UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Centre for Biological Sciences

Volume 1 of 1

Elucidating the genomics of nutritional and morphological traits in watercress (*Nasturtium officinale* R. Br.): The first genomic resources

by

Nikol Voutsina

Thesis for the degree of Doctor of Philosophy

November 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Centre for Biological Sciences

Thesis for the degree of Doctor of Philosophy

ELUCIDATING THE GENOMICS OF NUTRITIONAL AND MORPHOLOGICAL TRAITS IN WATERCRESS (*NASTURTIUM OFFICINALE* R. BR.): THE FIRST GENOMIC RESOURCES

Nikol Voutsina

Watercress (*Nasturtium officinale* R. Br.; Brassicaceae) has a long history of human use for medicine and consumption. In recent years, it has received a large deal of attention as one of the most nutrient-dense foods. Despite this, watercress remains largely underdeveloped with limited breeding resources through which to meet current and future intensifying market demands, such as for a more compact morphology, enhanced nutritional benefits and resource-use efficiency. The aim of this PhD has been to characterize the genetic structure of nutritional and morphological traits in watercress and develop molecular breeding tools that will inform and facilitate future work on this crop. To this end, Chapter 1 provides an overview of pre-existing knowledge on watercress and reviews the opportunities offered by Next Generation Sequencing tools for undeveloped crops. Chapter 2 describes the application of RNASeq towards *de novo* assembly and functional annotation the watercress transcriptome for the first time. Differential expression analysis resulted in a catalogue of significant genes for antioxidant capacity and glucosinolate content in watercress and identified orthologs to known phenylpropanoid and glucosinolate biosynthetic pathway genes. In Chapter 3, the first genetic linkage map and QTL analysis were completed for this crop, utilizing Genotyping-By-Sequencing for marker discovery. In a novel undertaking to identify QTL for chemopreventive qualities in a plant genome, the toxicity of watercress to human cancer cells was mapped successfully explaining 20 % of variation in this trait. As the development of new cultivars remains central to this work, Chapter 4 reports on the first commercial trials of the new ‘Boldrewood’ accession, aimed at informing its commercialization process. Excitingly, this study also highlighted previously unknown trends in phytonutrient character of the crops across a temporal gradient, which suggests the potential for increasing consumer health benefit by alternations to crop management practices. The sum of this work has resulted in significant advances in the understanding of watercress genetics and genomics and the production of valuable resources for its future preservation and advancement.

Table of Contents

List of tables

List of figures

List of accompanying materials

DECLARATION OF AUTHORSHIP

I, Nikol Voutsina, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Elucidating the genomics of nutritional and morphological traits in watercress (*Nasturtium officinale* R. Br.): The first genomic resources

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Chapter 2: **Voutsina N**, Payne AC, Hancock RD, Clarkson GJJ, Rothwell SD, Chapman MA & G Taylor (2016) Characterization of the watercress (*Nasturtium officinale* R. Br.; Brassicaceae) transcriptome using RNASeq and identification of candidate genes for important phytonutrient traits linked to human health. *BMC Genomics* 17:378, doi 10.1186/s12864-016-2704-4

Signed: ....

Date: ....

Acknowledgements

Life is full of unexpected twists and turns. I did not expect to find myself writing this text a few years ago. I am here doing so due to my supervisor Professor Gail Taylor, who from our first meeting has challenged me to aim higher than I would have for myself. I thank her for her hard work, encouragement, belief in me, and for all the opportunities she has put in front of me. From the lab bench to the farm, and from P.A.G. to the Chelsea Flower Show, it has been an unexpected exciting journey on which I have collected many precious lessons and memories. I also thank my second supervisor, Dr Mark Chapman, for his tireless explanations and sketches of seemingly impossible genomic scenarios, for getting me started in the lab and for always being available and supportive.

I would like to thank my funding body, Vitacress Salads Ltd, for the opportunity to carry out this research with a direct link and frequent contact with industry from the farm to management level. This PhD has been unique in that it has been framed through the prism of collaboration between academia and industry, which I feel is imperative to innovation and progress. Particular thanks to Dr Steve Rothwell for his feedback and support, and to Andy Elworthy and Mark Newton for welcoming us to the farms and taking such great care for my field trials.

I would like to extend my gratitude to Dr Rob Hancock at The James Hutton Institute, Dundee, U.K. for kindly hosting me in his lab on two separate occasions and patiently teaching me how to extract glucosinolates, run HPLC-MS and ascorbic acid assays. One of the highlights of my PhD was the two weeks I spent working on linkage mapping and QTL analysis at Wageningen University and Research. I thank Chaozhi Zheng for his time and efforts. I especially thank Joost Keurentjes and Erik Wijnker for their collaboration, excellent hosting and for teaching me so much. These trips were partially funded and possible due to a Training Grant for the Genetics Society.

Like a fish out of water, I had the rare opportunity to complete a substantial mammalian tissue culture project. Professor Graham Packham at the University of Southampton (UoS) was an important source of information, contacts and advice in designing the project. I thank Dr Manos Papadakis for introducing me to mammalian tissue culture. I am indebted to Dr Mark Coldwell and Dr Jo Cowan for adopting me, teaching me everything I needed to know and supporting me during this undertaking. They were an absolute joy to work with and an unexpected blessing in the face of a daunting task. I also thank Dr Coldwell for the kind contribution of the MCF-7 cells and his support during data analysis.

The excellent technical staff at UoS are instrumental to completing research at a high standard. I particularly thank Mike Cotton for his invaluable support while running long trials with large numbers of plants in the controlled environmental chambers or glasshouses.

My lab mates have made it possible to complete the ‘1,000+ x every step of every protocol’ challenge. For their help and support at various points, I thank: Dr Adrienne Payne, Dr Suzie Milner, Billy Valdes, Gwen Isaacs, Dr Zoe Harris, Dr Hazel Smith, Dr Rob Holland, Dr Mike Allwright, Dr Joe Jenkins, Dr Annabelle Damerum, Libby Rowland, Jasmine Saban, Libby Arnold, Ollie White, Anna Page and Rachael Graham. Special thanks to Dr Adrienne Payne for showing me the ropes and Dr Suzie Milner; with her knowledge, calm and strength everything is possible. Also, special thanks to my fantastic volunteer and friend, Felipe Becerra, who spent his days off in 2016 helping me process seemingly endless amounts of frozen or freeze-dried samples.

Thank you to my neighbours and friends for being amazing, providing me with lots of snacks, beds to sleep in, and for putting up with my humour and songs about liquid nitrogen or broken autoclaves. Annabelle, Zoe, Libby, Suzie, Ilektra, Tania, Felipe, Przemak, Ollie, Lauren and Beatrice, to name a few. Finally, I am grateful for my wonderful husband, Ben, who has supported me throughout – what shall we take on next? And to our wonderful families, who are never too far despite the distances.

My grandmother, Despina Voutsina (*yiayia*), passed away during this PhD. When I was growing up she always encouraged me to do my homework because she didn’t have the opportunity to go to school beyond the second grade. I dedicate this PhD thesis to her.

Definitions and Abbreviations

AFLP: Amplified Fragment Length Polymorphisms

ANOVA: Analysis of Variance

AO: Antioxidants

BSA: Bulk Segregant Analysis

BSRSeq: Bulk Segregant RNA Sequencing

CIM: Composite Interval Mapping

CYP: Cytochrome P450

CTAB: Cetyl Trimethylammonium Bromide

DAPI: 4’,6-Diamindino-2-Phenylindole nuclear stain

DE: Differential Expression/Differentially Expressed

DMEM: Dulbecco’s Modified Eagle’s Medium

eQTL: Expression Quantitative Trait Loci

ETI: Effector Triggered Immunity

FDR: False Discovery Rate

FPKM: Fragments Per Kilobase of exon per Million fragments

FRAP: Ferric Reducing Ability of Plasma

GBS: Genotyping-By-Sequencing

GLS: Glucosinolate

GM: Genetic Modification

GO: Gene Ontology

GSH: Glutathione

GST: Glutathione S-transferase

HIF: Hypoxia-Inducible Factor

HPLC-MS: High Performance Liquid Chromatography coupled with Mass Spectrophotometry

IC50: Half-Maximal Concentration of the Inhibitor

IPCC: Intergovernmental Panel on Climate Change

ITC: Isothiocyanate

JI2: John Innes Number 2 potting soil

LOD: Logarithm of Odds

MAS: Marker Assisted Selection

MIF: Migration Inhibition Factor

NILs: Near Isogenic Lines

NGS: Next Generation Sequencing

NHS: National Health Service

P: Phosphorus

PCA: Principle Component Analysis

PCR: Polymerase Chain Reaction

PEITC: Phenethyl Isothiocyanate

PTI: Pattern-triggered Immunity

PUE: Phosphate Use Efficiency

QC: Quality Control

QTL: Quantitative Trait Loci

RAD: Restriction-site Associated DNA

RE: Restriction Enzyme

RILs: Recombinant Inbred Lines

RNASeq: RNA Sequencing

RRL: Reduced Representation Library

SD: Standard Deviation

SE: Standard Error of the mean

SIM: Simple Interval Mapping

SLA: Specific Leaf Area

SNP: Single Nucleotide Polymorphism

SSR: Simple Sequence Repeat or microsatellite

UoS: University of Southampton

WX: Watercress

Chapter 1: The current state of knowledge regarding the watercress (*Nasturtium officinale* R. Br.; Brassicaceae) crop and prospects for breeding improved cultivars.

1.1 Topic Introduction

> 'Plant breeding must therefore focus on traits that improve nutritional quality, confer enhanced nutrients- and water-use efficiency (WUE), and those that enhance adaptation to abiotic and biotic stresses to increase yield'
> Dwivedi et al. 2013, Advances in Agronomy

Future food security is a critical issue in present-day research and policy. Global population is expected to reach 9.8 billion by 2050 (United Nations et al. 2017), increasing demands on food supplies. In addition, climate change is expected to challenge agricultural resources and practices with climate unpredictability, increased temperatures and frequency of water stress, salinization, pests and pathogens. The Intergovermental Panel on Climate Change (2014) anticipates that climate change is likely to impact food production, trade, prices and availability on a global scale. However, the impacts of climate change will not be evenly distributed amongst regions of the world, with poorer countries in a worse off position (IPCC 2014). These additive problems increase the difficulty of meeting international goals concerning food security, the elimination of hunger in poor countries and environmental conservation.

Food availability is not the only concerning aspect of food security. Malnutrition has severe social and economic consequences globally and is estimated to cost $3.5 trillion annually (Food and Agriculture Organization of the United Nations 2013). Even in nations where food is generally considered a broadly-accessible resource, chronic disease due to poor dietary habits is a substantial expenditure. In the U.K., the cost of poor nutrition exceeds £5 billion per annum and represents 21.6% of the National Health Service's (NHS) total expenditures, making it the most expensive cause of illness (Scarborough et al. 2011). A new report suggested a much higher impact by estimating the cost of malnutrition to the NHS in 2011 - 2012 to be £19.6 billion (Elia 2015). In the U.S.A., the financial cost of poor diet was estimated to be $70.9 billion annually (Frazão 1999). Confounded by climate change, poor health and death associated with malnutrition are likely to increase with reduced availability of nutritious food (Springmann et al. 2016) and potential loss of nutritional quality in certain food crops (Dwivedi et al. 2013).

Therefore, the enhancement of crop nutritional value is an important target that could lead to greater life quality and the liberation of resources that are currently tied into avoidable medical expenses.

Sustainable-intensification of crop production in the face of resource challenges caused by climate change, and the preservation or enhancement of the nutritional value of produced crops must become a priority defining future plant research. Innovative solutions are needed by farmers, breeders, and plant scientists to face the challenges of the future. Maximum yields and quality can be reached by optimisation of agricultural management and through the application of modern breeding techniques to deliver new crop varieties with desirable traits (Godfray et al. 2010). Thus, the ultimate goal of crop science becomes linked to ensuring the availability of sufficient and nutritious food for all individuals through sustainable means. These goals require food crop varieties that will produce larger and highly-nutritious yields with less resource inputs, while simultaneously tolerating abiotic and biotic stressors.

There are certain resources available to plant scientists and breeders that will contribute to meeting these global needs. They include agrobiodiversity, genetic modification (GM) technologies, and new technologies, such as Next Generation Sequencing (NGS) or genome editing, which facilitate molecular plant breeding, marker-assisted selection (MAS) and an array of new genomic approaches. Agrobiodiversity is a crucial source of genetic variation from which to draw for future breeding needs (Dwivedi et al. 2013). Domestication of crop species has naturally resulted in the loss of much genetic information through the crop selection process. The genetic information that is still present in the wild relatives of modern crop species is essential to breeding new varieties of cultivated crops, particularly in securing resilience to abiotic stressors and pathogens. MAS incorporates molecular techniques into traditional crop selection, thus increasing the speed and accuracy of the breeding process (Tester & Langridge 2010). New molecular technologies in MAS have the potential to greatly increase the success rates of crop improvement projects. GM takes crop development a step further enabling the incorporation of genetic material across species and is expected to contribute greatly to developing crops for the future (Tester & Langridge 2010). The cultivation of GM-produced maize, cotton, and soybean is widespread in America and Asia; however, GM crops are still considered controversial in parts of Europe (Baulcombe et al. 2014). The ground-breaking technologies known collectively as “genome editing” have as of yet not been classified as GM and could be a game-changer in crop breeding (Hartung & Schiemann 2014).

Within this global setting and with these values under consideration, this literature review is a compilation of the relevant information and tools required to advance the molecular knowledge and breeding of the highly-nutritious salad crop, watercress. The watercress market is an irreplaceable component of agriculture in southern England and its phytonutrient components are promising dietary combatants of numerous diseases that impact human health.

1.2 The Watercress Crop Today

1.2.1 An Introduction to Watercress:

Watercress, *Nasturtium officinale* R. Br., (previously known as, *Rorippa nasturtium-aquaticum* (L.) Hayek) is a perennial dicotyledonous herb. The plant's life cycle is depicted in Figure 1.1. It is coloured green/brown and is hairless; has pinnate leaves with small round leaflets; and produces small white flowers with four petals in the shape of a cross (Howard & Lyon 1952). It flowers in longer days, starting in May, and the flowers give rise to long siliques with two rows of seed (Howard & Lyon 1952; Palaniswamy & Mcavoy 2001). The seed is immediately viable but can be stored for 5 years (Howard & Lyon 1952). Watercress is adapted to an aquatic or semi-aquatic habitat with hollow 'floating' stems and extensive adventitious roots forming from stem nodes, which stabilise plants against the water's current (Howard & Lyon 1952; Cumbus et al. 1980).

Watercress belongs to the Brassicaceae family (or Crucifereae); an economically-valuable flowering plant family of the Capparales order, also known as the cabbage family or as cruciferous vegetables. It contains approximately 338 genera and 3700 species, including globally important food, oil, spice crops, as well as popular ornamental plants (for a review of Brassicaceae phylogeny, see Al-Shehbaz et al. 2006). For example, rapeseed (*Brassica napus* L.), is grown for animal feed, vegetable oil and biodiesel, and is the third largest vegetable oil crop globally (FAO 2011). The Brassicaceae family also includes major edible crops such as broccoli, cauliflower, cabbage, Brussels sprouts, mustard, radish, turnips, rocket and kale. Another particularly important member is thale cress, *Arabidopsis thaliana* (L.) Heynh, which has served as a model species for much contemporary plant research and was the first plant species to have its genome sequenced (The Arabidopsis Genome Initiative 2000). The Brassicaceae genus, *Brassica*, has also been studied extensively because of the wide-ranging list of common food-crop species it includes. However, not all food crop species of the Brassicaceae family have received sufficient attention. For watercress, a highly-nutritious food crop, there are limited resources informing industry and science about the genetic basis of important traits; a foundation through which to maintain and improve this salad crop in the future.

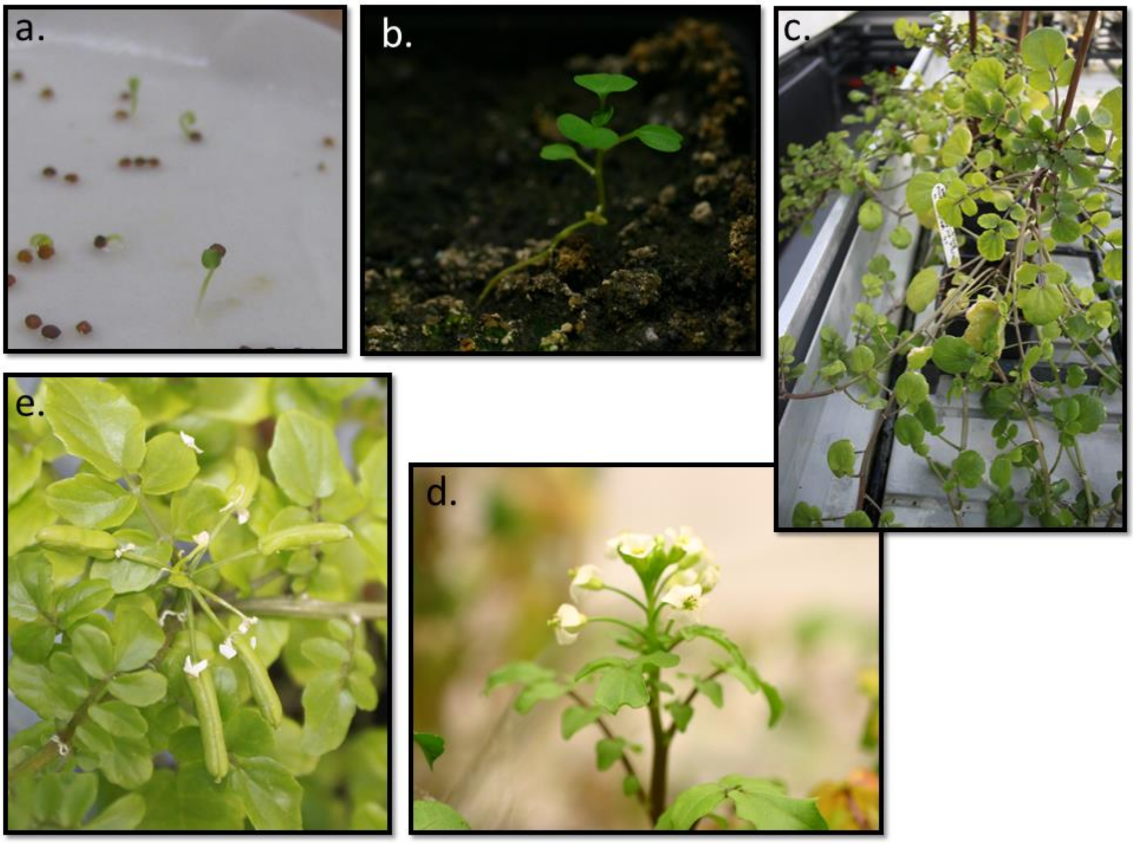


Figure 1.1 The life cycle of watercress is depicted clockwise from the top left corner: (a.) germinating seeds, (b.) a young seedling, (c.) an individual during the vegetative growth phase, (d.) a flower head during anthesis showing the characteristic white cross-shaped flower, and (e.) setting siliques.

The taxonomy and genetic origin of watercress is not clear. The genus of *Nasturtium* has often been merged to *Rorippa* or *Cardamine* - a historic cross from one of this group with an unknown ancestor described as a possible origin - but it is now maintained as a separate genus (Manton 1935; Al-Shehbaz & Price 1998). More recently, Sheridan et al. (2001) identified *Nasturtium* as more closely related to *Barbarea* than to either *Rorippa* or *Cardamine*. The *Nasturtium* genus consists of seven species: *N. officinale*, *Nasturtium africanum* Braun-Blanq., *Nasturtium floridanum* (Al-Shabaz & Rollins) Al-Shehbaz & R.A. Price, *Nasturtium gambelii* (S. Watson) O.E. Schulz, *Nasturtium groenlandicum* (Hornem.) Kuntze, *Nasturtium microphyllum* (Boenn. ex Rchb.) Rchb. which has also been called *Rorippa microphylla*, and *Nastutrium sordidum* (A. Gray) Kuntze. The approximate position of watercress within the Brassicaceae family is shown in Figure 1.2 (Bailey et al. 2006). A more recent study on *Arabidopsis* lineage suggests that its closest ancestor with watercress existed approximately 40 million years ago, making them close relatives within the broad Brassicaceae family (Beilstein et al. 2010).

N. officinale was previously considered a diploid (2n = 32) but is currently assumed to be a tetraploid species with unknown diploid ancestors (Manton 1935; Howard & Lyon 1952; Bleeker et al. 1999). However, occasional images of watercress nuclei with 16 chromosomes have been published suggesting that there could be varying ploidy in this species (Manton & Howard 1946; Jeelani et al. 2013). Brown watercress, *Nasturtium microphyllum,* is an octoploid with which *N. officinale* can cross, giving rise to the often sterile *N. x sterile* (Bleeker et al. 1999). Recently, flow cytometry has been applied to calculate the nuclear genome size of these species. *N. officinale* is reported to have a 2C = 0.76 pg (or approximately 743 Mbp) and *N. microphyllum* at 2C = 1.43 pg (Morozowska et al. 2010). Originally both types of watercress were grown commercially in England. *N. officinale* was grown in summer and *N. microphillum* was grown in winter; however only the first is now used due to its greater resistance to crook-root disease (Manton 1935; Palaniswamy & McAvoy 2001).

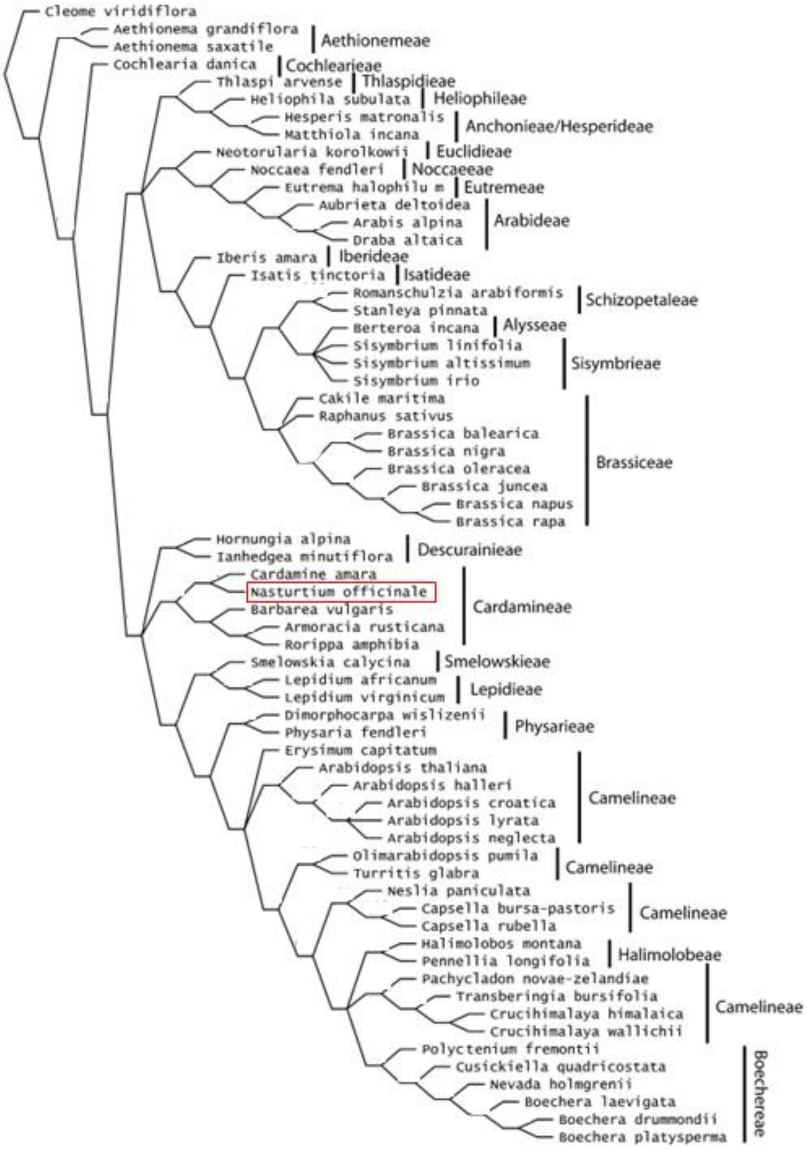


Figure 1.2 A basic Brassicaceae phylogeny with the estimated position of watercress, modified from Bailey et al. (2006)

Although the origin of watercress is likely to be southern Europe or the Middle East and it is firmly considered an introduced species to America, S. Africa, Australia and New Zealand (Howard & Lyon 1952), it has been difficult to pinpoint with certainty. This is because its artificial distribution due to historic human interference and use are impossible to distinguish from natural distribution (Manton 1935; Howard & Lyon 1952; Bleeker et al. 1999). Watercress has been collected or grown for medicinal and food purposes for over 2000 years. Its medicinal use is recorded from 77 A.D. to the 19th century, when it was administered as an antiscorbutic (Manton 1935). Historically collected from the wild and known as ‘cresson de fontaine’ in 14th century France, it was not commercially cultivated until later, with the first U.K. watercress farm opening in 1808 (Manton 1935). It is now grown at commercial scale in Europe and the U.S.A (primarily in Florida and Hawaii) with the presence of smaller growers hard to verify on a global scale. Commercial watercress production in the U.K. is currently focused in Hampshire, Dorset, and Wiltshire counties in the South of England. The U.K. watercress market, including commercially and organically-grown watercress, was worth approximately £25 million in 2013, contributing to the £536 million total leafy salads U.K. market (G Clarkson 2014, personal communication).

Commercial watercress practices have adapted from traditional watercress farming in the U.K., in order to accommodate the greater market demands, but remain much the same (Figure 1.3). Seeds are germinated in glasshouses or poly-tunnels and seedlings are later moved to shallow gravel purpose-built beds (Natural England 2009). The watercress beds are flood-irrigated with spring water, pumped from boreholes at a rate of 5,000 gallons per acre per hour, so that the crop is grown in a constant flowing water (Natural England 2009). Optimum flavour has been associated with plants having 12 to 15 internodes, and this usually occurs 6 to 7 weeks after germination (Palaniswamy & McAvoy 2001). Approximately 25-35 days after transplanting, the crop is harvested with special machinery, processed on site, and sold in bags as pre-washed watercress or in mixed salads (Natural England 2009; Palaniswamy & McAvoy 2001). Minimal changes have occurred over the years, namely the product transitioning from watercress bundles to salad bags; the mechanisation of harvesting; and the introduction of settling tanks to minimize the effect of the farms on freshwater wildlife (Dixon 2010).

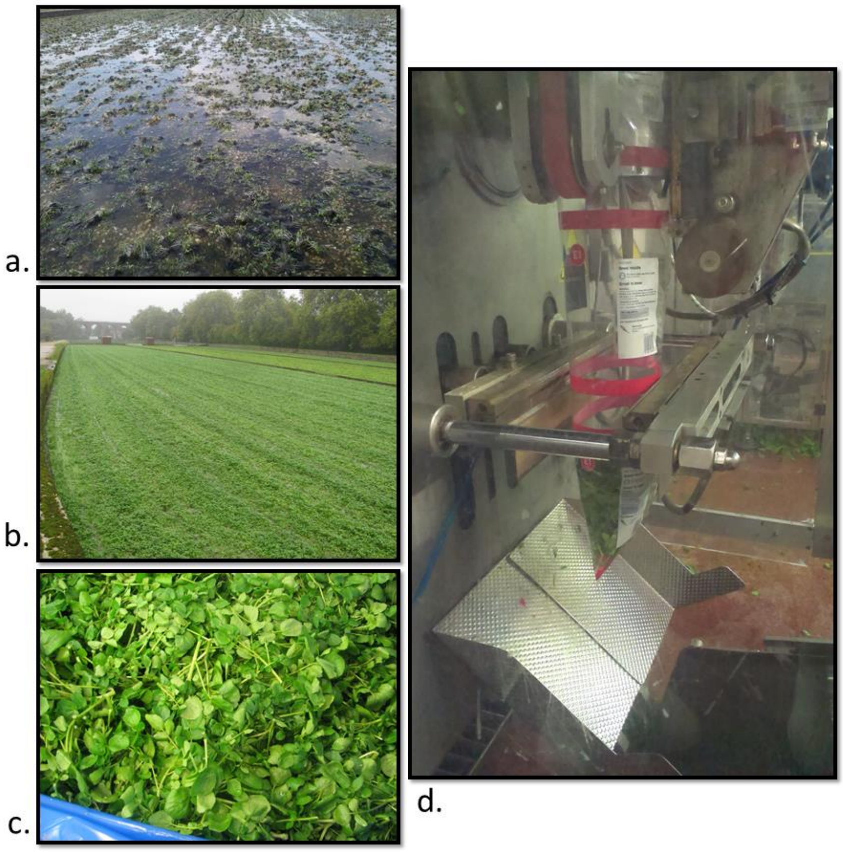


Figure 1.3 The commercial life cycle of watercress: (a.) plugs spread in the watercress beds in standing water and allowed to root, (b.) the growing crop in a typical watercress bed, (c.) leafy watercress harvested and transported to the factory, (d.) packaged in ready-to-eat bags before arriving at market.

1.3 Watercress and Human Health

1.3.1 The Dietary Contribution of Fruits and Vegetables:

It has become widely accepted that the consumption of fruits and vegetables is linked to better health (Block et al. 1992; Van Duyn & Pivonka 2000; Martin & Li 2017). Two recent studies, a large-scale epidemiology study and a widely-publicised meta-analysis, suggested that mortality and chronic disease decreases consistently with increasing intake of fresh fruit and vegetables (Oyebode et al. 2014; Aune et al. 2017). However, due to the complex nature of this hypothesis and the difficulties in conducting long-term studies, the benefits to be gained, portions required to gain these health benefits, and their specificity to categories of fruit and vegetables is not clear. Some studies did not identified a significant benefit of a vegetable-based diet for particular health conditions or have identified a threshold beyond which further vegetable consumption had no further impact (Hung et al. 2004; Wang et al. 2014).

A focus of these studies has often been specifically on the Brassicaceae family, which includes food crops of international popularity, such as broccoli, kale and cabbage. The consumption of Brassicaceae plants has been linked to a decreased occurrence of chronic diseases including cancer, cardiovascular disease, asthma, Alzheimer's disease, and diabetes (Verhoeven et al. 1996; Higdon et al. 2007; Manchali et al. 2012). Cruciferous vegetables rank high not only because they contain an array of essential macro and micronutrients, but because they are also rich in non-essential plant-derived compounds, often referred to as phytonutrients or phytochemicals (Jahangir et al. 2009; Björkman et al. 2011; Manchali et al. 2012; Avato & Argentieri 2015). These are typically actors in the plant's secondary metabolism and include glucosinolates, phenolics, phytoalexins, and plant pigments, such as carotenoids or anthocyanins (Jahangir et al. 2009; Björkman et al. 2011; Manchali et al. 2012; Avato & Argentieri 2015). Many appear to act as antioxidants, alongside certain vitamins, and reduce damage caused by oxidative stress in the cell, thus minimizing chronic damage (Kaur & Kapoor 2008; Martin et al. 2013). To be more specific, highly-reactive free radicals produced during oxygen-based metabolism, or by external exposure to radiation or toxins, can impact cells negatively by reacting with other cellular molecules by way of their unpaired electron and, thus causing undesirable alterations in molecules and harming DNA (Lobo et al. 2010). The main dietary antioxidants are ascorbic acid, tocopherols, carotenoids and phenolic compounds and their role in Brassicas is reviewed by Podsędek (2007).

Caution must be exercised, however, when making general assumptions about the bioavailability and nutritional value of plant-derived bioactive compounds for several reasons. Each step of the food production chain can affect the amount of phytonutrients that ultimately reach a consumer. The food production chain has the potential to introducing a cumulative effect of up to 100-fold change in phytonutrient concentrations from harvest to consumption (Dekker et al. 2000; K. Hennig et al. 2014). In addition, there are still gaps in our knowledge regarding the impact of plant-derived compounds on human health (Traka & Mithen 2011). Granado et al. (2006) found differences in the absorption of various broccoli-derived phytonutrients *in vitro* and *in vivo,* suggesting that study models may greatly impact the conclusions of research, and occasionally inconsistent results have been published. In part the lack of confidence in big health-related claims comes from current limitations in the quantification of the presence/absence and/or effect of phytochemicals on health. In the case of antioxidant capacity - or the ability to scavenge for reactive oxygen species as gained by the consumption of a dietary source of antioxidants - is measured vaguely through antioxidant assays, reviewed in Moon & Shibamoto (2009) and compared in Payne et al. (2013), which are performed on sap removed from the fresh plant material and do not have specificity to any particular class of antioxidants. In human-based studies, the amounts of particular antioxidants are quantified in the blood plasma (Gill et al. 2007) and taken as an indication of antioxidant capacity in the body. Positive contributions of such compounds to human health are not direct and, thus, their nutritional value cannot be easily shown or quantified beyond all doubt. Yet, as the broader epidemiological evidence suggests that a diet high in fruits and vegetables is associated with better health, the role of antioxidants in this observation is often assumed despite lack of direct evidence or possible synergistic or antagonistic interactions that may be overseen *in situ* (Patil et al. 2014).

Another consideration is the ability of Brassicaceaes to act as a vector for the consumption of toxic compounds under specific environmental conditions, such as heavy metals (Jahangir et al. 2009). This issue is linked to human activity and waste management, and is reviewed in Islam *et al.* (2007).

The overarching conclusion, however, is that the consumption of vegetables and Brassicaceae vegetables has a strong positive impact on human health.

1.3.2 Glucosinolates and Isothiocyanates:

Glucosinolates (GLS) in Brassicaceaes have been reviewed extensively and are of particular interest because of the health benefits attributed to their derivatives, the isothiocyanates (ITC) (Mithen et al. 2000; Cartea & Velasco 2008; Traka & Mithen 2008; Manchali et al. 2012; Avato & Argentieri 2015; Giacoppo et al. 2015). GLS are a secondary metabolite found in 16 plant families but with what appears to be a universal presence in the Brassicaceae family, where 96 species of glucosinolate have been identified (Fahey et al. 2001). Fahey *et al.* described the chemical structure of the stable and water-soluble glucosinolates as "… β-thioglucoside *N*-hydroxysulfates with a side chain (R) and a sulfur-linked β-D-glucopyranose moiety" and has been known since 1956, when it was correctly proposed by Ettlinger and Lindeen (Ettlinger & Lundeen 1956; Fahey et al. 2001).

There are several types of glucosinolates. The most well-known are aliphatic (carbons in straight or branched chains), aromatic (carbons forming aromatic ring), and indolic or heterocyclic (an aromatic compound structure with an additional ring including a nitrogen) glucosinolates (Fahey et al. 2001; Manchali et al. 2012; Ishida et al. 2014). They have a great diversity of side chains- over 120 have been identified- and more than one type of glucosinolate is usually found in a species of plant (Fahey et al. 2001). In plants, they are synthesized via the following main steps: first, amino acid side chain elongation; then the decarboxylation of the amino acid to form a glucone; followed by thiohydroxamic acid formation with the addition of a sulfur group; desulfoglucosinolate formation with the addition of UDP-glucose; and then a final sulfur is added to form the glucosinolate (Grubb & Abel 2006; Manchali et al. 2012). The glucosinolates found in cruciferous vegetables are commonly derived from tryptophan (indolmethyl and *N*-methoxyindolmethyl glucosinolates) and in lesser quantities, methionine and phenylalanine (Traka & Mithen 2008).

As previously noted, these compounds are secondary plant metabolites and play a key role in plant defense systems; and their role in the response of Brassicas to environmental stress was recently reviewed (Martínez-Ballesta et al. 2013). Their products are known to be antimicrobial (Tierens 2001), fungicidal (Angus et al. 1994), and nematocidal (Potter et al. 1998; Avato et al. 2013) and have been shown to influence feeding preferences of invertebrate herbivores (Newman et al. 1992). This defensive role is linked to the products of glucosinolate breakdown. Specifically, when the plant tissue is injured or bruised, as happens in mastication, glucosinolates are hydrolysed by the enzyme myrosinase (Bones & Iversen 1985; Bones & Rossiter 1996). The hydrolysis reaction, shown in Figure 1.4, results in the production of nitriles, thiocyanates, or isothiocyanates, in quantities which depend on the reaction conditions, including pH, available substrate, presence of ferrous ions and the epithiospecifier protein (Bones & Rossiter 1996; Fahey et al. 2001). Before plant tissue injury, myrosinase is stored separately in the cell in vacuoles referred to as ‘myrosin cells’ (Bones & Iversen 1985). The rupture of these vacuoles permits myrosinase to come into contact with glucosinolates which leads to their hydrolysis (Figure 1.4). The glucosinolate-myrosinase system is reviewed in Bones & Rossiter (1996). There are various forms of the enzyme myrosinase, which has also been identified in bacteria and fungi (Fahey *et al.* 2001), sequenced and cloned (Xue et al. 1992).

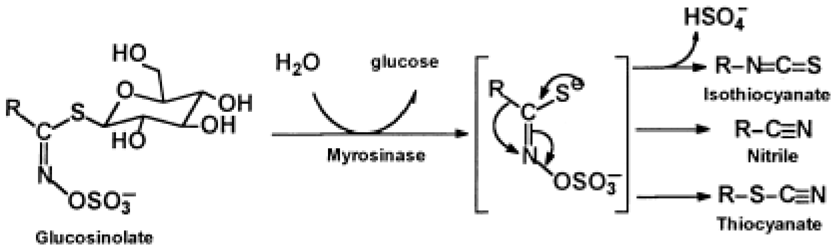


Figure 1.4 The hydrolysis of a glucosinolate, by the enzyme myrosinase, results in isothiocyanate, thiocyanate or nitrile products, depending on the reaction conditions. Figure from Fahey et al. (2001).

The chemical properties of ITC are responsible for their high molecular activity and their associated health benefits, which will be discussed in the next section. They are highly reactive compounds as the central carbon atom is electrophilic and will react with thiols, amines, and hydroxyl groups (Minarini et al. 2014). In addition to this, the particular side chain associated with an ITC type will determine its nature further, for example, how lipophilic the ITC is.

1.3.3 The Nutritional Composition of Watercress:

Watercress is often portrayed in the media as a ‘healthy’ or a ‘superfood’. These claims stem from the increasing number of biomedical studies focused on the health benefits of watercress and will be discussed further in this section. In fact, watercress was recently ranked as the top ‘powerhouse’ out of 47 fruits and vegetables based on its composition of essential and non-essential nutrients (Di Noia 2014). Table 1.1 shows vitamins and minerals found in 100g of fresh raw watercress, alongside broccoli and lettuce.

Table 1.1 Some of the nutritional components of watercress, broccoli and lettuce. Values are per 100g of fresh, raw material and units are listed in the table. Table modified from the USDA, 2008, National Database for Standard Reference.

Nutrient	**Value per 100g of raw fresh weight**			**Unit**
Vitamins and active compounds	*Watercress*	*Broccoli*	*Lettuce (Cos or Romaine)*	
Vitamin C (total ascorbic acid)	43	89.2	4	mg
Thiamin	0.09	0.07	0.07	mg

Riboflavin	0.12	0.12	0.07	mg
Niacin	0.2	0.64	0.31	mg
Pantothenic acid	0.31	0.57	0.14	mg
Vitamin B-6	0.13	0.18	0.07	mg
Folate total	0.009	0.06	0.14	mg
Vitamin A	3191	623	8710	IU
Vitamin E (α-tocopherol)	1	0.78	0.13	mg
β-carotene	1.91	0.62	5.23	mg
α-carotene	0	0.02	0	mg
Vitamin K (phylloquinone)	0.25	0.1	0.1	mg
Lutein and zeaxanthin	5.77	1.4	2.3	mg
Minerals				
Calcium	120	47	33	mg
Iron	0.2	0.73	0.97	mg
Magnesium	21	21	14	mg
Potassium	330	316	247	mg
Manganese	0.24	0.21	0.16	mg
Selenium	0.9	2.5	0.4	μg

Watercress has been consistently shown to have higher levels of antioxidants (AO) than other leafy green salads. Martínez-Sánchez et al. (2008) found watercress to have higher antioxindant capacity and ascorbic acid to other Brassicaceae leaves, namely mizuna and rocket, and Payne et al. (2013) found it to have higher AO capacity to rocket and spinach.The primary phenolic group in watercress has been described as flavonols and their derivatives, with various quercetin, kaempferol and isorhamnetin species (Martínez-Sánchez et al. 2008; Santos et al. 2014; Giallourou et al. 2016). Another study, focused specifically on organic baby-leaf watercress, identified high levels of the phenolics, including chlorogenic acid, quercetin-3-*O*-rutinoside, caffeoyltartaric acid, and isorhamnetin, with additional phenolics gallic acid and caffeic acid in lower concentrations (Aires et al. 2013). The presence of high levels of ascorbic acid in watercress has also been reported (Palaniswamy et al. 2003; Martínez-Sánchez et al. 2008; Santos et al. 2014).

Watercress is a good source of GLS and ITC. A number of GLS have been reported previously in watercress, specifically several aliphatic GLS: glucoibarin (7-methylsulfinylheptyl GLS), glucohirsutin (8-methylsulfinyloctyl GLS), 7-methylthioheptyl GLS, 8-methylthiooctyl GLS, 9-methylthiononyl GLS, and several aromatic GLS: gluconasturtiin (2-phenylethyl GLS), sinalbin (4-hydroxybenzyl GLS), and glucotropaeolin (benzyl GLS) (Kaoulla et al. 1980; Gil & MacLeod 1980; Rose et al. 2000; Fahey et al. 2001; Aires et al. 2013; Zeb 2015).

Gluconasturtiin appears to be the primary GLS in watercress with a presence in much higher concentrations that any of the others (Rose et al. 2000; Aires et al. 2013). It is derived from the amino acid phenylalanine, and its isothiocyanate derivative is β-phenethyl isothiocyanate, abbreviated PEITC and with a molecular formula C_9H_9NS (Figure 1.5). PEITC was found to be 3.1 times higher than other ITC in watercress (Rose et al. 2000) and is the cause of the crop's distinctive peppery flavour (Palaniswamy & McAvoy 2001).

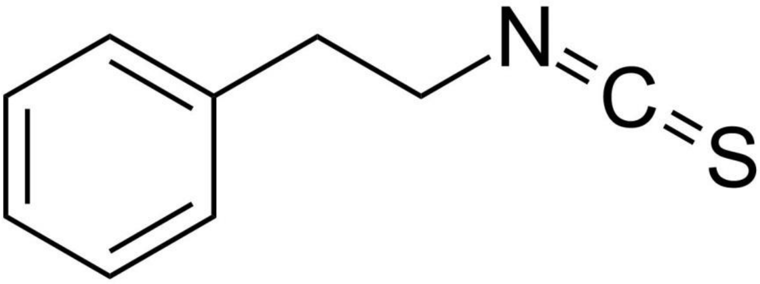


Figure 1.5 The structure of β-phenethyl isothiocyanate (PEITC), the primary isothiocyanate found in watercress, derived from gluconasturtiin

1.3.4 Health Benefits Associated with Watercress Consumption:

The consumption of Brassicaceae vegetables has been associated with reduced risk of diseases, including cancer, cardiovascular disease, diabetes, asthma, and neurodegenerative diseases (Verhoeven et al. 1996; Higdon et al. 2007; Manchali et al. 2012; Giacoppo et al. 2014). The benefits gained by consuming Brassicaceaes are delivered through a reduction in inflammation, an increase in antioxidant and detoxification capacities of cells, and chemopreventive power and are largely attributed to ITC (Cartea & Velasco 2008; Traka & Mithen 2008; Cavell et al. 2011; Wagner et al. 2013). Once consumed, ITC are absorbed through the epithelial cells of the intestines, where they are then bound by glutathione (GSH) - a thiol that plays a key role in maintaining cell oxidative stress levels by binding to target compounds and making them disposable - either directly or via glutathione-S-transferase (GSTs). The GSH-ITC complex is then digested via the mercapturic acid pathway and excreted in urine (Traka & Mithen 2008). This exposure leads to the activation of several pathways for which the key findings from cruciferous vegetables, and watercress in particular, are reviewed below.

1.3.4.1 *In vivo* studies on the effect of watercress on toxin metabolism:

The consumption of watercress is thought to have protective effects against toxins in humans. When were smokers were asked to consume watercress for 3 days, there was an increase in the concentration of deactivated tobacco carcinogens in the urine of subjects, suggesting an increase in the metabolism of these toxins with increased watercress consumption (Hecht et al. 1995). Gill *et al.* (2007) assessed the effect of watercress consumption on biomarkers thought to represent the risk of cancer in humans. The study found that men and women who consumed 85 g of watercress daily had less DNA damage and greater levels of antioxidant phytonutrients in their blood plasma, specifically 100% more lutein and 33% more β-carotene, than a control group and this effect was greater in smokers (Gill et al. 2007). More recently, a further two studies have further investigated the impact of watercress on DNA damage caused by toxins or oxidative stress. The first study identified that the effect of the teratogenic toxin cyclophosphamide was significantly reduced in the presence of watercress juice, suggesting a protective effect in mice (Casanova et al. 2013). Fogarty *et al.* (2013) found that various regiments of watercress consumption (short and long-term) reduced the DNA damage and H_2O_2 accumulations in healthy men after sessions of strenuous exercise (Fogarty et al. 2013). An increase in antioxidant tocopherols and xanthophyll in the subjects was also noted (Fogarty et al. 2013). These studies demonstrate a protective effect of watercress phytonutrients against toxic conditions in the cells in both humans and mice.

1.3.4.2 The antioxidant pathways of isothiocyanates:

It has become apparent that ITC play a large part in increasing the ability of cells to protect themselves against toxins and carcinogens. They do this by increasing antioxidant capabilities of cells directly and indirectly. As previously described, ITC become bound to glutathione (GSH) - a compound which binds to foreign toxins and makes then water-soluble- depleting cell reserves rapidly upon absorption (Traka & Mithen 2008). This depletion of GSH likely triggers an immediate up-regulation in GSH production, thus increasing the cells' ability to maintain an appropriate oxidation-reduction balance (Traka & Mithen 2008). A transcriptome analysis of human cells treated with sulforaphane also identified the up-regulation of certain heat shock proteins, which could boost a cell's ability to deal with misfolded or degraded proteins efficiently (Traka et al. 2005). This 'conditioning' of the cell is a likely pathway through which ITC boost cell performance.

1.3.4.2.1 Phase II enzymes:

There is much evidence that ITC enhance the activity of Phase II biotransformation enzymes (or drug-metabolising enzymes), such as glutathione-S-transferases (GSTs), quinone reductase, and glutamate cysteine ligase, all of which are active in the elimination of reactive oxygen species and carcinogens from the cell by converting them into inactive and water-soluble forms (Fahey et al. 1997; Rose et al. 2000; Cartea & Velasco 2008; Ernst et al. 2011). Human consumption of Brussels sprouts for 3 weeks led to a significant increase in plasma levels of GSTs in the blood (Bogaards et al. 1994). Another study showed the increase in GSTs in four different digestive tract tissues of rats after dietary treatment with PEITC (Van Lieshout et al. 1996). Watercress also contains methylsulfinylalkyl ITC which have been found to be 10 to 25 times more powerful inducers of Phase II enzymes than PEITC; however, they are present in much smaller quantities in the plant (Rose et al. 2000). Overall, as Phase II enzyme inducers, ITC increase the ability of an organism to deal with carcinogens quickly and efficiently.

Cellular uptake of ITC also appears to increase the cells antioxidant response by activation of the transcription factor Nfr2, possibly through signals transduction (Heiss et al. 2001; Cheung & Kong 2010; Wagner et al. 2013). This transcription factor is linked to a variety of antioxidant and anti-inflammatory genes and is bound in its inactive cytosolic state by the inhibitor protein Kelch-like ECH-associated protein 1 (Keap1). ITC appears to disrupt this complex, leading to the activation of Nfr2, after which Nfr2 moves to the nucleus and promotes the transcription of antioxidant genes, particularly the transcription of detoxifying Phase II enzymes (Cheung & Kong 2010; Wagner et al. 2013). This pathway is especially relevant to watercress nutrition, as PEITC was found to a very powerful inducer of Nfr2-led gene transcription in cells, certainly comparable to sulforaphane from broccoli and higher than other aromatic ITC (Ernst et al. 2011).

1.3.4.2.2 Phase I enzymes:

Phase I biotransformation enzymes, also drug-metabolising enzymes, are another component of the cell’s toxin metabolism. The Cytochrome P450 (CYP) family of Phase I enzymes make substances polar so that they can then be disposed of by Phase II enzymes (Traka & Mithen 2008). However, their activity often results in the activation of certain harmful carcinogens therefore compounds which activate CYP enzymes are assumed to contribute to carcinogenesis (Steinkellner et al. 2001; Cartea & Velasco 2008; Cheung & Kong 2010). ITC, such as sulforaphane and PEITC, have been shown to reduce the activity of rat and human Phase I CYP enzymes (Barcelo et al. 1996; Thapliyal & Maru 2001; Nakajima et al. 2001). On the other hand, PEITC was found to upregulate gene expression of both Phase II and Phase I enzymes in human hepatocytes, but the impact of this on the bigger picture is not clear (Gross-Steinmeyer et al. 2004). This pathway may be a powerful way through which ITC, like the ones found in watercress, can reduce the risk of cancer.

1.3.4.2.3 Consumer polymorphisms:

Interestingly, there is evidence that some consumer genotypes benefit more from ITC than others. Studies have focused on variations in GST genotypes which are involved in the ITC metabolism when it enters the cell. Homozygous individuals for the genotypes GSTM1-null and GSTT1-null are unable to make the particular GST enzyme, meaning that they are slower to dispose of ITC and have prolonged exposure to them (Higdon et al. 2007). London *et al.* (2000) examined the chemopreventive effects of ITC and found that there seemed to be a greater benefit in individuals with GST-deficient genotypes, suggesting that the longer exposure to ITC was beneficial (London et al. 2000). This evidence is particularly interesting when considering the possible applications of ITC in medicine and how the particular genetics of an individual may be involved.

To summarise, ITC increase the ability of cells to manage reactive oxygen species and toxins by amplifying a cell’s AO response, facilitating the activities of Phase II detoxifying enzymes and inhibiting the potentially unbeneficial activities of Phase I enzymes. In the long-term, the enhancement of cellular defences combats the occurrence of chronic disease in humans by minimising damage caused by oxidative stress and carcinogens.

1.3.4.3 The anti-inflammatory effect of isothiocyanates:

Chronic inflammation is another leading cause of many health problems and can be stimulated by oxidative stress. The consumption of vegetables high in ITC may be associated with benefits of reduced inflammatory response, such as offer a protective effect against neurodegenerative diseases (Giacoppo et al. 2015). In animal models, consumption and the topical application of watercress extracts was shown to decrease tissue swelling and damage in response to chemically-induced inflammation (Sadeghi et al. 2013).

1.3.4.3.1 Nuclear factor Kappa B:

Wagner *et al.* (2013) reviewed processes by which ITC preserve the transcription factor NfκB (Nuclear Factor Kappa B) in its inactive cytosolic form, which is bound to the inhibitory protein IκBa. Usually, an inflammatory signal will lead to NfκB- IκBa dissociation, allowing NfκB to initiate the expression of genes associated with inflammation, including angiogenesis genes (Wagner et al. 2013). ITC appear to inhibit the transition of NfκB to its active form. The pathways responsible for this are unclear, but may occur through down-regulation of some NfκB inducers, a stabilisation of IκBa, or an upstream control mechanism (Wagner et al. 2013). PEITC is an inhibitor of NfκB and has been noted to influence NfκB activity and gastric cancer invasiveness via controls on gene expression (Yang et al. 2010).

1.3.4.3.2 Migration Inhibitory Factor:

Recent work has identified another molecular target of ITC, specifically the pro-inflammatory cytokine macrophage migration inhibitory factor (MIF). Brown *et al.* (2009) investigated the molecular activity of PEITC and noted that it causes a significant conformational change to MIF, which likely leads to loss of catalytic function (Brown et al. 2009). MIF is involved in immune response, and increased levels in the cell are often associated with disease and tumorigenesis. Therefore, the fact that PEITC heavily targets this macrophage is a strong indication that it may reduce inflammation, and ultimately diseases associated with it.

