N). We considered as true negatives all markers outside a 10cM interval around our true QTL positions (5cM above and 5cM below). We then define as false positives (FP) those markers that are above the significance threshold (they have lower p-values, higher significance) and are outside the 10cM true interval. Lastly the false positive rate is calculated as *FP/N*.

The range of a QTL interval is defined by the positions of its leftmost and rightmost markers. QTL intervals will be considered *true positives* if the QTL range includes the simulated QTL position. All markers belonging to a true positive QTL interval are considered true positive markers, whereas the rest of significant markers present in other QTL intervals will be considered *false positives*. Isolated significant markers will be ignored.

Under this framework we can define detected QTLs, true QTLs, significant markers and true positive markers. We will use these values to calculate the precision (proportion of true positives over all positives) for both QTLs and markers.

$$QTL_{precision} = \frac{true\ positive\ QTLs}{detected\ QTLs}$$

$$marker_{precision} = \frac{true\ positive\ markers}{significant\ markers}$$

Finally, we considered the ability of a model to predict the position of QTL within an interval. We can define a QTL peak as the most significant marker within a QTL interval, as is done when applying logarithm of odds (LOD) scores. QTL accuracy can then be calculated as the average distance of a QTL peak in a true QTL to the true QTL position.

Power measures were calculated for each of the three models in 11 populations for each level of genetic diversity (total of 44 populations).

Implementation

All computations in this study were done in R (R Core Team 2016).

Ridge regression using a restricted maximum likelihood procedure was used to obtain the mixed model estimates, which in this context are equivalent to the Best Linear Unbiased Predictions (BLUP) (Whittaker et al. 2000; Meuwissen et al. 2001). Such calculations can be performed using the mpQTL package, where the solution algorithm, F-test approximation and p-value calculation where based on the mixed.solve() function of the rrBLUP package (Endelman 2011).

To improve computational efficiency, the EMMAX/P3D approach was applied (Kang et al. 2010; Zhang et al. 2010), which approximates variance components once, and recycles these components at each marker position, reducing the amount of large matrix multiplications that must be performed.

Simulation

Multiparental Population design and genotype Simulation

Nested Association Mapping (NAM) populations were generated using PedigreeSim V2.0, a simulation software that can simulate not only diploid but also polyploid meiosis (Voorrips and Maliepaard 2012). PedigreeSim generates genotypes given a genetic map, a pedigree and the genotypes of the first generation (founders) of that pedigree. Simulations were performed using Haldane's mapping function, allowing only bivalents with random pairing and the parameter "NATURALPAIRING" set to 1.

To speed up the calculations, an adapted tetraploid potato genetic map was used (Bourke et al. 2016) containing only the first five chromosomes (3509 markers representing 485 cM). The individuals used in this study were simulated in a two-stage process: firstly, ancestor individuals were generated and used to obtain ten separate populations (ancestral groups); secondly, from each ancestral group a set of NAM founders were chosen to obtain parallel NAM populations.

For each ancestral group (AG), 10 ancestor individuals were generated with random SNP scores at each marker. Each SNP position is also given an "IBD allele", unique to each homologue of each ancestor (even if the SNP state is the same). Each ancestral group has 10 founders, and thus a total of 40 IBD alleles will segregate in each AG. These alleles we will name *ancestral alleles.* Each ancestor is randomly crossed (without selfing, as potato is an outbreeder) to produce a first generation of 100 individuals, which will serve as parents of the second generation. This process was repeated for 50 generations, maintaining a constant generation size of 100 individuals. Finally, 100 individuals per AG were obtained as potential parents for the creation of NAM populations.

A NAM population consisted of one central parent crossed with nine peripheral parents, without any of the subsequent inbreeding that was originally proposed for NAM crossing scheme for selfing crops (McMullen et al. 2009). Each cross produced 50 offspring, thus totalling at 460 individuals per NAM. To simulate NAMs with different degrees of genetic diversity, parents were sampled from the same or from different AGs. A NAM1 contains parents from only one AG, while a NAM5 contains parents from 5 different AGs, with the same number of parents per group when possible. When the numbers of parents per AG was not equal the central parent always originated from the AG providing the most parents. For each level of genetic diversity, 11 populations were simulated. At the end of the process, the genotypes of each individual were obtained in terms of ancestral alleles (IBD alleles) and in terms of SNP dosages.

Phenotype Simulation

Phenotypes were simulated based on the simulated genotypes: genotypic values were obtained by assigning genetic effects to the ancestral alleles at pre-defined QTL positions. Each individual will then harbour four QTL alleles at each QTL position and the final phenotype is equal to the added effects of all QTL alleles plus a normally distributed noise. No interactions between alleles in one QTL or among QTL loci were simulated, and thus additive phenotypes were obtained.

We considered a situation where three unique QTL positions (at chromosome

1, 67.88 cM; chromosome 2, 61.2 cM and chromosome 4, 100.49 cM) were segregating. Each AG has a random allelic mean, and allele effects are drawn from a normal distribution around that mean. Additionally, 50 small-effect QTLs were added randomly across the genome to simulate a polygenic effect.

For further information see Appendix 1.

Results

Population Simulation

Ten Ancestral Groups (AGs) were simulated, each of them being founded with 40 different founder alleles. After 50 generations of random mating with a generation size of 100 individuals, each locus contained 8 to 20 founder alleles, with an average between 12.5 and 13.5 depending on the AG.

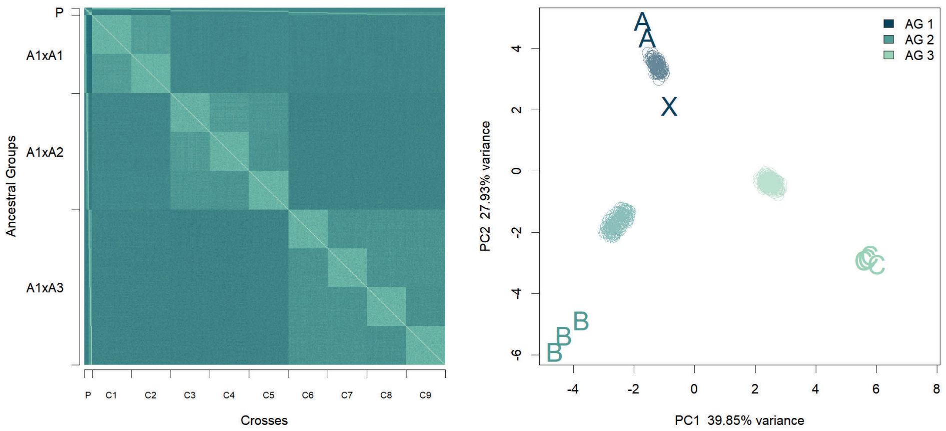


Figure 1. Visualization of genetic distance matrix K.

Left: Heatmap of K, where lighter colours indicate higher genetic similarity between individuals. (P, parents; An x Am, cross between AG n and AG m; Cn, cross n). **Right**: Individual genotypes plotted on the two first principal components of the K matrix. Dot clouds correspond to offspring of crosses 1 to 9 (X, central parent, of AG1; A, peripheral parents of AG1; B, peripheral parents of AG2; C, peripheral parents of AG3).

Parents from the last generation of AGs were used to obtain NAM populations. Different degrees of genetic diversity were simulated by sampling parents from the same or different AGs, thus producing genetic structure. This is visualized for one example in Figure 1, which shows a heatmap of the relatedness matrix K and a Principal Coordinate Analysis (PCoA) plot of the same matrix. On the left, we see how cross 3, 4 and 5, derived from crosses between AG1 and AG2 (A1 x A2 in Fig. 1), have a higher relatedness between them than with any other cross. Similarly, in the PCoA plot we observe how the individuals from these crosses (light blue dot cloud) cluster together in the midpoint between X (from AG1) and the three parents B (from AG2). These indications confirm that our two-step approach was successful in generating NAM populations with genetic structure. A similar outcome can be observed in the NAM1 to NAM10 simulations.

Population comparison

For each level of diversity, 11 populations were tested with the three proposed models. In almost all cases, al models were able to detect all QTL regions. Regardless of the linking distance used for QTL estimation, lower diversity resulted in higher detection power (Table 1). This can be observed at $l = 3$ using haplotype markers: NAM1 has a detection power of 1 (all QTLs were found in the 11 populations), but this power decreases to 0.818 in NAM10. Similarly, the false positive rate decreases as diversity increases and is lowest in the SNP model than in IBD or haplotype models. In Figure 2 the 99th percentile profiles also highlight the increased power in lower diversity populations, where the dark blue line representing NAM1 populations had higher significance values for all QTL peaks and for all models. As diversity increases, a similar decrease can be observed for QTL precision. Finally, the mean peak distance from the QTL peak to the true QTL position was also larger (lower accuracy) at a higher level of diversity in the populations (Table 1).

Marker comparison

Across NAM populations and at a linking distance (l) of 3 cM, detection power averaged at 0.74 for SNPs, 0.93 for IBD and 0.92 for haplotypes and was stable for $l > 1$ cM. The decrease in detection power as genetic diversity increased was markedly larger in the SNP models than in the multiallelic models (Table 1). This can be clearly observed in the 99th percentile lines in Figure 2: when diversity increases, the trend line is below the significance threshold in the SNP models, while for both multiallelic models all trend lines stay well above their respective thresholds. In Figure 3 left and centre panels, we can see how the proportion of true positives increases as the value of l increases. For $l > 1$ cM, QTL precision is on average higher for multiallelic models (0.91 IBD, 0.92 haplotype) than for the SNP model (0.86). Marker precision is also higher for the multiallelic models (0.99 IBD, 0.99 haplotype, 0.92 SNP). The choice of l has an impact on this difference, as for lower values of l (but above 1) precision is much lower for the SNP model. This is due to the presence of significant markers further away from the true QTL position in the SNP model than in the multiallelic models (Fig. 4).

	Detection power)			False positive rate			QTL precision			Accuracy (cM from true position)		
	SNP	IBD	Hap	SNP	IBD	Hap	SNP	IBD	Hap	SNP	IBD	Hap
NAM1	0.939	1	1	0.012	0.066	0.055	0.917	0.850	0.941	0.593	0.161	0.121
NAM3	0.909	0.970	0.970	0.008	0.065	0.054	0.865	0.814	0.886	0.550	0.192	0.130
NAM7	0.545	0.939	0.909	0.005	0.064	0.040	0.697	0.842	0.879	0.687	0.325	0.331
NAM10	0.606	0.848	0.818	0.005	0.055	0.038	0.773	0.932	0.850	0.665	0.312	0.621

Table 1: Power comparison across genetic diversity and marker types.
Each estimate is an average of 11 populations for each diversity level, with $l = 3$ cM. SNP refers to the biallelic model, IBD refers to the ancestral, multiallelic model and Hap refers to the haplotype-based approach. For detection power and QTL precision, higher numbers indicate a better model, while for false positive rate and accuracy, lower numbers indicate a better model.

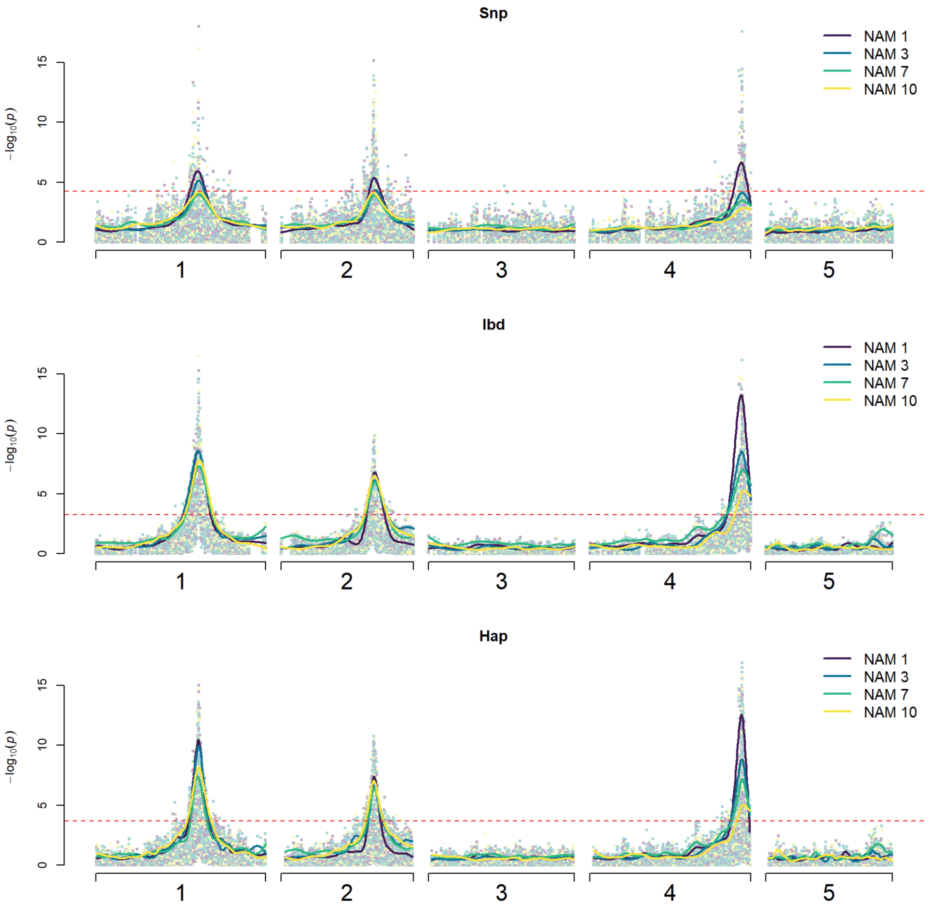


Figure 2: Overlap of p-value distribution across all populations in the three models.

Top, biallelic SNP model; middle, multiallelic IBD model; bottom, multiallelic haplotype model. Coloured solid lines represent the 99th percentile of all p-values observed in each genetic diversity level at a particular position. The red dotted line marks the estimated permutation threshold for each model (SNP: 10^{-422}, IBD: 10^{-327}, haplotype: 10^{-367}).

Peak accuracy (Fig. 3, right panel) is stable from $l > 1$ at 0.25 cM for IBD and 0.30 cM for haplotype models. In the SNP model, peak accuracy is lower and shows more variation. At $l = 1$ peak accuracy is similar to the IBD and haplotype models, yet many false positives are present in the QTL analysis (see Fig. 4). At higher *l*, average peak distance increases from 0.33 cM at *l* = 2 to 0.83 cM at *l* = 7.

Discussion

Model comparison

The essence of a QTL study is the genetic linkage between observed markers and unobserved QTL alleles. When dense genetic maps are used, the purpose of a QTL model should be to obtain an increasing marker significance as the analysis approaches a true QTL position. The definition of QTL interval used in this study stems from such reasoning: we expect a chain of contiguous significant markers that form a peak structure, pointing towards the true QTL position.

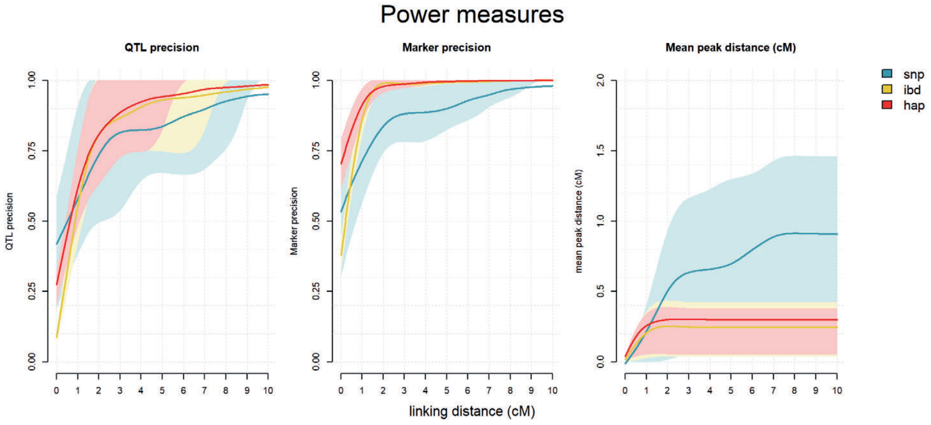


Figure 3: Power measures for each model with different values of l.

Power was calculated with l=0 to 10 in steps of 0.5 cM over 44 NAM populations (11 of each: NAM1, NAM3, NAM7 and NAM10), for the SNP dosage model (snp), true IBD model (ibd) and haplotype model (hap). Coloured areas represent the 20th to 80th percentile of power values for each model, and trend lines represent the average of each. Both lines and area edges where smoothed using a LOESS regression.

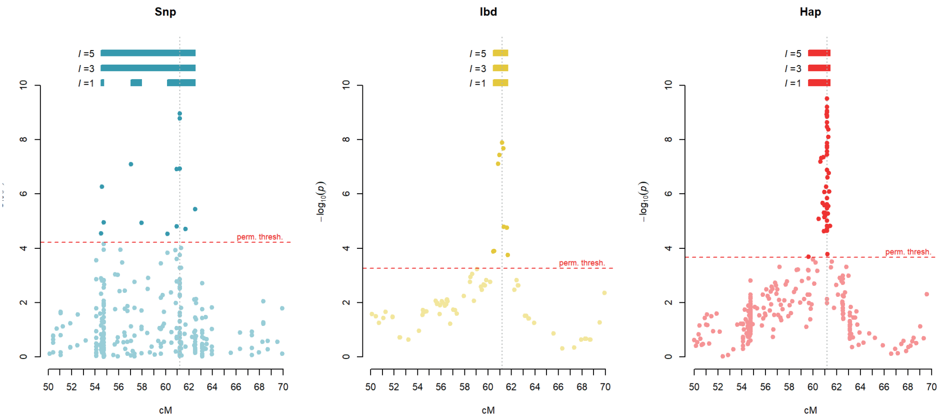


Figure 4: Example of QTL interval detection.

P-value distributions are shown for the same region in the same population using the three models (left, SNP; middle, IBD and right, haplotype). Detected QTLs are presented above each plot for three values of *l*: 1, 3 and 5 cM. Horizontal red dotted lines represent the permutation threshold for each model, and grey vertical lines highlight the true QTL position. In the SNP model with *l* = 1 cM, three QTL intervals are detected, of which only one contains the true QTL position, while with higher values of *l*, only one QTL interval is detected. In the IBD and haplotype models, a single true QTL is detected for all *l* values shown.

Classical QTL experiments were carried out on inbred diploid experimental crosses. In this setup one can expect only two alleles per QTL to segregate, and thus biallelic SNP markers are able to uniquely tag each allele. In this context, a SNP regression is equivalent to testing the difference in phenotype due to having 0, 1 or 2 copies of each marker allele (Lander and Botstein 1989; Haley and Knott 1992). However, when we move to scenarios where more than two alleles per QTL are expected to segregate at a single locus, for instance when heterozygosity is expected to be high or in multiparental populations, single SNPs no longer tag QTL alleles uniquely. Thus, each SNP allele might tag more than one functional QTL allele, creating a situation where the regression test is being performed between groups that do not represent a unique effect. Only if, by chance, those groups happen to divide functional alleles between those with large effects and those with small effects, will SNP markers be significant. Since two factors are affecting the significance of biallelic markers (i.e. distance to the true QTL position and the grouping of multiple effects), they become worse at estimating the true QTL position.

Figure 4 illustrates this situation. The three panels represent the same population being analysed with the three models presented in this study. It can be seen how in the SNP model there are three significant markers at approximately 54.5 cM, while the true QTL position is at 61.2 cM. Meanwhile there are quite some markers near the true position that are not significant. Such behaviour is not seen in the multiallelic models where markers near the true QTL position form a clear peak and more distant markers show no significance.

The consequences of this can be seen in Fig. 2 and Fig. 3. First, SNP models have overall lower significance at the QTL regions (Fig. 2), an effect that is increased when genetic diversity increases and biallelic markers become increasingly worse at tracking the multiple effects present in the population. This explains the lower detection power of biallelic models when genetic diversity is increased (Table 1). Secondly, we see how at low linking distances, SNP models have a high number of significant markers in false-positive QTL intervals (Fig. 3 middle). As l is increased, marker precision increases (there are less false-positive QTLs), but at the cost of accuracy (Fig. 3 left): the QTL intervals become larger (Fig. 4), including markers at some distance of the QTL position with higher significance than those at the simulated QTL position.

Thus, in a context of high genetic diversity, the usefulness of SNP models will depend on marker density, as higher density gives higher chances of having at least one marker at the QTL position that divides functional QTL alleles in two groups with statistically different means. Even if such a marker is found and the location of the QTL is detected, the effect estimated by a regression model does not realistically represent the true functional alleles present in the population.

Considering the lower detection power, lower accuracy and inability of biallelic QTL models to estimate effects for multiple alleles, it is clear that SNP-based biallelic models are a limited and limiting tool when applied to multiallelic populations.

Multiallelic markers

In order to apply multiallelic models, one must be able to obtain multiallelic genotypes. One possibility is to utilize markers that are multiallelic per se, such as SSR markers, but these markers are less common along the genome, their detection cannot be automated, and they are therefore hard to apply within high-throughput pipelines.

Alternatively, several studies have proposed the use of multiallelic haplotypes: groups of phased adjacent SNPs. This type of markers has the advantage of being predictive of two parts of IBD: family IBD, regions of chromosomes from two individuals that originate from the same *parental chromosome*; and ancestral IBD, chromosomal regions originating from the same *ancestral chromosome* that could occur in more than a single founder and that are broken down by recombination events (Browning and Browning 2011a).

While in our simulations haplotyping was trivial because the genotype of each individual was known, haplotyping of real SNP data requires *phasing.* For instance, if two adjacent marker genotypes of an individual are *AAAB* and *AAAB*, the underlying four haplotypes could be both *AA-AA-AB-BA* or *AA-AA-AA-BB*. Some approaches have been developed for haplotyping in polyploids (Browning and Browning 2011b; He et al. 2018; Thérèse Navarro et al. 2020) but regardless of the method, haplotype estimation from SNP data carries a certain degree of uncertainty due to the high number of possible solutions with similar probabilities. This uncertainty is not present in the haplotypes used in this study, meaning that the haplotype model here presented is performing better than what would be expected with real data, depending on the accuracy of haplotype estimation.

Nevertheless, sequencing technologies are becoming a mainstream approach for genotyping, and haplotypes can be directly observed in longer sequencing reads. Identifying haplotypes for different individuals given a set of reads is a complex mathematical problem that has spurred the development of a variety of tools (Berger et al. 2014; Garg et al. 2016; He et al. 2018; Motazedi et al. 2018). The haplotypes obtained from these methods could also be used with the multiallelic polyploid model introduced in this paper, allowing to perform QTL analysis in genetically diverse polyploid populations based on sequence data.

Lastly, in this simulated population each founder allele had a different QTL effect. In nature this might not be the case, as it is well known that many mutations are in fact neutral and thus do not change the QTL effect of that mutated allele. This could imply that the number of haplotypes would be higher than the number of QTL effects in a population, thus decreasing the usefulness of haplotype-based multiallelic markers.

Preparing multiparental populations

When organizing an MPP, the power to be able to detect the effects of an allele at a QTL depends on its frequency. The more individuals harbour one QTL allele, the more information the MPP provides about it. The expected frequency of founder alleles is directly affected by two factors: founder genetic diversity and offspring per founder.

The number of alleles segregating in a population is a direct reflection of the genetic diversity of its founders. When relatedness between founders is high, the chances of two founder chromosomes harbouring the same allele is also high. In MPPs where founders are very related, ultimately not many alleles can be expected to segregate. In contrast, when relatedness between founders is low, they have high chances to contribute unique alleles. The approach here presented estimates one parameter per each allele in the population, and thus, if population size is maintained constant, the power of the model decreases as the number of alleles increases. This hypothesis was confirmed by our simulation study where systematically, higher diversity populations, which require more allele effect parameters, presented lower QTL detection power, lower precision and lower QTL accuracy (Fig. 2, Table 1).

A second aspect to be considered is the number of offspring per founder. The larger the contribution of a founder to the individuals of the MPP, the higher the power to detect and estimate the effects of its alleles (Garin et al. 2020). For instance, using our NAM design, the alleles present in the central parent were present in all crosses. Alleles from peripheral parents not shared with the central parent had fewer individuals contributing to their effect estimation, meaning these estimations will be less powerful.

Considering the previous points, we suggest that MPPs should be developed with an intermediate diversity and ensuring that those alleles to be studied are kept at a relatively high frequency. Following this logic, a few parents from the same ancestral group (AG) can be selected (which likely share some alleles) and crossed with several other AGs. If all AGs are equally interesting for the QTL study, then all AGs should have a similar contribution to the offspring (Garin et al. 2020). If an MPP is designed from an already-existing set of connected F1 crosses, then each cross should be of similar size and the number of crosses per AG should be similar. When more complex pedigrees are used, ancestry coefficients can help guide the design of MPP.

Conclusion

Genetic diversity is the basis of breeding, and thus, characterizing it becomes essential in the development of new varieties. The methods developed within the “mpQTL” package add to the growing toolset for polyploid organisms. It is now possible to apply multiallelic models in polyploid organisms in the presence of genetic structure, which we have shown are more powerful, especially in the presence of high genetic diversity. Additionally, this study supports an alternative approach to the study of genetic diversity. Instead of using a diversity panel to perform a GWAS, a selection of these diverse accessions can be used as founders of an MPP. Each biparental cross within the MPP will add information to the QTL study, and future crosses can be added to the overall MPP analysis. This approach shows much promise in the context of breeding, particularly for its ability to connect and share information between crosses that in traditional approaches would remain separate.

Declarations

Data availability

The genotypes and phenotypes generated and analysed during the current study are available in the Figshare repository: doi.org/10.6084/m9.figshare.14315867

The QTL results generated and analysed during the current study are available in the Fighsare repository: QTL results, doi.org/10.6084/m9.figshare.14316068

Code availability

PedigreeSim can be found on GitHub: https://github.com/PBR/pedigreeSim

The package mpQTL together with vignette and example datasets can be found on CRAN or on Github (https://github.com/Alethere/mpQTL).

The code for producing the datasets used in this article can be found on Figshare: doi.org/10.6084/m9.figshare.14316107

Author's contributions

Conceptualization of the research was made by ATN, GT, REV, and CM. Data simulation and statistical analysis was performed by ATN with substantial help from GT. Supervision during the research was granted by GT, REV and CM. The first draft of this paper was written by ATN. Extensive review was granted by GT, REV, PA, MJMS, EvdW and CM. Funding was acquired by PA and CM. All authors have reviewed and approved this manuscript.

Acknowledgments

The contribution of all breeding companies involved in the project, and the Top sector Horticulture and Starting Materials (https://topsectortu.nl/nl) is highly appreciated.

Competing interests

The authors have no conflicts of interest regarding this article.

Funding

This research was funded under the TKI project: “Novel genetic and genomic tools for polyploid crops” (KV 1605-020; BO 26.03 009 004, BO 50 002 022). The funding committee evaluated the research proposal of which this study was part of.

Supplementary information

Genetic model selection and phenotype simulation

Simulating phenotypes is a challenging process as countless genetic situations might be generated: one big QTL effect and several small ones, many randomly sized QTLs, etc. An educated choice must be taken considering the type of research question to be addressed.

In this study, our objective was to characterize the statistical power of the IBD model in multiparental populations of polyploids. Ancestral groups represent closed populations from which parents are sampled and were created to simulate different degrees of genetic similarity between parents. This approach emulates common scenarios in plant breeding, where diversity is structured in gene pools due to geographic isolation and differential selection pressures between pools (e.g. Balfourier *et al.* 2018).

Genetic model

To describe our phenotypic simulation methods, let us consider a phenotype that is defined by:

$$y_i = G_i + E_i \quad E_i \sim N(0, \sigma_E^2)$$

Where y_i is the phenotype of individual i, with genetic effect G_i and random residual effects E_i, distributed as a normal random variable with mean 0 and variance σ_E^2. For simplicity, we simulated **additive QTLs** with no inter QTL interactions. Following the notation presented in Materials and Methods, the genetic effect can be defined as the sum of the allele effects multiplied by the dosage of each allele. Thus, for a phenotype defined by QTLs $l = \{1, \ldots, L\}$, each QTL containing $j = \{1, \ldots, k_l\}$ alleles and each allele having an effect α_{jl} (for the size of genetic effects, see following section), we can express the value y_i as:

$$y_i = \sum_{l=1}^{L} \left(\sum\nolimits_{j=1}^{k_l} \delta_{ijl} * \alpha_{jl} \right) + E_i$$

As heritability is a relevant parameter in QTL analysis, we might be interested in controlling it. Let us define a general heritability (of all QTLs combined) as $h^2 = \sigma_A^2 / (\sigma_A^2 +, \sigma_E^2)$ where σ_A^2 is the additive variance (in our case also the total genetic variance, as all genetic effects are additive). Assuming independence between genetic and environmental effects, we can rescale the genetic effects to achieve a heritability h^{2^*} using the following formula:

$$\alpha_{jl}^* = c\alpha_{jl} \qquad c = \sqrt{\frac{h^{2*}\sigma_E^2}{\left(1 - h^{2^*}\right)\sigma_A^2}}$$

Where α_{jl}^* represents the rescaled genetic effects. Additionally, a “polygenic term” has been added to the phenotypes, which is intended to increase family phenotypic resemblance. The polygenic term was generated by selecting 50 random positions along the genome, and assigning genetic effects distributed normally following $e_i \sim (\mu = 0, \sigma = 0.1)$ where e_i corresponds to the genetic effect on locus i and 0.1 is the standard deviation of the normal distribution. These genetic effects were assigned without taking into account the AG of the alleles, thus simulating family-relatedness rather than ancestral relatedness. As a result, genetic effects on polygenic loci ranged from -0.37 to 0.32, with an average of 0.002.

Choice of genetic effects

The number of genetic effects that must be obtained will depend on the number of different ancestral alleles present at each position, and thus is both position and population dependent. In a NAM3 population we have on average 11 alleles per locus. We simulated genetic effects in such a way that each ancestral group contributed functional alleles (alleles with a nonzero effect) for three QTLs positions. Once these alleles have been assigned we must assign allele effects. To do so, we considered three different scenarios:

1) Allelic effects within ancestral groups are very similar, but very different between ancestral groups. Those are genes that are very strongly selected (a trait that is essential for survival) but with a different effect (a different phenotypic value) being selected in each ancestral group.
2) Allelic effects with some variation within ancestral groups, but also some variation between ancestral groups. We could imagine a gene contributing to a non-essential trait of which multiple variants exist in each ancestral group, but that have different means for each ancestral group.
3) Allelic effects where the ancestral group has no influence on the effect distribution. Highly variable genes under diversifying selection would behave in such manner.

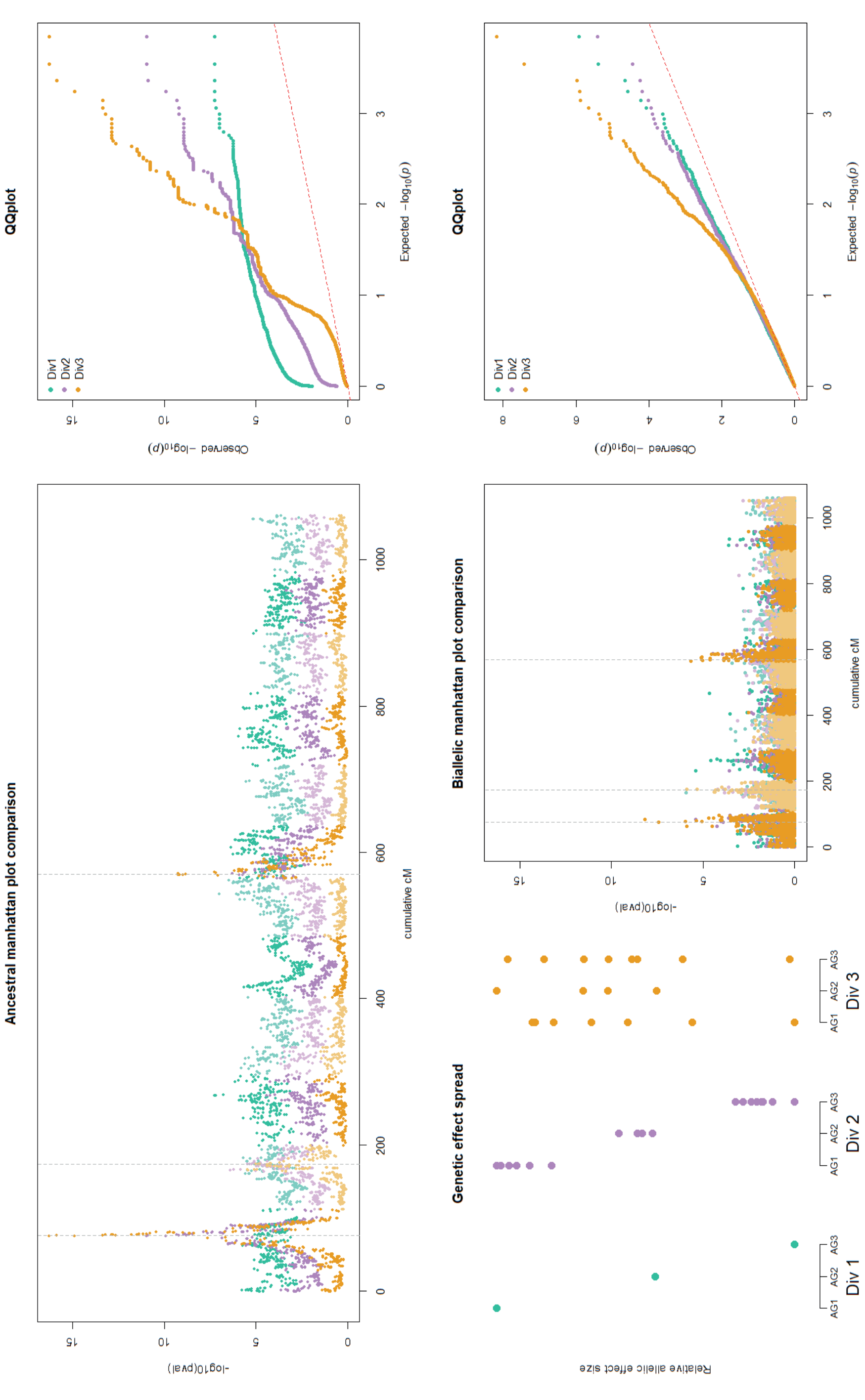
Once a scenario has been chosen, the allelic effect choice was performed as follows:

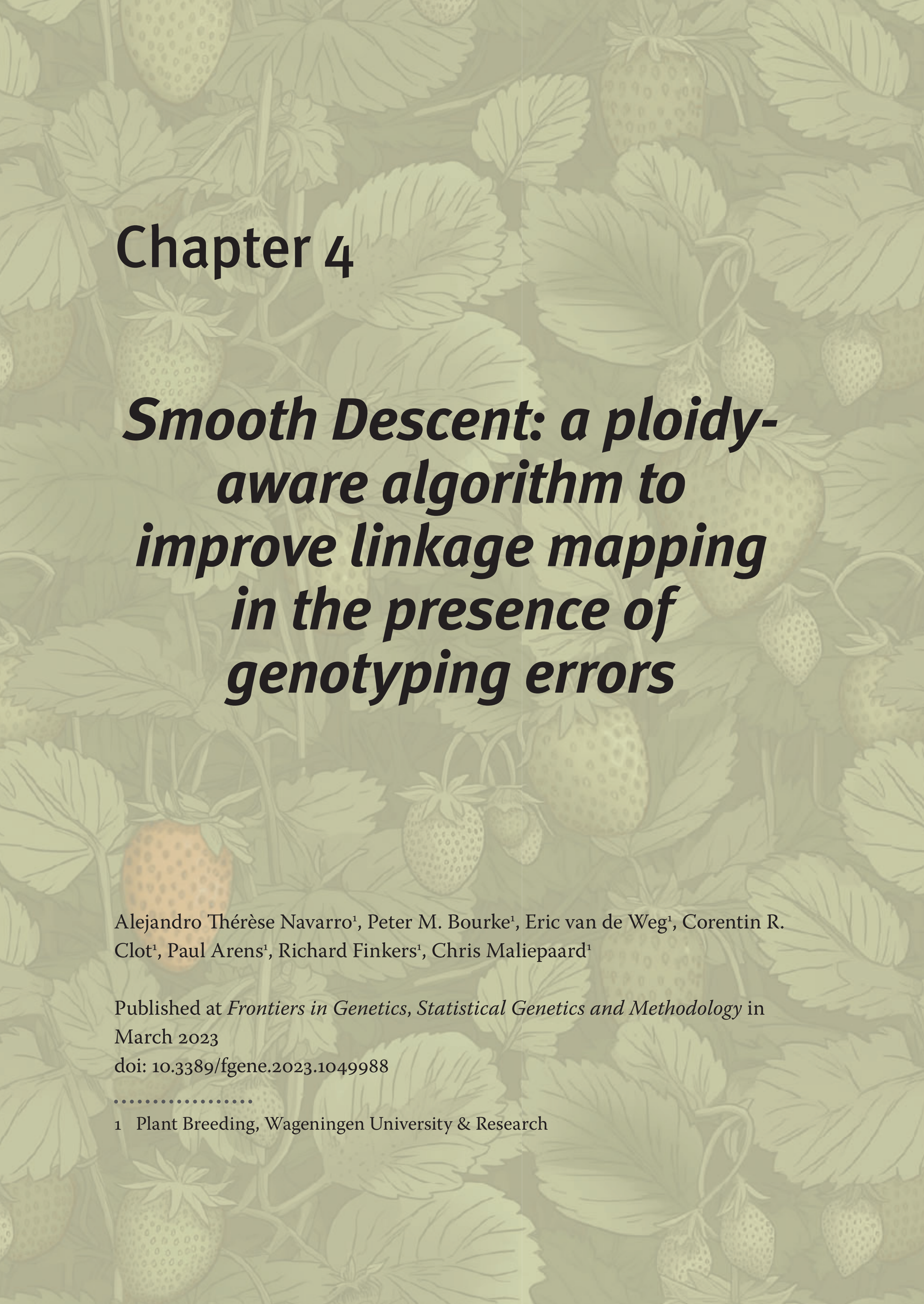
1) An ancestral mean μ_{AG} is chosen randomly from a uniform distribution. Since the effects will be scaled when the heritability control is performed, only the relative size between means is relevant.
2) Genetic effects are chosen from a normal distribution so that $\alpha \sim N(\mu_{AG},\sigma^2)$, where we consider all ancestral groups to have the same variance σ^2.

When σ^2 is much smaller than the difference between the μ_{AG}, we will simulate scenario 1. On the other extreme we find scenario 3, where the size of σ^2 is so large that differences between μ_{AG} are not meaningful. It is in the intermediate point between these two extremes where we can find scenario 2. Since it was not evident which scenario would be more interesting for our study, we simulated high heritability phenotypes with all of them and applied the ancestral model with true IBD alleles (Fig. S1)

We observed that in scenario 1, no peaks were found back, while scenario 3 had the strongest peaks and scenario 2 still detected peaks but of smaller size. Scenario 1, where all genetic diversity is *between* ancestral groups, rather than *within* ancestral groups has no detection power due to the structure correction we perform. Since we are eliminating structure-associated phenotype variation (in scenario 1 this is in fact all variation) the only phenotypic variation left is noise, and thus no QTL peaks are detected. While scenario 3 offers the most power, it seems unlikely that there would be no relationship between ancestral group and genetic effect. Thus, **we performed simulations using scenario 2.** In practice, that meant generating phenotypes using three random means between 0 and 1 with $\sigma^2 = 1$.

Supplementary Figure 1 (following page): Genetic effect diversity scenarios. When there is a high correlation between phenotype and genetic structure, that is, when certain effects are present only in subsets of the population, the models become insensitive. In green we see a case where all the effect variance is present between AGs (Div 1), in purple when variance is divided between and within AGs (Div 2) and in ochre, the case when effect variance is only present within AGs and all AGs harbour similar effects (Div 3). **Top left:** overlap of three effect models for an ancestral model. The diversity scenario 1 is the least powerful since the structure correction is elimination all population variation, in scenario 2 there is some detection although p-values remain inflated and in scenario 3 there is no inflation and QTL peaks are clearly detected. **Top right:** QQ-plot of the p-values in the top left panel, highlighting the p-value inflations seen in scenarios 2 and 3. **Bottom left:** graphical representation of the relative effect sizes for alleles of each AG in each of the diversity scenarios. **Bottom middle:** same Manhattan plot as in top left but using a biallelic model. In this case there is no inflation since alleles are not nested within certain parts of the population. **Bottom right:** QQ-plot of the bottom middle panel. We see how in this case there is no inflation.


QQplot
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
5
10
15
0
1
2
3
Div1
Div2
Div3
QQplot
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
2
4
6
8
0
1
2
3
Div1
Div2
Div3
Ancestral manhattan plot comparison
-log10(pval)
cumulative cM
0
5
10
15
0
200
400
600
800
1000
Biallelic manhattan plot comparison
-log10(pval)
cumulative cM
0
5
10
15
0
200
400
600
800
1000
Genetic effect spread
Relative allelic effect size
AG1
AG2
AG3
Div 1
AG1
AG2
AG3
Div 2
AG1
AG2
AG3
Div 3



Chapter 4

Smooth Descent: a ploidy-aware algorithm to improve linkage mapping in the presence of genotyping errors

Alejandro Thérèse Navarro[1], Peter M. Bourke[1], Eric van de Weg[1], Corentin R. Clot[1], Paul Arens[1], Richard Finkers[1], Chris Maliepaard[1]

Published at *Frontiers in Genetics, Statistical Genetics and Methodology* in March 2023
doi: 10.3389/fgene.2023.1049988

[1] Plant Breeding, Wageningen University & Research

Abstract

Linkage mapping is an approach to order markers based on recombination events. Mapping algorithms cannot easily handle genotyping errors, which are common in high-throughput genotyping data. To solve this issue, strategies have been developed, aimed mostly at identifying and eliminating these errors. One such strategy is SMOOTH, an iterative algorithm to detect genotyping errors. Unlike other approaches, SMOOTH can also be used to impute the most probable alternative genotypes, but its application is limited to diploid species and to markers heterozygous in only one of the parents. In this study we adapted SMOOTH to expand its use to any marker type and to autopolyploids with the use of identity-by-descent probabilities, naming the updated algorithm Smooth Descent (SD). We applied SD to real and simulated data, showing that in the presence of genotyping errors this method produces better genetic maps in terms of marker order and map length. SD is particularly useful for error rates between 5% and 20% and when error rates are not homogeneous among markers or individuals. With a starting error rate of 10%, SD reduced it to ~5% in diploids, ~7% in tetraploids and ~8.5% in hexaploids. Conversely, the correlation between true and estimated genetic maps increased by 0.03 in tetraploids and by 0.2 in hexaploids, while worsening slightly in diploids (~0.0011). We also show that the combination of genotype curation and map re-estimation allowed us to obtain better genetic maps while correcting wrong genotypes. We have implemented this algorithm in the R package SmoothDescent.

Keywords

Linkage mapping; genotyping error; identity-by-descent; imputation

Smooth Descent: a ploidy-aware algorithm to improve linkage mapping in the presence of genotyping errors

Introduction

Linkage mapping is the process by which a set of markers segregating in a population are grouped and ordered. Each marker is placed within a linkage group, oftentimes corresponding to a chromosome, and given a genetic position within that group. The usefulness of genetic mapping has made it a consistent tool during the past century: starting with the study of trait co-segregation in *Drosophila* (Sturtevant 1913), continuing to the proof of the linear structure of genes and chromosomes (Benzer 1959), and the first QTL analyses (Lander and Botstein 1989). Its relevance has not diminished nowadays, as it enables the study of genomic patterns of recombination, thereby highlighting the functional and structural properties of a genome. Linkage maps are also an essential tool for studies in organisms without a reference genome (e.g. (Hu et al. 2021a), in plant and animal QTL studies and in the assembly and improvement of genome sequences (Mascher and Stein 2014; Fierst 2015).

Genetic mapping algorithms have been greatly influenced by the progress of genotyping. As newer technologies provided larger marker sets, novel mapping algorithms had to be developed to handle growing numbers of markers (Cheema and Dicks 2009). The most recent genotyping techniques, sequencing-based methods such as genotyping by sequencing (Elshire et al. 2011) or whole genome sequencing (Varshney et al. 2014), are able to identify and genotype millions of variants in a single analysis but suffer from a common drawback: an increased proportion of genotyping errors. That is particularly problematic for the purpose of genetic mapping, since the ordering algorithms on which many mapping approaches rely are notoriously sensitive to errors (Hackett and Broadfoot 2003; van Os et al. 2005; Cartwright et al. 2007). Since most algorithms depend on pairwise recombination estimates, wrong genotypes can give the false estimate that a double recombination has occurred, producing sub-optimal map orders and inflated map lengths (i.e. >100 cM). The general strategy to deal with this problem has been to detect and eliminate highly spurious markers (Lincoln and Lander 1992; van Os et al. 2005; Cartwright et al. 2007; Wu et al. 2008; Cheema and Dicks 2009; Liu et al. 2014; Rastas et al. 2016), although the errors can also be explicitly modelled, increasing the number of retained markers (Bilton et al. 2018).

Polyploidy, the presence of more than two chromosome sets in an organism, is a relatively common condition in crop species (e.g. rose, potato, strawberry, sugarcane, wheat) that poses special challenges in linkage mapping. In autopolyploids, which usually originate from genome duplication within a single species, polysomic segregation and double reduction require specialized methods of linkage estimation (Bourke et al. 2018a). In allopolyploids, arising from interspecific hybrids, segregation usually follows a diploid pattern, but genotyping can be more inaccurate due to the difficulty of distinguishing between homoeologous sequences (Kaur et al. 2012). Although these issues have been addressed with specialized tools and approaches (Glover et al. 2016; Bourke et al. 2018b), these tools were not designed with consideration of the high error proportion in sequencing-based genotype data, and due to the unique challenges of polyploids, diploid-oriented tools cannot be used.

In this study, we aimed to develop a ploidy-aware approach that would help in using high-throughput genotyping information for genetic mapping, without discarding vast amounts of data due to an increased error rate. Therefore, we adapted SMOOTH (van Os et al. 2005), a simple and efficient method for error detection and correction based on the identification of unlikely genotype scores. The original algorithm was only applicable to diploids and to markers heterozygous in only one of the parents. By using identity-by-descent (IBD) probabilities, we extended this model to any ploidy and marker segregation type. Additionally, we changed the k-nearest neighbours approach used in SMOOTH to an interval-based approach, which improves identification and correction of errors in maps with a heterogeneous marker distribution. We term this updated method Smooth Descent, the IBD-based descendent of SMOOTH. Similar to the original algorithm, Smooth Descent requires a preliminary map to be applied, thus it should be thought of as part of an iterative mapping approach, so that with each round of mapping and smoothing a better map is obtained.

This algorithm has been implemented as an R package called 'SmoothDescent'. The package also generates so-called "graphical genotypes" that can be used as a quality assessment tool by researchers, along with visualizations of the iterative correction process and other diagnostic plots.

Materials and Methods

Smooth Descent approach

SMOOTH and Smooth Descent are both based on the same principle: comparing an observation (error sensitive) and expectation (error tolerant) matrix of genotypes and identifying as errors the inconsistencies between both matrices. The difference lies in the way genotypes are expressed in both approaches: as raw genotype scores in SMOOTH, and as Identity-by-Descent (IBD) probabilities in Smooth Descent. In Smooth Descent observed IBD is obtained through the naive IBD algorithm described below, while expected IBD can be obtained through two methods, weighted average IBD or hidden Markov model IBD. The three methods are described below.

Naive IBD probabilities

The algorithm begins with parental phasing and a preliminary map that indicates the order and distances of markers. A number of methods can be used, experimental and computational, to obtain parental phasing (Browning and Browning 2011b; He et al. 2018; Al Bkhetan et al. 2021) and a preliminary map (Rastas 2017; Bilton et al. 2018). In our software, mapping is performed by polymapR (Bourke et al. 2018a) and parental phasing is expected to be obtained by the researcher.

Phased parental genotypes are expressed using the homologue matrix H, in which columns represent parental homologues and rows are markers, ordered according to the preliminary map. The number of columns p will be the sum of parental ploidies. Thus, the matrix H is composed of columns H_1 to H_p. In a diploid cross $p = 2 + 2 = 4$, there would be 4 columns; in a tetraploid cross, 8 and in a cross between a diploid and a tetraploid, 6 columns would be specified. The first set of columns correspond to the homologues of the first parent, and the rest to the homologues of the second parent. Each cell of the H matrix contains a 0 when that homologue holds the reference allele A at that marker, and 1 if it holds the alternative allele B. Because of this, only biallelic markers can be used in Smooth Descent. The choice of reference allele will not influence IBD calculations, and thus it can be done at random. For a diploid cross, an example of H would be:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad [1]$$

In a tetraploid:

$$H = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \qquad [2]$$

First, we will calculate the error-sensitive, observed IBD probabilities or naïve IBD probabilities. For that we need to obtain all possible homologue combinations that can be inherited, which we denote as *configurations* with the symbol c_i. This will depend on the number of homologues that parent 1 and parent 2 pass on to the offspring, which in turn depends on their ploidy.

In the case of a diploid, parent 1 provides a single homologue, either H_1 or H_2; while parent 2 can provide H_3 or H_4. Although there can be recombinations along the inherited homologues (*e.g.* switching from H_1 to H_2), this does not affect our analysis since it is performed marker by marker. Thus, there are four configurations, c_1={H_1,H_3}, c_2={H_1,H_4}, c_3={H_2,H_3}, c_4={H_2,H_4}. On the other hand, in a tetraploid example, each parent will provide two homologues. Thus, a single parent can provide any of six pairs of homologues: (H_1,H_2), (H_1,H_3), (H_1,H_4), (H_2,H_3), (H_2,H_4) or (H_3,H_4). Moreover, due to multivalent formation, double reduction scenarios are possible, meaning that parent 1 could also contribute (H_1,H_1), (H_2,H_2), (H_3,H_3) or (H_4,H_4). If both parents are tetraploid, this amounts to 100 possible configurations. However, since double reduction is relatively rare, and for the sake of simplicity, it has not been considered in this implementation of Smooth Descent. Thus, we will only consider the 36 configurations possible, *i.e.* we assume that no double recombination occurs.

The next step is to determine the marker *dosage*, d_j, (of the alternative allele) of each configuration. This must be calculated independently for each marker. For one marker, matrix H assigns either 0 or 1 to each parental homologue. The inherited dosage of that configuration is simply the sum of the associated parental homologues. For instance, for the first marker (row) in the diploid example, c_1 = {H_1,H_3} thus d_1 = 1 + 0 = 1 while c_3 = {H_2,H_3} thus d_3 = 0 + 0 = 0. For the first marker of the tetraploid example, c_1 = {H_1,H_2,H_5,H_6} thus d_1 = 0 + 1 + 0 + 0 = 1 *etc.*

To obtain IBD probabilities for one individual, one must consider the observed genotype of that individual. Since an individual must hold one of the described configurations, only those configurations whose dosage matches the observed genotype are *possible configurations.* For each genotype g, we denote the set of possible configurations as C_g, where k_g the number of possible configurations. When no double reduction is considered, all configurations are equally probable, thus the IBD probability of H_i is:

$$p(H_i|g) = \frac{\sum_{j \in C_g} f(c_j, H_i)}{k_g} \qquad [3]$$

Where $f(c_j, H_i)$ is an indicator function that takes the value 1 if H_i belongs to

c_j and 0 otherwise.

$$f(c_j, H_i) : \begin{cases} if\ H_i \in c_j\ then\ 1 \\ if\ H_i \notin c_j\ \ then\ 0 \end{cases} \qquad [4]$$

For example, let us consider an offspring for the two parents represented in the homologue matrix in equation 1 with a genotype of 1, 0, 1. The possible inheritance configurations for a diploid parent are c_1={H_1,H_3}, c_2={H_1,H_4}, c_3={H_2,H_3}, c_4={H_2,H_4}. For the first marker H_1 = 1; H_2 = 0; H_3 = 0; H_4 = 0, meaning that each configuration has the following values: c_1 = 1, c_2 = 1, c_3 = 0 and c_4 = 0. Only two configurations, c_1 and c_2 are possible given that the genotype is 1, meaning that k_g = 2. Thus:

$$p(H_1|1) = \frac{f(c_1,H_1) + f(c_2,H_1)}{2} = \frac{2}{2} = 1$$

$$p(H_2|1) = \frac{f(c_1,H_2) + f(c_2,H_2)}{2} = \frac{0}{2} = 0$$

$$p(H_3|1) = \frac{f(c_1,H_3) + f(c_2,H_3)}{2} = \frac{1}{2} = 0.5$$

$$p(H_4|1) = \frac{f(c_1,H_4) + f(c_2,H_4)}{2} = \frac{1}{2} = 0.5$$

A similar process can be followed for the second marker. In that case H_1 = 0; H_2 = 1; H_3 = 1; H_4 = 0, meaning c_1 = 1, c_2 = 0, c_3 = 2 and c_4 = 1. Only one configuration is possible that the genotype is 0: c_2, thus k_g = 1. Applying equations 3 and 4 as done above yields the following results:

$$p(H_1|1) = 1 \quad p(H_2|1) = 0 \quad p(H_3|1) = 0 \quad p(H_4|1) = 1$$

Lastly, the third marker can be computed considering that H_1 = 0; H_2 = 0; H_3 = 0 and H_4 = 1. Thus, c_1 = 0, c_2 = 1, c_3 = 0 and c_4 = 1. In this case the genotype is also 1, meaning that k_g = 2, since only c_2 and c_4 are possible. This yields:

$$p(H_1|1) = 0.5 \quad p(H_2|1) = 0.5 \quad p(H_3|1) = 0 \quad p(H_4|1) = 1$$

If we combine these results, we can obtain the IBD matrix according to the naive model for this individual:

$$I_0 = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 1 & 0 & 0 & 1 \\ 0.5 & 0.5 & 0 & 1 \end{bmatrix}$$

This algorithm will be applied after each iteration of correction, as described below, to obtain matrix I_1, and subsequently to obtain matrix I_2, etc.

IBD prediction – weighted average

One of the two methods implemented for IBD prediction in Smooth Descent is based on a local weighted average of observed IBD around a marker, inspired by SMOOTH's proposal and similar to the procedure suggested by (Wu et al. 2008). This requires two steps: first, defining the set of local markers and second, estimating the weights to be applied to each marker.

Let's start with marker m_i. The set of local markers, L_i, are those markers closer than from m_i, where l is a chosen distance threshold (we chose l = 10 cM, but a different threshold can be provided). Additionally, low-informative markers will be excluded from the local set. We defined these as markers for which the observed IBD probability is within the 0.3 – 0.7 range (see Error Prediction section for more information). Since the predicted IBD is calculated per homologue, this means that L_i will differ slightly per homologue.

The weight for the observed IBD probability at marker m_j will be proportional to the chance that there is no recombination between m_i and m_j. This no-recombination probability can be obtained from the distance estimates:

$$1 - \rho_{ij} = 1 - f(d_{ij}) \qquad [5]$$

Where 1 - ρ_{ij} is the probability of no recombination and $f(d_{ij})$ is a reversed mapping function of the distance between m_i and m_j. Three functions have been implemented: Morgan's, Haldane's and Kosambi's. We can define the weights as:

$$w_j = \frac{1 - \rho_{ij}}{\sum_{k \in L_i}(1 - \rho_{ij})} \qquad [6]$$

For each individual, the predicted IBD probability for marker m_i will then be the weighted average of all the markers in L_i, for which $d_{ij} < 1$ and the observed IBD probability is informative. Applying this along the I_0 matrix will allow us to calculate the predicted IBD matrix $\widehat{I_0}$.

IBD prediction – hidden Markov model

The second model for IBD prediction is based on a hidden Markov model (HMM), a common approach to obtain error-tolerant IBD estimates (Zheng et al. 2016, 2021; Mollinari and Garcia 2019). We have included in Smooth Descent the HMM implemented within polyqtlR (Bourke et al. 2021), an expanded version of the TetraOrigin model (Zheng et al. 2016). This HMM uses a discrete-time Markov chain to model parental origins of chromosomes along the markers of each offspring. To do so, it models homologue pairing in the gamete's meiosis, including recombination probabilities and gamete fusion to constitute a zygote, thus closely modelling the biological reality of inheritance. By defining a series of likelihoods for the parental haplotypes conditional on the offspring genotypes, it provides a powerful tool for estimating IBD probabilities and recombination points.

Error prediction

In SD error estimation is performed by comparing an error-sensitive IBD matrix (naive IBD) with an error-tolerant matrix (weighted average IBD, or HMM IBD). Therefore, using SD one can obtain error estimates by comparing naïve probabilities to the weighted average probabilities, or to the HMM-based IBD probabilities.

Each IBD matrix, I_0 or $\widehat{I_0}$ is composed of IBD probabilities for each homologue and each marker, which we term i_0 and $\widehat{i_0}$ respectively. The principle of error prediction is to identify markers for which their observed and predicted IBD probabilities disagree strongly, meaning that the observed genotype clearly indicates a homologue inheritance that does not match the predicted IBD. More formally, an error can be identified if $|i_0 - \widehat{i_0}| > \delta$, where δ is an error threshold preferably above 0.7.

Due to this definition, low-informative markers (with observed probabilities between 0.3 and 0.7) must be excluded from the weighted-average IBD prediction step. The contrast $|i_0 - \widehat{i_0}|$ will not reach a high value if either i_0 or $\widehat{i_0}$ are close to 0.5. The observed IBD i_0 will be close to 0.5 if the observed inheritance is uncertain, which means we do not have enough information to discern whether that genotype is an error. The predicted IBD $\widehat{i_0}$, should be close to 0.5 if the set of local markers have both high and low IBD probabilities, indicating that there is a local disagreement on inheritance. If low-informative markers are kept, even if many informative markers exist that clearly indicate homologue inheritance, the presence of low-informative markers will centralize the local weighted average and prevent identification of putative errors. Thus, low-informative markers should be removed from IBD prediction.

Genotype correction and iteration

When a marker is detected as erroneous, a new genotype can be imputed by computing the most likely marker genotype according to the predicted IBD. The new set of genotypes can be used to calculate an improved map, and a corrected IBD matrix, I_1. The previous steps can then be repeated to obtain a new error matrix E_1 and further improved genotypes. Thus, an iterative approach emerges, where in each iteration the genotypes are further corrected. As iterations progress the genetic map is expected to change less, and thus we are more certain of the achieved order. In view of caution regarding the introduction of artefacts, the error threshold was set at $\delta = 0.9$ during the first iteration, and then slowly decreased to 0.7 as iterations progress.

Best iteration selection

When using Smooth Descent, we must choose the best iteration according to some criterion. We offer the R^2 estimate of the second-order polynomial relationship (i.e. $d = a + br + cr^2 + \varepsilon$) between inter-marker distance d, and the recombination frequency (not to be confused with distance-based recombination frequency ρ used for IBD prediction). Unlike ρ, r is calculated during the mapping process through a likelihood or Bayesian method and is the basis of the final map order. In a good map, the relationship between r_{ij} and d_{ij} should be mostly linear, where high recombination frequencies lead to high distances. Thus, the iteration with the highest R^2 can be considered the best.

Simulated data

PedigreeSim (Voorrips and Maliepaard 2012), a program that simulates meiotic pairing and recombination for a range of pedigrees and ploidies, was used to simulate genotype data. We simulated diploids, tetraploids and hexaploids. For each ploidy, ten F1 populations were simulated (30 in total) with 100 individuals each. Every individual had one single chromosome containing 200 segregating markers distributed at variable densities along the chromosome. Error rates were applied randomly by changing the genotypes of 1%, 5%, 10%, 20% of the markers.

Additionally, two special cases were designed to test the effect of variable error rates across individuals (special case A) and across markers (special case B). Special case A contained 80 individuals with an error rate of 0.02 and 20 individuals with an error rate of 0.3. Special case B had the same error rate for all individuals, but variable across markers, ranging in a continuous curve along the chromosomes. The curve was defined as a smooth spline passing through the error rates 0.02, 0.1, 0.3, 0.02 and 0.1 at approximately 25 cM intervals along the chromosome. Thus, high error rate markers were located close to one another and at the centre of the chromosome.

Each genotype dataset was mapped using Smooth Descent with 10 iterations and tested using the weighted average or HMM method for computing error-tolerant IBD probabilities. To evaluate the effectiveness of SD, as well as the additional tools tested, three parameters were used: genotyping error, the percentage of genotypes different from the true genotypes; position correlation, the correlation between the true map positions and estimated map positions; and map length, the size of the estimated genetic map.

Real data

Data from strawberry (*Fragaria* x *ananassa*) data was obtained from whole genome sequencing of 48 individuals from an F1 population. Variant discovery was performed using bcftools and genotyping with the R package "updog" (Gerard et al. 2018)' allowing to genotype ~10M markers. After filtering markers based on depth and genotyping quality, ~1.8M markers were kept and summarised into ~6500 unique markers across all chromosomes. Due to a skim sequencing strategy, many genotyping errors were expected and observed, which proved this dataset useful for testing our approach. Since strawberry is an allopolyploid with strict chromosomal pairing behaviour, the data could be treated as that of a diploid.

Data from sweet potato (*Ipomoea batatas*) was taken from (Mollinari et al. 2020). Sequencing was performed using the polyploidy-optimized method described in GBSpoly (Wadl et al. 2018). The obtained read counts were passed to SuperMASSA (Serang et al. 2012) and genotypes were filtered for quality. For chromosome 15, a final count of 1513 genotypes were obtained for 287 individuals. These genotypes were used with SD, creating a preliminary map *de novo* and performing genotype correction on the genotypes. A single iteration of SD was used since no more improvements could be made subsequently.

Data from diploid potato was taken from (Clot et al. 2022). The dataset consisted of 1536 full-sibs from a cross between two heterozygous clones C (USW5337.3) and E (77.2102.37). This population was skim sequenced to an average coverage of ~1.5x. Parent specific SNPs were called using bcftools v.1.13 and used to impute haplotypes in bins of 0.1Mbp resulting in 4893 female and 4735 male segregating markers. Smooth Decent was used based on physical position with five rounds at prediction interval of 1 Mbp and two final rounds with a prediction intervals of 5 and 10 Mbp respectively.

Software comparison

SD is a unique tool since it is the only available tool that aims at correcting polyploid (and diploid) linkage maps while simultaneously correcting genotyping errors. However, other tools exist that can perform one of the two functions. We have compared SD to polymapR (Bourke et al. 2018a), a polyploid linkage mapping approach that does not perform genotype correction; and to MAPpoly (Mollinari and Garcia 2019), a HMM approach that is able to correct genotypes and re-estimate marker positions but that does not re-compute linkage map orders.

Ten F1 populations equivalent to those described in the Simulated Data section were used. Genotyping errors were added at a rate of 1%, 5%, 10%, 15%, 20%, 25% and 30%. For each population and error rate four approaches were tested: polymapR, MAPpoly, SD using weighted average IBD prediction and SD with HMM IBD prediction. For both MAPpoly and SD the same preliminary map was provided. Additionally, the error prior provided to MAPpoly was the actual simulated error rate. Lastly, SD results were obtained with 5 iterations since previous results (see Simulation results) showed that iterating more than 5 times did not have a significant impact in the result.

After running each approach, position correlation (correlation between true and estimated map positions), map length and computational run-time were obtained. Genotyping error was only calculated for SD and MAPpoly methods, since polymapR does not perform genotype correction.

Results

Simulated data

A total of 10 populations per ploidy were tested with 6 different levels of genotyping error and two IBD prediction methods, showing the usefulness of Smooth Descent (SD) in correcting genotypes, improving map orders and shortening map lengths (fig. 1). It can be observed how the most impactful changes occur in the first few iterations: the biggest change in genotype correctness (fig. 1 top), the largest improvement in genetic map correctness (fig. 1 middle) and the biggest reduction in map length (fig. 1 bottom). Note that map length was particularly short in polyploids (~60 cM in tetraploids and ~45 cM in hexaploids), an issue that seems to stem from preliminary map calculation.

Ploidy is an important factor in the behaviour of SD, moving from a genotype corrector at lower ploidies to a map corrector in higher ploidies. In diploid cases (fig. 1 left column) SD is able to halve genotyping error (e.g. ~5% reduction in the 0.1 error rate scenario, fig. 1 left top, table 1) and to shorten map lengths, especially in the highest error rate cases (e.g. ~30 cM shortening,

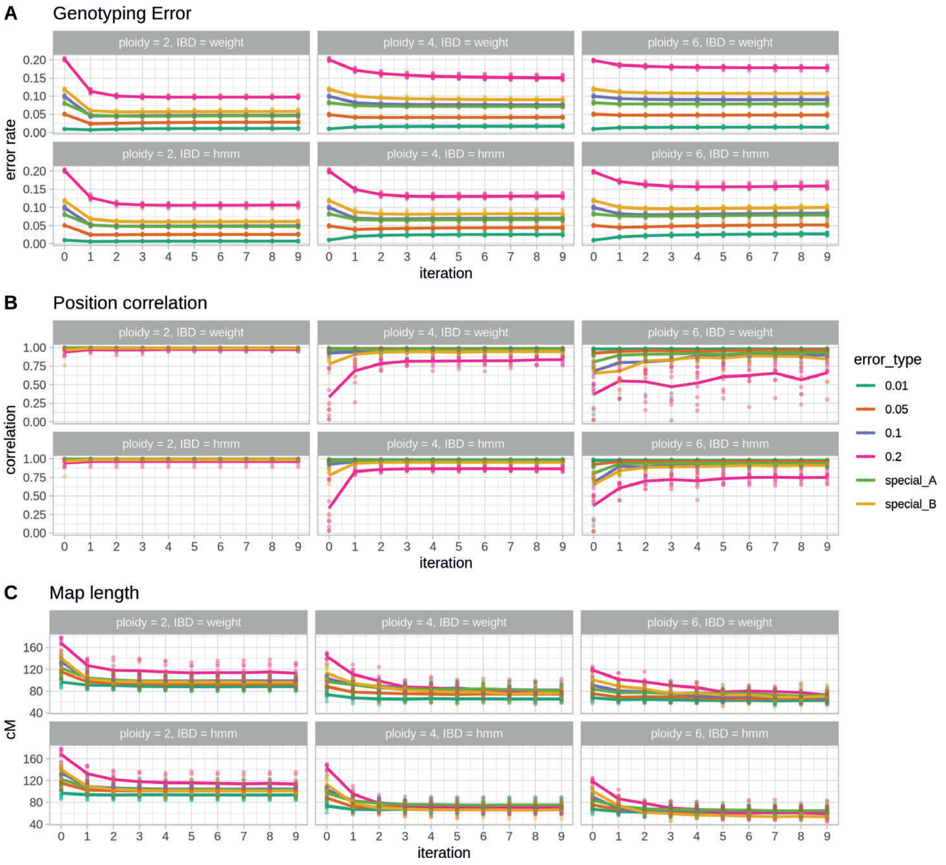


Figure 1: Results of 10 simulated populations across error rates and ploidies. Within each section, each column represents a ploidy and panel the top row shows the results for the IBD estimation with the weighted average procedure (IBD = weight) and the bottom row for the IBD estimation with the HMM (IBD = hmm). A) Genotyping error, the rate of genotypes that are different from the true genotypes. B) Position correlation, the correlation between true genetic positions and estimated positions in a genetic map. C) Map length, the size in cM of the estimated maps. In each plot, points represent individual observations and lines are the average. Each colour represents one simulated error type, with special A being heterogeneous rate across individuals and special B being heterogeneous rate along the map.

fig. 1 left bottom, table 1). Nevertheless, in diploids, SD does not significantly impact the correlation (there's a small decrease) between true and estimated map positions, since the preliminary map is already highly correlated to the true map, although longer. In contrast, in polyploid scenarios reduction of genotyping error is smaller (table 1, fig. 1 middle and right columns), but the correlation between true and estimated maps improves substantially, especially in the hexaploid case. Map size reduction is of the same order, about 30cM. Importantly, for lower error rate cases, there was a slight increase in genotyping errors, although this did not affect the correlation with the true map or map size. This can be attributed to incorrect imputations by the SD algorithm. Wrong imputations occur in all scenarios, but in most cases they represent a small fraction of the imputed genotypes, finally yielding an overall improved genotype correctness. Only when ploidy is high and genotyping error is low the number of correct genotypes decreases due to wrong imputations.

The two IBD prediction methods tested (weighted average and HMM) performed similarly in diploids but had some differences as ploidy increased. Genotyping error correction was better for the HMM as ploidy and initial error rate increased (table 1, error rate 10%). Consequently, estimated map positions and map sizes were also better for the HMM in high ploidy and high error rate cases. However, at lower error rates the HMM method produced a larger increase in genotyping errors (table 1, error rate 1%).

Real data

Two real datasets were tested using Smooth Descent, a low-depth dataset of garden strawberry (*Fragaria × ananassa*) (fig.2 A), chromosome 15 of *Ipomoea batatas* (fig. 2 B) and a low-depth dataset of a diploid potato (fig. 2 C). Each strawberry chromosome was mapped using a relatively small population genotyped at low depth. Smooth Descent corrected up to 13% of genotypes, largely correlating with depth so that samples sequenced at lower depth had more genotype corrections. About 3.5% of studied chromosomes had a depth above 10x and had more than 2% of genotypes corrected, an unexpected result probably caused by errors during mapping leading to overcorrection of some samples.

Error rate (%)	Ploidy	IBD method	Δ Error (%)	Δ Correlation	Δ Size (cM)
1	*2*	*hmm*	-0.27	-0.0008	-3.10
1	*2*	*weight*	0.16	-0.0032	-8.27
1	*4*	*hmm*	1.58	-0.0034	-6.05
1	*4*	*weight*	0.69	-0.0013	-7.94
1	*6*	*hmm*	1.76	-0.0031	-7.12
1	*6*	*weight*	0.53	0.0014	-4.70
10	*2*	*hmm*	-4.98	-0.0011	-29.41
10	*2*	*weight*	-5.15	-0.0020	-35.01
10	*4*	*hmm*	-2.94	0.0302	-29.00
10	*4*	*weight*	-2.33	0.0298	-22.44
10	*6*	*hmm*	-1.51	0.2354	-30.88
10	*6*	weight	-0.95	0.2083	-24.64

Table 1: Average change between preliminary map and last iteration of Smooth Descent.

Two error rate cases (0.01 and 0.1) are shown to illustrate the difference between the last iteration of SD and the preliminary error rate (Δ Error), correlation between the true map positions and estimated map positions (Δ Correlation) and map size (Δ Size). All values were calculated as last iteration – preliminary value (positive means increase, negative means decrease). Values are shown for all ploidies and IBD estimation methods (hmm is hidden Markov model and weight is weighted average method).

The dataset of autohexaploid *I. batatas* was used to test SD in a scenario with better genotype accuracy. SD corrected 7.38% of genotypes while maintaining an equivalent relationship between the physical and genetic maps (fig. 2B). This highlights the ability of SD to improve genotype accuracy even in situations where there have not been major issues in defining linkage map.

Lastly, a diploid dataset of potato was genotyped using very low sequencing coverage of ~1.5x, which suggested a low-quality genotypic dataset (Clot et al.

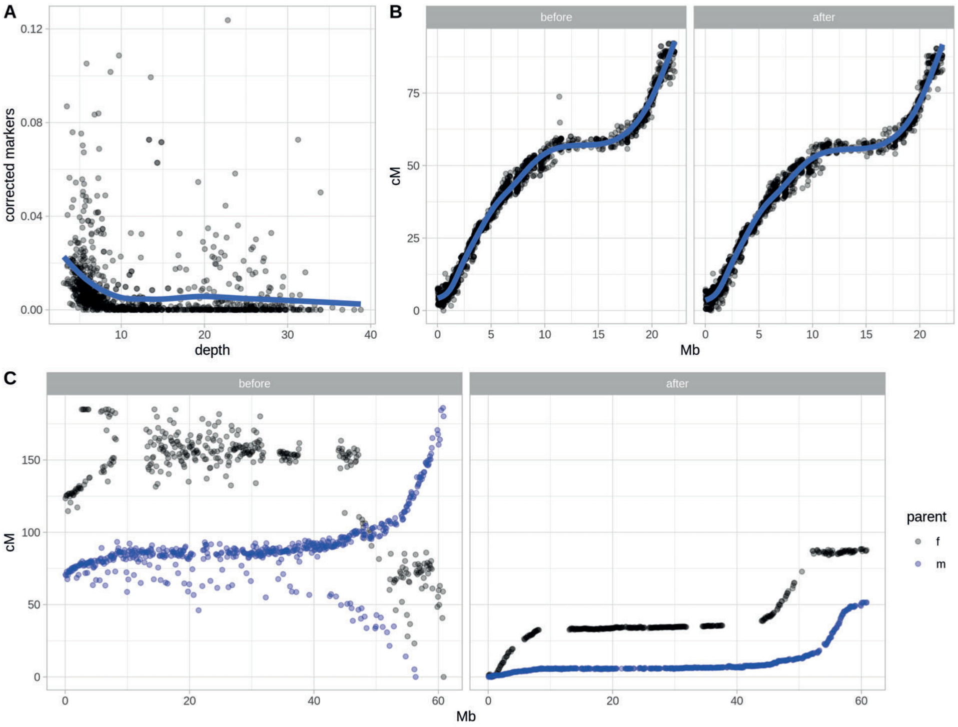


Figure 2: Error detection and marker order in two real datasets after applying Smooth Descent.

A) Relationship between sequencing depth and the rate of markers corrected by Smooth Descent for each chromosome of 52 individuals of strawberry (Fragaria x ananassa). **B)** Relationship between physical and genetic positions of 1513 markers in chromosome 15 of Ipomoea batatas, before and after correcting 7.38% of genotype calls using Smooth Descent. **C)** Relationship between physical and genetic positions of 1716 markers in chromosome 12 of Solanum tuberosum, before and after using Smooth Descent to correct low depth genotypes based on a physical order.

2022). Separate parental maps were generated and each group of markers was corrected using SD with physical order as an input, since a high-quality potato genome sequence was available. The results show a drastic improvement in the correlation between the physical and genetic maps before and after applying Smooth Descent.

Software comparison

The performance of Smooth Descent was compared to two similar software tools: polymapR (Bourke et al. 2018a) and MAPpoly (Mollinari and Garcia 2019). The former performs linkage mapping in polyploids without considering genotyping errors. The latter uses a pre-determined order and a HMM method to obtain new map distances and new genotypes.

In figure 3 we can see the improvements that SD brings. The reconstructed maps have better position correlation and shorter lengths with SD, particularly when the error rates increase. Importantly, only SD changes the order as genotyping errors are corrected, a feature that is clearly useful especially as the error rate and ploidy increases (fig. 3 top left). As expected, higher error rates lead to longer maps when using polymapR, but surprisingly, in MAPpoly that is the case with both very low or very high error rates. Note that polyploid map lengths are much shorter than expected, an issue that is common to polymapR and SD. In terms of genotyping error correction, MAPpoly is better than SD in diploids, but both perform equivalently well in polyploids, except in higher error rates where the HMM of SD is somewhat better. Lastly, the computation time needed for 5 iterations of SD is around 400s in diploids and tetraploids, and around 1000s or 2500s in hexaploids for the weighted average or HMM methods. In comparison, polymapR was always faster, which is to be expected since SD is iteratively running polymapR. MAPpoly time consumption was much higher as ploidy and error rate increased, with very long waiting times in hexaploids.

Overall, SD is better at recovering the correct order and shortening maps regardless of the situation. MAPpoly was better in the diploid scenario in terms of genotype correction and time consumption but became equivalent or worse than SD in tetraploids and hexaploids.