1.3.4.4 Chemopreventive and chemotherapeutic pathways:

The effect of ITC on cancer cells has been very well studied. ITC act to inhibit initiation and metastasis, retard cancer cell growth, and induce cell apoptosis (For reviews see: Higdon *et al.* 2007; Traka & Mithen 2008; Czapski 2009; Cheung & Kong 2010). Two of the pathways relevant to cancer development have already been discussed. The transcription factor, NfκB, which is associated with inflammation and immune response, is also linked with cancer development. Not only is chronic inflammation linked with cancer development, but activation of NfκB increases the ability of cancerous cells to resist treatment and grow (Karin 2006). Therefore, the inhibition of the NfκB transcription factor reduces risk of serious diseases but also weakens the defences of cancerous cells. The second pathway is the inhibition of Phase I and the enhancement of Phase II enzymes by ITC, via which ITC inhibit cancer initiation through the elimination of potential cancer-inducing stimuli. Overall, the inhibition of NfκB activation by ITC and detoxification of carcinogens are important mechanisms that reduce carcinogen exposure and therefore the risk of cancer development, and potentially reduce fitness of present cancer cells.

1.3.4.4.1 Isothiocyanates and cancer initiation:

Angiogenesis is the process through which a new blood supply is created from an existing source in order to support new growth. This process is essential to cancer proliferation and metastasis and is a mandatory requirement for tumours to develop beyond 1-2 mm in diameter

(Cavell et al. 2011). Hypoxia inducible factors (HIF) are a family of transcription factors that play a role in regulating angiogenesis initiation (Reviewed in Cavell *et al.*, 2011). Studies have shown that PEITC inhibits HIF pathways of angiogenesis induction in breast cancer cells by preventing translation of RNAs associated with cell growth (Cavell 2012; Cavell et al. 2012). A pilot study showed that watercress extracts high in PEITC inhibited cancer cell growth and HIF activation. This was later confirmed in the plasma of healthy adults who consumed watercress (Syed Alwi et al. 2010). These results highlight a strong connection between PEITC and the prevention of angiogenesis; an important mechanism in the initiation of tumours.

1.3.4.4.2 Isothiocyanates and cancer cell cycle arrest and apoptosis:

ITC have been shown to cause apoptosis, or programmed cell death, and cell cycle arrest, disrupting growth and proliferation of cancer cells. PEITC is a particularly strong apoptotic agent. It has been recorded to cause apoptosis in cells of human sarcoma, gastric cancer, leukaemia, melanoma cells and cholangiocarcinoma, though the mechanisms may vary depending on tissue type (Hu et al. 2003; Losso & Truax 2009; Wu et al. 2011; Tusskorn et al. 2013; Huang et al. 2014). PEITC caused cancer cell apoptosis by mobilising calcium in the cell and through oxidative stress induced by GSH depletion (Tusskorn et al. 2013; Huang et al. 2014). Other mechanisms through which PEITC may cause cancer cell apoptosis are a decrease in anti-apoptotic proteins, direct binding of PEITC to tubulin, and inhibition of translation necessary to prevent apoptosis (Cheung & Kong 2010). The case of cholangiocarcinoma is particularly important, as this rare form of liver cancer is most often unresponsive to current chemotherapy drugs. This is an example of a situation that could greatly benefit from better understanding of these important chemopreventive agents.

Cell cycle arrest was first observed in cancer cell with PEITC exposure by Hasegawa et al. (1993). The processes involved are down-regulation of cyclins, a family of proteins involved in the cell cycle, and inhibition of cyclin-dependent kinases, in addition to epigenetic modifications (Cheung & Kong 2010). Histone modifications are an important post-translation control and there is evidence that ITC inhibit the deacetylation of histones changing the expression of growth genes in cancer cells. Research in human prostate cancer cells showed significant cell cycle arrest on exposure to PEITC and suggested that, via the inhibition histone deacetylase enzymes, PEITC led to histone acetylation and chromatin restructuring, which in turn made the tumour-suppressant p21 promoter available for transcription (Wang et al. 2008). Thakur et al. (2014) reviewed the epigenetic modifications associated with PEITC and sulforaphane (Thakur et al. 2014). Substances that alter the activity of histone deacetylases are particularly promising for applications in cancer-treatment pharmacology.

1.3.4.4.3 Isothiocyanates and cancer invasiveness:

Besides angiogenesis, ITC have also been found to block the activity of enzymes that are essential for cancer cell migration. Specifically, the activity of matrix metalloproteinases (MMP), which are elevated in cancer cells and thought to be important for invasiveness, was inhibited in human breast cancer cells by the ITC 7-methylsulfinylheptyl from watercress extracts (Rose et al. 2005). The restriction of MMP action by PEITC has been shown in human gastric and colon cancer cells (Lai et al. 2010; Yang et al. 2010); and the invasiveness of breast cancer has been limited *in vivo* by PEITC in mice (Gupta et al. 2013). Most recently, a second inhibition pathway was suggested. Watercress extracts and synthetic PEITC were shown to decrease the metastatic ability of colorectal cancer cells by impacting cell survival, colony formation and size, and anchoring ability by downregulation of the gene *CTNNB1* (Pereira et al. 2017). This gene encodes for β – catenin, a key protein in cell to cell adhesion and a component of a major carcinogenesis signalling pathway (Pereira et al. 2017). These studies are evidence of the existence of a strong inhibitory effect of PEITC and other ITC on cancer spread.

Through inhibition of initiation, cancer cell growth, angiogenesis and new colony anchoring, ITC are a very important nutritional compound but also offer insight into chemopreventive pathways for future treatment design.

1.3.4.5 The antibiotic properties of isothiocyanates:

ITC have also been attributed antibiotic properties that would seem to play a role in plant defences against pathogenesis (Tierens 2001), although the level of inhibition depended on the ITC and bacteria combination (Wilson et al. 2013). PEITC inhibited the growth of *Escherichia coli*, *Pseudomonas aeruginosa*, *Staphylococcus aureus,* and *Listeria monocytogenes* (Borges et al. 2014) and was the second most potent antimicrobial ITC tested after benzyl ITC (Wilson et al. 2013). In the first study, PEITC was able to reduce biofilm growth by 60%. Interestingly, sulforaphane was able to inhibit three types of antibiotic-resistant strains of *Helicobacter pylori*, a common source of gastric problems (Fahey et al. 2002). These findings suggest not only another potential health benefit of ITC but also another potential application as antibiotic agents.

1.3.4.6 Consideration of potential negative effects:

Beyond the protective properties of these compounds, there are questions about their potential toxicity. Many secondary metabolites are intended for plant defences and herbivory deterrence, and glucosinolate degradation products have been found to be antimicrobial (Tierens 2001), fungicidal (Angus et al. 1994), and nematocidal (Potter et al. 1998). There are seemingly few cases where normal cell lines were subject to ITC treatment along with cancerous cells. In this valuable example, the cytotoxicity of sulforaphane is examined in both healthy and abnormal prostate epithelial cells, showing that cell cycle arrest and apoptosis were only induced in the cancerous cells (Clarke et al. 2011). In another example, Telang et al. (2009) compared the effect of PEITC and sulforaphane treatment on the expression gene associated with estrogen receptors. The work was completed in both normal and cancerous human breast cell lines. Interestingly, there were greater changes in gene expression of normal cells than in cancerous cells, including the upregulation of tumour suppressant or cancer cell apoptosis genes and the authors concluded that this created a protective effect for the healthy cells (Telang et al. 2009). Overall, any potential negative effects or exposure thresholds of Brassicaceae compounds are not well investigated but there currently no known negative impacts in humans.

Negative effects have not been noted with watercress consumption either. Watercress extracts did not comprise the integrity of DNA in mice (Casanova et al. 2013). Addressing the issue of contaminant accumulation by watercress, watercress was not categorised as a “hyper-accumulator” of arsenic (Falinski et al. 2014). A second recent study exposed watercress to increased levels of arsenic, cadmium and lead but noted that the metals were not freely translocated to above ground structures and that phytotoxic symptoms expressed in the plant were such that it would not likely survive to reach the consumer (Gounden et al. 2015). On occasion, raw watercress has been associated with fascioliasis (liver fluke disease) and *Escherichia coli* infections (Launders et al. 2013) in humans, which are keenly managed against in the salads industry.

1.3.5 Applications of Isothiocyanates in Medicine:

In 2012, there were 14.1 million new cancer diagnoses worldwide and 8.2 million deaths (Ferlay et al. 2013). The need for chemopreventive and chemotherapeutic knowledge and medications continues to be a priority. ITC have been identified as compounds involved in anticancer pathways, and our understanding of these pathways is vital to the accumulation of knowledge that may lead to the development of treatment applications. Many chemotherapy drugs are greatly toxic and lack of target specificity, causing a large amount of damage to produce a small effect. The parallel use of ITC and chemotherapy medication not only increases the effectiveness of the medication by weakening the cancer cells, but also could allow for decreases in the dosage of toxic chemotherapy drugs (Reviewed in Minarini et al. 2014). Recent work showed that the common treatment for human cholangiocarcinoma cancer, crisplatin, was more effective when *in vitro* cells were treated with PEITC as well (Li et al. 2016). Using ITC in natural approaches or preventatively has also been suggested. For example, the use of PEITC or sulforaphane in combination with the anti-inflammatory component of turmeric, curcumin, improved the effect of these compounds on inflammation biomarkers (Cheung et al. 2009). In addition, synthetic ITC have been developed and tested; showing promising results of anti-inflammatory action (Prawan et al. 2009).

These findings suggest great potential for the application of ITC to be applied in the field of pharmacology and cancer treatment.

1.3.6 Growing Crops for Nutrition:

Backed by the strong foundation of biomedical evidence and a motivation to improve human health, producing crops with enhanced nutritional value is in the forefront of breeding and growing. Genotype will define an amount of the phenotypic variation in the trait of a crop (Kushad et al. 1999; Charron et al. 2005). The remaining deviation from the expected mean phenotype will be explained by the environment and the environment x genotype interaction (El-Soda et al. 2014). In its simplest form, cultivation practices can alter the ‘environment’ which the plant experiences; and, therefore, understanding how external factors affect a crop can help to shape management practices for optimized nutritional qualities.

As secondary metabolites play an important role in stress and defense responses, there are many other factors which, when varied, affect the amount of GLS and other phytonutrients present in plant tissue at any one time. These include the tissue or organ, developmental stage, environmental conditions, soil nutrients, and the extent of pest and herbivory threat (Booth et al. 1991; Brown et al. 2003; Yan & Chen 2007; Velasco et al. 2007; Payne 2011). For example, Velasco and co-workers investigated how the amount of glucosinolates and variation in glucosinolate types related to various conditions in kale, a *Brassica oleracea* member. They found that total GLS increased from seedling to sexual maturity and that composition of total GLS also changed with age, pest threat, and environmental variation (Velasco et al. 2007).

Variation in concentrations of secondary metabolite has been identified across plant organs, Brown et al. (2003) described reproductive organs as having higher amounts of GLS than other plant organs and younger leaves more than older leaves in *Arabidopsis*. Similarly, broccoli florets contained 10-100 times more glucoraphane that older mature broccoli heads (Fahey et al. 1997). In watercress, leaves had 5.6 times greater AO capacity than the stems (Payne 2011). Watercress tissue content of ascorbic acid and PEITC levels increased across the life cycle, reaching their maximum concentrations at 40 days (Palaniswamy et al. 2003). As the plant tissue harvested for human use will vary for each crop, determining how best to make use of this information towards improved phytonutrient content will also vary.

Growth environment and conditions clearly play a definitive role as well. Payne et al. (2015) found that watercress AO capacity was much greater in field-grown samples that were exposed to varied natural growth conditions and pests than in samples grown in controlled laboratory conditions (Payne 2011). Watercress was also found to contain higher amounts of gluconasturtiin when grown in longer days (16 h versus 8 h) and at moderate (10 – 15 ^{o}C) versus warmer (20 – 25 ^{o}C) temperatures (Engelen-Eigles et al. 2006). Soil nutrient availability has also been shown to affect phytonutrient accumulation in Brassicaceae crops. In rocket, lower nitrogen availability equated with higher levels of GLS (Chun et al. 2017), whereas gluconasturtiin concentrations in watercress, the precursor to PEITC, were increased with both nitrogen or sulfur application (Kopsell et al. 2007). Tissue concentrations of the antioxidants lutein, zeaxanthin and carotenoids in watercress also increased with increasing nitrogen concentrations (Kopsell et al. 2007). These types of variations are likely to be crop and environment-specific, but also could be useful tools in optimization of the phytonutrient profile of a crop.

These studies demonstrate that the crop’s phytonutrient profile will vary based on several variables, such as tissue type, environmental conditions, plant nutrition, and developmental stage. Ideally, all these components and their interactions will need to be elucidated or considered in order to how to achieve optimal nutritional values in a replicable and marketable manner.

1.3.7 Effect of Cooking and Storage on Nutritional Quality:

The impact of food preparation and storage on the availability and conversion of GLS and ITC is important as most Brassicaceaes are purchased, stored for a time, and prepared at home through some degree of cooking before consumption. There are three known issues to affect nutritional quality of Brassicaceae food during cooking: 1) the denaturation of heat-sensitive myrosinase, halting of GLS conversion to ITC, 2) the leaching of the GLS from the plant into the cooking media (Getahun & Chung 1999; Cartea & Velasco 2008; Traka & Mithen 2008), and 3) the degradation of GLS by heat, although this varies by GLS chemical structure (Hanschen et al. 2012). However, it appears that the hydrolysis reaction may not by limited to within the plant tissue itself. There is evidence that the conversion from glucosinolates to their isothiocyanate also occurs within mammalian digestive tracts by the present microflora (Getahun & Chung 1999; Shapiro et al. 2001). Degradation of GLS by several bacterial strains has been observed and is reviewed by Traka & Mithen (2008); but, as the authors point out, whether bacteria regularly facilitate the benefits of isothiocyanates to human health and to what extent is unknown. Research has also identified a potential benefit of limited cooking. A mild heating appears to denature the epithiospecifier protein (ESP), a protein which guides the glucosinolate conversion toward nitriles instead of ITC (Matusheski et al. 2004). Thus, mild cooking could improve the conversion to ITC instead of nitriles, assuming myrosinase has not been denatured.

As a leafy crop, watercress is most commonly consumed raw as a salad or garnish and, therefore, no heating takes place. However, cooking does occur in Asian-style cuisine and briefly in classic British watercress soup. In a study, cooked watercress did not show conversion of GLS to ITC, however there was evidence of some ITC production after consumption, presumably by human gut microflora (Getahun & Chung 1999). Steaming and microwaving watercress have been recommended over boiling in terms of preserving phenolic content and AO capacity (Giallourou et al. 2016). However, the benefit of mild cooking suggested earlier for other Brassicaceaes may not be relevant to watercress because it appears to lack ESP activity (Kaoulla et al. 1980). Eating watercress as a fresh salad is likely to yield the highest health benefit to the consumer.

Watercress tends to have a short shelf life. It has been determined to keep for approximately 6 days (Payne 2011). It is particularly susceptible to moisture loss but not cold damage, and these factors guide the processes that follow commercial watercress harvest, where it is immediately cooled to 0°C, packaged and shipped (Palaniswamy & McAvoy 2001). Some nutritional attributes are maintained well over time, specifically total phenolic content and levels of various derivatives did not change greatly over a ten day period of cold storage at 3 °C (Santos et al. 2014). This is consistent with commercial watercress post-harvest practises.

1.3.8 Summary:

To summarise, ITC in watercress have been shown to interact with human cells on many different levels and through these processes provide several health benefits. Understanding the pathways and molecules involved in what ultimately becomes a human health benefit is valuable in terms of improving human health and identifying what qualities to breed for. However, equally important is the understanding of the genes behind nutritional traits. In order to breed more nutritious foods, or maintain these traits while improving other aspects of a crop, breeders must be able to identify and preserve the relevant genetic information, as well as understand the environmental influences on expression. Watercress breeding efforts would benefit from knowledge about what genes and pathways contribute to the high nutritional quality of watercress in order to sustain and/or breed for them.

1.4 The Watercress Crop in the Future

1.4.1 Breeding Targets:

As the publication of research on the health benefits associated with a diet high in fruit and vegetables has increased consumer awareness, the demand for watercress production has increased along with the industry's desire for a larger market share. Pressures for sustainable intensification, while simultaneously producing a nutritious crop that visually fulfils the consumer's sensory expectations, are growing. In collaboration with a major watercress grower, the following 'wish list' of breeding targets has been formulated and includes breeding for consumer health benefits, improved crop morphology, and crop resource-use efficiency.

Breeding for improved consumer health benefit is thought to be a major selling point for the watercress crop. A variety with higher concentrations of phytonutrients could entice a larger consumer base to regularly include watercress in their diet, particular consumers interested in healthy eating. Arguably, on the fruit and vegetable scale of nutritional benefit, watercress already is ranked highly (Di Noia 2014); so would an improved variety merely act as a marketing ploy? Though interest in healthy diets seems to be on the rise, public campaigns to raise awareness do not have a large impact and many adults are still not meeting nutritional guidelines (Martin 2013; K. Hennig et al. 2014). Small portions of 'ultra' nutrient-dense food stuff may be an effective way of delivering important nutrition to a greater portion of the population. Successful examples of crops bred for improved nutrition are glucosinolate-enriched Beneforte® broccoli (Traka et al. 2013) and 'Golden Rice' with increased concentrations of β-carotene to tackle Vitamin A deficiency (Paine et al. 2005). The case for bio-fortified foods, current standings, and how to move forward are reviewed in (Martin & Li 2017).

Work toward developing an enhanced watercress cultivar would also elucidate the genetic bases of these benefits, leading to new knowledge of how to improve practices and how best to grow a crop that consistently meets high standards regardless of environmental variation or other limiting factors. Finally, this knowledge will be critical to preserving this trait while breeding for others as there may be unforeseen links between traits or loss of desirable alleles during breeding efforts that would result in an unexpected losses in nutritional quality. Care also needs to be taken with regards to the effect of any phytonutrient intensification on the crop's palatability.

The current morphology of the watercress crop is consistent with its traditional image of bunched watercress sold in London markets during the 1800's and closely resembles wild watercress as it has not been altered by long or intense domestication pressures (Manton 1935). The breeding target of producing a new watercress variety with improved morphology is guided directly by the industry's consumer research. A popular request is for a watercress crop that more resembles popular baby-leaf salad types and is bite-sized or easier to eat with a fork (S Rothwell 2014, personal communication). This task would consist of reducing the length of the watercress stem, possibly accompanied with an increase of leaf production. As mentioned previously, it would be important to not lose any of the crop's nutritious value in this process.

Finally, aspects of sustainable intensification of production are a critical target for most crops in light of climate change and limited resources. For watercress production, a specific and significant management issue has arisen in recent years. Effective January 2016, abstractions of phosphorous (P), the principle fertilizer supplement used in watercress farming, to chalk stream watershed will be limited to 20 mg/L of suspended solids and 0.065 mg/L of total reactive P (Environmental Agency, 2014). The new regulations limit fertilization options for watercress growers and will affect crop productivity (Figure 1.6). Currently, the only possible adaptations to these changes are compromises in management practices, such as taking farms off the river systems or moving the production abroad. Both options would come at the loss of traditional practices, a significant monetary cost, and potential domestic job loss (Vitacress Conservation Trust Meetings 2014-2016). It is also important to note that watercress farming has been linked with chalk stream ecosystems for many decades and is a major contributor to water levels in the summer months, so this transition would not be without effect (Casey & Smith 1994; Natural England 2009). A balance between meeting the needs of the watercress crop and protecting the surrounding habitat is desirable. The development watercress crop with improved P-use efficiency would be a promising avenue forward. As this target has been developed more recently, it will not be investigated further in the work presented here.

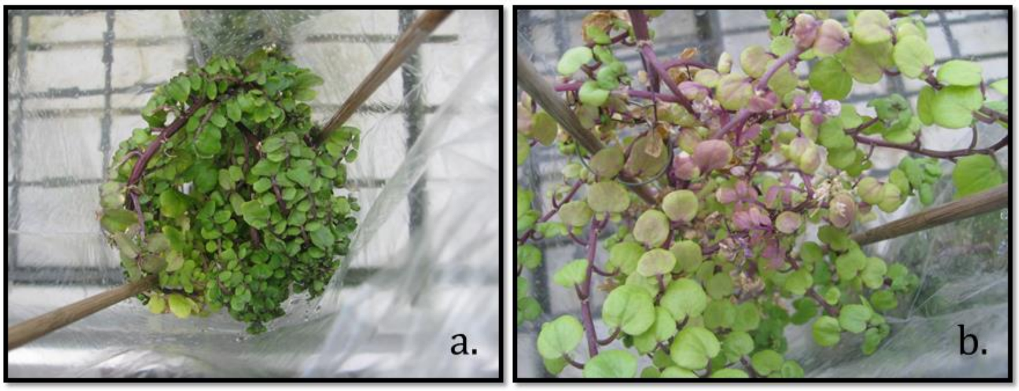


Figure 1.6 Images of a normal (a.) and a phosphorus-deficient (b.) watercress plant, as indicated by the purple/red coloration of the leaves

1.4.2 Current Breeding and Resources for Watercress:

Over the years, some instances and efforts to select for traits associated with commercial cultivation, such as frost or disease resistance, have been made within watercress companies; but there are no cultivars or varieties specifically bred for commercial production (Rothwell & Robinson 1986; Palaniswamy et al. 2003). The lack of knowledge regarding the genetic origin and diversity of *N. officinale* no doubt limits the potential development of new stock. For example, when watercress cultivation globally was affected by crook-root disease, caused by the pathogen *Spongospora*, interest in breeding a resistant cultivar was halted due to lack of ability to move this target forward (Sheridan et al. 2001). This pathogen was dealt with primarily through compromises in management practices, specifically discontinuing the cultivation of the winter crop, brown watercress. And although disease resistance is not a major concern to the industry presently, the interest in breeding improved watercress varieties that meet market demands exists as discussed previously.

To fill this gap, a watercress research and breeding program has been established at the University of Southampton (UoS) with collaboration and funding from a significant industry partner, in the laboratory of Professor Gail Taylor. Firstly, a germplasm collection was initiated and currently contains nearly 50 accessions. The collection was established through donations from Warwick Horticultural Research International, local and international growers, and collected from the wild by members of the UoS team. Figure 1.7 visualizes the collection's current global reach.

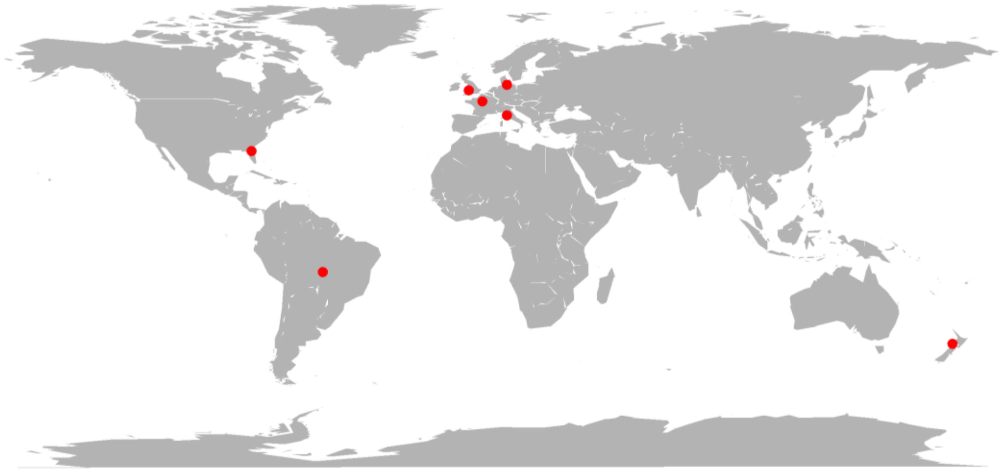


Figure 1.7 A global map indicating the current global reach of the UoS watercress germplasm collection

This collection represents a valuable resource for watercress and facilitated an initial assessment of the phenotypic diversity currently available for watercress breeding (Payne 2011; Payne et al. 2015). Payne (2011) surveyed differences in above-ground characters, specifically stem length, stem diameter, and number of leaves, and found promising variation in the first two traits. The collection was also screened for phytonutrient qualities. Antioxidant (AO) assays were also optimized and tested, finding significant differences in AO capacity between accessions in the collection (Payne 2011; Payne et al. 2013). The glucosinolates (GLS), isothiocyanates (ITC), and cancer cell cytotoxicity were quantified in a few accessions of interest (Payne et al. 2011). This work suggested promising potential for breeding a watercress variety with improved health benefits. In addition, a particular accession was marked of commercial interest and was later named 'Boldrewood'.

As promising phenotypic diversity was found within the germplasm collection for traits of interest, the genetic diversity within the collection was also examined. This was done using Amplified Fragment Length Polymorphisms (AFLPs) – a technique which creates a unique genetic fingerprint based on the difference in restriction enzyme digests that arise from polymorphisms in the DNA of organisms (Savelkoul et al. 1999). The AFLP analysis showed greater genetic diversity within accession replicates than between; leading the authors to believe that a genetic bottleneck, or a loss of the genetic variation in the genome due to domestication (Doebley et al. 2006), has not yet been reached in watercress possibly due to limited breeding selection (Payne et al. 2011). This was seen as a positive indicator for watercress breeding, suggesting a promising amount of allelic variation within the UoS collection from which to draw when developing an ideal new variety.

An effort to explain the potential role of gene expression in the phenotypic variation was also made. Gene expression analysis, through quantitative Polymerase Chain Reaction (qPCR), was completed to compare differential expression in three accessions, including Boldrewood, that were found to vary with a marked down-regulation of gene groups associated with plant growth and up-regulation of gene groups associated with plant defences in the dwarf and high AO Boldrewood accession (Payne et al. 2015).

This work established the foundations needed to move forward and already highlighted an accession of potential commercial value. However, the genetic architecture of a trait, such as the number of genes involved and the nature and magnitude of their effect greatly impacts the ability of breeders to make gains (Moose & Mumm 2008). With the advent of Next Generation Sequencing and other modern molecular technologies, it has become possible to develop this knowledge and genomic resources currently absent in watercress breeding, something not previously attainable for a crop with a modest market and financial backing. Much could be gained by watercress breeders and researchers from the pursuit of further molecular knowledge regarding this unique crop.

1.5 Research Tools: Next Generation Sequencing for Crop Breeding

1.5.1 Molecular Markers in Traditional Crop Breeding:

Traditional plant breeding involves discovery of an individual or plant group with a desirable trait, careful crossing of parents carrying the desirable phenotype, cultivation of progeny, and selection of the individuals that show this phenotype. This procedure would need to be repeated over many cycles to lock the phenotype into the plant material and required a large investment of time, capital, and labour with breeders often facing the loss of valuable genetic variation in non-target traits and the parallel inheritance of undesirable traits or deleterious genes (Collard et al. 2005). In addition, traits vary in their level of genetic complexity. Many traits that are important in agriculture, such as yield or tolerance to a stressor, do not follow a ‘Mendelian’ or qualitative trait pattern of simple presence or absence of a phenotype. Instead they are complex and exhibit a continuous phenotype, for example, a range of high to low tolerance to a stressor. Such traits are termed quantitative traits and their phenotype will be associated with multiple regions in a genome: the quantitative trait loci (QTL) (Collard et al. 2005).

Identifying and breeding for QTL is complex but the application of improved theoretical understanding and molecular techniques has revolutionized crop breeding and improved selection (Moose & Mumm 2008). The use of genetic markers to tag traits of interest allows breeders to screen for the presence of a trait in the DNA. Genetic markers are genetic differences, as small as a Single Nucleotide Polymorphism (SNPs), between species or individuals and are attributed to mutations (Collard et al. 2005; Zargar et al. 2015). When linked to a trait of interest, markers can be used to identify genetic variation within a germplasm collection and select for the presence of one or many desirable traits during breeding, a practice termed Marker Assisted Selection (MAS) (Xu & Crouch 2008). Through the application of MAS, breeders can use the markers to select for these traits at the seedling stage, thus reducing time, money, and space required in plant breeding (Varshney et al. 2009); breed for multiple desirable traits at once, called ‘pyramiding’; and inform breeding decisions with important facts about the location and recombination rates of undesirable or lethal genes (Collard et al. 2005; Xu & Crouch 2008).

Key applications that involve genetic markers are the development of mapping populations, linkage maps and the identification of QTL. A mapping population provides the data needed to understand the genomic location and inheritance patterns of key traits. Mapping populations are developed by crossing two parents that differ for one or more traits of interest, and then inbreeding or backcrossing for a number generations (Doerge 2002; Collard et al. 2005). During meiosis, a cross-over of genetic material occurs between homologous chromosomes, meaning that the chromosomes exchange some homologous genetic material, which allows for the progeny to inherit a unique combination of genetic information from the two parental chromosomes. Genes and markers that are located close together on a chromosome will tend to be inherited together more often than ones that are located further apart (Mauricio 2001).Based on this principle, the frequency of recombination of parental markers can be observed in the offspring and a recombination fraction can be computed (Mauricio 2001; Collard et al. 2005; Langridge & Fleury 2011). A recombination fraction is an estimation of the genetic distance between two markers. For example, markers with a recombination frequency of 50 % are unrelated; whereas markers inherited together 99 % are likely to be close together.

With the information about how a complete set of markers are related, a linkage map can be generated. A linkage map shows the theoretical position of markers relative to each other (Collard et al. 2005; Langridge & Fleury 2011). Linkage maps can then be used to identify the location of a gene or a QTL of interest. It should be noted that linkage groups do not necessarily correlate with the true physical positions of markers on chromosomes, only how markers in the dataset are related; and that linkage maps are unique to the particular mapping population and experiment from which they were constructed.

QTL analysis is a method used to identify regions controlling a trait of interest. It creates the link between the genotype and the phenotype which is the foundation of molecular plant breeding. For QTL analysis, the markers and their relative location on a linkage map must be complimented by a set of phenotypic data (Mauricio 2001). The offspring can then be grouped according to their expressed phenotype. Any markers strongly associated with this phenotype can be identified because they will have been consistently inherited with each phenotypic group (Collard et al. 2005; Langridge & Fleury 2011). Appropriate statistical tools are in place to test the significance and value of the QTL analysis. These have evolved from single marker testing and simple interval mapping (SIM) to composite interval mapping (CIM) and multi-QTL mapping (MQM) (Doerge 2002). Various factors can critically affect QTL results, depending on the complexity and nature of the trait being tagged; and they include the size of the population screened, environmental conditions that influence expressed phenotype, inaccurate evaluation of phenotype, and interactive effects between QTL (Collard et al. 2005).

The overall aim for plant breeding is to build enough information regarding QTLs, such as its location, the size of its effect, its strength in different environments, and, if possible, candidate causal genes in order to facilitate selection and breeding steps for the trait of interest (Moose & Mumm 2008). QTL analysis has become a broadly-used tool in the identification of the genetic basis of important agronomic traits. It has been applied to most major crops, including rice, soybean and oilseed rape, and across traits ranging from stress tolerance, such as salinity, drought and disease, to yield and nutritional composition (Toroser et al. 1995; Du et al. 2009; Marathi et al. 2012; Leonforte et al. 2013; Jeennor & Volkaert 2013; Tan et al. 2013; Bian et al. 2014; Yuste-Lisbona et al. 2014; D'hoop et al. 2014; Obara et al. 2014; Fang et al. 2014; Pan et al. 2017).

1.5.2 Next Generation Sequencing:

A limiting factor of QTL and candidate gene discovery traditionally was adequate marker development and genotyping. New technologies are seemingly exponentially expanding modern capabilities however. The major technological innovations behind Next Generation Sequencing (NGS) were driven by the commitment to sequence the human genome and understand disease. These new technological capabilities have significantly increased the number of markers that can be located and the number of individuals that can be screened (Moose & Mumm 2008). Three of the NGS technologies currently applied in crop breeding include: 454 sequencing by Roche Applied Science, Illumina's Solexa sequencing and SOLiD by Applied Bioscience. These technologies use read-by-synthesis or read-by-ligation during emulsion Polymerase Chain Reaction (PCR) to amplify DNA fragments and 'read' the sequence nucleotides (Reviewed in Ansorge 2009; Deschamps & Campbell 2009).

With NGS, library preparation can be streamlined, samples can be sequenced in parallel with the use of barcoding for retrospective sample identification, and the sequencing itself is a matter of hours. Compared to Sanger sequencing, these technologies produce massive amounts of sequence data – in the range of 21 million DNA fragments in a lane with Illumina HiSeqX (Zargar et al. 2015) – at a cheaper overall cost and in much less time; but the reads tend to be shorter and of lower quality than with Sanger sequencing and require a much greater computational investment and know-how to interpret and manage outputs (Ansorge 2009; Varshney et al. 2009; Zargar et al. 2015). However, the field of bioinformatics is growing constantly with efforts to develop new accurate and straight-forward pipelines to make the most of NGS outputs.

NGS is a game-changer for crop breeding because it makes acquiring knowledge about the genetics of a crop and developing molecular tools feasible, even for crops that are not heavily invested in and for which there is no background genetic information (Varshney et al. 2009), such as the understudied watercress.

1.5.2.1 NGS tools for crop improvement:

1.5.2.1.1 Whole transcriptome sequencing:

RNA Sequencing (RNASeq), also known as Whole Transcriptome Shotgun Sequencing, is a method developed to generate a snap-shot of the expressed genome and expression levels within a sample in a particular set of conditions (Wang et al. 2009). Naturally, gene expression can vary in tissue depending on developmental stage, environmental conditions, and tissue type, which is why RNASeq profiles expression under a specific set of parameters. Through this method, the active genes in a watercress can be identified and compared amongst different phenotypes. For example, what genes are active at high levels in high GLS watercress phenotypes that are not in low phenotypes, or vice versa?

During RNASeq, RNA is extracted and a cDNA fragment library is created with adaptors at each end, then each fragment is either single-end or pair-end sequenced (Wang et al. 2009). The resulting reads are then aligned to a reference genome or assembled *de novo,* thus creating a complete map of the expressed genome (Wang et al. 2009). Further annotation of these reads can provide a catalogue of expressed genes in a species. RNASeq also provides a method of quantifying gene expression. It is a particularly important tool because unlike previous methods, such as microarrays, it is not limited to known gene sequences (Wang et al. 2009). Pair-end sequencing is generally preferable because the overlap can help identify errors and align reads in *de novo* assembly (Haas et al. 2013). The high resolution output from RNASeq output can be used to detect polymorphisms down to the base pair level providing a functional collection of molecular markers for breeding selection. Using RNASeq for marker discovery applications produces a smaller number of markers than DNA-based methods. These are highly specific to the study; however, it does have the benefit of naturally excluding large repetitive sections of DNA and so reducing the complexity of large genomes (Zargar et al. 2015).

The application of this method has provided some successful results so far. RNASeq and *de novo* transcriptome assembly were applied to the coconut tree (*Cocos nucifera*), a valuable crop for which very little genomic information is available, resulting in the identification of 57,304 unigenes, 68.2% of which were successfully annotated (Fan et al. 2013). Transcriptome analysis for blueberries (*Vaccinium* section *Cyanococcus*) facilitated the identification of cold acclimation and fruit development genes (Rowland et al. 2012). As blueberries are an important nutritional super-food, this information will contribute to breeding of varieties that contain both climate adaptation and nutritional quality genes. In another applied example, the Illumina platform was used to genotype sorghum varieties with high and low tolerance to nitrogen stress, in order to identify the genes associated with low nitrogen tolerance and contribute markers for breeding through MAS (Gelli et al. 2014). Efforts to unravel P deficiency response, via transcriptome sequencing, in white lupin produced a list of candidate genes to improve this important trait through MAS (O’Rourke et al. 2014; O’Rourke et al. 2013). Applying this tool to sequence the watercress transcriptome and to compare the watercress germplasm collection would identify variation hot spots within a germplasm collection that are of commercial significance.

1.5.2.1.2 Reduced representation libraries:

The discovery of markers is an essential part of molecular plant breeding. Markers are used to assess differences between individuals or varieties and to identify the presence or absence of a particular allele. Until recently, applications such as SNP chip microarrays were the main tools available; but they require prior knowledge of the genome and are expensive (Peterson et al. 2012; Zargar et al. 2015). In addition, if used to genotype a new population, using pre-established markers from a different population could bias the results (Davey et al. 2011). With NGS technologies, there are new alternatives for marker discovery and genotyping.

Sequencing whole genomes is ideal because all variation would be covered, but cost is a limiting factor. An alternative approach has been to sequence a fraction of the genome, known as a reduced representation library (RRL). These types of approaches sequence samples at a low coverage (compared to whole-genome sequencing) by employing restriction enzymes (REs) to fragment the DNA at corresponding restriction sites. REs are very useful tools and, due to their great diversity in restriction sites and their methylation sensitivity, can be appropriately selected on a project-specific basis to meet particular requirements (Davey et al. 2011). Two methods that employ REs and are now widely used in marker discovery across species are Restriction-site Associated DNA Sequencing (RADSeq) and Genotyping-By-Sequencing (GBS) (Davey et al. 2010; Elshire et al. 2011).

In these procedures, DNA is extracted from each sample and fragmented with a suitable RE (Figure 1.8). The fragments from each sample are then ligated with a unique adaptor (barcode) and pooled, amplified, and sequenced together (Davey et al. 2010; Davey et al. 2011; Elshire et al. 2011). RADSeq and GBS are the same in principle but differ in certain library preparation steps. GBS has a simplified library preparation protocol which means that samples are handled less and thus cost and time is further reduced (Davey et al. 2011; Elshire et al. 2011). RADSeq includes fragment size selection and additional purification steps which GBS does not (Elshire et al. 2011). RADSeq has been further developed into double-digest RADSeq (ddRADSeq) which introduces further fragment selectivity by using both a rare and common cutting RE (Peterson et al. 2012).

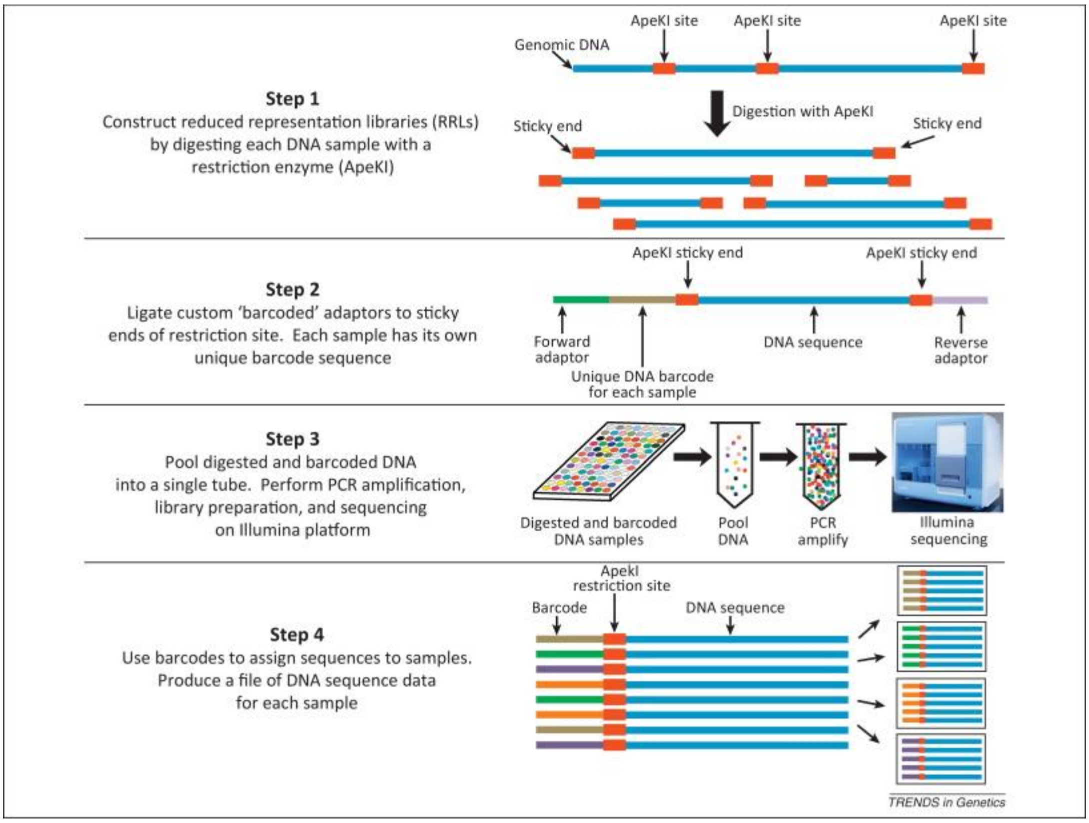

ApeKI site
ApeKI site
ApeKI site
Genomic DNA
Step 1
Construct reduced representation libraries (RRLs) by digesting each DNA sample with a restriction enzyme (ApeKI)
Digestion with ApeKI
Sticky end
Sticky end
ApeKI sticky end
ApeKI sticky end
Step 2
Ligate custom 'barcoded' adaptors to sticky ends of restriction site. Each sample has its own unique barcode sequence
Forward adaptor
DNA sequence
Reverse adaptor
Unique DNA barcode for each sample
Step 3
Pool digested and barcoded DNA into a single tube. Perform PCR amplification, library preparation, and sequencing on Illumina platform
Digested and barcoded DNA samples
Pool DNA
PCR amplify
Illumina sequencing
ApeKI restriction site
Barcode
DNA sequence
Step 4
Use barcodes to assign sequences to samples. Produce a file of DNA sequence data for each sample
TRENDS in Genetics

Figure 1.8 Genotyping-by-sequencing (GBS) protocol as depicted in Myles (2013). DNA is fragmented with a RE and adaptors are ligated to provide unique identification for each sample. This allow for samples to be pooled reducing cost of sequencing.

RADSeq and GBS offer several advantages. Firstly, because of the consistency in the how REs work, the same fragments of the genome will be sequenced across samples and the fragments sequenced can be reliably reproduced (Elshire et al. 2011). Secondly, as the sequenced fragments are comparable, loci can be assembled *de novo* and no reference genome is required, meaning that these methods can be applied in species that have no molecular tools available already, like watercress (Davey et al. 2011). And although it is critical to select a suitable RE, barcode use allows for pooling of numerous samples into the same sequencing channel leading to workable sequencing cost (Davey et al. 2011; Elshire et al. 2011). For example, Elshire et al. (2011) produced 200,000 markers for maize at a cost of $ 8,000. Finally, these approaches are ideal for candidate gene mapping because they can capture variation in regions of the DNA with regulatory function, unlike RNASeq which would only capture coding regions (Elshire et al. 2011). Table 1.2 compiles recent studies that applied RADSeq and GBS for marker discover and genetic linkage map construction in various plant species, many of which are not are not well studied.

Table 1.2 A selection of current literature using high-throughput sequencing for marker discovery and linkage mapping in plant species. The method used to sequence the markers is Restriction-site Associated DNA (RAD), double-digested Restriction-site Associated DNA (ddRAD), or Genotyping-By-Sequencing (GBS). The mapping method is denoted '*de novo*' if SNPs and map were assembled *de novo* from the sequencing outputs or 'combined' if previously available resources were also used.

Species	**Population size**	**Generation**	**Method**	**Markers produced**	**Markers mapped**	**Mean distance (cM)**	**Mapping method**	**Citation**	**Journal**
jujube	107	F1	RAD	42,784	2748	0.34	*de novo*	(Zhao et al. 2014)	PLoS ONE
chickpea	92	F1	RAD	51,632	604	0.5	combined	(Deokar et al. 2014)	BMC Genomics
peanut	166	F9	ddRAD	14,663	1685	1.95	combined	(Zhou et al. 2014)	BMC Genomics
raspberry	74	F1	GBS	9,143	4521	0.1	combined	(Ward et al. 2013)	BMC Genomics
grape	100	F1	RAD	1,814	1646	1.71	*de novo*	(Wang et al. 2012)	BMC Plant Biology
eggplant	156	F2	RAD	Unknown	431	3.8	combined	(Barchi et al. 2012)	PLoS ONE
wheat	96	F10	GBS	Unknown	972	1.86	combined	(Talukder et al. 2014)	BMC Genetics
watermelon	91	F2	GBS	379,125	266	Unknown	combined	(Lambel et al. 2014)	Theor Applied Genetics
cotton	170	F2	GBS	23,576	3,978	0.62	*de novo*	(Li et al. 2017)	PLoS ONE
zucchini	120	F8	GBS	26,430	7,718	0.4	*de novo*	(Montero-Pau et al. 2017)	BMC Genomics
cowpea	170	F2	RAD	34,868	17,996	0.066	*de novo*	(Pan et al. 2017)	Frontiers in Plant Science
oil palm tree	108	F2	GBS	21,471	1,085	1.26	*de novo*	(Pootakham, Jomchai, et al. 2015)	Genomics
rubber tree	118	F1	GBS	21,353	2,052	0.89	*de novo*	(Pootakham, Ruang-Areerate, et al. 2015)	Frontiers in Plant Science
apple	89	F1	GBS	270,000	2436	0.68	*de novo*	(Gardner et al. 2014)	Genes, Genomes, Genetics
olive	121	F1	GBS	10,941	5,643	0.53	*combined*	(Ipek et al. 2016)	Biochemical Genetics
coffee	278	F2	GBS	Unknown	848	4.52	combined	(Moncada et al. 2016)	Tree Genetics & Genomics

1.5.3 Applications in Brassicaceae crop breeding:

The Brassicaceae family contains numerous valuable crop species and the model research species *Arabidopsis*, and so a great number of genomic studies and resources have already been published for several of its members. *Arabidopsis* was the first plant to be fully sequenced (The Arabidopsis Genome Initiative 2000) and has various dedicated, open-access and curated databases, such as TAIR (Swarbreck et al. 2008), Araport (Krishnakumar et al. 2015) and ThaleMine (Krishnakumar et al. 2016). Whole genome sequencing and draft genomes exist for members of the *Brassica* genus also, such as *B. rapa* (X. Wang et al. 2011), *B. napus* (Chalhoub et al. 2014) and *B. oleracea* (Liu et al. 2014), accompanied by databases, such as BRAD (Cheng et al. 2011), with useful tools such as genome browsers, gene annotations and marker lists. Excitingly, a draft genome for *Barbarea vulgaris* was recently published (Byrne et al. 2017), which is thought to be closely related to watercress (Sheridan et al. 2001; Beilstein et al. 2006).

Molecular tools have been broadly applied in the Brassicaceae family with breeding targets in mind. QTL mapping has been applied in efforts to breed high oil content and weight in oilseed rape (Fan et al. 2010; Cai et al. 2012; Jiang et al. 2014), and pathogen and pest resistance in various species (Kuzina et al. 2011; Chen et al. 2013). Recently, GBS was used to map heat tolerance in broccoli in light of rising global temperatures (Branham et al. 2017). A particular focus has been the identification of candidate genes for GLS content for use in MAS in Brassicaceaes (Toroser et al. 1995; Uzunova et al. 1995; Heidel et al. 2006; Lou et al. 2008; Kuzina et al. 2011; Sotelo et al. 2014). RNASeq has also been applied toward understanding glucosinolate (GLS) gene expression and pathways in Brassicaceae species (Harper et al. 2012; Zhang et al. 2016; Lee et al. 2017).

These and similar works have greatly expanded our understanding of glucosinolate biosynthesis and metabolism in Brassicaceaes, such as the important role of MYB transcriptome factors reviewed in Seo & Kim (2017). Zhang et al. (2015) generated transgenic *B. napus* lines for three genes previously shown to be important in GLS pathways of *Arabidopsis*: a methylthioalkylmalate synthase (MAM), which control side chain elongation; a CYP83 which acts in core structure formation; and a glucosyltransferase UGT74. These transgenic plants had better resistance to pathogenic fungi and greater GLS content, showing the feasibility of creating improved varieties with the knowledge gained from molecular work (Zhang et al. 2015).

Probably the most notable example of the application of molecular techniques and MAS in Brassicaceae is that of nutritionally-enhanced Beneforte® broccoli (Mithen et al. 2003; Traka et al. 2013). Initially, a 10-fold increase in the GLS glucoraphanin and a 100-fold increase in the activation of the phase II enzymes were observed in hybrids made by crossing cultivated broccoli with a wild relative, *Brassica villosa* (Faulkner et al. 1998). This cross was taken forward and used to identify three QTLs with segments introgressed from the *B. villosa* genome into the cultivated variety that were likely responsible for the phytonutrient enrichment (Mithen et al. 2003; Sarikamis et al. 2006). This work led to three independent breeding efforts; and the hybrids produced were shown to share seven *B. villosa* SNPs, some collocating with the previously-identified QTL and a common candidate gene of *myb28* (Traka et al. 2013). One of the outputs of these efforts is the now commercial product known as Beneforte® broccoli, or super broccoli.

Resources for Brassicaceaes are only likely to increase further in the near future, through the ease of application of NGS technologies coupled with improvements in software pipelines for polyploids (Wei et al. 2013). As aforementioned, tools are limited for watercress, yet tools developed for closely-related Brassicaceaes may be helpful. For example, watercress genes were successfully hybridised to an *Arabidopsis* gene chip suggesting that the fully sequenced *Arabidopsis* genome may serve as a useful resource in the identification of putative genes function through sequence similarity (Payne 2011). However, there are still large gaps to be filled. There are no molecular markers, linkage maps or known QTLs for the phytonutrient or agronomic traits discussed in this review. The application of NGS techniques to development of such genomic resources would be an important undertaking for watercress and would make these tasks achievable in reasonable time and budget. Ultimately, a better understand of what genes are associated with various desirable phenotypes in watercress would inform future breeding decisions and assist in meeting commercial and environmental needs.

1.6 Research Aims and Objectives

This literature review has endeavoured to gather the knowledge relevant to watercress research and breeding. It has also suggested certain gaps in this knowledge which hinder future progress in this field of research and improvements to the commercial crop. The overall aim of this PhD is to elucidate the genetic bases of important agronomic traits in watercress, specifically the establishment of the necessary genetic knowledge to sustain and/or improve crucial nutritional and morphological attributes of the crop.

With the discussed traits of interest in mind, the specific objectives include:

1.) the further evaluation of commercially-relevant material in the University of Southampton's watercress germplasm collection;

2.) the completion of a whole-transcriptome study in watercress, aimed at identifying the genes responsible for nutritional benefits in watercress;

3.) the construction of the first watercress genetic linkage map;

4.) the detection of quantitative trait loci for the traits of interest; and

5.) the development of molecular markers datasets, which may then enable further watercress breeding

1.7 References

Aires, A., Carvalho, R., Rosa, E. a. S., & Saavedra, M. J. (2013). Phytochemical characterization and antioxidant properties of baby-leaf watercress produced under organic production system. *CyTA - Journal of Food*, *11*(4), 343–351.

Al-Shehbaz, I. A., & Price, R. A. (1998). Delimitation of the genus Nasturtium (Brassicaceae). *Novon*, *8*(2), 124–126.

Angus, J. F., Gardner, P. A., Kirkegaard, J. A., & Desmarchelier, J. M. (1994). Biofumigation: Isothiocyanates released frombrassica roots inhibit growth of the take-all fungus. *Plant and Soil*, *162*(1), 107–112. https://doi.org/10.1007/BF01416095

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology, 25*(4), 195–203. https://doi.org/10.1016/j.nbt.2008.12.009

Aune, D., Giovannucci, E., Boffetta, P., Fadnes, L. T., Keum, N., Norat, T., … Tonstad, S. (2017). Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—a systematic review and dose-response meta-analysis of prospective studies. *International Journal of Epidemiology*, *46*(3), 1029–1056. https://doi.org/10.1093/ije/dyw319

Avato, P., & Argentieri, M. P. (2015). Brassicaceae: a rich source of health improving phytochemicals. *Phytochemistry Reviews, 14*(6), 1019-1033. https://doi.org/10.1007/s11101-015-9414-4

Avato, P., D'Addabbo, T., Leonetti, P., & Argentieri, M. P. (2013). Nematicidal potential of Brassicaceae. *Phytochemistry Reviews, 12*(4), 791-802. https://doi.org/10.1007/s11101-013-9303-7

Bailey, C. D., Koch, M. A., Mayer, M., Mummenhoff, K., O'Kane, S. L., Warwick, S. I., … Al-Shehbaz, I. A. (2006). Toward a global phylogeny of the Brassicaceae. *Molecular Biology and Evolution*, *23*(11), 2142–60. https://doi.org/10.1093/molbev/msl087

Barcelo, S., Gardiner, J. M., Gescher, A., & Chipman, J. K. (1996). CYP2E1-mediated mechanism of anti-genotoxicity of the broccoli constituent sulforaphane. *Carcinogenesis, 17*, 277–282. https://doi.org/10.1093/carcin/17.2.277

Beilstein, M. A., Al-Shehbaz, I. A., & Kellogg, E. A. (2006). Brassicaceae phylogeny and trichome evolution. *American Journal of Botany, 93*(4), 607–19. https://doi.org/10.3732/ajb.93.4.607

Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., & Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(43), 18724–8. https://doi.org/10.1073/pnas.0909766107

Bian, Y., Sun, D., Gu, X., Wang, Y., Yin, Z., Deng, D., … Li, G. (2014). Identification of QTL for stalk sugar-related traits in a population of recombinant inbred lines of maize. *Euphytica*, *198*(1), 79–89. https://doi.org/10.1007/s10681-014-1085-5

Björkman, M., Klingen, I., Birch, A. N. E., Bones, A. M., Bruce, T. J. A., Johansen, T. J., … Stewart, D. (2011). Phytochemicals of Brassicaceae in plant protection and human health--influences of climate, environment and agronomic practice. *Phytochemistry*, *72*(7), 538–656. https://doi.org/10.1016/j.phytochem.2011.01.014

Bleeker, W., Huthmann, M., & Hurka, H. (1999). Evolution of the hybrid taxa in Nasturtium R.BR. (Brassicaceae). *Folia Geobotanica*, *34*, 421–433.