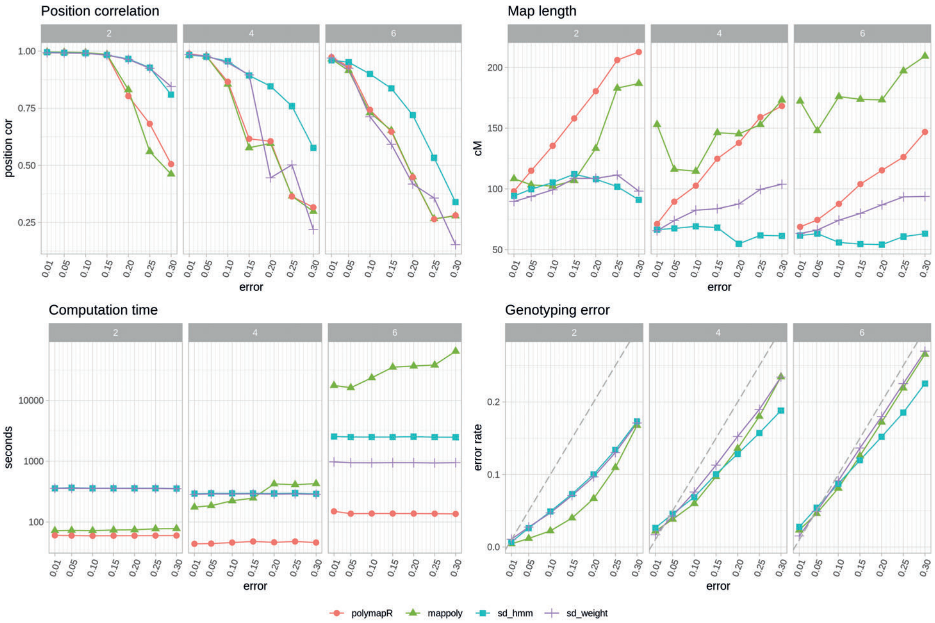


Figure 3: Software comparison with other tools. Average observations of 10 populations per ploidy with simulated genotyping errors.

Top left, correlation between true map position and estimated map position. **Top right**, map length in cM. **Bottom left**, time spent in seconds, note logarithmic y axis. **Bottom right**, genotyping error of corrected genotypes. Grey dashed line indicates the starting error rate. Note that polymapR does not produce corrected genotypes and thus is not included in this panel. Each color and shape corresponds to a different approach: red circle, polymapR, green triangle MAPpoly, blue square (sd hmm) SD with HMM approach, purple cross (s _weight) SD with weighted average approach.

Discussion

In this study we have shown that Smooth Descent is able to substantially reduce genotyping errors, particularly in diploids, and to greatly improve marker order in polyploid linkage mapping, especially using the HMM approach. Moreover, when compared to related tools, SD computes better linkage maps with an equivalent or better level of genotype correction. Our findings are supported by analysis of real data: there was a clear correlation between sequencing depth and estimated genotyping errors in a low-depth strawberry dataset, and an accurate genetic map was obtained after correcting around 7.4% of genotyping errors in hexaploid sweet potato. Thus, we have shown that genotype correction is a useful method to improve linkage mapping in the presence of genotyping errors.

In contrast, the most popular strategy of error management in current genetic mapping software is marker or genotype removal. In JoinMap this is achieved through a Bayesian parameter (Liu et al. 2014), while Lep-Map2 does so through a Hidden Markov Model (HMM) (Broman et al. 2003; Rastas et al. 2016). GUSMap, on the other hand, does not remove errors but compensates their impact in map length, also within an HMM model (Bilton et al. 2018). Finally, HighMap uses SD's predecessor SMOOTH (van Os et al. 2005), and thus could benefit from the developments presented here (Liu et al. 2014).

The genotype correction approach presented in this article depends on transmitting confident parental information to uncertain offspring genotypes. Essentially, if most local markers indicate that one chromosomal region of a parent has been inherited, the offspring genotypes should match parental haplotypes. This rationale, and therefore the accuracy of SD, depends on two important factors: marker order and parental phasing.

Marker order

The set of local markers used to identify wrong genotypes is clearly defined by marker order. It is not crucial that marker order is exact, but the overall preliminary order should be correlated to the true order. In instances where the provided preliminary order is very far off from the true order, SD will not be able to impute genotypes correctly and any map improvement will be spurious.

Marker order can be determined by a linkage mapping procedure where a measure of linkage and an ordering algorithm is used to obtain a genetic map. In our implementation of SD these correspond to polymapR (Bourke et al. 2018a) and MDSMap (Preedy and Hackett 2016) respectively. Both processes are sensitive to genotyping errors, meaning that as errors increase, the accuracy of the estimated linkage map will decrease. Consequently, there is a natural upper limit to the level of genotyping error that SD can tolerate: once the error rate impedes the calculation of a relatively good preliminary genetic map, SD stops being useful. This also means that if different methods were designed that could compute marker orders independently of genotyping errors, SD applicability would be expanded.

Linkage mapping is not the only way to determine marker order. As reference genomes are built, it is increasingly common to obtain physical positions for markers. If such information is available, one could apply SD using physical, instead of genetic positions. This opens the possibility of using SD to datasets that are too large to be mapped using linkage techniques, but that could benefit from an error-cleaning algorithm. Moreover, since the order would not need to be re-calculated after genotype correction, only a single iteration of the algorithm would be necessary. Nevertheless, particularly for the weighted-average IBD estimation procedure, the usage of physical positions rather than genetic positions could be problematic since physical distances do not represent the same recombination probabilities along the genome. In centromeres a distance of 100,000 bp will include less recombinations than 100,000 bp in the chromosome arms. This should not be a major problem in the application of SD though, since the recombination frequencies are used relative to each other within small local intervals. Furthermore, if the locations of the pericentromeric regions are known (which they often are), then it would be possible to generate pseudo-cM positions of markers to circumvent this issue.

Parental phasing

To calculate identity-by-descent (IBD) probabilities, the backbone of genotype error detection and correction in SD, accurate parental phases or parental haplotypes are required. In this study we have not aimed at characterizing the effects of parental phasing in SD, as there has been much research dedicated to this complex issue (Browning and Browning 2011b; He et al. 2018; Al Bkhetan et al. 2021), both in diploids and in polyploids. Currently, there are two types of approaches that can be used to establish parental phasing: based on marker scores or on sequence reads.

Marker scores have been used within several Hidden Markov Models (HMM) to obtain accurate phases. Recent studies in diploid data showed that consensus haplotyping approaches are the most accurate (Al Bkhetan et al. 2021), although individually tools like SHAPEIT4 (Delaneau et al. 2019) and BEAGLE5 (Browning et al. 2018) have the best performances in terms of time efficiency and accuracy. Several HMM have also been developed focused on polyploid data which can estimate phases: MAPpoly (Mollinari and Garcia 2019), polyOrigin (Zheng et al. 2021), and polyqtlR (Bourke et al. 2021). Although many of these methods consider genotyping error in their estimations, since phasing depends on marker segregation, an increased genotyping error rate in the target population can decrease phasing accuracy.

Alternatively, reads can be used to perform haplotype assembly: by observing multiple polymorphisms in a single read one can infer the most likely haplotype phases. Multiple tools have been developed to produce long-range haplotypes using short reads, long reads or a combination of both (Garg 2021). In diploids, WhatsHap (Patterson et al. 2015) and HapCut2 (Edge et al. 2017) are the most popular methods, being able to produce chromosome-level haplotypes when combining short and long read data (Garg 2021). In polyploids, the assembly problem is more complex, which has required the development of specific tools such as HapCompass (Aguiar and Istrail 2012), HapTree (Berger et al. 2014) and SDhaP (Das and Vikalo 2015). Although useful, the accuracy of these tools is quite variable depending on depth and ploidy (Motazedi et al. 2017), never reaching the performance of their diploid counterparts. More recent developments like WhatsHap polyphase (Schrinner et al. 2020), based on long-read sequencing or Hap10 (Majidian et al. 2020), oriented to link-read data, are promising in closing the gap between diploid and polyploid haplotype assembly.

Application of Smooth Descent

The original idea behind the development of SD was to create a tool that would be able to utilize low-depth, inaccurate genotypes to obtain accurate linkage maps. Intuitively, we expected that confident parental phasing would be enough to create such an approach. We have shown that indeed, if parental information is accurate and marker order is well established, genotype correction can be performed, and accurate linkage maps obtained. Thus, we can imagine the following genotyping setup for an F1 population. First, the two parents are sequenced at high depth using long-read sequencing, in order to compute parental haplotype phases. Secondly, the F1 population is genotyped using low-depth short reads. If a marker order is not established yet, SD can be used iteratively to improve genotypes and obtain an accurate linkage map. Otherwise, a single iteration of SD is used to eliminate as many genotyping errors as possible. If the marker number is relatively small, the HMM method of SD is applied, if the dataset is larger the more efficient, although less accurate, weighted average method is used. Finally, a set of corrected genotypes is obtained. In this manner, SD would reduce genotyping costs by allowing a lower depth of sequencing in the F1 offspring.

Overall, SD is a simple and informative software tool. It estimates IBDs, calculates error rates per marker and individual and can impute corrected genotypes. Our implementation, together with MDSmap (Preedy and Hackett 2016) and polymapR (Bourke et al. 2018a) allows SD to work in multiple ploidies and with large datasets. We also provide many visualization tools which will help uncover the hidden information within genotyping data and turn Smooth Descent into SMOOTH's descendent.

Declarations

Data availability

The data used on this article can be found on FigShare: https://figshare.com/articles/dataset/Smooth_Descent_results/20038589

Code availability

R package SmoothDescent can be found on github: github.com/Alethere/SmoothDescent

Author's contributions

ATN: designed software, performed research, drafted and reviewed manuscript. PB, EvdW: proposed research, supervised research, reviewed manuscript. CC: contributed to manuscript and results. PA: obtained funding, supervised research, reviewed manuscript. RF: supervised research, reviewed manuscript. CM: proposed research, obtained funding, supervised research, reviewed manuscript.

Acknowledgements

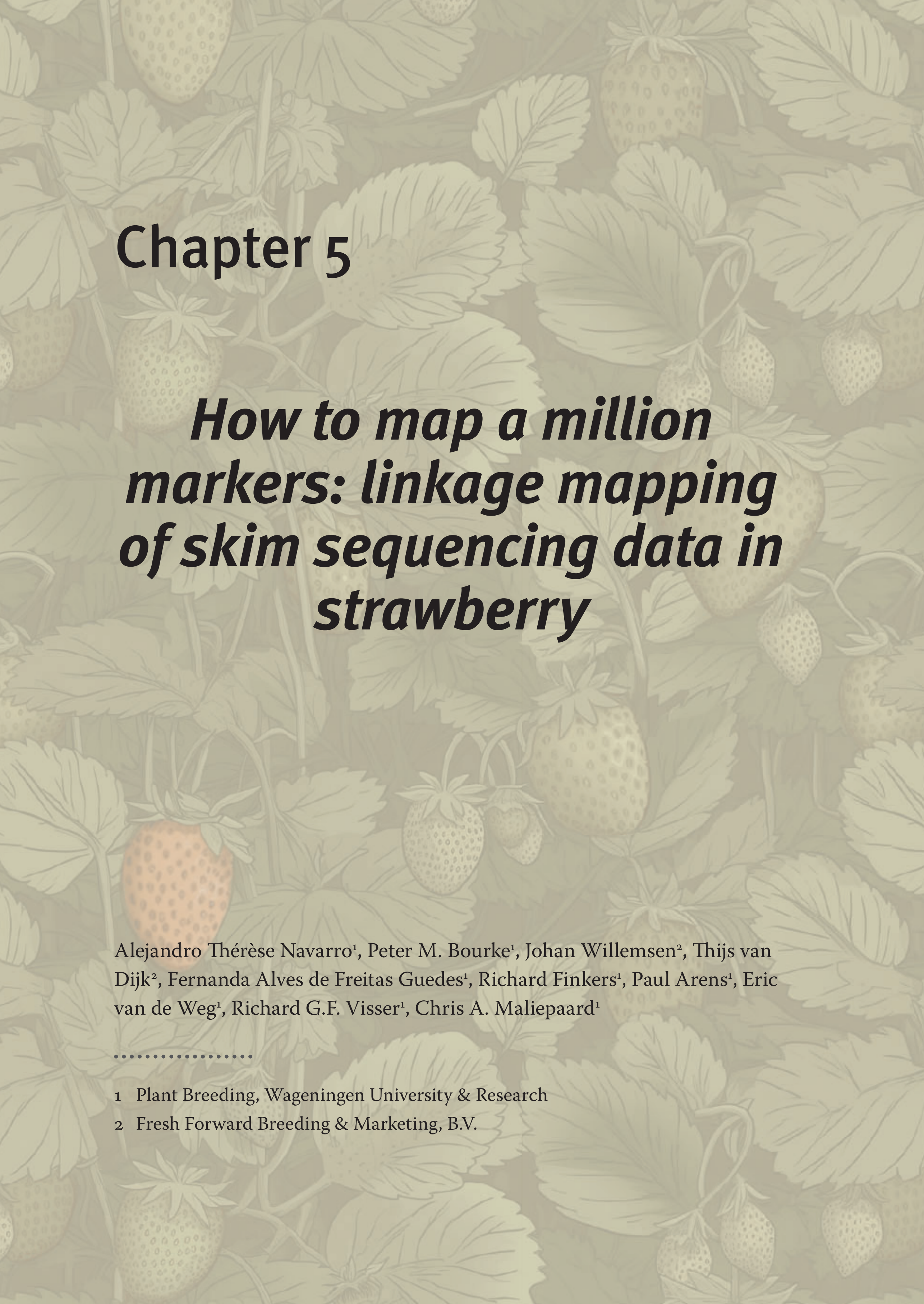
The authors would like to acknowledge the help of the breeding company Fresh Forward by providing the data that spurred the development of this approach.

Competing interests

The authors have no conflicts of interest regarding this article.

Funding

This work was supported by the TKI projects LWV20.112 Application of sequence-based multi-allelic markers in genetics and breeding of polyploids (BO-68-001-042-WPR) LWV20.085 Genetics Assisted Assembly of Complex Genomes (BO-68-001-033-WPR), Novel genetic and genomic tools for polyploid crops (KV 1605-020/TU-16006; BO-68-001-001), and by USDA's National Institute of Food and Agriculture (NIFA) Speciality Crop Research Initiative project 'Tools for genomics- assisted breeding in polyploids: Development of a community resource' [2020-51181-32156/SCRI].



Chapter 5

How to map a million markers: linkage mapping of skim sequencing data in strawberry

Alejandro Thérèse Navarro[1], Peter M. Bourke[1], Johan Willemsen[2], Thijs van Dijk[2], Fernanda Alves de Freitas Guedes[1], Richard Finkers[1], Paul Arens[1], Eric van de Weg[1], Richard G.F. Visser[1], Chris A. Maliepaard[1]

1 Plant Breeding, Wageningen University & Research

2 Fresh Forward Breeding & Marketing, B.V.

Abstract

Next generation sequencing technologies are revolutionizing the way in which we study genotypes. They provide much larger genotype datasets than SNP arrays, without the problem of ascertainment bias. However, they also bring with them a higher error rate. Data volume and this increased error rate are both big problems for linkage mapping, a technique that remains a crucial source of information of marker location in the genome. Especially in complex crops like allopolyploid strawberry, containing multiple highly homologous subgenomes, it is possible that sequencing cannot provide data of high enough quality (at a reasonable cost) to produce linkage maps as good as those obtained with the more accurate SNP array genotypes. In this study we aimed to produce such linkage maps, starting from whole genome skim resequencing data of 48 strawberry (*Fragaria* x *ananassa*) individuals of the Holiday x Korona cross. By the combination of binning, iterative error correction and the physical positions of markers, we were able to produce linkage maps of 27 of the 28 chromosomes of comparable quality to a linkage map obtained with SNP array technology. We placed 1.85M markers in 2434 unique positions and detected a small portion of the genome that is likely to have been wrongly phased during the subgenome assembly. With this, it is clear that even in such a complex situation as an allo-octoploid crop, skim sequencing can be used to produce good quality linkage maps, provided that adequate analytical techniques are applied.

Keywords

strawberry, whole-genome sequencing, high-density linkage map

How to map a million markers: linkage mapping of skim sequencing data in strawberry

Introduction

In the year 2000, Genbank stored 11.1 Gbp of sequences; by 2010 that number had increased tenfold (112.1 Gbp) and it rose to another order of magnitude by 2021 (1053.3 Gbp) (NCBI 2022). These numbers reflect an undeniable rise in the popularity and usage of sequencing data: every day we sequence more biological samples. With this transition, we move away from SNP arrays and the well-known ascertainment bias they introduce (Lachance and Tishkoff 2013; Geibel et al. 2021). Combined with lower prices, longer reads and increased accuracy, sequencing technologies are certain to become -if they are not already- the backbone of modern genetic research. As this technological shift unfolds, many analytical techniques originally designed for marker-based data must be revisited and adapted to handle the new properties of sequence information: very large datasets with potentially higher error rates.

Linkage mapping, the grouping and ordering of markers based on recombination frequencies, has allowed the study of chromosomal structure since its first application in *Drosophila* (Sturtevant 1913). However, Sturtevant only had to order 6 polymorphic markers, while modern sequencing-based genotyping studies routinely discover hundreds of thousands or millions of polymorphisms. With more than a dozen markers it is impossible to compare all possible marker orders to find the optimal solution, since the number of possible orders is prohibitively large (1.21x10^{18} for 20 markers). Instead, several algorithms have been proposed to find approximate solutions for hundreds or a few thousand markers (Cheema and Dicks 2009; Liu et al. 2014; Rastas et al. 2016; Preedy and Hackett 2016). When marker numbers rise beyond the thousands, and with limited sized populations, many markers segregate identically and become redundant, a property that can be exploited to reduce the effective number of markers used during mapping (Rastas 2017). Thus, even when millions of markers are genotyped, only a small subset of these need to be directly considered for mapping.

Thus far, genotyping errors are more common in sequencing data than in other genotyping technologies, an issue with well-described impacts in linkage mapping (Hackett and Broadfoot 2003; van Os et al. 2005; Cartwright et al. 2007). This problem is exacerbated when sequencing with low depth, or skim sequencing, the main factor influencing sequence-based genotyping accuracy (Chan et al. 2016; Gerard et al. 2018). In linkage mapping studies, marker removal is the most usual way to deal with erroneous genotypes (Lincoln and Lander 1992; van Os et al. 2005; Cartwright et al. 2007; Wu et al. 2008; Cheema and Dicks 2009; Liu et al. 2014; Rastas et al. 2016). It would then seem that the advantage of genotyping hundreds of thousands of markers is then countered by the need to remove many of them due to genotyping errors. An alternative is to use imputation or genotype correction to improve sequencing-based genotypes (Chan et al. 2016; Torkamaneh et al. 2018; Zheng et al. 2018; Malmberg et al. 2018; Thérèse Navarro et al. 2022). However, it remains unclear whether such an approach would provide the same level of linkage mapping accuracy that can be achieved with SNP arrays.

Strawberry has been a challenging crop to work with, particularly regarding linkage mapping. Due to its allo-octoploidy, it is difficult to distinguish the subgenome origin of markers based on sequence, and segregation information is also difficult to interpret (Edger et al. 2018). In spite of this, several linkage maps have been produced over the years (Rousseau-Gueutin et al. 2009; Spigler et al. 2010; van Dijk et al. 2014; Tennessen et al. 2014; Sargent et al. 2015; Hardigan et al. 2020), with varying nomenclature. Recently, a subgenome-resolved genome assembly has been published (Edger et al. 2019). With it, it becomes possible to use read mapping as a means to discover markers in strawberry. However, due to high sequence similarity between subgenomes, it is unclear whether this approach will yield markers useful for linkage mapping.

In this study we used whole genome skim resequencing of an F1 population of strawberry to produce linkage maps. With this setup we aimed to answer a multifaceted question: is it possible to use whole genome skim resequencing data (instead of SNP array data) to generate linkage maps in an allopolyploid crop? The issues of data-volume, sequencing errors and mis-mappings due to subgenome homology will need to be overcome. Thus, our analysis provides both a protocol to use this type of data as well as an evaluation of the quality that can be expected. Finally, we contribute to the growing number of analytical approaches to use sequencing data in genetic studies.

Materials and methods

Marker discovery

Strawberry varieties Holiday and Korona were crossed and 46 individuals of their offspring, as well as the parents, were whole genome resequenced using Illumina 150 paired-end technology. Sequencing depth (reads bp per haplotype bp) was variable across samples, between 25x and 5x. Variant discovery was performed de novo by aligning reads against the Camarosa v1.01.a reference genome (Edger et al. 2019) using the Burrows-Wheeler Aligner (Li and Durbin 2009). Using the program bcftools (Li 2011) 10.24M polymorphic sites were detected. Markers with average read counts below 20 reads per individual or not segregating in the progeny were discarded, resulting in 4.04M kept markers. Allelic read counts were extracted and the R package updog was used to call genotypes (Gerard et al. 2018), which estimated diploid genotype scores as well as miss-genotyping probability and other parameters.

Binning and linkage group assignment

Markers were grouped in bins of co-segregating markers, that is, markers with identical genotypes across individuals. Note the difference with other binning approaches where markers are grouped by estimating recombination points (Rastas 2017). The number of markers within a bin was used as a criterion to filter erroneous genotypes. We filtered those with <40 markers. This value was obtained empirically during our research but can be derived by calculating the number of expected recombinations and segregation types per chromosome.

For example, in chromosome 1A we initially detected ~127K markers. Assuming that only 5% of markers have some genotyping error, we expect 120.65K to be correctly genotyped. We expect between 1 and 3 recombinations per individual in each chromosome pair, with 46 individuals that is equivalent to 46 to 138 recombination bins. Since strawberry is a heterozygous diploid, each recombination bin can contain three different segregation marker types. Thus, 120.65/46/3 = 0.87K markers per bin if there is one recombination per individual, 0.29K markers per bin if there are three recombinations. Thus, we could expect between 870 and 290 markers per bin to contain only correctly typed markers. This can be generalized into the following formula:

$$expected\ binsize = \frac{m(1-\varepsilon)}{nr(s)}$$

Where m is the number of markers, ε is the expected error rate, n is the number of individuals and r the number of recombinations per individual, and lastly s is the number of segregation marker types, which will depend on ploidy and heterozygosity. Due to a large variation in the number of markers per chromosome, in our case the bin sizes ranged between 130 for chromosome 7C and 1601 for chromosome 6C. This calculation assumes that markers and recombination points are homogeneously distributed across the chromosomes, which is far from true. This may explain why despite choosing a much lower value than our lower margin of 130 we still obtained good results.

Bins were assigned to chromosomes by analysing the physical position of markers within each bin. A binomial test was used to assign bins to a chromosome if at least 70% of the markers within a bin belonged to the same chromosome in the physical sequence. Bins that failed to be assigned were discarded. The true region of a bin was computed using the range of marker positions within the assigned chromosome, thus defining the *true bin region* that was used further in the study.

Conflict marker analysis and blast test

After assignment to a chromosome, some markers showed a conflict between physical and genetic information: while the bin had been assigned to one chromosome, some of its component markers originated from other genome sequences. These were named *conflict markers*. To test the origin of these conflict markers we proceeded to analyse the homology between the sequences around each conflict marker and the overall bin sequence for their respective bins.

For each conflict marker, a sequence of 500bp was extracted (250bp at each side of the polymorphism). Each sequence was then blasted against the whole Camarosa genome using default parameter settings. If a hit of any quality was found within the range of the true bin region plus an additional 100kb on each side, the conflict was considered as “blast positive”, meaning the sequence of the conflict marker shares homology with some portion of the corresponding bin sequence.

Linkage mapping and error detection

Bins were mapped using a two-step approach. First, preliminary maps were built using polymapR (Bourke et al. 2018a). Those maps were used to obtain parental marker phases by analysing the segregation of simplex markers for each parent, as described in the vignette of polymapR. The preliminary maps and parental phases obtained were then used to apply Smooth Descent, a tool that is able to correct the order and estimate genotyping errors based on identity-by-descent probabilities (Thérèse Navarro et al. 2022).

All bins assigned to the same chromosome were mapped as one linkage group. Smooth Descent was iterated 10 times, and the iteration with the highest R^2 between pairwise recombination frequencies and pairwise marker distances was taken as the best iteration. Through this iterative process a set of corrected genotypes was obtained. Comparing the initial genotypes and the corrected genotypes allowed us to estimate the number of genotyping errors per chromosome and individual.

A new set of bins was obtained by joining those bins that, after genotype correction, contained the same genotypes across all individuals (*i.e.* were identical). These new set of bins was mapped to obtain the final genetic maps. The physical range of each bin was computed using the 5% and 95% position quantiles to minimise the influence of potential outliers. The final bin and map composition is shown in supplementary data 1 and 2.

SNP array map

The obtained map was compared with a genetic map developed in the same population using SNP array-derived genotypes of 50 individuals, the 46 individuals used for skim-sequencing plus an additional four individuals. Since an unambiguous relationship between the SNPs of the SNP array dataset and the genotypes used in this study could not be established, comparison between the two maps was based on genotype correlations between markers. Two markers were considered equivalent, and thus amenable to be placed in the same genetic location if the correlation between their genotypes exceeded 95%.

Results

Marker filtering and bin analysis

The polymorphism discovery pipeline found 10.24M markers in the Holiday x Korona population when compared to the Camarosa reference genome. Several criteria were used to filter low-quality markers: read count below an average of 20 per individual, mis-genotyping probability above 5% and not

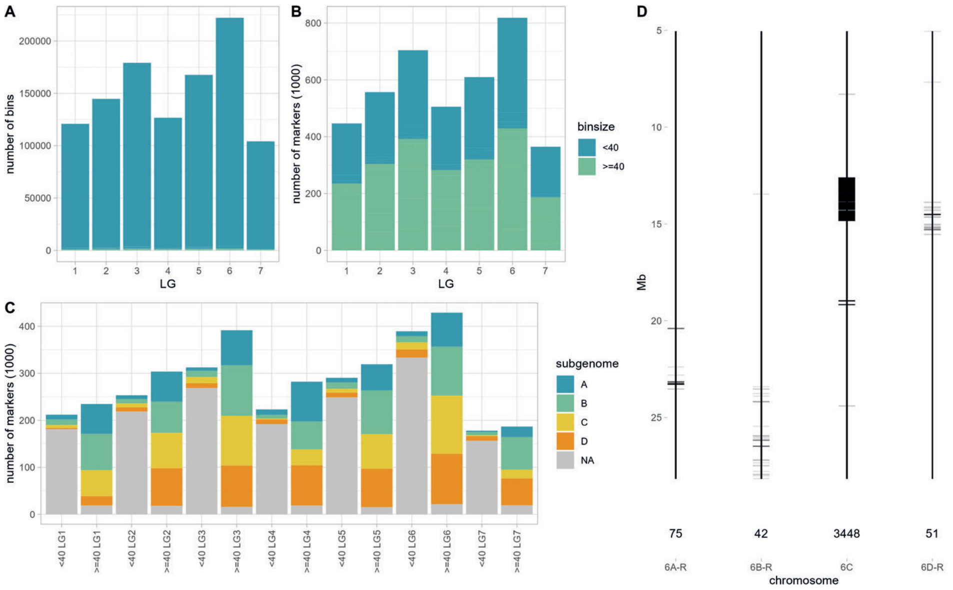


Figure 1: Overview of co-segregating marker bins.

A) Number of bins found within each linkage group. A very small number of bins contain ≥40 markers (green portion of each bar is barely visible). **B)** Number of markers found in bins of <40 markers (blue) or ≥40 markers (green) per linkage group. About 50% of markers within each linkage group are found in bins with ≥40 markers. **C)** Number of markers that can be clearly assigned to a subgenome split according to bin size. **D)** Example of a single bin of co-segregating markers, physical positions of contained markers. The number above each sequence label indicates the number of markers in that sequence.

segregating in offspring. After this process, 4.05M markers were retained. Co-segregation bins were subsequently computed, which resulted in 1.06M bins of uniquely segregating markers.

Bin size, the number of markers within each bin, was greatly variable. The number of markers in a bin can be used as an indicator of quality since error-containing bins are expected to be smaller. We obtained a lower threshold by roughly estimating the number of expected markers of the same segregation within a bin delimited by recombination events. We chose to consider

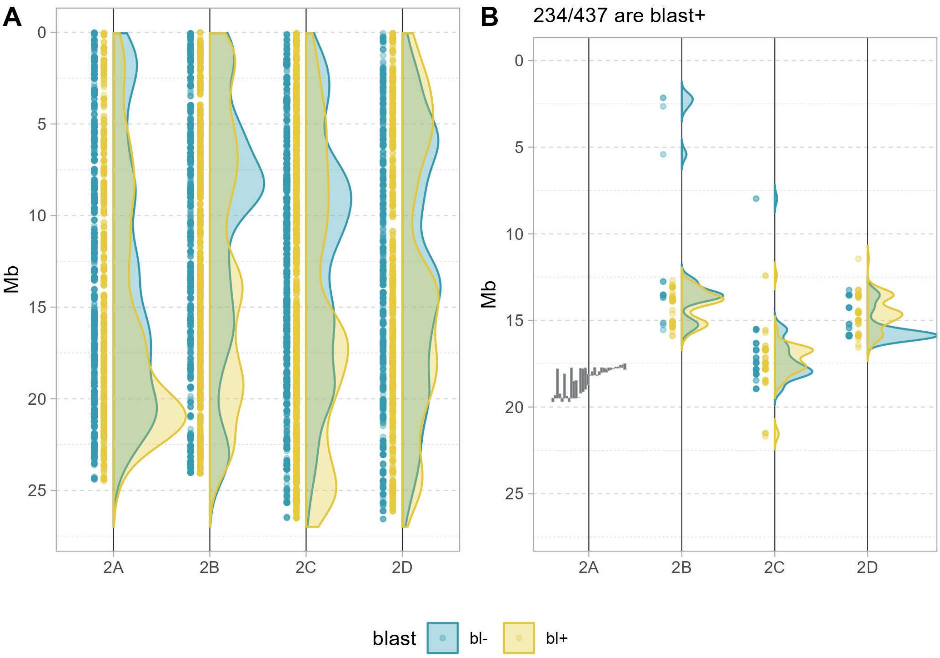


Figure 2: Distribution of conflict markers on linkage group 2.

Markers with homology to the true bin region are shown as bl+, and markers without homology are shown as bl-. **A)** Overall distribution of conflict markers across all subgenomes of linkage group 2. **B)** Distribution of conflict markers identified in bins located in a specific area of chromosome 2A.

confident bins those with 40 markers or more. A large proportion of bins contained less than 40 markers (fig. 1A). Of the 4.05M markers, about half was found in the 0.76% bins with more than 40 markers (fig. 1B). This puts into question the reliability of the rest of the markers, since we expect a high number of identically segregating markers due to an oversaturation of polymorphisms, it is unlikely that any marker has less than 40 replicates (see Methods for details).

Once markers are grouped into bins of co-segregation, we assume that each bin corresponds to a unique position in the genome. Each marker within a bin had a position in the *F.* x *ananassa* genome “Camarosa”, and as such we expected co-segregating markers to all originate from a single region of a single sequence. However, in many bins that was not the case, with markers assigned to different chromosomes (often different subgenomes of the same linkage group) being binned together (fig. 1D). A proportion test was used to assess whether a bin was formed of more than 70% of markers belonging to a single chromosome. The result showed that most bins containing ≥40 markers could be assigned to a sequence, while most bins <40 markers could not be confidently assigned, *i.e.* less than 70% of markers belonged to the same chromosome (fig. 1C). The following analyses were performed using 6567 bins (representing 2.01M markers) which contained ≥40 markers and could be unequivocally assigned to one chromosome.

Origin of conflict markers

The retained bins were composed of a large proportion of markers from one sequence but with a few markers (average of 5%) originating from the homoeologous chromosomes (fig. 1D). We hypothesized that those out-of-place markers, the conflict markers, corresponded to wrongly mapped reads due to a high sequence similarity between the true bin region (where most markers within a bin are found) and the sequence around the conflict marker on another subgenome. A blast search was performed, which showed that across the whole genome, only 50.11% of conflict markers shared sequence similarity with the true bin region (blast positive, bl+), indicating that at least half of the cases could not be explained due to read mapping to homologous regions of a different subgenome. There was no particular distribution of blast positive or blast negative (bl-) markers across the genome (fig. 2A) nor within specific bins (fig. 2B). This suggests that the conflict markers cannot be easily attributed to mis-mapping. Nevertheless, we removed the conflict markers from further analysis, since whatever the cause, their physical positions in the genome sequence were not reliable. In total, 1.85M markers remained, distributed across 6567 bins of co-segregating markers.

Error correction

On average, 1.8% of all genotype calls were identified as genotyping errors across the whole dataset. However, error rates were highly variable among

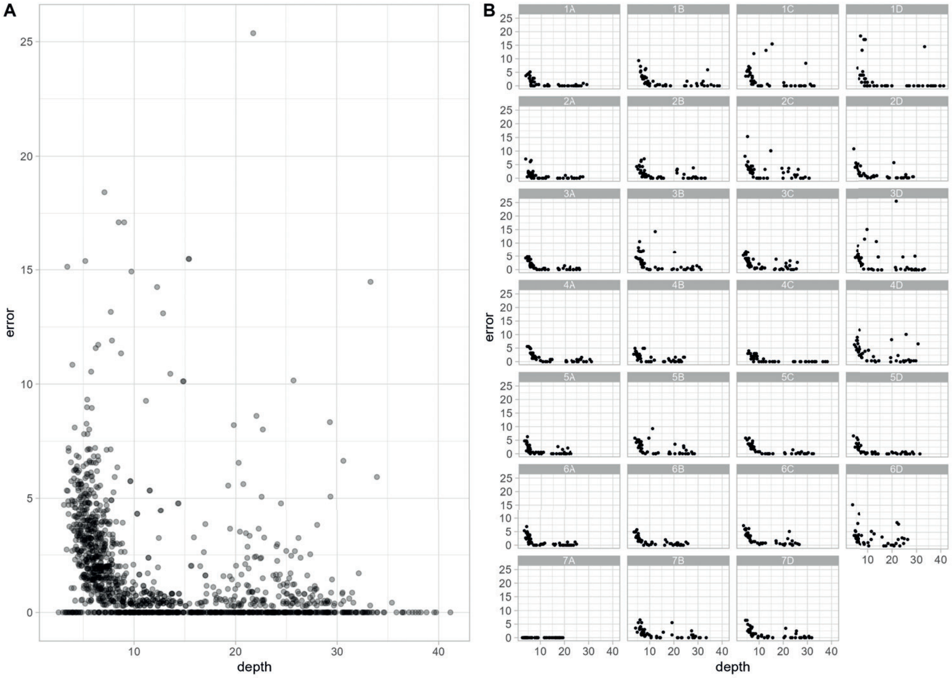


Figure 3: Relationship between sequencing depth and error rate per chromosome and individual.

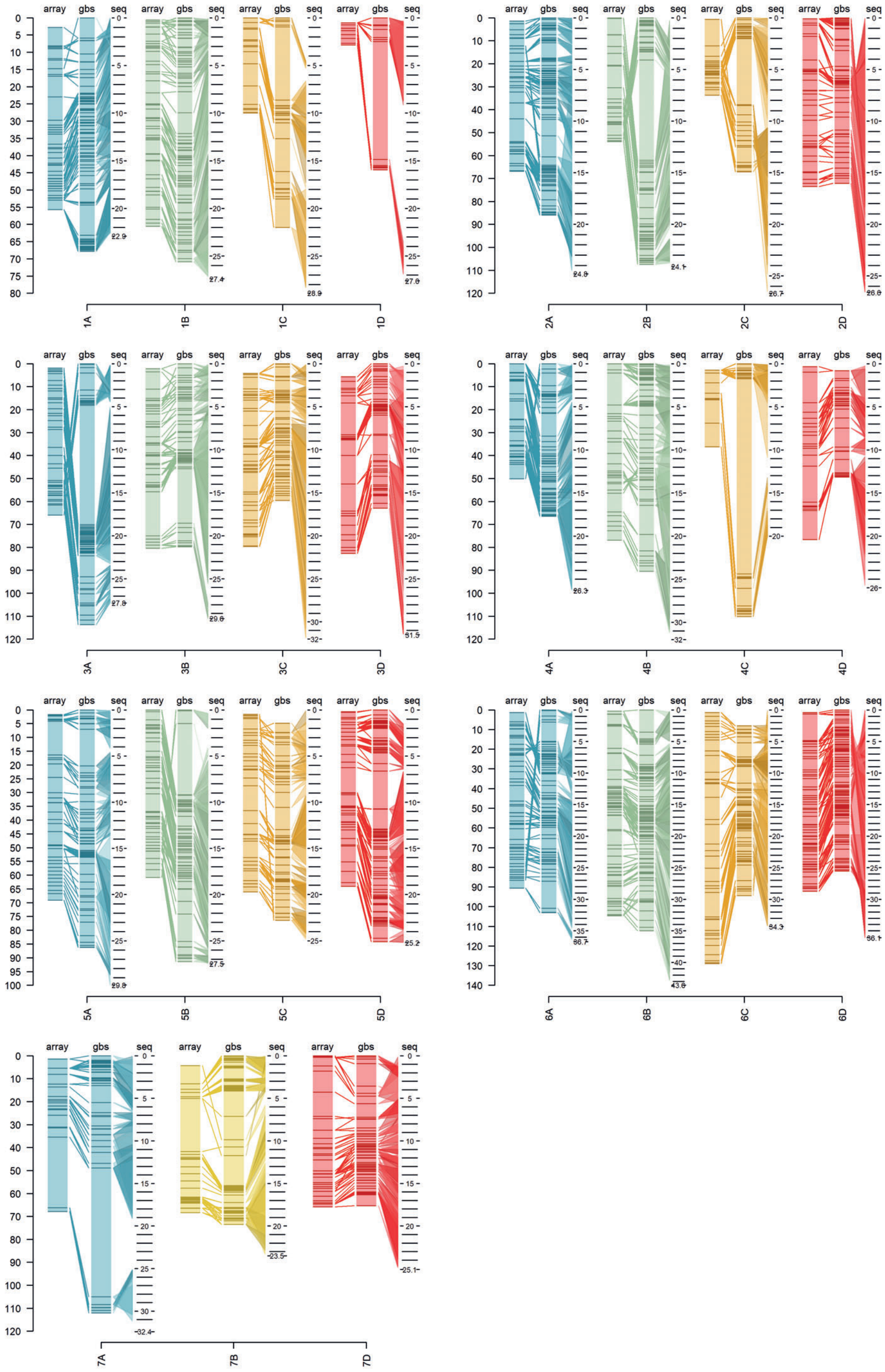
A) Overall depth to error distribution. **B)** Depth to error per individual, plotted separately for each mapped chromosome. Note the absence of 7C which could not be mapped. The lack of errors of chromosome 7A is due to the choice of no correction as the best possible map according to the Smooth Descent algorithm.

individuals depending on their sequencing depth (fig. 3). Over all chromosomes, individuals with an average sequencing depth below 10x had elevated error percentages. There are also some chromosomes that, particularly for a few individuals, seem to have unusually large percentages of error, e.g. ~25% in 3D or~20% in 1D. It seems likely that these errors are due the relatively small mapping population and regions of low marker density that cannot be adequately mapped. Nevertheless, the overall relationship between genotyping errors and depth is, as expected, abundantly clear. It is noteworthy that the markers were already filtered, and thus the observed genotyping error rates do not reflect the true number of genotyping errors produced, but rather those errors that were kept after filtering based on marker quality parameters.

Linkage maps

Using Smooth Descent, errors were detected and corrected, which allowed to re-compute the number of bins. Finally, the 1.85M markers were grouped into 2434 bins and mapped into 27 chromosomes (7C could not be mapped due to a lack of polymorphisms in one parent). The obtained map (sequencing map) was compared with the Camarosa genome assembly (physical map) and with a linkage map of the same population produced using SNP array genotypes (array map). The collinearity between the sequencing map with the array and physical maps was on average 0.76 and 0.71 respectively (fig. S1). The physical sequence and array maps show the largest discordance, with an average collinearity of 0.64. The most common errors in the sequencing-based map are large gaps due to low marker density (e.g fig. 4 chromosomes 4C, 5B, 7A), and small inversions of a few markers (e.g. fig. 4, 0-20cM of 3A, 15-30cM of 6A). Nevertheless, it seems that even though there was an abundance of genotyping errors the produced genetic maps have a similar quality to the array-based genetic map.

The comparison between sequences helps us pinpoint a few chromosomes that were likely wrongly assembled in the reference genome: the sequencing and array maps are largely collinear, while the physical sequence is not (fig. S1). Such is the case for chromosomes 2C, 2D and 6D. This might have to do with the reference variety used to assemble the genome, Camarosa, since it is known that it harbours a translocation on chromosome 6D (this has been fixed in the following genome version, Royal Royce). For 2D a similar situation might be the case. However, the linkage maps comparisons show that for 2C, even if the array and sequencing maps are mostly collinear, there is still quite some disagreement between the two maps. The difficulty on mapping this chromosome might be due to a lack of markers in a relatively large region, resulting in a gap in the sequencing-based linkage map (fig. 4).


array
gbs
seq
1A
1B
1C
1D
2A
2B
2C
2D
3A
3B
3C
3D
4A
4B
4C
4D
5A
5B
5C
5D
6A
6B
6C
6D
7A
7B
7D
22.9
27.4
28.9
27.0
24.8
24.1
26.7
26.8
27.8
29.6
32
31.5
26.3
32
26
29.8
27.5
25
25.2
36.7
43.6
34.3
36.1
32.4
23.5
25.1

Discussion

We were able to produce linkage maps using low-depth whole genome resequencing data, with a quality comparable to linkage maps produced using SNP array technology. The crucial difference, however, is the final number of mapped markers: 13384 in the SNP array map (in 1587 bins) and 1.85M in the sequencing-based map (in 2434 bins). With this there is a clear confirmation that, even when using low-depth data, sequencing can substitute SNP arrays in the linkage mapping arena. However, there are a couple of caveats to be considered when using sequencing data.

Imputation and genotyping errors

A main driver of cost in sequencing-based genotyping of a population is depth. Consequently, low-depth approaches, also known as skim sequencing, have been proposed and applied in a variety of settings (Malmberg et al. 2018; Kumar et al. 2021; Adhikari et al. 2022). A main consequence of skim sequencing is the large proportion of missing or incorrect genotypes obtained, due to low coverage and incomplete allele sampling. To address this issue, imputation has been shown to be highly effective, particularly in populations with low genetic diversity such as biparental populations (Xu and Bai 2015; Chan et al. 2016; Fuentes-Pardo and Ruzzante 2017; Torkamaneh et al. 2018; Zheng et al. 2018; Malmberg et al. 2018; Thérèse Navarro et al. 2022). What remains unclear is the level at which a sample can be considered “skim sequenced”. While some researchers claim between 1x and 2x to be sufficient for diploids, others

Figure 4: Comparison between linkage maps.

For each chromosome, three maps and their relationships are shown. The left band represents the SNP array-based map, the middle band is the sequencing-based map produced in this study (note that each marker represents a bin of co-segregating markers), the rightmost band indicates the length of each chromosome sequence in the Camarosa genome. The relationship between sequencing map (gbs) and chromosome sequence (seq) shows the range of physical positions within a bin of co-segregating markers.

point towards a per-sample average of 5x (Malmberg et al. 2018; Adhikari et al. 2022). The true required depth is possibly related to genetic diversity of the population in question: higher heterozygosity and rare variants decrease the accuracy of imputation (Malmberg et al. 2018). Naturally, baseline sequencing levels should also be increased as the ploidy of the organism increases.

Notwithstanding the usefulness of imputing missing genotypes, erroneous calls are also a crucial aspect of low-depth sequencing data. Most imputation tools reach accuracies above 80% under different circumstances, but they only take into account imputation of missing values, not correcting wrongly scored genotypes (Malmberg et al. 2018). A genotyping pipeline that hinges on skim sequencing should also consider methods to detect and correct erroneous calls. Detection can be based on genotyping quality scores provided by the genotype calling software. Alternatively, a specific algorithm to detect and correct erroneous genotypes can be applied. Some such methods have already been developed and show significant promise (Money et al. 2015; Zheng et al. 2018; Browning et al. 2018; Thérèse Navarro et al. 2022).

Linkage mapping with millions of points

Data volume is the second aspect that critically defines sequencing data and genotypes derived from it. Where SNP array datasets often produce somewhere between 10k and 100k SNPs usable for downstream analysis, sequencing-based datasets produce millions of points. Importantly, this allows to filter the data stringently, only keeping the most reliable genotypes. In our study, this meant moving from the initial 10.24M SNPs discovered to the final 1.85M SNPs in the linkage map. On the other hand, such staggering numbers of data-points easily overwhelm genetic analytical methods. The case is quite clear in linkage mapping, where no more than a few thousand markers can be ordered in a reasonable amount of time using modern algorithms. To deal with this limitation, we reduced the dataset using the co-segregation binning approach described, similar to that proposed by (Rastas 2017). Other methodologies of which the time cost scales based on the number of markers suffer from a similar problem, e.g. GWAS analyses. The latest developments in the field deal not as much with increased power as they do with increased performance in terms of time and memory efficiency (Liu et al. 2016; Huang et al. 2019). Thus, the surprisingly voluminous flow of data that sequencing provides could as well be a wellspring of knowledge as a flood of noisy information.

A reference genome can greatly influence the outcome

In most cases, sequencing data are analysed by comparing them to a reference genome. What such a reference contains can vary greatly: a genome sequence of a single individual, a linear consensus of multiple genomes or even a pan-genomic set of sequences (Hurgobin and Edwards 2017; Marschall et al. 2018; Tao et al. 2019). Alternatively, reference-free marker discovery methods can be used to avoid reference bias (Leggett and MacLean 2014), although these methods usually do not work well in allopolyploids due to high sequence similarity between subgenomes (Edger et al. 2018). In an organism like strawberry, where four subgenomes inhabit the nucleus, the construction of a subgenome-phased reference genome (in which the four subgenomes have already been separated) seems almost indispensable for sequencing-based genotyping. In that regard, there has been some concern that high sequence similarity would prevent accurate read mapping in allopolyploids. In strawberry, this does not seem to be a problem: only a small fraction of markers (~5%) were putatively the result of mis-mapping, and of those, only half could be explained by mis-mapping. We think it is more likely that such conflict markers were the result of incorrect subgenome phasing of small sequences during genome assembly. Some of these errors, such as the large translocation in 6D, have been addressed in the Royal Royce genome sequence (Hardigan et al. 2021a). However, it seems evident that further improvements on the genome assembly of *F.* x *ananassa* can still be achieved, especially now that a solid subgenome-phased blueprint is available.

Computational resources become necessary

Finally, although skim sequencing generates large datasets for a small price per sample, it requires a substantial computational investment. While SNP arrays produce genotype data that is relatively easy to process, skim sequencing requires imputation, data reduction and an adequate reference (or reference-free method) in order to produce genotype information of equivalent quality. Thus, one cost is substituted by another. Whether this trade-off is worth it will likely depend on the research aim, the size and genetic diversity of the population and the computational resources already available in each case.

What is undeniable is that the shift towards sequencing technologies is unavoidable and using them to their full potential will require more computational resources, techniques and expertise than ever before.

Declarations

Data availability

Data can be found in the online repository Zenodo with doi: doi.org/10.5281/zenodo.7978963.

Supplementary data 1: Linkage map of binned markers (2434 bins).
Supplementary data 2: Linkage map of all markers (1.85M markers).
Supplementary data 3: Corrected genotypes.
Supplementary data 4: Reproducible code and results (produces all outputs and figures).

Code availability

Reproducible code can be found in the online repository Zenodo with the doi: doi.org/10.5281/zenodo.7978963.

Author's contribution

Research was proposed by EvdW, TvD and JW. TvD produced the population. JW genotyped the samples. ATN carried out research and wrote the manuscript. PB, EvdW and CM provided regular supervision. FAdFG, JW, PA and RF discussed the approaches and results. CAM, PA obtained funding. All co-authors reviewed the manuscript.

Acknowledgements

The authors would like to thank the contribution of Fresh Forward in preparing, maintaining and genotyping this population, as well as their support and advice while developing this research.

Competing interests

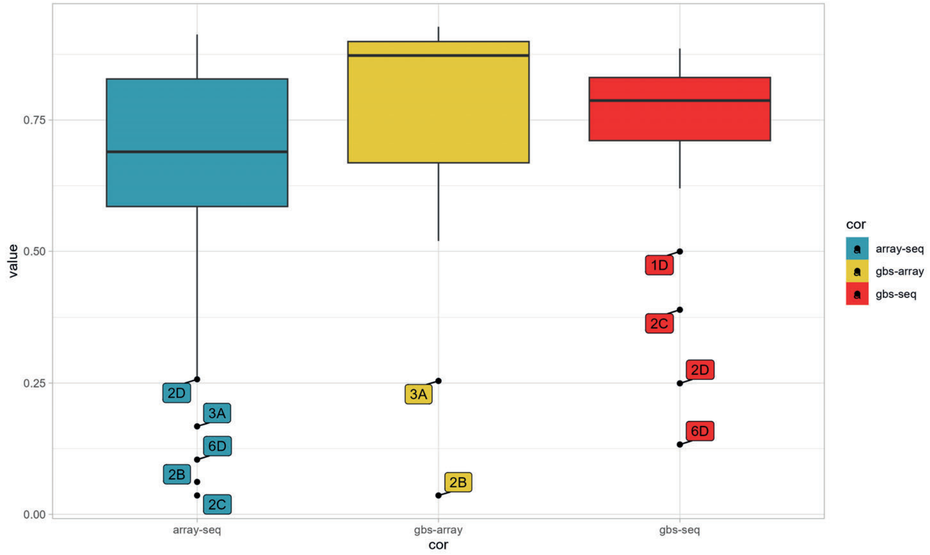
The authors have no conflicts of interest regarding this article.

Funding

This work was supported by the TKI projects LWV20.112 Application of sequence-based multi-allelic markers in genetics and breeding of polyploids

(BO-68-001-042-WPR) LWV20.085 Genetics Assisted Assembly of Complex Genomes (BO-68-001-033-WPR), Novel genetic and genomic tools for polyploid crops (KV 1605-020/TU-16006; BO-68-001-001), and by USDA's National Institute of Food and Agriculture (NIFA) Speciality Crop Research Initiative project 'Tools for genomics- assisted breeding in polyploids: Development of a community resource' [2020-51181-32156/SCRI].

Supplementary Figures



Supplementary figure 1: Boxplot of order correlations between array (array), sequencing-based map (gbs) of the Holiday x Korona population and the F. x ananasa "Camarosa" genome sequence (seq).

Chromosomes with a correlation below 0.5 were labelled to highlight the maps showing low collinearity.

Chapter 6

Multivariate QTL approach reveals a major regulator of terpenoid production and other volatiles in strawberry

Alejandro Thérèse Navarro[1], Olga Zafra Delgado[1,2], Yury Tikunov[1], Carel Peeters[3], Johan Willemsen[4], Thijs van Dijk[4], Peter M. Bourke[1], Eric van de Weg[1], Richard G.F. Visser[1], Chris A. Maliepaard[1]

1 Plant Breeding, Wageningen University & Research
2 Volatile Biosynthesis, Max Planck Institute for Chemical Ecology
3 Mathematical and Statistical Methods, Wageningen University & Research
4 Fresh Forward Breeding & Marketing, B.V.

Abstract

Garden strawberries, *Fragaria* x *ananasa,* are loved for their unique aroma and taste. Their fragrance can be attributed to one of the most complex repertoires of volatile organic compounds (VOC), comprising esters, terpenoids, aldehydes, alcohols, furans, lactones, benzenoids and sulfuric compounds. While the aroma profile has been the object of study for a long time, the genetic control of this complex trait is still mostly unknown. In this study, we analysed a diversity panel and a biparental population of strawberries using gas-chromatography mass-spectrometry. We detected 125 compounds of which 96 were identified, comprised mostly of esters and terpenoids. To simplify these complex datasets, we applied multivariate transformations to obtain multivariate phenotypes that summarise phenotypic information of groups of correlated metabolites. Combined with a QTL analysis, this approach allowed us to identify major QTLs regulating terpenoid production. We were not able, however, to detect general QTLs for ester compounds, suggesting that this compound class is not as co-regulated as terpenoids. Finally, we compared our results with those published in four previous studies, confirming some of our results but with little overall overlap. Although the populations of these studies are quite different, the lack of overlapping QTL may suggest large non-genetic effects on VOC production that need to be considered to obtain more reproducible results.

Keywords

aroma; metabolomics; multivariate QTL; strawberry; terpenoid

Multivariate QTL approach reveals a major regulator of terpenoid production and other volatiles in strawberry

Introduction

Volatile organic compounds (VOC) are a large group of molecules produced by the secondary metabolism of plants. Due to their variability and abundance, they are used by all kinds of organisms to detect and identify plants. For instance, pollinators use them to find flowers (Schiestl and Johnson 2013; Raguso 2016). They also mediate attraction of herbivores and their predators (de Boer et al. 2004; Bruce et al. 2005; Clavijo McCormick et al. 2012) and they are an essential tool in competition among plants (Effah et al. 2019). In general, VOCs are detected through the sense of taste and smell, thus they also play a major role in determining consumer liking of plant parts, or any other food (Fan et al., 2021; Pavan et al., 2021; Torri et al., 2021).

Consumer preference studies have shown multiple times that aroma, together with sweetness, are the most important indicators for overall liking in strawberry (Schwieterman et al. 2014; Fan et al. 2021). The importance of its aroma stands out even in its genus name, *Fragaria,* from the Latin *fragans*, meaning "fragrant". Indeed, the chemical composition of strawberry aroma is surprisingly complex: over almost three decades of study, 978 different VOC compounds have been identified in *Fragaria* fruits (both wild and cultivated), with around 300 being reported more than once (Ulrich et al. 2018). They can be summarized into a few chemical families: terpenoids, esters, aldehydes, lactones, furans, alcohols and ketones. The wide variety of compounds reported highlights a complex aroma metabolism that has proven a significant challenge for biologists trying to elucidate its genetic control. Aroma volatiles are often produced by complex biosynthetic pathways – chains of reactions mediated by multiple enzymes and controlled by regulators, each susceptible to genetic variability, resulting in quantitative variation of VOCs. Moreover, fruit physiology (*e.g.* the ripening process) and environmental factors also affect VOC production. It is probably due to this combination of effects that only four volatile-related enzymes have been functionally validated; FaNES (Aharoni et al. 2004), FaOMT (Zorrilla-Fontanesi et al. 2012), FaFAD (Zorrilla-Fontanesi et al. 2012; Oh et al. 2021) and FaSAAT (Aharoni et al. 2004; Leonardou et al. 2021). While more is being discovered about the fruit physiology of strawberry and its regulators, what is abundantly clear is that the ripening process is heavily influenced by environmental factors (Leonardou et al. 2021).

Due to the high chemical complexity of strawberry aroma, a comprehensive analytical approach such as metabolomics is a useful tool to understand its biochemistry. Analysis of metabolomic datasets, however, is far from simple. They are typically highly collinear, due to the underlying metabolic networks controlling compound abundance, resulting in many metabolites displaying similar patterns of variation across samples. For this reason, multivariate statistical methods are a powerful tool to study and characterize metabolomic datasets, shedding light on the relationships between metabolites (Hendriks et al. 2011; Worley and Powers 2013; Saccenti et al. 2014; Debik et al. 2022). Consequently, when performing QTL studies using metabolomic data, multivariate QTL models are more powerful than univariate ones (Galesloot et al. 2014). The simplest approach to a multivariate QTL study is to perform a multivariate transformation of the original traits into variables representing distinct features of the whole dataset. These multivariate phenotypes can then be used in a univariate QTL analysis without the need of using cumbersome multivariate regression models. In a metabolomic context, this translates to using the correlation between metabolites to obtain a multivariate phenotype representing a subset of correlated metabolites. Thus, the obtained QTL would identify loci controlling groups of metabolites, rather than individual ones. Moreover, we expect that by capturing shared information across variables we can diminish the influence of noise present in each individual metabolite measurement thus enhancing the signal to noise ratio, yielding clearer QTL profiles. Particularly in strawberry VOC studies, where so many metabolites are identified with relatively low repeatability across studies (Ulrich and Olbricht 2016), this technique could have the potential to simplify the interpretation of results and help clarify on an otherwise obscure topic.

In this study we present the volatile repertoire of 345 strawberry individuals included in a biparental cross and a diverse population. The two datasets provide an interesting look into metabolic diversity of strawberries, highlighting the variability within a single family versus the variability of a wider germplasm. We applied both network-based correlation clusters and factor analysis (Peeters et al. 2019) in order to understand the relationship between compounds and to compute multivariate phenotypes. Two association models implemented in the software GAPIT3 (Wang and Zhang 2021), namely the mixed linear model (MLM) and Blink (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway) (Huang et al. 2019) were used for QTL detection. Additionally, we collected QTL markers previously reported (Zorrilla-Fontanesi et al. 2012; Barbey et al. 2021; Rey-Serra et al. 2022; Fan et al. 2022) and imputed their locations in the novel Royal Royce genome assembly (Hardigan et al. 2021a). To our knowledge, this represents the most comprehensive list of volatile QTLs in strawberry to date.

Materials & Methods

Plant materials

Two strawberry populations were used in this study, both genotyped using the iStraw90K array (Bassil et al. 2015). The first was a biparental population between two Junebearing varieties Rumba and Malwina, consisting of 259 samples representing 147 unique genotypes. Plants were grown in Stevensbeek, The Netherlands in outdoor tabletops. Fruits were collected during a 2-week period in June of 2013 with most genotypes being sampled at least twice. The second dataset consisted of 486 samples representing 198 genotypes from a diversity panel, varying from experimental clones to released cultivars, including everbearers, June-bearers and Mediterranean varieties. These were grown in the autumn in a greenhouse in Stevensbeek, the Netherlands, and harvested during the first week of August 2019. Core samples from fruits of both trials were obtained by diagonal sampling using a fruit corer. About 20g of fruit cores were obtained per genotype, frozen using liquid nitrogen and ground to a fine powder. The samples were then stored at -80°C before analysis, for a few days up to two months, depending on the order of sample analysis. Within each trial, a series of quality control (QC) mixtures were obtained by mixing core samples from all varieties.

Volatile extraction, identification and quantification

Before volatile extraction, 1g of frozen fruit powder was incubated in a water bath at 30°C for 10 minutes. Then, 1 mL EDTA-NaOH solution (100mM EDTA, pH 7.5) was added to the sample. Shortly after, 2.2g of solid $CaCl_2$ (Sigma-Aldrich) was added. The samples were shaken thoroughly before being exposed for 20 min at 45°C to a 65-mm polydimethylsiloxane-divinylbenzene (SPME) fibre, fused silica (Supelco). The extracted volatiles were inserted in the injection port and desorbed for 1 min at 250°C in splitless mode.

The biparental population and the diversity panel were analysed using different chromatographic columns. The former was separated using a Thermo TR-5ms SQC column, with helium as carrier gas and a constant flow of 2.0 mL/min, which resulted in low chromatographic resolution of esters. To improve this, volatiles from the diversity panel were analysed using a polar chromatographic column, Stabilwax®-DA. Differences in the chromatographic column hampers the combined raw data files pre-processing, and the quantitative comparison across datasets. Therefore, the two volatile datasets were independently pre-processed using Metalign (Lommen 2009), MSClust (Tikunov et al. 2012) and annotated matching mass spectra and retention indices to the

NIST14 mass spectra library using NIST MS Search software (NIST).

Multivariate analysis

Two multivariate techniques were used to understand the relationship between metabolite intensities. Each technique resulted in a grouping and a multivariate phenotype summarising the group. All volatile intensities were expressed as log10 transformations. All analyses were performed in R (R Core Team 2016).

Correlation networks were computed using the igraph package (Csardi and Nepusz 2006). To construct the graph, each metabolite was used as a vertex. Edges were drawn between vertices when the absolute correlation between the two metabolites was above 0.5. Community detection was performed using a greedy modularity optimization algorithm (Clauset et al. 2004). Each cluster was afterwards summarised using principal components decomposition, only considering the two first components per cluster.

Factor analyses were performed using a regularized maximum likelihood approach implemented in the R package FMradio (Peeters et al. 2019). To determine the optimal number of latent factors, the Guttman bound was used which seeks to separate signal factors from noise factors. The factor loadings were used to determine the most important compounds for each latent feature. All compounds with an absolute factor loading >0.3 were considered as important metabolites for that factor.

The results of both multivariate approaches were used to inform QTL analysis twofold. First, the QTL results of those metabolites grouped together (either in a network community or within a factor), were analysed together. Secondly, multivariate phenotypes (either principal components or factors) were used to perform a QTL analysis, and their results were contrasted with those of the individual compounds.

Quantitative Trait Loci identification

In both datasets QTL analyses were performed using the package GAPIT3 (Wang and Zhang 2021). In particular, the general mixed model (GLM), mixed linear model (MLM), FarmCPU (Liu et al. 2016) and Blink (Huang et al. 2019) models were applied. We report the results of MLM and Blink since they are the most interesting, although the full set of results can be found in the supplementary data 3 and 4. In all models and traits, the Bonferroni threshold was applied. Although this threshold might be considered strict, some permutation tests indicated that such threshold was reasonable (fig. 4). Since our final number of markers with physical positions in the Royal Royce genome was 25400 this was equivalent to 1.97 x 10^{-6}.

Labelling of QTL peaks was done to simplify position reporting. To identify each locus, we first pooled all identified QTL across populations and models within our own study. We then grouped those positions within 500kb of each other to create QTL groups representing a locus (fig. S1). Each locus was labelled according to the chromosome number and its position in the chromosome.

Imputation of physical positions in the 'Royal Royce' assembly

The genetic markers used in this study, as well as those used in previous studies (Zorrilla-Fontanesi et al. 2012; Barbey et al. 2021; Fan et al. 2021; Rey-Serra et al. 2022) were placed in the *F.* x *ananasa* "Royal Royce" genome assembly (Hardigan et al. 2021a). To obtain physical locations of these markers BLAST was used to find regions of high similarities of the marker sequences (either microarray probes or primers) to the Royal Royce genome. Since all markers had perfect hits in multiple subgenomes we had to use additional information to resolve the ambiguities. To that end, we used three linkage maps: (Zorrilla-Fontanesi et al. 2012), a consensus linkage map developed as part of the OctoSeq project (Vickerstaff and Harrison 2017) and the physical map of the 850K SNP array (Hardigan et al. 2020). Together, these maps covered the iS-

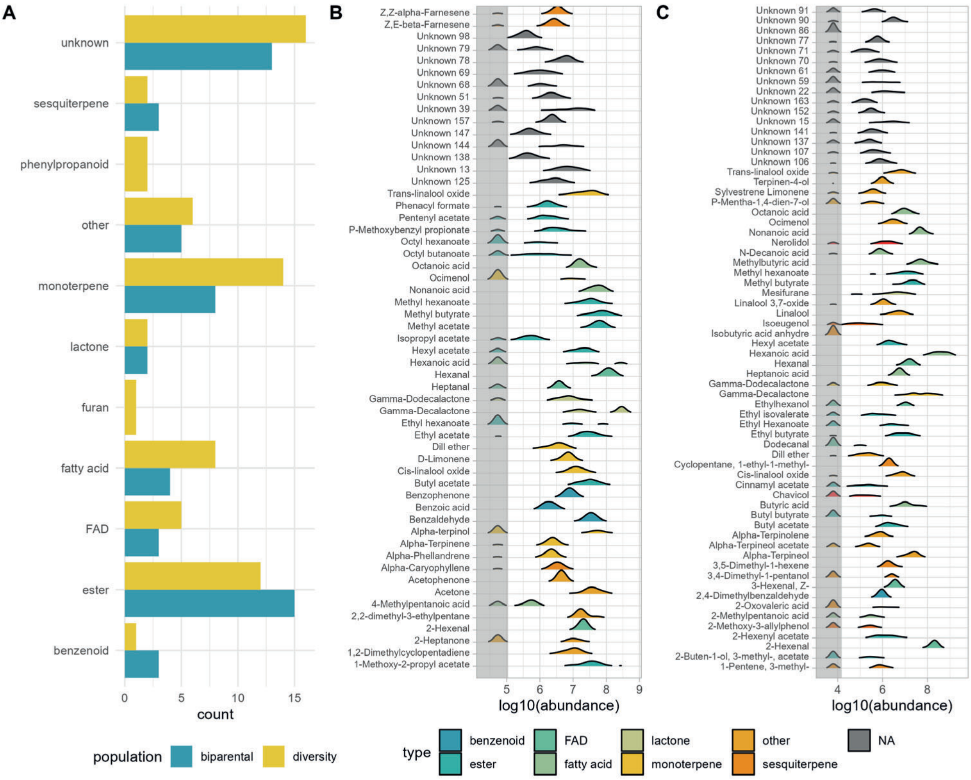


Figure 1: Summary of compound identification and abundances.

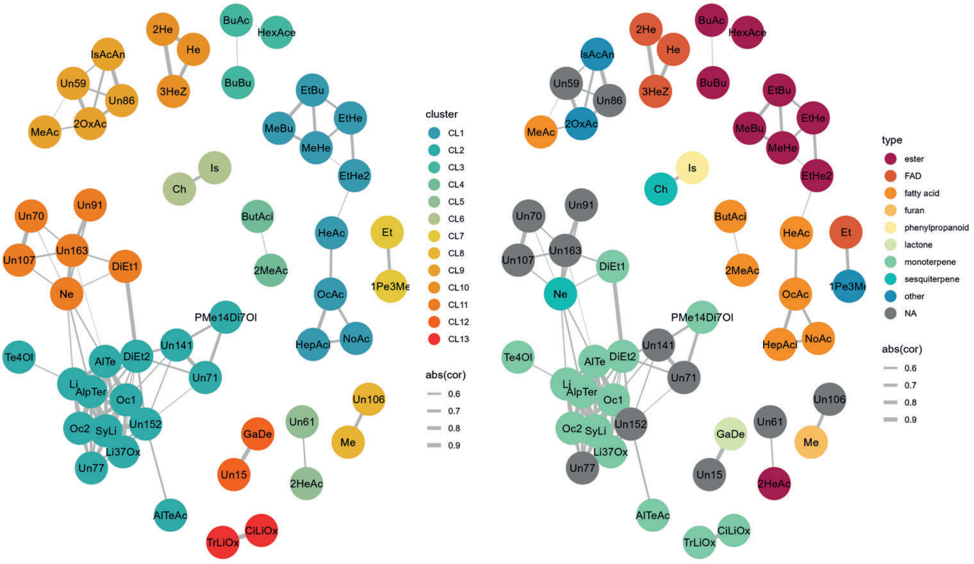
A) Barplot showing the number of identified compounds of each type per population. FAD refers to fatty-acid derived compounds. **B)** Distribution of compound abundances for the Rumba x Malwina biparental population. Grey area indicates the threshold of non-detection, i.e. samples with abundances below the threshold do not have detectable traces of the compound. **C)** Same as B but for the diversity panel.

traw SNP array probes (Bassil et al. 2015), the FanaSNP probes (Hardigan et al. 2020) and the SSR markers used in previous studies (Zorrilla-Fontanesi et al. 2012). By comparing the subgenome blast hits and the linkage groups of each map, we were able to unequivocally assign each linkage group (and hence the markers within that group) to a subgenome, and thus we could select an adequate position for each marker.

Results

Identified compounds

A total of 125 compounds were detected across both populations, of which 96 could be confidently identified. In both datasets esters and terpenoids were the most common type of compound, although in the diversity panel many more monoterpenes were identified than in the biparental dataset (fig. 1A). The diversity panel harboured a higher number of chemical compounds, particularly monoterpenes, fatty acids, fatty-acid derived (FAD) and unidentified compounds. However, more esters and benzenoid compounds were found in the biparental population. In both datasets hexanoic acid was one of the most abundant compounds followed by gamma-decalactone (fig. 1B, 1C). Hexanal was also very abundant in the biparental population, while in the diversity panel 2-hexanal was more abundant. The most striking quality of both datasets is the great variation in abundances between compounds, with some compounds being 10.000 times more abundant than others (e.g. isopropyl acetate and hexanoic acid, fig. 1B).



Terpenoids and esters form independent clusters

The range of compound abundance correlations within the diversity panel (DP) and biparental (BP) datasets was from -0.51 to 0.96 and -0.41 to 0.90 respectively. Many compounds exhibited high correlations among them, which suggests a genetic co-regulation of their metabolism. Terpenoid compounds (monoterpene and sesquiterpene) and ester compounds exhibited the highest levels of correlations within each biochemical class. In both datasets, terpenoid compounds formed the largest clusters (15 compounds in DP, fig. 2; and 9 in BP, fig. S1). Interestingly, in both cases 4 unidentified compounds were grouped within the terpenoid cluster. In the case of esters, in the DP there is a clear relationship between the abundance of some esters (methyl and ethyl butanoate, and methyl and ethyl hexanoate) and of C6 to C9 fatty acids. That

Figure 2: Correlation network of volatile abundance in the diversity panel. Lines between nodes represent (absolute) correlations above 0.5 and the width of the line is proportional to the correlation. **Right)** Colours correspond to clusters identified using a greedy modularity identification algorithm. **Left)** Colours correspond to the chemical group of each metabolite (FAD: fatty-acid derived). Metabolite names have been summarized: 1Pe3Me: 1-Pentene, 3-methyl-; 2He: 2-Hexenal; 2HeAc: 2-Hexenyl acetate; 2MeAc: 2-Methylpentanoic acid; 2OxAc: 2-Oxovaleric acid; 3HeZ: 3-Hexenal, Z-; AlpTer: alpha-Terpineol; AlTe: alpha-Terpinolene; AlTeAc: alpha-Terpineol acetate; BuAc: Butyl acetate; BuBu: Butyl Butyrate; ButAci: Butyric acid; Ch: Chavicol; CiLiOx: cis-Linalool oxide; DiEt1: Dill ether 1; DiEt2: Dill ether 2; Et: Ethylhexanol; EtBu: Ethyl butyrate; EtHe: Ethyl Hexanoate; EtHe2: Ethyl Hexanoate 2; GaDe: gamma-Decalactone; He: Hexanal; HeAc: Hexanoic acid; HepAci: Heptanoic acid; HexAce: Hexyl acetate; Is: isoeugenol; IsAcAn: Isobutyric acid anhydre; Li: Linalool; Li37Ox: Linalool 3,7-oxide; Me: Mesifurane; MeAc: Methylbutyric acid; MeBu: Methyl butyrate; MeHe: Methyl hexanoate; Ne: Nerolidol; NoAc: Nonanoic acid; Oc1: Ocimenol 1; Oc2: Ocimenol 2; OcAc: Octanoic Acid; PMe14Di7Ol: p-Mentha-1,4-dien-7-ol; SyLi: Sylvestrene Limonene; Te4Ol: Terpinen-4-ol; TrLiOx: trans-linalool oxide; UnXXX: Unknown compound.

is not the case in the biparental dataset (fig. S1), where a single cluster of 7 compounds (6 esters and 1 unknown) contains 6 out of 9 identified esters. The remaining clusters are relatively small, but almost all of them contain chemical compounds belonging to a single chemical class.

Factor analysis and correlation clusters compute similar groups

We used factor analysis as described in (Peeters et al. 2019) in order to com-

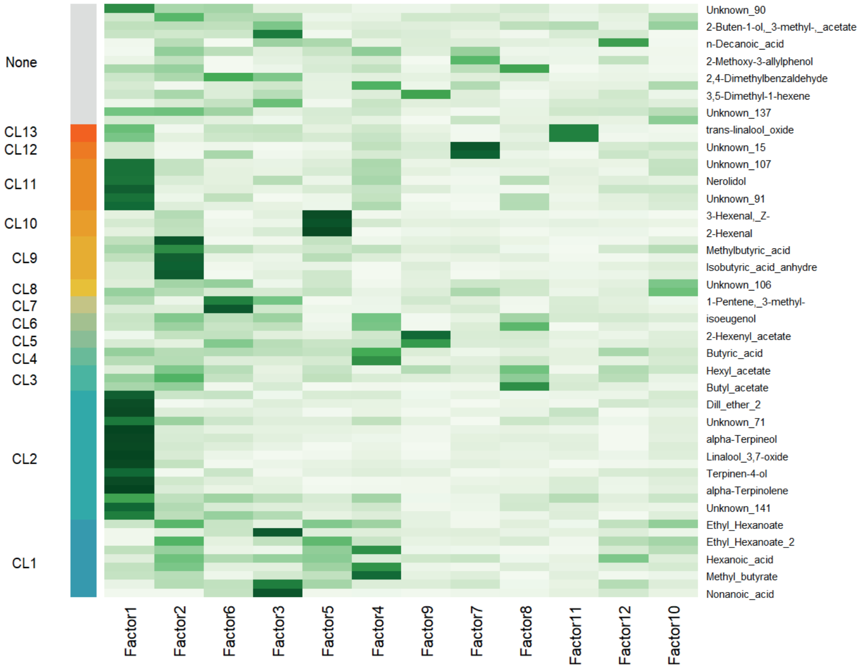


Figure 3: Heatmap of absolute loadings for factor analysis of metabolite abundances in the diversity panel, compared to correlation clusters.

Each row corresponds to one compound and each column to one factor. Darker colours represent a higher importance of that compound for that factor, measured by the absolute loading value. The left, coloured band indicates which rows belong to which cluster in the network-based correlation analysis.

pute multivariate phenotypes. For each factor a set of loadings was obtained for each metabolite which indicates the importance of each metabolite for that factor. A comparison of factor loadings with the correlation clusters (fig. 3) shows mostly a clear correspondence between clusters and factors (*i.e.* metabolites that had been clustered together in the network analysis all have high loadings in a single factor, thus a factor represents a cluster and vice-versa). Some factors are represented by single clusters: factor 2 corresponds to CL9, factor 5 to CL10, etc. In other cases, multiple clusters are represented by one factor, as happens with the terpenoid clusters CL2 and CL11 which are encompassed by factor 1. Some clusters are split for different factors, e.g. CL1 into factors 3 and 4. Lastly, factor analysis allows us to include a series

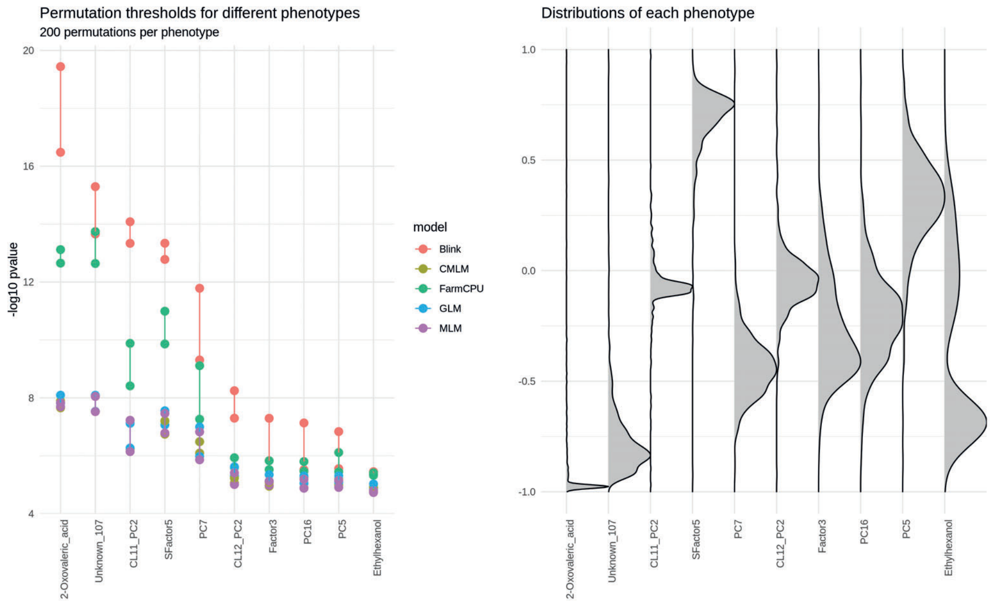


Figure 4: Permutation thresholds for different phenotypes using 200 permutations.