Block, G., Patterson, B., & Subar, A. (1992). Fruit, vegetables, and cancer prevention: a review of the epidemiological evidence. *Nutrition and Cancer*, *18*(1), 1–29. https://doi.org/10.1080/01635589209514201

Bogaards, J. J., Verhagen, H., Willems, M. I., van Poppel, G., & van Bladeren, P. J. (1994). Consumption of Brussels sprouts results in elevated alpha-class glutathione S-transferase levels in human blood plasma. *Carcinogenesis*, *15*(5), 1073–1075.

Bones, A. M., & Iversen, T.-H. (1985). Myrosin cells and myrosinase. *Israel Journal of Botany*, *34*(2–4), 351–376.

Bones, A. M., & Rossiter, J. T. (1996). The myrosinase-glucosinolate system, its organisation and biochemistry. *Physiologia Plantarum*, *97*(1), 194–208.

Booth, E. J., Walker, K. C., & Griffiths, D. W. (1991). A time-course study of the effect of sulphur on glucosinolates in oilseed rape (*Brassica napus*) from the vegetative stage to maturity. *Journal of the Science of Food and Agriculture*, *56*(4), 479–493.

Borges, A., Simões, L. C., Saavedra, M. J., & Simões, M. (2014). The action of selected isothiocyanates on bacterial biofilm prevention and control. *International Biodeterioration & Biodegradation*, *86*, 25–33. https://doi.org/10.1016/j.ibiod.2013.01.015

Branham, S. E., Stansell, Z. J., Couillard, D. M., & Farnham, M. W. (2017). Quantitative trait loci mapping of heat tolerance in broccoli (*Brassica oleracea* var. italica) using genotyping-by-sequencing. *Theoretical and Applied Genetics*, *130*(3), 529–538. https://doi.org/10.1007/s00122-016-2832-x

Brown, K. K., Blaikie, F. H., Smith, R. a J., Tyndall, J. D. a, Lue, H., Bernhagen, J., … Hampton, M. B. (2009). Direct modification of the proinflammatory cytokine macrophage migration inhibitory factor by dietary isothiocyanates. *The Journal of Biological Chemistry*, *284*(47), 32425–32433. https://doi.org/10.1074/jbc.M109.047092

Brown, P., Tokuhisa, J., Reichelt, M., & Gershenzon, J. (2003). Variation of glucosinolate accumulation among different organs and developmental stages of *Arabidopsis thaliana*. *Phytochemistry*, *62*(3), 471–481. https://doi.org/10.1016/S0031-9422(02)00549-6

Byrne, S. L., Erthmann, P. Ø., Agerbirk, N., Bak, S., Hauser, T. P., Nagy, I., … Asp, T. (2017). The genome sequence of *Barbarea vulgaris* facilitates the study of ecological biochemistry. *Scientific Reports*, *7*, 40728. https://doi.org/10.1038/srep40728

Cai, G., Yang, Q., Yang, Q., Zhao, Z., Chen, H., Wu, J., … Zhou, Y. (2012). Identification of candidate genes of QTLs for seed weight in Brassica napus through comparative mapping among Arabidopsis and Brassica species. *BMC Genetics*, *13*(1), 105. https://doi.org/10.1186/1471-2156-13-105

Cartea, M. E., & Velasco, P. (2008). Glucosinolates in Brassica foods: bioavailability in food and significance for human health. *Phytochemistry Reviews*, *7*(2), 213–229. https://doi.org/10.1007/s11101-007-9072-2

Casanova, N. A., Ariagno, J. I., López Nigro, M. M., Mendeluk, G. R., de los A Gette, M., Petenatti, E., … Carballo, M. A. (2013). In vivo antigenotoxic activity of watercress juice (*Nasturtium officinale*) against induced DNA damage. *Journal of Applied Toxicology : JAT*, *33*(9), 880–885. https://doi.org/10.1002/jat.2746

Casey, H., & Smith, S. M. (1994). The effects of watercress growing on chalk headwater streams in Dorset and Hampshire. *Environmenral Pollution*, *85*, 217–228.

Cavell, B. E. (2012). *In vitro analysis of potential anticancer effects associated with watercress*. *PhD Thesis*. University of Southampton.

Cavell, B. E., Syed Alwi, S. S., Donlevy, A. M., Proud, C. G., & Packham, G. (2012). Natural product-derived antitumor compound phenethyl isothiocyanate inhibits mTORC1 activity via TSC2. *Journal of Natural Products*, *75*(6), 1051–1057.

Cavell, B. E., Syed Alwi, S. S., Donlevy, A., & Packham, G. (2011). Anti-angiogenic effects of dietary isothiocyanates: mechanisms of action and implications for human health. *Biochemical Pharmacology, 81*(3), 327–336.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., … Wincker, P. (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*, *345*(6199), 950–953. https://doi.org/10.1126/science.1253435

Charron, C. S., Saxton, A. M., & Sams, C. E. (2005). Relationship of climate and genotype to seasonal variation in the glucosinolate-myrosinase system. I. Glucosinolate content in ten cultivars of *Brassica oleracea* grown in fall and spring seasons. *Journal of the Science of Food and Agriculture*, *85*(4), 671–681. https://doi.org/10.1002/jsfa.1880

Chen, J., Jing, J., Zhan, Z., Zhang, T., Zhang, C., & Piao, Z. (2013). Identification of novel QTLs for isolate-specific partial resistance to *Plasmodiophora brassicae* in *Brassica rapa*. *PloS One*, *8*(12), e85307. https://doi.org/10.1371/journal.pone.0085307

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., … Wang, X. (2011). BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biology*, *11*(1), 136. https://doi.org/10.1186/1471-2229-11-136

Cheung, K. L., Khor, T. O., & Kong, A.-N. (2009). Synergistic effect of combination of phenethyl isothiocyanate and sulforaphane or curcumin and sulforaphane in the inhibition of inflammation. *Pharmaceutical Research*, *26*(1), 224–231. https://doi.org/10.1007/s11095-008-9734-9

Cheung, K. L., & Kong, A.-N. (2010). Molecular targets of dietary phenethyl isothiocyanate and sulforaphane for cancer chemoprevention. *The AAPS Journal*, *12*(1), 87–97. https://doi.org/10.1208/s12248-009-9162-8

Chun, J.-H., Kim, S., Arasu, M. V., Al-Dhabi, N. A., Chung, D. Y., & Kim, S.-J. (2017). Combined effect of Nitrogen, Phosphorus and Potassium fertilizers on the contents of glucosinolates in rocket salad (*Eruca sativa* Mill.). *Saudi Journal of Biological Sciences*, *24*(2), 436–443. https://doi.org/10.1016/j.sjbs.2015.08.012

Clarke, J. D., Hsu, A., Yu, Z., Dashwood, R. H., & Ho, E. (2011). Differential effects of sulforaphane on histone deacetylases, cell cycle arrest and apoptosis in normal prostate cells versus hyperplastic and cancerous prostate cells. *Molecular Nutrition & Food Research*, *55*, 999–1009. https://doi.org/10.1002/mnfr.201000547

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, *142*(1–2), 169–196. https://doi.org/10.1007/s10681-005-1681-5

Czapski, J. (2009). Cancer preventing properties of cruciferous vegetables. *Vegetable Crops Research Bulletin*, *70*(1), 5–18. https://doi.org/10.2478/v10032-009-0001-3

D'hoop, B. B., Keizer, P. L. C., Paulo, M. J., Visser, R. G. F., van Eeuwijk, F. A., & van Eck, H. J. (2014). Identification of agronomically important QTL in tetraploid potato cultivars using a marker-trait association analysis. *Theoretical and Applied Genetics*, *127*(3), 731–748. https://doi.org/10.1007/s00122-013-2254-y

Davey, J. W., Davey, J. L., Blaxter, M. L., & Blaxter, M. W. (2010). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, *9*(5–6), 416–423. https://doi.org/10.1093/bfgp/elq031

Davey, J. W., Hohenlohe, P. a, Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

Dekker, M., Verkerk, R., & Jongen, W. (2000). Predictive modelling of health aspects in the food production chain: a case study on glucosinolates in cabbage. *Trends in Food Science & Technology*, *11*(4–5), 174–181. https://doi.org/10.1016/S0924-2244(00)00062-5

Deschamps, S., & Campbell, M. a. (2009). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding*, *25*(4), 553–570. https://doi.org/10.1007/s11032-009-9357-9

Di Noia, J. (2014). Defining powerhouse fruits and vegetables: a nutrient density approach. *Prev Chronic Dis*, *11*, 130390.

Dixon, M. J. (2010). *The sustainable use of water to mitigate the impact of watercress farms on chalk streams in southern England*. University of Southampton.

Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006). The Molecular Genetics of Crop Domestication. *Cell*, *127*, 1309–1321.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet*, *3*(1), 43–52. https://doi.org/10.1038/nrg703

Du, W., Yu, D., & Fu, S. (2009). Detection of quantitative trait loci for yield and drought tolerance traits in soybean using a recombinant inbred line population. *Journal of Integrative Plant Biology*, *51*, 868–878. https://doi.org/10.1111/j.1744-7909.2009.00855.x

El-Soda, M., Malosetti, M., Zwaan, B., Koornneef, M., & Aarts, M. (2014). Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends in Plant Science*, *19*(6), 390–398. https://doi.org/10.1016/J.TPLANTS.2014.01.001

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a, Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Engelen-Eigles, G., Holden, G., Cohen, J. D., & Gardner, G. (2006). The effect of temperature, photoperiod, and light quality on gluconasturtiin concentration in watercress (*Nasturtium officinale* R. Br.). *Journal of Agricultural and Food Chemistry*, *54*(2), 328–334.

Environmental Agency. (2014). *Notice of variation and consolidation with introductory note: The environmental permitting (England & Wales) regulations 2010*.

Ernst, I. M. a, Wagner, A. E., Schuemann, C., Storm, N., Höppner, W., Döring, F., … Rimbach, G. (2011). Allyl-, butyl- and phenylethyl-isothiocyanate activate Nrf2 in cultured fibroblasts. *Pharmacological Research : The Official Journal of the Italian Pharmacological Society*, *63*(3), 233–240. https://doi.org/10.1016/j.phrs.2010.11.005

Ettlinger, M. G., & Lundeen, A. J. (1956). The structure of sinigrin and sinalbin; an enzymatic rearrangement. *Journal of the American Chemical Society*, *78*(16), 4172–4173. https://doi.org/10.1021/ja01597a090

Fahey, J. W., Haristoy, X., Dolan, P. M., Kensler, T. W., Scholtus, I., Stephenson, K. K., … Lozniewski, A. (2002). Sulforaphane inhibits extracellular, intracellular, and antibiotic-resistant strains of *Helicobacter pylori* and prevents benzo[a]pyrene-induced stomach tumors. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(11), 7610–7615. https://doi.org/10.1073/pnas.112203099

Fahey, J. W., Zalcmann, A. T., & Talalay, P. (2001). The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry*, *56*(1), 5–51.

Fahey, J. W., Zhang, Y., & Talalay, P. (1997). Broccoli sprouts: An exceptionally rich source of inducers of enzymes that protect against chemical carcinogens. *Proceedings of the National Academy of Sciences*, *94*(19), 10367–10372. https://doi.org/10.1073/pnas.94.19.10367

Falinski, K. A., Yost, R. S., Sampaga, E., & Peard, J. (2014). Arsenic accumulation by edible aquatic macrophytes. *Ecotoxicology and Environmental Safety*, *99*, 74–81.

Fan, C., Cai, G., Qin, J., Li, Q., Yang, M., Wu, J., … Zhou, Y. (2010). Mapping of quantitative trait loci and development of allele-specific markers for seed weight in *Brassica napus*. *Theoretical and Applied Genetics*, *121*(7), 1289–1301. https://doi.org/10.1007/s00122-010-1388-4

Fan, H., Xiao, Y., Yang, Y., Xia, W., Mason, A. S., Xia, Z., … Tang, H. (2013). RNA-Seq analysis of *Cocos nucifera*: transcriptome sequencing and de novo assembly for subsequent functional genomics approaches. *PloS One*, *8*(3), e59997. https://doi.org/10.1371/journal.pone.0059997

Fang, D. D., Jenkins, J. N., Deng, D. D., McCarty, J. C., Li, P., & Wu, J. (2014). Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in Upland cotton (*Gossypium hirsutum* L.). *BMC Genomics*, *15*(1), 397. https://doi.org/10.1186/1471-2164-15-397

Faulkner, K., Mithen, R., & Williamson, G. (1998). Selective increase of the potential anticarcinogen 4-methylsulphinylbutyl glucosinolate in broccoli. *Carcinogenesis*, *19*(4), 605–609. https://doi.org/10.1093/carcin/19.4.605

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., … Bray, F. (2013). *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]*. Lyon, France.

Fogarty, M. C., Hughes, C. M., Burke, G., Brown, J. C., & Davison, G. W. (2013). Acute and chronic watercress supplementation attenuates exercise-induced peripheral mononuclear cell DNA damage and lipid peroxidation. *The British Journal of Nutrition*, *109*(2), 293–301.

Gelli, M., Duo, Y., Konda, A. R., Zhang, C., Holding, D., & Dweikat, I. (2014). Identification of differentially expressed genes between sorghum genotypes with contrasting nitrogen stress tolerance by genome-wide transcriptional profiling. *BMC Genomics*, *15*(1), 179. https://doi.org/10.1186/1471-2164-15-179

Getahun, S., & Chung, F. (1999). Conversion of glucosinolates to isothiocyanates in humans after ingestion of cooked watercress. *Cancer Epidemiol Biomarkers Prev*, *8*, 447–451.

Giacoppo, S., Galuppo, M., Iori, R., De Nicola, G. R., Bramanti, P., & Mazzon, E. (2014). (RS)-glucoraphanin purified from Tuscan black kale and bioactivated with myrosinase enzyme protects against cerebral ischemia/reperfusion injury in rats. *Fitoterapia*, *99C*, 166–177. https://doi.org/10.1016/j.fitote.2014.09.016

Giacoppo, S., Galuppo, M., Montaut, S., Iori, R., Rollin, P., Bramanti, P., & Mazzon, E. (2015). An overview on neuroprotective effects of isothiocyanates for the treatment of neurodegenerative diseases. *Fitoterapia*, *106*. https://doi.org/10.1016/j.fitote.2015.08.001

Giallourou, N., Oruna-Concha, M. J., & Harbourne, N. (2016). Effects of domestic processing methods on the phytochemical content of watercress (*Nasturtium officinale*). *Food Chemistry, 212*, 411–419. https://doi.org/10.1016/j.foodchem.2016.05.190

Gil, V., & MacLeod, A. J. (1980). Degradation of glucosinolates of *Nasturtium officinale* seeds. *Phytochemistry*, *19*(8), 1657–1660. https://doi.org/10.1016/S0031-9422(00)83788-7

Gill, C. I. R., Haldar, S., Boyd, L. a, Bennett, R., Whiteford, J., Butler, M., … Rowland, I. R. (2007). Watercress supplementation in diet reduces lymphocyte DNA damage and alters blood antioxidant status in healthy adults. *The American Journal of Clinical Nutrition*, *85*(2), 504–510.

Gounden, D., Kisten, K., Moodley, R., Shaik, S., & Jonnalagadda, S. B. (2015). Impact of spiked concentrations of Cd, Pb, As and Zn in growth medium on elemental uptake of *Nasturtium officinale* (Watercress). *Journal of Environmental Science and Health. Part. B, Pesticides, Food Contaminants, and Agricultural Wastes*, 1–7. https://doi.org/10.1080/03601234.2015.1080477

Granado, F., Olmedilla, B., Herrero, C., Pérez-Sacristán, B., Blanco, I., & Blázquez, S. (2006). Bioavailability of carotenoids and tocopherols from broccoli: in vivo and in vitro assessment. *Experimental Biology and Medicine (Maywood, N.J.)*, *231*(11), 1733–1738.

Gross-Steinmeyer, K., Stapleton, P. L., Liu, F., Tracy, J. H., Bammler, T. K., Quigley, S. D., … Eaton, D. L. (2004). Phytochemical-induced changes in gene expression of carcinogen-metabolizing enzymes in cultured human primary hepatocytes. *Xenobiotica; the Fate of Foreign Compounds in Biological Systems*, *34*, 619–632. https://doi.org/10.1080/00498250412331285481

Grubb, C. D., & Abel, S. (2006). Glucosinolate metabolism and its control. *Trends in Plant Science*, *11*(2), 89–100. https://doi.org/10.1016/j.tplants.2005.12.006

Gupta, P., Adkins, C., Lockman, P., & Srivastava, S. K. (2013). Metastasis of breast tumor cells to brain is suppressed by phenethyl isothiocyanate in a novel in vivo metastasis model. *PloS One*, *8*(6), e67278. https://doi.org/10.1371/journal.pone.0067278

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*, 1494–1512.

Hanschen, F. S., Platz, S., Mewis, I., Schreiner, M., Rohn, S., & Kroh, L. W. (2012). Thermally induced degradation of sulfur-containing aliphatic glucosinolates in broccoli sprouts (*Brassica oleracea* var. *italica*) and model systems. *Journal of Agricultural and Food Chemistry*, *60*(9), 2231–2241. https://doi.org/10.1021/jf204830p

Harper, A. L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., … Bancroft, I. (2012). Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nature Biotechnology*, *30*(8), 798–802. https://doi.org/10.1038/nbt.2302

Hasegawa, T., Nishino, H., & Iwashima, A. (1993). Isothiocyanates inhibit cell cycle progression of HeLa cells at G2/M phase. *Anti-Cancer Drugs*, *4*(2), 273–279.

Hecht, S., Chung, F., Richie, J. J., Akerkar, S., Borukhova, A., Skowronski, L., & Carmella, S. (1995). Effects of watercress consumption on metabolism of a tobacco-specific lung carcinogen in smokers. *Cancer Epidemiol. Biomarkers Prev.*, *4*(8), 877–884.

Heidel, A. J., Clauss, M. J., Kroymann, J., Savolainen, O., & Mitchell-Olds, T. (2006). Natural variation in MAM within and between populations of *Arabidopsis lyrata* determines glucosinolate phenotype. *Genetics*, *173*(3), 1629–36. https://doi.org/10.1534/genetics.106.056986

Heiss, E., Herhaus, C., Klimo, K., Bartsch, H., & Gerhäuser, C. (2001). Nuclear factor kappa B is a molecular target for sulforaphane-mediated anti-inflammatory mechanisms. *The Journal of Biological Chemistry*, *276*, 32008–32015. https://doi.org/10.1074/jbc.M104794200

Hennig, K., Verkerk, R., van Boekel, M. A. J. S., Dekker, M., & Bonnema, G. (2014). Food science meets plant science: A case study on improved nutritional quality by breeding for glucosinolate retention during food processing. *Trends in Food Science & Technology*, *35*(1), 61–68. https://doi.org/10.1016/j.tifs.2013.10.006

Higdon, J. V, Delage, B., Williams, D. E., & Dashwood, R. H. (2007). Cruciferous vegetables and human cancer risk: epidemiologic evidence and mechanistic basis. *Pharmacological Research : The Official Journal of the Italian Pharmacological Society*, *55*(3), 224–236. https://doi.org/10.1016/j.phrs.2007.01.009

Howard, A. H. W., & Lyon, A. G. (1952). *Nasturtium officinale* R. Br. (Rorippa Nasturtium-Aquaticum (L.) Hayek). *Journal of Ecology*, *40*(1), 228–245.

Hu, R., Kim, B. R., Chen, C., Hebbar, V., & Kong, A.-N. T. (2003). The roles of JNK and apoptotic signaling pathways in PEITC-mediated responses in human HT-29 colon adenocarcinoma cells. *Carcinogenesis, 24*(8), 1361–1367. https://doi.org/10.1093/carcin/bgg092

Huang, S., Hsu, M., Hsu, S., Yang, J., Huang, W., Huang, A., … Chung, J.-G. (2014). Phenethyl isothiocyanate triggers apoptosis in human malignant melanoma A375.S2 cells through reactive oxygen species and the mitochondria-dependent pathways. *Human & Experimental Toxicology*, *33*(3), 270–283. https://doi.org/10.1177/0960327113491508

Hung, H.-C., Joshipura, K. J., Jiang, R., Hu, F. B., Hunter, D., Smith-Warner, S. A., … Willett, W. C. (2004). Fruit and vegetable intake and risk of major chronic disease. *Journal of the National Cancer Institute, 96*(21), 1577–1584. https://doi.org/10.1093/jnci/djh296

Ishida, M., Hara, M., Fukino, N., Kakizaki, T., & Morimitsu, Y. (2014). Glucosinolate metabolism, functionality and breeding for the improvement of Brassicaceae vegetables. *Breeding Science, 64*(1), 48–59. https://doi.org/10.1270/jsbbs.64.48

Islam, E. ul, Yang, X., He, Z., & Mahmood, Q. (2007). Assessing potential dietary toxicity of heavy metals in selected vegetables and food crops. *Journal of Zhejiang University. Science. B, 8*(1), 1–13. https://doi.org/10.1631/jzus.2007.B0001

Jahangir, M., Kim, H. K., Choi, Y. H., & Verpoorte, R. (2009). Health-affecting compounds in Brassicaceae. *Comprehensive Reviews in Food Science and Food Safety*, *8*, 31–43. https://doi.org/10.1111/j.1541-4337.2008.00065.x

Jeelani, S. M., Rani, S., Kumar, S., Kumari, S., & Gupta, R. C. (2013). Cytological studies of Brassicaceae Burn. (*Cruciferae Juss*.) from Western Himalayas. *Cytology and Genetics*, *47*(1), 20–28. https://doi.org/10.3103/S0095452713010076

Jeennor, S., & Volkaert, H. (2013). Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (*Elaeis guineensis* Jacq.). *Tree Genetics & Genomes*, *10*(1), 1–14. https://doi.org/10.1007/s11295-013-0655-3

Jiang, C., Shi, J., Li, R., Long, Y., Wang, H., Li, D., … Meng, J. (2014). Quantitative trait loci that control the oil content variation of rapeseed (*Brassica napus* L.). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *127*, 957–68. https://doi.org/10.1007/s00122-014-2271-5

Kaoulla, N., MacLeod, A. J., & Gil, V. (1980). Investigation of *Brassica oleracea* and *Nasturtium officinale* seeds for the presence of epithiospecifier protein. *Phytochemistry*, *19*(6), 1053–1056. https://doi.org/10.1016/0031-9422(80)83055-X

Karin, M. (2006). Nuclear factor-kappaB in cancer development and progression. *Nature*, *441*, 431–436. https://doi.org/10.1038/nature04870

Kaur, C., & Kapoor, H. C. (2008). Antioxidants in fruits and vegetables - the millennium's health. *International Journal of Food Science & Technology*, *36*(7), 703–725. https://doi.org/10.1111/j.1365-2621.2001.00513.x

Kopsell, D. a, Barickman, T. C., Sams, C. E., & McElroy, J. S. (2007). Influence of nitrogen and sulfur on biomass production and carotenoid and glucosinolate concentrations in watercress (*Nasturtium officinale* R. Br.). *Journal of Agricultural and Food Chemistry*, *55*(26), 10628–10634.

Krishnakumar, V., Contrino, S., Cheng, C.-Y., Belyaeva, I., Ferlanti, E. S., Miller, J. R., … Chan, A. P. (2016). ThaleMine: A warehouse for Arabidopsis data integration and discovery. *Plant and Cell Physiology*, *58*(1), pcw200. https://doi.org/10.1093/pcp/pcw200

Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M., … Town, C. D. (2015). Araport: the Arabidopsis Information Portal. *Nucleic Acids Research*, *43*(D1), D1003–D1009. https://doi.org/10.1093/nar/gku1200

Kushad, M., Brown, A., Kurilich, A., Juvik, J., Klein, B., Wallig, M., & Jeffery, E. (1999). Variation of Glucosinolates in Vegetable Crops of *Brassica oleracea*. https://doi.org/10.1021/JF980985S

Kuzina, V., Nielsen, J. K., Augustin, J. M., Torp, A. M., Bak, S., & Andersen, S. B. (2011). Barbarea vulgaris linkage map and quantitative trait loci for saponins, glucosinolates, hairiness and resistance to the herbivore *Phyllotreta nemorum*. *Phytochemistry*, *72*(2–3), 188–198. https://doi.org/10.1016/j.phytochem.2010.11.007

Lai, K.-C., Hsu, S.-C., Kuo, C.-L., Ip, S.-W., Yang, J.-S., Hsu, Y.-M., … Chung, J.-G. (2010). Phenethyl isothiocyanate inhibited tumor migration and invasion via suppressing multiple signal transduction pathways in human colon cancer HT29 cells. *Journal of Agricultural and Food Chemistry*, *58*(20), 11148–11155.

Langridge, P., & Fleury, D. (2011). Making the most of 'omics' for crop breeding. *Trends in Biotechnology*, *29*(1), 33–40.

Launders, N., Byrne, L., Adams, N., Glen, K., Jenkins, C., Tubin-Delic, D., … Morgan, D. (2013). Outbreak of Shiga toxin-producing E. coli O157 associated with consumption of watercress, United Kingdom, August to September 2013. *Euro Surveillance : Bulletin Européen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *18*(44), 1–5.

Lee, Y.-S., Ku, K.-M., Becker, T. M., & Juvik, J. A. (2017). Chemopreventive glucosinolate accumulation in various broccoli and collard tissues: Microfluidic-based targeted transcriptomics for by-product valorization. *PLOS ONE*, *12*(9), e0185112. https://doi.org/10.1371/journal.pone.0185112

Leonforte, A., Sudheesh, S., Cogan, N. O., Salisbury, P. A., Nicolas, M. E., Materne, M., … Kaur, S. (2013). SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity tolerance in field pea (*Pisum sativum* L.). *BMC Plant Biology*, *13*(1), 161. https://doi.org/10.1186/1471-2229-13-161

Li, Q., Zhan, M., Chen, W., Zhao, B., Yang, K., Yang, J., … Wang, J. (2016). Phenylethyl isothiocyanate reverses cisplatin resistance in biliary tract cancer cells via glutathionylation-dependent degradation of Mcl-1. *Oncotarget*. https://doi.org/10.18632/oncotarget.7171

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A. P., … Paterson, A. H. (2014). The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*, *5*, 3930. https://doi.org/10.1038/ncomms4930

Lobo, V., Patil, A., Phatak, A., & Chandra, N. (2010). Free radicals, antioxidants and functional foods: Impact on human health. *Pharmacognosy Reviews*, *4*(8), 118–26. https://doi.org/10.4103/0973-7847.70902

London, S. J., Yuan, J. M., Chung, F. L., Gao, Y. T., Coetzee, G. A., Ross, R. K., & Yu, M. C. (2000). Isothiocyanates, glutathione S-transferase M1 and T1 polymorphisms, and lung-cancer risk: a prospective study of men in Shanghai, China. *Lancet*, *356*(9231), 724–729. https://doi.org/10.1016/S0140-6736(00)02631-3

Losso, J. N., & Truax, R. E. (2009). Comparative inhibitory activities of sulforaphane and phenethyl isothiocyanate against leukemia resistant CEM/C2 cancer cells. *Journal of Functional Foods*, *1*(2), 229–235. https://doi.org/10.1016/j.jff.2009.02.003

Lou, P., Zhao, J., He, H., Hanhart, C., Del Carpio, D. P., Verkerk, R., … Bonnema, G. (2008). Quantitative trait loci for glucosinolate accumulation in *Brassica rapa* leaves. *The New Phytologist*, *179*(4), 1017–1032. https://doi.org/10.1111/j.1469-8137.2008.02530.x

Manchali, S., Chidambara Murthy, K. N., & Patil, B. S. (2012). Crucial facts about health benefits of popular cruciferous vegetables. *Journal of Functional Foods*, *4*(1), 94–106. https://doi.org/10.1016/j.jff.2011.08.004

Manton, I. (1935). The cytological history of Watercress (*Nasturtium officinale* R. Br.). *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, *69*(1), 132–157.

Manton, I., & Howard, H. W. (1946). Autopolyploid and allopolyploid watercress with the description of a new species. *Annals of Botany*, *10*, 1–13.

Marathi, B., Guleria, S., Mohapatra, T., Parsad, R., Mariappan, N., Kurungara, V. K., … Singh, A. K. (2012). QTL analysis of novel genomic regions associated with yield and yield related traits in new plant type based recombinant inbred lines of rice (*Oryza sativa* L.). *BMC Plant Biology*, *12*(1), 137. https://doi.org/10.1186/1471-2229-12-137

Martin, C. (2013). The interface between plant metabolic engineering and human health. *Current Opinion in Biotechnology*, *24*(2), 344–353. https://doi.org/10.1016/J.COPBIO.2012.11.005

Martin, C., & Li, J. (2017). Medicine is not health care, food is health care: plant metabolic engineering, diet and human health. *New Phytologist*, *216*(3), 699–719. https://doi.org/10.1111/nph.14730

Martin, C., Zhang, Y., Tonelli, C., & Petroni, K. (2013). Plants, diet, and health. *Annual Review of Plant Biology*, *64*, 19–46.

Martínez-Ballesta, M. del C., Moreno, D. A., & Carvajal, M. (2013). The physiological importance of glucosinolates on plant response to abiotic stress in Brassica. *International Journal of Molecular Sciences*, *14*(6). https://doi.org/10.3390/ijms140611607

Martínez-Sánchez, A., Gil-Izquierdo, A., Gil, M. I., & Ferreres, F. (2008). A comparative study of flavonoid compounds, vitamin C, and antioxidant properties of baby leaf Brassicaceae species. *Journal of Agricultural and Food Chemistry*, *56*(7), 2330–2340.

Matusheski, N. V, Juvik, J. A., & Jeffery, E. H. (2004). Heating decreases epithiospecifier protein activity and increases sulforaphane formation in broccoli. *Phytochemistry*, *65*(9), 1273–1281. https://doi.org/10.1016/j.phytochem.2004.04.013

Mauricio, R. (2001). Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nature Reviews Genetics*, *2*(5), 370–381. https://doi.org/10.1038/35072085

Minarini, A., Milelli, A., Fimognari, C., Simoni, E., Turrini, E., & Tumiatti, V. (2014). Exploring the effects of isothiocyanates on chemotherapeutic drugs. *Expert Opinion on*

Drug Metabolism & Toxicology, *10*(1), 25–38.

Mithen, R. F., Dekker, M., Verkerk, R., Rabot, S., & Johnson, I. T. (2000). The nutritional significance, biosynthesis and bioavailability of glucosinolates in human foods. *Journal of the Science of Food and Agriculture*, *80*(7), 967–984.

Mithen, R., Faulkner, K., Magrath, R., Rose, P., Williamson, G., & Marquez, J. (2003). Development of isothiocyanate-enriched broccoli, and its enhanced ability to induce phase 2 detoxification enzymes in mammalian cells. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *106*(4), 727–34.

Moon, J.-K., & Shibamoto, T. (2009). Antioxidant Assays for Plant and Food Components. *Journal of Agricultural and Food Chemistry*, *57*(5), 1655–1666. https://doi.org/10.1021/jf803537k

Moose, S. P., & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology*, *147*(3), 969–77. https://doi.org/10.1104/pp.108.118232

Morozowska, M., Czarna, A., & Jędrzejczyk, I. (2010). Estimation of nuclear DNA content in Nasturtium R. Br. by flow cytometry. *Aquatic Botany*, *93*(4), 250–253.

Myles, S. (2013). Improving fruit and wine: what does genomics have to offer? *Trends in Genetics : TIG*, *29*(4), 190–196. https://doi.org/10.1016/j.tig.2013.01.006

Nakajima, M., Yoshida, R., Shimada, N., Yamazaki, H., & Yokoi, T. (2001). Inhibition and inactivation of human cytochrome P450 isoforms by phenethyl isothiocyanate. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *29*, 1110–1113.

Natural England. (2009). *Watercress growing and its environmental impacts on chalk rivers in England (NECR027)*. *www.naturalengland.org.uk*.

Newman, R. M., Hanscom, Z., & Kerfoot, W. C. (1992). The watercress glucosinolate-myrosinase system: a feeding deterrent to caddisflies, snails and amphipods. *Oecologia*, *92*, 1–7.

O'Rourke, J. A., Bolon, Y.-T., Bucciarelli, B., & Vance, C. P. (2014). Legume genomics: understanding biology through DNA and RNA sequencing. *Annals of Botany*, *113*(7), 1107–1120. https://doi.org/10.1093/aob/mcu072

O'Rourke, J. A., Yang, S. S., Miller, S. S., Bucciarelli, B., Liu, J., Rydeen, A., … Vance, C. P. (2013). An RNA-Seq transcriptome analysis of orthophosphate-deficient white lupin reveals novel insights into phosphorus acclimation in plants. *Plant Physiology*, *161*(2), 705–724. https://doi.org/10.1104/pp.112.209254

Obara, M., Ishimaru, T., Abiko, T., Fujita, D., Kobayashi, N., Yanagihara, S., & Fukuta, Y. (2014). Identification and characterization of quantitative trait loci for root elongation by using introgression lines with genetic background of Indica-type rice variety IR64. *Plant Biotechnology Reports*, *8*(3), 267–277. https://doi.org/10.1007/s11816-014-0320-9

Oyebode, O., Gordon-Dseagu, V., Walker, A., & Mindell, J. S. (2014). Fruit and vegetable consumption and all-cause, cancer and CVD mortality: analysis of Health Survey for England data. *Journal of Epidemiology & Community Health*, 1–7. https://doi.org/10.1136/jech-2013-203500

Paine, J. A., Shipton, C. A., Chaggar, S., Howells, R. M., Kennedy, M. J., Vernon, G., … Drake, R. (2005). Improving the nutritional value of Golden Rice through increased pro-vitamin A content. *Nature Biotechnology*, *23*(4), 482–487. https://doi.org/10.1038/nbt1082

Palaniswamy, U. R., & McAvoy, R. J. (2001). Watercress: A salad crop with chemopreventive potential. *HortTechnology*, *11*(4), 622–626.

Palaniswamy, U. R., McAvoy, R. J., Bible, B. B., & Stuart, J. D. (2003). Ontogenic variations of ascorbic acid and phenethyl isothiocyanate concentrations in watercress (*Nasturtium officinale* R.Br.) leaves. *Journal of Agricultural and Food Chemistry*, *51*(18), 5504–5509.

Pan, Q., Xu, Y., Li, K., Peng, Y., Zhan, W., Li, W., … Yan, J. (2017). The Genetic Basis of Plant Architecture in 10 Maize Recombinant Inbred Line Populations. *Plant Physiology*, *175*(2), 858–873. https://doi.org/10.1104/pp.17.00709

Patil, B. S., Crosby, K., Byrne, D., & Hirschi, K. (2014). The intersection of plant breeding, human health, and nutritional security: Lessons learned and future perspectives. *HortScience*, *49*(2), 116–127.

Payne, A. C. (2011). *Harnessing the Genetic Diversity of Watercess (Rorippa nasturtium - aquaticum) for Improved Morphology and Anti- cancer Benefits: Underpinning Data for Molecular Breeding [doctoral thesis]*. Southampton: University of Southampton.

Payne, A. C., Clarkson, G. J. J., Rothwell, S., & Taylor, G. (2015). Diversity in global gene expression and morphology across a watercress (*Nasturtium officinale* R. Br.) germplasm collection: first steps to breeding. *Horticulture Research*, *2*, 15029. https://doi.org/10.1038/hortres.2015.29

Payne, A. C., Mazzer, A., Clarkson, G. J. J., & Taylor, G. (2013). Antioxidant assays - consistent findings from FRAP and ORAC reveal a negative impact of organic cultivation on antioxidant potential in spinach but not watercress or rocket leaves. *Food Science & Nutrition*, *1*(6), 439–44.

Pereira, L., Silva, P., Duarte, M., Rodrigues, L., Duarte, C., Albuquerque, C., & Serra, A. (2017). Targeting colorectal cancer proliferation, stemness and metastatic potential using Brassicaceae extracts enriched in isothiocyanates: a 3D cell model-based study. *Nutrients*, *9*(4), 368. https://doi.org/10.3390/nu9040368

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Podsędek, A. (2007). Natural antioxidants and antioxidant capacity of Brassica vegetables: A review. *LWT - Food Science and Technology*, *40*(1), 1–11. https://doi.org/10.1016/j.lwt.2005.07.023

Pootakham, W., Jomchai, N., Ruang-areerate, P., Shearman, J. R., Sonthirod, C., Sangsrakru, D., … Tangphatsornruang, S. (2015). Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics*, *105*(5), 288–295. https://doi.org/10.1016/j.ygeno.2015.02.002

Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., … Tangphatsornruang, S. (2015). Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). *Frontiers in Plant Science*, *6*, 367. https://doi.org/10.3389/fpls.2015.00367

Potter, M. J., Davies, K., & Rathjen, A. J. (1998). Suppressive impact of glucosinolates in Brassica vegetative tissues on root lesion nematode *Pratylenchus neglectus*. *Journal of Chemical Ecology*, *24*(1), 67–80. https://doi.org/10.1023/A:1022336812240

Prawan, A., Saw, C. L. L., Khor, T. O., Keum, Y.-S., Yu, S., Hu, L., & Kong, A.-N. (2009). Anti-NF-kappaB and anti-inflammatory activities of synthetic isothiocyanates: effect of chemical structures and cellular signaling. *Chemico-Biological Interactions*, *179*(2–3), 202–211. https://doi.org/10.1016/j.cbi.2008.12.014

Rose, P., Faulkner, K., Williamson, G., & Mithen, R. (2000). 7-Methylsulfinylheptyl and 8-methylsulfinyloctyl isothiocyanates from watercress are potent inducers of phase II enzymes. *Carcinogenesis*, *21*(11), 1983–1988.

Rose, P., Huang, Q., Ong, C. N., & Whiteman, M. (2005). Broccoli and watercress suppress matrix metalloproteinase-9 activity and invasiveness of human MDA-MB-231 breast cancer cells. *Toxicology and Applied Pharmacology*, *209*(2), 105–113.

Rothwell, S. D., & Robinson, L. W. (1986). Cold acclimation potential of watercress in relation to growing season and nutrient status. *Journal of Horticultural Science*, *61*(3), 373–378.

Rowland, L. J., Alkharouf, N., Darwish, O., Ogden, E. L., Polashock, J. J., Bassil, N. V, & Main, D. (2012). Generation and analysis of blueberry transcriptome sequences from leaves, developing fruit, and flower buds from cold acclimation through deacclimation. *BMC Plant Biology*, *12*(1), 46. https://doi.org/10.1186/1471-2229-12-46

Sadeghi, H., Mostafazadeh, M., Sadeghi, H., Naderian, M., Barmak, M. J., Talebianpoor, M. S., & Mehraban, F. (2013). In vivo anti-inflammatory properties of aerial parts of *Nasturtium officinale*. *Pharmaceutical Biology*, (2008), 1–6. https://doi.org/10.3109/13880209.2013.821138

Santos, J., Oliveira, M. B. P. P., Ibáñez, E., & Herrero, M. (2014). Phenolic profile evolution of different ready-to-eat baby-leaf vegetables during storage. *Journal of Chromatography. A*, *1327*, 118–131.

Sarikamis, G., Marquez, J., Maccormack, R., Bennett, R. N., Roberts, J., & Mithen, R. (2006). High glucosinolate broccoli: a delivery system for sulforaphane. *Molecular Breeding*, *18*(3), 219–228. https://doi.org/10.1007/s11032-006-9029-y

Savelkoul, P. H., Aarts, H. J., de Haas, J., Dijkshoorn, L., Duim, B., Otsen, M., … Lenstra, J. A. (1999). Amplified-fragment length polymorphism analysis: the state of an art. *Journal of Clinical Microbiology*, *37*(10), 3083–91.

Seo, M.-S., & Kim, J. (2017). Understanding of MYB transcription factors involved in glucosinolate biosynthesis in Brassicaceae. *Molecules*, *22*(9), 1549. https://doi.org/10.3390/molecules22091549

Shapiro, T. A., Fahey, J. W., Wade, K. L., Stephenson, K. K., & Talalay, P. (2001). Chemoprotective glucosinolates and isothiocyanates of broccoli sprouts: Metabolism and excretion in humans. *Cancer Epidemiol. Biomarkers Prev.*, *10*(5), 501–508.

Sheridan, G. E. C., Claxton, J. R., Clarkson, J. M., & Blakesley, D. (2001). Genetic diversity within commercial populations of watercress (Rorippa nasturtium-aquaticum), and between allied Brassicaceae inferred from RAPD-PCR. *Euphytica*, *122*, 319–325.

Sotelo, T., Soengas, P., Velasco, P., Rodríguez, V. M., & Cartea, M. E. (2014). Identification of metabolic QTLs and candidate genes for glucosinolate synthesis in Brassica oleracea leaves, seeds and flower buds. *PloS One*, *9*(3), e91428. https://doi.org/10.1371/journal.pone.0091428

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., … Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, *36*(Database issue), D1009-14. https://doi.org/10.1093/nar/gkm965

Syed Alwi, S. S., Cavell, B. E., Telang, U., Morris, M. E., Parry, B. M., & Packham, G. (2010). In vivo modulation of 4E binding protein 1 (4E-BP1) phosphorylation by watercress: a pilot study. *The British Journal of Nutrition*, *104*(9), 1288–1296.

Tan, C., Han, Z., Yu, H., Zhan, W., Xie, W., Chen, X., … Xing, Y. (2013). QTL scanning for rice yield using a whole genome SNP array. *Journal of Genetics and Genomics*, *40*(12), 629–638.

Telang, U., Brazeau, D. A., & Morris, M. E. (2009). Comparison of the effects of phenethyl isothiocyanate and sulforaphane on gene expression in breast cancer and normal mammary epithelial cells. *Experimental Biology and Medicine (Maywood, N.J.)*, *234*(3), 287–95. https://doi.org/10.3181/0808-RM-241

Thakur, V. S., Deb, G., Babcook, M. a, & Gupta, S. (2014). Plant phytochemicals as epigenetic modulators: role in cancer chemoprevention. *The AAPS Journal*, *16*(1), 151–163. https://doi.org/10.1208/s12248-013-9548-5

Thapliyal, R., & Maru, G. . (2001). Inhibition of cytochrome P450 isozymes by curcumins in vitro and in vivo. *Food and Chemical Toxicology*, *39*(6), 541–547. https://doi.org/10.1016/S0278-6915(00)00165-4

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*(6814), 796–815. https://doi.org/10.1038/35048692

Tierens, K. F. M.-J. (2001). Study of the role of antimicrobial glucosinolate-derived isothiocyanates in resistance of Arabidopsis to microbial pathogens. *PLANT PHYSIOLOGY*, *125*(4), 1688–1699. https://doi.org/10.1104/pp.125.4.1688

Toroser, D., Thormann, C. E., Osborn, T. C., & Mithen, R. (1995). RFLP mapping of quantitative trait loci controlling seed aliphatic-glucosinolate content in oilseed rape

(*Brassica napus* L). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *91*(5), 802–808. https://doi.org/10.1007/BF00220963

Traka, M. H., Gasper, A. V, Smith, J. A., Hawkey, C. J., Bao, Y., & Mithen, R. F. (2005). Transcriptome analysis of human colon Caco-2 cells exposed to sulforaphane. *The Journal of Nutrition*, *135*(8), 1865–1872.

Traka, M. H., & Mithen, R. (2008). Glucosinolates, isothiocyanates and human health. *Phytochemistry Reviews*, *8*(1), 269–282.

Traka, M. H., & Mithen, R. F. (2011). Plant science and human nutrition: challenges in assessing health-promoting properties of phytochemicals. *The Plant Cell*, *23*, 2483–2497. https://doi.org/10.1105/tpc.111.087916

Traka, M. H., Saha, S., Huseby, S., Kopriva, S., Walley, P. G., Barker, G. C., … Mithen, R. F. (2013). Genetic regulation of glucoraphanin accumulation in Beneforté broccoli. *The New Phytologist*, *198*(4), 1085–1095.

Tusskorn, O., Senggunprai, L., Prawan, A., Kukongviriyapan, U., & Kukongviriyapan, V. (2013). Phenethyl isothiocyanate induces calcium mobilization and mitochondrial cell death pathway in cholangiocarcinoma KKU-M214 cells. *BMC Cancer*, *13*, 571. https://doi.org/10.1186/1471-2407-13-571

Uzunova, M., Ecke, W., Weissleder, K., & Robbelen, G. (1995). Mapping the genome of rapeseed (*Brassica napus* L.). I. Construction of an RFLP linkage map and localization of QTLs for seed glucosinolate content. *Theoretical and Applied Genetics*, *90*(2), 194–204.

Van Duyn, M. A., & Pivonka, E. (2000). Overview of the health benefits of fruit and vegetable consumption for the dietetics professional: selected literature. *Journal of the American Dietetic Association*, *100*(12), 1511–1521. https://doi.org/10.1016/S0002-8223(00)00420-X

Van Lieshout, E. M. M., Peters, W. H. M., & Jansen, J. B. M. J. (1996). Effect of oltipraz, a-tocopherol, ^-carotene and phenethylisothiocyanate on rat oesophageal, gastric, colonic and hepatic glutathione, glutathione S-transferase and peroxidase. *Carcinogenesis*, *17*(7), 1439–1445.

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. a. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, *27*(9), 522–530.

Velasco, P., Cartea, M. E., Gonzalez, C., Vilar, M., & Ordas, A. (2007). Factors affecting the glucosinolate content of kale (*Brassica oleracea* acephala group). *Journal of Agricultural and Food Chemistry*, *55*(3), 955–962.

Verhoeven, D., Goldbohm, R., van Poppel, G., Verhagen, H., & van den Brandt, P. (1996). Epidemiological studies on brassica vegetables and cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, *5*(9), 733–748.

Wagner, A. E., Terschluesen, A. M., & Rimbach, G. (2013). Health promoting effects of brassica-derived phytochemicals: from chemopreventive and anti-inflammatory activities to epigenetic regulation. *Oxidative Medicine and Cellular Longevity*, *2013*, 964539.

Wang, L. G., Liu, X. M., Fang, Y., Dai, W., Chiao, F. B., Puccio, G. M., … Chiao, J. W. (2008). De-repression of the p21 promoter in prostate cancer cells by an isothiocyanate via inhibition of HDACs and c-Myc. *International Journal of Oncology*, *33*(2), 375–380.

Wang, X., Ouyang, Y., Liu, J., Zhu, M., Zhao, G., Bao, W., & Hu, F. B. (2014). Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. *BMJ*, *349*(jul29 3), g4490–g4490. https://doi.org/10.1136/bmj.g4490

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., … Zhang, Z. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*. https://doi.org/10.1038/ng.919

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Wei, L., Xiao, M., Hayward, A., & Fu, D. (2013). Applications and challenges of next-generation sequencing in Brassica species. *Planta*, *238*(6), 1005–1024. https://doi.org/10.1007/s00425-013-1961-6

Wilson, A. E., Bergaentzlé, M., Bindler, F., Marchioni, E., Lintz, A., & Ennahar, S. (2013). *In vitro* efficacies of various isothiocyanates from cruciferous vegetables as antimicrobial agents against foodborne pathogens and spoilage bacteria. *Food Control*, *30*(1), 318–324. https://doi.org/10.1016/j.foodcont.2012.07.031

Wu, C.-L., Huang, A.-C., Yang, J.-S., Liao, C.-L., Lu, H.-F., Chou, S.-T., … Chung, J.-G. (2011). Benzyl isothiocyanate (BITC) and phenethyl isothiocyanate (PEITC)-mediated generation of reactive oxygen species causes cell cycle arrest and induces apoptosis via activation of caspase-3, mitochondria dysfunction and nitric oxide (NO) in human osteogenic. *Journal of Orthopaedic Research : Official Publication of the Orthopaedic Research Society*, *29*(8), 1199–209. https://doi.org/10.1002/jor.21350

Xu, Y., & Crouch, J. H. (2008). Marker-assisted selection in plant breeding: From publications to practice. *Crop Science*, *48*(2), 391. https://doi.org/10.2135/cropsci2007.04.0191

Xue, J., Lenman, M., Falk, A., & Rask, L. (1992). The glucosinolate-degrading enzyme myrosinase in Brassicaceae is encoded by a gene family. *Plant Molecular Biology*, *18*(2), 387–398. https://doi.org/10.1007/BF00034965

Yan, X., & Chen, S. (2007). Regulation of plant glucosinolate metabolism. *Planta*, *226*(6), 1343–1352.

Yang, M.-D., Lai, K.-C., Lai, T.-Y., Hsu, S.-C., Kuo, C.-L., Yu, C.-S., … Chung, J.-G. (2010). Phenethyl isothiocyanate inhibits migration and invasion of human gastric cancer AGS cells through suppressing MAPK and NF-kappaB signal pathways. *Anticancer Research*, *30*(6), 2135–2143.

Yuste-Lisbona, F. J., González, A. M., Capel, C., García-Alcázar, M., Capel, J., De Ron, A. M., … Lozano, R. (2014). Genetic variation underlying pod size and color traits of common bean depends on quantitative trait loci with epistatic effects. *Molecular Breeding*, *33*(4), 939–952. https://doi.org/10.1007/s11032-013-0008-9

Zargar, S. M., Raatz, B., Sonah, H., MuslimaNazir, Bhat, J. A., Dar, Z. A., … Rakwal, R. (2015). Recent advances in molecular marker techniques: Insight into QTL mapping, GWAS and genomic selection in plants. *Journal of Crop Science and Biotechnology*, *18*(5), 293–308. https://doi.org/10.1007/s12892-015-0037-5

Zeb, A. (2015). Phenolic profile and antioxidant potential of wild watercress (*Nasturtium officinale* L.). *SpringerPlus*, *4*, 714. https://doi.org/10.1186/s40064-015-1514-5

Zhang, X., Liu, T., Duan, M., Song, J., & Li, X. (2016). De novo transcriptome analysis of *Sinapis alba* in revealing the glucosinolate and phytochelatin pathways. *Frontiers in Plant Science*, *7*, 259. https://doi.org/10.3389/fpls.2016.00259

Zhang, Y., Huai, D., Yang, Q., Cheng, Y., Ma, M., Kliebenstein, D. J., & Zhou, Y. (2015). Overexpression of three glucosinolate biosynthesis genes in *Brassica napus* identifies enhanced resistance to *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLOS ONE*, *10*(10), e0140491. https://doi.org/10.1371/journal.pone.0140491

Chapter 2: Characterization of the watercress (*Nasturtium officinale* R. Br.; Brassicaceae) transcriptome using RNASeq and identification of candidate genes for important phytonutrient traits linked to human health

Nikol Voutsina[1], Adrienne C. Payne[1], Robert D. Hancock[2], Graham J. J. Clarkson[3], Steve D. Rothwell[3], Mark A. Chapman[1], Gail Taylor[1]

[1] Centre for Biological Sciences, University of Southampton, Southampton SO17 1BJ, UK
[2] Cell and Molecular Sciences, The James Hutton Institute, Dundee DD2 5DA, UK
[3] Vitacress Salads Ltd, Lower Link Farm, St Mary Bourne, Andover SP11 6DB, UK

This chapter has been published as:

Voutsina N, Payne AC, Hancock RD, Clarkson GJJ, Rothwell SD, Chapman MA & G Taylor (2016) Characterization of the watercress *(Nasturtium officinale* R. Br.; Brassicaceae) transcriptome using RNASeq and identification of candidate genes for important phytonutrient traits linked to human health. *BMC Genomics* 17:378, doi 10.1186/s12864-016-2704-4

Author Thesis Contributions: The plants described in this chapter were grown and phenotyped for antioxidant capacity by A. Payne for her PhD Thesis. N. Voutsina utilized this data to select the best samples and used the frozen samples from these plants to perform all further work detailed here.