Left) Permutation thresholds given as confidence intervals as suggested by Nettleton and Doerge. Each colour represents the threshold for a different model. **Right)** Phenotypic distributions for each phenotype with the spread of the distribution scaled between 1 and -1. The presence of long tails, for instance in 2-oxovaleric acid or SFactor5, indicates the presence of outliers.

of compounds that were not in the network analysis due to their overall low correlation with the rest of the compounds. Although there are some discrepancies, both methods are largely equivalent: terpenoids (and some unknown compounds) are grouped in a single factor, esters are grouped into two factors, and the remaining factors represent mostly two or three compounds.

Blink and FarmCPU are less tolerant to non-normal distributions than mixed models

We were interested in applying the novel models implemented in the Blink and FarmCPU software due to their enhanced performance and statistical power (Liu et al. 2016; Huang et al. 2019). Since the authors of these models did not provide a specific method of establishing marker significance, we considered computing empirical significance thresholds through permutation tests (Nettleton and Doerge 2000). We selected a subset of phenotypes with clearly different distributions, many deviating significantly from a normal distribution. Our permutations show a clear relationship between the lack of normality and the presence of false positives. Particularly in phenotypes where a few samples were strong outliers we could observe how the permutation thresholds reached levels of ~10^{-16} to ~10^{-20} for Blink, and around 10^{-13} for FarmCPU (fig. 4). In contrast, mixed linear models had a permutation threshold of ~10^{-8} for the same trait. This should signal that although Blink and FarmCPU can be more powerful models, they are more prone to extremely significant false-positive results in the absence of normality. This suggests that data with extreme distributions such as those found in metabolomics data might not be suitable for this family of GWAS models. For this reason, we will focus mostly on the results of mixed linear models for the QTL analysis.

Multivariate phenotypes help summarise the QTL results of related compounds

Three groups of compounds defined through the multivariate analysis contained more than three compounds. The first corresponds to terpenoid compounds (Factor1, CL2 and CL11), the second to compounds related to butyric acid (Factor 2 and CL9) and the third to ester compounds (Factor 3, Factor 4 and CL1). We performed a GWAS analysis on the multivariate phenotypes, as well as on the metabolites that composed each group.

For the terpenoid groups (fig. 5), we can clearly identify three major QTL

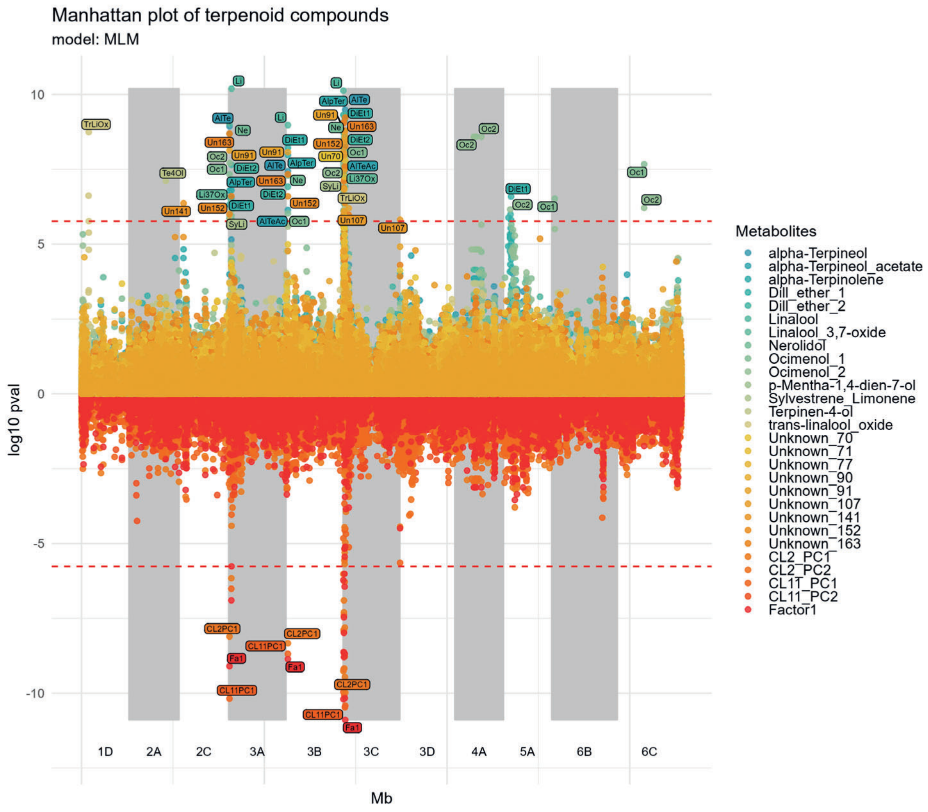


Figure 5: Overlapped Manhattan plot of terpenoid compounds represented by Factor1, CL2 and CL11.

Only chromosomes with a significant marker in at least one trait are shown. The log10 p-values of multivariate phenotypes are plotted below the x-axis (log10(pval)), while log 10 p-values obtained from metabolites are plotted above the x-axis (-log10(pval)). The three multivariate groups represented summarise the same subset of compounds in this study: terpenoids. The red dashed lines represent the Bonferroni significance threshold. All significant points were labelled for easier interpretation.

positions shared across many compounds of this class on chromosomes 3A, 3B and 3C. The signal on chromosome 3A can be further separated into two nearby positions, which we labelled as loci 3A.1 and 3A.2. These QTLs are detected regardless of the multivariate phenotype used and are shared across 15, 8, 12 and 17 terpenoids respectively (table 1), clearly pointing towards major regulators of terpenoid production. Their positions in equivalent areas of each subgenome and high correlation among the genotypes of the peaks on loci 3A.2, 3B.1 and 3C.1 suggest that positions on chromosomes 3A and 3B might be artifacts due to wrong imputation of the physical location in the genome. As for 3A.1, the correlation of its peak marker with the others is lower, suggesting a putative second QTL. Besides these major QTLs, others were detected for individual metabolites or smaller groups of metabolites. For instance, in chromosome 5A a QTL for dill ether and ocimenol was detected. However, these QTL were not detected using multivariate phenotypes. Thus, the multivariate phenotypes do not encompass the complete set of QTL detected within the group, but rather the main QTL affecting most compounds. It is important to note that although all terpenoid compounds have correlated abundances, not all of them had significant QTL.

For the other two groups (ester compounds and butyric acid-related compounds) there were fewer QTL detected (Table 1) and little overlap between compounds (Supplementary fig. S3, S4). As a result, there were no major regulatory QTL identified. Moreover, the results from MLM and Blink models were very different, with many more QTL identified using the Blink model, suggesting a high number of false positives among the results. In these groups there were some QTL detected only when using the multivariate phenotypes. Although these QTL could point towards regulators that are only detected when combining information across phenotypes, the lack of repeatability of these QTL across models suggests they might be false positives. The same can be said of the multivariate groups that represented smaller sets of metabolites.

QTL signals are not repeatable across studies

We gathered QTL positions from four additional studies (Zorrilla-Fontanesi et al. 2012; Barbey et al. 2021; Rey-Serra et al. 2022; Fan et al. 2022) and imputed their positions in the Royal Royce genome assembly (Hardigan et al. 2021a) (Supplementary table 3). We compared the QTLs reported for terpenoid, ester and butyric-acid-related compounds across studies and found little overlap. Of the three possible major regulators of terpenoid biosynthesis found in our study, only one (locus 3C.1) co-located with QTL from three other studies: Barbey *et al.* 2021, Fan *et al.* 2022 and Rey-Serra *et al.* 2022 (fig. 6). Of the many QTL reported on chromosome 4A for terpenoid compounds, few of them had positional overlap across studies. The picture is not much better for esters, where there is also little overlap (Supplementary fig. S6): There are plenty of QTL reported across the genome, with two locations containing QTL from more than two studies: 3B 24Mb and 6A 17.5-21Mb. For the butyric compounds, no other study detected significant QTL. However, we did find some common QTL across studies for two furan compounds: furaneol on 1C and mesifurane on 7D (Supplementary fig. S7)

It is important to note that each study identified different compounds, with only 7 out of 102 compounds identified in more than two studies (alpha-terpineol, butyl acetate, gamma-decalactone, mesifurane, methyl anthranilate, nerolidol, octanol) and 76 compounds were identified only in a single study. Previous studies were all based on biparental crosses, and thus the lack of overlapping volatiles might be due to the comparison with our diversity panel, which is more diverse in metabolites. Nevertheless, common QTL for groups of metabolically related compounds would be expected.

Discussion

Multivariate analysis of metabolic data

One of the main challenges of this study was the organisation of results into a sensible and comprehensive summary. High-throughput datasets such as the metabolomic dataset here studied offer a substantial amount of information but extracting and understanding this information is not a trivial task. To address this issue, we classified our metabolites using two different multivariate approaches: correlation network communities and factor analysis. While the

Model	MV group	MV QTLs in metabolites	MV-only QTLs	Met-only QTLs	Compound types
Blink	CL1	2B.1 (1), 2C.5 (2)	6A.10, 6A.4	10	ester, fatty acid
Blink	CL2	3C.1 (9)		21	monoterpene, unknown
MLM	CL2	3A.1 (9), 3A.2 (5), 3B.1 (8), 3C.1 (10)		8	monoterpene, unknown
Blink	CL5	6A.9 (1)	1A.5, 2A.12, 7A.1	0	ester
MLM	CL5		6A.9	0	
Blink	CL6		5B.10	5	phenylpropanoid, sesquiterpene
Blink	CL7		1C.2, 7D.7	1	FAD
MLM	CL7		7D.7	0	
Blink	CL8		2B.6, 3C.1, 5A.3, 5C.4	2	unknown
MLM	CL8		3A.2, 3B.1, 3C.1	1	furan
Blink	CL9	7A.8 (2)	4B.1	6	other, unknown
MLM	CL9		4B.1	2	unknown
Blink	CL10	7B.1 (3)	6C.11, 7A.10, 7B.11	2	FAD
MLM	CL10	7A.2 (1)		0	FAD
Blink	CL11	3C.1 (3)	6B.12	9	monoterpene, unknown
MLM	CL11	3A.1 (6), 3A.2 (3), 3B.1 (4), 3C.1 (6)		4	monoterpene, sesquiterpene, unknown
Blink	CL12	3B.11 (2)	5B.5, 7C.8	0	lactone, unknown
MLM	CL12	3A.20 (2), 3B.11 (2)		0	lactone, unknown
Blink	CL13	1D.2 (2), 3A.17 (1), 3C.1 (2), 4A.10 (2)	5A.3, 6A.6	2	monoterpene
MLM	CL13	1D.2 (2), 3C.1 (1)		0	monoterpene
Blink	Factor1	2A.16 (1), 3C.1 (13)		32	monoterpene, unknown
MLM	Factor1	3A.1 (15), 3A.2 (8), 3B.1 (12), 3C.1 (17)		12	monoterpene, sesquiterpene, unknown

Model	MV group	MV QTLs in metabolites	MV-only QTLs	Met-only QTLs	Compound types
Blink	Factor2		5A.13	9	ester, other, unknown
Blink	Factor4	3D.6 (1), 6D.5 (4)	3A.10, 4C.1	7	ester, fatty acid
Blink	Factor5	6A.6 (1), 7B.1 (3)	3B.7, 5D.9	2	ester, FAD
Blink	Factor7		3C.4, 6A.3, 6A.9	1	lactone, unknown
Blink	Factor8		3B.11, 7A.6	7	ester, sesquiterpene
MLM	Factor8		3A.20, 3B.11	2	ester, sesquiterpene
Blink	Factor9	6A.9 (1)	1D.4, 2A.9, 6D.13, 6D.4	0	ester
Blink	Factor11	1D.2 (2)	1C.4	5	monoterpene
MLM	Factor11	1D.2 (2)		1	monoterpene
Blink	Factor12		3D.4, 6D.8	1	fatty acid

Table 1: Summary of QTL found organised by multivariate groups.

For each multivariate group (MV group) we show the loci of multivariate QTL confirmed by individual metabolite QTL (MV QTL in metabolites), with the number of supporting metabolites between parentheses. We show the loci of QTL found only using the multivariate phenotypes (MV-only QTL). Lastly, we report the number of metabolite QTL that were not found in the multivariate phenotypes (Met-only QTL). Only those multivariate groups for which some significant QTL were found are reported. The description of each QTL locus (size, markers included, location) can be found in supplementary table 2.

former technique is more commonly used in metabolomics studies (Toubiana et al. 2013; Debik et al. 2022; Rey-Serra et al. 2022) the latter is still gaining traction in natural sciences as a promising method to simplify multivariate datasets (Reyment and Jöreskog 1993; Brzozowski et al. 2022). Although our study suggests that both techniques provide mostly equivalent groupings and subsequent multivariate phenotypes, we favour factor analysis as the most robust of the two techniques as justified below.

In this study, we were interested in producing multivariate phenotypes that

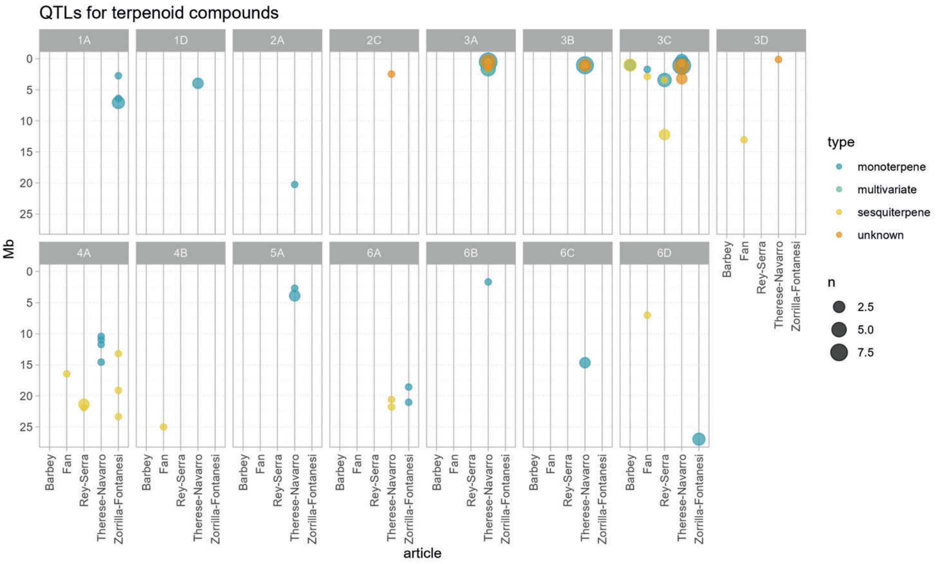


Figure 6: QTL positions of terpenoid compounds across studies.

Positions were imputed on the Royal Royce genome assembly. The size of each dot corresponds to the number of significant associations at a specific location. The colour corresponds to the type of phenotype. Compound QTLs shown include: alpha-curcumene, alpha-farnesene, alpha-terpineol, alpha-terpineol acetate, alpha-terpinolene, beta-bisabolene, beta-farnesene, beta-myrcene, beta-pinene, cl11 pc1, cl2 pc1, d-limonene, dill ether 1, dill ether 2, factor1, geraniol, linalool, linalool 3,7-oxide, mesifurane, myrtenol, nerolidol, ocimenol 1, ocimenol 2, p-mentha-1,4-dien-7-ol, sylvestrene limonene, terpinen-4-ol, trans-linalool oxide, unknown 107, unknown 141, unknown 152, unknown 163, unknown 70, unknown 71, unknown 77, unknown 90, unknown 91. For detailed positions check supplementary table 3.

would summarise information within groups of related compounds. Using the network-based approach, first we must compute correlation communities (*i.e.* clusters). To that end, an arbitrary threshold (to compute the network) as well as a clustering algorithm must be selected. The two choices can greatly impact the resulting groups. Moreover, metabolites with correlations below the chosen threshold will not be considered during the analyses, even though they might contain relevant information. This was the case for unknown compound 90, which only through factor analysis can be assigned to the terpenoid group. Since the network approach is not a dimensionality reduction technique, we must use a different method to obtain multivariate phenotypes. We used principal component decomposition, which forces us to choose a number of components to keep, yet another arbitrary choice. The combination of threshold, clustering algorithm and decomposition method yields an approach the outcome of which will greatly depend on the decisions made by the researcher, rather than on the data. In contrast, factor analysis takes all metabolites into account, regardless of overall correlation and calculates a set of reduced dimensions that can be directly used as multivariate traits. While it is true that factor analysis is not a clustering method, analysing the loadings of each factor revealed the most important metabolites for that factor. We have shown that this is mostly equivalent to the network-based clusters while considering more compounds and obtaining fewer summary phenotypes to analyse. Overall, this tilts the scale towards using factor analysis in future studies, particularly to generate multivariate phenotypes.

Our main interest was uncovering major regulatory QTLs that could explain metabolite abundance for many compounds at once. Analysing the QTLs for multivariate phenotypes greatly clarified the interpretation of individual compound QTL. Importantly, multivariate QTL (mvQTL) did not encompass the totality of metabolite QTL (metQTL) and therefore we recommend using multivariate analysis as an additional layer of information that is able to summarise and coordinate QTL results for many traits. We obtained the metabolite groups presented in this study using multivariate techniques, and thanks to these we discovered the common QTL shared within groups –as well as those for unique metabolites, independent of the group. We also identified some QTL that were only present in the multivariate traits, not in the underlying metabolites. It remains to be seen whether these QTL have a true biological origin and reflect the higher power of multivariate QTL studies or whether they should be considered an artifact. Given the results shown in the terpenoid group it seems that multivariate traits are more useful to summarise overlapping QTL, rather than to find new, multivariate-only QTL.

Groups of co-regulated metabolites

In this study we focused on three groups of correlated compounds which could be controlled by major regulators. According to our results, that is only the case for terpenoid compounds. While esters are known as major compounds produced by Rosaceae fruits (Schwab et al. 2008; Defilippi et al. 2009), we did not find any major regulator. This was also the case for the group of butyrate-related compounds.

Terpenoid compounds are a highly diverse chemical group found in all plants with roles in plant growth, development and environmental interactions, both biotic and abiotic. They are also highly appreciated for their aromatic properties, adding floral and herbal aromas to many essential oils and fruits, including strawberries (Aharoni et al. 2004; Schwab et al. 2008; Ulrich and Olbricht 2016). Although diverse, all terpenoids originate from the same five-carbon isoprenoids: isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). The flux of these precursors is a critical determinant of terpenoid biosynthesis, and as such, plant cells heavily regulate the core isoprenoid biosynthetic pathways (Hemmerlin 2013; Tholl 2015; Zhou and Pichersky 2020). Plants use two pathways to produce IPP and DMAPP: the cytosolic mevalonic acid pathway (MVA) and the plastidial methylerythritol phosphate pathway (MEP). Precursors from MVA produce mostly sesquiterpenes, while isoprenoids produced through MEP are mostly turned into monoterpenes. In our study, most of the identified terpenoid compounds and QTLs are for monoterpenes, suggesting that the major regulators found have to do with the plastidial synthesis of terpenoids. The location of our detected QTL coincides with those reported in other studies, and it is possible that our signal coincides with the FaNES gene suggested by Aharoni *et al.* 2004, and putatively located in 3C (Barbey et al. 2021; Fan et al. 2021). However, in *Arabidopsis thaliana* crucial regulatory enzymes of the two terpenoid pathways have been suggested: isopentenyl phosphate kinase (Henry et al. 2015) and Nudix hydrolases (Henry et al. 2018). By controlling the amount of available precursors, these enzymes have been shown to impact overall terpenoid abundance. Our QTLs could also reflect one of such regulators. A candidate gene study of the

QTL regions would provide much insight into this genetic regulator.

Esters are considered the key compounds, or character impact compounds, of strawberry fruits (Zabetakis and Holden 1997; Ulrich et al. 2018). Due to their high concentration and early detection, they have traditionally been the main focus of aroma research in strawberries (Aharoni et al. 2004; Dong et al. 2013; Rey-Serra et al. 2022). Esters are produced through the esterification of an alcohol and an acid through the action of an alcohol-acyl transferase (AAT) (Schwab et al. 2008; Defilippi et al. 2009). There exists a wide range of variability in AAT enzymes, both in structure and substrate specificity (Defilippi et al. 2009). Due to the lack of specificity of the identified strawberry AAT (SAAT) enzymes, the ester profile is mostly determined by the availability of alcohol and acid substrates (Aharoni et al. 2000). For this reason, it is not surprising that we could not identify any major regulator for all esters: some are determined by limiting concentrations of alcohols or acids, some are determined by the (in)activity or (in)specificity of ester-producing enzymes (Defilippi et al. 2009). For example, fatty acid metabolism was directly linked to ester production in pear fruits, where lipoxygenase activity could be linked to the resulting ester profiles (Luo et al. 2021).

Lastly, a group of butyrate-related compounds was found, formed by isobutyric acid anhydre, methylbutyric acid, 2-oxovaleric acid and two unknown compounds. Of these, two are clear derivates of butyric acid, while the third is of valeric acid. In cashew apple extract such compounds were found as crucial aroma active compounds adding cheesy, sweaty and rancid notes (Filomena Valim et al. 2003). While these can be an important part of a well-rounded aroma in cheese and alcoholic beverages, it tends to be less desirable in fresh fruits and thus can be considered off-flavours. Although biosynthesis of these compounds could also be derived from fatty acid or branched-chain amino acid degradation, butyric acid derivates such as indole butyric acid (IBA) or gamma-amino butyric acid (GABA) are well known plant hormones and thus the presence of butyric acid derivates could be more related to developmental processes. Regardless of their origin, our study suggests that the observed correlation between these compounds cannot be attributed to a single or few genetic factors since no QTLs were found for them.

Inconsistency across studies

Volatile research of strawberry aroma has been of interest since there has been the possibility of extracting and studying individual metabolite compounds (Zabetakis and Holden 1997). Over the years, the advent of metabolomic techniques that allowed the identification of multiple compounds at once has allowed the discovery of a great variety of compounds. However, as noted in (Ulrich et al. 2018) there is little overlap among the detected volatiles. Our results confirm that the same is true for QTL signals. Across four previous studies and this one, there was a remarkably low overlap between the detected QTLs, and even the detected metabolites were mostly study-specific. Several reasons have been suggested for this seeming lack of consensus, some experimental and some biological (Ulrich et al. 2018).

Experimentally, a diverse set of methods is used to prepare and store samples, extract VOCs, and identify and quantify them; but one would expect that regardless of technique, there would be some level of agreement across studies beyond the identification of broad chemical classes such as esters, terpenoids or lactones. More likely, biological differences might be the real culprit behind the chemical variability observed. It has been reported on several occasions that strawberry genotypes have a wide chemical diversity (Song et al. 2011; Dong et al. 2013; Ulrich et al. 2018; Yan et al. 2018). As a result, the choice of individuals could greatly affect the identified compounds in each study. Our panel contained a wide range of genetic diversity, unlike previous studies that detected QTLs in biparental crosses. Furthermore, volatile abundance within the same variety can vary greatly within the ripening process of a single fruit, and throughout the fruiting period of a strawberry plant. Consequently, the choice of harvesting time can greatly impact the final set of volatiles discovered. On top of that, any environmental factor that influences ripening or fruiting can have a clear measurable effect on VOC abundance and composition, and such has been shown in strawberry (Schwieterman et al. 2014; Leonardou et al. 2021). Together, methodological differences, identification limitations, biological diversity, developmental processes and environmental impact on volatile production may be behind the apparent extreme chemical variability observed in strawberries.

A different approach is necessary

This study contributes to a growing corpus of research studying the production of volatiles in strawberry, but somehow it seems unable to clarify what genetic elements control the wide diversity of strawberry volatiles more deeply. Previous research has pre-eminently focused on ester production, probably due to their high abundance, early detection and known relationship with alcohol and fatty acid catabolization during ripening (Zabetakis and Holden 1997; Aharoni et al. 2000; Defilippi et al. 2009; Schwieterman et al. 2014; Fan et al. 2021). As others have before (Ulrich et al. 2018; Barbey et al. 2021), we point towards terpenoids as an additional group to consider as relevant. Our study suggests that their variation is mostly influenced by few QTL controlling many compounds at once and confirms some loci across other research, suggesting thus an important target for future breeding efforts. Previous studies have also shown the importance of a high terpenoid abundance to improve sweetness perception (Ulrich and Olbricht 2016), although given their wide variety of smells, more detailed studies might be required. However, our analysis also highlights the complexity of volatile metabolomics in strawberry and its dependence on genotype-specific and environmental factors. With this in mind, it might be relevant to question the open-ended discovery approach used in previous research and our own. If the focus of the study is to detect as many metabolites as possible in as many samples as possible, one cannot expect to be able to understand the dynamics of volatile production during ripening or its reaction to environmental cues. Perhaps a better approach would be to limit sampling to a few compounds within a chemical class and increase the number of samples along the ripening stages of a fruit. With such an approach, a clearer relationship between metabolism, development and environment could be drawn, moving us closer to understanding how volatiles are produced and what influences their abundance.

Conclusion

In this study we aimed and achieved to identify QTL for volatile compounds by applying multivariate methods with the hopes of finding major regulatory QTL. We identified many individual QTLs, some unique to the biparental and diversity panels, highlighting the relevance of different population types when studying strawberry metabolomics. In the diversity panel, we found a clear major QTL affecting terpenoid compounds, mostly monoterpenes, suggesting that terpenoid variability is produced by differential regulation of the plastidial MEP pathway. Our analysis also reveals that compound correlation does not necessarily reflect a common set of QTL, suggesting that other factors such as ripening and response to environmental factors could be controlling volatiles beyond what our research could clarify. To characterize the metabolic networks controlling volatile production, especially for esters, further studies will be required, hopefully targeting more specifically the role of developmental processes in the metabolic activity of strawberry fruits. In the meantime, further research into the terpenoid regulatory networks of strawberry could present an interesting target for fundamental research into the MEP and MAV pathways, and for applied research in breeding for strawberry aroma.

Declarations

Data availability

All code and supplementary data can be found in the repository Zenodo.org with the DOI doi.org/10.5281/zenodo.7974239

Supplementary Data 1. Reproducible results file (includes all supplementary data, tables and figures).
Supplementary Data 2. Volatile data table
Supplementary Data 3. Total p-value table of biparental population
Supplementary Data 4. Total p-value table of GWAS population
Supplementary Table 1. Total QTL found
Supplementary Table 2. Summary of loci labels

Supplementary Table 3. Literature QTLs
Supplementary Table 4. Compound ID table
Supplementary Table 5. Physical positions of SNP data in Royal Royce

Code availability

All code and supplementary data can be found in the repository Zenodo.org with the DOI doi.org/10.5281/zenodo.7974239

Supplementary Data 1 contains all code and data required to reproduce all results from this study.

Author's Contributions

TvD grew strawberries and collected samples. YT prepared samples and performed metabolomic analysis. JW genotyped the samples. ATN and OZD performed the multivariate analysis with support from CP. ATN performed the QTL analysis and wrote the manuscript. PB, EvdW and CAM provided supervision. CAM obtained funding. All co-authors provided revisions on the manuscript and accepted the final version.

Acknowledgements

This work was made possible through a multi-year collaboration between researchers at Wageningen University & Research and at Fresh Forward Breeding and Marketing B.V. We would like to thank specially the many greenhouse technicians that contributed to raising and sampling the strawberry populations at the facilities of Fresh Forward. We would also like to acknowledge researchers of other institutions that were kind enough to complement their publicly available QTL data by private communication: Iraida Amaya, Cristopher Barbey, Amparo Monfort and Pol Rey-Serra. We would also like to acknowledge the work on creating a consensus map for strawberry by Richard Vickerstaff, Richard Harrison, Matt Clark and Bernardo Clavijo, who were supported by the OctoSeq project funded by grants BB/N006682/2 and BB/N009819/1. To all, thanks for your help and spirit of open scientific collaboration.

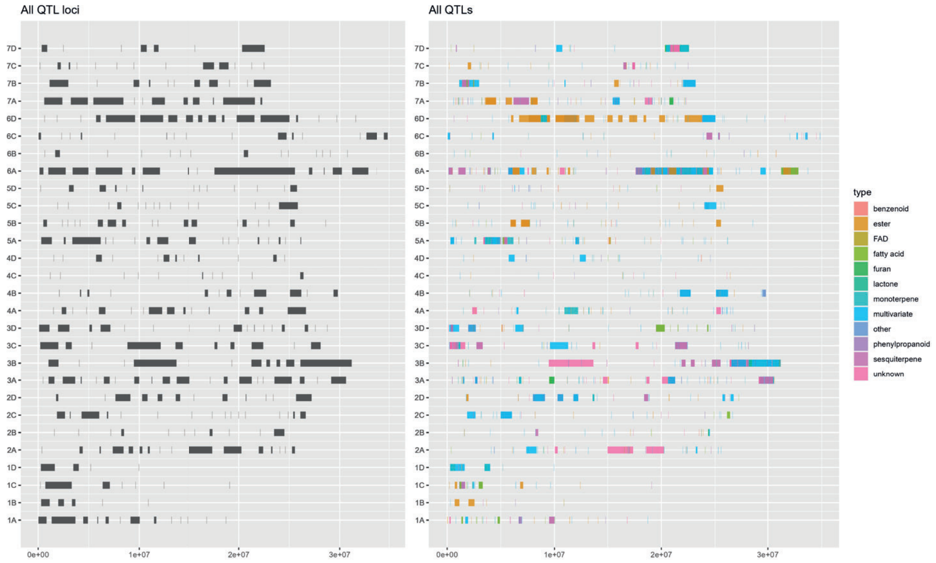
Competing interests

This work was supported by Fresh Forward Breeding and Marketing B.V., an active breeding company of strawberry varieties. Since the genotype data used in this study contains sensitive information from their breeding activities, we will refrain from sharing the complete genotype dataset. For specific inquiries please contact the authors.

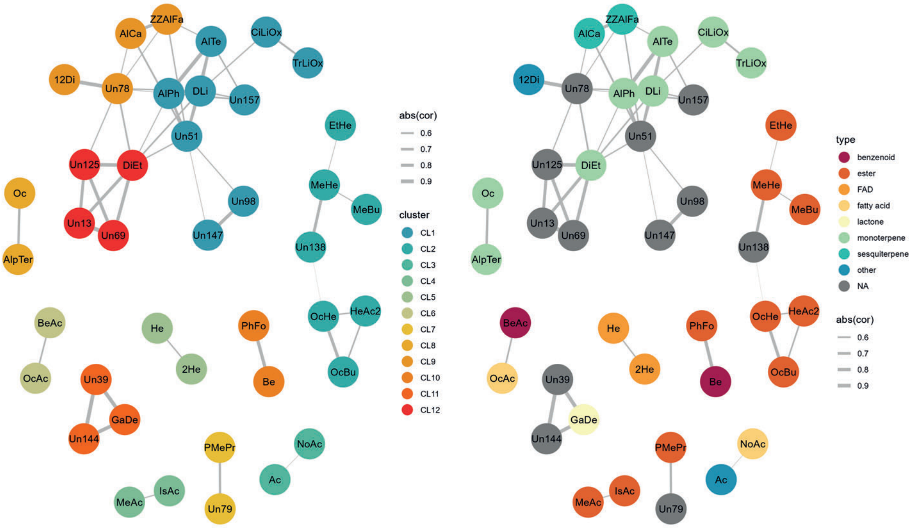
Funding

This work was supported by the TKI projects LWV20.085 Genetics Assisted Assembly of Complex Genomes (BO-68-001-033-WPR), Novel genetic and genomic tools for polyploid crops (KV 1605-020/TU-16006; BO-68-001-001), and by USDA's National Institute of Food and Agriculture (NIFA) Speciality Crop Research Initiative project 'Tools for genomics- assisted breeding in polyploids: Development of a community resource' [2020-51181-32156/SCRI].

Supplementary Figures

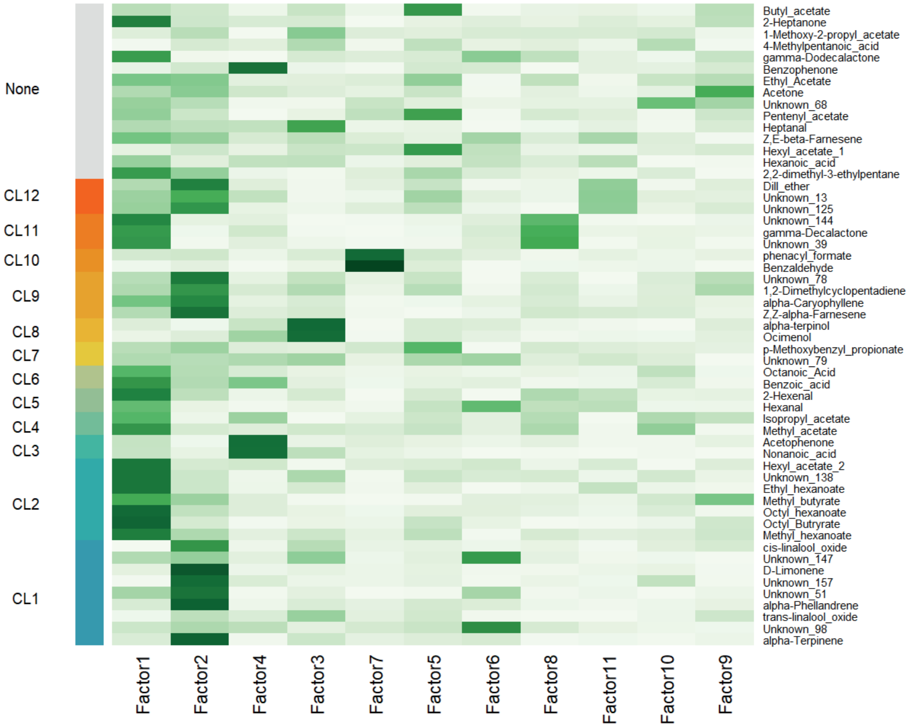


Supplementary figure 1: Representation of QTL loci detected in our analysis. **Left panel,** all QTL loci detected in the analysis, each black box representing the range of each locus. **Right panel**, individual QTL ranges for traits, color coded according to the compound type of each trait. Unknown compounds and multivariate traits are also included.



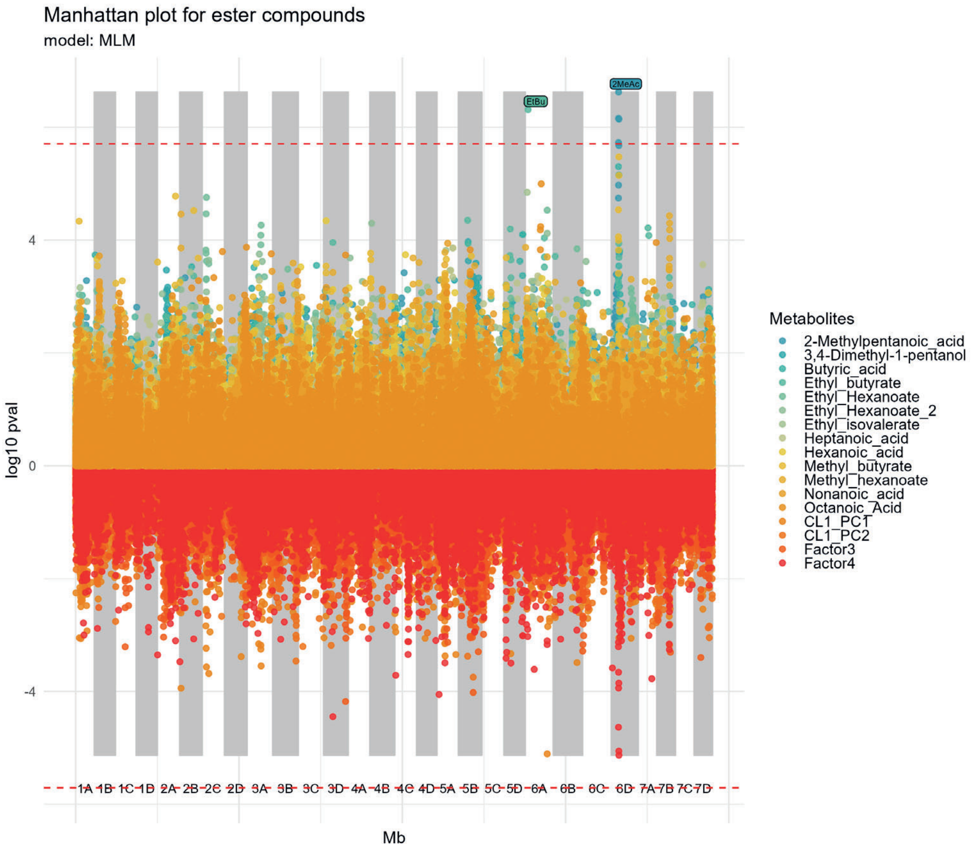
Supplementary figure 2: Correlation network of volatile expression in Rumba x Malwina population.