2.1 Abstract

Consuming watercress is thought to provide health benefits as a consequence of its phytonutrient composition. However, for watercress there are currently limited genetic resources underpinning breeding efforts for either yield or phytonutritional traits. In this paper, we use RNASeq data from twelve watercress accessions to characterize the transcriptome, perform candidate gene mining and conduct differential expression analysis for two key phytonutritional traits: antioxidant (AO) capacity and glucosinolate (GLS) content.

The watercress transcriptome was assembled to 80,800 transcripts (48,732 unigenes); 71 % of which were annotated based on orthology to *Arabidopsis*. Differential expression analysis comparing watercress accessions with ‘high’ and ‘low’ AO and GLS resulted in 145 and 94 differentially expressed loci for AO capacity and GLS respectively. Differentially expressed loci between high and low AO watercress were significantly enriched for genes involved in plant defence and response to stimuli, in line with the observation that AO are involved in plant stress-response. Differential expression between the high and low GLS watercress identified links to GLS regulation and also novel transcripts warranting further investigation. Additionally, we successfully identified watercress orthologs for *Arabidopsis* phenylpropanoid, GLS and shikimate biosynthesis pathway genes, and compiled a catalogue of polymorphic markers for future applications.

Our work describes the first transcriptome of watercress and establishes the foundation for further molecular study by providing valuable resources, including sequence data, annotated transcripts, candidate genes and markers.

2.2 Introduction

Watercress, *Nasturtium officinale* R. Br. (Brassicaceae), is a perennial dicotyledonous herb usually found in close proximity to water (Howard & Lyon 1952). As a member of the Brassicaceae, it is related to several popular food and spice crops, such as broccoli, cabbage, kale, radish and mustard, as well as the model plant *Arabidopsis thaliana* (L.) Heynh. The consumption of Brassicaceae vegetables is suggested to benefit human health as a consequence of their phytochemical composition, which includes high concentrations of glucosinolates (GSL) (Verhoeven et al. 1996; Manchali et al. 2012; Wagner et al. 2013). In particular, watercress has been used as a medicinal and food crop for over 2000 years (Manton 1935). Over the past few decades, a growing number of studies suggest that watercress consumption supports health by providing chemopreventive, antioxidant and anti-inflammatory benefits. Specifically, several studies have shown that watercress extracts can act *in vitro* to combat the growth and metastasis of cancer cells (Rose et al. 2000; Rose et al. 2005; Lai et al. 2010; Syed Alwi et al. 2010; Cavell et al. 2011). The consumption of watercress by adults also limited exercise-induced DNA damage (Fogarty et al. 2013) and increased blood antioxidants (Hecht et al. 1995; Gill et al. 2007). Recently, it was ranked as the top “powerhouse fruit and vegetable” with the strongest link to decreased occurrence of chronic disease (Di Noia 2014), ranking highly because it contains an array of both essential nutrients as well as non-essential health-promoting phytochemicals.

Two pivotal traits contributing to the watercress phytonutritient profile are antioxidant (AO) capacity and GLS content. As plant-derived AOs are thought to be an important source of health benefits associated with vegetable and fruit consumption (Martin et al. 2013), maintaining or increasing AO capacity of food crops is the principal aim of several research and breeding programs (Prohens et al. 2007; Cantín et al. 2009; Kavitha et al. 2014; Vaz Patto et al. 2014). Several types of dietary AOs are derived from the phenylpropanoid pathway, such as phenolic acids and flavonoids (Pandey & Rizvi 2009) and this pathway has been well described in *Arabidopsis* (Fraser & Chapple 2011). Three studies have recently described phenolic compounds present in watercress. Santos et al. (2014) observed that the major phenolic group in watercress are the flavonols, primarily quercetin, kaempferol and isorhamnetin species. A second study, on baby-leaf watercress, identified chlorogenic acid, quercetin-3-O-rutinoside, caffeoyltartaric acid and isorhamnetin as the most abundant phenolic components (Aires et al. 2013). Finally, Martínez-Sánchez et al. (2008) demonstrated that watercress leaves contain almost double the amount of polyphenols found in other leafy Brassicaceae crops, namely mizuna, rocket and wild rocket.

GLS, which are secondary plant metabolites with anti- herbivory properties (Newman et al. 1992), are thought to be responsible for the health benefits and characteristic strong mustard flavour associated with several Brassicaceaes (Traka & Mithen 2008; Manchali et al. 2012). Upon injury of the plant tissue, GLS are hydrolysed by the enzyme myrosinase to nitriles, thiocyanates and isothiocyanates, the quantities of each dependent on reaction conditions (Bones & Rossiter 1996; Fahey et al. 2001). Isothiocyanates have been studied extensively and are thought to have chemopreventive properties (Traka & Mithen 2008; Wagner et al. 2013). In addition, evidence suggests that the use of these compounds in association with chemotherapy drugs could increase their effectiveness (Minarini et al. 2014). Thus, the GLS phenotype is an integral part of the nutritional profile in watercress, as well as contributing to the potent peppery flavour of the crop.

Despite its unique nutritional profile and its global market as a food crop, there is no watercress breeding programme and no genetic and genomic resources are available. Research to date has focused primarily on the biomedical implications of watercress consumption and little is known about the watercress crop as a source of germplasm for breeding and improvement. Particularly limited are the genetic resources available to inform industry and science in future improvement or preservation of these important nutritional traits in the crop. To date, selection for important agronomic traits, such as frost or disease resistance, has been conducted on a small scale by growers in-house and there no varieties specifically bred for commercial production (Rothwell & Robinson 1986; Palaniswamy & McAvoy 2001). In fact, little genetic variation appears to exist amongst commercial watercress (Sheridan et al. 2001). Recently, Payne et al. (2015) surveyed differences in morphology of above-ground characteristics in 25 accessions of watercress from the University of Southampton germplasm collection, which maintains germplasm from growers around the world. The research identified promising range in agronomic characters but limited accession specificity and suggested that breeding could lead to great improvements through selection and the development of varieties. High precision molecular breeding tools could make significant contributions to this crop, especially for the preservation and improvement of traits associated with the high nutritional profile and unique flavour of this crop in future breeding.

Next Generation Sequencing (NGS) technologies provide an opportunity for accelerated crop breeding, even for crops that are considered ‘specialist’ and for which there is no genetic and genomic underpinning knowledge (Varshney et al. 2009). RNA Sequencing (RNASeq), also known as Whole Transcriptome Shotgun Sequencing, is a method developed to generate a snap-shot of the expressed genome and expression levels within a tissue under a particular set of conditions (Wang et al. 2009). This tool can be applied to reveal differences in gene expression under varying environmental conditions, developmental stages, or between phenotypes.

In this study, we present the development of a set of genomic tools for watercress breeding. Specifically, the watercress transcriptome was sequenced using NGS-based Illumina paired-end reads and assembled using the software, Trinity. An annotated catalogue of watercress transcripts was created and differential expression (DE) analysis completed to investigate the genetic basis of two key watercress nutritional attributes: AO capacity and GLS content. Candidate gene mining was also conducted to identify watercress orthologs of known genes in the phenylpropanoid and GLS biosynthetic pathways, and a catalogue of polymorphic markers assembled.

2.3 Methods and Materials

2.3.1 Plant Material and Phenotyping:

Twenty five watercress accessions, from the University of Southampton germplasm collection, were grown side by side at a field site in Spetisbury (50°48’46.8”N, 2°08’47.9”W), Dorset U.K., under standard watercress commercial cultivation conditions, as described previously Payne et al. (2015). Specifically, watercress is traditionally grown in shallow gravel beds with flowing spring water. After seven weeks, the time at which the crop would typically be harvested for market, leaf and stem tissue was collected from all watercress accessions. Tissue was snap frozen in liquid nitrogen, ground and stored at -80°C until further use. The antioxidant (AO) capacity of each sample was evaluated using an adapted Ferric Reducing Ability of Plasma (FRAP) protocol (Benzie & Strain 1996), described by Payne et al. (2013). Sap was extracted using a QiaShredder homogenizer tube (Qiagen, www.qiagen.com) and spun at 13,000 for 5 min at 4 °C. Samples were plated in 96-well plate alongside a serial dilution of iron sulphate heptahydrate. FRAP reagent mix, containing acetate buffer, TPTZ (2,4,6-tripyrid-s-triazine/hydrochloric acid) and ferric chloride hexahydrate, was added and the plate read immediately on a spectrophotometer (Anthos Labtec Instruments) at 620nm. The FRAP assay utilizes the colour change which occurs during the reduction of ferric to ferrous ion to quantify the AO capacity of a sap sample (Benzie & Strain 1996).

Glucosinolates (GLS) were extracted from snap-frozen and ground tissue in 10 volumes of 70% methanol at 70°C. Sinigrin, a GLS not found in watercress (Agerbirk et al. 2014) was added as an internal standard at a concentration of 10 µg ml^{-1}. Samples were incubated at 70°C for 30 minutes with periodic mixing. The liquid phase was removed and centrifuged at 1° C at 16000 g for 5 minutes. Supernatants were transferred to amber vials and analysed by HPLC-MS. 10 µl of each extract was injected onto a Synergi Hydro-RP 150 x 2.0 mm column

(Phenomenex, Macclesfield, UK) using an Accela autosampler (Thermo Fisher Scientific, UK). The mobile phase comprised 0.1% (v/v) formic acid in water (solvent A) and 0.1% formic acid in methanol (solvent B) pumped at 200 µl min^{-1} using an Accela 600 pump. The mobile gradient comprised an isocratic phase of 100% solvent A for 4 min then a ramp to 20% B over the next 10 min which was held for a further 6 min. A second ramp increased solvent B to 50% over 5 min and was held for a further 10 min. Finally, solvent B was increased to 80% over 5 min and held for a further 2 min prior to requilibration at 100% before injection of the following sample. Column eluent was monitored using an Accela PDA and GLS were identified and quantified by ESI-MS and MS2 in negative ion mode using an LCQ fleet ion trap mass spectrometer according to Rochfort et al. (2008). The mass spectrometer was tuned against sinigrin using a sheath gas flow of 25, an auxiliary gas flow of 5, a spray voltage of 5 kV and a capillary temperature 275^{o}C.

Phenotype data from the above procedures was used to categorize ‘extreme’ samples for differential expression analysis and is shown in Table 2.1. The five samples with the highest and lowest AO capacity were selected for sequencing. From these, the samples with the three highest and three lowest concentrations of gluconasturtiin were used for differential expression analysis. Two control accessions were also sequenced but not used in gene expression analysis. The first, NAS080, is a Vitacress Salads Ltd commercially-active accession that is widely sold across the U.K. and grown in the U.K., U.S.A., Portugal and Australia. The second control accession was NAS065, an accession from the University of Southampton germplasm collection which is of breeding interest because it exhibits the desirable phenotypes of high phytonutrient content and dwarf size.

Table 2.1 Phenotypic data describing the watercress sequenced in this project. Antioxidant (AO) capacity was assessed using the FRAP antioxidant assay. Gluconasturtiin concentration, the primary glucosinolate (GLS) in watercress, was assessed using HPLC-MS. Concentration of gluconasturtiin was then quantified for this study as the ratio of the compounds peak area over the peak area of the internal standard (sinigrin). NA specifies that no data or classification is available for that sample.

Sample	**Antioxidant capacity***	**AO Group**	**Gluconasturtiin**	**GLS Group**
	(mmol Fe^{2+} equivalent/g fresh weight)		(Peak Area Ratio)	
NAS080	501	Control	11.1	Control
NAS081	837	High	7.4	Low
NAS057	808	High	15.9	High
NAS092	803	High	12.9	NA
NAS095	841	High	12.0	NA
NAS058	903	High	15.1	High
NAS061	373	Low	14.5	High
NAS068	185	Low	9.4	Low
NAS066	405	Low	11.4	NA
NAS070	271	Low	11.0	NA
NAS093	327	Low	7.0	Low
NAS065	NA	NA	NA	NA

*Antioxidant data modified from Payne et al. (2015)

2.3.2 RNA Extraction and Illumina Sequencing:

RNA was extracted with the RNeasy Mini kit (Qiagen, www.qiagen.com) and tested by NanoDrop (Thermo Scientific ND-1000) and Agilent 2200 TapeStation (Agilent Technologies, www.agilent.com) for concentration, purity and integrity. Samples were sent to the Wellcome Trust Centre for Human Genetics, where they were converted to cDNA, A-tailed and adapter-ligated. The 12 samples were individually barcoded, combined and then pair-end sequenced in one lane of an Illumina Hiseq2500 (Illumina, www.illumina.com) producing 100nt length reads. Initial quality checks were carried out using the standard Illumina pipeline.

2.3.3 Processing and *de novo* Assembly:

Sample barcodes and poor quality reads (Q <15) were removed using cutadapt v1.5. NAS080 was chosen to assemble the reference transcriptome for watercress because it is an important commercial line. We used *de novo* assembly software, Trinity, version 20140717 (Grabherr et al. 2011). Data was *in silico* normalised to limit copies of each k-mer to 30, increasing the efficiency of the assembly by reducing run time and memory requirements (Haas et al. 2013). The normalised reads were then assembled multiple times using various settings. As watercress is a tetraploid (Bleeker et al. 1999), four alleles could potentially be determined for each gene. Assemblies which permit greater numbers of differences per path could potentially collapse these paralogous genes or multiple alleles into one assembled component. We tested assemblies which allowed from 0 to 8 differences (SNPs) per path to expose such patterns in our data. Of these, we took forward the assembly with minimum k-mer coverage of 2, maximum gap allowed per path of 15 bases, and 2 differences allowed per path (see results).

The resulting assembly was then trimmed to filter out low count transcripts that are likely to be errors. The 12 sample libraries were mapped back to the reference transcriptome using RSEM in Trinity to determine FPKM (Fragments per kilobase of exon per million fragments mapped – a standardized value of expression) for each isoform of each gene. These data were then used to examine the effects of various trimming parameters. A trim with settings of minimum IsoPct (% expression of a transcript compared to other transcripts) of 1 % and a minimum FPKM of 1 was selected to be carried forward. The 12 individuals were then mapped back to the trimmed reference assembly, using RSEM, in order to examine gene expression variation between the individuals.

2.3.4 Classification of Transcripts and Candidate Gene Identification:

As watercress is closely related to *Arabidopsis* (Bailey et al. 2006; Beilstein et al. 2010), the trimmed transcriptome was in first instance annotated using the *Arabidopsis* genome which has been fully sequenced and well-described (Lin et al. 1999; Theologis et al. 2000). We used the software BioEdit Sequence Alignment Editor version 7.2.5 (Hall 1999) to perform a BLASTx peptide search against current *Arabidopsis* peptide sequences, available from the TAIR database (file name: TAIR10_pep_20101214.fas). A cut-off e-value of e^{-20} was applied. Locus identifiers were then used to retrieve GO terms using the GO Annotation tool on the TAIR website (http://www.arabidopsis.org/tools/bulk/go/index.jsp). The transcriptome was further annotated using BLASTx search via Trinotate (http://trinotate.sourceforge.net/) against the UniProtKB/SWISS-PROT database (The UniProt Consortium 2014) to annotate transcripts without a match and the UniProt ID mapping tool was used to retrieve gene identifiers and protein descriptions (http://www.uniprot.org/uploadlists/).

The mitochondrial and chloroplast genome of *Arabidopsis* was also compared to the transcripts in order to identify any transcripts that may have originated from non-nuclear DNA. Transcripts with a 95% or higher sequence match and minimum 100 bases hit length were identified and not included in further analyses or interpretation.

A literature search was conducted to compile a list of genes involved in the phenylpropanoid pathway and in GLS biosynthesis, two major pathways directly linked with the traits of interest. For the GLS pathway, the gene list compiled in Wang et al. (2011) for the watercress relatives *Arabidopsis* and *Brassica rapa* L. var. *silvestris* [Lam.] Briggs was used. In addition, we searched for matches to myrosinase (thioglucoside glucohydrolase), which is responsible for the conversion of GLS to the beneficial isothiocyanates upon consumption, in the annotated watercress transcriptome. The *Arabidopsis* gene sequences were retrieved from NCBI and then used in a BLASTn search of the watercress transcriptome. Orthology of best matches was further confirmed by a reciprocal BLAST of each watercress candidate sequence using the NCBI online BLAST tool (blast.ncbi.nlm.nih.gov/Blast.cgi).

2.3.5 Differential Expression Analysis:

Differential expression (DE) analysis was completed on standardized abundance estimates of transcripts for traits underpinning AO capacity and GLS content in Trinity, which utilises edgeR (Robinson et al. 2010). The five watercress samples with the highest and lowest AO capacity were used in DE analysis for the AO trait and the three highest and lowest GLS concentration were used for DE analysis for the GLS trait. After correction for false discovery due to multiple hypotheses testing, DE loci with FDR ≤ 0.05 are reported as significant in this study. Using AgriGO (0.05 significance with chi-squared test and Bonferroni correction), the GO terms of DE genes were compared to those of the reference transcriptome in order to identify over-represented GO categories. Fixation index (F_{ST}) was also calculated between groups, using ProSeq3 (Filatov 2009), to guide the identification of potential polymorphisms associated with each trait.

2.3.6 Genetic Relatedness and Polymorphic Marker Development:

Following RSEM alignment of reads to the reference transcriptome, .bam files were exported. SAMtools (H. Li et al. 2009) (settings: mpileup -q 3 -Q 20 -D –u), bcftools, vcfutils.pl (setting: -d 3) and seqtk were used to score polymorphisms relative to the reference transcriptome and create fasta files for polymorphism assessment. Polymorphisms were identified within ProSeq3. ProSeq3 was also used to calculate nucleotide diversity indices: π (Nei & Li 1979) and θ (Watterson 1975; Halushka et al. 1999). In addition, the script misa.pl (http://pgrc.ipk-gatersleben.de/misa/) was applied to search for SSRs in the reference transcriptome, with the minimum repeat number of 8, 6, and 4 di-, tri- and tetranucleotides, respectively. A phylogenetic analysis of the accessions was completed using phyml (http://www.atg-montpellier.tr/phyml/) and based on 7,000 SNPs.

2.4 Results

2.4.1 Sequencing and *de novo* Assembly:

Watercress accessions from the University of Southampton germplasm collection were grown under standard commercial conditions in the U.K, as described previously (Payne et al. 2015). Tissue samples were collected at the time of commercial harvest and evaluated for antioxidant (AO) capacity and glucosinolate (GLS) content (Table 2.1). RNA was extracted from the highest and lowest five samples, as well as two controls of commercial significance. The resulting twelve watercress accessions were sequenced on an Illumina Hiseq2500 generating a total of 323,827,923 paired-end fragments, thus producing an average of 26,985,660 reads per sample: documented in detail in Table 2.2. Reads have been deposited in the National Center for Biotechnology Information Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) under SRA accession number SRP058520 and BioProject: PRJNA284126. For the commercial watercress accession chosen for the reference assembly (NAS080), 28,128,352 paired-end reads were sequenced. Following quality check and normalization of data, the initial transcriptome was *de novo* assembled using Trinity (Grabherr et al. 2011) and contained 87,844 transcripts, which correspond to 48,732 components or “unigenes” (further statistics in Table 2.3). These numbers did not change greatly when Trinity assembly settings were altered to allow reads with more single nucleotide polymorphisms (SNPs) to be assembled together (See Table 2.3). A reduction in the allowed gap, to 10 bases between sequences of the same transcript, increased the number of transcripts by 3,258 (i.e. there are 3,258 transcripts which are merged when a 15 base gap is allowed). The permission of single copy k-mers increased the number of transcripts by 31,672 and genes by 30,469 however these will be enriched for those with little support.

The selected assembly (k2g15d2; Table 2.3) was then trimmed to further remove transcripts with low support, reducing the total transcript number to 80,800 (8 % of total transcripts trimmed). The distribution of transcript lengths is shown in Figure 2.1. The reference individual's original reads were mapped back to reference transcriptome, resulting in successful alignment of 68.9 % of reads (19,294,839 of 27,988,115 reads). Alignment success was consistent across samples sequenced and ranged from 61.4 % to 69.6 %, with a mean of 67.4 %. The assembled transcriptome has been submitted to DDBJ/EMBL/GenBank under accession number GEMC00000000.

Table 2.2 Per sample returns from RNA sequencing on an Illumina Hiseq2500 of twelve samples. This table indicates the total number of fragments sequenced per sample, the number of reads remaining after removal of poor quality reads (Q <15), and the percentage of total reads removed.

Sample	**Total fragments sequenced**	**Reads with Q >15**	**% Reads removed**
NAS080	28128352	27988115	0.5
NAS081	25972028	25863143	0.4
NAS057	24014626	23897299	0.5
NAS092	23467409	23383732	0.4
NAS095	24974526	24847043	0.5
NAS058	28504130	28299106	0.7
NAS061	27430238	27263143	0.6
NAS065	30038254	29802658	0.8
NAS068	30021260	29843121	0.6
NAS066	26151350	25992461	0.6
NAS070	28834483	28627104	0.7
NAS093	26291267	26205166	0.3

Table 2.3 Descriptors of the assemblies completed using differing settings to assess the nature of the data. The assemblies use RNASeq data from a commercially active watercress line, NAS080. Underlined assembly k2g15d2 (k-mer overlap: 2, maximum gap permitted within path: 15 bases, maximum differences allowed within a path: 2) was taken forward as the reference transcriptome for watercress.

Assembly	**Min k-mer coverage**	**Max gap allowed**	**Max differences allowed**	**Total transcripts**	**Total components**	**% GC**	**N50**
k2g10d2	2	10	2	91102	48635	41.12	1587
k2g15d0	2	15	0	87823	48709	41.09	1574
k2g15d2	**2**	**15**	**2**	**87844**	**48732**	**41.08**	**1571**
k2g15d4	2	15	4	87945	48717	41.09	1574
k2g15d8	2	15	8	87923	48701	41.08	1575
k1g15d2	1	15	2	119516	79201	40.72	1534
k1g15d4	1	15	4	119564	79225	40.74	1532

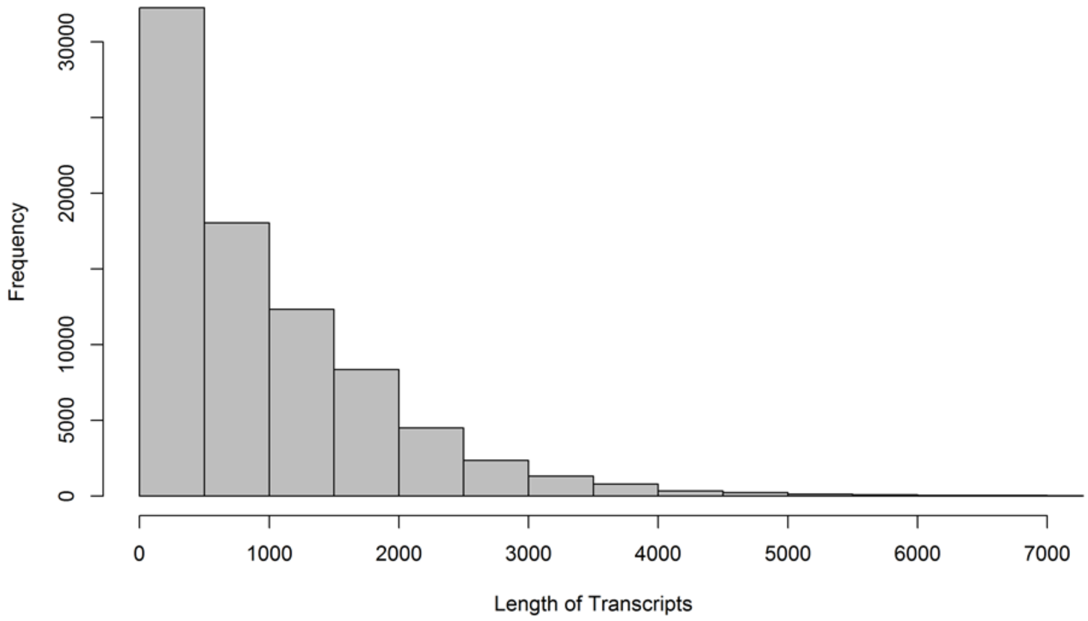


Figure 2.1 Assembled transcript length distribution. Frequency histogram showing the distribution of transcript length in the watercress reference transcriptome.

2.4.2 Annotation of the Watercress Transcriptome:

Of the 80,800 watercress transcripts, 54,595 (67.6 %) were annotated using a BLASTx search against *Arabidopsis* directly (Table S2.1) and mean hit match of watercress to *Arabidopsis* sequences was 84.9 %. An additional 3% of transcripts were annotated from the UniProtKB/SWISS-PROT database, a further 2,480 hits in *Arabidopsis* and 274 hits in other plant species. Throughout the whole transcriptome, the most represented Gene Ontology (GO) categories were ‘other cellular processes’, ‘other binding’, and ‘nucleus’ under each the GO categories biological process, molecular function and cellular component, respectively (Figure 2.2). A check for non-nuclear DNA contamination revealed 179 transcripts to be at least 95 % similar to mitochondrial or chloroplast DNA, which were flagged as such (0.2 % of all transcripts).

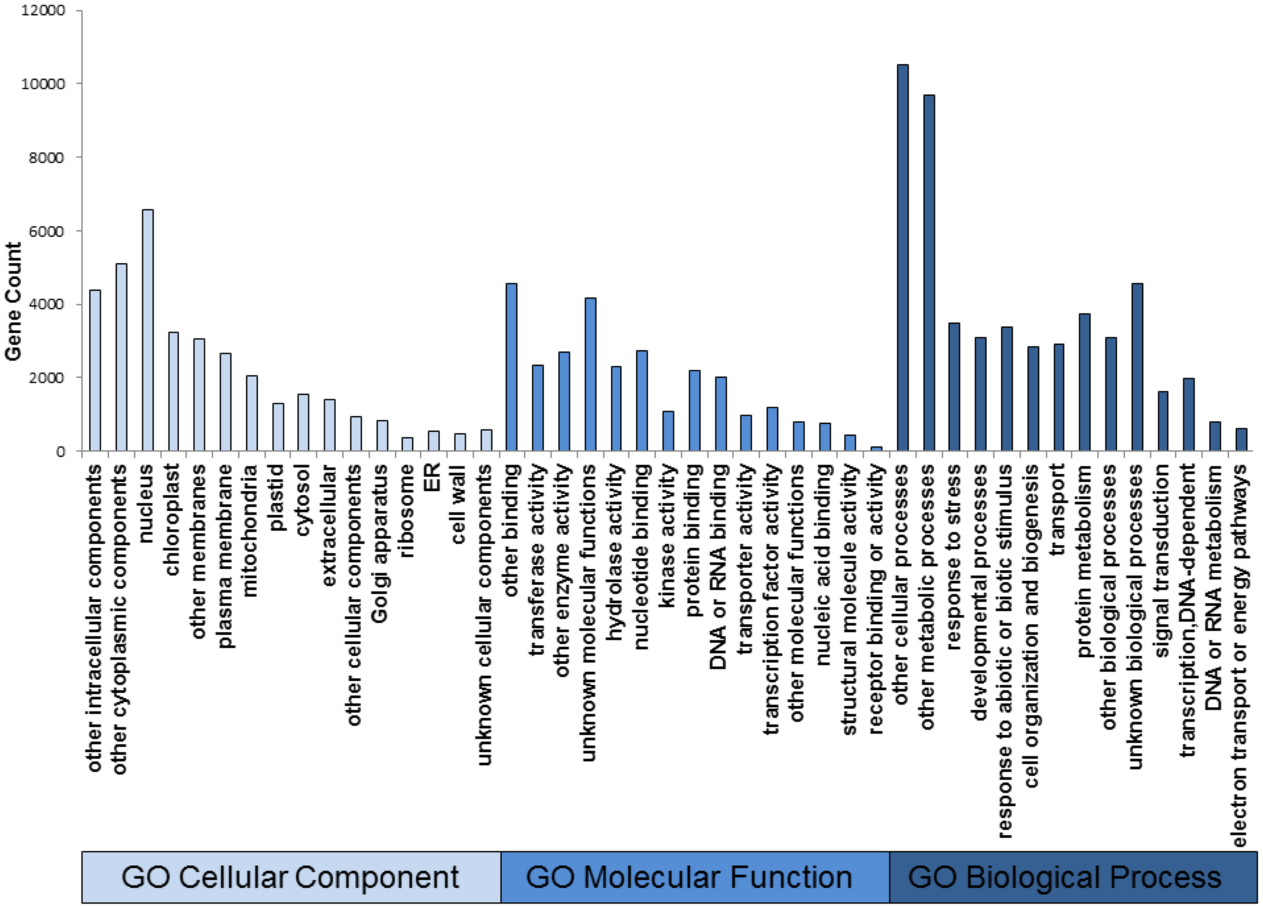


Figure 2.2 Gene ontology description of the watercress transcriptome. Histogram illustrating the number of genes in the reference watercress transcriptome belonging to GO terms for Biological Process, Molecular Function or Cellular Component categories.

2.4.3 Identification of Candidate Genes:

Known *Arabidopsis* AO and GLS biosynthesis pathway gene sequences were queried against the watercress transcriptome using BLASTn. The sequences for several *Arabidopsis* phenylpropanoid pathway enzymes had orthologs; specifically 19 of 24 phenylpropanoid genes queried had at least one close match in the watercress transcriptome (25 transcripts in total). Of the 19 hits, 14 were true orthologs as confirmed by a reciprocal best match BLAST query. The watercress transcripts were an 80.2 – 94.3 % (mean = 89.0 %) match to the *Arabidopsis* gene sequences.

For the GLS biosynthesis gene queried, 54 of 54 gene sequences were successfully matched to at least one watercress transcript (63 transcripts in total). For these 54 genes, the top hit was further confirmed as an orthologs by reciprocal BLAST query. These transcripts ranged from 81.0 % to 94.2 % (mean = 88.5 %) match to the queried *Arabidopsis* sequences. Four *Arabidopsis* loci identifiers (*AT1G62570, AT1G62540, AT1G62560,* and *AT1G65860*), all corresponding to sequences for the enzyme glucosinolate S-oxygenase, hit the same transcript in this search. In addition, three annotated transcripts (belonging to the same unigene) were identified as a match for the coding sequence of the enzyme myrosinase and three additional transcripts (two unigenes) are described as coding for myrosinase-like proteins.

2.4.4 Genetic Relatedness and Polymorphic Marker Development:

There were 46,078 (57.0 % of total transcripts produced) loci with at least 100 bases of sequence without missing data present in all twelve accessions, and these were compared using the software ProSeq3 (Filatov 2009) for the presence of polymorphisms. The number of transcripts containing at least one SNP was 10,134 (22 % of 46,078 transcripts) and 2,129 (4.6 %) contained 5 or more SNPs. Nucleotide diversity indices π (Nei & Li 1979) and θ (Watterson 1975; Halushka et al. 1999) were calculated across the dataset, and excluding sites with missing data, the mean π was 0. 78 and the mean θ was 0.87 per kilobase (Kb). In the reference transcriptome, 4,972 loci contained at least one Simple Sequence Repeat (SSR) with a total of 5,277 SSRs identified. Of these, 54 were compound (two SSRs within 50 bases of each other), 2,250 were dinucleotide repeat SSRs, 2,448 were trinucleotide repeat SSRs, and 525 were tetranucleiotide repeat SSRs. Seven thousand SNPs were used to draw the phylogenetic relationship between the accessions and is presented in Figure S2.1.

2.4.5 Differential Expression between High and Low Antioxidant Watercress:

Differential expression analysis was conducted to compare gene expression between five high and low AO watercress previously identified using edgeR (Robinson et al. 2010). For the AO trait, 145 transcripts (corresponding to 134 genes) were DE at a significance level of FDR ≤ 0.05 (60 transcripts at FDR ≤ 0.01, n = 10) (See Table S2.1 & Table S2.4). Fourteen transcripts did not have a BLAST hit and remain of unknown function. Many of the annotated DE loci are associated with mechanisms of stress tolerance, wounding, or response to threat and external stimulus (Table S2.1). The AgriGO pipeline confirmed this by highlighting 23 significantly over-represented GO categories in the DE loci, related to immune system response, response to biotic stimulus and stress response functions (See Figure 2.3).

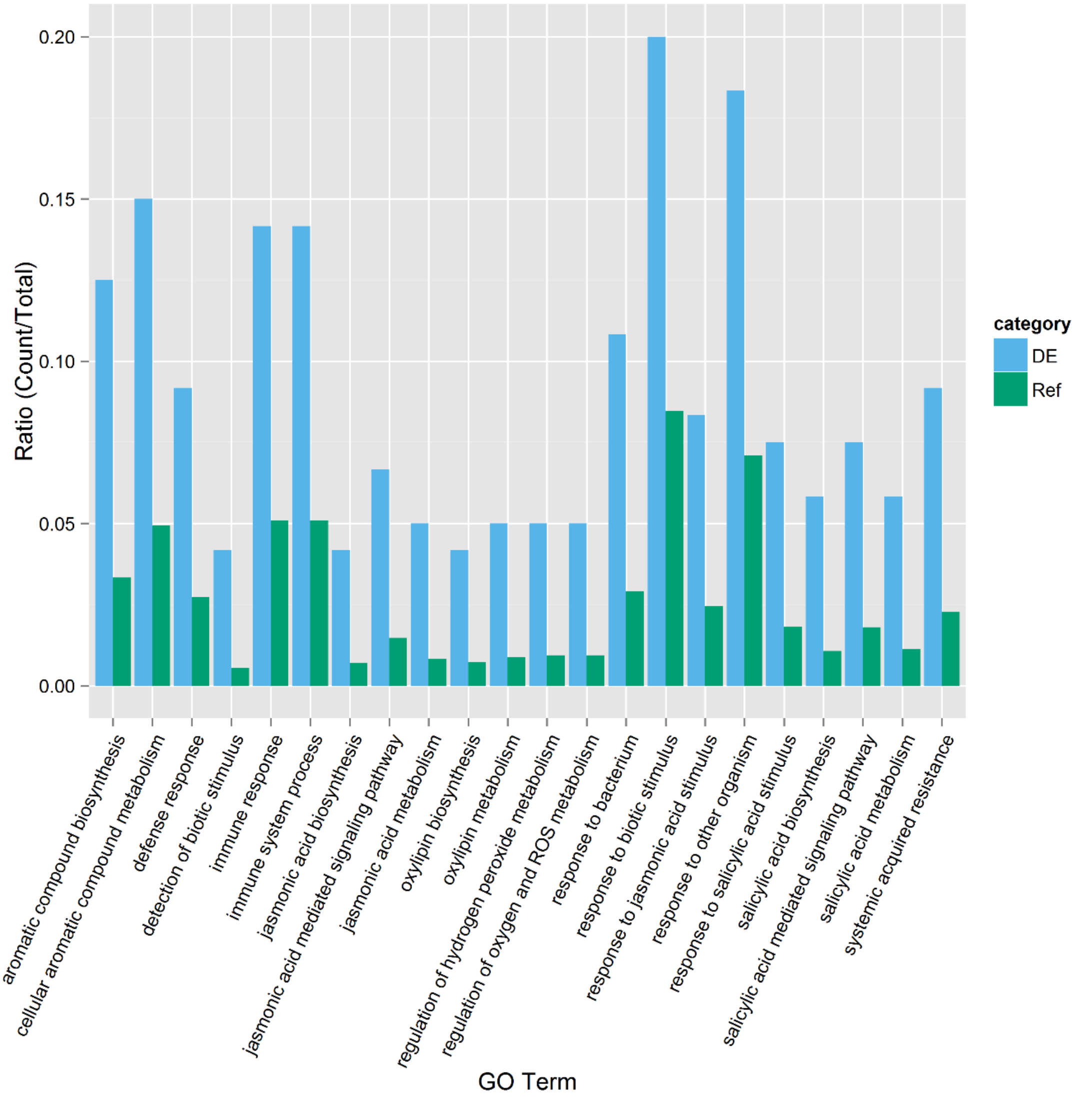


Figure 2.3 Highlighted gene ontology categories in high antioxidant watercress. Barplot depicting standardized gene count (ratio of gene count in that category over total gene count) of significantly overrepresented GO terms in the AO DE genes in comparison the reference transcriptome.

2.4.6 Differential Expression between High and Low Glucosinolate Watercress:

The DE analysis for GLS content yielded 94 DE loci at a significance level of FDR ≤ 0.05, corresponding to 93 different genes (50 transcripts at FDR ≤ 0.01, n = 6) (See Table S2.1 & Table S2.5). Twenty four of these transcripts did not have a BLAST hit (Table S2.5). The functional classification of the 70 annotated loci was completed using AgriGO and yielded only one significant GO term: exopeptidase activity. The DE results revealed several genes with putative functions related to GLS biosynthesis; including 13 stress response genes and two genes associated with the shikimate pathway (*AT3G06350* & *AT2G35500*). This pathway results in the production of chorismate which is then converted to phenylalanine (Tzin & Galili 2010), the precursor to aromatic GLS, including the most abundant GLS in watercress: gluconasturtiin (Macleod & Islam 1975; Kopsell et al. 2007; Manion et al. 2014). The shikimate pathway produces chorismate through seven steps involving six enzymes, chorismate is then converted to L-phenylalanine primarily via an arogenate intermediate (Tzin & Galili 2010). We used BLASTn to identify equivalent transcripts to the genes in these pathways, the *Arabidopsis* sequences of which were obtained from NCBI database. This resulted in discovery of 112 transcripts which matched the 18 known shikimate and phenylalanine biosynthesis pathway gene sequences. The total of standardised expression counts of all transcript isoforms for the best match watercress unigene (lowest e-value and highest score) are shown in Figures 2.4 and 2.5. We also compared expression levels (standardized mean count) of the GLS biosynthesis candidate genes identified previously (Figure 2.6). Although the expression of these transcripts was not significantly different between high and low GLS concentration watercress based on the transcriptome-wide analysis (ca. 80,000 loci), there was a noticeable trend of up-regulation of genes involved in the shikimate (15/17) and GLS (39/54) biosynthetic pathways in the high GLS watercress.

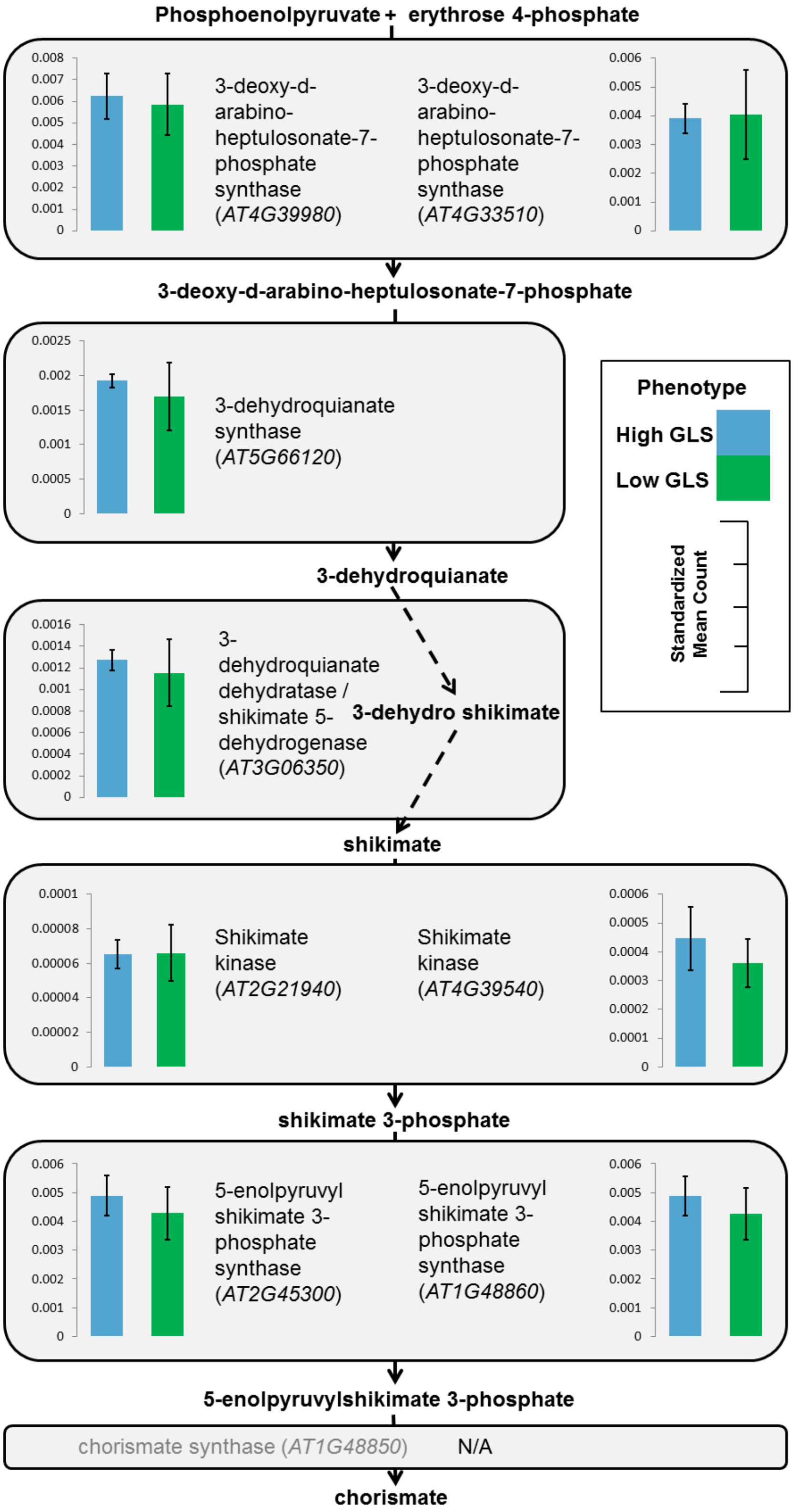


Figure 2.4 Expression levels throughout the shikimate pathway in high and low GLS watercress. Representation of the shikimate biosynthesis pathway with expression levels, as standardized mean counts (± standard error of the mean), of the best match transcript for high and low glucosinolate accessions. Chorismate synthase (AT1G48850) did not have a BLAST hit to the watercress transcriptome.

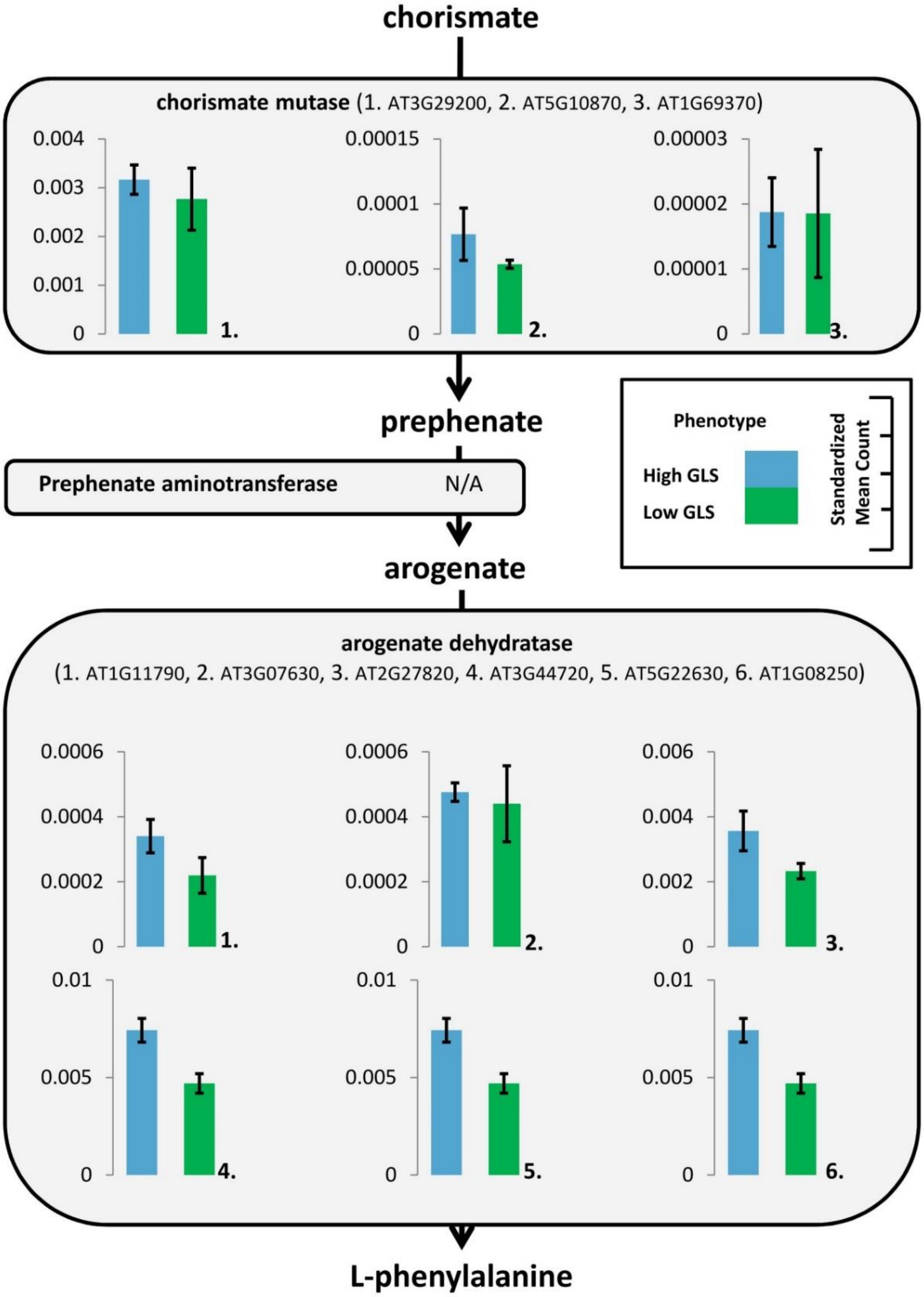


Figure 2.5 Expression levels throughout phenylalanine biosynthesis in high and low GLS watercress. Representation of the most common phenylalanine biosynthesis pathway in plants with expression levels, as standardized mean counts (± standard error of the mean), of the best match transcript for high and low glucosinolate accessions. Prephenate aminotransferase did not have an available consensus sequence currently.

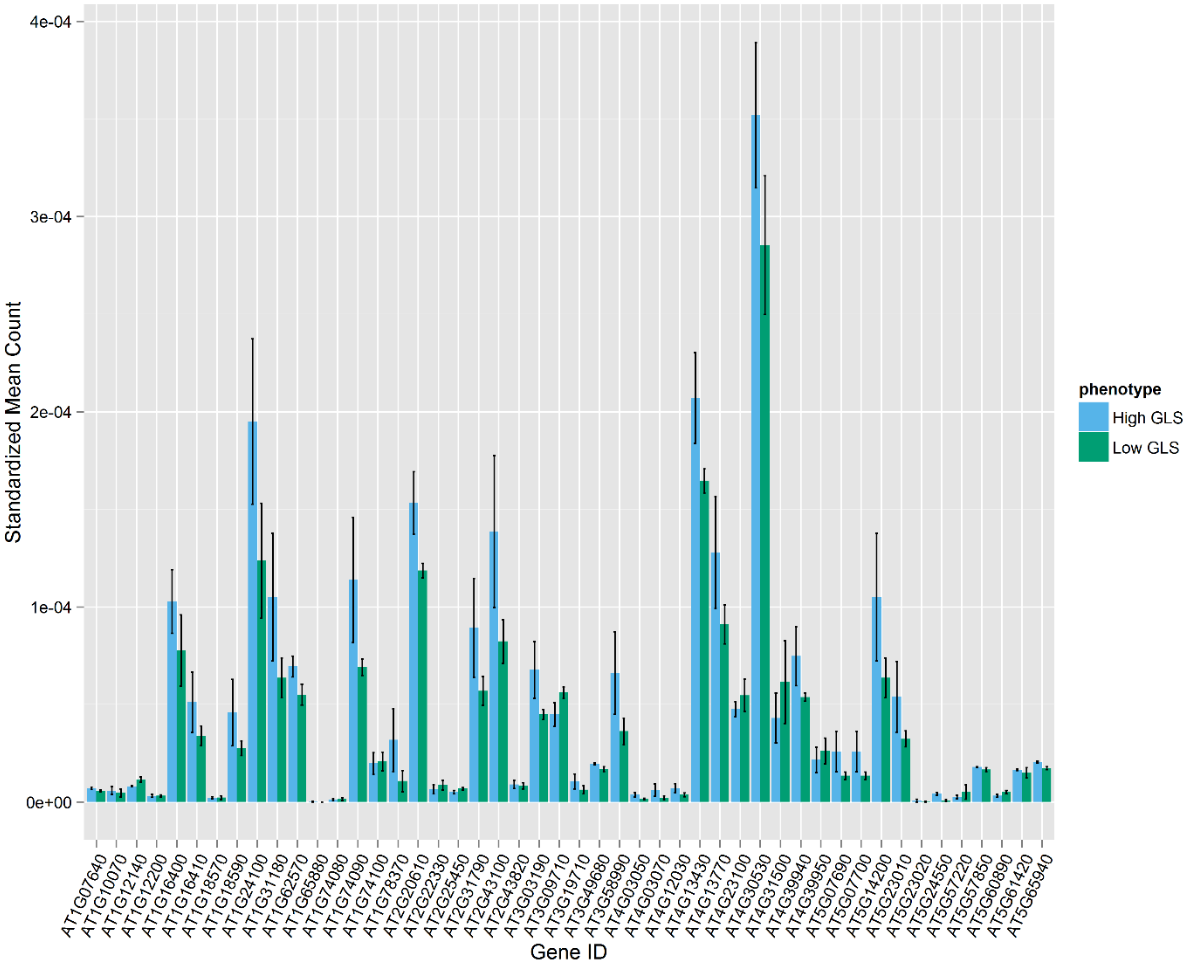


Figure 2.6 Expression levels of genes in GLS biosynthesis in high and low GLS watercress. Mean expression levels (± standard error of the mean), as standardized mean counts, of watercress transcripts similar to known glucosinolate biosynthesis genes in high and low glucosi nolate accessions.

2.4.7 Sequence Divergence between High and Low Accessions:

Strong sequence divergence between high and low AO and GLS accessions would be expected for loci involved in these pathways. We therefore calculated F_{ST} for all loci in ProSeq3. At a cut-off of $F_{ST} = 0.5$, 608 (of 19,229) and 306 (of 21,924) loci showed high sequence divergence between high/low AO and high/low GLS groups, respectively. Some of the loci in this subset may play a role in governing AO or GLS biosynthesis.

This comparison between high/low AO groups revealed only three loci with fixed differences between the high and low accessions. These transcripts matched those for a sucrose/ferredoxin-like protein, a putative RING-H2 finger protein associated with abscisic acid signalling, and a chloroplast-specific heat shock protein.

The comparison between high/low GLS accessions yielded five transcripts with at least one fixed site. These five loci corresponded to two beta glucosidases (*AT2G25630* & *AT2G44450*), involved in carbohydrate metabolic processes; a protein kinase with potential function in salicylic acid biosynthesis (*AT5G47070*); a methyltransferase (*AT1G50000*); an MEK kinase (*AT1G53570*); and an unknown protein (*AT1G50020*).

2.5 Discussion

Watercress is recognised as a crop with especially high concentrations of certain phytonutrients. These compounds not only confer the characteristic peppery taste associated with watercress, but are now considered to also offer important health benefits. However, limited knowledge exists on watercress genetics and genomics hindering efforts to preserve or select for these key traits. In this paper, we present the first transcriptome sequence for watercress that has utilised a unique germplasm resource collected globally and held currently at The University of Southampton (Payne et al. 2015). We compiled a catalogue of over 80,000 watercress transcripts (57,349 annotated), described and compared the gene expression profile of ready-for-market watercress with contrasting antioxidant (AO) and glucosinolate (GLS) phenotypes and identified candidate genes for follow-up work, a subset of which may be useful in future watercress breeding. Some of the candidate genes identified in this analysis correspond to known metabolite pathways as well as others which require further investigation.

2.5.1 Watercress Transcriptome *de novo* Assembly:

Plants used in this study were harvested at the time point when the crop would be sent to market. The ten watercress samples with ‘extreme’ phytonutritional phenotypes and two control accessions were extracted for RNA and sequenced. NAS080, a commercial accession, was used to assemble a watercress reference transcriptome which comprised of 87,844 transcripts (trimmed to 80,800) and 48,732 corresponding “unigenes” (Table 2.3). For the *de novo* transcriptome assembly of the allohexaploid *Spartina* species, Ferreira de Carvalho et al. (2013) applied a less stringent assembly in order to accommodate for up to six different alleles per sequenced locus. As watercress is thought to be tetraploid (Bleeker et al. 1999; Morozowska et al. 2010), we also applied this approach. We conducted a variety of different assemblies which would potentially allow for the collapse, if present, of four alleles per locus into one. However, allowing for 0 to 8 differences (SNPs) within a path made no notable differences in transcripts or genes compiled among assemblies (Table 2.3). This would suggest that the watercress genome, if polyploidy, is likely to be autopolyploid, which would allow for duplicate polyploid genes (if expressed) to collapse into one regardless of assembly allowances.

By BLAST query against *Arabidopsis,* a close relative to watercress, coding regions we were able to annotate 70.6 % of the transcripts (57,075 of 80,800 transcripts) with an *Arabidopsis* locus identifier. Only 0.4 % of transcripts had a top hit in other plant species. For broccoli, another member of the Brassicaceae, 77.0% of *de novo* assembled transcripts were annotated based on homology to *Arabidopsis* using an e-value of e^{-5} (Gao et al. 2014). In our analysis, there were several cases were multiple watercress transcripts matched the same *Arabidopsis* locus identifier. This is likely to be a result of different fragments of the same transcript not being joined into a single transcript during assembly and/or gene duplication or loss in the lineage leading to one of the species. The transcripts that were not successfully annotated could be transcripts not shared with *Arabidopsis*, unique to watercress or incompletely assembled.

Watercress is assumed to be primarily self-fertilizing and spreads through clonal growth and root expansion. Commercial watercress is clonally propagated or selfed, since there is no current selection and breeding programme globally, so it is considered that watercress should have little genetic diversity. Thus, we would expect low polymorphism between accessions. Our results are consistent with this hypothesis, with 22 % of transcripts containing a polymorphic site. Nucleotide diversity was low across the entire data set (mean π = 0.78 and mean θ = 0.87 per Kb). For comparison, transcriptome nucleotide diversity θ in cultivated and wild carrot roots was 0.56 and 0.64 per Kb respectively (Rong et al. 2014). The common bean transcriptome nucleotide diversity was greater than watercress and, in a comparison of Mesoamerican wild and cultivated beans, the wild variety (π = 2.11, θ = 2.08 per Kb) had higher diversity than its cultivated counterpart (π = 0.85, θ = 0.83 per Kb) (Bellucci et al. 2014).

2.5.2 Gene Expression and Antioxidant Capacity in Watercress:

The AO trait is desirable in crops cultivated for human consumption and is of particular interest in leafy salads, with the links between consumption of high AO leaves and their disease-preventing properties now becoming established. The phenylpropanoid pathway is an important and well-characterized pathway associated with the production of secondary plant metabolites and dietary AO compounds (Fraser & Chapple 2011) and here, thirty six transcripts matched 21 of 24 phenylpropanoid pathway sequences queried. Although considered at the gene sequence level and not taken through to translation, our findings suggest well-conserved gene sequences between *Arabidopsis* and watercress in the phenylpropanoid pathway and represent an immediately useful catalogue of important genes likely contributing to the AO crop trait.

We also completed DE analysis on five high and five low AO 'extreme' samples to describe the character of this trait at the whole-transcriptome level. DE analysis between high and low AO watercress returned 145 DE transcripts from 23 GO categories which were significantly associated with plant immunity, response to stimuli and stress. This direct link between plant defences and AO profile is not surprising, considering most compounds contributing to plant AO capacity are secondary plant metabolites associated with the very plant functions highlighted by the GO results. This link has been confirmed in field conditions. A multi-year field study on cauliflower showed annually variable phytochemical and AO contents which were linked to climate and rainfall (Lo Scalzo et al. 2013), also confirming a significant environment component to this trait. In our laboratory, a significant difference between AO capacity (FRAP assay) of watercress grown in the field and in controlled environments has been identified, with field samples being overall higher (Payne et al. 2015). These studies confirm that the synthesis and accumulation of secondary metabolites, underpinning the increase in AO capacity, is linked to plant response to external stimuli and stress (i.e. abiotic environmental stress or biotic stress through predation or pathogens). Thus, plant stress and immunity response genes and pathways should be considered strong candidates for breeding high AO food crops.

Although AO assays such as FRAP and ORAC provide a consistent measure of total AO capacity (Payne et al. 2013), the assays are unable to provide significant details on the specific compounds present that underpin AO. This is a disadvantage when seeking a particular compound or pathway to attribute this health benefit but useful in overall characterisation of the consumer benefit derived from a crop. Thus, the phenotype we have assessed here represents a combination of multiple compounds with AO properties, and may include polyphenols (anthocyanins, flavonols, isoflavonoids, catechins, caffeoylquinic acid), carotenoids (lycopene, β-carotene, lutein), tocotrienols, tocopherols and ascorbic acid (Martin et al. 2013). Indeed, several of the DE transcripts corresponded to elements of these AO compound biosynthetic pathways. For example, ferulate 5-hydroxylase (*AT4G36220*) is a phenylpropanoid pathway enzyme involved in lignin biosynthesis (Franke et al. 2000; Fraser & Chapple 2011), three transcripts (*AT5G41040*, *AT2G28630*, *AT2G28670*) are associated with suberin biosynthesis, a cell wall polymer containing phenolic components (Soler et al. 2007), a putative carotenoid hydrolase (*AT4G15110*), and tyrosine aminotransferase (*AT2G24850*) which is involved in tocopherol synthesis (Riewe et al. 2012).

2.5.3 Genes and Pathways Associated with GLS Content of Watercress:

GLS are secondary plant metabolites utilized in plant defences against herbivory and have been the subject of many studies in the Brassicaceae. They contribute to the peppery flavour as well as the strong phytonutritional profile associated with watercress, thus the pathways and genes involved in the biosynthesis and processing of these compounds are an important research and breeding target for this crop. GLS biosynthesis is well-studied and the enzymes and genes involved in these steps are well-described in *Arabidopsis* and *Brassica rapa* for aliphatic and indolic GLS (Fahey et al. 2001; Wang et al. 2013). Here, sequences of known GLS pathway genes in *Arabidopsis* were successfully identified in watercress. Wang et al. (2013) used RNASeq to identify GLS biosynthesis genes in radish taproots as, similarly to watercress, these compounds contribute to the dietary and flavour profile of the crop. The authors identified sequences in radish that matched *Arabidopsis* and *B. rapa* GLS gene sequences and suggested that these genes are well-conserved in the Brassicaceae family (Wang et al. 2013). Our findings support this hypothesis, as all GLS pathway gene sequences were also identified in watercress. In addition, we identified transcripts in watercress matching the *Arabidopsis* myrosinase coding sequence. This catalogue is immediately useful for further study of GLS biosynthesis in watercress, as well as in breeding, for hunting allelic variation in germplasm collections.

In addition, we compared whole transcriptome gene expression of three high and three low GLS watercress. A total of 94 transcripts were DE for this phenotype. Twenty four of these did not have a BLAST hit in *Arabidopsis.* Although the DE genes for this trait did not contain any GO categories with immediately obvious connection to GLS biosynthesis and regulation, there were several DE genes which were interesting on a gene-by-gene basis. Specifically, two

DE transcripts belonged to the shikimate pathway (*c33663_g1_i2* –similar to shikimate kinases, *c37926_G1_i6* – dehydroquinate-shikimate dehydrogenase). The shikimate pathway leads to the synthesis of chorismate which is the precursor to phenylalanine, from which gluconasturtiin is derived (see results). This direct link prompted a further investigation of the shikimate and phenylalanine biosynthetic pathways genes for which we used the known *Arabidopsis* sequences to mine for orthologs in watercress. These results are depicted in Figures 2.4, 2.5 and 2.6 and show greater expression of 15 out of 17 genes in the high GLS watercress suggesting increased flux through this pathway in the high GLS plants. The potential connection between the shikimate pathway output and GLS levels in a plant provides a direct and appealing link for further investigation and would be of particular breeding interest, as phenylalanine also feeds into the AO phenylpropanoid pathway.

The GLS content of any plant tissue is under both genetic and environmental controls and depends on a variety of factors and conditions, including developmental stage (Booth et al. 1991; Velasco et al. 2007; Gao et al. 2014), environmental conditions (Lo Scalzo et al. 2013), and pest/herbivore exposure (Engelen-Eigles et al. 2006; Velasco et al. 2007). For watercress, studies have shown GLS content variation in response to soil nitrogen and sulphur (Kopsell et al. 2007), selenium (Manion et al. 2014), as well as light and temperature (Engelen-Eigles et al. 2006). In another study, 62 varieties of Chinese cabbage assessed were found to vary ca. 20-fold in GLS content, suggesting an effect of genotype on GLS production and accumulation (Lee et al. 2014). Despite this, the variation in germplasm collection reported here, when all material was grown under identical environmental conditions, suggests there is potential for selective breeding for higher GLS. In fact, such a breeding endeavour has been undertaken successfully in broccoli, where an enriched GLS crop was produced through molecular breeding techniques and was shown to be associated with enhanced chemopreventive activity (Mithen et al. 2003). More recently, Beneforte broccoli has been released to market having 2.5 - 3 times higher GLS content than other broccoli varieties (Traka et al. 2013).

It is clear that the controls involved in the regulation GLS biosynthesis and accumulation in plants are complex and interdependent (Yan & Chen 2007). Several DE genes in this study could be linked to relevant regulatory pathways, such as stress and immune response, development and life stage, and ion or light response. Interestingly, 13 of DE 93 loci identified were linked with stress or immune response in plants, including genes associated with abscisic acid, jasmonic acid and salicylic acid signalling; an ethylene response transcription factor; a heat shock protein; glutathione-S-transferase, which is involved in cell detoxification; and a carotenoid biosynthesis enzyme.

As discussed previously, watercress GLS concentrations have been shown to respond to certain soil nutrients (Kopsell et al. 2007; Manion et al. 2014; Thiruvengadam & Chung 2015). We identified two genes involved in cadmium ion response (*AT4G08790* & *AT4G10320*) that were DE between the high and low GLS plants. Watercress GLS content has also been shown to respond to light (Engelen-Eigles et al. 2006) and our list of DE loci included a carotenoid biosynthesis enzyme (*AT4G25700*), carotenoids play a key role in photosynthesis and protects plant photosynthetic machinery from light damage (Young 1991; Cazzonelli 2011), and a phototropic-response protein (*AT3G44820*).

Finally, there were several DE elements for the GLS phenotype that were related to developmental processes. We resolved two MYB transcription factors; Circadian 1 (AT5G37260) and the circadian rhythm putative transcription factor LHY (*AT1G01060*). Certain MYB transcriptional factors have been suggested to act in GLS biosynthesis regulation (Celenza et al. 2005; Yan & Chen 2007). However, these transcription factors do not appear to fit previously suggested MYB links to GLS regulation, instead both are involved in circadian rhythms. An additional two transcription factors were DE here: *AT1G11950*, which contains a jumonji domain and is associated with flowering time, and a transcription factor of unknown function (*AT2G42780*). A pectin lyase-like protein was also differentially expressed (*AT1G19170*). Pectin lyases, which are cell wall components, are thought to act in fruit ripening and senescence amongst other plant developmental processes (Marin-Rodriguez 2002). These findings are in support of previous field results showing differences in tissue GLS concentration over time and plant maturity (Booth et al. 1991; Velasco et al. 2007).

2.6 Conclusions

In conclusion, we present the first fully annotated whole transcriptome sequencing of the highly nutritious leafy crop, watercress. Differential expression analysis of ‘extreme’ samples was used to detect genes potentially important to key nutritional traits and identified transcripts pertaining to the shikimate, phenylpropanoid and GLS biosynthetic pathways. The transcriptome of watercress offers a valuable resource for comparative study of the Brassicaceae which contains many crops, several of which have unique nutrient qualities which benefit humans. This work furthers our understanding of key genes and pathways associated with phytonutrient phenotypes in watercress and the genomic resources gathered will allow for the development of markers for marker assisted selection and further molecular studies on watercress, with aims to inform industry and research.

2.7 Author Contributions

NV contributed through data collection, analysis and interpretation, and wrote the manuscript. AP contributed to data collection. RH contributed to data collection and revisions to the manuscript. GJJC and SR contributed to project design. MC and GT conceived of the study, guided data analysis and interpretation, and revised the manuscript. All authors have read and approved of the final manuscript.

2.8 References

Agerbirk, N. et al., 2014. Specific glucosinolate analysis reveals variable levels of epimeric glucobarbarins, dietary precursors of 5-phenyloxazolidine-2-thiones, in watercress types with contrasting chromosome numbers. *Journal of agricultural and food chemistry*, 62(39), pp.9586–96.

Aires, A. et al., 2013. Phytochemical characterization and antioxidant properties of baby-leaf watercress produced under organic production system. *CyTA - Journal of Food*, 11(4), pp.343–351.

Bailey, C.D. et al., 2006. Toward a global phylogeny of the Brassicaceae. *Molecular biology and evolution*, 23(11), pp.2142–60.

Beilstein, M.A. et al., 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, 107(43), pp.18724–8.

Bellucci, E. et al., 2014. Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. *The Plant cell*, 26(5), pp.1901–1912.

Benzie, I.F. & Strain, J.J., 1996. The ferric reducing ability of plasma (FRAP) as a measure of “antioxidant power”: the FRAP assay. *Analytical biochemistry*, 239(1), pp.70–76.

Bleeker, W., Huthmann, M. & Hurka, H., 1999. Evolution of the hybrid taxa in Nasturtium R.BR. (Brassicaceae). *Folia Geobotanica*, 34, pp.421–433.

Bones, A.M. & Rossiter, J.T., 1996. The myrosinase-glucosinolate system, its organisation and biochemistry. *Physiologia Plantarum*, 97(1), pp.194–208.

Booth, E.J., Walker, K.C. & Griffiths, D.W., 1991. A time-course study of the effect of sulphur on glucosinolates in oilseed rape (Brassica napus) from the vegetative stage to maturity. *Journal of the Science of Food and Agriculture*, 56(4), pp.479–493.

Cantín, C.M., Moreno, M.A. & Gogorcena, Y., 2009. Evaluation of the antioxidant capacity, phenolic compounds, and vitamin C content of different peach and nectarine [*Prunus persica* (L.) Batsch] breeding progenies. *Journal of agricultural and food chemistry*, 57(11), pp.4586–4592.

Cavell, B.E. et al., 2011. Anti-angiogenic effects of dietary isothiocyanates: mechanisms of action and implications for human health. *Biochemical pharmacology*, 81(3), pp.327–336.

Cazzonelli, C.I., 2011. Goldacre Review: Carotenoids in nature: insights from plants and beyond. *Functional Plant Biology*, 38(11), p.833.

Celenza, J.L. et al., 2005. The Arabidopsis ATR1 Myb transcription factor controls indolic glucosinolate homeostasis. *Plant physiology*, 137(1), pp.253–262.

Engelen-Eigles, G. et al., 2006. The effect of temperature, photoperiod, and light quality on gluconasturtiin concentration in watercress (*Nasturtium officinale* R. Br.). *Journal of agricultural and food chemistry*, 54(2), pp.328–334.

Fahey, J.W., Zalcmann, A.T. & Talalay, P., 2001. The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry*, 56(1), pp.5–51.

Ferreira de Carvalho, J. et al., 2013. Transcriptome *de novo* assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). *Heredity*, 110, pp.181–193.

Filatov, D.A., 2009. Processing and population genetic analysis of multigenic datasets with ProSeq3 software. *Bioinformatics*, 25(23), pp.3189–3190.

Fogarty, M.C. et al., 2013. Acute and chronic watercress supplementation attenuates exercise-induced peripheral mononuclear cell DNA damage and lipid peroxidation. *The British journal of nutrition*, 109(2), pp.293–301.

Franke, R. et al., 2000. Modified lignin in tobacco and poplar plants over-expressing the Arabidopsis gene encoding ferulate 5-hydroxylase. *The Plant Journal*, 22(3), pp.223–234.

Fraser, C.M. & Chapple, C., 2011. The phenylpropanoid pathway in Arabidopsis. *The Arabidopsis book / American Society of Plant Biologists*, 9, p.e0152.

Gao, J. et al., 2014. RNA-Seq analysis of transcriptome and glucosinolate metabolism in seeds and sprouts of broccoli (*Brassica oleracea* var. italic). *PloS one*, 9(2), p.e88804.

Gill, C.I.R. et al., 2007. Watercress supplementation in diet reduces lymphocyte DNA damage and alters blood antioxidant status in healthy adults. *The American journal of clinical nutrition*, 85(2), pp.504–510.

Grabherr, M. et al., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature biotechnology*, 29(7), pp.644–652.

Haas, B.J. et al., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8, pp.1494–1512.

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, pp.95–98.

Halushka, M.K. et al., 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature genetics*, 22(3), pp.239–247.

Hecht, S. et al., 1995. Effects of watercress consumption on metabolism of a tobacco-specific lung carcinogen in smokers. *Cancer Epidemiol. Biomarkers Prev.*, 4(8), pp.877–884.

Howard, A.H.W. & Lyon, A.G., 1952. *Nasturtium officinale* R. Br. (*Rorippa Nasturtium-Aquaticum* (L.) Hayek). *Journal of Ecology*, 40(1), pp.228–245.

Kavitha, P. et al., 2014. Genotypic variability for antioxidant and quality parameters among tomato cultivars, hybrids, cherry tomatoes and wild species. *Journal of the science of food and agriculture*, 94(5), pp.993–999.

Kopsell, D. a et al., 2007. Influence of nitrogen and sulfur on biomass production and carotenoid and glucosinolate concentrations in watercress (*Nasturtium officinale* R. Br.). *Journal of agricultural and food chemistry*, 55(26), pp.10628–10634.

Lai, K.-C. et al., 2010. Phenethyl Isothiocyanate Inhibited Tumor Migration and Invasion via Suppressing Multiple Signal Transduction Pathways in Human Colon Cancer HT29 Cells. *Journal of agricultural and food chemistry*, 58(20), pp.11148–11155.

Lee, M.-K. et al., 2014. Variation of glucosinolates in 62 varieties of Chinese cabbage (*Brassica rapa* L. ssp. pekinensis) and their antioxidant activity. *LWT - Food Science and Technology*, 58(1), pp.93–101.

Li, H. et al., 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, pp.2078–2079.

Lin, X. et al., 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, 402(6763), pp.761–768.

Macleod, A.J. & Islam, R., 1975. Volatile flavour components of watercress. *Journal of the Science of Food and Agriculture*, 26(10), pp.1545–1550.

Manchali, S., Chidambara Murthy, K.N. & Patil, B.S., 2012. Crucial facts about health benefits of popular cruciferous vegetables. *Journal of Functional Foods*, 4(1), pp.94–106.

Manion, L.K. et al., 2014. Selenium fertilization influences biomass, elemental accumulations, and phytochemical concentrations in watercress. *Journal of Plant Nutrition*, 37(3), pp.327–342.

Manton, I., 1935. The cytological history of watercress (*Nasturtium officinale* R. Br.). *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, 69(1), pp.132–157.

Marin-Rodriguez, M.C., 2002. Pectate lyases, cell wall degradation and fruit softening. *Journal of Experimental Botany*, 53(377), pp.2115–2119.

Martin, C. et al., 2013. Plants, diet, and health. *Annual review of plant biology*, 64, pp.19–46.

Martínez-Sánchez, A. et al., 2008. A comparative study of flavonoid compounds, vitamin C, and antioxidant properties of baby leaf Brassicaceae species. *Journal of agricultural and food chemistry*, 56(7), pp.2330–2340.

Minarini, A. et al., 2014. Exploring the effects of isothiocyanates on chemotherapeutic drugs. *Expert opinion on drug metabolism & toxicology*, 10(1), pp.25–38.

Mithen, R. et al., 2003. Development of isothiocyanate-enriched broccoli, and its enhanced ability to induce phase 2 detoxification enzymes in mammalian cells. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 106(4), pp.727–34.

Morozowska, M., Czarna, A. & Jędrzejczyk, I., 2010. Estimation of nuclear DNA content in Nasturtium R. Br. by flow cytometry. *Aquatic Botany*, 93(4), pp.250–253.

Nei, M. & Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76, pp.5269–5273.

Newman, R.M., Hanscom, Z. & Kerfoot, W.C., 1992. The watercress glucosinolate-myrosinase system: a feeding deterrent to caddisflies, snails and amphipods. *Oecologia*, 92, pp.1–7.

Di Noia, J., 2014. Defining powerhouse fruits and vegetables: a nutrient density approach. *Prev Chronic Dis*, 11, p.130390.

Palaniswamy, U.R. & Mcavoy, R.J., 2001. Watercress: A salad crop with chemopreventive potential. *HortTechnology*, 11(4), pp.622–626.

Pandey, K.B. & Rizvi, S.I., 2009. Plant polyphenols as dietary antioxidants in human health and disease. *Oxidative medicine and cellular longevity*, 2(5), pp.270–278.

Payne, A.C. et al., 2013. Antioxidant assays - consistent findings from FRAP and ORAC reveal a negative impact of organic cultivation on antioxidant potential in spinach but not watercress or rocket leaves. *Food science & nutrition*, 1(6), pp.439–44.

Payne, A.C. et al., 2015. Diversity in global gene expression and morphology across a watercress (*Nasturtium officinale* R. Br.) germplasm collection: first steps to breeding. *Horticulture Research*, 2, p.15029.

Prohens, J. et al., 2007. Total phenolic concentration and browning susceptibility in a collection of different varietal types and hybrids of eggplant: Implications for breeding for higher nutritional quality and reduced browning. *J. Amer. Soc. Hort. Sci.*, 132(5), pp.638–646.

Riewe, D. et al., 2012. A tyrosine aminotransferase involved in tocopherol synthesis in Arabidopsis. *The Plant journal : for cell and molecular biology*, 71(5), pp.850–859.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), pp.139–140.

Rochfort, S.J. et al., 2008. Class targeted metabolomics: ESI ion trap screening methods for glucosinolates based on MSn fragmentation. *Phytochemistry*, 69(8), pp.1671–1679.

Rong, J. et al., 2014. New insights into domestication of carrot from root transcriptome analyses. *BMC genomics*, 15(1), p.895.

Rose, P. et al., 2000. 7-Methylsulfinylheptyl and 8-methylsulfinyloctyl isothiocyanates from watercress are potent inducers of phase II enzymes. *Carcinogenesis*, 21(11), pp.1983–1988.

Rose, P. et al., 2005. Broccoli and watercress suppress matrix metalloproteinase-9 activity and invasiveness of human MDA-MB-231 breast cancer cells. *Toxicology and applied pharmacology*, 209(2), pp.105–113.

Rothwell, S.D. & Robinson, L.W., 1986. Cold acclimation potential of watercress in relation to growing season and nutrient status. *Journal of Horticultural Science*, 61(3), pp.373–378.

Santos, J. et al., 2014. Phenolic profile evolution of different ready-to-eat baby-leaf vegetables during storage. *Journal of chromatography. A*, 1327, pp.118–131.

Lo Scalzo, R. et al., 2013. Variations in the phytochemical contents and antioxidant capacity of organically and conventionally grown Italian cauliflower (*Brassica oleracea* L. subsp. botrytis): Results from a three-year field study. *Journal of agricultural and food chemistry*, 61(43), pp.10335–10344.

Sheridan, G.E.C. et al., 2001. Genetic diversity within commercial populations of watercress (*Rorippa nasturtium-aquaticum*), and between allied Brassicaceae inferred from RAPD-

PCR. *Euphytica*, 122, pp.319–325.

Soler, M. et al., 2007. A genomic approach to suberin biosynthesis and cork differentiation. *Plant physiology*, 144(1), pp.419–431.

Syed Alwi, S.S. et al., 2010. In vivo modulation of 4E binding protein 1 (4E-BP1) phosphorylation by watercress: a pilot study. *The British journal of nutrition*, 104(9), pp.1288–1296.

The UniProt Consortium, 2014. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(Database issue), pp.D204-12.

Theologis, A. et al., 2000. Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. *Nature*, 408(6814), pp.816–820.

Thiruvengadam, M. & Chung, I.-M., 2015. Selenium, putrescine, and cadmium influence health-promoting phytochemicals and molecular-level effects on turnip (*Brassica rapa* ssp. rapa). *Food chemistry*, 173, pp.185–193.

Traka, M. & Mithen, R., 2008. Glucosinolates, isothiocyanates and human health. *Phytochemistry Reviews*, 8(1), pp.269–282.

Traka, M.H. et al., 2013. Genetic regulation of glucoraphanin accumulation in Beneforté broccoli. *The New phytologist*, 198(4), pp.1085–1095.

Tzin, V. & Galili, G., 2010. The biosynthetic pathways for shikimate and aromatic amino acids in *Arabidopsis thaliana*. *The Arabidopsis book / American Society of Plant Biologists*, 8, p.e0132.

Varshney, R.K. et al., 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in biotechnology*, 27(9), pp.522–530.

Vaz Patto, M.C. et al., 2014. Achievements and challenges in improving the nutritional quality of food legumes. *Critical Reviews in Plant Sciences*, 34(1–3), pp.105–143.

Velasco, P. et al., 2007. Factors affecting the glucosinolate content of kale (*Brassica oleracea* acephala group). *Journal of agricultural and food chemistry*, 55(3), pp.955–962.

Verhoeven, D. et al., 1996. Epidemiological studies on brassica vegetables and cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, 5(9), pp.733–748.

Wagner, A.E., Terschluesen, A.M. & Rimbach, G., 2013. Health promoting effects of brassica-derived phytochemicals: from chemopreventive and anti-inflammatory activities to epigenetic regulation. *Oxidative medicine and cellular longevity*, 2013, p.964539.

Wang, H. et al., 2011. Glucosinolate biosynthetic genes in *Brassica rapa*. *Gene*, 487(2), pp.135–142.

Wang, Y. et al., 2013. *De novo* transcriptome sequencing of radish (*Raphanus sativus* L.) and analysis of major genes involved in glucosinolate metabolism. *BMC genomics*, 14(1), p.836.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), pp.57–63.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2), pp.256–276.

Yan, X. & Chen, S., 2007. Regulation of plant glucosinolate metabolism. *Planta*, 226(6), pp.1343–1352.

Young, A.J., 1991. The photoprotective role of carotenoids in higher plants. *Physiologia Plantarum*, 83(4), pp.702–708.

Chapter 3: The development of a mapping population and linkage map for watercress (*Nasturtium officinale* R. Br.; Brassicaceae) and their application in mapping QTL for traits of crop morphology and consumer health

3.1 Abstract

Watercress (*Nasturtium officinale* R. Br.) is a highly nutritious leafy crop with a long history of cultivation and a growing market due to increasing interest in healthy diets. However, currently there is no breeding programme for this crop aimed at meeting current and future market demands. Next Generation Sequencing (NGS) now allows for rapid expansion of molecular resources in any crop and is particularly useful for understudied species. Here we describe the development of the first 259 F_2 progeny mapping population, the use of Genotyping-by-sequencing (GBS) in marker discovery and the construction of a genetic linkage map. Sequencing produced 940 million reads from which 9,689 loci were assembled and filtered to 561 informative SNP markers. Of these, 321 markers were assembled into a 1,733 cM linkage map with a mean inter-marker distance of 5.8 cM. This was used to identify novel QTL in two separate experiments in contrasting controlled-environment and field conditions. Seventeen QTL were identified for traits of stem length, diameter, number of nodes, mean internode distance, specific leaf area (SLA), antioxidant capacity and cytotoxicity against human cancer cells (IC_{50}, the concentration required to kill 50 % of cells in culture); with the largest effect QTL accounting for 10 % of phenotypic variation in SLA. For cytotoxicity to cancer cells, IC_{50}, we identified three QTL explaining 20 % of phenotypic variation in total. We undertook a bioinformatics analysis, using the publicly available watercress transcriptome and closely-related *Arabidopsis thaliana* (L.) Heynh, which revealed a set of candidate genes underpinning QTL including a variety of cell wall components for stem morphology and a potential defence cluster for IC_{50}. This work represents a significant step forward for this high nutrient dense, leafy crop and will facilitate further research and breeding, targeted on improved traits for growth and health.

3.2 Introduction

Watercress (*Nasturtium officinale* R. Br.) is a perennial semi-aquatic herb and member of the Brassicaceae plant family which contains many important food crops, including oilseed rape, broccoli, kale, cabbage, mustard and also the model research species *Arabidopsis thaliana* (L.) Heynh (Al-Shehbaz et al. 2006). Modern day watercress is often commercially cultivated adjacently to freshwater streams but its use as a medicinal crop dates back 2,000 years (Manton 1935; Palaniswamy & McAvoy 2001). This leafy crop, a component of salads, soups and garnishes, is recognized by its characteristic small pinnate leaves and strong peppery flavour. This flavour is derived from the hydrolysis of gluconasturtiin, the primary glucosinolate (GLS) in watercress, to phenethyl isothiocyanate (PEITC) during mechanical damage to the plant tissue, such as mastication (Macleod & Islam 1975; Fahey et al. 2001). Although the main function of this system is defence against herbivory for the plant (Newman et al. 1996), isothiocyanates have been recognized as potent chemopreventive agents in the human consumer (Traka & Mithen 2008; Wagner et al. 2013).

Poor diet is listed as the most expensive cause of illness to the U.K.'s National Health Service (Scarborough et al. 2011; Elia 2015) and costs the U.S.A. $71 billion annually (Frazão 1999). On the other hand, the role of a diet rich in Brassicaceae plants in reducing the occurrence of choric disease has been discussed extensively (Verhoeven et al. 1996; Higdon et al. 2007; Manchali et al. 2012; Wagner et al. 2013). In watercress, studies have highlighted the health benefits associated with watercress *in vitro*, in animal models and in human studies. Daily watercress intake increased toxin deactivation and reduced human cancer risk, particularly in smokers (Hecht et al. 1995; Gill et al. 2007). Watercress-derived isothiocyanates had a potent chemopreventive effect and even reduced the invasiveness of cancer cell in tissue culture and in mice (Rose et al. 2000; Rose et al. 2005; Yang et al. 2010; Syed Alwi et al. 2010; Cavell et al. 2012; Gupta et al. 2013). The consumption of watercress also increased the concentration of antioxidants (AO) in the blood of human subjects (Gill et al. 2007; Fogarty et al. 2013); had a protective effect against DNA damage in mice and in humans (Casanova et al. 2013; Fogarty et al. 2013) and reduced injury-related inflammation in rats (Sadeghi et al. 2013).

Increased research effort into these health-promoting crops seem more imperative than ever and for many Brassicaceaes, numerous resources and databases have already been developed to inform crop breeding and research (e.g. Snowdon and Friedt, 2004; Shi et al., 2013; Wang et al., 2013; Sotelo et al., 2014; Branham et al. 2017) and to the development of improved varieties is pursued, such as the nutritionally-enhanced Beneforté® broccoli (Traka et al. 2013). For watercress however, research has been focused primarily on biomedical aspects, with no development of pre-breeding resources in this leafy crop.

Watercress is widely thought to be a tetraploid because of the large number of chromosomes observed (2n=4x=32) (Bleeker et al. 1999) and has a genome size of approximately 750 Mbp (Morozowska et al. 2010). A few studies have also imaged watercress samples with 16 chromosomes (2n) (Manton & Howard 1946; Jeelani et al. 2013) suggesting that there may be varying ploidy levels within the species. Cultivated watercress is inbred and lacking in genetic diversity as a consequence of regular selfing, with no provision, globally of new genetically diverse material with any novel traits or crop properties (Palaniswamy & McAvoy 2001; Sheridan et al. 2001). To address this, a global germplasm collection was recently established at the University of Southampton, the watercress transcriptome was sequenced and assembled, and gene expression was compared in germplasm accessions of interest (Payne et al. 2015; Voutsina et al. 2016).

Watercress crop quality is defined by several market standards, including peppery flavour, rich-green colour and size specifications. Improved morphology of the watercress crop without loss to yield is one breeding target. Driven by consumer preference, a leafier crop morphology with less stem thus making the crop more fork-friendly is seen as desirable. At the same time, improve nutritional quality is also desirable, in particular optimized AO capacity, GLS content and chemopreventive potency.

Genetic linkage maps have been utilized extensively in plant breeding to identify molecular markers linked to traits of interest, requiring a mapping population with polymorphic markers in the progeny (Collard et al. 2005). The relevant genomic ‘distance’ of markers is estimated based on the likelihood of their co-inheritance of markers during chromosome recombination events across the generations (Langridge & Fleury 2011). Linking phenotype to genotype can lead to the identification of QTL: hot spots in the genome containing markers that are consistently inherited alongside the phenotypic trait of interest. In recent years, the use of Next Generation Sequencing (NGS) and restriction enzymes (RE) to develop and sequence restricted representation libraries (RRLs) has increased the efficiency and reduced the cost of genome-wide marker production (Davey et al. 2011). To date, two such techniques have been used broadly for genotyping of plant traits: Restriction-site Associated DNA (RADSeq) sequencing (Davey et al. 2010) and Genotyping-By-Sequencing (GBS) (Elshire et al. 2011). Essentially, both methods utilize restriction enzymes to fragment the genome and produced fragments are sequenced. This results in consistent coverage of specific regions of the genome across samples, massively reducing the cost of producing reliable markers (Davey et al. 2011). The number of studies utilizing GBS for linkage mapping and QTL analysis in various plant species is increasing in both major and minor crops, e.g. wheat, blackcurrant, apple, watermelon, oil palm, broccoli, and zucchini (Poland et al. 2012; Russell et al. 2014; Gardner et al. 2014; Lambel et al. 2014; Pootakham, Jomchai, et al. 2015; Branham et al. 2017; Montero-Pau et al. 2017).

In this study, we report on the development of the first mapping population for watercress and the use of GBS to create a polymorphic marker database and construct the first linkage map for this crop based on an F_2 family. In addition, we present QTL analysis results on crop morphology and phytonutrient profile, collected across two trials and discuss the potential relevance of the results to watercress breeding and research.

3.3 Aims and Objectives

The overarching aim of this work is to elucidate the genetic basis of important agronomic and nutritional traits in watercress.

The specific objectives were:

1) To produce a watercress mapping population from parents that differ in traits of interest (particularly crop morphology and phytonutrient profile);
2) To produce genetic marker database through NGS of sufficient density and coverage to describing inheritance patterns in the offspring;
3) To develop the first genetic linkage map for watercress;
4) To compile a database of phenotypic data for the mapping population and use this to identify QTL in the genome.

3.4 Methods and Materials

3.4.1 Plant Material and Trait Evaluation:

3.4.1.1 Production of the mapping population:

Parental lines were selected based on their contrasting morphology and phytonutrient profile, as reported by Payne et al. (2015). Parent A was identified as a line of particular commercial interest with a dwarf morphology and high antioxidant (AO) capacity, both highly valued traits by growers (Payne et al. 2015). Parent B was characterized by longer stems and lower nutritional trait values in both field and controlled conditions (Payne 2011). In addition to being selected as extremes of the phenotypic range, these two lines were also found to have a weaker phylogenetic relationship to each other than other accessions studied based on markers from RNASeq (Chapter 2, Figure S2.1). Therefore, they were marked as the most suitable parents for the mapping population.

Seeds from the parents were germinated in a controlled-environment chamber (23 ºC day/21 ºC night, 18 h day length, and 60 % RH) and planted on John Innes Potting Compost No. 2 (JI2) in 10 cm diameter circular pots. This process was repeated for Parent A at ten day intervals for a month to ensure overlap of flowering in the two accessions. Plants were later enclosed in flowering bags to ensure flower isolation and fed with a 3:1:6 (N:P:K) soluble fertilizer every two weeks. Once flowering commenced, suitable flowers were crossed following standard practices for *Arabidopsis*. Siliques were harvested when dry, seeds were cleaned of debris and stored at -20 ºC. Several F_1 seeds were later germinated, genotyped to confirm a true cross (Appendix B.1), and allowed to self and set seed. Three hundred F_2 seeds from a selfed F_1 individual were germinated, of which a total of 259 F_2 plants survived to maturity and were used in this study. The mapping population structure is depicted in Figure 3.1.

To determine the most suitable mapping approach and software, chromosome counts and ploidy level analysis were kindly conducted by Dr E. Wijnker (Wageningen University and Research, The Netherlands) on early morning fixed F_2 buds (Appendix B.2).

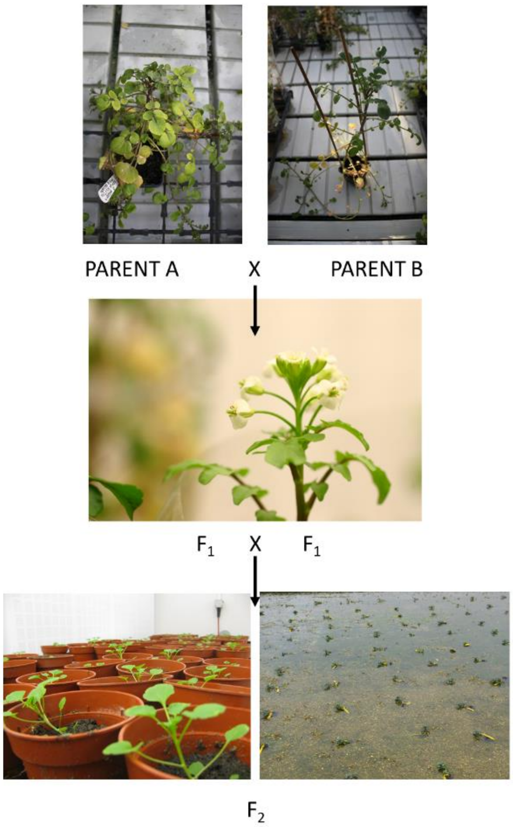


Figure 3.1 The design of the watercress mapping population developed in this work. Inbred parents of contrasting phenotypes were crossed to produce a heterozygous F_1. The F_1 plant was selfed leading to a F_2 progeny of 259 individuals.

3.4.1.2 Phenotype characterization:

3.4.1.2.1 Trials and morphology:

The F_2 population was phenotyped in two environments – a) in a controlled environment chamber and b) on a commercial watercress farm, called ‘Control’ and ‘Field’ respectively. In the Control environment, 259 F_2 plants and five replicates of each parent were grown in 22 °C day/20 °C night temperatures and with 16 h day length, and phenotyped at seven weeks of growth. The length (cm) and diameter (mm, ten cm above the base) of the primary stem was measured and the number of nodes counted. The mean internode distance was used as an inverse indicator of crop ‘leafiness’ and was calculated as the number nodes to the stem length. The second mature leaf (defined as ≥ 20 mm diameter) was photographed, dried to constant weight and its weight recorded. Leaf area was quantified from the images captured using ImageJ (Abràmoff et al. 2004). Specific Leaf Area (SLA, cm^2/g) was calculated as the ratio of leaf area to leaf mass and regarded as an indicator of leaf quality. Tissue was snap-frozen in liquid nitrogen and stored at -80 ºC for AO and cancer cell cytotoxicity assays. Plants were then allowed to flower and set seed. Collected seed was cleaned and stored at -20 °C for future propagation as recombinant inbred lines (RILs).

In July 2015, the population was taken to the Field in order to assess the impact of environment in a relevant commercial setting on the progeny. Triplicate 10 cm cuttings were taken from each of 196 viable F_2s and from replicates of each of the parents, tagged and planted in fine peat (Sinclair, U.K.). A week later, the cuttings were transplanted into a watercress bed located on a commercial farm (50°48’46.8”N, 2°08’47.9”W) in Dorset, U.K. Cuttings were planted within the gravel beds in a random design, each at the centre of a 50 cm x 50 cm square, and with a guard plant layer on the perimeter of the study to reduce any edge effects. After 6 weeks of commercial cultivation, the population was phenotyped. Each plant was photographed, a main stem removed and measurements of length, diameter, number of nodes, leaf area and weight, were taken as previously described. Tissue was snap-frozen for AO and cancer cell cytotoxicity assays. Additional notes were taken as needed to record particularly interesting features, such as flea-beetle damage or distinct crop morphology.

3.4.1.2.2 Antioxidant capacity assays:

AO capacity of each F_2 sample was quantified using the Ferric Reducing Ability of Plasma (FRAP) assay, modified for plant sap (Benzie & Strain 1996; Payne et al. 2013). Frozen plant tissue was ground to powder and homogenized with a pre-chilled mortar and pestle. Sap was extracted from weighed amounts of ground material (~350 g) by centrifuging in QIAshredders tubes (QIAGEN, www.qiagen.com) at 4 °C for 20 minutes (11,000 rpm) and weighed. Three technical replicates of each sample were randomly positioned in flat-bottom 96-well plates alongside a triplicated serial dilution of iron sulphate heptahydrate ranging from 0.5 to 8 mM concentrations and water blanks. The addition of FRAP reagent (300 mM acetate buffer, 10 mM 2,4,6-tripyrid-s-triazine in 40mM hydrochloric acid, and 20mM ferric chloride hexahydrate) produced a colour reaction indicative of the reduction of ferric to ferrous ions, which was read on a spectrophotometer (FLUOstar Optima Microplate reader, BMG Labtech) at 584 nm. AO capacity, as the mmol of ferrous (Fe^{2+}) ion conversion equivalent per gram of fresh plant weight, was extrapolated based on the standard curve of the serial dilutions per plate and the weight of plant material and sap used for each sample.

3.4.1.2.3 Cancer cell cytotoxicity assays:

The cytotoxicity to cancer cells of each F_2 individual was quantified as an IC_{50} value, the inhibitor concentration required to kill 50 % of cells present in culture, using an MTS Cell Proliferation colorimetric assay (Cory et al. 1991; Riss et al. 2004) in a modified version of the protocol used by Cavell et al. (2012). This assay estimates the number of living cells in a cell culture by the addition of tetrazolium [3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl-2H-tetrazolium salt], which metabolically active cells will convert to purple-coloured formazan. The assay does not account for cell division or cell death, both of which would influence the amount of formazan produced, but is a widely-accepted method for estimating the approximate number of living cells (Riss et al. 2004).

In this study, MCF-7 human breast cancer cells (American Type Culture Collection, Manassas, VA, U.S.A.) were seeded in 96-well plates at a density of 2,000 cells per well in 50 µl of Dulbeccos' modified Eagle's Medium, or DMEM (with L-glutamine, high glucose and pyruvate, Thermo Fisher Scientific, U.K.), supplemented with 10 % (v/v) fetal bovine serum (Thermo Fisher Scientific, U.K.). Cells were incubated at 37° C with 10% (v/v) CO_2. Twenty four hours later, the cells were treated with a serial dilution of 0.22 µm filter sterilized (Corning® Costar® Spin-X® 0.22 µm column filter, Merck, U.K.) watercress sap at six concentrations, ranging from 1.6 µl/ml of media to 50 µl/ml of media, and in duplicate from each watercress F_2 sample. Each plate also contained replicates of untreated controls and positive controls, treated with another 50 µl of complete DMEM and 0.5 µM of the kinase-inhibitor staurosporine (Fisher Scientific, U.K.) respectively. Moated 96-well plates (CytoOne, StarLab, U.K.) were used and the intra-well space was filled with phosphate-buffered saline solution (Thermo Fisher Scientific, U.K.) to minimize any edge effect and temperature differences between wells. The cells were then incubated for a further five days. On the fifth day, all DMEM media was aspired off and replaced with 100 µl of Rosewell Park Memorial Institute (RPMI) media without phenol red (Thermo Fisher Scientific, U.K.). Five µl of CellTiter 96® AQ_{ueous} One Solution reagent (Promega, Southampton, U.K.) were then added to each well. Plates were then incubated for 90 minutes and the colour change was read on a plate reader (FLUOstar Optima Microplate reader, BMG Labtech) at 490 nm.

The effect of the watercress sap on the MCF-7 cancer cells was quantified after first eliminating background noise from the data by removing the average value of the blank controls for each plate. For each F_2 and parental sample, survival curves were produced as a percentage of the untreated control cells (100 %). The R package 'drc' (Ritz et al. 2015) was used to fit a log-logistic three parameter dose-response model to each individual and calculate its IC_{50} value (Appendix B.3). Due to the volume of samples, both negative and positive control readings were quality checked to ensure there was no effect of time, cell maturation and plate on the results.

3.4.2 Development of a Linkage Map:

3.4.2.1 DNA extractions and Genotyping-By-Sequencing:

Genotyping-by-sequencing (GBS) was employed to develop novel markers for watercress. Genomic DNA was extracted from leaf tissue using a modified centyltrimethylammonium bromide (CTAB) protocol (Doyle & Doyle 1987). To ensure adequate DNA quality and quantity samples were assessed 1) on a Thermo Scientific NanoDrop 1000 (Thermo Fisher Scientific, U.K.), 2) by gel electrophoresis on a 0.8% agarose gel, alongside a λ HindiIII size/mass standard, and 3) by the Quant-iT™ double-stranded DNA Assay Kit (Thermo Fisher Scientific, U.K.). In addition, ten percent of samples were digested using HindiIII to trial viability of restriction enzyme (RE) digestion. Sequencing library preparation and GBS was completed by the Genomic Diversity Facility (Cornell University, Ithaca, U.S.A.) according to the protocol by Elshire et al. (2011). The RE *PstI* (CTGCA/G) was selected for this project, based on deeper coverage per locus than the other REs tested, and parents were sequenced in triplicate to achieve greater sequencing depth.

3.4.2.2 SNP discovery and linkage mapping:

Stacks was utilized to de-multiplex, quality-check (Q > 10, 90% base call accuracy), and *de novo* assemble the raw GBS reads into loci and then to identify markers (Catchen et al. 2013). *De novo* assembly parameters of m = 5, M =2, n = 1 were applied (m – minimum number of raw reads required, M – number of SNPs allowed within an individual, n – number of SNPs allowed for loci to be merged). The 'genotype' pipeline in Stacks was then used to produce a comprehensive SNP catalogue. As there in no physical map in watercress for comparison, a strict filtration of markers was applied to ensure an accurate map was produced and reliable, robust QTL could be detected. Catalogued markers that were either monomorphic or multi-allelic in the parents were excluded. Linkage mapping was performed in R/qtl (Broman et al. 2003) and markers were further quality checked through this package for: individuals with large amounts of missing data (< 50 markers, 10%), likely genotyping errors, duplicated individuals, or individuals with irregular number of crossover events, and markers with high levels of segregation distortion (χ^2 test for a 1:2:1 ratio, P <1e^{-5}). The final step also acted as a filter to the majority of markers with more than 20% missing data because missing data was the primary cause of segregation distortion in this dataset. The functions available in R/qtl were fully utilized to explore linkage between markers, form linkage groups with various combinations of logarithm of odds (LOD) scores and maximum recombination fractions, compare alternative markers orders, and drop problematic markers that introduced questionable gaps in linkage groups. The final linkage map was constructed with a minimum LOD score of 5 and maximum recombination fractions of 0.25.

3.4.3 Phenotype and QTL Analysis:

For both Control and Field plants, the normality of data distribution for each trait was inspected by frequency histograms, QQ plots and calculation of skewness and kurtosis, and transformations were applied as needed. Data was further explored for interesting trait relationships by calculating Spearman's rank correlation coefficients and by Principle Component Analysis (PCA). To help estimate whether QTL might be detectable for each trait, the replicated Field trial was used to calculate heritability of traits for this population in two ways. Firstly, using Analysis of Variance (ANOVA), where the F statistic represents the ratio of the variance between the groups (genetic component) to the variance within the groups (residual variance, or environmental component). Therefore, an F statistic that is >1 and a significant p value would suggest a large role of the genetic component in explaining overall phenotypic variance in the trait. In addition, broad-sense heritability (H^2) was also calculated as the variance explained by the genetic component over the overall phenotypic variance. Finally, transgressive segregation was tested for by a one-way ANOVA and Tukey's Honest Significant Difference test for the five highest and five lowest values of each trait against each of the five parental replicates. All statistical analysis and visualisation was completed in R (R Core Team 2013).

QTL analysis completed for each trait using R/qtl (Broman et al. 2003) separately for the two trials. Composite Interval Mapping (CIM) was used to detect QTL with a walk speed of 1 cM, a maximum of 5 cofactor markers, and a scan window size of 10 cM. Data distribution for the number of branches was not improved by transformation therefore this trait was analysed using the non-parametric model of single QTL mapping in the R/qtl function *scanone*. For each trait/environment, 1,000 permutation tests were carried out to estimate a genome-wide significant LOD threshold score. QTL with LOD scores above 10% threshold ($P \leq 0.1$) are reported. The QTL additive and dominance effects, % phenotypic variation (% R^2) explained and 1.5 LOD support intervals were also calculated for each significant peak. The linkage map and significant QTL were graphically represented using MapChart 2.3 (Voorrips 2002). The highest and lowest ten F_2 genotypes were examined using Flapjack genotype visualizer to further explore the structure of the QTL (Milne et al. 2010). The sequences of markers within the 1.5 LOD support intervals were checked for similarity to known plant genes using the online NCBI BLASTn tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi) against the nucleotide database (nr/nt) with default search settings (Altschul et al. 1990). In addition, the markers were also aligned locally to the watercress transcriptome (Voutsina et al. 2016) using BioEdit Sequence Alignment Editor version 7.2.5 (Hall 1999) through a BLASTn search with a cut-off value of e^{-30}.

Figure 3.2 visualizes the data collection and analysis described in this Methods section.

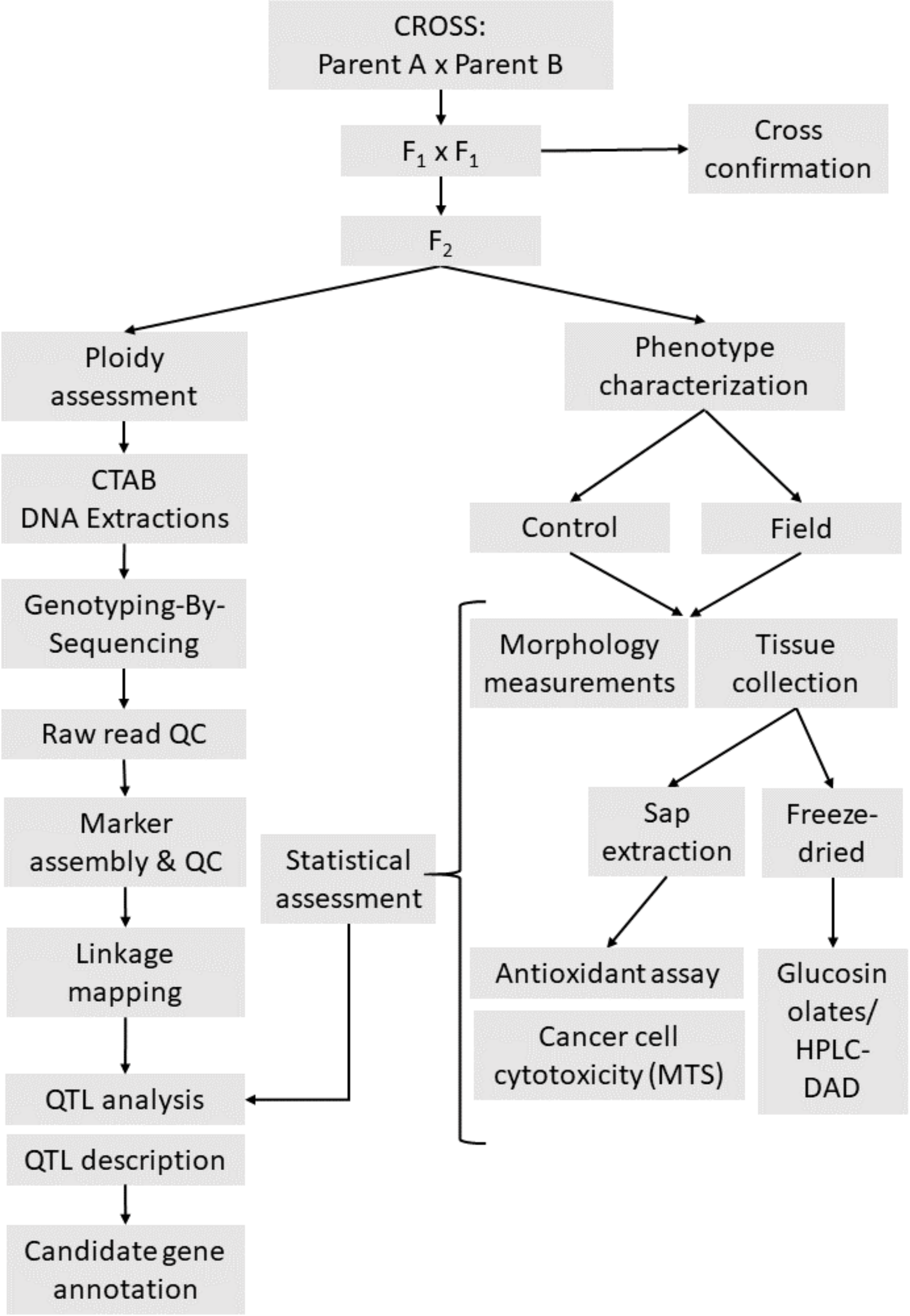


Figure 3.2 Flowchart of data collection and analysis procedures undertaken in the development of a linkage map and QTL for watercress. An confirmed initial cross was taken forward to F_2 population of 259 offspring which were genotyped and

phenotyped towards the development of markers, a linkage map and QTL discovery for traits of morphology and phytonutrient content.