Lines between nodes represent (absolute) correlations above 0.5 and the width of the line is proportional to the correlation. **Right)** Colours correspond to clusters identified using a greedy modularity identification algorithm. **Left)** Colours correspond to the chemical group of each metabolite (FAD: fatty-acid derived). Metabolite names have been summarized: 12Di: 1,2-Dimethylcyclopentadiene; 2He: 2-Hexenal; Ac: Acetophenone; AlCa: alpha-Caryophyllene; AlPh: alpha-Phellandrene; AlpTer: alpha-terpinol; AlTe: alpha-Terpinene; Be: Benzaldehyde; BeAc: Benzoic acid; CiLiOx: cis-linalool oxide; DiEt: Dill ether; DLi: D-Limonene; EtHe: Ethyl hexanoate; GaDe: gamma-Decalactone; He: Hexanal; HeAc2: Hexyl acetate 2; IsAc: Isopropyl acetate; MeAc: Methyl acetate; MeBu: Methyl butyrate; MeHe: Methyl hexanoate; NoAc: Nonanoic acid; Oc: Ocimenol; OcAc: Octanoic Acid; OcBu: Octyl Butryrate; OcHe: Octyl hexanoate; PhFo: phenacyl formate; PMePr: p-Methoxybenzyl propionate; TrLiOx: trans-linalool oxide; ZZAlFa: Z,Z-alpha-Farnesene; UnXXX: Unknown XXX



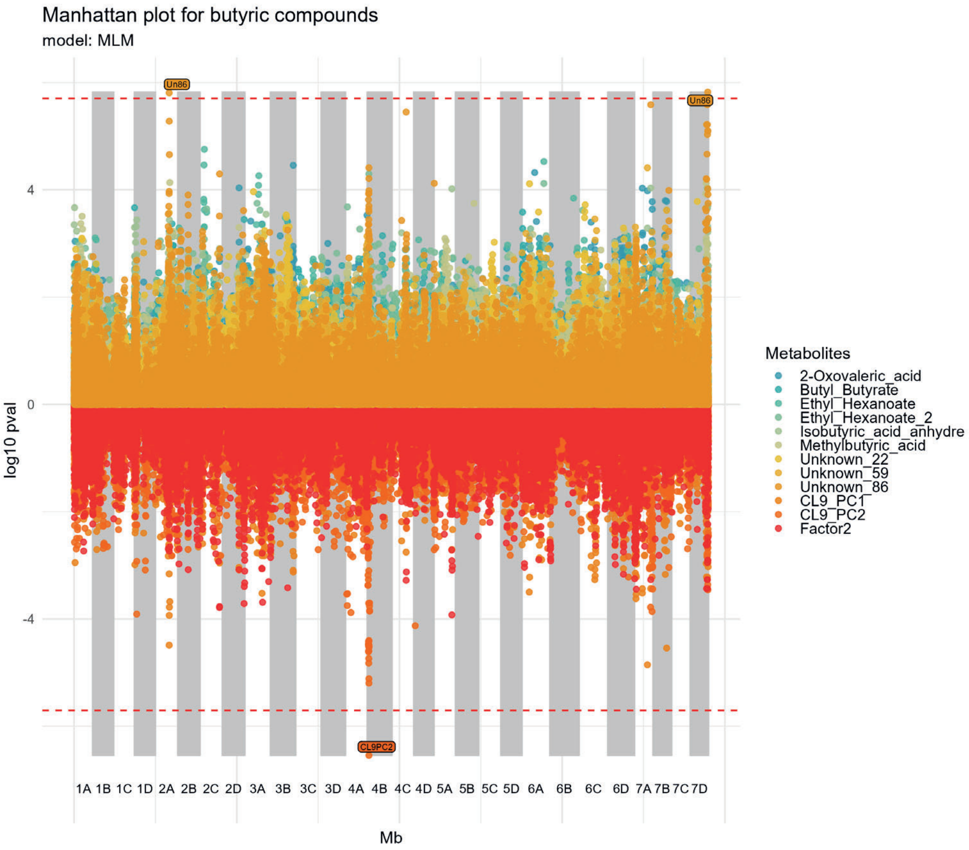
Supplementary figure 3: Heatmap of absolute loadings in a factor analysis of metabolite abundances in the Rumba x Malwina population, compared to correlation clusters.

Each row is one compound and each column is one factor. Darker colours represent a higher importance of that compound for that factor, measured by the absolute loading value. The left, coloured band indicates which rows belong to which cluster in the network-based correlation analysis.



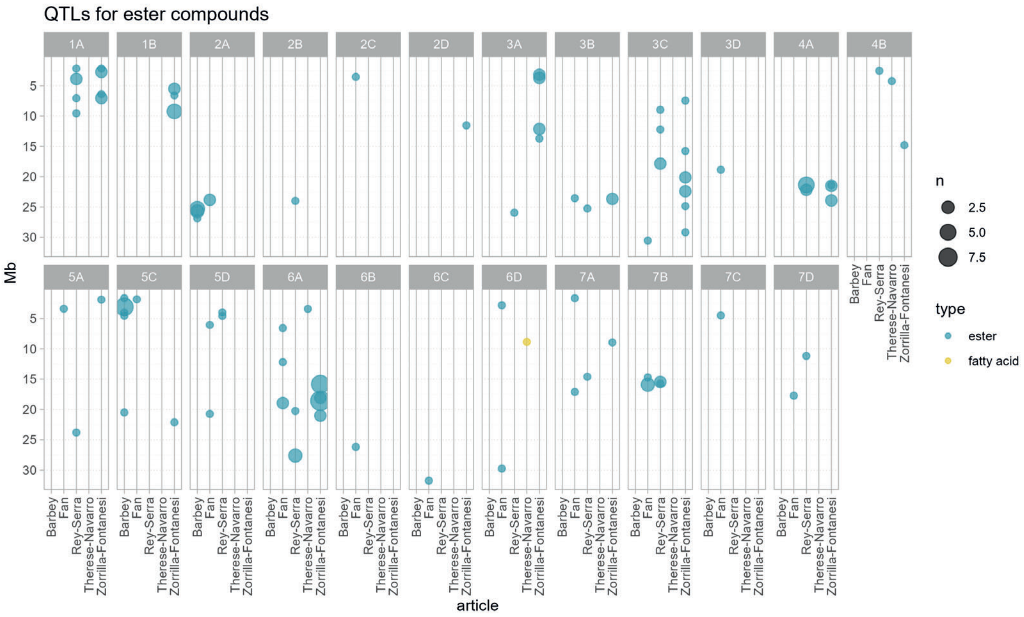
Supplementary figure 4: Overlapped Manhattan plots of ester compounds represented by Factor3, Factor4 and CL1.

The log10 p-values of multivariate phenotypes are expressed in negative (log10(pval)), while p-values obtained from real metabolites are expressed in positive (-log10(pval)). The three multivariate groups represented summarise a group of compounds mostly composed of esters, although other compounds are present. The red dashed lines represent the Bonferroni significance threshold. All significant points were labelled for easier interpretation.

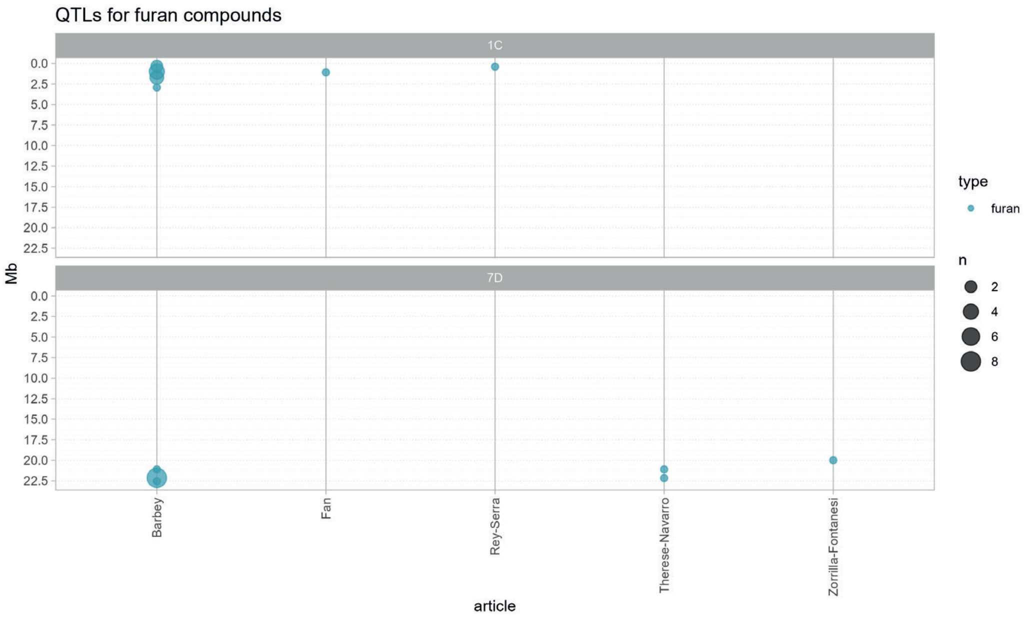


Supplementary figure 5: Overlapped Manhattan plot of butyric compounds represented by Factor2 and and CL9.

The log10 p-values of multivariate phenotypes are expressed in negative (log10(pval)), while p-values obtained from real metabolites are expressed in positive (-log10(pval)). The two multivariate groups represented summarise groups of compounds mostly composed of butyric-related compounds, although others are also present. The red dashed lines represent the Bonferroni significance threshold. All significant points were labelled for easier interpretation.



Supplementary figure 6: QTL positions of ester compounds across studies. Positions were imputed on the Royal Royce genome assembly. Compound QTL shown include: 2-butenoic acid, methyl ester, (e)-, 2-hexen-1-ol, acetate, (z)-, 2-hexenoic acid, methyl ester, 2-methylpentanoic acid, 2-metyhlbutyl acetate, 3-methylbutyl acetate, 3,4-dimethyl-1-pentanol, butanoic acid, 1-methyloctyl ester, butanoic acid, 2-methyl-, octyl ester, butanoic acid, 3-methyl-, octyl ester, butyl acetate, butyl butanoate, butyl hexanoate, butyl propanoate, butyl acetate, butyric acid, cinnamyl acetate, decyl acetate, decyl hexanoate, ethyl acetate, ethyl decanoate, ethyl hexanoate, ethyl butyrate, ethyl hexanoate, ethyl hexanoate 2, ethyl isovalerate, heptanoic acid, hexanethioic acid, s-methyl ester, hexanoic acid, hexyl acetate, hexyl butanoate, hexyl hexanoate, isoamyl hexanoate, isopropyl butanoate, isopropyl hexanoate, methyl 2-hexenoate, methyl 2-methylbutyrate, methyl anthranilate, methyl benzeneacetate, methyl benzoate, methyl butanoate, methyl decanoate, methyl hexanoate, methyl isovalerate, methyl nicotinate, methyl octanoate, methyl pentanoate, methyl salicylate, methyl thiolacetate, methyl butyrate, methyl hexanoate, metyhl acetate, myrtenil acetate, nonanoic acid, octanoic acid, octyl acetate, octyl butanoate, octyl hexanoate, propyl butanoate. For detailed positions refer to supplementary table 3.



Supplementary figure 7: QTL positions of furan compounds across studies.

Positions were imputed on the Royal Royce genome assembly. Compound QTL shown include furaneol (location on 1C) and mesifurane (location on 7D). For detailed positions refer to supplementary table 3.

Chapter 7

General Discussion

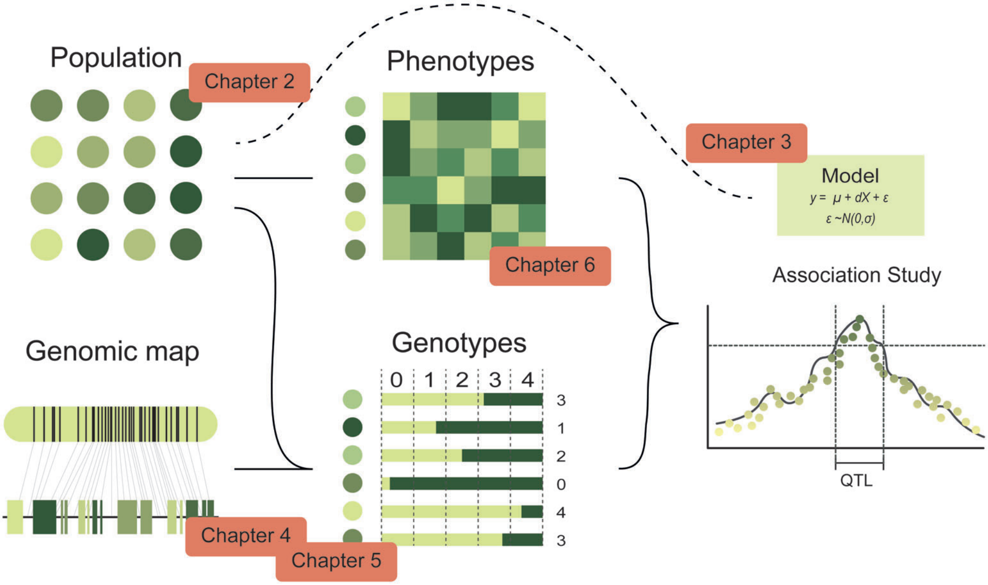


Figure 1: Visual summary of the contributions of this thesis within a typical quantitative genetics study.

A quantitative genetics study is composed of a population, its phenotypes, some genomic map and genotypes associated with such a map. These sources of information come together in the form of a statistical model, usually influenced by the population type, which can be leveraged to discover quantitative trait loci (QTL). In Chapter 2, multiparental population structures are reviewed, in particular for polyploid organisms. In Chapter 3 a new model is proposed that can harness information from a polyploid multiparental population to find association signals within the genome. In Chapter 4, Smooth Descent is developed, a new method that can improve genotyping accuracy and linkage map construction using skim-sequencing, a new type of genotyping approach. In Chapter 5, Smooth Descent is applied to the skim-sequencing genotype data of allopolyploid strawberry, highlighting that the obtained quality is equivalent to that of maps constructed with SNP arrays while increasing the number of mapped markers. In Chapter 6, metabolic profiles of strawberry aroma are analysed with multivariate methods and used for a QTL study using two multivariate approaches, showing how such an analysis can help identify major regulators of phenotypes, particularly in multivariate datasets.

General Discussion

A data-driven quantitative genetics study is based on a population of individuals, some type of map of the genome (usually a linkage map or genome sequence), and phenotypes and genotypes for those individuals (Fig. 1). A statistical model is chosen, partly informed by the population structure, which allows to perform an association study and find those genomic regions most associated with the examined phenotypes. The chapters presented in this thesis contribute to this growing array of analytical approaches, statistical methodologies that open doors to types of data that were previously not possible to study.

In Chapter 2 I discuss multiparental population structures for polyploid organisms, a type of population that is seldomly used in quantitative genetics, although it is very commonly found in breeding and research populations. To effectively use such populations, a statistical model that can adequately model such data needs to be developed. I present such model under the name mpQTL in Chapter 3, showing the increased accuracy of multiallelic markers compared to biallelic ones. In Chapter 4, I move towards linkage mapping, a method to obtain genetic maps from markers. Although linkage mapping is a well-known technique developed over a hundred years ago, it is still useful and informative, yet current methodologies are easily challenged by sequencing-based methods to obtain genetic markers. I address these challenges and propose a methodology suitable for polyploids, Smooth Descent, in Chapter 4. In Chapter 5, I use this same approach in a real population of strawberry, an allopolyploid with challenging genetics. Using error-prone skim-sequencing data I obtain genetic maps and compare their quality with maps obtained using SNP array markers (usually considered more accurate). The maps I produced were of similar quality to those using SNP array data, but with a hundredfold increase in the number of mapped markers. Lastly, in Chapter 6 I move towards QTL analyses again, this time focusing on multivariate phenotypes: the metabolic profile of strawberry aroma. I show that by adequately harnessing the multivariate information within this dataset, overarching regulatory trends can be confidently identified, helping organise an otherwise difficult study. With this approach we found and confirmed a major regulator of terpenoid biosynthesis and located it in the most recent genome assembly of strawberry, a crucial piece of information that was hitherto not reported.

In the section that follows I will dive into each of these topics in more detail, outlining the main findings of this thesis and current state of the art, while introducing those questions that in my view remain to be answered.

Moving towards multiallelic markers

In Chapter 2 I covered the application of QTL models to multiparental polyploid populations, a type of population that combines multiple linked families, akin to the complex pedigree structures of breeding programmes. I highlighted the need to track identity by descent (IBD) as the main innovation needed to move from biparental to multiparental populations, especially in polyploids. In Chapter 3 I present a new model and tool, mpQTL, which adapts the common mixed model framework for QTL analysis to the polyploid scenario. I tackled IBD-tracking by proposing the usage of a multiallelic model (using IBD-based alleles), in contrast to the classical biallelic SNP markers that are commonly used. Leveraging a simulation study, I was able to prove that multiallelic models are more precise in QTL studies than biallelic ones, an observation that has been echoed in similar research in the past years (Wang et al. 2016; Sallam et al. 2020; Bajgain and Anderson 2021; Li et al. 2021). Nevertheless, it was recently found that this improvement is not constant and depends on the structure of the multiparental population, sometimes being equivalent to the biallelic model (Li et al. 2021). Regardless, multiallelic models represent an interesting movement forward in QTL analysis, allowing to link biological alleles with their statistical effects on traits of interest. I anticipated that there was a crucial factor affecting the increase in accuracy of mpQTL: allele estimation. High throughput multiallelic genotyping is a challenging endeavour, more so in polyploid organisms. Now, it remains as the biggest hurdle to the application of multiallelic markers, which in turn are necessary for accurate multiparental analysis.

For the moment, the main method to estimate alleles in a population of related individuals is by IBD estimation. Obtaining IBD estimates in polyploid biparental populations can be readily achieved, with methods such as that used in Chapter 4, or other Hidden Markov Model (HMM) approaches (Mollinari and Garcia 2019; Zheng et al. 2021). In diploids, many IBD calculation tools are available. While most are oriented to specific population designs, R/qtl2 implements most crossing schemes (Broman et al. 2019), and RABBIT can use any multiparental design (Zheng et al. 2015). In contrast, for polyploids only polyOrigin is available, and its usage is limited to interconnected F1 populations of autotetraploids (Zheng et al. 2021). More complex multiparental designs like Multiparent Advanced Generation Inter-Cross (MAGIC), or pedigreed families common in breeding programmes remain out of the possibilities for current polyploid IBD estimation tools.

An alternative to probabilistic calculation of IBD is the usage of haplotypes. In Chapter 3 I have shown that concatenated biallelic markers can be turned into multiallelic markers that can closely predict IBD segregation. We might then turn to multiallelic haplotype markers in order to apply mpQTL to polyploid populations, but then the question follows, how does one obtain these haplotypes?

Haplotyping methods

Searching for the word “haplotyping” in any literature database yields hundreds of papers spanning multiple decades. Haplotyping has been a topic of interest for a long time. Unsurprisingly, there is an abundance of methodologies, problem formulations and tools, many applicable to polyploids. However, no golden standard or preferred method has appeared yet. Moreover, the nomenclature of these approaches is quite diffuse, with terms like haplotype phasing, haplotype inference or plain haplotyping being used interchangeably in vastly different contexts. Attempting to obtain haplotypes that would be suitable for multiallelic analysis in a multiparental polyploid population easily turns into the task of finding the best tool in this vast library of methods. To assess their suitability, I propose to divide haplotyping approaches based on the type of data they require. According to this, we find *marker phasing, read phasing, haplotype assembly* and *haplotype inference.*

The principle of **marker phasing** is to use only marker information and their position on a genome to obtain haplotypes. More often than not this approach targets populations of siblings, in which case the problem is resolved by phasing the parents, usually employing a Hidden Markov Model (Saada et al. 2022). This technique is in fact equivalent to the IBD estimation tools discussed above, which indeed can also be used to obtain haplotypes (Mollinari and Garcia 2019; Zheng et al. 2021). Other types of models leverage pedigree information outside the HMM framework (Motazedi et al. 2018, 2019; Voorrips and Tumino 2022), or simply use marker dosages from unrelated individuals to restrict the search space (Neigenfind et al. 2008; Su et al. 2008; Voorrips and Tumino 2022). Although the implementations differ, the results across this type of tools are similar. Haplotypes can be resolved locally, thus resulting in multiallelic markers, but they are not totally accurate, producing either missing values or incorrectly estimated haplotypes (Voorrips and Tumino 2022). Moreover, their limitations are equivalent to those described for IBD-based approaches. No tool is able to analyse complex pedigree structures. Nevertheless, specific parts of breeding programmes might be possible to haplotype using some of these tools, making mpQTL applicable in those cases.

Read phasing is another popular method of developing haplotypes. The process starts by comparing reads to a reference genome, obtaining then a set of polymorphic sites. Unlike in marker phasing, haplotypes are obtained by observing multiple polymorphisms in the same read. While in marker phasing haplotypes are built probabilistically, with read phasing the real haplotypes are observed within a read. As such, these methods can produce haplotypes for single individuals without the need of a population, although some can improve estimation by leveraging family or population information (e.g. Motazedi et al., 2018). **Haplotype assembly** approaches are similar to read phasing, but instead of simply phasing polymorphisms on a reference genome these techniques attempt to assemble the whole genome but obtaining one sequence per haplotype -instead of a single consensus sequence (Zhang et al. 2020; Michael and VanBuren 2020). The division between read phasing and haplotype assembly is often tenuous, since detecting read overlap is the main source of information in both approaches. Moreover, if a putative genome is *de novo* assembled with reads, any read phasing method could be used as a haplotype assembly. Many tools exist for diploids and polyploids. The main differences among them are the type of data used: low-coverage short reads, high-coverage short reads, long reads, chromatin contact data or a combination of these sources (Zhang et al. 2020; Michael and VanBuren 2020; Saada et al. 2022; Guk et al. 2022).

Both for read phasing and haplotype assembly approaches, one might question their applicability in the context of a multiparental QTL study. On the one hand, a multiparental population does contain family inheritance information that can be leveraged to improve haplotype estimation. Both methods tend to ignore such information, favouring the within-individual information only and thus do not seem the most appropriate in a multiparental context. On the other hand, applying these methods across a large population would require sequencing many samples and processing each set of reads, leading to a substantial increase in the experimental and computational cost. With much higher expenses, these approaches are unfeasible if large economic resources are not available.

Lastly, **haplotype inference** is another method of obtaining haplotypes. In this case, a haplotype library or reference panel is constructed first, and marker data are used to infer the most likely haplotype. This approach has had special relevance in human genetics, where exceptionally large reference panels of haplotype-phased genomes are already available (Browning et al. 2018; Ebler et al. 2022). Whether these methods are relevant to plant breeding is debatable. Firstly, plant breeding programmes usually have a relatively narrow set of founders. Assembling haplotypes for these founders could be enough to infer the haplotypes of all their descendants. However, it is often common to introduce exotic germplasm into breeding programmes in order to bring novel alleles. How would inference methods behave with such unknown haplotypes? Obtaining a haplotype confidence estimate would likely be crucial in these cases, but such parameter is not trivial to calculate.

Applicability: a need for standards

So, the question remains, which of these methods would be best for obtaining multiallelic markers? The most complete is haplotype assembly, but its excessive cost makes it unreasonable for large populations. The cheapest is marker phasing, and current tools seem to be usable in a breeding context with some interconnected families (although not with more complex designs). The accuracy of current tools has not been deeply studied and thus it is unclear how applicable they would be in a real multiparental context. Read phasing seems promising, but it is particularly difficult in polyploid organisms without expensive long-read sequencing. Overall, it remains unclear what tool performs best, an issue worsened by the lack of consensus on appropriate metrics to evaluate haplotyping methods (Saada et al. 2022). Some researchers have proposed to develop a benchmarking dataset to test all tools in similar conditions (Garg 2021; Saada et al. 2022; Guk et al. 2022). Considering the needs of multiparental polyploid populations and what has been highlighted by other authors, such dataset should include repetitive sequences, structural variation, polyploid organisms, a pedigree structure, and some population diversity, possibly simulating a wide variety of genotyping and sequencing technologies. An attempt to build such dataset has already been started with the "Phasing toolkit" of Saada *et al.* but will need to be expanded much further to encompass the wide range of tools currently available (Saada et al. 2022). Only when such a wide study is performed there will be clear answers as to when to apply each tool.

This type of deeply analytical and methodological research contrasts heavily with the applied interests of this field. In a recent haplotyping study, transcriptome analysis of autotetraploid potato revealed substantial allele-specific expression in 11% of genes (Sun et al. 2022). This is equivalent to specific haplotype alleles having a large effect on the phenotype, while others have a smaller effect, a situation that can be modelled within mpQTL. The situation is similar in allopolyploids for alleles across subgenomes: strawberry alleles have been shown to be differentially expressed, particularly across subgenomes (Lee et al. 2021). This is in line with the subgenome dominance hypothesis, which states that one subgenome dominates over others (Cheng et al. 2018; Bird et al. 2018). Although there is a clear trend, it is unclear whether allelic differences may change the directionality of subgenome dominance, a question that can only be answered with genome-wide studies of allelic effects. With mpQTL such research could be easily applied in a multiparental setting. Allelic effect estimations produced through the mixed model framework presented would highlight the importance of intra- and inter-subgenome variation, while also enabling the analysis of a wider range of alleles that traditional QTL studies under biparental populations ignore. Nevertheless, application of mpQTL to its full extent requires multiallelic genotyping that currently seems out of the realm of possibility, particularly for large populations. With current methods, haplotype estimation depends on obtaining large datasets of expensive long-read sequences for each individual. The most cost-effective paradigm would be to obtain short blocks of phased markers instead, with unclear and possibly heterogeneous accuracies across the genome. Alternatively, investing in the development of a representative haplotype panel to be used with an inference method would likely yield satisfactory results. Until a good benchmarking and evaluation method exist, the best tools to obtain multiallelic markers to apply in mpQTL will need to be studied case by case, as is being currently done.

Allopolyploid linkage mapping

In Chapters 4 and 5 we developed and applied a novel approach to use skim-sequencing in an allopolyploid linkage mapping context. The most crucial question for this piece of research was whether the increased error rate of skim-sequencing data would impede assigning markers to the linkage groups of each subgenome. Reads were aligned to a reference genome and there was a general concern that such alignment would not be accurate due to the high similarity between subgenomes, an issue that had clearly been a major obstacle in previous research (Kaur et al. 2012; Bassil et al. 2015; Edger et al. 2018). In that regard, the release of the allo-octoploid reference genome presented a unique opportunity (Edger et al. 2019). Aligning reads to the diploid *F. vesca* genome, as was planned before the release of the *F.* x *ananassa* genome, would have greatly complicated this research. An additional challenge was the great volume of data. Sequencing-based linkage maps require specific methodologies that can handle large or very large datasets, as well as large error rates (Liu et al. 2014; Rastas 2017; Bilton et al. 2018). Exactly how to perform such analysis was not clear from literature, especially considering the allopolyploid nature of strawberry.

Handling of error-prone genotype data

In Chapter 5 I proved that, with the appropriate cautions, linkage mapping can be performed with skim-sequencing data as accurately as with SNP array data. The key, however, was having a suitable method to handle genotyping errors since skim-sequencing data is extremely error prone, particularly in polyploid organisms. I opted to first eliminate highly spurious markers through a bin-based approach, followed by genotype correction using the Smooth Descent algorithm presented in Chapter 4. Although in practice genotype correction was performed, *i.e.,* a dosage score was changed to another score, the word "correction" might be a misnomer. It would be more accurate to call Smooth Descent an error detection and genotype imputation algorithm, since in essence potential genotyping errors are first detected, implicitly made missing, and substituted by an imputed genotype. Except for the error detection aspect, the rest of the approach is similar to other imputation methods usually applied to improve sequencing-based genotypes, especially those that capitalize on family relationships (Deschamps et al. 2012; Huang and Han 2014; Chung et al. 2017; Torkamaneh et al. 2018).

Imputation is sometimes perceived as an untrustworthy method to improve data, at least requiring further confirmation. However, there has been much consideration to its accuracy and the effects it can have on false positives (e.g. Hickey et al., 2012). When adequately applied, imputation is currently considered to be an accurate approach to improve genotyping data. In reference panel imputation, where a set of references are genotyped at high density and target samples are genotyped at much lower marker density, it has been shown that great accuracy can be achieved in diploids (Halperin and Stephan 2009; Das et al. 2018). Considering the research presented in Chapters 4 and 5, I agree with these observations. In the simulations and real data used to test Smooth Descent, it was clear that the algorithm is able to improve genotyping accuracy and linkage map reconstruction. The linkage maps produced in Chapter 5 showed the application of Smooth Descent in practice. I was able to obtain linkage maps that were highly colinear with the strawberry genome assembly and an equivalent linkage map obtained with higher accuracy SNP array genotypes. Thus, the ability of Smooth Descent to impute markers in this context is clearly established, as well as the procedure to generate linkage maps using skim-sequencing data.

As mentioned, the application of Smooth Descent in Chapter 5 was preceded by a bin-based removal of genotyping errors *i.e.,* markers with identical segregation across individuals were binned, and bins with few markers removed. The reasoning was that since we expect a great number of markers with identical segregation, those markers that are (close to) unique most probably contain many genotyping errors. Although we did not try, it would seem possible to apply Smooth Descent differently, as an imputation method. Instead of removing the most spurious markers from the analysis, those in bins with few markers, we could include them and attempt to use Smooth Descent to improve their genotyping. The only limitation to this approach is the estimation of marker order and position. In the research presented in Chapter 5, the marker order was obtained from genotype scores and thus highly spurious markers would probably not be correctly placed in these maps, nor correctly imputed. With a different source for marker order, however, such imputation would be possible, and this was in fact the approach that was taken in the research of Clot *et al.* (2022). This highlights that as an imputation algorithm Smooth Descent can probably be used beyond the application given to it in this thesis.

Subgenome sequence differentiation

Crucial for allopolyploid genetics is the differentiation between subgenomes. In the research conducted in Chapter 5 this was based on the analysis of co-segregating, fully linked markers and aided by the chromosome sequences of the "Camarosa" octoploid genome assembly (Edger et al. 2019). Within each group of co-segregating markers (within each bin) the proportion of physical positions assigned to each chromosome was tested. This revealed an interesting phenomenon. Firstly, bins containing more markers had confident assignments, while smaller bins had similar proportions of markers across chromosomes, leading to ambiguous assignments. This aligned well with our assumption that bins with more markers represented correct genotypes. Secondly, we also saw that within bins, even in those that were confidently assigned and that contained hundreds or thousands of identically segregating markers, we could find some "conflict" markers that originated from other chromosomes, according to the genome assembly. These markers represent incongruencies between the genetic data and physical data. Our analysis suggests that the conflicts originated in the genome assembly due to its inaccurate phasing. There seemed to be small regions that had been wrongly assigned to a homoeologous chromosome of a different subgenome.

The inaccuracies shown in Chapter 5 are not surprising given the complexity of haplotype assembly in allopolyploids (Zhang et al. 2020). Such issues have been recognised by the group that produced the "Camarosa" genome when they published a new *F.* x *ananassa* sequence named "Royal Royce" (Hardigan et al. 2021a). Since then, two other genome sequences have been released, "Wongyo" (Lee et al. 2021) and "Yanli" (Mao et al. 2023). The main strategy to avoid subgenome chimeras during the assembly of these genomes was the usage of long-read sequencing and chromatin contact data (HiC). For Royal Royce, trio-binning was applied to separate high-accuracy long reads (HiFi) into two haplotype bins; for Yanli, long reads, short reads, and HiC were combined to confidently phase haplotype sequences. As Wongyo is a highly homozygous variety, haplotype phasing was not performed and instead a single consensus sequence was obtained for each of the 28 chromosomes. In all cases, linkage mapping approaches akin to that described in Chapter 5 were used in order to identify putative chimeras and improve assembly quality. This highlights the importance of linkage mapping even in the era of haplotype-phased genome assemblies. Although a hefty investment in multiple sequencing approaches seems unavoidable when trying to assemble polyploid haplotypes, linkage mapping is clearly a uniquely useful source of information to understand genome structure. Once such genomes are available, I have shown that even skim-sequencing data can be used to obtain accurate linkage maps, even helping to point out inaccuracies in the reference genomes.

Remaining questions on allopolyploidy

After all the progress made in recent years on allopolyploid genomics, some biological questions remain poorly understood. The presence of orthologous genes in different subgenomes leads to an unavoidable question: which genes are controlling the phenotypes. Some argue that one subgenome contributes above all others, a hypothesis known as *subgenome dominance* (Alger and Edger 2020). In strawberry, it seems that subgenome A (from *F. vesca*) could behave as dominant, according to some expression data (Edger et al. 2019). However, this is not the case for all transcripts, with some subgenomes being dominant over others in different developmental stages (Lee et al. 2021). The underlying reasons for such dominance are poorly understood, although it is possible that they are part of the "genome shock" produced when two different species generate an allopolyploid hybrid (Soltis et al. 2012, 2016; Cheng et al. 2018). Edger *et al.* 2019 proposed that the dominance had a phylogenetic basis, meaning that all alleles originating from *F. vesca* would dominate over others. Others think this apparent bias might be due to a more accurate and complete reference genome for *F. vesca* and thus more reliable transcript identification from that subgenome. These hypotheses add another layer of complexity to the subgenome ancestry question, which remains unresolved.

Genome shock after allopolyploidization includes several symptoms. Transcriptional de-regulation is one of them, but also transposable element activation, chromosome rearrangements within and between subgenomes and overall meiotic instability (Ramsey and Schemske 2002; Soltis et al. 2012, 2015; van de Peer et al. 2017). This phenomenon can be clearly observed in triploid bananas since the ancestors are known (Baurens et al. 2019). Additionally, it has been well described that when *Fragaria* species meet in the wild they readily hybridize, producing fertile hybrids and more surprisingly *fertile aneuploids*, each with its own genomic rearrangement landscape (Bringhurst 1990; Liston et al. 2014). This type of polyploid evolution has been referred to as a "polyploid complex", local, temporary, unstable, multi-ploidy populations, from which a single stable polyploid emerges and expands to a wider ecological niche (Stebbins 1940, 1942). Nowadays this type of phenomenon is referred to as reticulate evolution and can be readily studied in natural populations of other species, such as the Eurasian goldilocks buttercup, *Ranunculus auricormus* (Karbstein et al., 2022). These observations would point to a mosaic ancestry hypothesis for strawberry, contradicting the notion that each subgenome must have a unique ancestor. Indeed, multiple authors have already hinted towards this (Liston et al. 2020; Hardigan et al. 2020). This was recently shown by Feng *et al.* (2021) using a variety of phylogenetic approaches, showing a heterozygous phylogeny within each chromosome, although no particular "ancestry mosaic map" was obtained. For this topic, it seems that more complete genome sequences of wild *Fragaria* species are required, since no phylogenetic study has been able to compare phylogeny across the >20 *Fragaria* species described to date. The haplotype-phased genome assemblies recently released, including that of the wild octoploid *F. chiloensis* (Cauret et al. 2022), will likely be helpful in this regard since they could help clarify whether mosaic ancestries are similar across individuals.

Fruit ripening and volatile production

One of the most important traits of strawberry is its unique aroma, produced by a mixture of volatile compounds. Due to the high natural variability of volatile production in strawberries, it is relevant to find key genetic regulators that control aroma. Moreover, as a complex multivariate trait, it offers an interesting challenge for traditional QTL analysis. In Chapter 6 of this thesis, I analysed such a metabolic dataset in strawberry, both for a biparental and a diverse population. I also gathered results across other similar studies in order to assess the reproducibility and validity of our own findings. By leveraging multivariate techniques, I was able to find an interesting locus on chromosome 3C that seems to control overall terpenoid abundance, that is, I found a major regulator of terpenoid production. I could confirm this finding in other studies after estimating the positions of all previously reported QTLs in the Royal Royce genome. This was in stark contrast to the results for esters, a compound class that is often considered the most important in strawberry (Dong et al. 2013; Ulrich and Olbricht 2016; Fan et al. 2021; Rey-Serra et al. 2022). We did not find any meaningful QTL signals for esters in our study and across the literature there was little agreement. The issues behind this lack of repeatability have already been reviewed and are most likely due to biological, experimental, and analytical reasons (Ulrich et al. 2018). Variability of fruit ripening and flowering times might heavily contribute to the apparent lack of relevant QTL signals for esters.

Although we did find a major regulator, I believe the analysis performed and reported on in Chapter 6 suggests that this type of metabolic QTL study is not well suited for understanding the dynamic process of strawberry ripening and its secondary metabolism. Instead, an approach that focuses on studying metabolic profiles through the development and growth of the strawberry fruit and its shifting metabolism might be more appropriate. Ripening has been heavily studied, although much remains to be uncovered in strawberry.

The physiology of ripening

As a horticultural crop, the importance of fruit set and ripening in strawberry for agriculture is self-evident. The edible portion of strawberry is in fact not a fruit, but an enlarged receptacle. Technically, the fruits are the seeds that cover strawberries, the achenes. Strawberry ripening is a dynamic process that starts with auxin production and transport in the achenes, followed by a decrease in auxin levels and an increase in abscisic acid (ABA) (Cherian et al. 2014; Li et al. 2022a). Different processes must be synchronised to obtain a ripe strawberry: fruit enlargement, sugar accumulation, fruit softening through cell wall degradation and production of secondary metabolites like flavonoids, esters or terpenoid compounds (McAtee et al. 2013). Strawberry is a non-climacteric fruit, meaning that ethylene does not play a significant role in strawberry ripening – although it is not entirely absent. There has been much research into the regulation and control of ripening in strawberry, owing in part to the simplicity of transformation and gene silencing techniques (Folta and Davis 2007; Guidarelli and Baraldi 2015). Despite extensive knowledge of up- and down-regulated genes during fruit ripening, the network of molecular regulation of strawberry is largely unknown (Cherian et al. 2014). The interplay of such regulation network with environmental factors has not been explored, despite the clear observations that light and temperature affect strawberry shape and flavour (Carbone et al. 2009; Tulipani et al. 2011; Alvarez-Suarez et al. 2014; Warner et al. 2021; Leonardou et al. 2021). Importantly, virtually all studies have highlighted that environmental impact on quality traits is genotype dependent. This is unsurprising, given that adaptation to climatic conditions has always been an important goal of strawberry breeding programmes.

As previously pointed out, the availability of an octoploid genome assembly will clearly aid in the discovery of ripening-related genes, facilitating RNA expression studies and the subsequent inference of regulatory networks. The datasets generated in Chapter 6 could easily be linked to transcriptomic data to predict regulatory networks much more effectively than through QTL analysis. Moreover, some of the analytical methods published to reconstruct regulatory networks are well-suited for the study of environmental effects on gene expression (e.g. Li et al. 2015; Jones and Vandepoele 2020). Octoploid strawberry would be an interesting model to study regulatory network evolution, since the presence of the four subgenomes has likely contributed to sub-functionalization and specialization of the key regulators of ripening. Such studies would require extensive temporal, environmental and transcriptomic data, paired with a sophisticated analysis that could link findings to the newly published genome. The recent advances in strawberry genomics will undoubtedly propel more sophisticated molecular research that, until now, has only been possible in better characterized model species.

Multivariate analysis of plant molecular data

Chapter 6 introduced multivariate analysis as a useful way to reduce and study otherwise large and obscure datasets. A not-so-common method in the metabolomics field was applied, *factor analysis*, which showed much promise. Unlike network-based approaches that are more commonly applied in metabolomic studies, factor analysis is simple and reproducible, with a single hyperparameter that needs to be tuned (the number of latent factors). The technique proved useful in identifying groups of correlated variables and their common genetic associations and thus seems promising for future analyses. I have clearly shown the usefulness of multivariate approaches to dissecting phenotypic traits, an interest that is growing in recent years.