3.5 Results

3.5.1 Ploidy Assessment:

Chromosome imaging and ploidy analysis concluded that the mapping population has 32 chromosomes and is either a diploid (2n = 32), or an allotetraploid (2n = 4x = 32) which behaves like a large diploid (Appendix B.2). Thus, linkage mapping and QTL analysis could be approached with diploid recombination assumptions.

3.5.2 Marker and Linkage Map Development:

A total of 940,208,464 reads were returned from GBS sequencing of 280 samples, with a range of 50,000 to 4,100,000 raw reads per sample. Of these reads, 93.3 % were retained past the initial quality check (Q > 10). A total of 9,689 loci were assembled by the Stacks genotype pathway. However, 87.4 % (8,464) loci were discarded as ‘consensus’ across the population and, thus, not useful for mapping. An additional 6.6 % (666) markers were dropped because they were multi-allelic in one or both of the parents, suggesting either complex repetitive regions that had been grouped together in *de novo* assembly or sequencing errors. The remaining 561 markers (6 % of the original total) were taken forward (Genotype coverage per marker and individual shown in Appedix B.6). For this marker data set, SNP type frequencies are shown in Table 3.1 with the most common nucleotide substitutions being A/G and C/T.

A final filtration was applied in R/qtl during the linkage map construction to remove markers with significant levels of segregation distortion and missing data which might bias the linkage relationship between markers. Since missing data and consensus loci limited the number of markers discovered, 65 markers with >20% missing data were retained in the dataset because they did not exhibit segregation distortion and appeared to be reasonable candidates for mapping. Linkage groups were formed with a required max recombination fraction of 0.25 and minimum LOD score of 5, producing 35 linkage groups (Appendix B.7). Small linkage groups containing less than 3 markers were dropped. The final map grouped 321 markers into 21 linkage groups with an overall length of 1733.2 cM and mean marker spacing of 5.8 cM. Table 3.2 contains detailed descriptions of the linkage groups and overall map.

Table 3.1 SNP type number of occurrences and percentage overall for marker dataset produced from this mapping population.

SNP	Occurrence	Percentage (%)
A/C	65	10.0
A/G	177	27.2
A/T	81	12.4
C/G	37	5.7
C/T	146	22.4
G/T	55	8.4

Table 3.2 The number of markers, length in cM, mean and maximum marker spacing in cM for each linkage group assembled and for the map overall.

Linkage group (LG)	Number of markers	Length (cM)	Mean marker spacing (cM)	Max marker spacing (cM)
1	32	151.2	4.9	21.4
2	29	160.2	5.7	15.5
3	27	173.1	6.7	22.4
4	25	75.4	3.1	16.4
5	23	170.6	7.8	26.1
6	21	147.7	7.4	24.1
7	20	87.6	4.6	19.7
8	20	119.4	6.3	18.2
9	19	70.1	3.9	11.3
10	15	92.7	6.6	21.1
11	13	136.1	11.3	24.6
12	11	31	3.1	8.3
13	10	39.4	4.4	17.9
14	10	45.1	5	11.8
15	9	38.9	4.9	11
16	7	25.2	4.2	8.7
17	7	52.1	8.7	17
18	6	22.6	4.5	16.1
19	6	36.4	7.3	16
20	6	43.9	8.8	20.6
21	5	14.6	3.6	5.2
Total	321	1733.2	5.8	26.1

3.5.3 Phenotype and QTL Analysis:

Figure 3.3 shows the frequency distribution for morphological traits across the F_2 family in the two trials respectively. Figure 3.4 illustrates the frequency distribution of phytonutrient traits in both trials. The morphology of the parental accessions ranked overall as predicted from previous research with the ‘dwarf’ Parent A having shorter and thinner stems with shorter mean internode distance (Figure 3.5). The parental phenotypes for AO capacity did not differ strikingly and, although, Parent A was the most potent cancer cell inhibitor in the Control environment, this was not the case in the Field where Parent B was in fact the more potent of the two (Figure 3.5).

F_2 plants morphology differed between the two experiments, as might be expected due to developmental stage and environment. Specifically, as Control plants were grown directly from seed, they consisted of one obvious primary stem and occasional branching, whereas plants grown in the Field were propagated from cuttings of the Control plants and grew out from the central cutting with reduced apical dominance. SLA, a commonly-used indicator of leaf quality, was much lower in the Field (mean ± SD: 141.8 ± 64, cm^2 /g) than in the Control plants (500 ± 115.7), suggesting a thicker leaf in the field experiment The AO capacity of the Field samples (mean ± SD: 770.4 ± 116.4, Fe^{2+} equivalent per g fresh weight) was notably higher than the Control samples (267.9 ± 78.27), whereas IC_{50} values were within a similar range in both trials (Figure 3.4) and comparable to values quantified for watercress with a similar protocol by Cavell (2012). Transgressive segregation was visually apparent in most traits, with the phenotypic range of F_2 progeny reaching beyond that exhibited by the parents (Figures 3.3 and 3.4). The difference between each parental and F2 high or low group was statistically significant in most cases (Table 3.3), suggesting transgressive segregation was indeed achieved for all traits examined here.

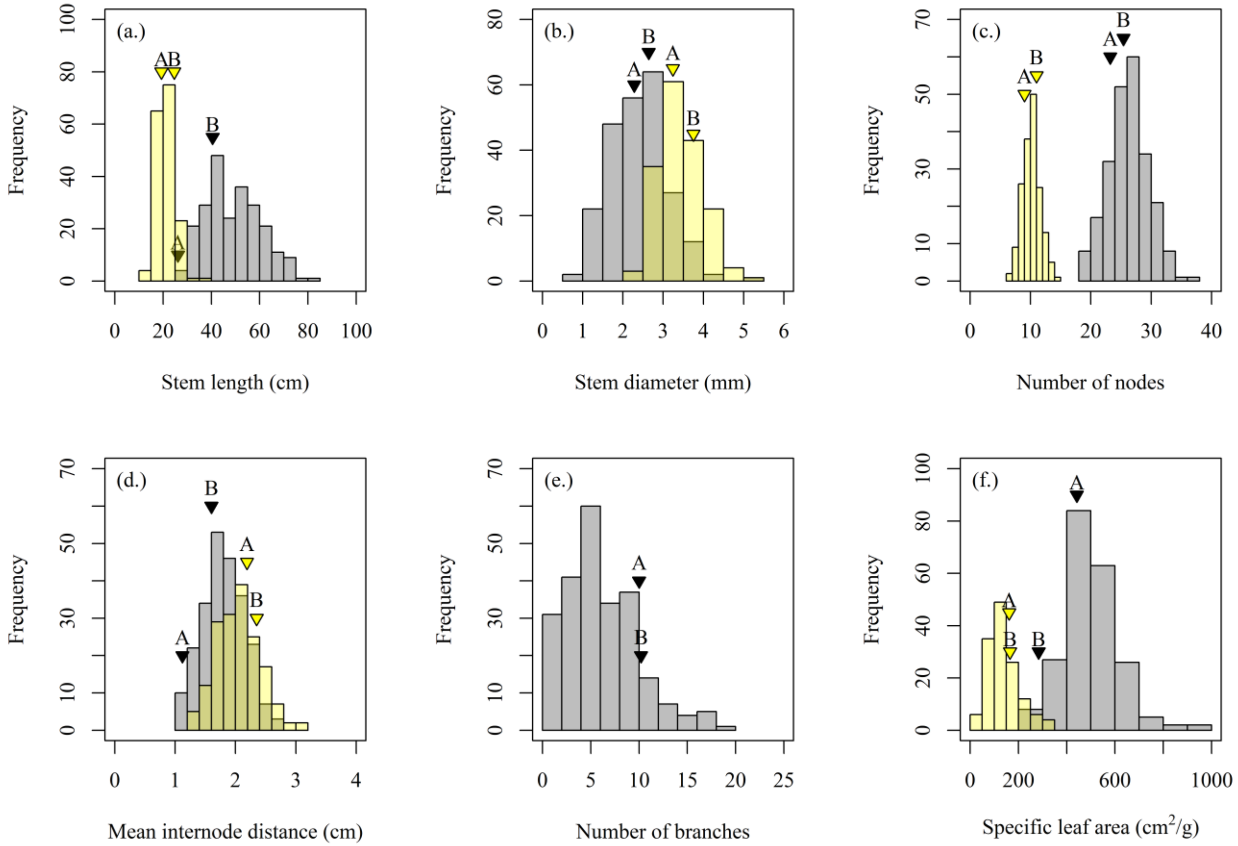


Figure 3.3 Stem length (a, cm), stem diameter (b, mm), number of nodes (c), mean internode distance (d, cm), number of branches (e) and specific leaf area (f, cm^2/g) in F$_2$ plants grown under Control (grey, *n* = 1) and Field (yellow, *n* = 3) conditions. Parental means (*n* = 5 & *n* = 15 respectively) are displayed as ‘A’ for Parent A and ‘B’ for Parent B.

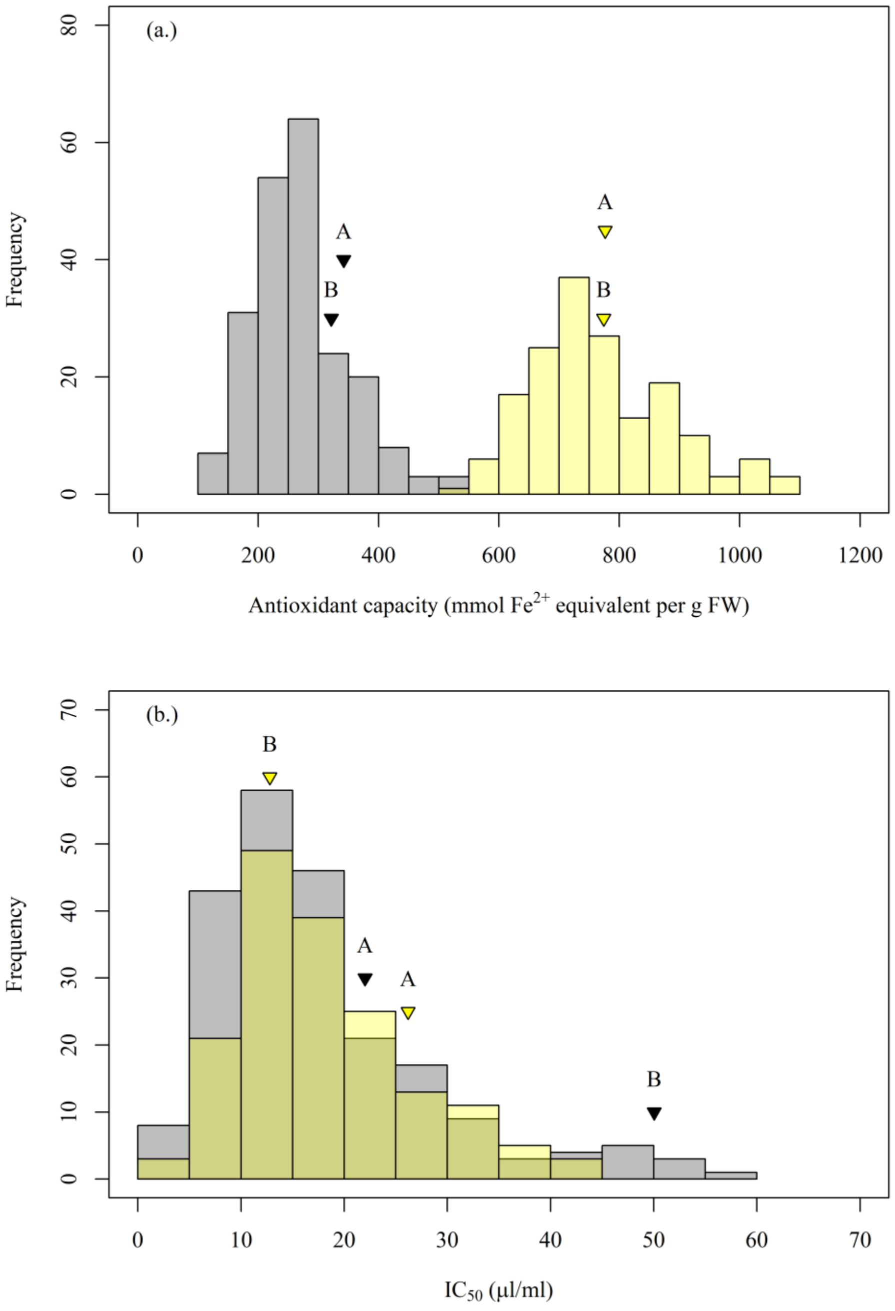


Figure 3.4 Phytonutrient phenotypes in both trials: (a.) antioxidant capacity as mmol Fe^{2+} equivalent per g of fresh weight and (b.) IC_{50} of sap to cancer cells, for F_2 plants grown under Control (grey, *n* = 1), and Field (yellow, *n* = 3) conditions. Parental means (*n* = 5 & *n* = 15 respectively) are displayed as ‘A’ for Parent A and ‘B’ for Parent B.

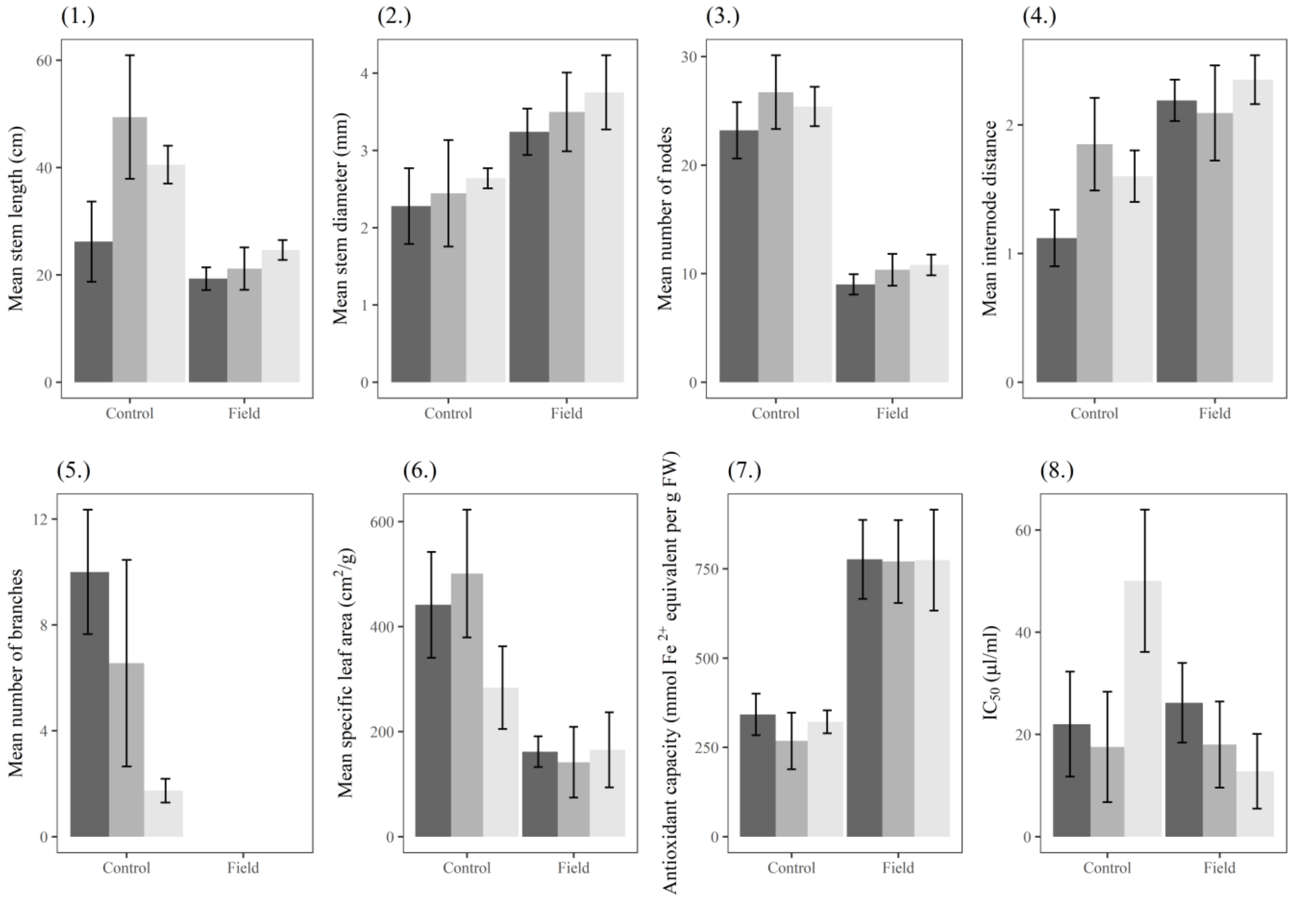


Figure 3.5 The mean (± standard deviation) phenotypic value for Parent A (dark grey), F_2 samples (intermediate grey) and Parent B (light grey) for stem length (1, cm), stem diameter (2, mm), number of nodes (3), internode distance (4, cm), number of branches (5), specific leaf area (6, cm^2/g), antioxidant capacity of tissue (7, mmol Fe^{2+} equivalent per g of fresh weight) and inhibitor concentration for a 50 % cancer cell kill (8, µl/ml) in Control and Field experiments.

Table 3.3 Significance levels of transgressive segregation in the Field and Control trials of highest and lowest ranging phenotypes recorded for the F_2 population in comparison by ANOVA and Tukey's HSD testing to the parental replicates (asterisks indicate the level of significance: * ≥ 0.01, ** ≥ 0.001, *** ≥ 0.0001, **** < 0.0001).

Control

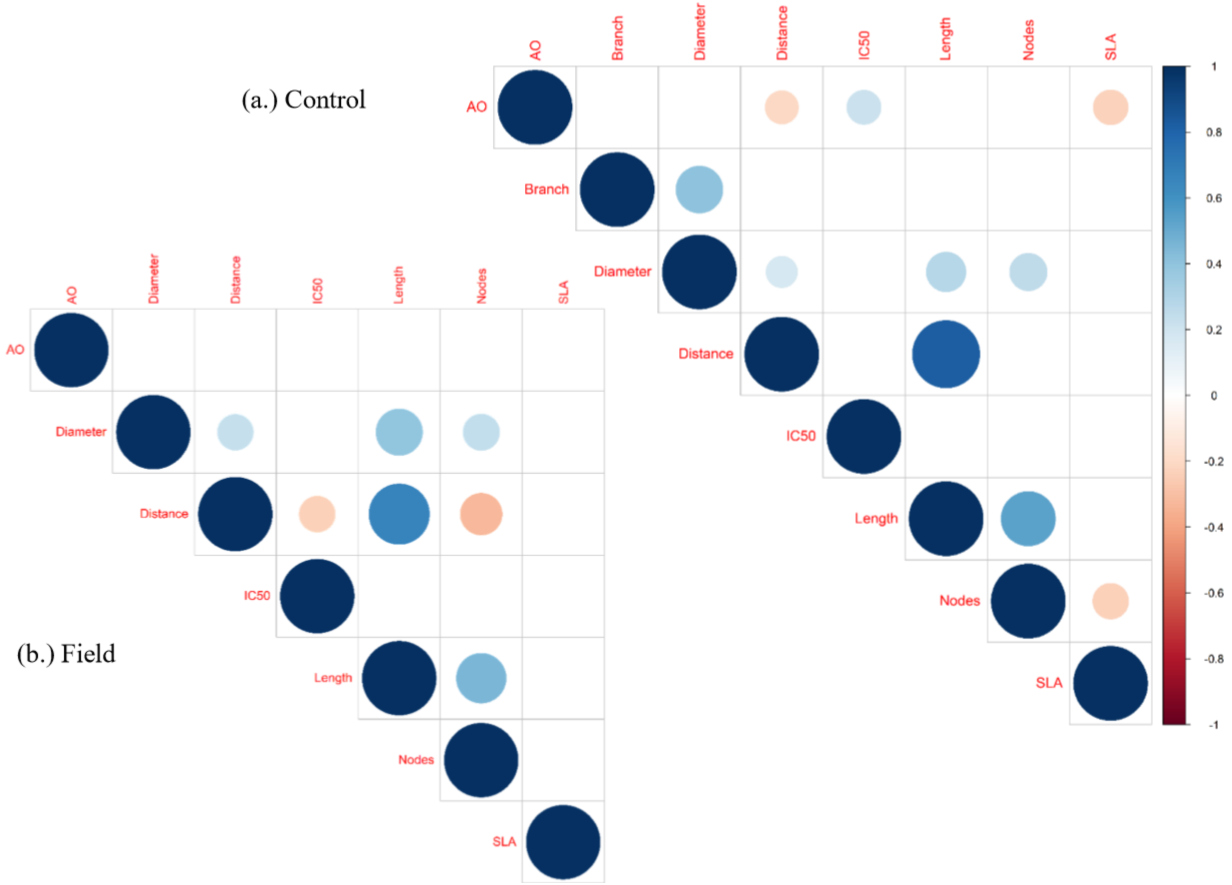
Trait	Phenotype	Parent A		Parent B	
Stem length	High	0	****	0	****
	Low	0.98	ns	0.0018	**
Stem diameter	High	0.0000018	****	0.0000214	****
	Low	0.00013	***	0.000009	****
Number of nodes	High	0.0000001	****	0.0000021	****
	Low	0.0058424	**	0.0001188	***
Mean internode distance	High	0	****	0	****
	Low	0.99	ns	0.0005	***
Antioxidant capacity	High	0.0000239	****	0.0000012	****
	Low	0.0000015	****	0.0000009	****
IC50	High	0.0002476	***	0.9695257	ns
	Low	0.0152542	*	0.0000013	****

Field

Trait	Phenotype	Parent A		Parent B	
Stem length	High	0.0000012	****	0.0004938	****
	Low	0.0116637	*	0.0000151	****
Stem diameter	High	0.0000076	****	0.0005945	***
	Low	0.0094417	**	0.0000859	****
Number of nodes	High	0.0000001	****	0.0000546	****
	Low	0.009529	**	0.0000083	****
Mean internode distance	High	0.0000017	****	0.00003	****
	Low	0.0000003	****	0	****
Antioxidant capacity	High	0.0033532	**	0.0031476	**
	Low	0.0205754	*	0.0221442	*
IC50	High	0.0040816	**	0.0000031	****
	Low	0.000038	****	0.0764067	ns

Spearman's rank correlation between traits was examined in each environment and results are show as heatmaps in Figure 3.6. For both environments, positive correlations existed between many morphology traits. Specifically, plants with larger stems had larger stem diameter and a greater number of nodes. Correlations with phytonutrient traits were not as consistent across environments. In the Control trial, greater AO capacity was negatively correlated with mean internode distance and specific leaf area but positively correlated with IC_{50} potency. However, in the Field trial, there were no significant correlations to AO capacity and IC_{50} data was negatively correlated to mean internode distance.

Principle component analysis (PCA) was completed on trait data from each trialled environment separately. For the Control conditions, Principle Component (PC) 1 explained 31.5 % of variation in the dataset and was primarily comprised of two linked traits: stem length and mean internode distance, PC2 explained 23 % of variation in the dataset was composed of SLA and AO (Figure S3.3, Appendix B.4). In the Field data, PC1 explained 29.6 % of data variation and its strongest contributions came from, again, stem length and mean internode distance but also stem diameter (Figure S3.4, Appendix B.5). The second PC explained 20.3 % of variation and was made of the number of nodes and internode distance, which are linked traits.


(a.) Control

(b.) Field

Figure 3.6 Heatmaps illustrating significant correlations (Spearman’s) between traits in the (a.) Control and (b.) Field trials. The blue and red colours signify positive and negative correlations respectively and the size of the circle relates to the significance of the correlation.

An ANOVA and broad sense heritability (H^2) were used to determine whether variation in the replicated Field data was better explained by the genetic or environmental component, thus indicating whether it is likely to detect any QTL for this population and this trial (Table 3.4). For all traits, the F_2 line was a significant factor ($p \leq 0.05$) and H^2 was greater than 0.5, suggesting a moderate to high role of genetic component in explaining phenotypic variation and suggesting that QTL should be detectable. Heritability was lowest for AO capacity and number of nodes and highest for stem length and IC_{50}.

Table 3.4 Broad-sense heritability values for traits measured in the Field environment: the F ratio and p value of an ANOVA comparing the amount of variance explained by the genetic component versus the environmental component (asterisks indicate the level of significance: * ≥ 0.01, ** ≥ 0.001, *** ≥ 0.0001, **** < 0.0001), and broad-sense heritability (H^2).

Trait	**F-ratio (V_g/V_e)**	***P***		**H^2**
Stem length	1.91	4.02E-08	****	0.66
Stem diameter	1.52	0.000266	***	0.60
Number of nodes	1.23	0.045	*	0.55
Mean internode distance	1.40	0.00304	**	0.58
Specific leaf area	1.43	0.00339	**	0.59
Antioxidant capacity	1.31	0.0153	*	0.57
IC_{50}	1.86	5.88E-07	****	0.65

QTL analysis, for the traits described in this mapping population, was completed using Composite Interval Mapping (CIM) in R/qtl separately for each environment (models and QTL effect plots included as Appendices B.8-B.21). Following 1,000 permutations per trait/environment, 17 significant QTL identified at 10 % genome-wide LOD threshold and 12 at a 5 % LOD threshold (Table 3.5). The effect of each QTL was also quantified (Table 3.6). A single QTL explaining from 1.6 to 10 % of variance within a trait and jointly QTLs explained 6.6 to 20 % of phenotypic variation within a trait. The majority of significant markers exhibited an additive effect over the trait in question. Overall, QTL were independently located except for LG 16 were a QTL for number of nodes collocated with a QTL for IC_{50} cytotoxicity (Figure 3.7) and with internode distance QTL *dis16* separately.

Table 3.5 QTL peaks and descriptive statistics, including the QTL name, 5 % genome-wide LOD threshold for significance, linkage group, position in cM, QTL peak LOD score and significance, 1.5 LOD support interval in position (cM) and in flanking marker names.

Trait	Environment	QTL name	5% LOD threshold	Linkage group	Position (cM)	QTL peak LOD	P value	1.5 LOD support interval (cM)	Flanking markers
AO	Field	ao1	3.73	1	51	3.58	0.064	45 - 56	13593 - 12741
Branch	Control	brnp6	3.46	6	75.2	3.36	0.067	47.76 - 81.7	4509 - 24452
Diameter	Control	dmt10	3.64	10	53	5	0.003	52 - 63	3576 - 11145
Diameter	Field	dmt13	3.73	13	8	3.63	0.059	0 - 9	11434 - 21365
Distance	Control	dis11	3.69	11	89	4.23	0.018	84 - 95	4904 - 4228
Distance	Field	dis16	3.76	16	14	6.68	0.001	10 - 19	24756 - 4131
Distance	Field	dis17	3.76	17	9	4.91	0.008	7 - 18	14068 - 12241
IC50	Field	ic2	3.85	2	141	3.5	0.09	48 - 146	13347 - 19822
IC50	Field	ic11	3.85	11	50	4.31	0.016	49 - 60	5068 - 21621
IC50	Field	ic16	3.85	16	25.2	3.72	0.062	20 - 25	6992 - 13595
Length	Control	lgth4	3.8	4	75.4	4.56	0.013	70 - 75.42	13196 - 13988
Length	Field	lgth6	3.78	6	131	4.13	0.015	127 - 136	18842 - 13758
Length	Field	lgth7	3.78	7	15	4.18	0.024	19 - 59	14902 - 10627
Nodes	Control	nd3	3.69	3	173	4.77	0.006	141 - 173.1	2472 - 5677
Nodes	Field	nd16	3.69	16	25.2	3.77	0.043	20 - 25.2	6992 - 13595
SLA	Field	sla1	3.67	1	17.5	3.88	0.026	13 - 22.6	20675-24040
SLA	Control	sla5	3.67	5	78	7.8	0	70 - 81	24266-20888

Table 3.6 The effect of the underlying marker of each significant QTL peak detected in this study: the proportion of phenotypic variance explained by the QTL (% R^2), the additive and dominance effect, and the parent sourcing the favourable allele for that trait.

Trait	Environment	QTL name	Marker	% R^2	Additive effect	Dominance effect	Favourable allele
AO	Field	ao1	20152	6.8	28.3	-45.6	Parent B
Branch	Control	brnp6	5629	6.6	1.1	-1.1	Parent A
Diameter	Control	dmt10	24070	7.4	-0.2	0.2	Parent B
Diameter	Field	dmt13	14771	5.8	-0.2	0.1	Parent B
Distance	Control	dis11	2092	7.2	0.1	0.0	Parent A
Distance	Field	dis16	6992	9.0	-0.1	0.1	Parent B
Distance	Field	dis17	18458	8.1	0.2	0.0	Parent A
IC50	Field	ic2	24736	7.8	-0.4	0.0	Parent B
IC50	Field	ic11	5068	5.0	0.3	0.3	Parent A
IC50	Field	ic16	13595	6.7	0.3	0.1	Parent B
Length	Control	lgth4	13988	5.3	-3.6	1.3	Parent B
Length	Field	lgth6	20669	8.5	-1.4	-1.5	Parent B
Length	Field	lgth7	14902	3.0	-1.0	0.2	Parent B
Nodes	Control	nd3	5677	1.6	-0.5	-0.3	Parent A
Nodes	Field	nd16	23478	6.6	0.4	-0.4	Parent A
SLA	Field	sla1	13308	8.9	-1.0	0.7	Parent B
SLA	Control	sla5	7475	10.0	-1.1	-0.1	Parent B

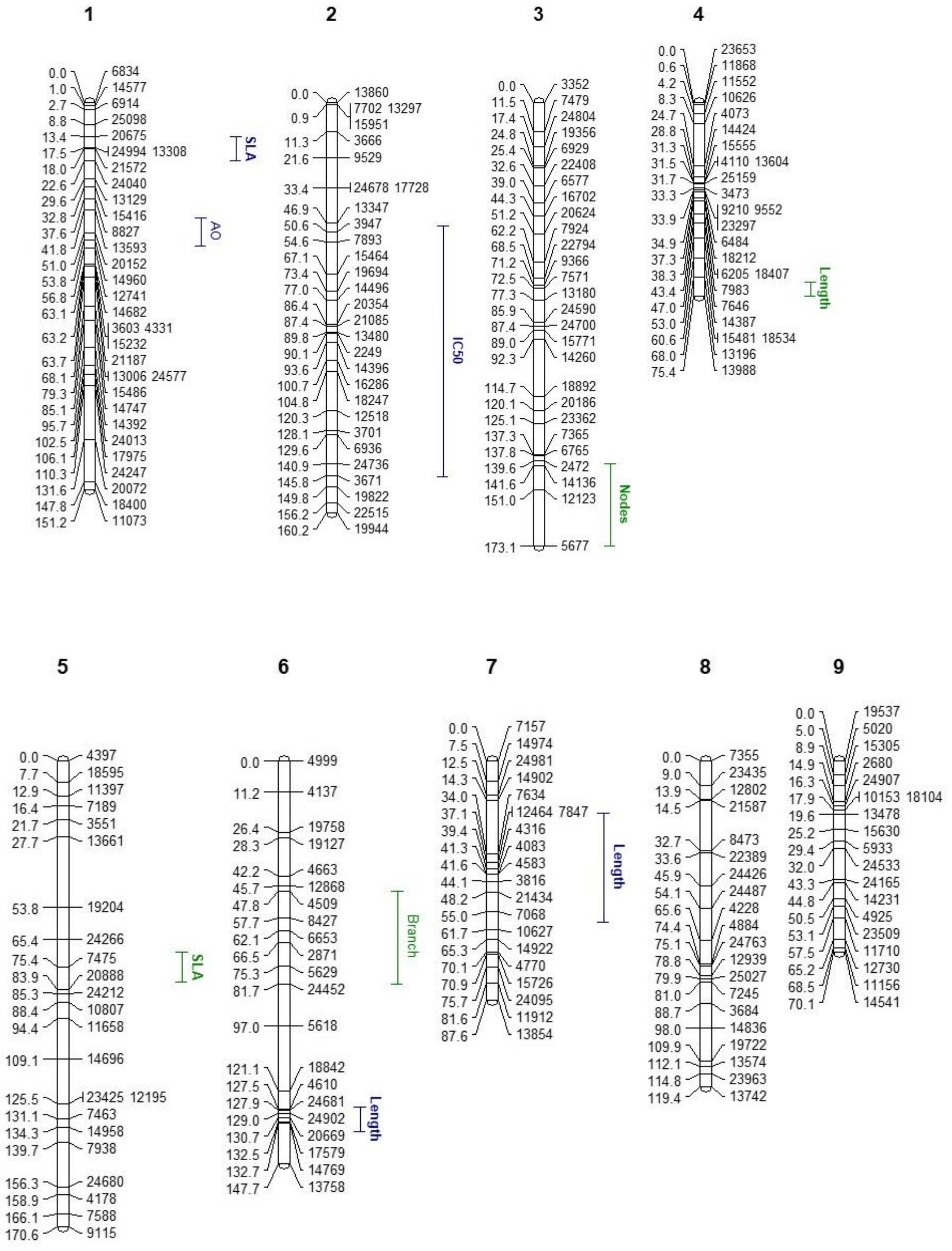


Figure 3.7 Visualization of the significant QTL (1.5 LOD support interval) identified in this study. Phenotyping for numerous traits was carried out in two locations: a controlled environment (Control) in green and a watercress farm (Field) in blue. Continued on adjacent page.

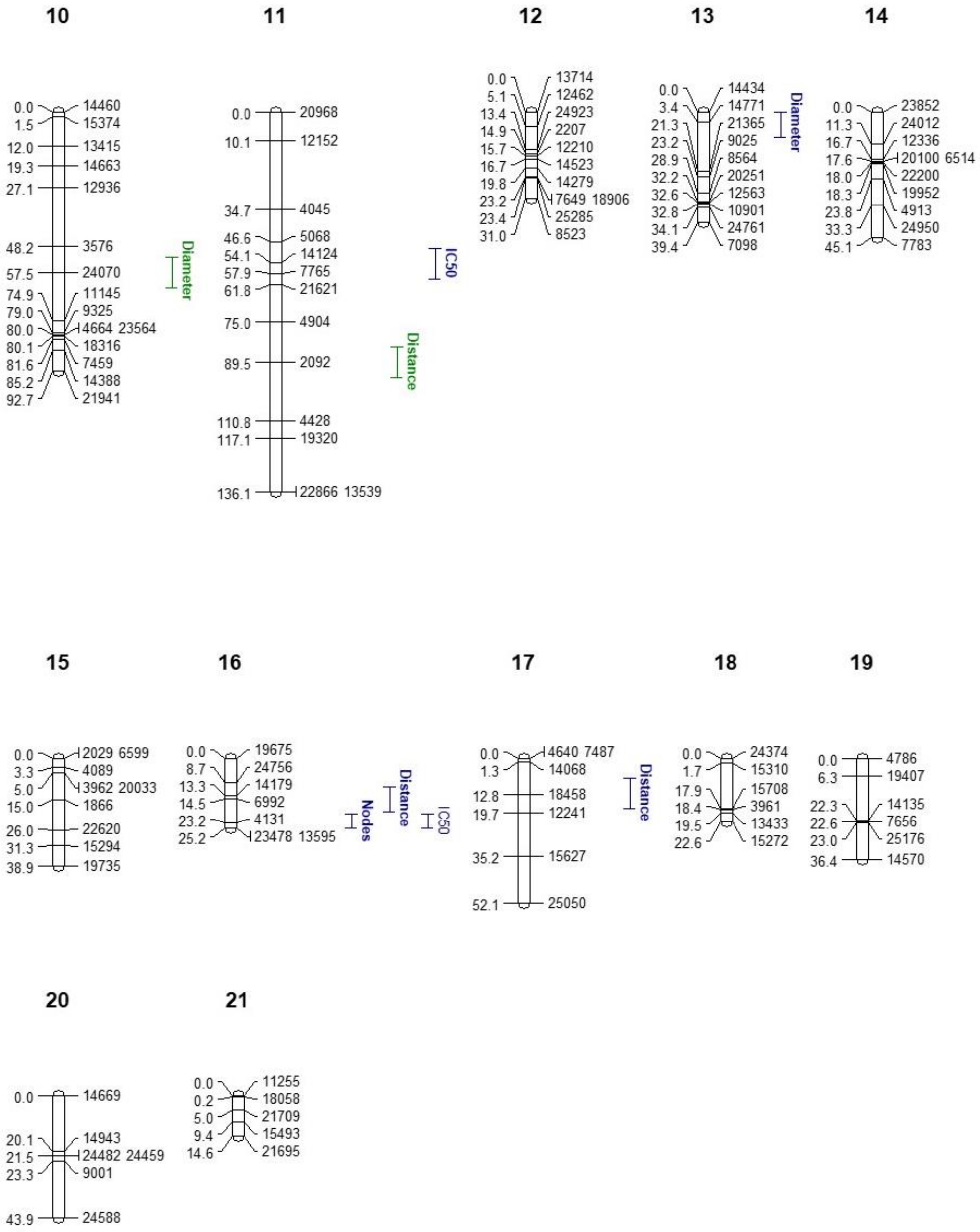


Figure 3.7 Continued on previous page

Watercress does not have a sequenced genome, however the transcriptome was sequenced and *de novo* assembled in Chapter 2 and it is related to many well-studied Brassicaceae species. The markers underlying each significant QTL peak where compared with the watercress transcriptome and the NCBI nucleotide collection (nr/nt) database to check for any similarity with known genes of relevant function. Of these markers, 55 % had a close match in either or both of the resources, suggesting a marker in an expressed part of the genome (Appendix B.22, Table S3.2).

3.6 Discussion

Watercress is a peppery leafy crop with a global market and a wide range of associated health benefits, including chemopreventive, antioxidant (AO) and anti-inflammatory properties (Hecht et al. 1995; Rose et al. 2000; Gill et al. 2007; Fogarty et al. 2013; Sadeghi et al. 2013). For the first time in watercress, we applied Genotyping-by-sequencing (GBS) producing the first genetic linkage map for watercress. From this, we identified 17 QTL of interest for further study in a range of morphological and nutritional traits, including cytotoxicity to cancer cells.

3.6.1 SNP Discovery and Linkage Mapping:

The discovery of single nucleotide polymorphisms (SNP) by the sequencing of reduced representation libraries (RRLs) has enabled the study of genetic variation in species where no previous marker data or reference genome existed (Varshney et al. 2009; Langridge & Fleury 2011). In this study, a total of over 940 million reads were produced for 280 samples utilizing GBS technology. Despite this extraordinary number of reads produced, under 10,000 loci were *de novo* assembled due to large amounts of missing data per individual (Appendix B.6, Figure S3.5). The GBS issue of patchy coverage across individual samples and loci has been discussed previously and attributed to inconsistent amounts or quality of DNA in samples, PCR bias to GC content or particular fragment sizes, or potentially the selection of an unsuitable restriction enzyme (RE) for the particular project (Ward et al. 2013; Li et al. 2014). Ward et al. (2013) suggested this issue may be overcome by sequencing better quality DNA, greater depth of coverage per sample, or using a rare cutter RE to increase coverage per locus. In this study, DNA extractions were optimized for purity and two extractions were made per sample to ensure adequate quantity. Digest effectiveness was tested according to the protocol of the sequencing company and the samples were approved as a candidate for successful GBS. The RE used, *Pst1*, is a rare cutter and was selected out of three tested by the sequencing company as the enzyme which would produce the most coverage per locus, although possibly less SNPs in total. Through these efforts, the project was optimized as effectively as possible, and although parents were sequenced in triplicate to ensure maximum coverage, large amounts of missing data still plagued the dataset. The final suggestions of sequencing each individual F_2 in multiples would impact greatly on the cost-effectiveness of the project, and hence some of the appeal of the protocol and feasibility of the project. Thus, it is possible that GBS does not deliver the vast amounts of markers expected for every project. Despite these limitations, we were still able to produce a set of high quality new resources for watercress with relative efficiency.

Markers with an AA x BB configuration in the parents were used in this project, as parental accessions were expected to be inbred. It became evident in marker development stages that the parents were not separated by great genetic divergence as 87.4% loci assembled were monomorphic across the parents and F_2 progeny, despite being the farthest removed accessions available. This is in line with observations by Sheridan et al. (2001) of overall low genetic diversity within commercial watercress.

In light of the low numbers of SNPs available, 65 SNPs with missing data in a large number of the samples were maintained in the analysis (missing in > 20% of the samples). This cut off has varied in recent literature from 10 – 80%, for GBS-based linkage map construction in crops without a reference genome, but it is becoming regular practice to filter out SNPs missing in more than 50% of the samples (Marchese et al. 2016; Velmurugan et al. 2016; Hussain et al. 2017; Saxena et al. 2017; Yang et al. 2017; Jo et al. 2017; Goonetilleke et al. 2018). The inclusion of large amounts of missing data in GBS-based analyses has an effect on downstream results, namely making the assumption that the samples are homogenous because the data is lacking where indeed polymorphic sites exist, and in this particular study this bias would affect the detection and power of existing QTL. Imputation of missing genotypes based on the position of markers in the linkage map is an option, however, imputation of large numbers of missing genotypes without an accurate reference genome, from which the position of a marker can be extracted in relation to known genotypes, can also introduce bias to the data (Fu 2014). Wickland et al. (2017) suggest that when an approach is sensitive to polymorphism and there is low coverage, as is often the case in such studies as workers tend to opt for higher number of samples over depth of sequencing, then it is inevitable that the amount of missing genotypes is high and the best approach is to maximize the number of SNPs produced. The authors also recommend using multiple GBS data pipelines in SNP calling in order to produce the best quality SNPs, this may have been an approach that could optimize the results of this study.

Despite these limitations, a good quality linkage map was produced using 321 polymorphic markers, grouped into a 21 linkage groups with an average marker spacing of 5.8 cM and a total length of 1,733 cM (Table 3.2). No previous attempt at a linkage map for watercress exists and this mapping population and linkage map constitute important new resources for this understudied crop.

3.6.2 The Phenotype of the Mapping Population:

The parental accessions of this population were grown in multiple replicates alongside the F_2 progeny in both trials. Based on previous characterization of these accessions (Payne et al. 2015), Parent A was expected to display a ‘dwarf’ phenotype with shorter stems and higher phytonutrient content than Parent B. Means of phenotypes for traits related to morphology followed expected trends however this was not the case for AO capacity and IC_{50} values (Figure 3.5). The mean AO value of the parents were very similar in both trials (Figure 3.5.7). And, although the parents ranked as expected for IC_{50} value in the Control trial and mean values for Parent A did not vary greatly between trials, samples from our expected ‘low phytonutrient’ Parent B producing a more potent cytotoxic effect on cancer cells in the Field than did Parent A (Figure 3.5.8). It would appear that environmental conditions in the Field induced a very different phenotype in Parent B, which was not triggered in the invariable conditions of the controlled environmental chamber. These results would suggest a strong environmental component to the phenotype of this crop and that further evidence is needed to secure a consistent phenotypic classification or a reaction norm for these accessions (El-Soda et al. 2014); particularly in light of the low levels of genetic diversity identified between the seemingly divergent parents used here.

This is not the first QTL study to find low trait heritability and a large G x E effect specifically in secondary metabolites. Similar patterns have been reported for antioxidant compounds in oilseed rape (Marwede et al. 2005), soybean (Li et al. 2010), tomatoes (Rousseaux et al. 2005), and lettuce (Hayashi et al. 2012). Difference in phytonutrient potency of plants across environments has been noted before in watercress and reflects phenotypic plasticity of the species and critical role of environmental stimuli in the production and accumulation of secondary metabolites in watercress (Payne et al. 2015). In fact, AO capacity shifted dramatically across the entire F_2 progeny with Control mean ± SD value of 267.9 ± 78.27 mmol of Fe^{2+} equivalent per g fresh weight to 770.4 ± 116.4 mmol Fe^{2+} equivalent per g fresh weight in the Field (Figure 3.4.a). The AO trait had the lowest heritability and only one QTL with a low LOD score. A much smaller shift to higher potency values occurred in the IC_{50}

range of Field samples when compared to the Control samples (Figure 3.4.b), where heritability was higher and three significant QTL explained 20 % of phenotypic variance in the Field trial, suggesting the size of this effect is trait-specific (Tables 3.3 & 3.5). However, differences in the quality of traits between Field and Control environments extended to morphology traits as well. In the majority of traits examined in this study, the phenotypic range was much narrower in the Field than in the Control environment (Figures 3.3 & 3.4).

In the case of the mapping populations, these results highlight the importance of multiple cross-environmental trials in identifying the source of variation in phenotypic traits of interest and breeding decisions. However, they also highlight the importance of studying the genetic component of these traits, particularly as breeding progress could not be made reliably through selection based on phenotypes alone. Defining the G x E or QTL x E effects involved in nutritional quality trait would determine the value and effectiveness of a marker (Marwede et al. 2005; Li et al. 2010).

Another interesting point highlighted by the phenotypic data is the transgressive segregation which appears to exist across the F_2 progeny in most traits examined (Table 3.3, Figures 3.3 & 3.4). This would suggest that the cross may have unlocked variation beyond that exhibited by the parental accessions through combining alleles from both parents. Interesting work by Kirk et al. into the occurrence of transgressive segregation for plant defence compounds during hybridization showed that new defence profiles were achievable in backcrossed hybrids within the Senecio genus (Kirk et al. 2004) but also that transgressive segregation for some polyphenols- antioxidant compounds for the consumer- occurred in F_2 hydrid crosses of the *Jacobaea sp*. These finds would suggest that breeding in watercress for secondary metabolites, or nutritional quality traits, could lead to the production of RILs of commercial interest through exhibiting enhanced qualities of current stock. It is a promising results for the breeding potential of food stuff with greater nutritional quality for the consumer.

3.6.3 QTL in Morphological Traits and Candidate Genes:

Moderate to high levels of heritability (0.55-0.66) were found in all traits (Table 3.4), with 17 QTL detected for the seven traits. Markers underlying these QTL were examined for similarity to known genes in plants. The RE employed for this study was the rare cutter, *Pst1*, which is methylation-sensitive and thought to increase the coding regions sequenced by GBS and avoid highly repetitive parts of the genome (Emberton et al. 2005; Pootakham, Ruang-Areerate, et al. 2015). This may explain the reasonable number of markers with a BLAST hit – over half the markers underlying a significant QTL peak (Appendix B.22 Table S3.2). Despite the limited amount of information on watercress, this enabled a preliminary exploration for potential candidate genes or gene clusters within these QTL and provide an interesting set of candidates for further exploration.

The plant cytoskeleton - composed of microtubules and actin microfilaments - enables plant growth through cell wall formation and alteration (Kost & Chua 2002; Wasteneys & Galway 2003). QTL regions for traits of morphology in this population appeared to be enriched for markers with sequence similarity to genes associated with cell growth and cytoskeleton synthesis, function and relevant signalling (Table 3.6). Stem length QTL *lgth7* contained a marker (7634) with similarity to an actin-binding formin homology 2 (FH2) protein (*AT3G07540*), involved in actin organization (Winder & Ayscough 2005). A potential pectin acetylesterase 7 (*AT4G19410*) lay within the branching QTL on linkage group (LG) 6, which appears to be highly expressed in most phases of plant growth and development (Philippe et al. 2017). Two genes connected to cellulose biosynthesis were also found in these QTLs. The underlying marker to diameter QTL *dmt13* shows strong sequence similarity to cellulose synthase interactive 1 protein (CSI1), which is thought to play a role in cellulose organization in cell walls and the mutants of which display ‘swollen’ phenotypes with shorter hypocotyls with wider stem diameters (Gu et al. 2010). A second closely-related protein, cellulose synthase interactive 3 (CSI3) was also matched by peak marker 23478 in *nd16* for number of nodes. This marker had a dominant effect of lesser stem nodes from the dwarf Parent A allele and explained 6.6 % of phenotypic variance in the trait.

An important function of the cytoskeleton in plant growth is the facilitation of exocytosis and relevant genes were found twice amongst markers underlying QTL for morphology in this population. The marker 4131 within LG 16, where QTL collocate for mean internode distance (*dis16*) and number of nodes (*nd16*), has a close sequence match to the exocyst complex component SEC5A. SEC5A mutants in *Arabidopsis* exhibited problematic hypocotyl elongation (Hala et al. 2008). Calcium ion fluxes have been highlighted as a key signalling pathway for cellular response to developmental but also external stimuli (Ranty et al. 2006; Zeng et al. 2015) and are essential triggers for exocytosis initiation (Wasteneys & Galway 2003). It was interesting that several peak QTL markers in the population’s morphology traits showed sequence similarity with calmodulin (calcium-modulated messenger)-binding proteins. Peak QTL marker 5629 for branching on LG 6 matched a calmodulin-binding protein in *Arabidopsis*. Similarly, peak marker 2092 for QTL *dis11*, for internode distance, showed a strong similarity to the calmodulin-binding gene *AT1G13210* and explained 7.2 % of phenotypic variance. The phenotype associated with this marker had shorter intermodal distance in the dwarf Parent A and exhibited an additive effect over the phenotype.

In the QTL for stem length, *lgth7*, peak marker 14902 showed high sequence similarity to a calossin-like auxin transporter protein (BIG) (*AT3G02260*). As auxin is a key plant growth hormone, it is not surprising that BIG mutants exhibited reduced organelle growth and elongation but a potential involvement of BIG in calmodulin signalling has also been suggested (Gil et al. 2001; Luschnig 2001). Several steps in the biosynthetic pathway of brassinosteroids another set of hormones critical to plant growth, have been shown to depend on functional interaction with calmodulin (Du & Poovaiah 2005). Polymorphisms with effect on brassinosteroid biosynthesis may have particular relevance in this population as Payne et al. (2015) found a down-regulation in genes related to brassinosteroid biosynthesis in the dwarf watercress, used here as Parent A. An overall reduction in expression of genes involved in the control of plant growth, such as the phenylpropanoid and lignin biosynthetic pathways and in cell wall components, were identified in this transcription study (Payne et al. 2015). Therefore, such combinations of genes involved in the physical components or signalling of plant growth could be interesting targets when breeding for a dwarf watercress. As strong correlations were found between many morphology traits in both trials (Figure 3.6), attention must be paid to their effect of breeding on the overall yield of the crop. In explanation, we found a positive correlation between stem length and number of nodes so breeding for a shorter stem could also produce a crop with less leaves, affecting yield negatively.

Shelf life and leaf processability are particularly important for leafy crops and have been found to correlate well with leaf weight, suggesting that a denser leaf would be desirable (Zhang et al. 2007; Wagstaff et al. 2010). Although no particular BLAST hits stood out for markers underlying QTL for specific leaf area (SLA), the QTL on LG 1 and 5 together explained approximately 20% of variation and could therefore make an important contribution to crop quality. The favourable allele in both QTL originated from the large Parent B and had an additive effect over SLA. This might suggest that the 'dwarf' Parent A could be improved in leaf hardiness through these QTL.

3.6.4 QTL in Phytonutrient Traits and Candidate Genes:

The health benefit gained from a nutrient-dense crop, such as watercress, is an important marketable quality for growers and a valuable target for breeding and research. In this analysis we have focused on two potential consumer benefits: the AO and chemopreventive benefits of watercress consumption. Both were labelled here as phytonutrient traits, reflecting the accumulation of secondary plant metabolites, such as polyphenols and glucosinolates (GLS).

Such compounds are involved in plant response to external stimuli, i.e. herbivore or pathogen defence, and therefore candidate genes with related functions would be of particular interest.

There are a number of compounds that have been identified in watercress and contribute to its phytonutrient profile: phenolic compounds, ascorbic acid, and GLS. Flavonols were found to be prevalent in watercress, particularly quercetin, kaempferol, isorhamnetin and their derivatives (Martínez-Sánchez et al. 2008; Aires et al. 2013; Santos et al. 2014; Zeb 2015). Ascorbic acid, or Vitamin C, has been quantified in watercress on several occasions and it was recorded at higher concentrations than in other leafy greens to which it was compared (Palaniswamy et al. 2003; Martínez-Sánchez et al. 2008; Santos et al. 2014). A number of aliphatic and aromatic GLS have been recorded in watercress, with the primary GLS being gluconasturtiin and its derivative phenethyl isothiocyanate being a potent chemopreventive agent (Rose et al. 2000; Fahey et al. 2001; Zeb 2015; Voutsina et al. 2016). Martínez-Sánchez et al. (2008) found a correlation of AO capacity in watercress with ascorbic acid and phenols.