The need to apply a dimension-reduction tool to understand a plant dataset is not unique to metabolomic datasets. There is a growing field of research that is oriented towards generating large datasets in an automated fashion by using sensors and imaging technology (Coppens et al. 2017; Pieruschka and Schurr 2019; Yang et al. 2020). Such big datasets present several challenges, among them the need to somehow reduce their complexity to essential features that are more explanatory than the hundreds of variables gathered in automated phenotyping experiments. The QTL analysis of Chapter 6 presented a relatively traditional approach to phenotype-genotype association. Although we added a multivariate aspect to it, the nature of the association model was in essence identical to that proposed by Lander and Botstein in 1989 –albeit with adaptations and improvements developed later (Lander and Botstein 1989; Yu and Buckler 2006; Wang and Zhang 2021). More modern approaches to detect association that intelligently leverage the multivariate properties of large datasets are englobed in *genomic prediction.* The main principle behind genomic prediction is to use the entire genome to predict traits, comparing the predictions to true, evaluated datasets. The multivariate approaches used to dissect large genotypic data in genomic prediction can contribute many lessons to multivariate analysis of phenotypes. However, as larger phenotyping datasets are made available, such prediction models also need to tackle with the increased phenotypic dimensionality problem. Such a combination of multivariate genotypes and phenotypes has been reviewed in detail in a recent statistical manual (Montesinos López et al. 2022). Through a multivariate lens, the statistical models developed in the genomic prediction field will likely play a key role in understanding increasingly complex phenotypic datasets, including but not limited to metabolomic datasets.

One might wonder what the relevance of traditional QTL studies in this genomic prediction context is. I would argue that the findings produced by the analysis of Chapter 6 could not be produced by a genomic prediction approach. Only with such a QTL analysis, powered by multivariate techniques, we can leverage phenotype information to find major genetic regulators that point to *specific loci*, which in turn open the door to further molecular research. In the case of Chapter 6, a locus that controls the production of flowery, herbal and citrussy aromas in strawberry, that controls the production of terpenoids.

Computational genetics

In the past years, we have experienced a veritable social revolution. With the expansion of the internet, the popularization of personal computers and the generation of ever-larger datasets that carefully describe our societies, no aspect of our lives has been left untouched by information technologies. The thesis you have in your hands (or perhaps in front of you through a screen) is clearly a consequence of these changes in the world, particularly in the field of plant genetic research applied to plant breeding. Vast amounts of genetic, phenotypic, and environmental data are being acquired, sprouting in us feelings of possibility, of that scientific hope one feels when a discovery seems at the tip of your fingers. The road towards these discoveries is paved with quantitative genetics, the hybrid child of applied statistics, genetics, and molecular biology. Unsurprisingly, the tools of quantitative genetics now heavily rely on the same information systems: databases to hold and distribute data, algorithms to efficiently resolve mathematical problems and programs to integrate algorithms into useful statistical tools, all are becoming the standard toolset of any quantitative geneticist. This expansion and transformation of our methods gives us, the statistically minded geneticists, a dynamic and exhilarating work environment to apply our skills. Naturally, such changes bring with them important challenges, particularly as we transition from classical to data-driven plant breeding.

There are plenty of excellent reviews on the topic of large data in plant breeding. Some focus on the statistical methods needed to utilize this data, particularly in the genomic prediction context (Azodi et al. 2019; Tong and Nikoloski 2021). Others highlight the need for adequate databases for data management: since expensive experimental data usually needs to be stored, accessed, and shared across multiple parties, including for publication, an adequate database infrastructure seems indispensable in order to easily manage data (Pieruschka and Schurr 2019). These observations have propelled international coordination, with initiatives like ELIXIR, a European institution that aims to provide best practices, tool repositories and general data literacy education to researchers of the life sciences (Harrow et al. 2021). A new interdisciplinary field of science is emerging, where data collection, management, analysis, and dissemination meet. To fulfil those needs, new types of research roles are emerging, like that of the *data steward* (Arend et al. 2022). As a librarian or archivist did in the past, a steward's job is to ensure data quality, manage its organisation and promote its storage on standardised systems that facilitate data integration. Also, the *scientific programmer* has appeared, the researcher whose main job is to develop software tools, usually including experimental approaches that implement new analytical ideas. Best practices have been published regarding this type of software development, focused on program design and structure (Wilson et al. 2014; Artaza et al. 2016) and on software management and dissemination (Alves et al. 2021). It will not be difficult to notice that this thesis greatly hinges between these roles, managing and using data and developing and storing software. Proofs of this are the presented software tools, mpQTL and Smooth Descent, that will need to be adequately maintained in the future. The existence of these new roles and their required skills also highlights the importance of education renewal, adding learning goals to curriculums that reflect the changing needs of the field (Arend et al. 2022).

Finally, to complete this thesis, I can say that I have managed to answer what is likely the most puzzling question of any PhD research: what was I researching? I am confident to say that computational genetics has been my specialization. I have worked to expand and upgrade statistical methodologies to new computational practices, addressing the gaps left by their original proponents regarding the usage of polyploid data. For this, I have learned programming, high-efficient computing, new statistical methods, automated data analysis and other topics that are probably closer to computer sciences than to biology. I have also deeply dived into more classically genetic topics, the polyploidization history of strawberry and its unclear ancestry, subgenome differentiation and linkage mapping, the physiology of fruit ripening and the incredibly diverse metabolic products of this little red fruit. As my research has been interdisciplinary, so has my focus shifted between fundamental research on the mathematical properties of multiallelic, polyploid QTL models and applied research on the suitability of skim-sequencing data and interpretability of big phenotypic data. The application of multiallelic models in complex population structures or the clear differentiation between allopolyploid subgenomes and their relevance to complex traits like metabolic profiles will require further research to be understood. With this thesis, I hope to have contributed to resolve some uncertainties, and more importantly, I hope I have opened new mysteries for the broader scientific community. I can certainly say that after all this learning I have more questions than ever before. How exciting!

References

Abbott S, Fairbanks DJ (2016) Experiments on Plant Hybrids by Gregor Mendel. Genetics 204:407. https://doi.org/10.1534/GENETICS.116.195198

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101. https://doi.org/10.1038/ng786

Adhikari L, Shrestha S, Wu S, et al (2022) A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations. Scientific Reports 2022 12:1 12:1–12. https://doi.org/10.1038/s41598-022-19858-2

Aguiar D, Istrail S (2012) HapCompass: A Fast Cycle Basis Algorithm for Accurate Haplotype Assembly of Sequence Data. Journal of Computational Biology 19:577–590. https://doi.org/10.1089/cmb.2012.0084

Aharoni A, Giri AP, Verstappen FWA, et al (2004) Gain and loss of fruit flavour compounds produced by wild and cultivated strawberry species. Plant Cell 16:3110–3131. https://doi.org/10.1105/TPC.104.023895

Aharoni A, Keizer LCP, Bouwmeester HJ, et al (2000) Identification of the SAAT Gene Involved in Strawberry Flavor Biogenesis by Use of DNA Microarrays. Plant Cell 12:647. https://doi.org/10.2307/3870992

Akond Z, Alam MdJ, Hasan MN, et al (2019) A Comparison on Some Interval Mapping Approaches for QTL Detection. Bioinformation 15:90. https://doi.org/10.6026/97320630015090

Al Bkhetan Z, Chana G, Ramamohanarao K, et al (2021) Evaluation of consensus strategies for haplotype phasing. Brief Bioinform 22:1–12. https://doi.org/10.1093/BIB/BBAA280

Alger EI, Edger PP (2020) One subgenome to rule them all: underlying mechanisms of subgenome dominance. Curr Opin Plant Biol 54:108–113. https://doi.org/10.1016/J.PBI.2020.03.004

Alvarez-Suarez JM, Mazzoni L, Forbes-Hernandez TY, et al (2014) The effects of pre-harvest and post-harvest factors on the nutritional quality of strawberry fruits: A review. J Berry Res 4:1–10. https://doi.org/10.3233/JBR-140068

Alves R, Bampalikis D, Castro LJ, et al (2021) ELIXIR Software Management Plan for Life Sciences. https://doi.org/10.37044/osf.io/k8znb

Anciro A, Mangandi J, Verma S, et al (2018) FaRCg1: a quantitative trait locus conferring resistance to Colletotrichum crown rot caused by Colletotrichum gloeosporioides in octoploid strawberry. Theoretical and Applied Genetics 131:2167–2177. https://doi.org/10.1007/s00122-018-3145-z

Arend D, Psaroudakis D, Memon JA, et al (2022) From data to knowledge – big data needs stewardship, a plant phenomics perspective. The Plant Journal 111:335–347. https://doi.org/10.1111/TPJ.15804

Artaza H, Hong NC, Corpas M, et al (2016) Top 10 metrics for life science software good practices. F1000Res 5:. https://doi.org/10.12688/F1000RESEARCH.9206.1

Azodi CB, Bolger E, McCarren A, et al (2019) Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. G3: Genes|Genomes|Genetics 9:3691. https://doi.org/10.1534/G3.119.400498

Bajgain P, Anderson JA (2021) Multi-allelic haplotype-based association analysis identifies genomic regions controlling domestication traits in intermediate wheatgrass. Agriculture (Switzerland) 11:667. https://doi.org/10.3390/AGRICULTURE11070667/S1

Bajgain P, Rouse MN, Tsilo TJ, et al (2016) Nested Association Mapping of Stem Rust Resistance in Wheat Using Genotyping by Sequencing. PLoS One 11:e0155760. https://doi.org/10.1371/journal.pone.0155760

Barbey CR, Hogshead MH, Harrison B, et al (2021) Genetic Analysis of Methyl Anthranilate, Mesifurane, Linalool, and Other Flavor Compounds in Cultivated Strawberry (Fragaria × ananassa). Front Plant Sci 12:718. https://doi.org/10.3389/fpls.2021.615749

Bardol N, Ventelon M, Mangin B, et al (2013) Combined linkage and linkage disequilibrium QTL mapping in multiple families of maize (Zea mays L.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. Theoretical and Applied Genetics 126:2717–2736. https://doi.org/10.1007/s00122-013-2167-9

Bassil N v, Davis TM, Zhang H, 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. https://doi.org/10.1186/s12864-015-1310-1

Baurens FC, Martin G, Hervouet C, et al (2019) Recombination and Large Structural Variations Shape Interspecific Edible Bananas Genomes. Mol Biol Evol 36:97–111. https://doi.org/10.1093/molbev/msy199

Bayer PE, Golicz AA, Scheben A, et al (2020) Plant pan-genomes are the new reference. Nature Plants 2020 6:8 6:914–920. https://doi.org/10.1038/s41477-020-0733-0

Benzer S (1959) ON THE TOPOLOGY OF THE GENETIC FINE STRUCTURE. Proceedings of the National Academy of Sciences 45:1607–1620. https://doi.org/10.1073/pnas.45.11.1607

Berger E, Yorukoglu D, Peng J, Berger B (2014) HapTree: A Novel Bayesian Framework for Single Individual Polyplotyping Using NGS Data. PLoS Comput Biol 10:e1003502. https://doi.org/10.1371/journal.pcbi.1003502

Bilton TP, Schofield MR, Black MA, et al (2018) Accounting for errors in low coverage high-throughput sequencing data when constructing genetic maps using Biparental outcrossed populations. Genetics 209:65–76. https://doi.org/10.1534/genetics.117.300627

Bink MCAM, Jansen J, Madduri M, et al (2014) Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. Theoretical and Applied Genetics 127:1073–1090. https://doi.org/10.1007/s00122-014-2281-3

Bink MCAM, Totir LR, ter Braak CJF, et al (2012) QTL linkage analysis of connected populations using ancestral marker and pedigree information. Theoretical and Applied Genetics 124:1097–1113. https://doi.org/10.1007/s00122-011-1772-8

Birchler JA (2012) Genetic Consequences of Polyploidy in Plants. In: Polyploidy and Genome Evolution. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 21–32

Bird KA, VanBuren R, Puzey JR, Edger PP (2018) The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytologist 220:87–93

Blanc G, Charcosset A, Mangin B, et al (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: An application in maize. Theoretical and Applied Genetics 113:206–224. https://doi.org/10.1007/s00122-006-0287-1

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314

Bouchet S, Olatoye MO, Marla SR, et al (2017) Increased Power To Dissect Adaptive Traits in Global Sorghum Diversity Using a Nested Association Mapping Population. Genetics 206:573–585. https://doi.org/10.1534/genetics.116.198499

Bourke PM, van Geest G, Voorrips RE, et al (2018a) polymapR—linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. Bioinformatics 34:3496–3502. https://doi.org/10.1093/bioinformatics/bty371

Bourke PM, Voorrips RE, Hackett CA, et al (2021) Detecting quantitative trait loci and exploring chromosomal pairing in autopolyploids using polyqtlR. Bioinformatics 37:3822–3829. https://doi.org/10.1093/bioinformatics/btab574

Bourke PM, Voorrips RE, Kranenburg T, et al (2016) Integrating haplotype-specific linkage maps in tetraploid species using SNP markers. Theoretical and Applied Genetics 129:2211–2226. https://doi.org/10.1007/s00122-016-2768-1

Bourke PM, Voorrips RE, Visser RGF, Maliepaard C (2018b) Tools for Genetic Studies in Experimental Populations of Polyploids. Front Plant Sci 9:513. https://doi.org/10.3389/fpls.2018.00513

Bringhurst RS (1990) Cytogenetics and evolution in American Fragaria. HortScience 25:879–881. https://doi.org/10.1016/j.scitotenv.2014.02.075

Broman KW, Gatti DM, Simecek P, et al (2019) R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. Genetics 211:495–502. https://doi.org/10.1534/GENETICS.118.301595

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890. https://doi.org/10.1093/bioinformatics/btg112

Browning BL, Browning SR (2011a) A Fast, Powerful Method for Detecting Identity by Descent. The American Journal of Human Genetics 88:173–182. https://doi.org/10.1016/J.AJHG.2011.01.010

Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics 103:338–348. https://doi.org/10.1016/J.AJHG.2018.07.015

Browning SR, Browning BL (2011b) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12:703–714. https://doi.org/10.1038/nrg3054

Bruce TJA, Wadhams LJ, Woodcock CM (2005) Insect host location: a volatile situation. Trends Plant Sci 10:269–274. https://doi.org/10.1016/J.TPLANTS.2005.04.003

Brůna T, Hoff KJ, Lomsadze A, et al (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform 3:1–11. https://doi.org/10.1093/NARGAB/LQAA108

Brzozowski LJ, Campbell MT, Hu H, et al (2022) Generalizable approaches for genomic prediction of metabolites in plants. Plant Genome 15:e20205. https://doi.org/10.1002/TPG2.20205

Carbone F, Preuss A, de Vos RCH, et al (2009) Developmental, genetic and environmental factors affect the expression of flavonoid genes, enzymes and metabolites in strawberry fruits*. Plant Cell Environ 32:1117–1131. https://doi.org/10.1111/J.1365-3040.2009.01994.X

Carey SB, Aközbek L, Harkess A (2022) The contributions of Nettie Stevens to the field of sex chromosome biology. Philosophical Transactions of the Royal Society B 377:. https://doi.org/10.1098/RSTB.2021.0215

Cartwright DA, Troggio M, Velasco R, Gutin A (2007) Genetic Mapping in the Presence of Genotyping Errors. Genetics 176:2521–2527. https://doi.org/10.1534/genetics.106.063982

Castillejo C, Waurich V, Wagner H, et al (2020) Allelic Variation of MYB10 Is the Major Force Controlling Natural Variation in Skin and Flesh Color in Strawberry (Fragaria spp.) Fruit. Plant Cell 32:3723–3749. https://doi.org/10.1105/TPC.20.00474

Cauret CMS, Mortimer SME, Roberti MC, et al (2022) Chromosome-scale assembly with a phased sex-determining region resolves features of early Z and W chromosome differentiation in a wild octoploid strawberry. G3 Genes|Genomes|Genetics 12:. https://doi.org/10.1093/G3JOURNAL/JKAC139

Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221. https://doi.org/10.1016/j.pbi.2008.01.002

Chan AW, Hamblin MT, Jannink JL (2016) Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data. PLoS One 11:e0160733. https://doi.org/10.1371/journal.pone.0160733

Cheema J, Dicks J (2009) Computational approaches and software tools for genetic linkage map estimation in plants. Brief Bioinform 10:595–608. https://doi.org/10.1093/bib/bbp045

Cheng F, Wu J, Cai X, et al (2018) Gene retention, fractionation and subgenome differences in polyploid plants. Nat Plants 4:258–268

Cherian S, Figueroa CR, Nair H (2014) 'Movers and shakers' in the regulation of fruit ripening: a cross-dissection of climacteric versus non-climacteric fruit. J Exp Bot 65:4705–4722. https://doi.org/10.1093/JXB/ERU280

Chung YS, Choi SC, Jun TH, Kim C (2017) Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. Hortic Environ Biotechnol 58:425–431. https://doi.org/10.1007/S13580-017-0297-8/METRICS

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971. https://doi.org/10.1007/s11703-007-0022-y

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 70:6. https://doi.org/10.1103/PHYSREVE.70.066111/FIGURES/3/MEDIUM

Clavijo McCormick A, Unsicker SB, Gershenzon J (2012) The specificity of herbivore-induced plant volatiles in attracting herbivore enemies. Trends Plant Sci 17:303–310. https://doi.org/10.1016/J.TPLANTS.2012.03.012

Clevenger JP, Korani W, Ozias-Akins P, Jackson S (2018) Haplotype-Based Genotyping in Polyploids. Front Plant Sci 9:564. https://doi.org/10.3389/fpls.2018.00564

Clot CR, Wang X, Koopman J, et al (2022) High-density linkage map constructed from a skim sequenced diploid potato population reveals transmission distortion and QTLs for tuber and pollen production. https://doi.org/10.21203/RS.3.RS-2302091/V1

Colle M, Leisner CP, Wai CM, et al (2019) Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. Gigascience. https://doi.org/10.1093/gigascience/giz012

Comai L (2005) The advantages and disadvantages of being polyploid. Nat Rev Genet 6:836–846

Coppens F, Wuyts N, Inzé D, Dhondt S (2017) Unlocking the potential of plant phenotyping data through integration and data-driven approaches. Curr Opin Syst Biol 4:58–63. https://doi.org/10.1016/J.COISB.2017.07.002

Corcos AF, Monaghan F v. (2008) Mendel's work and its rediscovery: A new perspective. https://doi.org/101080/07352689009382287 9:197–212. https://doi.org/10.1080/07352689009382287

Creighton HB, McClintock B (1931) A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. Proceedings of the National Academy of Sciences 17:492–497. https://doi.org/10.1073/PNAS.17.8.492/ASSET/1B33E98F-B797-4CAB-A0EE-00035DC80D28/ASSETS/PNAS.17.8.492.FP.PNG

Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Sy:1695

Darrow GM (1966) The Strawberry: history, breeding and Physiology. THE NEW ENGLAND INSTITUTE FOR MEDICAL RESEARCH

Das S*, *Abecasis GR*, *Browning BL (2018) Genotype Imputation from Large Reference Panels. https://doi.org/101146/annurev-genom-083117-021602 19:73–96. https://doi.org/10.1146/ANNUREV-GENOM-083117-021602

Das S*, *Vikalo H (2015) SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. BMC Genomics 16:260. https://doi.org/10.1186/s12864-015-1408-5

Davis TM*, *DiMeglio LM*, *Yang R, et al (2006) Assessment of SSR Marker Transfer from the Cultivated Strawberry to Diploid Strawberry Species: Functionality, Linkage Group Assignment, and Use in Diversity Analysis. Journal of the American Society for Horticultural Science 131:506–512. https://doi.org/10.21273/JASHS.131.4.506

de Boer JG*, *Posthumus MA*, *Dicke M (2004) Identification of Volatiles That Are Used in Discrimination Between Plants Infested with Prey or Nonprey Herbivores by a Predatory Mite. Journal of Chemical Ecology 2004 30:11 30:2215–2230. https://doi.org/10.1023/B:JOEC.0000048784.79031.5E

Debik J*, *Sangermani M*, *Wang F, et al (2022) Multivariate analysis of NMR-based metabolomic data. NMR Biomed 35:e4638. https://doi.org/10.1002/NBM.4638

Defilippi BG*, *Manríquez D*, *Luengwilai K*, *González-Agüero M (2009) Chapter 1 Aroma Volatiles: Biosynthesis and Mechanisms of Modulation During Fruit Ripening. Adv Bot Res 50:1–37. https://doi.org/10.1016/S0065-2296(08)00801-X

Delaneau O*, *Zagury JF*, *Robinson MR, et al (2019) Accurate, scalable and integrative haplotype estimation. Nature Communications 2019 10:1 10:1–10. https://doi.org/10.1038/s41467-019-13225-y

Deschamps S*, *Llaca V*, *May GD (2012) Genotyping-by-Sequencing in Plants. Biology 2012, Vol 1, Pages 460-483 1:460–483. https://doi.org/10.3390/BIOLOGY1030460

Doerge RW*, *Craig BA (2000) Model selection for quantitative trait locus analysis in polyploids. Proc Natl Acad Sci U S A 97:7951–6. https://doi.org/10.1073/pnas.97.14.7951

Dominique DD*, *Poorten TJ*, *Hardigan MA, et al (2018) Genome-Wide Association Mapping Uncovers Fw1, a Dominant Gene Conferring Resistance to Fusarium Wilt in Strawberry. G3 Genes|Genomes|Genetics 8:1817–1828. https://doi.org/10.1534/G3.118.200129

Dong J*, *Zhang Y*, *Tang X, et al (2013) Differences in volatile ester composition between Fragaria × ananassa and F. vesca and implications for strawberry aroma patterns. Sci Hortic 150:47–53. https://doi.org/10.1016/J.SCIENTA.2012.11.001

Ebler J*, *Ebert P*, *Clarke WE, et al (2022) Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nature Genetics 2022 54:4 54:518–525. https://doi.org/10.1038/s41588-022-01043-w

Edge P*, *Bafna V*, *Bansal V (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res 27:801–812. https://doi.org/10.1101/GR.213462.116

Edger PP*, *McKain MR*, *Bird KA*, *VanBuren R (2018) Subgenome assignment in allopolyploids: challenges and future directions. Curr Opin Plant Biol 42:76–80

Edger PP*, *McKain MR*, *Yocca AE, et al (2020) Reply to: Revisiting the origin of octoploid strawberry. Nat Genet 52:5–7

Edger PP*, *Poorten TJ*, *VanBuren R, et al (2019) Origin and evolution of the octoploid strawberry genome. Nat Genet 51:541–547. https://doi.org/10.1038/s41588-019-0356-4

Effah E*, *Holopainen JK*, *McCormick AC (2019) Potential roles of volatile organic compounds in plant competition. Perspect Plant Ecol Evol Syst 38:58–63. https://doi.org/10.1016/J.PPEES.2019.04.003

Elshire RJ*, *Glaubitz JC*, *Sun Q, et al (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS One 6:e19379. https://doi.org/10.1371/journal.pone.0019379

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal 4:250. https://doi.org/10.3835/plantgenome2011.08.0024

Fan Z, Hasing T, Johnson TS, et al (2021) Strawberry sweetness and consumer preference are enhanced by specific volatile compounds. Hortic Res 8:66. https://doi.org/10.1038/S41438-021-00502-5/42042927/41438_2021_ARTICLE_502.PDF

Fan Z, Tieman DM, Knapp SJ, et al (2022) A multi-omics framework reveals strawberry flavor genes and their regulatory elements. New Phytologist. https://doi.org/10.1111/nph.18416

Feng C, Wang J, Harris AJ, et al (2021) Tracing the Diploid Ancestry of the Cultivated Octoploid Strawberry. Mol Biol Evol 38:478–485. https://doi.org/10.1093/MOLBEV/MSAA238

Ferrão LF V., Benevenuto J, Oliveira I de B, et al (2018) Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context. Front Ecol Evol 6:107. https://doi.org/10.3389/fevo.2018.00107

Fierst JL (2015) Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Front Genet 6:220. https://doi.org/10.3389/fgene.2015.00220

Filomena Valim M, Rouseff RL, Lin J (2003) Gas Chromatographic–Olfactometric Characterization of Aroma Compounds in Two Types of Cashew Apple Nectar. J Agric Food Chem 51:1010–1015. https://doi.org/10.1021/jf025738+

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–74. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Folta KM, Davis TM (2007) Strawberry Genes and Genomics. https://doi.org/101080/07352680600824831 25:399–415. https://doi.org/10.1080/07352680600824831

Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol Ecol 26:5369–5406. https://doi.org/10.1111/MEC.14264

Galesloot TE, Van Steen K, Kiemeney LALM, et al (2014) A Comparison of Multivariate Genome-Wide Association Methods. PLoS One 9:e95923. https://doi.org/10.1371/JOURNAL.PONE.0095923

Galletta GJ, Maas JL (1990) Strawberry Genetics. HortScience 25:871–879. https://doi.org/10.21273/HORTSCI.25.8.871

Ganal MW, Polley A, Graner EM, et al (2012) Large SNP arrays for genotyping in crop plants. J Biosci 37:821–828. https://doi.org/10.1007/S12038-012-9225-3/TABLES/1

Garg S (2021) Computational methods for chromosome-scale haplotype reconstruction. Genome Biology 2021 22:1 22:1–24. https://doi.org/10.1186/S13059-021-02328-9

Garg S, Martin M, Marschall T (2016) Read-based phasing of related individuals. Bioinformatics 32:i234–i242. https://doi.org/10.1093/bioinformatics/btw276

Garin V, Wimmer V, Borchardt D, et al (2020) The influence of QTL allelic diversity on QTL detection in multi-parent populations: A simulation study in sugar beet. bioRxiv 2020.02.04.930677

Garin V, Wimmer V, Malosetti M (2015) mppR : An R Package for QTL Analysis in Multi-parent Populations using Linear Mixed Models

Garin V, Wimmer V, Mezmouk S, et al (2017) How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. Theoretical and Applied Genetics 130:1753–1764. https://doi.org/10.1007/s00122-017-2923-3

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. 1–9. https://doi.org/arXiv:1207.3907 [q-bio.GN]

Gayon J (2016) From Mendel to epigenetics: History of genetics. C R Biol 339:225–230. https://doi.org/10.1016/J.CRVI.2016.05.009

Geibel J, Reimer C, Weigend S, et al (2021) How array design creates SNP ascertainment bias. PLoS One 16:e0245178. https://doi.org/10.1371/JOURNAL.PONE.0245178

Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. Theoretical and Applied Genetics 46:319–330. https://doi.org/10.1007/BF00281673

Gerard D, Ferrão LFV, Garcia AAF, Stephens M (2018) Genotyping Polyploids from Messy Sequencing Data. Genetics 210:789–807. https://doi.org/10.1534/genetics.118.301468

Gilchrist E, Haughn G (2010) Reverse genetics techniques: engineering loss and gain of gene function in plants. Brief Funct Genomics 9:103–110. https://doi.org/10.1093/BFGP/ELP059

Giraud H, Lehermeier C, Bauer E, et al (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. Genetics 198:1717–1734. https://doi.org/10.1534/genetics.114.169367

Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: What Are They and How Do We Infer Them? Trends Plant Sci 21:609–621. https://doi.org/10.1016/j.tplants.2016.02.005

Griffing B (1956) Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems. Aust J Biol Sci 9:463. https://doi.org/10.1071/BI9560463

Guidarelli M, Baraldi E (2015) Transient transformation meets gene function discovery: The strawberry fruit case. Front Plant Sci 6:444. https://doi.org/10.3389/FPLS.2015.00444/BIBTEX

Guk JY, Jang MJ, Choi JW, et al (2022) De novo phasing resolves haplotype sequences in complex plant genomes. Plant Biotechnol J 20:1031–1041. https://doi.org/10.1111/PBI.13815

Hackett CA, Bradshaw JE, Bryan GJ (2014) QTL mapping in autotetraploids using SNP dosage information. Theoretical and Applied Genetics 127:1885–1904. https://doi.org/10.1007/s00122-014-2347-2

Hackett CA, Bradshaw JE, McNicol JW (2001) Interval mapping of quantitative trait loci in autotetraploid species. Genetics 159:1819–1832

Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity (Edinb) 90:33–38. https://doi.org/10.1038/sj.hdy.6800173

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity (Edinb) 69:315–24

Hall RD, D'Auria JC, Silva Ferreira AC, et al (2022) High-throughput plant phenotyping: a role for metabolomics? Trends Plant Sci 27:549–563. https://doi.org/10.1016/J.TPLANTS.2022.02.001

Halperin E*, *Stephan DA (2009) SNP imputation in association studies. Nature Biotechnology 2009 27:4 27:349–351. https://doi.org/10.1038/nbt0409-349

***Han S*, *Utz HF*, *Liu W*,** et al (2016) Choice of models for QTL mapping with multiple families and design of the training set for prediction of Fusarium resistance traits in maize. Theoretical and Applied Genetics 129:431–444. https://doi.org/10.1007/s00122-015-2637-3

Hancock JF*, *Luby JJ (1993) Genetic Resources at Our Doorstep: The Wild Strawberries. Bioscience 43:141–147. https://doi.org/10.2307/1312017

***Hardigan MA*, *Feldmann MJ*, *Lorant A*,** et al (2020) Genome Synteny Has Been Conserved Among the Octoploid Progenitors of Cultivated Strawberry Over Millions of Years of Evolution. Front Plant Sci 10:1789. https://doi.org/10.3389/fpls.2019.01789

***Hardigan MA*, *Feldmann MJ*, *Pincot DDA*,** et al (2021a) Blueprint for Phasing and Assembling the Genomes of Heterozygous Polyploids: Application to the Octoploid Genome of Strawberry. bioRxiv 2021.11.03.467115

***Hardigan MA*, *Lorant A*, *Pincot DDA*,** et al (2021b) Unraveling the Complex Hybrid Ancestry and Domestication History of Cultivated Strawberry. Mol Biol Evol 53:1689–1699. https://doi.org/10.1093/molbev/msab024

***Harrow J*, *Drysdale R*, *Smith A*,** et al (2021) ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics 37:2506–2511. https://doi.org/10.1093/BIOINFORMATICS/BTAB481

He D*, *Saha S*, *Finkers R*, *Parida L (2018) Efficient algorithms for polyploid haplotype phasing. BMC Genomics 19:110. https://doi.org/10.1186/s12864-018-4464-9

Heather JM*, *Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107:1–8. https://doi.org/10.1016/J.YGENO.2015.11.003

Hemmerlin A (2013) Post-translational events and modifications regulating plant enzymes involved in isoprenoid precursor biosynthesis. Plant Science 203–204:41–54. https://doi.org/10.1016/J.PLANTSCI.2012.12.008

Hendriks MMWB, Eeuwijk FA van, Jellema RH, et al (2011) Data-processing strategies for metabolomics studies. TrAC Trends in Analytical Chemistry 30:1685–1698. https://doi.org/10.1016/J.TRAC.2011.04.019

Henry LK, Gutensohn M, Thomas ST, et al (2015) Orthologs of the archaeal isopentenyl phosphate kinase regulate terpenoid production in plants. Proc Natl Acad Sci U S A 112:10050–10055. https://doi.org/10.1073/PNAS.1504798112/-/DCSUPPLEMENTAL

Henry LK, Thomas ST, Widhalm JR, et al (2018) Contribution of isopentenyl phosphate to plant terpenoid metabolism. Nature Plants 2018 4:9 4:721–729. https://doi.org/10.1038/s41477-018-0220-z

Hershey AD, Chase M (1952) INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. J Gen Physiol 36:39. https://doi.org/10.1085/JGP.36.1.39

Hickey JM, Crossa J, Babu R, de los Campos G (2012) Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci 52:654–663. https://doi.org/10.2135/CROPSCI2011.07.0358

Hokanson SC (2001) SNiPs, Chips, BACs, and YACs: Are Small Fruits Part of the Party Mix? HortScience 36:859–871. https://doi.org/10.21273/HORTSCI.36.5.859

Hu Q, Liu Y, Liao X, et al (2021a) A high-density genetic map construction and sex-related loci identification in Chinese Giant salamander. BMC Genomics 22:230. https://doi.org/10.1186/s12864-021-07550-0

Hu T, Chitnis N, Monos D, Dinh A (2021b) Next-generation sequencing technologies: An overview. Hum Immunol 82:801–811. https://doi.org/10.1016/J.HUMIMM.2021.02.012

Huang BE, Verbyla KL, Verbyla AP, et al (2015) MAGIC populations in crops: current status and future prospects. Theoretical and Applied Genetics 128:999–1017

Huang M, Liu X, Zhou Y, et al (2019) BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 8:. https://doi.org/10.1093/GIGASCIENCE/GIY154

Huang X, Han B (2014) Natural Variations and Genome-Wide Association Studies in Crop Plants. https://doi.org/101146/annurev-arplant-050213-035715 65:531–551. https://doi.org/10.1146/ANNUREV-ARPLANT-050213-035715

Hurgobin B, Edwards D (2017) SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? Biology 2017, Vol 6, Page 21 6:21. https://doi.org/10.3390/BIOLOGY6010021

Jacquin L, Elsen JM, Gilbert H (2014) Using haplotypes for the prediction of allelic identity to fine-map QTL: Characterization and properties. Genetics Selection Evolution 46:. https://doi.org/10.1186/1297-9686-46-45

Jannink JL, Jansen R (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 157:445–454

Jansen RC, Jannink J-L, Beavis WD (2003) Mapping Quantitative Trait Loci in Plant Breeding Populations: Use of Parental Haplotype Sharing. Crop Sci 43:829. https://doi.org/10.2135/cropsci2003.0829

Jetti RR, Yang E, Kurnianta A, et al (2007) Quantification of Selected Aroma-Active Compounds in Strawberries by Headspace Solid-Phase Microextraction Gas Chromatography and Correlation with Sensory Descriptive Analysis. J Food Sci 72:S487–S496. https://doi.org/10.1111/J.1750-3841.2007.00445.X

Jones DM, Vandepoele K (2020) Identification and evolution of gene regulatory networks: insights from comparative studies in plants. Curr Opin Plant Biol 54:42–48. https://doi.org/10.1016/J.PBI.2019.12.008

Kang HM, Sul JH, Service SK, et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354. https://doi.org/10.1038/ng.548

Karbstein K, Tomasello S, Hodač L, et al (2022) Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large Ranunculus auricomus species complex. New Phytologist 235:2081–2098. https://doi.org/10.1111/NPH.18284

Kaur S, Francki MG, Forster JW (2012) Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. Plant Biotechnol J 10:125–138. https://doi.org/10.1111/J.1467-7652.2011.00644.X

Kempthorne O (1957) An introduction to genetic statistics. New York: John Wiley & Sons, Inc.; London : Chapman & Hall Ltd.