Several studies have previously mapped either total AO capacity or individual compounds with AO properties in other species. Studies in oilseed, tomato and lettuce have reported complex G x E interactions and also low heritability in trait (Marwede et al. 2005; Rousseaux et al. 2005; Hayashi et al. 2012). However, AO QTL for raspberries were found to be relatively consistent across several years and sites tested (Dobson et al. 2012). Despite the complex nature of this trait, several studies have reported candidate genes or transcription factors in directly relevant pathways, such as pigment or lignin biosynthesis, reactive oxygen species metabolism, or the phenylpropanoid pathway (Jin et al. 2009; Chagné et al. 2012; Sotelo et al. 2014; Damerum et al. 2015).

To the best of our knowledge, investigation into QTL associated directly with the effect of plant extracts on human cancer cells has never been undertaken previously. In this study, we identified three significant QTL for the cytotoxicity of watercress to cancer cells. The QTL were located in LG 2, 11 and 16 and collectively explained 20% of variation in IC_{50} values of the population under conditions of commercial cultivation (Table 3.6, Figure 3.7), where Parent B had the highest toxicity to cancer cells and from which two out of three favourable alleles originated. This would suggest that the chemopreventive benefit of the commercially-desirable ‘dwarf’ watercress could be improved further by breeding for these QTL. F_2 individuals with desirable phenotypes for both these significant breeding targets could become interesting commercial material once taken forward to later fixed generations.

Although the majority of markers underlying QTL for phytonutrient traits did not have an annotation match in current databases, a few links were identified to pathways with potential effects in these traits. The QTL for IC_{50}, *ic16*, had an additive effect (with a more potent cytotoxic effect from Parent A) and explained 7% of phenotypic variation for this trait (Figure 3.8). The marker underlying this QTL (13595) matched a phospholipid- transporting ATPase. The importance of phospholipid signalling in both plant development and stress responses has been documented and reviewed (Laxalt & Munnik 2002; Cowan 2006; Ruelland et al. 2015). Phospholipid signalling cascades play a role in pathogen immunity, including both nonspecific pathogen pattern-triggered immunity (PTI) and pathogen-specific effector-triggered immunity (ETI), which together increase plant defence gene expression and induce targeted cell death to stem the spread of a biotic threat (Ruelland et al. 2015).

Interestingly, the adjacent marker 6992 within *ic16* was found to be similar to leucine-rich repeat protein kinase PEPR1 in *Arabidopsis*, another receptor involved in plant defences and PTI (Krol et al. 2010; Liu et al. 2013). The transcription of both PEPR1 and several components of the phospholipid signalling pathway are upregulated by stress-signalling jasmonates (Yamaguchi et al. 2010; Profotová et al. 2006). The proximity of these two genes involved in plant defences and the significance of this QTL could suggest the potential presence of a resistance gene cluster in this location (Michelmore & Meyers 1998). Resistance gene clusters have been described in the genomes of various plant species, including soybean (Graham et al. 2002), coffee trees (Ribas et al. 2011), lettuce (Meyers et al. 1998) and common bean (Geffroy et al. 1999), and are important targets for breeding hardier crops (Shen et al. 1998). When the distribution of resistance genes was investigated in *Brassica napus*, it was noted that the genes of this family that were clustered together tended to contain more polymorphisms than those outside clusters (Alamery et al. 2018). If in fact these clusters can also be linked to the consumer’s health benefit, as suggested by the QTL found in this study, then such clusters could be particularly valuable breeding targets for the breeding of crops with greater nutritional quality.

The *ic16* QTL also collocated with QTL *nd16* on LG16 for which phospholipid signalling could also be relevant though its role in plant development (Figure 3.7 & 3.8). The collocation of QTL for a phytonutrient and a morphology trait is particularly interesting and could have implications for breeding. Previously, a relationship between phytonutrient accumulation and plant age or developmental stage has been identified in watercress both under controlled and variable field conditions. Firstly, in work completed by this group the concentration of phytonutrients, in this case glucosinolates, was found to increase over time in two commercial trials (see Chapter 4). A second example showed an increase in ascorbic acid concentrations (an AO) and PEITC, a derivative of the GLS gluconasturtiin, in watercress from seedling to 40 days of age when grown under controlled conditions (Palaniswamy et al. 2003). Two similar example have been reported in Brassicas where the concentration of GLS was found to increase in younger leaves for 40 days post-planting and from seedling to flowering stages (Porter et al. 1991; Velasco et al. 2007). Overall, the phytonutrient qualities of the vegetative tissue over time does not seem to be a broadly-studied topic, likely due to its particular relevance and application in only the small pool of leafy crops where the plant leaf itself is the marketable product.

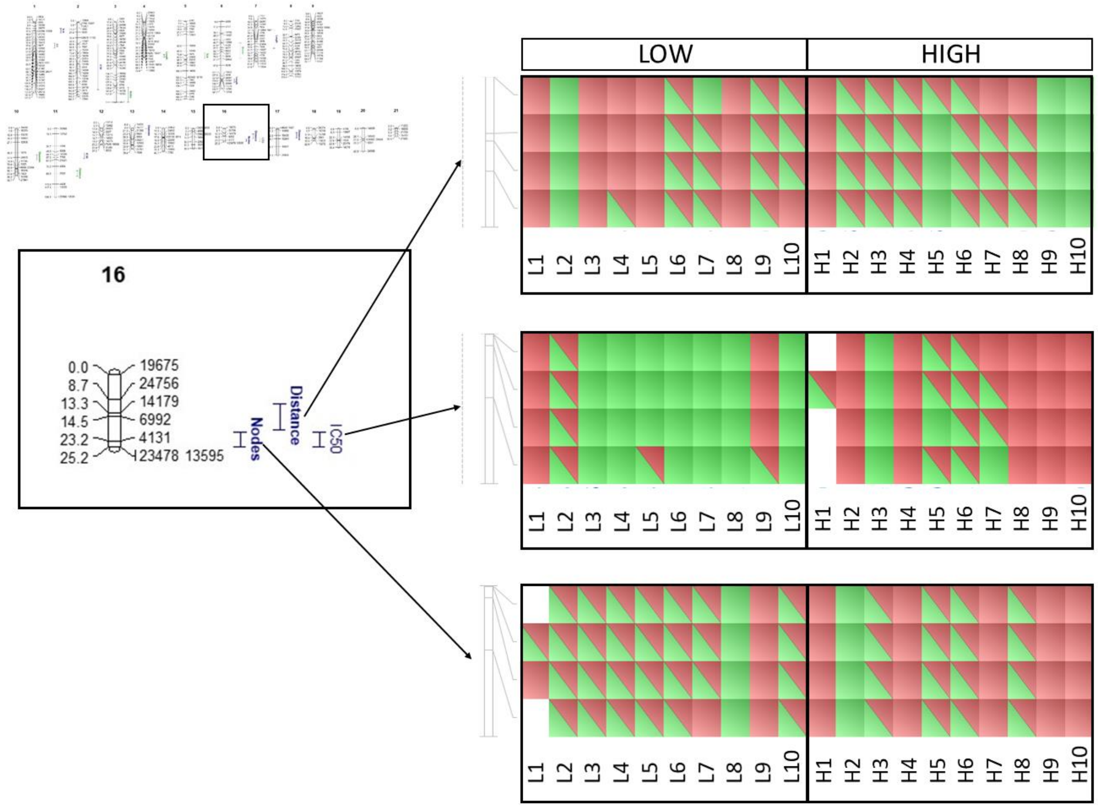


Figure 3.8 The SNP allele inherited by the ten highest and ten lowest F_2 individuals for number of nodes, mean internode distance and IC_{50} (where low equates with higher potency) at collocating QTL on linkage group 16. Each row indicates a marker with the allele from Parent A in green, the allele from Parent B in red, heterozygous individuals with both, and missing data left blank. The QTL are shown on linkage group 16 which illustrates genetic distance on the left in centimorgans (cM) and marker name on the right.

In a potential second link of morphology and phytonutrient profile, marker 13593 underlying AO QTL *ao1*, which explained 6.8 % of phenotypic variance for this trait, showed a strong similarity to a cyclin (*AT5G11300*). This family of proteins regulate the progression of the cell life cycle through their interaction with cyclin-dependent kinases (Hemerly et al. 1992; Mironov et al. 1999). It has been suggested that a cyclin-induced arrest in the Gap 1 phase of the cell cycle could induce an increase in production of secondary metabolites (Nejad et al. 2012). It Cross talk has also been identified between cell growth, specifically cyclin-dependent kinases, and secondary plant metabolites in the light of response to plant wounding (Neubauer et al.

2012). It is therefore conceivable that a polymorphism in a pathway regulating cell division, or other aspects of plant growth and development, would result in greater AO capacity of the plant material through a relationship with the accumulation of secondary metabolites. Proteins with roles in both plant growth and resistance pathways are also good breeding targets. Differential expression analysis for the AO trait in watercress, showed several leucine-rich repeat kinase receptors (Voutsina et al. 2016) and these receptors have been shown to play an important role in both plant defences and development, such brassinosteroid signalling (Li & Chory 1997; Goff & Ramonell 2007).

In watercress breeding, the implications of a strong link between plant morphology and phytonutrient value could be significant. If traits associated with watercress development and, ultimately architecture, are linked with secondary metabolite accumulation, then breeding for a morphologically different crop could also have direct impact on health benefit traits. It will thus be important to consider the effect of selection for morphology traits on phytonutrient traits and vice versa. Ideally, a genotype could be selected that simultaneously improved both qualities. The stability of the QTL identified here and further candidate gene mapping would improve further our understanding of the genetic architecture of these traits. The genotyped markers and linkage map created here could be tested in an F2:3 family experiment (Austin & Lee 1996) and permanent RILs resource could be developed and revisited. Fine mapping of the QTL could be accomplished by cloning the regions and identifying causal gene relationships to the traits based on sequence similarity with closely-related *Arabidopsis* at first instance (Salvi & Tuberosa 2005). However with the high levels of phenotypic plasticity in mind, approaches that characterize differential expression of differing genotypes for a QTL may produce very interesting results in watercress (Gelli et al. 2014; Gelli et al. 2017; Jian et al. 2017).

The findings of this study may also inform broader plant breeding efforts beyond watercress. Firstly, the results reported here contribute to the general understanding of the design and application of techniques utilizing reduced representation of the genome in the resolution genomic questions, particularly in light of the design and limitations present when working with understudied and/or non-model species for which DNA extraction, reference genomes or optimal RE may not already exist or have established protocols. Secondly, as traits linked to human health often involve secondary metabolites or other transient plant phenotypes which can vary vastly from one state of the subject plant to the next, the results of this study also highlight the need for particular attention to phenotypic data collection when working with such traits. F_2 or RIL phenotype can vary greatly from field to control and year to year so testing the genetic component across a large number of conditions is essential to the correct characterization of G x E, QTL or QTL x E effects. Towards plant breeding for leaf and stem component morphology, the QTL identified in this work points to the contribution of cell growth and cytoskeleton synthesis in developing and delivering the desirable morphology to salad leaf crops. Such breeding targets, as the brassinosteroid link to dwarf phenotype, may be targeted in breeding for similar leafy crop varieties. Finally, the identification of QTL for traits linked with human health in this work, and in particular the cytotoxicity to cancer cells, shows that QTL can be mapped directly to plant genomes and result in meaningful breeding targets, such as the potential disease resistance gene cluster. This direct approach could be valuable to breeding healthier food stuffs across crops and may point to key genes with consistent roles in the phytonutrient trait across species.

Table 3.7 Candidate genes underlying QTL identified for various traits, based on marker sequence similarity to the watercress transcriptome or the NCBI database.

Trait	Linkage Group	Marker	Annotation Source	Gene	Functional description	Reference
AO	1	13593	Wx Transcript	Mitotic-like cyclin (AT5G11300)	Cell cycle regulation	(Hemerly et al. 1992; Mironov et al. 1999
Branches	6	5629	NCBI	*Arabidopsis thaliana* chromosome 4 sequence	Calmodulin binding	
Branches	6	24452	NCBI	Pectin acetylesterase 7-like (AT4G19410)	Pectin modification in cell walls	Philippe et al. (2017)
Diameter	13	14771	Wx Transcript & NCBI	Cellulose interactive 1 protein (CSI1)	Cell elongation	Gu et al. (201)
Distance	11	2092	Wx Transcript	Autoinhibited Ca2+/ATPase II (AT1G13210)	Calmodulin binding	
Distance/ Nodes/ IC50	16	4131	Wx Transcript & NCBI	Exocyst complex gene SEC5A (AT1G76850)	Regulated or polarized secretion from the plasma membrane	Hala et al. (2008)
Distance/ Nodes/ IC50	16	6992	NCBI	Leucine-rich repeat receptor-like protein kinase PEPR1	PEP defense peptide receptor	Krol et al. 2010; Liu et al. 2013
IC50	2	2249	Wx Transcript	NADPH--cytochrome P450 reductase 1 (AT4G24520)	Involved in phenylpropanoid metabolism	
Length	7	14902	Wx Transcript & NCBI	Auxin transport protein (BIG) (AT3G02260)	Auxin regulated stem internode elongation	Gil et al. (2001)
Nodes/ IC50	16	13595	NCBI	Phospholipid-transporting ATPase 11	Phospholipid transport	Laxalt & Munnik 2002; Cowan 2006; Ruelland et al. 2015
Nodes/ IC50	16	23478	Wx Transcript & NCBI	Cellulose synthase iInteractive 3 (CSI3)	Cytoskeleton regulation	Gu et al. (201)

3.7 Conclusions

The first mapping population and molecular genetic map was developed in watercress and used to identify a number of significant QTL. These were enriched with markers showing sequence similarity to genes in cytoskeleton structure, modification and developmental signalling pathways. In a novel effort, we mapped the inhibitory effect of watercress on human cancer cells and found three QTL which together explained 20 % of the phenotypic variation associated with this trait. This work has produced valuable genomic resources and will make a significant impact on future research and breeding in this species.

3.8 References

Abràmoff, M. D., Magalhães, P. J., & Ram, S. J. (2004). Image processing with ImageJ. *Biophotonics International*. https://doi.org/10.1117/1.3589100

Aires, A., Carvalho, R., Rosa, E. a. S., & Saavedra, M. J. (2013). Phytochemical characterization and antioxidant properties of baby-leaf watercress produced under organic production system. *CyTA - Journal of Food*, *11*(4), 343–351.

Al-Shehbaz, I. A., Beilstein, M. A., & Kellogg, E. A. (2006). Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Systematics and Evolution*, *259*(2–4), 89–120. https://doi.org/10.1007/s00606-006-0415-z

Alamery, S., Tirnaz, S., Bayer, P., Tollenaere, R., Chaloub, B., Edwards, D., & Batley, J. (2018). Genome-wide identification and comparative analysis of NBS-LRR resistance genes in *Brassica napus*. *Crop and Pasture Science*, *69*(1), 72. https://doi.org/10.1071/CP17214

Altschul, S. F., Gish, W., Miller, W., Myers, E. E. W. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. https://doi.org/10.1016/S0022-2836(05)80360-2

Austin, D. F., & Lee, M. (1996). Comparative mapping in F2:3 and F6:7 generations of quantitative trait loci for grain yield and yield components in maize. *Theoretical and Applied Genetics*, *92*(7), 817–826. https://doi.org/10.1007/BF00221893

Benzie, I. F., & Strain, J. J. (1996). The ferric reducing ability of plasma (FRAP) as a measure of “antioxidant power”: the FRAP assay. *Analytical Biochemistry*, *239*(1), 70–76.

Bleeker, W., Huthmann, M., & Hurka, H. (1999). Evolution of the hybrid taxa in Nasturtium R.BR. (Brassicaceae). *Folia Geobotanica*, *34*, 421–433.

Branham, S. E., Stansell, Z. J., Couillard, D. M., & Farnham, M. W. (2017). Quantitative trait loci mapping of heat tolerance in broccoli (*Brassica oleracea* var. italica) using genotyping-by-sequencing. *Theoretical and Applied Genetics, 130*(3), 529–538. https://doi.org/10.1007/s00122-016-2832-x

Broman, K. W., Wu, H., Sen, S., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics, 19*(7), 889–890. https://doi.org/10.1093/bioinformatics/btg112

Casanova, N. A., Ariagno, J. I., López Nigro, M. M., Mendeluk, G. R., de los A Gette, M., Petenatti, E., … Carballo, M. A. (2013). In vivo antigenotoxic activity of watercress juice (*Nasturtium officinale*) against induced DNA damage. *Journal of Applied Toxicology : JAT*, *33*(9), 880–885. https://doi.org/10.1002/jat.2746

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Cavell, B. E. (2012). *In vitro analysis of potential anticancer effects associated with watercress*. *PhD Thesis*. University of Southampton.

Cavell, B. E., Syed Alwi, S. S., Donlevy, A. M., Proud, C. G., & Packham, G. (2012). Natural product-derived antitumor compound phenethyl isothiocyanate inhibits mTORC1 activity via TSC2. *Journal of Natural Products*, *75*(6), 1051–1057.

Chagné, D., Krieger, C., Rassam, M., Sullivan, M., Fraser, J., André, C., … Laing, W. A. (2012). QTL and candidate gene mapping for polyphenolic composition in apple fruit. *BMC Plant Biology*, *12*(1), 12. https://doi.org/10.1186/1471-2229-12-12

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, *142*(1–2), 169–196. https://doi.org/10.1007/s10681-005-1681-5

Cory, A. H., Owen, T. C., Barltrop, J. A., & Cory, J. G. (1991). Use of an aqueous soluble tetrazolium/formazan assay for cell growth assays in culture. *Cancer Communications*, *3*(7), 207–12.

Cowan, A. K. (2006). Phospholipids as Plant Growth Regulators. *Plant Growth Regulation*, *48*(2), 97–109. https://doi.org/10.1007/s10725-005-5481-7

Damerum, A., Selmes, S. L., Biggi, G. F., Clarkson, G. J., Rothwell, S. D., Truco, M. J., … Taylor, G. (2015). Elucidating the genetic basis of antioxidant status in lettuce (*Lactuca sativa*). *Horticulture Research*, *2*, 15055. https://doi.org/10.1038/hortres.2015.55

Davey, J. W., Davey, J. L., Blaxter, M. L., & Blaxter, M. W. (2010). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, *9*(5–6), 416–423. https://doi.org/10.1093/bfgp/elq031

Davey, J. W., Hohenlohe, P. a, Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

Dobson, P., Graham, J., Stewart, D., Brennan, R., Hackett, C. A., & McDougall, G. J. (2012). Over-seasons analysis of quantitative trait loci affecting phenolic content and antioxidant capacity in raspberry. *Journal of Agricultural and Food Chemistry*, *60*(21), 5360–5366. https://doi.org/10.1021/jf3005178

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, *19*, 11–15. https://doi.org/10.2307/4119796

Du, L., & Poovaiah, B. W. (2005). Ca2+/calmodulin is critical for brassinosteroid biosynthesis and plant growth. *Nature*, *437*(7059), 741–745. https://doi.org/10.1038/nature03973

El-Soda, M., Malosetti, M., Zwaan, B., Koornneef, M., & Aarts, M. (2014). Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends in Plant Science*, *19*(6), 390–398. https://doi.org/10.1016/J.TPLANTS.2014.01.001

Elia, M. (2015). *The cost of malnutrition in England and potential cost savings from nutritional interventions (full report). BAPEN.*

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a, Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Emberton, J., Ma, J., Yuan, Y., SanMiguel, P., & Bennetzen, J. L. (2005). Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Research*, *15*(10), 1441–1446. https://doi.org/10.1101/gr.3362105

Fahey, J. W., Zalcmann, A. T., & Talalay, P. (2001). The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry*, *56*(1), 5–51.

Fogarty, M. C., Hughes, C. M., Burke, G., Brown, J. C., & Davison, G. W. (2013). Acute and chronic watercress supplementation attenuates exercise-induced peripheral mononuclear cell DNA damage and lipid peroxidation. *The British Journal of Nutrition*, *109*(2), 293–301.

Frazão, E. (1999). America’s eating habits: changes and consequences. *Agriculture Information Bulletin*, *494*, 5–32.

Fu, Y.-B. (2014). Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. *G3 (Bethesda, Md.)*, *4*(5), 891–900. https://doi.org/10.1534/g3.114.010942

Gardner, K. M., Brown, P., Cooke, T. F., Cann, S., Costa, F., Bustamante, C., … Myles, S. (2014). Fast and cost-effective genetic mapping in apple using next-generation sequencing. *Genes, Genomes, Genetics*, *4*(9), 1681–1687. https://doi.org/10.1534/g3.114.011023

Geffroy, V., Sicard, D., de Oliveira, J. C., Sévignac, M., Cohen, S., Gepts, P., … Dron, M. (1999). Identification of an ancestral resistance gene cluster involved in the coevolution process between *Phaseolus vulgaris* and its fungal pathogen Colletotrichum lindemuthianum. *Mpmi*. https://doi.org/10.1094/MPMI.1999.12.9.774

Gelli, M., Duo, Y., Konda, A. R., Zhang, C., Holding, D., & Dweikat, I. (2014). Identification of differentially expressed genes between sorghum genotypes with contrasting nitrogen stress tolerance by genome-wide transcriptional profiling. *BMC Genomics*, *15*(1), 179. https://doi.org/10.1186/1471-2164-15-179

Gelli, M., Konda, A. R., Liu, K., Zhang, C., Clemente, T. E., Holding, D. R., & Dweikat, I. M. (2017). Validation of QTL mapping and transcriptome profiling for identification of candidate genes associated with nitrogen stress tolerance in sorghum. *BMC Plant Biology*, *17*(1), 123. https://doi.org/10.1186/s12870-017-1064-9

Gil, P., Dewey, E., Friml, J., Zhao, Y., Snowden, K. C., Putterill, J., … Chory, J. (2001). BIG: a calossin-like protein required for polar auxin transport in Arabidopsis. *Genes & Development*, *15*(15), 1985–97. https://doi.org/10.1101/gad.905201

Gill, C. I. R., Haldar, S., Boyd, L. a, Bennett, R., Whiteford, J., Butler, M., … Rowland, I. R. (2007). Watercress supplementation in diet reduces lymphocyte DNA damage and alters blood antioxidant status in healthy adults. *The American Journal of Clinical Nutrition*, *85*(2), 504–510.

Goff, K. E., & Ramonell, K. M. (2007). The role and regulation of receptor-like kinases in plant defense. *Gene Regulation and Systems Biology*, *1*, 167–75.

Goonetilleke, S. N., March, T. J., Wirthensohn, M. G., Arús, P., Walker, A. R., & Mather, D. E. (2018). Genotyping by sequencing in almond: SNP discovery, linkage mapping, and marker design. *G3&amp;#58; Genes|Genomes|Genetics*, *8*(1), 161–172. https://doi.org/10.1534/g3.117.300376

Graham, M. A., Marek, L. F., & Shoemaker, R. C. (2002). Organization, expression and evolution of a disease resistance gene cluster in soybean. *Genetics*.

Gu, Y., Kaplinsky, N., Bringmann, M., Cobb, A., Carroll, A., Sampathkumar, A., …

Somerville, C. R. (2010). Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(29), 12866–71. https://doi.org/10.1073/pnas.1007092107

Gupta, P., Adkins, C., Lockman, P., & Srivastava, S. K. (2013). Metastasis of breast tumor cells to brain is suppressed by phenethyl isothiocyanate in a novel in vivo metastasis model. *PloS One*, *8*(6), e67278. https://doi.org/10.1371/journal.pone.0067278

Hala, M., Cole, R., Synek, L., Drdova, E., Pecenkova, T., Nordheim, A., … Zarsky, V. (2008). An exocyst complex functions in plant cell growth in Arabidopsis and Tobacco. *THE PLANT CELL ONLINE*, *20*(5), 1330–1345. https://doi.org/10.1105/tpc.108.059105

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series, 41*, 95–98.

Hayashi, E., You, Y., Lewis, R., Calderon, M. C., Wan, G., & Still, D. W. (2012). Mapping QTL, epistasis and genotype × environment interaction of antioxidant activity, chlorophyll content and head formation in domesticated lettuce (Lactuca sativa). *Theoretical and Applied Genetics, 124*(8), 1487–1502. https://doi.org/10.1007/s00122-012-1803-0

Hecht, S., Chung, F., Richie, J. J., Akerkar, S., Borukhova, A., Skowronski, L., & Carmella, S. (1995). Effects of watercress consumption on metabolism of a tobacco-specific lung carcinogen in smokers. *Cancer Epidemiol. Biomarkers Prev.*, *4*(8), 877–884.

Hemerly, A., Bergounioux, C., Van Montagu, M., Inzé, D., & Ferreira, P. (1992). Genes regulating the plant cell cycle: isolation of a mitotic-like cyclin from Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America, 89*(8), 3295–9.

Higdon, J. V, Delage, B., Williams, D. E., & Dashwood, R. H. (2007). Cruciferous vegetables and human cancer risk: epidemiologic evidence and mechanistic basis. *Pharmacological Research : The Official Journal of the Italian Pharmacological Society, 55*(3), 224–236. https://doi.org/10.1016/j.phrs.2007.01.009

Hussain, W., Baenziger, P. S., Belamkar, V., Guttieri, M. J., Venegas, J. P., Easterly, A., … Poland, J. (2017). Genotyping-by-Sequencing derived high-density linkage map and its application to QTL mapping of flag leaf traits in bread wheat. *Scientific Reports, 7*(1), 16394. https://doi.org/10.1038/s41598-017-16006-z

Jeelani, S. M., Rani, S., Kumar, S., Kumari, S., & Gupta, R. C. (2013). Cytological studies of Brassicaceae Burn. (Cruciferae Juss.) from Western Himalayas. *Cytology and Genetics*, *47*(1), 20–28. https://doi.org/10.3103/S0095452713010076

Jian, H., Yang, B., Zhang, A., Zhang, L., Xu, X., Li, J., & Liu, L. (2017). Screening of candidate leaf morphology genes by integration of QTL mapping and RNA sequencing technologies in oilseed rape (*Brassica napus* L.). *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0169641

Jin, L., Xiao, P., Lu, Y., Shao, Y., Shen, Y., & Bao, J. (2009). Quantitative trait loci for brown rice color, phenolics, flavonoid contents, and antioxidant capacity in rice grain. *Cereal Chemistry Journal*, *86*(6), 609–615. https://doi.org/10.1094/CCHEM-86-6-0609

Jo, J., Purushotham, P. M., Han, K., Lee, H.-R., Nah, G., & Kang, B.-C. (2017). Development of a genetic map for onion (*Allium cepa* L.) using reference-free genotyping-by-sequencing and SNP assays. *Frontiers in Plant Science*, *8*, 1606. https://doi.org/10.3389/fpls.2017.01606

Kirk, H., Máčel, M., Klinkhamer, P. G. L., & Vrieling, K. (2004). Natural hybridization between *Senecio jacobaea* and *Senecio aquaticus*: Molecular and chemical evidence. *Molecular Ecology*. https://doi.org/10.1111/j.1365-294X.2004.02235.x

Kost, B., & Chua, N. (2002). The Plant Cytoskeleton: Vacuoles and cell walls make the difference. *Cell*, *108*(1), 9–12. https://doi.org/10.1016/S0092-8674(01)00634-1

Krol, E., Mentzel, T., Chinchilla, D., Boller, T., Felix, G., Kemmerling, B., … Hedrich, R. (2010). Perception of the Arabidopsis danger signal peptide 1 involves the pattern recognition receptor AtPEPR1 and its close homologue AtPEPR2. *The Journal of Biological Chemistry*, *285*(18), 13471–9. https://doi.org/10.1074/jbc.M109.097394

Lambel, S., Lanini, B., Vivoda, E., Fauve, J., Patrick Wechter, W., Harris-Shultz, K. R., … Levi, A. (2014). A major QTL associated with *Fusarium oxysporum* race 1 resistance identified in genetic populations derived from closely related watermelon lines using selective genotyping and genotyping-by-sequencing for SNP discovery. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *127*(10), 2105–2115. https://doi.org/10.1007/s00122-014-2363-2

Langridge, P., & Fleury, D. (2011). Making the most of ‘omics’ for crop breeding. *Trends in Biotechnology*, *29*(1), 33–40.

Laxalt, A., & Munnik, T. (2002). Phospholipid signalling in plant defence. *Current Opinion in Plant Biology*, *5*(4), 332–338. https://doi.org/10.1016/S1369-5266(02)00268-6

Li, H., Liu, H., Han, Y., Wu, X., Teng, W., Liu, G., & Li, W. (2010). Identification of QTL underlying vitamin E contents in soybean seed among multiple environments. *Theoretical and Applied Genetics*, *120*(7), 1405–1413. https://doi.org/10.1007/s00122-010-1264-2

Li, J., & Chory, J. (1997). A putative leucine-rich repeat receptor kinase involved in Brassinosteroid signal transduction. *Cell*, *90*(5), 929–938. https://doi.org/10.1016/S0092-8674(00)80357-8

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., & Brummer, E. C. (2014). A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. *G3 (Bethesda, Md.)*, *4*(10), 1971–1979. https://doi.org/10.1534/g3.114.012245

Liu, Z., Wu, Y., Yang, F., Zhang, Y., Chen, S., Xie, Q., … Zhou, J.-M. (2013). BIK1 interacts with PEPRs to mediate ethylene-induced immunity. *Proceedings of the National Academy of Sciences*, *110*(15), 6205–6210. https://doi.org/10.1073/pnas.1215543110

Luschnig, C. (2001). Auxin transport: Why plants like to think BIG. *Current Biology*, *11*(20), R831–R833. https://doi.org/10.1016/S0960-9822(01)00497-3

Macleod, A. J., & Islam, R. (1975). Volatile flavour components of watercress. *Journal of the Science of Food and Agriculture*, *26*(10), 1545–1550.

Manchali, S., Chidambara Murthy, K. N., & Patil, B. S. (2012). Crucial facts about health benefits of popular cruciferous vegetables. *Journal of Functional Foods*, *4*(1), 94–106. https://doi.org/10.1016/j.jff.2011.08.004

Manton, I. (1935). The cytological history of Watercress (*Nasturtium officinale* R. Br.). *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, *69*(1), 132–157.

Manton, I., & Howard, H. W. (1946). Autopolyploid and allopolyploid watercress with the description of a new species. *Annals of Botany*, *10*, 1–13.

Marchese, A., Marra, F. P., Caruso, T., Mhelembe, K., Costa, F., Fretto, S., & Sargent, D. J. (2016). The first high-density sequence characterized SNP-based linkage map of olive (*Olea europaea* L. subsp. europaea) developed using genotyping by sequencing. *Australian Journal of Crop Science*, *10*(6), 857–863. https://doi.org/10.21475/ajcs.2016.10.06.p7520

Martínez-Sánchez, A., Gil-Izquierdo, A., Gil, M. I., & Ferreres, F. (2008). A comparative study of flavonoid compounds, vitamin C, and antioxidant properties of baby leaf Brassicaceae species. *Journal of Agricultural and Food Chemistry*, *56*(7), 2330–2340.

Marwede, V., Gul, M. K., Becker, H. C., & Ecke, W. (2005). Mapping of QTL controlling tocopherol content in winter oilseed rape. *Plant Breeding*, *124*(1), 20–26. https://doi.org/10.1111/j.1439-0523.2004.01050.x

Meyers, B. C., Chin, D. B., Shen, K. A., Sivaramakrishnan, S., Lavelle, D. O., Zhang, Z., & Michelmore, R. W. (1998). The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. *Plant Cell*. https://doi.org/http://www.plantcell.org/content/10/11/1817.short

Michelmore, R. W., & Meyers, B. C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research*, *8*(11), 1113–30. https://doi.org/10.1101/GR.8.11.1113

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W. T. B., … Marshall, D. (2010). Flapjack - Graphical Genotype Visualization. *Bioinformatics (Oxford, England)*, *26*(24), 3133–3134. https://doi.org/10.1093/bioinformatics/btq580

Mironov, V., De Veylder L, L. De, Van Montagu M, M. Van, & Inze, D. (1999). Cyclin-dependent kinases and cell division in plants- the nexus. *The Plant Cell*, *11*(4), 509–22. https://doi.org/10.1105/TPC.11.4.509

Montero-Pau, J., Blanca, J., Esteras, C., Martínez-Pérez, E. M., Gómez, P., Monforte, A. J., … Picó, B. (2017). An SNP-based saturated genetic map and QTL analysis of fruit-related traits in Zucchini using Genotyping-by-sequencing. *BMC Genomics*, *18*(1), 94. https://doi.org/10.1186/s12864-016-3439-y

Morozowska, M., Czarna, A., & Jędrzejczyk, I. (2010). Estimation of nuclear DNA content in Nasturtium R. Br. by flow cytometry. *Aquatic Botany*, *93*(4), 250–253.

Nejad, E. S., Askari, H., & Soltani, S. (2012). Regulatory TGACG-motif may elicit the secondary metabolite production through inhibition of active Cyclin-dependent kinase/Cyclin complex. *POJ*, *5*(6), 553–558.

Neubauer, J. D., Lulai, E. C., Thompson, A. L., Suttle, J. C., & Bolton, M. D. (2012). Wounding coordinately induces cell wall protein, cell cycle and pectin methyl esterase genes involved in tuber closing layer and wound periderm development. *Journal of Plant Physiology*, *169*(6), 586–595. https://doi.org/10.1016/j.jplph.2011.12.010

Newman, R. M., Kerfoot, W. C., & Iii, Z. H. (1996). Watercress allelochemical defends high-nitrogen foliage against consumption : Effects on freshwater invertebrate herbivores. *Ecology*, *77*(8), 2312–2323.

Palaniswamy, U. R., & McAvoy, R. J. (2001). Watercress: A salad crop with chemopreventive potential. *HortTechnology*, *11*(4), 622–626.

Palaniswamy, U. R., McAvoy, R. J., Bible, B. B., & Stuart, J. D. (2003). Ontogenic variations of ascorbic acid and phenethyl isothiocyanate concentrations in watercress (*Nasturtium officinale* R.Br.) leaves. *Journal of Agricultural and Food Chemistry*, *51*(18), 5504–5509.

Payne, A. C., Clarkson, G. J. J., Rothwell, S., & Taylor, G. (2015). Diversity in global gene expression and morphology across a watercress (*Nasturtium officinale* R. Br.) germplasm collection: first steps to breeding. *Horticulture Research*, *2*, 15029. https://doi.org/10.1038/hortres.2015.29

Payne, A. C., Mazzer, A., Clarkson, G. J. J., & Taylor, G. (2013). Antioxidant assays - consistent findings from FRAP and ORAC reveal a negative impact of organic cultivation on antioxidant potential in spinach but not watercress or rocket leaves. *Food Science & Nutrition*, *1*(6), 439–44.

Philippe, F., Pelloux, J., & Rayon, C. (2017). Plant pectin acetylesterase structure and function: new insights from bioinformatic analysis. *BMC Genomics*, *18*(1), 456. https://doi.org/10.1186/s12864-017-3833-0

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS One*, *7*(2), e32253. https://doi.org/10.1371/journal.pone.0032253

Pootakham, W., Jomchai, N., Ruang-areerate, P., Shearman, J. R., Sonthirod, C., Sangsrakru, D., … Tangphatsornruang, S. (2015). Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics*, *105*(5), 288–295. https://doi.org/10.1016/j.ygeno.2015.02.002

Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., … Tangphatsornruang, S. (2015). Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). *Frontiers in Plant Science*, *6*, 367. https://doi.org/10.3389/fpls.2015.00367

Porter, A. J. R., Morton, A. M., Kiddle, G., DoughtyY, K. J., & Wallsgrove, R. M. (1991). Variation in the glucosinolate content of oilseed rape (*Brassica napus* L.) leaves. *Annals of*

Applied Biology, *118*(2), 461–467. https://doi.org/10.1111/j.1744-7348.1991.tb05647.x

Profotová, B., Burketová, L., Novotná, Z., Martinec, J., & Valentová, O. (2006). Involvement of phospholipases C and D in early response to SAR and ISR inducers in *Brassica napus* plants. *Plant Physiology and Biochemistry*, *44*(2–3), 143–151. https://doi.org/10.1016/J.PLAPHY.2006.02.003

Ranty, B., Aldon, D., & Galaud, J.-P. (2006). Plant calmodulins and calmodulin-related proteins: multifaceted relays to decode calcium signals. *Plant Signaling & Behavior*, *1*(3), 96–104.

Ribas, A. F., Cenci, A., Combes, M. C., Etienne, H., & Lashermes, P. (2011). Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. *BMC Genomics*. https://doi.org/10.1186/1471-2164-12-240

Riss, T. L., Moravec, R. A., Niles, A. L., Duellman, S., Benink, H. A., Worzella, T. J., & Minor, L. (2004). *Cell Viability Assays*. *Assay Guidance Manual*. Eli Lilly & Company and the National Center for Advancing Translational Sciences.

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-response analysis using R. *PLoS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0146021

Rose, P., Faulkner, K., Williamson, G., & Mithen, R. (2000). 7-Methylsulfinylheptyl and 8-methylsulfinyloctyl isothiocyanates from watercress are potent inducers of phase II enzymes. *Carcinogenesis*, *21*(11), 1983–1988.

Rose, P., Huang, Q., Ong, C. N., & Whiteman, M. (2005). Broccoli and watercress suppress matrix metalloproteinase-9 activity and invasiveness of human MDA-MB-231 breast cancer cells. *Toxicology and Applied Pharmacology*, *209*(2), 105–113.

Rousseaux, M. C., Jones, C. M., Adams, D., Chetelat, R., Bennett, A., & Powell, A. (2005). QTL analysis of fruit antioxidants in tomato using *Lycopersicon pennellii* introgression lines. *Theoretical and Applied Genetics*, *111*(7), 1396–1408. https://doi.org/10.1007/s00122-005-0071-7

Ruelland, E., Kravets, V., Derevyanchuk, M., Martinec, J., Zachowski, A., & Pokotylo, I. (2015). Role of phospholipid signalling in plant environmental responses. *Environmental and Experimental Botany*, *114*, 129–143. https://doi.org/10.1016/J.ENVEXPBOT.2014.08.009

Russell, J., Hackett, C., Hedley, P., Liu, H., Milne, L., Bayer, M., … Brennan, R. (2014). The use of genotyping by sequencing in blackcurrant (*Ribes nigrum*): developing high-resolution linkage maps in species without reference genome sequences. *Molecular Breeding*, *33*(4), 835–849. https://doi.org/10.1007/s11032-013-9996-8

Sadeghi, H., Mostafazadeh, M., Sadeghi, H., Naderian, M., Barmak, M. J., Talebianpoor, M. S., & Mehraban, F. (2013). In vivo anti-inflammatory properties of aerial parts of *Nasturtium officinale*. *Pharmaceutical Biology*, (2008), 1–6. https://doi.org/10.3109/13880209.2013.821138

Salvi, S., & Tuberosa, R. (2005). To clone or not to clone plant QTLs: Present and future challenges. *Trends in Plant Science*. https://doi.org/10.1016/j.tplants.2005.04.008

Santos, J., Oliveira, M. B. P. P., Ibáñez, E., & Herrero, M. (2014). Phenolic profile evolution of different ready-to-eat baby-leaf vegetables during storage. *Journal of Chromatography. A*, *1327*, 118–131.

Saxena, R. K., Singh, V. K., Kale, S. M., Tathineni, R., Parupalli, S., Kumar, V., … Varshney, R. K. (2017). Construction of genotyping-by-sequencing based high-density genetic maps and QTL mapping for fusarium wilt resistance in pigeonpea. *Scientific Reports*, *7*(1), 1911. https://doi.org/10.1038/s41598-017-01537-2

Scarborough, P., Bhatnagar, P., Wickramasinghe, K. K., Allender, S., Foster, C., & Rayner, M. (2011). The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006-07 NHS costs. *Journal of Public Health (Oxford, England)*, *33*(4), 527–535. https://doi.org/10.1093/pubmed/fdr033

Shen, K. A., Meyers, B. C., Islam-Faridi, M. N., Chin, D. B., Stelly, D. M., & Michelmore, R. W. (1998). Resistance gene candidates identified by PCR with degenerate oligonucleotide primers map to clusters of resistance genes in lettuce. *Molecular Plant-Microbe Interactions*. https://doi.org/10.1094/MPMI.1998.11.8.815

Sheridan, G. E. C., Claxton, J. R., Clarkson, J. M., & Blakesley, D. (2001). Genetic diversity within commercial populations of watercress (Rorippa nasturtium-aquaticum), and between allied Brassicaceae inferred from RAPD-PCR. *Euphytica*, *122*, 319–325.

Shi, J., Huang, S., Zhan, J., Yu, J., Wang, X., Hua, W., … Wang, H. (2014). Genome-wide microsatellite characterization and marker development in the sequenced Brassica crop species. *DNA Research*, *21*(1), 53–68. https://doi.org/10.1093/dnares/dst040

Snowdon, R. J., & Friedt, W. (2004). Molecular markers in Brassica oilseed breeding: Current status and future possibilities. *Plant Breeding*. https://doi.org/10.1111/j.1439-

0523.2003.00968.x

Sotelo, T., Soengas, P., Velasco, P., Rodríguez, V. M., & Cartea, M. E. (2014). Identification of metabolic QTLs and candidate genes for glucosinolate synthesis in *Brassica oleracea* leaves, seeds and flower buds. *PloS One*, *9*(3), e91428. https://doi.org/10.1371/journal.pone.0091428

Syed Alwi, S. S., Cavell, B. E., Telang, U., Morris, M. E., Parry, B. M., & Packham, G. (2010). In vivo modulation of 4E binding protein 1 (4E-BP1) phosphorylation by watercress: a pilot study. *The British Journal of Nutrition*, *104*(9), 1288–1296.

Team, R. (2013). R Development Core Team. *R: A Language and Environment for Statistical Computing*, *55*, 275–286.

Traka, M. H., & Mithen, R. (2008). Glucosinolates, isothiocyanates and human health. *Phytochemistry Reviews*, *8*(1), 269–282.

Traka, M. H., Saha, S., Huseby, S., Kopriva, S., Walley, P. G., Barker, G. C., … Mithen, R. F. (2013). Genetic regulation of glucoraphanin accumulation in Beneforté broccoli. *The New Phytologist*, *198*(4), 1085–1095.

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. a. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, *27*(9), 522–530.

Velasco, P., Cartea, M. E., Gonzalez, C., Vilar, M., & Ordas, A. (2007). Factors affecting the glucosinolate content of kale (*Brassica oleracea* acephala group). *Journal of Agricultural and Food Chemistry*, *55*(3), 955–962.

Velmurugan, J., Mollison, E., Barth, S., Marshall, D., Milne, L., Creevey, C. J., … Milbourne, D. (2016). An ultra-high density genetic linkage map of perennial ryegrass (*Lolium perenne*) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly. *Annals of Botany*, *118*(1), 71–87. https://doi.org/10.1093/aob/mcw081

Verhoeven, D., Goldbohm, R., van Poppel, G., Verhagen, H., & van den Brandt, P. (1996). Epidemiological studies on Brassica vegetables and cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, *5*(9), 733–748.

Voorrips, R. E. (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *Journal of Heredity*, *93*(1), 77–78. https://doi.org/10.1093/jhered/93.1.77

Voutsina, N., Payne, A. C., Hancock, R. D., Clarkson, G. J., Rothwell, S. D., Chapman, M. A.,

& Taylor, G. (2016). Characterization of the watercress (*Nasturtium officinale* R. Br.; Brassicaceae) transcriptome using RNASeq and identification of candidate genes for important phytonutrient traits linked to human health. *BMC Genomics*, *17*, 378. https://doi.org/DOI 10.1186/s12864-016-2704-4

Wagner, A. E., Terschluesen, A. M., & Rimbach, G. (2013). Health promoting effects of brassica-derived phytochemicals: from chemopreventive and anti-inflammatory activities to epigenetic regulation. *Oxidative Medicine and Cellular Longevity*, *2013*, 964539.

Wagstaff, C., Clarkson, G. J. J., Zhang, F., Rothwell, S. D., Fry, S. C., Taylor, G., & Dixon, M. S. (2010). Modification of cell wall properties in lettuce improves shelf life. *Journal of Experimental Botany*, *61*(4), 1239–1248. https://doi.org/10.1093/jxb/erq038

Wang, Y., Pan, Y., Liu, Z., Zhu, X., Zhai, L., Xu, L., … Liu, L. (2013). De novo transcriptome sequencing of radish (*Raphanus sativus* L.) and analysis of major genes involved in glucosinolate metabolism. *BMC Genomics*, *14*(1), 836.

Ward, J. A., Bhangoo, J., Fernández-Fernández, F., Moore, P., Swanson, J. D., Viola, R., … Sargent, D. J. (2013). Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. *BMC Genomics*, *14*(1), 2. https://doi.org/10.1186/1471-2164-14-2

Wasteneys, G. O., & Galway, M. E. (2003). Remodeling the cytoskeleton for growth and form : An overview with some new views. *Annual Review of Plant Biology*, *54*(1), 691–722. https://doi.org/10.1146/annurev.arplant.54.031902.134818

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*, *18*(1), 586. https://doi.org/10.1186/s12859-017-2000-6

Winder, S. J., & Ayscough, K. R. (2005). Actin-binding proteins. *Journal of Cell Science*, *118*(4), 651–654. https://doi.org/10.1242/jcs.01670

Yamaguchi, Y., Huffaker, A., Bryan, A. C., Tax, F. E., & Ryan, C. A. (2010). PEPR2 is a second receptor for the Pep1 and Pep2 peptides and contributes to defense responses in Arabidopsis. *The Plant Cell*, *22*(2), 508–22. https://doi.org/10.1105/tpc.109.068874

Yang, M.-D., Lai, K.-C., Lai, T.-Y., Hsu, S.-C., Kuo, C.-L., Yu, C.-S., … Chung, J.-G. (2010). Phenethyl isothiocyanate inhibits migration and invasion of human gastric cancer AGS cells through suppressing MAPK and NF-kappaB signal pathways. *Anticancer Research*, *30*(6), 2135–2143.

Yang, Z., Chen, Z., Peng, Z., Yu, Y., Liao, M., & Wei, S. (2017). Development of a high-density linkage map and mapping of the three-pistil gene (Pis1) in wheat using GBS markers. *BMC Genomics*, *18*(1), 567. https://doi.org/10.1186/s12864-017-3960-7

Zeb, A. (2015). Phenolic profile and antioxidant potential of wild watercress (*Nasturtium officinale* L.). *SpringerPlus*, *4*, 714. https://doi.org/10.1186/s40064-015-1514-5

Zeng, H., Xu, L., Singh, A., Wang, H., Du, L., & Poovaiah, B. W. (2015). Involvement of calmodulin and calmodulin-like proteins in plant responses to abiotic stresses. *Frontiers in Plant Science*, *6*, 600. https://doi.org/10.3389/fpls.2015.00600

Zhang, F. Z., Wagstaff, C., Rae, A. M., Sihota, A. K., Keevil, C. W., Rothwell, S. D., … Taylor, G. (2007). QTLs for shelf life in lettuce co-locate with those for leaf biophysical properties but not with those for leaf developmental traits. Journal of Experimental Botany, 58(6), 1433–1449. https://doi.org/10.1093/jxb/erm006

Chapter 4: Characterization of a new dwarf watercress ‘Boldrewood’ in commercial trials reveals consistent increase in chemopreventive properties in a longer-grown crop

4.1 Abstract

We describe the characterization of ‘Boldrewood’ a new accession of watercress (*Nasturtium officinale* R. Br.), that was initially found to be a short stature plant with high nutritional antioxidant capacity (Payne et al. 2015). This was of particular commercial interest because it offered the potential to develop a novel watercress product with fork-friendly size and improved health-benefits. In two commercial trials comparing Boldrewood to a commercial control, we confirmed that Boldrewood exhibits a dwarf phenotype with a significantly shorter stem and consistently produced more leaves per stem area alongside comparable crop biomass. The antioxidant and chemopreventive capacity of Boldrewood were comparable to the commercial crop. For the first time, we observed a novel increase in glucosinolate concentrations and cytotoxicity to cancer cells, characterised as decreased IC_{50} (half-maximal concentration of an inhibitor) associated with increased crop age at harvest, from 21 to 35 d post-planting. This suggests that a slower-growing and longer to harvest crop provides a significant improvement in health benefits gained in this leafy crop which is already known to be highly nutrient dense and with considerable chemopreventive ability.

4.2 Introduction

Watercress (*Nasturtium officinale* R. Br.) has a long history of medicinal and food crop use by humans, detailed by Manton (1935). It is produced on a large scale as a leafy green for salads and soups across the globe although exact market data is hard to find. In the U.K., watercress cultivation is inescapably linked with the chalk streams of the southern counties, particularly those of Hampshire, Dorset, Devon and Wiltshire. Typically, seedlings are produced in glasshouses or polytunnels and then transferred to purpose-built gravel beds (Natural England, 2009). The crop is flood-irrigated for several weeks until it reaches the desirable size, at which point it is mechanically harvested and sent to market, most commonly as part of pre-washed and bagged salad products (Natural England 2009).

Watercress is a member of the *Brassicaceae* plant family, which includes a large range of economically-important edible plants, such as broccoli, cabbage, and mustard; oilseed rape, one of the largest oil crops; and the prominent research model species, *Arabidopsis thaliana* (L.) Heynh. Many of these *Brassicaceae* plants have significant markets or applications driving the development of extensive resources to facilitate the production of new and improved cultivars. For example, Beneforte broccoli was recently developed and released to market as a more nutrient-dense cultivar (Traka et al., 2013). However, this is not the case with watercress, which in the UK holds a modest 5% of the total salads market (G Clarkson 2014, personal communication). To date, limited resources exist to enable crop improvement in response to consumer demands or increased production pressures.

Overall, efforts in crop breeding are driven by a variety of targets, such as increasing yield - for example, the development of modern high-yielding bread wheat cultivars (Slafer et al. 1994) - or introducing cosmetic improvement to a crop to increase sales, as with the standardization of the orange colour in carrots (Simon 2010). The UoS watercress breeding program aims to improve the current watercress cultivar and discover new cultivars that might be successful in market. Guided by the grower input, the focus has been set on three particular breeding targets: 1) above-ground morphology, 2) consumer health benefit, and 3) resource-use efficiency. Here we are concerned with targets 1) and 2) – developing a novel watercress crop of improved morphology and health benefits. The first is focused on the developing a leafier product with a shorter stem, more suited to bagged salads as opposed to the traditional product of bunched watercress. The second breeding target is to capture, preserve and potentially enhance the high phytonutrient profile in the commercial cultivar. Watercress is known to carry a considerable health benefit for the human consumer and was recently ranked as the most powerful fruit or vegetable to consume for health (Di Noia 2014). Research has shown the crop to be linked with detoxification of carcinogens (Hecht et al. 1995); reduction of cancer risk (Gill et al. 2007); decreased damage by reactive oxygen species during strenuous exercise (Fogarty et al. 2013); and reduced inflammation (Sadeghi et al. 2013). In this report, we will focus on two key aspects of the phytonutrient profile of watercress: antioxidant capacity (AO) and the anti-cancer properties, quantified as glucosinolate (GLS) content and the cytotoxicity of the plants to cancer cells (IC_{50} – the concentration required to kill 50 % of cells present in culture).

The AO profile of watercress is thought to be composed primarily of phenolic compounds, carotenoids and ascorbic acid (commonly Vitamin C), and several studies have looked at the concentrations and composition of the crop for these compounds in watercress (Martínez-Sánchez et al. 2008; Payne et al. 2013; Santos et al. 2014). In a recent publication comparing AO capacity of a number of *Brassicaceae* leafy crops, watercress contained the highest concentrations in both AO and ascorbic acid content, highlighting the important role this crop can play in delivering human nutrition (Martínez-Sánchez et al. 2008). GLS are thought to be particularly beneficial to humans (Traka & Mithen 2008; Manchali et al. 2012) and are produced in response to invertebrate predation on watercress (Newman et al. 1992). Upon tissue damage to the plant, glucosinolates are hydrolysed by the enzyme myrosinase and one of the products of this reaction are isothiocyanates (Bones & Iversen 1985; Bones & Rossiter 1996). In watercress, primarily gluconasturtiin is converted to phenethyl isothiocyanate (PEITC) (Newman et al. 1992), which is not only considered to have potent cancer-fighting properties (Traka & Mithen 2008; Manchali et al. 2012), but is also responsible for the highly-marketable peppery flavour of this leafy salad.