Klein RJ, Zeiss C, Chew EY, et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science (1979) 308:385–389. https://doi.org/10.1126/SCIENCE.1109557/SUPPL_FILE/KLEIN_SOM.PDF

Koltsov N (1927) Physical-chemical fundamentals of morphology. Progress in Experimental Biology 3–31

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29. https://doi.org/10.1186/1746-4811-9-29

Kumar P, Choudhary M, Jat BS, et al (2021) Skim sequencing: an advanced NGS technology for crop improvement. J Genet 100:1–10. https://doi.org/10.1007/S12041-021-01285-3/TABLES/3

Labadie M, Vallin G, Petit A, et al (2020) Metabolite Quantitative Trait Loci for Flavonoids Provide New Insights into the Genetic Architecture of Strawberry (Fragaria × ananassa) Fruit Quality. J Agric Food Chem 68:6927–6939. https://doi.org/10.1021/ACS.JAFC.0C01855/ASSET/IMAGES/LARGE/JF0C01855_0003.JPEG

Lachance J, Tishkoff SA (2013) SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays 35:780–786. https://doi.org/10.1002/BIES.201300014

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199. https://doi.org/10.1093/genetics/121.1.185

Lee HE, Manivannan A, Lee SY, et al (2021) Chromosome Level Assembly of Homozygous Inbred Line 'Wongyo 3115' Facilitates the Construction of a High-Density Linkage Map and Identification of QTLs Associated With Fruit Firmness in Octoploid Strawberry (Fragaria × ananassa). Front Plant Sci 12:1337. https://doi.org/10.3389/FPLS.2021.696229/BIBTEX

Leggett RM, MacLean D (2014) Reference-free SNP detection: Dealing with the data deluge. BMC Genomics 15:1–7. https://doi.org/10.1186/1471-2164-15-S4-S10/COMMENTS

Leonardou VK, Doudoumis E, Tsormpatsidis E, et al (2021) Quality traits, volatile organic compounds, and expression of key flavor genes in strawberry genotypes over harvest period. Int J Mol Sci 22:. https://doi.org/10.3390/IJMS222413499/S1

Leroux D, Rahmani A, Jasson S, et al (2014) Clusthaplo: A plug-in for MCQTL to enhance QTL detection using ancestral alleles in multi-cross design. Theoretical and Applied Genetics 127:921–933. https://doi.org/10.1007/s00122-014-2267-1

Li BJ, Grierson D, Shi Y, Chen KS (2022a) Roles of abscisic acid in regulating ripening and quality of strawberry, a model non-climacteric fruit. Hortic Res 9:. https://doi.org/10.1093/HR/UHAC089

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987. https://doi.org/10.1093/BIOINFORMATICS/BTR509

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Li W, Boer MP, van Rossum BJ, et al (2022b) statgenMPP: an R package implementing an IBD-based mixed model approach for QTL mapping in a wide range of multi-parent populations. Bioinformatics 38:5134–5136. https://doi.org/10.1093/BIOINFORMATICS/BTAC662

Li W, Boer MP, Zheng C, et al (2021) An IBD-based mixed model approach for QTL mapping in multiparental populations. Theoretical and Applied Genetics 134:3643–3660. https://doi.org/10.1007/S00122-021-03919-7/FIGURES/4

Li Y, Pearl SA, Jackson SA (2015) Gene Networks in Plant Biology: Approaches in Reconstruction and Analysis. Trends Plant Sci 20:664–675. https://doi.org/10.1016/J.TPLANTS.2015.06.013

Li Y, Willer CJ, Ding J, et al (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–34. https://doi.org/10.1002/gepi.20533

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics 14:604–610. https://doi.org/10.1016/S0888-7543(05)80158-2

Liston A, Cronn R, Ashman TL (2014) Fragaria: A genus with deep historical roots and ripe for evolutionary and ecological insights. Am J Bot 101:1686–1699. https://doi.org/10.3732/AJB.1400140

Liston A, Wei N, Tennessen JA, et al (2020) Revisiting the origin of octoploid strawberry. Nat Genet 52:2–4. https://doi.org/10.1038/s41588-019-0543-3

Liu D, Ma C, Hong W, et al (2014) Construction and analysis of high-density linkage map using high-throughput sequencing data. PLoS One 9:. https://doi.org/10.1371/journal.pone.0098855

Liu W, Reif JC, Ranc N, et al (2012) Comparison of biometrical approaches for QTL detection in multiple segregating families. Theoretical and Applied Genetics 125:987–998. https://doi.org/10.1007/s00122-012-1889-4

Liu X, Huang M, Fan B, et al (2016) Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLoS Genet 12:e1005767. https://doi.org/10.1371/JOURNAL.PGEN.1005767

Lommen A (2009) Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem. https://doi.org/10.1021/ac900036d

Luo M, Zhou X, Sun H, et al (2021) Insights into profiling of volatile ester and LOX-pathway related gene families accompanying post-harvest ripening of 'Nanguo' pears. Food Chem 335:127665. https://doi.org/10.1016/J.FOODCHEM.2020.127665

Luo ZW, Maliepaard CA, Leach L, et al (2005) Constructing Genetic Linkage Maps Under a Tetrasomic Model. Genetics 172:2635–2645. https://doi.org/10.1534/genetics.105.052449

Majidian S, Kahaei MH, de Ridder D (2020) Hap10: Reconstructing accurate and long polyploid haplotypes using linked reads. BMC Bioinformatics 21:1–18. https://doi.org/10.1186/S12859-020-03584-5/FIGURES/6

Malmberg MM, Barbulescu DM, Drayton MC, et al (2018) Evaluation and recommendations for routine genotyping using skim whole genome re-sequencing in canola. Front Plant Sci 871:1809. https://doi.org/10.3389/FPLS.2018.01809/BIBTEX

Malzahn A, Lowder L, Qi Y (2017) Plant genome editing with TALEN and CRISPR. Cell & Bioscience 2017 7:1 7:1–18. https://doi.org/10.1186/S13578-017-0148-4

Mangandi J, Verma S, Osorio L, et al (2017) Pedigree-Based Analysis in a Multiparental Population of Octoploid Strawberry Reveals QTL Alleles Conferring Resistance to Phytophthora cactorum. G3: Genes, Genomes, Genetics 7:1707–1719. https://doi.org/10.1534/G3.117.042119

Mao J, Wang Y, Wang B, et al (2023) High-quality haplotype-resolved genome assembly of cultivated octoploid strawberry. Hortic Res 10:. https://doi.org/10.1093/HR/UHAD002

Marschall T, Marz M, Abeel T, et al (2018) Computational pan-genomics: Status, promises and challenges. Brief Bioinform 19:118–135. https://doi.org/10.1093/bib/bbw089

Marta AE, Camadro EL, Díaz-Ricci JC, Castagnaro AP (2004) Breeding barriers between the cultivated strawberry, Fragaria x ananassa, and related wild germplasm. Euphytica 136:139–150. https://doi.org/10.1023/B:EUPH.0000030665.95757.76/METRICS

Mascher M, Stein N (2014) Genetic anchoring of whole-genome shotgun assemblies. Front Genet 5:. https://doi.org/10.3389/fgene.2014.00208

Maurer A, Draba V, Jiang Y, et al (2015) Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC Genomics 16:290. https://doi.org/10.1186/s12864-015-1459-7

McAtee P, Karim S, Schaffer R, David K (2013) A dynamic interplay between phytohormones is required for fruit development, maturation, and ripening. Front Plant Sci 4:79. https://doi.org/10.3389/FPLS.2013.00079/BIBTEX

McMullen MD, Kresovich S, Villeda HS, et al (2009) Genetic Properties of the Maize Nested Association Mapping Population. Science (1979) 325:737–740. https://doi.org/10.1126/science.1174320

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/11290733

Michael TP, VanBuren R (2020) Building near-complete plant genomes. Curr Opin Plant Biol 54:26–33. https://doi.org/10.1016/J.PBI.2019.12.009

Mitteroecker P, Cheverud JM, Pavlicev M (2016) Multivariate analysis of genotype–phenotype association. Genetics 202:1345–1363. https://doi.org/10.1534/GENETICS.115.181339/-/DC1

Mollinari M, Garcia AAF (2019) Linkage Analysis and Haplotype Phasing in Experimental Autopolyploid Populations with High Ploidy Level Using Hidden Markov Models. G3 Genes|Genomes|Genetics 9:3297–3314. https://doi.org/10.1534/G3.119.400378

Mollinari M, Olukolu BA, Da Pereira GS, et al (2020) Unraveling the Hexaploid Sweetpotato Inheritance Using Ultra-Dense Multilocus Mapping. G3 Genes|Genomes|Genetics 10:281–292. https://doi.org/10.1534/G3.119.400620

Money D, Gardner K, Migicovsky Z, et al (2015) LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3 (Bethesda) 5:2383–90. https://doi.org/10.1534/g3.115.021667

Montesinos López OA, Montesinos López A, Crossa J (2022) Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer International Publishing, Cham

Motazedi E, de Ridder D, Finkers R, et al (2018) TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. Bioinformatics 34:3864–3872. https://doi.org/10.1093/bioinformatics/bty442

Motazedi E, Finkers R, Maliepaard C, de Ridder D (2017) Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. Brief Bioinform 19:bbw126. https://doi.org/10.1093/bib/bbw126

Motazedi E, Maliepaard C, Finkers R, et al (2019) Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads. Front Genet 10:335. https://doi.org/10.3389/FGENE.2019.00335/BIBTEX

NCBI (2022) GenBank and WGS Statistics. In: National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/genbank/statistics/. Accessed 17 Jan 2022

Neigenfind J, Kersten B, Basekow R, et al (2008) Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. BMC Genomics 9:356. https://doi.org/10.1186/1471-2164-9-356

Nettleton D, Doerge RW (2000) Accounting for Variability in the Use of Permutation Testing to Detect Quantitative Trait Loci. Biometrics 56:52–58. https://doi.org/10.1111/j.0006-341X.2000.00052.x

NIST Mass Spectrometry Data Center. https://chemdata.nist.gov/. Accessed 9 Jan 2023

Oh Y, Barbey CR, Chandra S, et al (2021) Genomic Characterization of the Fruity Aroma Gene, FaFAD1, Reveals a Gene Dosage Effect on γ-Decalactone Production in Strawberry (Fragaria × ananassa). Front Plant Sci 12:. https://doi.org/10.3389/FPLS.2021.639345/FULL

Ozaki K, Ohnishi Y, Iida A, et al (2002) Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nature Genetics 2002 32:4 32:650–654. https://doi.org/10.1038/ng1047

Patterson M, Marschall T, Pisanti N, et al (2015) WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. Journal of Computational Biology 22:498–509. https://doi.org/10.1089/cmb.2014.0157

Pavan E, Ye Y, Eyres GT, et al (2021) Relationships among Consumer Liking, Lipid and Volatile Compounds from New Zealand Commercial Lamb Loins. Foods 2021, Vol 10, Page 1143 10:1143. https://doi.org/10.3390/FOODS10051143

Peeters CFW, Übelhör C, Mes SW, et al (2019) Stable prediction with radiomics data. https://doi.org/10.48550/arXiv.1903.11696

Pieruschka R, Schurr U (2019) Plant Phenotyping: Past, Present, and Future. Plant Phenomics 2019:. https://doi.org/10.34133/2019/7507131

Piet Q, Droc G, Marande W, et al (2022) A chromosome-level, haplotype-phased Vanilla planifolia genome highlights the challenge of partial endoreplication for accurate whole-genome assembly. Plant Commun 3:100330. https://doi.org/10.1016/J.XPLC.2022.100330

Pook T, Schlather M, De Los Campos G, et al (2019) HaploBlocker: Creation of subgroup specific haplotype blocks and libraries. https://doi.org/10.1534/genetics.119.302283

Preedy KF, Hackett CA (2016) A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. Theor Appl Genet 129:2117–2132. https://doi.org/10.1007/s00122-016-2761-8

R Core Team (2016) R: A Language and Environment for Statistical Computing

Raguso RA (2016) More lessons from linalool: insights gained from a ubiquitous floral volatile. Curr Opin Plant Biol 32:31–36. https://doi.org/10.1016/J.PBI.2016.05.007

Ramsey J, Schemske DW (2002) Neopolyploidy in Flowering Plants. Annu Rev Ecol Syst 33:589–639. https://doi.org/10.1146/annurev.ecolsys.33.010802.150437

Rastas P (2017) Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics 33:3726–3732. https://doi.org/10.1093/BIOINFORMATICS/BTX494

Rastas P, Calboli FCF, Guo B, et al (2016) Construction of ultradense linkage maps with Lep-MAP2: Stickleback F2 recombinant crosses as an example. Genome Biol Evol 8:78–93. https://doi.org/10.1093/gbe/evv250

Raymond O, Gouzy J, Just J, et al (2018) The Rosa genome provides new insights into the domestication of modern roses. Nat Genet 50:772–777. https://doi.org/10.1038/s41588-018-0110-3

Reyment RA, Jöreskog KG (1993) Applied Factor Analysis in the Natural Sciences. Cambridge University Press

Rey-Serra P, Mnejja M, Monfort A (2022) Inheritance of esters and other volatile compounds responsible for the fruity aroma in strawberry. Front Plant Sci 0:2959. https://doi.org/10.3389/FPLS.2022.959155

Rosyara UR, De Jong WS, Douches DS, Endelman JB (2016) Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. Plant Genome 9:. https://doi.org/10.3835/plantgenome2015.08.0073

Rousseau-Gueutin M, Gaston A, Aïnouche A, et al (2009) Tracking the evolutionary history of polyploidy in Fragaria L. (strawberry): New insights from phylogenetic analyses of low-copy nuclear genes. Mol Phylogenet Evol 51:515–530. https://doi.org/10.1016/j.ympev.2008.12.024

Saada OA, Friedrich A, Schacherer J (2022) Towards accurate, contiguous and complete alignment-based polyploid phasing algorithms. Genomics 114:110369. https://doi.org/10.1016/J.YGENO.2022.110369

Saccenti E, Hoefsloot HCJ, Smilde AK, et al (2014) Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics 10:361–374. https://doi.org/10.1007/S11306-013-0598-6/FIGURES/6

Sallam AH, Conley E, Prakapenka D, et al (2020) Improving Prediction Accuracy Using Multi-allelic Haplotype Prediction and Training Population Optimization in Wheat. G3 Genes|Genomes|Genetics 10:2265–2273. https://doi.org/10.1534/G3.120.401165

Sargent DJ, Fernandéz-Fernandéz F, Ruiz-Roja JJ, et al (2009) A genetic linkage map of the cultivated strawberry (Fragaria × ananassa) and its comparison to the diploid Fragaria reference map. Molecular Breeding 24:293–303. https://doi.org/10.1007/s11032-009-9292-9

Sargent DJ, Yang Y, Šurbanovski N, et al (2015) HaploSNP affinities and linkage map positions illuminate subgenome composition in the octoploid, cultivated strawberry (Fragaria×ananassa). Plant Science 242:140–150. https://doi.org/10.1016/j.plantsci.2015.07.004

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–44. https://doi.org/10.1086/502802

Schiestl FP, Johnson SD (2013) Pollinator-mediated evolution of floral signals. Trends Ecol Evol 28:307–315. https://doi.org/10.1016/J.TREE.2013.01.019

Schmitz Carley CA, Coombs JJ, Douches DS, et al (2017) Automated tetraploid genotype calling by hierarchical clustering. Theoretical and Applied Genetics 130:717–726. https://doi.org/10.1007/s00122-016-2845-5

Schrinner SD, Mari RS, Ebler J, et al (2020) Haplotype threading: accurate polyploid phasing from long reads. Genome Biol 21:252. https://doi.org/10.1186/s13059-020-02158-1

Schrödinger E (1945) What is life? The physical aspect of the living cell. University Press, Cambridge

Schwab W, Davidovich-Rikanati R, Lewinsohn E (2008) Biosynthesis of plant-derived flavor compounds. The Plant Journal 54:712–732. https://doi.org/10.1111/J.1365-313X.2008.03446.X

Schwieterman ML, Colquhoun TA, Jaworski EA, et al (2014) Strawberry Flavor: Diverse Chemical Compositions, a Seasonal Influence, and Effects on Sensory Perception. PLoS One 9:e88446. https://doi.org/10.1371/journal.pone.0088446

Serang O, Mollinari M, Garcia AAF (2012) Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. PLoS One 7:e30906. https://doi.org/10.1371/journal.pone.0030906

Shulaev V, Sargent DJ, Crowhurst RN, et al (2010) The genome of woodland strawberry (Fragaria vesca). Nature Genetics 2010 43:2 43:109–116. https://doi.org/10.1038/ng.740

Sims D, Sudbery I, Ilott NE, et al (2014) Sequencing depth and coverage: Key considerations in genomic analyses. Nat Rev Genet 15:121–132. https://doi.org/10.1038/nrg3642

Smith HO (1979) Nucleotide sequence specificity of restriction endonucleases. Science (1979) 205:455–462. https://doi.org/10.1126/SCIENCE.377492/ASSET/BC1311B3-DD38-4192-8A55-56251FD34FF2/ASSETS/SCIENCE.377492.FP.PNG

Soltis DE, Buggs RJA, Barbazuk WB, et al (2012) The Early Stages of Polyploidy: Rapid and Repeated Evolution in Tragopogon. In: Polyploidy and Genome Evolution. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 271–292

Soltis DE, Visger CJ, Marchant DB, Soltis PS (2016) Polyploidy: Pitfalls and paths to a paradigm. Am J Bot 103:1146–1166. https://doi.org/10.3732/ajb.1500501

Soltis PS, Marchant DB, van de Peer Y, Soltis DE (2015) Polyploidy and genome evolution in plants. Curr Opin Genet Dev 35:119–125

Soltis PS, Soltis DE (eds) (2012) Polyploidy and Genome Evolution. Springer Berlin Heidelberg, Berlin, Heidelberg

Song J, Forney CF, Song J, Forney CF (2011) Flavour volatile production and regulation in fruit. https://doi.org/104141/CJPS07170 88:537–550. https://doi.org/10.4141/CJPS07170

Song Q, Yan L, Quigley C, et al (2017) Genetic Characterization of the Soybean Nested Association Mapping Population. Plant Genome 10:0. https://doi.org/10.3835/plantgenome2016.10.0109

Soyfer VN (2001) The consequences of political dictatorship for Russian science. Nature Reviews Genetics 2001 2:9 2:723–729. https://doi.org/10.1038/35088598

Spigler RB, Lewers KS, Johnson AL, Ashman TL (2010) Comparative Mapping Reveals Autosomal Origin of Sex Chromosome in Octoploid Fragaria virginiana. Journal of Heredity 101:S107–S117. https://doi.org/10.1093/JHERED/ESQ001

Sprague GF, Tatum LA (1942) General vs. Specific Combining Ability in Single Crosses of Corn1. Agron J 34:923. https://doi.org/10.2134/agronj1942.00021962003400100008x

Stebbins G (1940) The Significance of Polyploidy in Plant Evolution. Am Nat 74:54–66. https://doi.org/10.1086/280872

Stebbins G (1942) Polyploid Complexes in Relation to Ecology and the History of Floras. https://doi.org/101086/281012 76:36–45. https://doi.org/10.1086/281012

Stevens NM (1905) Studies in Spermatogenesis, Parts 1-2

Sturtevant A (1913) The linear arrangement of six sex-linked factors in Drosophila as shown by their mode of association. Z Indukt Abstamm Vererbungsl 10:293–294. https://doi.org/10.1007/bf01943452

Su SY, White J, Balding DJ, Coin LJM (2008) Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. BMC Bioinformatics 9:1–9. https://doi.org/10.1186/1471-2105-9-513/FIGURES/4

Sun H, Jiao WB, Krause K, et al (2022) Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nature Genetics 2022 54:3 54:342–348. https://doi.org/10.1038/s41588-022-01015-0

Tam V, Patel N, Turcotte M, et al (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20:467–484. https://doi.org/10.1038/s41576-019-0127-1

Tao Y, Zhao X, Mace E, et al (2019) Exploring and exploiting pan-genomics for crop improvement. Mol Plant. https://doi.org/10.1016/j.molp.2018.12.016

Tapia RR, Barbey CR, Lee S, et al (2021) Evolution of the MLO gene families in octoploid strawberry (Fragaria × ananassa) and progenitor diploid species identified potential genes for strawberry powdery mildew resistance. Hortic Res. https://doi.org/10.1038/s41438-021-00587-y

Tate JA, Soltis PS, Soltis DE (2005) Polyploidy in Plants. The Evolution of the Genome 371–426. https://doi.org/10.1016/B978-012301463-4/50009-7

Tennessen JA, Govindarajulu R, Ashman TL, Liston A (2014) Evolutionary origins and dynamics of octoploid strawberry subgenomes revealed by dense targeted capture linkage maps. Genome Biol Evol 6:3295–3313. https://doi.org/10.1093/gbe/evu261

Thérèse Navarro A, Bourke PM, Arens P, et al (2022) Smooth Descent: a Ploidy-Aware Algorithm To Improve Linkage Mapping In The Presence of Genotyping Errors. https://doi.org/10.21203/RS.3.RS-1165750/V1

Thérèse Navarro A, Tumino G, Visser RGF, et al (2020) Multiparental QTL analysis: can we do it in polyploids? In: Acta Horticulturae. International Society for Horticultural Science, pp 55–64

Tholl D (2015) Biosynthesis and Biological Functions of Terpenoids in Plants. 63–106. https://doi.org/10.1007/10_2014_295

Tikunov YM, Laptenok S, Hall RD, et al (2012) MSClust: A tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. Metabolomics. https://doi.org/10.1007/s11306-011-0368-2

Tong H, Nikoloski Z (2021) Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. J Plant Physiol 257:153354. https://doi.org/10.1016/J.JPLPH.2020.153354

Torkamaneh D, Boyle B, Belzile F (2018) Efficient genome-wide genotyping strategies and data integration in crop plants. Theoretical and Applied Genetics 131:499–511. https://doi.org/10.1007/S00122-018-3056-Z/FIGURES/5

Torkamaneh D, Laroche J, Belzile F (2016) Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. PLoS One 11:e0161333. https://doi.org/10.1371/journal.pone.0161333

Torri L, Aprea E, Piochi M, et al (2021) Relationship between sensory attributes, (Dis) liking and volatile organic composition of gorgonzola PDO cheese. Foods 10:2791. https://doi.org/10.3390/FOODS10112791/S1

Toubiana D, Fernie AR, Nikoloski Z, Fait A (2013) Network analysis: Tackling complex data to study plant metabolism. Trends Biotechnol 31:29–36. https://doi.org/10.1016/j.tibtech.2012.10.011

Tulipani S, Marzban G, Herndl A, et al (2011) Influence of environmental and genetic factors on health-related compounds in strawberry. Food Chem 124:906–913. https://doi.org/10.1016/J.FOODCHEM.2010.07.018

Ulrich D, Kecke S, Olbricht K (2018) What Do We Know about the Chemistry of Strawberry Aroma? J Agric Food Chem 66:3291–3301. https://doi.org/10.1021/acs.jafc.8b01115

Ulrich D, Olbricht K (2016) A search for the ideal flavor of strawberry - Comparison of consumer acceptance and metabolite patterns in Fragaria × ananassa Duch. Journal of Applied Botany and Food Quality 89:223–234. https://doi.org/10.5073/JABFQ.2016.089.029

van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nature Reviews Genetics 2017 18:7 18:411–424. https://doi.org/10.1038/nrg.2017.26

van Dijk T, Pagliarani G, Pikunova A, et al (2014) Genomic rearrangements and signatures of breeding in the allo-octoploid strawberry as revealed through an allele dose based SSR linkage map. BMC Plant Biol 14:1–16. https://doi.org/10.1186/1471-2229-14-55/FIGURES/6

van Huylenbroeck J (2018) Ornamental Crops. Springer International Publishing, Cham

van Os H, Stam P, Visser RGF, van Eck HJ (2005) SMOOTH: a statistical method for successful removal of genotyping errors from high-density genetic linkage data. Theoretical and Applied Genetics 112:187–194. https://doi.org/10.1007/s00122-005-0124-y

Varn DP, Crutchfield JP (2016) What did Erwin mean? The physics of information from the materials genomics of aperiodic crystals and water to molecular information catalysts and life. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374:. https://doi.org/10.1098/RSTA.2015.0067

Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. PLoS Biol 12:e1001883. https://doi.org/10.1371/JOURNAL.PBIO.1001883

Verhoeven KJF, Jannink J-L, McIntyre LM (2006) Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity (Edinb) 96:139–149. https://doi.org/10.1038/sj.hdy.6800763

Verma S, Zurn JD, Salinas N, et al (2017) Clarifying sub-genomic positions of QTLs for flowering habit and fruit quality in U.S. strawberry (Fragaria×ananassa) breeding populations using pedigree-based QTL analysis. Hortic Res 4:17062. https://doi.org/10.1038/hortres.2017.62

Vickerstaff RJ, Harrison RJ (2017) Crosslink: A fast, scriptable genetic mapper for outcrossing species. bioRxiv 135277. https://doi.org/10.1101/135277

Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics 12:172. https://doi.org/10.1186/1471-2105-12-172

Voorrips RE, Maliepaard CA (2012) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinformatics 13:248. https://doi.org/10.1186/1471-2105-13-248

Voorrips RE, Tumino G (2022) PolyHaplotyper: haplotyping in polyploids based on bi-allelic marker dosage data. BMC Bioinformatics 23:1–16. https://doi.org/10.1186/S12859-022-04989-0/TABLES/4

Wadl PA, Olukolu BA, Branham SE, et al (2018) Genetic diversity and population structure of the usda sweetpotato (ipomoea batatas) germplasm collections using gbspoly. Front Plant Sci 9:1166. https://doi.org/10.3389/FPLS.2018.01166/BIBTEX

Wang J, Zhang Z (2021) GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics Proteomics Bioinformatics 19:629–640. https://doi.org/10.1016/J.GPB.2021.08.005

Wang SB, Feng JY, Ren WL, et al (2016) Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 6:. https://doi.org/10.1038/SREP19444

Warner R, Wu B sen, MacPherson S, Lefsrud M (2021) A Review of Strawberry Photobiology and Fruit Flavonoids in Controlled Environments. Front Plant Sci 12:33. https://doi.org/10.3389/FPLS.2021.611893/BIBTEX

W***atson JD, Crick FHC*** (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 1953 171:4356 171:737–738. https://doi.org/10.1038/171737a0

Whitaker VM (2011) Applications of molecular markers in strawberry. J Berry Res 1:115–127. https://doi.org/10.3233/BR-2011-013

Whitaker VM, Knapp SJ, Hardigan MA, et al (2020) A roadmap for research in octoploid strawberry. Hortic Res 7:33

Whittaker IC, Thompson R, DENHAM MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252

Wilson G, Aruliah DA, Brown CT, et al (2014) Best Practices for Scientific Computing. PLoS Biol 12:. https://doi.org/10.1371/JOURNAL.PBIO.1001745

Worley B, Powers R (2013) Multivariate Analysis in Metabolomics. Curr Metabolomics 1:92. https://doi.org/10.2174/2213235X11301010092

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. PLoS Genet 4:e1000212. https://doi.org/10.1371/journal.pgen.1000212

Würschum T (2012) Mapping QTL for agronomic traits in breeding populations. Theoretical and Applied Genetics 125:201–210

Xie C, Xu S (2000) Mapping quantitative trait loci in tetraploid populations. Genet Res 76:105–15

Xu X, Bai G (2015) Whole-genome resequencing: changing the paradigms of SNP detection, molecular mapping and gene discovery. Molecular Breeding 35:1–11. https://doi.org/10.1007/S11032-015-0240-6/FIGURES/2

Yan J, Ban Z, Lu H, et al (2018) The aroma volatile repertoire in strawberry fruit: a review. J Sci Food Agric 98:4395–4402. https://doi.org/10.1002/jsfa.9039

Yang W, Feng H, Zhang X, et al (2020) Crop Phenomics and High-Throughput Phenotyping: Past Decades, Current Challenges, and Future Perspectives. Mol Plant 13:187–214. https://doi.org/10.1016/J.MOLP.2020.01.008

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160. https://doi.org/10.1016/j.copbio.2006.02.003

Yu J, Pressoir G, Briggs WH, et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. https://doi.org/10.1038/ng1702

Zabetakis I, Holden MA (1997) Strawberry Flavour: Analysis and Biosynthesis. J Sci Food Agric 74:421–434. https://doi.org/10.1002/(SICI)1097-0010(199708)74:4

***Zargar SM*, *Raatz B*, *Sonah H*,** et al (2016) Recent advances in molecular marker techniques: Insight into QTL mapping, GWAS and genomic selection in plants. Journal of Crop Science and Biotechnology 2015 18:5 18:293–308. https://doi.org/10.1007/S12892-015-0037-5

***Zhang X*, *Wu R*, *Wang Y*,** et al (2020) Unzipping haplotypes in diploid and polyploid genomes. Comput Struct Biotechnol J 18:66–72. https://doi.org/10.1016/J.CSBJ.2019.11.011

***Zhang Z*, *Ersoz E*, *Lai C-Q*,** et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360. https://doi.org/10.1038/ng.546

Zheng C*, *Amadeu RR*, *Munoz PR*, *Endelman JB (2021) Haplotype reconstruction in connected tetraploid F1 populations. Genetics 219:2020.12.18.423519. https://doi.org/10.1093/genetics/iyab106

Zheng C*, *Boer MP*, *van Eeuwijk FA (2018) Accurate genotype imputation in multiparental populations from low-coverage sequence. Genetics 210:71–82. https://doi.org/10.1534/genetics.118.300885

Zheng C*, *Boer MP*, *van Eeuwijk FA (2015) Reconstruction of genome ancestry blocks in multiparental populations. Genetics 200:1073–1087. https://doi.org/10.1534/GENETICS.115.177873/-/DC1

***Zheng C*, *Voorrips RE*, *Jansen J*,** et al (2016) Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. Genetics 203:119–131. https://doi.org/10.1534/genetics.115.185579

Zhou F*, *Pichersky E (2020) More is better: the diversity of terpene metabolism in plants. Curr Opin Plant Biol 55:1–10. https://doi.org/10.1016/J.PBI.2020.01.005

***Zorrilla-Fontanesi Y*, *Rambla J-L*, *Cabeza A*,** et al (2012) Genetic Analysis of Strawberry Fruit Aroma and Identification of O - Methyltransferase FaOMT as the Locus Controlling Natural Variation in Mesifurane Content. Plant Physiol 159:851–870. https://doi.org/10.1104/pp.111.188318

***Zych K*, *Gort G*, *Maliepaard CA*,** et al (2019) FitTetra 2.0 – improved genotype calling for tetraploids with multiple population and parental data support. BMC Bioinformatics 20:148. https://doi.org/10.1186/s12859-019-2703-y

Acknowledgements

Strangely enough, a doctoral thesis awards the title of doctor to a single person (hopefully me). This recognition comes after hard work and tedious effort is put into the research enclosed in a thesis, but such work would be of little use if not for the support offered by the many people surrounding the doctor-to-be. To all of you, who helped me in many ways, I would like to express my gratitude in helping my completing this thesis.

To my parents, **Rocío** and **Lluís**, who taught me the value of knowledge and a good education, who helped me follow my passion towards science and research and who supported me emotionally and economically throughout many years of education and exploration, thank you. If it was not for you, I would have never come to Wageningen and this thesis would not be in your hands. For these and many more reasons, this achievement is as much mine as it is yours. I will be forever grateful of all you did for me and for the person you have helped me become. Also to the rest of my family, **àvia Rosa, Rosamaría, Mariano,** I would like to give my thanks. When life was hard you welcomed me in your family and showed me what it meant to have love. Love for each other and love for the world, love for food and for books, love for discussions and for traditions, love for complaining and for making amends. This love I have not forgotten and has kept me alight in the moments of stress and hardship that come in every doctoral research. For the strength you have lent me, thank you.

To all my supervisors, thank you for your help and guidance in shaping my many chaotic ideas into sensible research. Without you this thesis would be a pool of incomplete and incoherent thoughts. **Chris**, you have given me several opportunities that have defined my career, supporting me along each of them, something I appreciate dearly and remember often. I have always felt welcome to come to you as I am, even when my emotions got the best of me. For all of this, my most heartfelt thanks. **Peter**, I have always admired your tremendous intelligence and wide knowledge and I greatly appreciate your many contributions to my thesis. You always found ways to improve my work and taught me to look deeper and more thoroughly to my own research. Thank you for all the effort you have put into this and for teaching me to do better. **Richard,** as the head of Plant Breeding I've always felt lucky when you found time to discuss my progress and personal interests. You've always listened with attention and have given insightful advice regarding my thesis and my personal development. For your help in this thesis and the opportunity you have given me in my future career, thank you. **Eric**, you are a wonderful person with an enormous heart and a clever eye for detail. I have greatly enjoyed working with you and learning from you and I hope you realise the positive impact you have had in me. You have never spared kind words towards me, an encouragement that I have often needed. My confidence in this research and myself is partly thanks to you, and for that I will be forever grateful. **Giorgio** and **Roeland,** thank you for your friendly disposition and insightful remarks that have given me the chance to learn and grow as a researcher working on such challenging topics. My technical skills have greatly improved thanks to your supervision and patience, for that I thank you. **Paul** and **René,** although not officially my supervisors you have also contributed to this thesis and my development. Be it through opportunities to contribute to projects, during personal and scientific discussions or while visiting collaborating companies, I have always felt welcomed when working with you and you are definitely part of the reason I feel so comfortable in our department. To you, and to all of those who make Plant Breeding a place where excellent science is carried out with positivity and camaraderie, my most sincere thank you.

I would also like to thank Fresh Forward and the people working in it for their contributions to this thesis. Although you are already acknowledged in many of my chapters, I would like to thank you personally. **Johan,** your kindness and helpfulness have been very appreciated. Although you are not my supervisor I think you've contributed much to this thesis and its underlying ideas, but most of all I would like to show my appreciation for your friendliness and openness. It has been great working with you and I hope we can continue to do so in the future. **Thijs,** I would also like to thank you for the work you have put into strawberries and for your openness when showing me your greenhouses and answering my questions. Your passion and intuition are admirable and a unique talent that will certainly propel strawberry breeding at Fresh Forward.

Also to my students I would like to give thanks, for showing me how to be a better teacher and contributing to my research with their own original ideas. **Jonathan,** through our collaboration we found an exciting and stimulating friendship which has brought me much intellectual growth (and interesting nights), for all this and for what is yet to come, thank you. **Olga Zafra,** we met well before we ever collaborated, and our friendship made our work together more fruitful and exciting. For your contributions to this thesis and my time in Wageningen, thank you. To **Mariaan Bas** and **Kevin Feng**, your will to learn was unbeatable and your personal development during our time together was inspiring. For the interesting discussions on the work of this thesis, thank you.

To my PhD colleagues and friends, thank you all for your support. **Miguel** and **Marcella**, you have been my closest friends, through ease and hardship, mania and depression, when the papers got rejected and the analysis failed, when the models worked and when we partied to forget the stress, you were always there to laugh, to joke, to complain, to listen… In short, thanks for being there and for your friendship. **Xulan**, one could not ask for a better office mate, whether to gossip or to remind me to keep working, to share stories or ask for advice, I could always count on you. Thank you for putting up with me during these years. **William**, we've known each other for many years now and although sometimes this closeness has brought friction, your sardonic humour and clever remarks have always brought a good laugh to me and all our friends. Sometimes, in the face of hardship, the best you can do is laugh, and for helping me realise that, thank you. **Yanlin,** thank you as well for being a friend and colleague, always with open arms and ears to discuss ideas and collaborate. You have been a great colleague to work with. To the rest of my

PhD colleagues at Plant Breeding, who I won't name individually for fear of forgetting anyone (but you all know who you are), thank you all for making our department fun and engaging, especially after more than two years of social pause due to the COVID pandemic. If I had not felt comfortable in our department this thesis would have not been completed, so for that I thank you. I hope those new PhDs that come after me can continue to enjoy such a nice atmosphere as the one you have created.

I would also like to thank the passionate and extremely helpful staff of the EPS PhD office. Especially **Juliane** and **Susan.** I know many PhDs must pass by your side every year and you lend support to all, but for me you have really been essential. Although not many appreciate it, you carry on your shoulders the weight of a whole community of PhD candidates. I doubt there are any other people that would do this work with the amount of passion and dedication that you put into it. For this, I thank you wholeheartedly.

Also the wonderful secretaries of our department deserve a wholehearted thank you. **Nicole** and **Daniëlle**, always ready to help me find my way through paperwork or academic protocols with a smile on your face. Thank you for helping me and all those who pass through the halls of Plant Breeding, without you this department (and the monthly coffees) would not be the same.

I would also like to thank my friends outside the PhD world. You have been my support during the worst times, my reminders that a world outside the university exists and you have helped me maintain my (sometimes scarce) sanity. To the wonderful family of de Wilde Wereld, in no particular order, **Loes, Judith, Daniël, Jostijn, Saskia, Margot, Brent, Rens, Max, Vinnie**, thank you for teaching so much about life in community and about getting my hands dirty when they need to be. You would be surprised of how similar life in an anarchist house and in a research department can be. Thank you all for putting up with my rants and for opening my heart to the Netherlands and its beautiful and colourful people. I would not feel at home in this country if it was not for your help and friendship. To my dear neighbours and friends, **Daria** and **Joris,** thank you for fun afternoons and great evenings, for cosy dinners and beers full of laughter and for taking care of Rigoberto when I

needed a long holiday away of my computer. To my friends in Spain too, for their long-standing friendship and emotional support, **Marta, Neus, Marc, Jaume,** thank you. With you I know that no matter how long I leave my homeland, I will always have a friendly face to come back to.

Last, but most definitely not least, to my boyfriend **Sergio**, thanks. Thank you for putting up with me, thank you for helping me deal with my emotions, thank you for building a dream life with me, thank you for cooking when my brain is so fried I cannot stand up from the sofa, thank you for forgiving my late hours and working weekends, thank you a thousand times for all you have been through during this time and for your love and support. As you approach your own thesis deadline, I hope I can repay all of your efforts to the best of my abilities.

And to you the reader, if you have gotten this far into the acknowledgements, thank you for reading them. Now would you like to read one of my chapters?

With my most heartfelt gratitude, see you anytime soon,

Alejandro

About the author

Alejandro Thérèse Navarro was born on the 29th of January of 1994 in Barcelona (Spain), although soon after his family moved to Tarragona and later, at the age of 6, to the village of Altafulla, the place he feels as his hometown. His interest in the natural world started at his hometown beach, where he became fascinated by the uniqueness of marine organisms. In high school he learned about genetics for the first time, which made him wonder about the mechanisms underlying the bewildering diversity of lifeforms found on Earth. As a curious tinkerer he started thinking about the creatures that could be created by modifying the genetics of organisms, a question that led him to want to study "genetic engineering" (with little understanding of what such a thing would entail).

With a strong desire to explore the world beyond his small village, he pursued his interests by enrolling in 2012 in a Bachelor in Genetics at the Autonomous University of Barcelona, coming back to his birthplace and moving away from his parents. There he explored the many topics covered by genetics: cancer, microbiology, population evolution, biomedicine, bioinformatics, applied statistics... Yet he still did not find his one true calling, until his first class on plant biotechnology. Plants combined the complexity of animals, with the feasibility of genetic engineering of microbes and with far fewer ethical dilemmas. With the hopes of pursuing his genetic engineering dream he moved in 2016 to Wageningen, where he enrolled into the MSc Plant Biotechnology, with a specialization on Plant Breeding.

Unexpectedly, the curricular freedom offered by Wageningen University would lead him astray from molecular genetics towards a more intriguing and exciting path: statistics and quantitative genetics. He realised the potential of this science to solve the many questions that still plagued him about genetic diversity, plant evolution and true genetic causes of many interesting traits. Slowly but surely, his curriculum moved from entirely biological to mostly statistical, finishing up with an MSc thesis and internship focused entirely on computational and analytical genetics. Shortly after the end of his internship, Dr. Ir. Chris Maliepaard offered him a PhD position to further his knowledge into this field, which he gladly accepted.

All this led to the current thesis, a detailed account of his attempts at exploring and pushing forward the mathematical lens by which all modern sciences must inevitably pass (to the dismay of many of his colleagues). In the future, he will continue to work on expanding tools and methodologies from the department of Plant Breeding, furthering his career into the academic world.

The research described in this thesis was financially supported by the Netherlands Enterprise Agency through multiple Top consortia for Knowledge and Innovation (TKI): “Novel genetic and genomic tools for polyploid crops” (BO-68-001-001, BO-50-002-022, BO-26.03-009-004), “Application of sequence-based multi-allelic markers in genetics and breeding of polyploids” (BO-68-001-042-WPR) and “Genetics assisted assembly of complex genomes” (BO-68-001-033-WPR). The contribution of the following companies to these projects and this thesis was greatly appreciated: Anthura B.V., Averis Seeds B.V., BASF Vegetable Seeds – Nunhems, Bejo Zaden B.V., Dekker Chrysanten B.V., Deliflor Chrysanten B.V., Dümmen Orange, Forage Genetics International L.L.C., Fresh Forward B.V., HZPC Holland B.V., Roath B.V. and Royal van Zanten.

The National Institute of Food and Agriculture (NIFA) from the United States Department of Agriculture (USDA) also contributed financially through the Specialty Crop Research Initiative (SCRI) project “Tools for genomics-assisted breeding in polyploids: Development of a community resource” (2020-51181-32156/SCRI).

Financial support from the Laboratory of Plant Breeding, Wageningen University & Research and the Experimental Plant Sciences graduate school for printing of this thesis is gratefully acknowledged.