The third breeding target of the program is improving the resource use efficiency of the crop, with particular and urgent focus on developing or identifying material with increased phosphate (P) use efficiency. The ‘as needed’ addition of P to the substrate and water on watercress farms has been utilized traditionally to optimize crop performance. However, P concentrations are typically very low in chalk streams and an increase in association to watercress farm outflow has contributed to the introduction of government limits on abstraction (Casey 1981; Casey & Smith 1994; Natural England 2009). Therefore, the development of P use efficient watercress cultivar would be particularly desirable however; this target was identified more recently and will be the target of a new PhD project.

Despite the outstanding promise of modern agrobiotechnology, such as Next Generation Sequencing and genome editing, pre-existing genetic diversity in the natural world remains central for crop improvement and is the single largest, and arguably the most straightforward, resource to draw from in crop breeding (Dwivedi et al. 2013). A germplasm collection, containing known variants and wild relatives or progenitors of a crop, holds the potential not only to discover pre-existing variants, but also the genetic information to begin understanding a crop and building the resources necessary to establish a breeding program. Therefore, for the understudied watercress the establishment of a germplasm collection, beginning in 2009, by the Taylor Lab group was a major stepping stone (Payne 2011). The collection contains approximately 50 accessions of seed and clonal material, collected globally by individuals from watercress growers and donated by Warwick HRI. The collection was assessed for variation in above-ground morphology and tissue AO capacity and showed promising variation across the 48 accessions (Payne et al. 2015). It has been utilized further to sequence the transcriptome of watercress and perform differential expression analysis for AO capacity and GLS content (Voutsina et al. 2016).

A particular accession, named ‘Boldrewood’, exhibited the top ranking AO capacity of the collection (Payne 2011) leading to its selection for further comparisons. Boldrewood has also repeatedly produced a dwarf phenotype. The gluconasturtiin and PEITC content in Boldrewood were examined and although the concentration of gluconasturtiin was lower than that of the commercial cultivar, the concentration of PEITC was similar, leading the researchers to suggest a more efficient conversion in the Boldrewood material (Payne 2011). Payne (2015) examined differences in gene expression between the two accessions and concluded that the Boldrewood watercress showed a reduction in gene expression of biochemical pathways critical to plant growth, specifically brassinosteroids and phenylpropanoid/lignin synthesis, which could explain its dwarf phenotype. In addition, Payne et al. (2015) noted an increase in expression of genes associated with plant defences, such as GLS and ascorbic acid biosynthesis, which could contribute to the consistent high phytonutrient profile identified by the researcher.

The potential of this accession to address two of the breeding targets of the program have led to commercialization trials to test practical aspects of its cultivation and marketing. Based on the current evidence detailed above, we hypothesize that the Boldrewood crop would be shorter and leafier with comparable or higher nutritional value to the current commercial crop. However, further work is also required to confirm this phenotype in a commercial field setting, because the work detailed above was conducted in controlled laboratory conditions lacking natural variation or stressors encountered by a commercial crop. Thus, the aim of this research was to compare the performance for marketable traits of the Boldrewood accession to the current cultivar under commercial conditions.

4.3 Materials and Methods

The ‘Boldrewood’ accession was compared to an active commercial U.K. cultivar, referred to as ‘Control’ hereafter, across two consecutive summer growing seasons on commercial farms in Dorset, U.K. (Appendix C.1 Figure S4.1). In both trials, the seed used had been fixed through several generations of selfing and the seedlings were produced through the standard commercial procedures, thus a consistent material quality can be assumed that is comparable to the usual commercial production for this particular grower. In addition, both trials were grown to organic farming specifications for the commercial production of organic watercress.

4.3.1 Trial 1 – A commercial trial in 2015:

Boldrewood was commercially tested during the U.K. growing season in 2015 on a commercial watercress farm located in Dorset, U.K. (50°44’24.7”N, 2°12’40.5”W). In late June, the following material was harvested from watercress beds that were close in proximity: 1) a Boldrewood at 28 d post planting, 2) a Control at 28 d post planting, and 3) a second Boldrewood at 42 d post planting. The final group represents the size at which the commercial grower classified the Boldrewood plants as ‘ready to harvest’ and was, thus, comparable to the ‘ready to harvest’ Control crop at 28 d.

Five 30 cm by 30 cm quadrats were randomly sampled and photographed for Boldrewood and Control crops. Five stems were randomly taken from each quadrat, cut from the base and measured immediately for stem length, stem diameter (in the middle of the stem), and number of leaves. Leaf number index was calculated as the ratio of leaves to stem. Within each quadrat, replicates were averaged to produce a quadrat mean value for further analysis. As an assessment of yield, the entire biomass in each quadrat was collected, dried at 80° C to constant weight, and weighed.

Five foil pouches were filled with tissue, randomly from across the bed, and snap-frozen for the evaluation of AO capacity and GLS concentration. Tissue AO capacity was assessed using a ferric reducing ability of plasma (FRAP) assay modified for plant sap (Benzie & Strain 1996; Payne et al. 2013). Quantitative GLS analysis was carried out on lyophilized and finely ground tissue as follows by Dr R. Hancock (The James Hutton Institute, Dundee, U.K.): Approximately 0.02 g of lyophilized material was rehydrated in 1 ml solution of 75 % methanol and 0.1 % formic acid, then briefly vortexed, sonicated for 15 min and mixed for 10 min on a shaker. Samples were then centrifuged (21000 g, 5 min) and the supernatant transferred to High

Performance Liquid Chromatography (HPLC) vials. Samples were placed in a random order and then sequentially injected (5 ml) onto a Phenomenex Luna 3 m C18 (2) 150 x 2 mm column. Solvents were 0.1 % formic acid in water (A) and 0.1 % formic acid in acetonitrile (B), flow rate was 190 ml/min and the gradient was 0 min, 5 % B; 45 min, 5 % B; 47 min, 75 % B; 52 min, 75 % B; 54 min, 5 % B; 60 min, 5 % B. GLS were identified by parent and fragment mass spectra and in the case of gluconasturtiin and glucobrassicin by coelution with commercial standards. GLS concentration was estimated using the peak area of the parent ion in the mass spectrometer compared to a standard curve of the same compound (for 4-Me glucobrassicin the standard curve of glucobrassicin was used to estimate concentration).

As the aim was to compare the groups for performance in particular marketable traits, a single-factor Analysis of Variance (ANOVA) was used to test group differences ($\alpha \leq 0.05$), with Accession treated as a fixed factor. All data were examined for normality and homogeneity of variance and an example ANOVA Sum of Squares (SS) table is included as Table 4.1. Where ANOVA assumptions were not met, specifically for the glucobrassicin and gluconasturtiin concentrations, a Welch's Test and a pairwise comparison were used instead. Further comparisons were made using a Tukeys's HSD test. Statistical analysis and graphs were completed in R, version 3.2.2 (R Development Core Team 2013).

Table 4.1 Example Sums of Squares table for Trial 1 in 2015. The dependent variable in this case was biomass (g).

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Accession	2	201.4754	100.7377	34.66559	1.03E-05
Residuals	12	34.87182	2.905985		

4.3.2 Trial 2 – A commercial trial in 2016:

Trial 2 was undertaken at a commercial watercress farm in Dorset U.K (50°48’46.8”N, 2°08’47.9”W). A watercress bed was cleared and dressed, according to standard commercial practice for organic watercress. Both Control and Boldrewood accessions were germinated and transplanted to the bed as week-old seedlings. A randomized plot design was used containing ten 4 m x 4 m plots of each accession and equal volumes of seedlings were transplanted to each plot by hand (Appendix C.2 Figure S4.2).

After 21 d of growth (3 weeks) the farm manager deemed the Control accession was "ready to harvest" and the following measurements were taken: stem length, stem diameter, and number of leaves, dry weight (of a 30 cm x 30 cm quadrat), and tissue was snap-frozen in liquid nitrogen for AO and cancer cell cytotoxicity assays. The morphological measurements were taken in the same way as in Trial 1: 5 replicate plant stems were harvested from within the plot, staying at least 50 cm within the plot boundary to avoid edge effects, and the mean was used for statistical comparisons. AO capacity was measured as in Trial 1. Additionally, the toxicity of the watercress sap to cancer cells was quantified as an IC_{50} value (half-maximal inhibitory concentration). For this, the survival of MCF-7 breast cancer cells in serial dilutions of watercress sap from each sample was assessed by MTS cell proliferation assays (Cell Titer 96® Aqueous One Solution, Promega, U.K.) and survival curves were fitted, similarly Cavell et al. (2012). To make this dataset relevant to the grower and because the two accessions vary in size, these measurements were repeated the following two weeks, at 28 d and 35 d, to cover the entire period during which either one or the other accession were deemed as "ready to harvest" by the farm manager.

Data were analysed using a two-factor Analysis of Variance (ANOVA), investigating differences in group means of fixed factors Accession and Date, and their interaction. Table 4.2 is an example of the resulting Sum of Squares (SS) tables. For most variables, the dataset was balanced and so they were analysed using Type I sum of squares (SS). The IC_{50} and AO capacity datasets had missing values and were analysed as a Type III SS. Normality and homogeneity of variances was checked and data was log transformed as necessary. Further comparisons were made using a Tukeys's HSD test. Statistical analysis and graphs were completed in R, version 3.2.2 (R Development Core Team 2013).

Table 4.2 Example Sums of Squares table for Trial 2 analysis in 2016. The dependent variable in this case was biomass (g).

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Accession	1	144.6775	144.6775	15.26597	0.000262
Date	2	570.7413	285.3707	30.11152	1.64E-09
Accession : Date	2	4.108663	2.054332	0.216767	0.805814
Residuals	54	511.7648	9.477126		

4.4 Results

4.4.1 Trial 1 in 2015

Table 4.5 shows the mean and standard error for agronomic traits collected in this study: stem length (cm), stem diameter (mm), number of leaves, leaf number index, and biomass (g). For stem length, the Control accession had longer stems ($p < 0.0001$) than both Boldrewood age groups; however stem diameter was not different. The number of leaves present was higher in the 42 d Boldrewood plants ($p = 0.0002$) but the 28 d plants of both accessions were not different to each other. Leaf number index was also found to be different between the accessions ($p = 0.0014$) with both Boldrewood age groups showing more leaves per unit of stem than the Control. This mirrors the difference in stem length between the two accessions and suggests that the 'dwarf' Boldrewood plants are indeed producing a comparable amount of leaves on a smaller stem. However, biomass (Table 4.5) was taken as an indicator of crop yield and showed a significant difference between the 28 d Boldrewood and the other two treatments ($p < 0.0001$). This suggests that Boldrewood would require an additional 14 d of growth to produce the same biomass yield as the Control accession.

Table 4.3 The mean and standard error of the mean (n = 5) for morphological traits and biomass for Control and Boldrewood accessions at 28 and 42 days post-planting in Trial 1. Statistical significance of the factor Accession is denoted as * p ≤ 0.05, ** p ≤ 0.01. *** p ≤ 0.001, and **** p ≤ 0.0001.

		28 d		***42 d***	
Trait		**Control**	**Boldrewood**	**Control**	**Boldrewood**
Stem length (cm)	****	9.32 (± 0.73)	2.88 (± 0.45)	N/A	3.85 (± 0.92)
Stem diameter (mm)	ns	1.35 (± 0.23)	1.27 (± 0.16)	N/A	1.47 (± 0.15)
No of leaves	***	13.94 (± 4.86)	15.28 (± 8.71)	N/A	28.84 (± 11.18)
Leaf number index	**	1.48 (± 0.05)	6.02 (± 1.50)	N/A	7.64 (± 0.54)
Biomass (g)	****	14.03 (± 0.48)	7.23 (± 0.56)	N/A	15.71 (± 1.09)

For phytonutrient content and the key benefits to consumers, an AO assay was conducted and specific GLS concentrations were assessed using HPLC-MS. The AO potential of the Boldrewood accession, shown in Figure 4.1, decreased significantly from 28 to 42 d of growth (*p = 0.01*) but the Control accession was not different to either of the other groups. An opposing trend was noted for gluconasturtiin, the primary glucosinolate found in watercress (Figure 4.2), which increased for Boldrewood at 28 to 42 d of age (*p = 0.006*). In this case again, the gluconasturtiin concentrations of the Control accession fell between and did not vary from either Boldrewood age group. The concentrations of glucobrassicin were not statistically different, however the concentrations of 4-methoxy-glucobrassicin were higher in both ages of the Boldrewood accession to those measured in the Control accession (*p = 0.0016*).

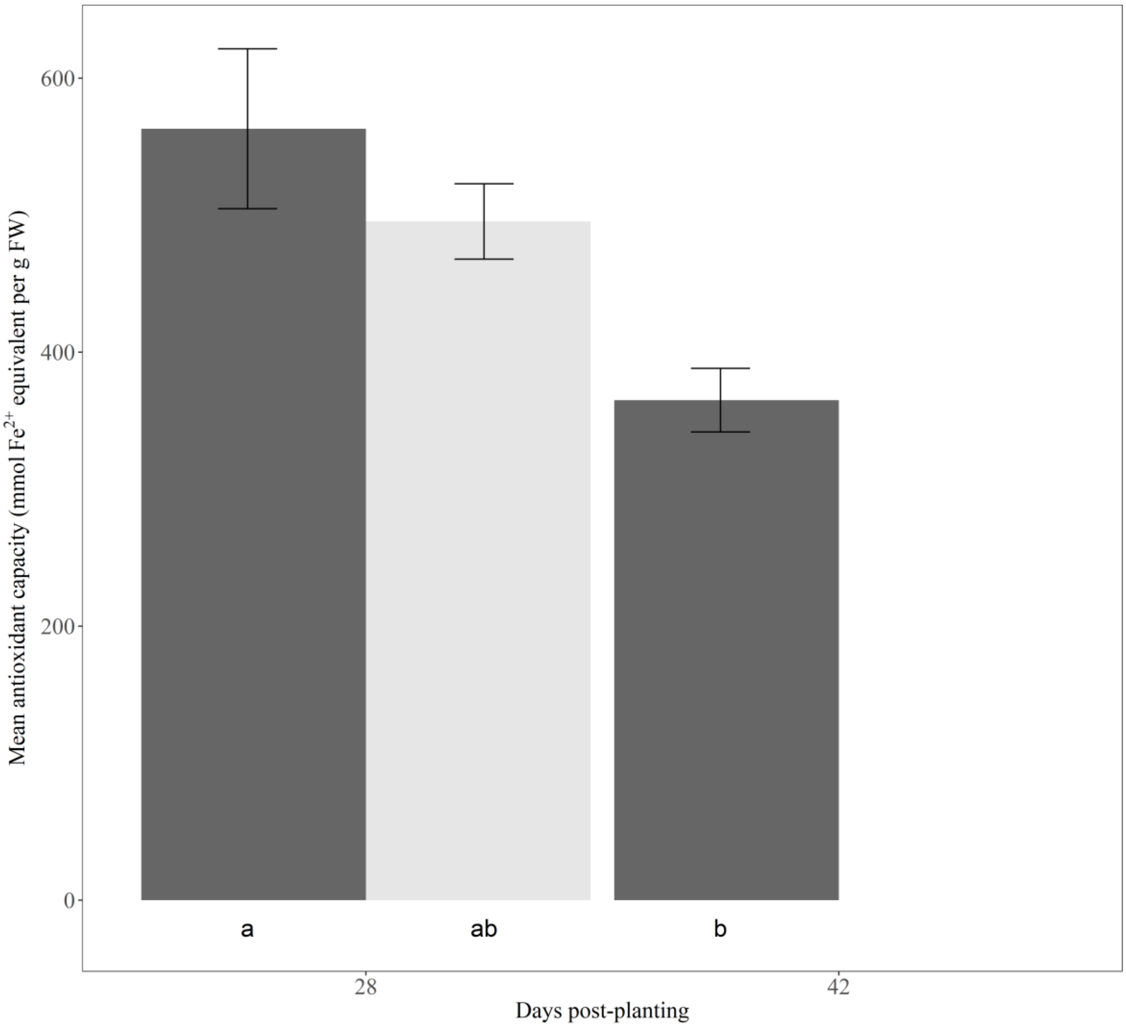


Figure 4.1 Mean antioxidant capacity (mmol of Fe^{2+} equivalent per g fresh weight) of Boldrewood (dark) and Control (light) accessions in Trial 1 at 28 and 42 days post planting. Values are as measured by FRAP antioxidant assay. Error bars indicate the standard deviation (n = 5) and letters denote statistically significant differences ($p \leq 0.05$).

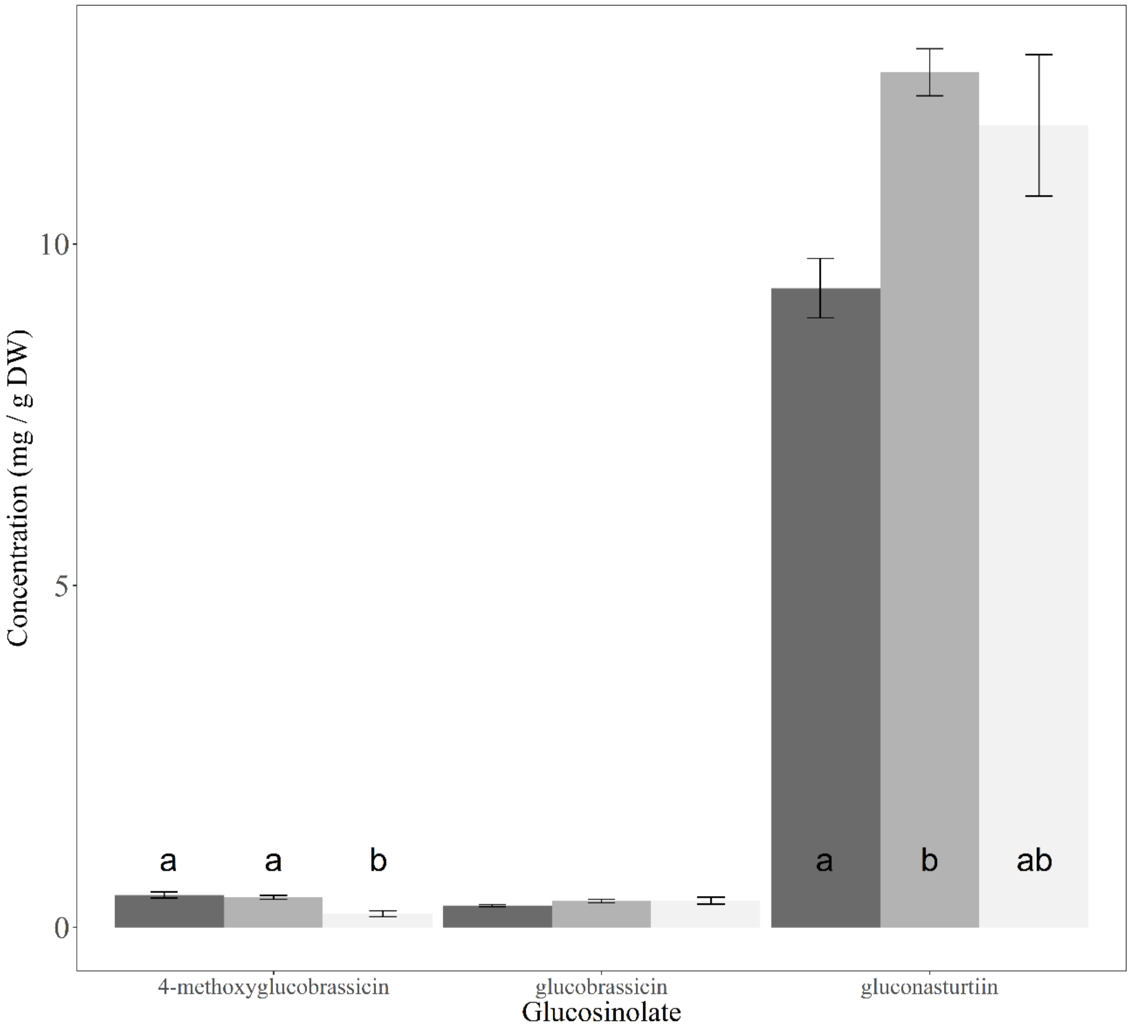


Figure 4.2 Mean specific glucosinolate concentrations (mg per gram of dry weight) for each accession in Trial 1. Boldrewood at 28 days post planting is represented in dark, Boldrewood at 42 days in medium, and Control at 28 days in light grey. Error bars indicate the standard deviation (n = 5) and letters denote statistically significant differences ($p \leq 0.05$).

In summary, during Trial 1 the Control plants had longer stems but Boldrewood plants produced more leaf per stem unit. Despite this, the Boldrewood biomass lagged approximately 14 d behind the Control. The accessions did not differ for the primary glucosinolate, gluconasturtiin, but Boldrewood samples contained more 4-methoxy-glucobrassicin.

4.4.2 Trial 2 in 2016

The agronomic traits assessed during this trial are presented in Figure 4.3. Stem length was consistently longer in Control plants (Accession, $p < 0.0001$; Date, $p < 0.0001$), confirming the results of Trial 1. Stem diameter also tested for significant ANOVA factors (Accession, $p = 0.001$; Date, $p < 0.0001$) but post hoc testing showed that true differences were due to an increase in stem diameter in both accessions at 35 d. Therefore, the results remain consistent with Trial 1; the two accessions do not differ significantly in stem diameter. There were significant differences in both factors for the number of leaves (Accession, $p = 0.03$; Date, $p < 0.0001$), however after post hoc testing, the significant p values are explained by an increase in leaf number for both accessions from 21 to 28 d. As in Trial 1, leaf number index was consistently higher for the Boldrewood plants (Accession, $p < 0.0001$; Date, $p < 0.0001$) and it universally decreased over time, as stem elongation was faster than new leaf formation and expansion. Crop productivity was compared again through an assessment of biomass set by each accession. Mean dry weight, shown in Figure 4.3, was significantly different for both factors Accession ($p = 0.00026$) and Date ($p < 0.0001$), but post hoc analysis revealed that differences were relevant to time and that the accessions yielded comparable biomass at each time point.

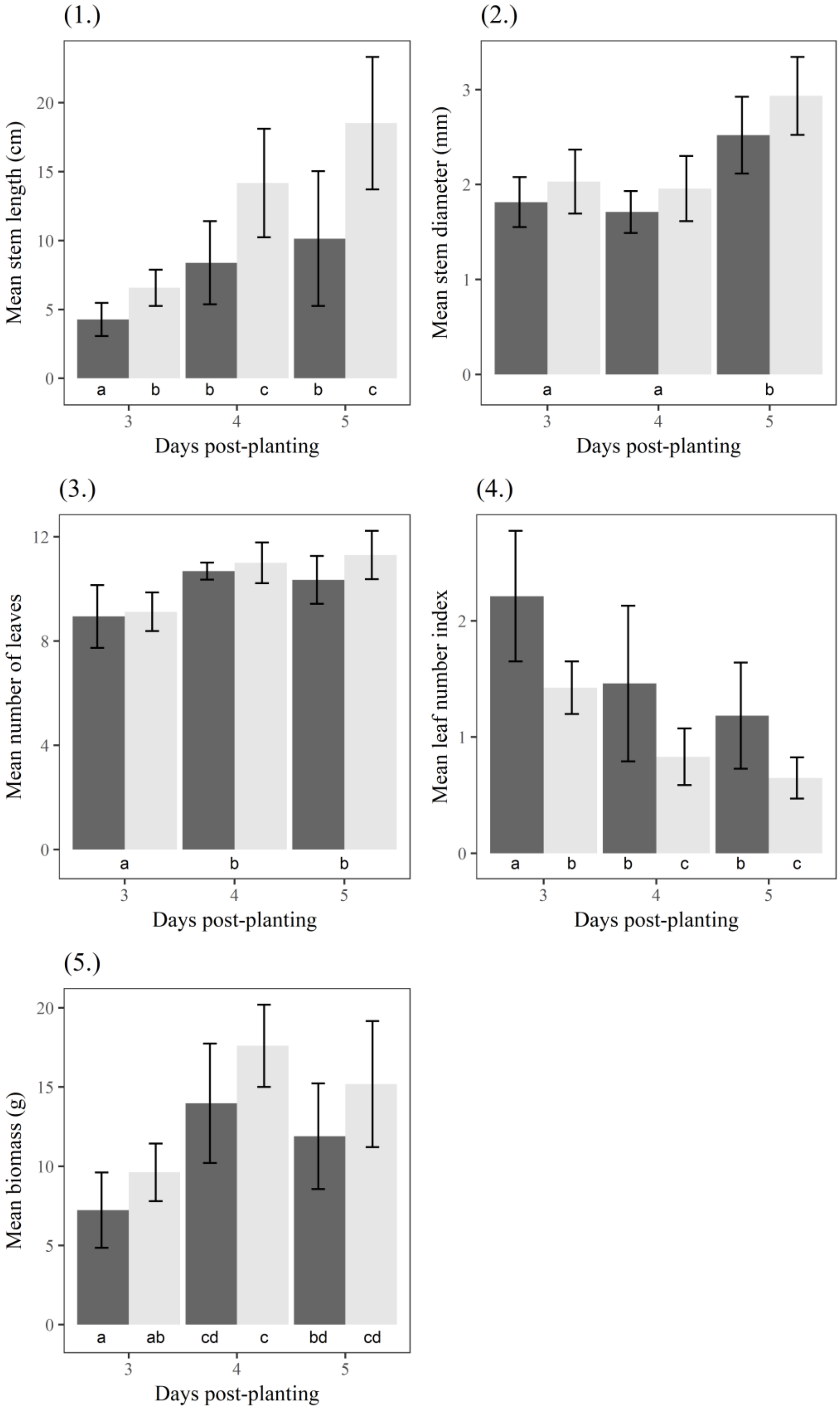


Figure 4.3 Mean values for stem length (1., cm), stem diameter (2., mm), number of leaves (3.), leaf number index (4.) and biomass (5., g) for Boldrewood (dark grey) and Control (light grey) accessions at 21, 28 and 35 days post-planting. Error bars signify the standard deviation and *n*=10

The phytonutrient quality of the crops was also tested. Sap was extracted and used in a serial dilution to treat cancer cells in order to estimate the concentration required to inhibit cancer cell growth by 50%, or IC_{50}. The results from this assay are shown in Figure 4.4, and although Accession was not a significant factor, Date was highly significant for this variable ($p = 0.00016$). The IC_{50} concentration decreased with time implying that the watercress extract increased in potency against cancer cell growth as the plants aged. The potent chemopreventive compound in watercress is phenethyl isothiocyanate (PEITC) which is derived from gluconasturtiin (Newman et al. 1992). As gluconasturtiin was found to increase with age in Trial 1, we can conclude that this result is consistent to Trial 1. For AO, the assay revealed no differences between Accessions or Date in Trial 2.

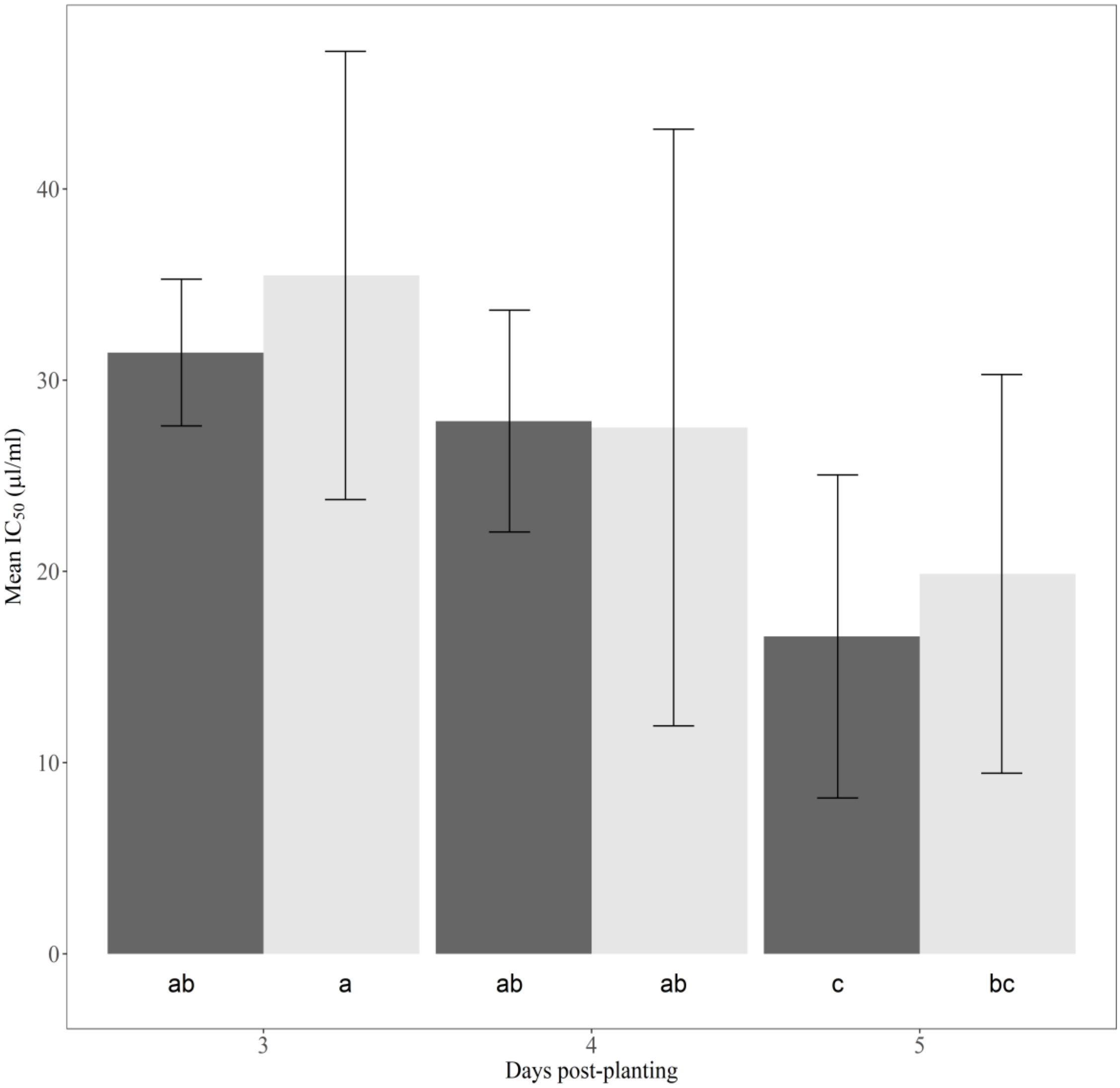


Figure 4.4 Mean inhibitor concentration required to cause lethal toxicity to 50 % of MCF7 breast cancer cells present (IC_{50}) for Boldrewood (dark) and Control (light) watercress accessions in Trial 2 at 21, 28 and 35 days post planting. Error bars indicate the standard deviation (n=10) and letters denote statistically significant differences ($p \leq 0.05$)

In summary for Trial 2, the Control accession maintained a longer stem than Boldrewood and Boldrewood plants again produced more leaves per stem unit. In contrast to Trial 1, biomass did not differ significantly between the two accession suggesting a minimal loss in yield if replacing one accession with the other in production. An interesting increase in cytotoxic capacity against cancer cells with crop age was noted for a second time here linked to the GLS/isothiocyanate system.

4.5 Discussion

Our results show that the Boldrewood accession exhibits a dwarf phenotype in comparison with the commercial Control watercress. The Control used in this research is a watercress cultivar that is grown widely across the south of England by many commercial companies and which each company continues to produce yearly from selfed seed. Thus, Boldrewood represents a new source of novel germplasm with a distinct dwarf phenotype now observed over several years and in contrasting environments. Boldrewood plants produced a higher ratio of desirable leaves to a smaller ratio of stem which is in line with the breeding targets set by this program. Thus, this new watercress has the more 'forkable' phenotype' of particular interest for use in bagged salads.

For phytonutrients, Control and Boldrewood appeared to be mostly comparable. Antioxidant (AO) capacity did not vary between accessions in either trial and this is in agreement with the previous comparisons (Payne et al., 2015). The indolic glucosinolate (GLS) 4-methoxy-glucobrassicin was significantly higher in Boldrewood than in Control watercress. This glucobrassicin derivative has been reported in watercress previously and is involved in plant defences (Kopsell et al. 2007). However, across accessions, gluconasturtiin concentrations were decisively higher compared to any other GLS detected and should be the focus of such comparisons in watercress. This observation is in agreement with the results of Rose et al. (2000), where the derivative of gluconasturtiin phenethyl isothiocyanate (PEITC) was found to be in concentrations three times higher than any other isothiocyanate. In human nutrition, both these GLS appear to play a role in cancer cell apoptosis and prevention (Cartea & Velasco 2008) and are therefore important breeding targets for watercress. The results presented here suggest that Boldrewood shows no loss of gluconasturtiin compared to the Control and has higher amounts of indolic GLS, suggesting that the health benefits to the consumer are at the very least preserved.

Perhaps our most important finding is the observation that stage of crop maturity has a significant effect on the phytonutritional profile. In Trial 1, AO capacity declined in the older Boldrewood samples, whereas gluconasturtiin concentrations increased. This increase in GLS was mirrored with a significant increase in the potency of the watercress as an inhibitor of cancer cell growth as the crop aged in Trial 2 – a novel finding which was unknown in watercress. Previously, studies have primarily linked AO capacity to polyphenolic compounds and ascorbic acid in watercress (Ninirola et al., 2014). AO capacity has shown variation in response to a number of factors. In watercress, AO capacity differed between spring and winter growing seasons (Niňirola et al. 2014); between parts of the plant and increased across 6 post-harvest days while refrigerated (Payne, 2011). GLS concentrations in watercress have been shown to increase in response to stressors/stimulants, specifically longer days, poor light quality, higher temperatures and soil nutrients (Engelen-Eigles et al. 2006; Kopsell et al. 2007).

A single study alone has measured change in growing watercress phytonutrients over time and charted a consistent increase in both ascorbic acid (an AO) and PEITC to a plateau at 40 days (Palaniswamy et al. 2003). This study however was limited to controlled conditions, which do not accurately reflect the commercial conditions (Payne et al., 2015). The current study is the first to examine change in AO and chemopreventive capacities over time in the watercress crop within a commercial setting. As the variation in GLS concentrations in plant tissue seems particularly species-specific across Brassicaceaes examined (Booth et al. 1991; Velasco et al. 2007; Wentzell & Kliebenstein 2008), this is an important finding for watercress. The implication is that eating watercress harvested from a more mature crop, versus the popular baby-leaf salads, could be linked to a greater chemopreventive effect. We speculate that extending the growth period could be further investigated as a method to increase the health benefit seen by the consumer for watercress.

The economic cost of growing Boldrewood on a longer rotation, instead of the current shorter cycle crop, would need to be examined in combination with market research to determine whether Boldrewood could be used singly to meet production demands or whether it would be better sold as a speciality product. This thesis cannot advise business on this further however, as the growth period of the crop watercress is already rather variable (4-7 weeks in the beds and defined by the farmer according to size specifications: baby leaf or regular) depending on location and weather conditions at that time, it would be advisable to explore time required to achieve a consistent nutritional quality across the entire growing season and in several locations.

While the focus of plant breeding has expanded in recent times, beyond aiming to benefit the farmer with a greater yield or less risky crop, to benefiting the consumer either nutritionally or functionally (Newell-McGloughlin 2008); breeding and releasing new crop variants to market requires a substantial investment. This study highlights the need for studies that examine and model the effect of environmental conditions and the G x E effect in nutritional quality of crops. Greater definition in our understanding of how phytonutrients vary throughout the food production chain based on environmental conditions and management practices could prove a cost-effective and quicker avenue than undertaking extensive breeding projects. A few interesting examples already available in the literature include improving nutritional quality by enhancing root and soil microbe interactions (Goicoechea & Antolín 2017), organic instead of conventional farming practices (Worthington 2001; Mitchell et al. 2007), and the application of certain stressors at particular developmental time points (Wang &

Frei 2011). This would be a timely approach as our understanding of the biosynthetic pathways delivering target compounds and the environmental, development and genetic factors that influence these has never been better and can inform the development of reasonable and testable hypotheses in the applied agricultural setting.

4.6 Conclusions

In conclusion, through extensive testing, we have established the dwarf growth form and consistently high and beneficial phytonutrient character of the new Boldrewood watercress. Boldrewood may replace the current commercial watercress to meet consumer demand for a salad leaf with less stem. In the commercial environment, Boldrewood was often harvested up to 14 d later that the usual crop, due to its smaller size and that these additional days of growth resulted in a crop with increased chemopreventive benefits in trials using human cancer cells. This new watercress offers significant potential to grow a slower crop, for up to two weeks longer than the traditional harvest date, but with a significant improvement in phytonutrition, whilst conforming to current market size specifications.

4.7 References

Benzie, I. F., & Strain, J. J. (1996). The ferric reducing ability of plasma (FRAP) as a measure of “antioxidant power”: the FRAP assay. *Analytical Biochemistry*, *239*(1), 70–76.

Bones, A. M., & Iversen, T.-H. (1985). Myrosin cells and myrosinase. *Israel Journal of Botany*, *34*(2–4), 351–376.

Bones, A. M., & Rossiter, J. T. (1996). The myrosinase-glucosinolate system, its organisation and biochemistry. *Physiologia Plantarum*, *97*(1), 194–208.

Booth, E. J., Walker, K. C., & Griffiths, D. W. (1991). A time-course study of the effect of sulphur on glucosinolates in oilseed rape (*Brassica napus*) from the vegetative stage to maturity. *Journal of the Science of Food and Agriculture*, *56*(4), 479–493.

Cartea, M. E., & Velasco, P. (2008). Glucosinolates in Brassica foods: bioavailability in food and significance for human health. *Phytochemistry Reviews*, *7*(2), 213–229. https://doi.org/10.1007/s11101-007-9072-2

Casey, H. (1981). Discharge and chemical changes in a chalk stream headwater affected by the outflow of a commercial watercress-bed. *Enviromental Pollution*, *2*, 373–385.

Casey, H., & Smith, S. M. (1994). The effects of watercress growing on chalk headwater streams in Dorset and Hampshire. *Environmenral Pollution*, *85*, 217–228.

Cavell, B. E., Syed Alwi, S. S., Donlevy, A. M., Proud, C. G., & Packham, G. (2012). Natural product-derived antitumor compound phenethyl isothiocyanate inhibits mTORC1 activity via TSC2. *Journal of Natural Products*, *75*(6), 1051–1057.

Di Noia, J. (2014). Defining powerhouse fruits and vegetables: a nutrient density approach. *Prev Chronic Dis*, *11*, 130390.

Dwivedi, S., Sahrawat, K., Upadhyaya, H., & Otriz, R. (2013). Food, Nutrition and Agrobiodiversity Under Global Climate Change. In D. L. Sparks (Ed.), *Advances in Agronomy* (pp. 1–128). Elsevier Inc.

Engelen-Eigles, G., Holden, G., Cohen, J. D., & Gardner, G. (2006). The effect of temperature, photoperiod, and light quality on gluconasturtiin concentration in watercress (*Nasturtium officinale* R. Br.). *Journal of Agricultural and Food Chemistry*, *54*(2), 328–334.

Fogarty, M. C., Hughes, C. M., Burke, G., Brown, J. C., & Davison, G. W. (2013). Acute and chronic watercress supplementation attenuates exercise-induced peripheral mononuclear cell DNA damage and lipid peroxidation. *The British Journal of Nutrition*, *109*(2), 293–301.

Gill, C. I. R., Haldar, S., Boyd, L. a, Bennett, R., Whiteford, J., Butler, M., … Rowland, I. R. (2007). Watercress supplementation in diet reduces lymphocyte DNA damage and alters blood antioxidant status in healthy adults. *The American Journal of Clinical Nutrition*, *85*(2), 504–510.

Goicoechea, N., & Antolín, M. C. (2017). Increased nutritional value in food crops. *Microbial Biotechnology*, *10*(5), 1004–1007. https://doi.org/10.1111/1751-7915.12764

Hecht, S., Chung, F., Richie, J. J., Akerkar, S., Borukhova, A., Skowronski, L., & Carmella, S. (1995). Effects of watercress consumption on metabolism of a tobacco-specific lung carcinogen in smokers. *Cancer Epidemiol. Biomarkers Prev.*, *4*(8), 877–884.

Kopsell, D. a, Barickman, T. C., Sams, C. E., & McElroy, J. S. (2007). Influence of nitrogen and sulfur on biomass production and carotenoid and glucosinolate concentrations in watercress (*Nasturtium officinal*e R. Br.). *Journal of Agricultural and Food Chemistry*, *55*(26), 10628–10634.

Manchali, S., Chidambara Murthy, K. N., & Patil, B. S. (2012). Crucial facts about health benefits of popular cruciferous vegetables. *Journal of Functional Foods*, *4*(1), 94–106. https://doi.org/10.1016/j.jff.2011.08.004

Manton, I. (1935). The cytological history of Watercress (*Nasturtium officinale* R. Br.). *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, *69*(1), 132–157.

Martínez-Sánchez, A., Gil-Izquierdo, A., Gil, M. I., & Ferreres, F. (2008). A comparative study of flavonoid compounds, vitamin C, and antioxidant properties of baby leaf Brassicaceae species. *Journal of Agricultural and Food Chemistry*, *56*(7), 2330–2340.

Mitchell, A. E., Hong, Y.-J., Koh, E., Barrett, D. M., Bryant, D. E., Denison, R. F., & Kaffka, S. (2007). Ten-year comparison of the influence of organic and conventional crop management practices on the content of flavonoids in tomatoes. https://doi.org/10.1021/JF070344+

Natural England. (2009). *Watercress growing and its environmental impacts on chalk rivers in England (NECR027)*. *www.naturalengland.org.uk*.

Newell-McGloughlin, M. (2008). Nutritionally improved agricultural crops. *Plant Physiology*, *147*(3), 939–53. https://doi.org/10.1104/pp.108.121947

Newman, R. M., Hanscom, Z., & Kerfoot, W. C. (1992). The watercress glucosinolate-myrosinase system: a feeding deterrent to caddisflies, snails and amphipods. *Oecologia*, *92*, 1–7.

Niňirola, D., Fernández, J. A., Conesa, E., Martínez, J. A., & Egea-Gilabert, C. (2014). Combined effects of growth cycle and different levels of aeration in nutrient solution on productivity, quality, and shelf life of watercress (*Nasturtium officinale* R. Br.) plants. *HortScience*, *49*(5), 567–573.

Palaniswamy, U. R., McAvoy, R. J., Bible, B. B., & Stuart, J. D. (2003). Ontogenic variations of ascorbic acid and phenethyl isothiocyanate concentrations in watercress (*Nasturtium officinale* R.Br.) leaves. *Journal of Agricultural and Food Chemistry*, *51*(18), 5504–5509.

Payne, A. C. (2011). *Harnessing the Genetic Diversity of Watercess (Rorippa nasturtium - aquaticum) for Improved Morphology and Anti- cancer Benefits: Underpinning Data for Molecular Breeding [doctoral thesis]*. Southampton: University of Southampton.

Payne, A. C., Clarkson, G. J. J., Rothwell, S., & Taylor, G. (2015). Diversity in global gene expression and morphology across a watercress (*Nasturtium officinale* R. Br.) germplasm collection: first steps to breeding. *Horticulture Research*, *2*, 15029. https://doi.org/10.1038/hortres.2015.29

Payne, A. C., Mazzer, A., Clarkson, G. J. J., & Taylor, G. (2013). Antioxidant assays - consistent findings from FRAP and ORAC reveal a negative impact of organic cultivation on antioxidant potential in spinach but not watercress or rocket leaves. *Food Science & Nutrition*, *1*(6), 439–44.

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

Rose, P., Faulkner, K., Williamson, G., & Mithen, R. (2000). 7-Methylsulfinylheptyl and 8-methylsulfinyloctyl isothiocyanates from watercress are potent inducers of phase II enzymes. *Carcinogenesis*, *21*(11), 1983–1988.

Sadeghi, H., Mostafazadeh, M., Sadeghi, H., Naderian, M., Barmak, M. J., Talebianpoor, M. S., & Mehraban, F. (2013). In vivo anti-inflammatory properties of aerial parts of *Nasturtium officinale*. *Pharmaceutical Biology*, (2008), 1–6. https://doi.org/10.3109/13880209.2013.821138

Santos, J., Oliveira, M. B. P. P., Ibáñez, E., & Herrero, M. (2014). Phenolic profile evolution of different ready-to-eat baby-leaf vegetables during storage. *Journal of Chromatography. A*, *1327*, 118–131.

Simon, P. W. (2010). *Domestication, Historical Development, and Modern Breeding of Carrot*. *Plant Breeding Reviews* (Vol. 19). https://doi.org/10.1002/9780470650172.ch5

Slafer, G. A., Satorre, E. H., & Andrade, F. H. (1994). Increases in grain yield in bread wheat from breeding and associated physiological changes. In *Genetic Improvement of Field Crops* (pp. 1–68).

Traka, M. H., & Mithen, R. (2008). Glucosinolates, isothiocyanates and human health. *Phytochemistry Reviews, 8*(1), 269–282.

Traka, M. H., Saha, S., Huseby, S., Kopriva, S., Walley, P. G., Barker, G. C., … Mithen, R. F. (2013). Genetic regulation of glucoraphanin accumulation in Beneforté broccoli. *The New Phytologist*, *198*(4), 1085–1095.

Velasco, P., Cartea, M. E., Gonzalez, C., Vilar, M., & Ordas, A. (2007). Factors affecting the glucosinolate content of kale (*Brassica oleracea* acephala group). *Journal of Agricultural and Food Chemistry*, *55*(3), 955–962.

Voutsina, N., Payne, A. C., Hancock, R. D., Clarkson, G. J., Rothwell, S. D., Chapman, M. A., & Taylor, G. (2016). Characterization of the watercress (*Nasturtium officinale* R. Br.; Brassicaceae) transcriptome using RNASeq and identification of candidate genes for important phytonutrient traits linked to human health. *BMC Genomics*, *17*, 378. https://doi.org/DOI 10.1186/s12864-016-2704-4

Wang, Y., & Frei, M. (2011). Stressed food - The impact of abiotic environmental stresses on crop quality. *Agriculture, Ecosystems and Environment*. https://doi.org/10.1016/j.agee.2011.03.017

Wentzell, A. M., & Kliebenstein, D. J. (2008). Genotype, age, tissue, and environment regulate the structural outcome of glucosinolate activation. *Plant Physiology*, *147*(1).

Worthington, V. (2001). Nutritional quality of organic versus conventional fruits, vegetables, and grains. *The Journal of Alternative and Complementary Medicine*, *7*(2), 161–173. https://doi.org/10.1089/107555301750164244

Chapter 5: General Discussion

5.1 Progress in Watercress Research and Breeding

Watercress (*Nasturtium officinale* R. Br.) is an aquatic herbaceous plant and member of the prominent Brassicaceae family (Howard & Lyon 1952). It is a tetraploid with 32 chromosome (n=8) of unknown diploid ancestry from within the Brassicaceaes (Bleeker et al. 1999) and a genome size of 743 Mbp (Morozowska et al. 2010). Human use of watercress as a medicinal or food crop dates back 2,000 years (Manton 1935). In modern times, it is consumed typically in fresh salads or soup in Europe and Asia but with increasing global popularity due to reports of the wealth of health benefits associated with its consumption (Palaniswamy & McAvoy 2001). Watercress is rich in essential dietary vitamins and minerals, and non-essential antioxidants and glucosinolates, and was recently ranked as top ‘powerhouse’ fruit or vegetable to consume (Di Noia 2014). These non-essential components of watercress are thought to grant antioxidant (Gill et al. 2007; Fogarty et al. 2013), anti-inflammatory (Sadeghi et al. 2013), and chemopreventive (Cavell et al. 2012) benefits to the consumer but are also responsible for the crop’s particular peppery flavour (Palaniswamy & McAvoy 2001). Despite this, watercress is a small, non-model crop dependent solely on traditional expertise and experience and with no active breeding or resources. This has been a hindrance to the industry’s ability to adapt to market changes or face new challenges (Sheridan et al. 2001).

Fuelled by the commercial sector’s interest in developing new and improved cultivars and the scientific motivation to understand the genetic properties of this unique crop, a breeding and research program was set up at the University of Southampton (UoS), U.K. (Payne 2011). A germplasm collection was established, from sources around the world, and set the foundation for the study and breeding of watercress. Breeding targets have been defined as: enhanced nutrient-density, an updated crop morphology with less stem for the salad market, and improved resource-use efficiency. The UoS collection was phenotyped and contained significant variation for agronomic traits of interest, such as stem morphology and antioxidant capacity (Payne et al. 2015). Genotyping of the collection, using Amplified Fragment Length Polymorphisms (AFLPs), showed that genetic variation was still present within each accessions of the collections and had not been reduced by selection pressures (Payne 2011). The comparison of gene expression between three outstanding accessions also identified exciting variation in the expression of pathways linked to stem morphology and antioxidant capacity. An accession, named ‘Boldrewood’, was singled out as a promising candidate for commercial application as a dwarf variety (Payne et al. 2015).

These insights, coupled with the onset of broadly-accessible Next Generation Sequencing (NGS) services for rapid genetic marker discovery, unlocked the possibilities for development of novel and valuable genomic resources for the description and breeding of watercress. With this backdrop, the work in this thesis was set out and has achieved the following novel outputs for watercress: the sequencing, assembly and annotation of the watercress transcriptome; the production of a mapping population and completion of comprehensive genotyping and phenotyping, which identified transgressive segregation in the F_2 offspring; the construction of a linkage map; the application of Quantitative Trait Loci (QTL) and Differential Expression analysis to identify the QTL and differentially expressed (DE) genes underpinning targeted traits of interest; and the development of two extensive marker datasets through RNASeq and Genotyping-By-Sequencing (GBS) that will facilitate future studies and the identification of allelic variation in the crop. In addition, the inhibitory effect of watercress on breast cancer cells was also mapped to the watercress genome. To the best of our knowledge, this is the first attempt to directly map this important health benefit to the genome of the source plant. Finally, commercial trials were conducted to test the suitability of ‘Boldrewood’ for commercialization, providing direct insight into the character of this ‘new’ watercress cultivar for growers. During these trials, previously unknown temporal variation in phytonutrient traits was described for the first time in the commercial setting with implications for the relationship between cultivation practices and consumer health benefit.

5.2 Broader Topics Touched Upon During These Studies

The ability to understand the relationship between phenotype and genotype in order to manipulate the first is at the heart of plant breeding. This has become a great deal more feasible with the development of modern genetic markers and genomics- based approaches. NGS in particular has made it possible to sequence and genotype large numbers of markers inexpensively, proving an avenue for exponential development of resources across species but especially for small-scale and non-model crops (Varshney et al. 2009). Many crops were till now considered difficult to work with because of large genome size, polyploidy, long life cycles, lack of a reference genome and financial backing to tackle these issues. In this context, the resources presented here for watercress consist of a significant advancement of breeding potential through genomics-led and marker-based pathways, such as Marker Assisted Selection (MAS), and showcase what is possible with modern technologies when starting from scratch and despite a polyploid genome. An extensive amount of sequencing was completed here for Chapters 2 & 3. In Chapter 2, a total of 323, 827,923 fragments of RNA were sequenced (approximately 100 bp in length) in twelve individuals using RNASeq. In Chapter 3, a total of 940,208,464 fragments of DNA (approximately 85 bp in length) were sequenced across 280 individuals using GBS. The grand total surpassed 112 billion bases of watercress sequenced. Although in the past the lack of a reference genome was limiting to such works, bioinformatic pathways are now available to make the most of this data *de novo*. In terms of markers developed, the watercress transcriptome produced here was assembled into 80,800 transcripts (48,732 unigenes) and GBS produced a total of 9,689 loci. Within these contigs, candidates for genes in pathways of known significance to traits of interest were mined for and markers with a strong statistical relationship to these traits have been highlighted. These tools could be further explored in research or put to use immediately in breeding through MAS.

Despite the great promise of advances in plant genomics and the outstanding figures listed above, species genetics remains a limiting factor. The genetic diversity across watercress commercial stock was found to be limited by Sheridan et al. (2001) but the UoS germplasm collection was expected to contain greater variability based on a broader range of origin and previous phenotypic/genotypic comparison within the collection (Payne 2011). However, only 22 % of RNASeq transcripts, originating from twelve individuals from different accessions, were polymorphic and nucleotide diversity indices were also low across the dataset (Chapter 2). And, despite the parents of the mapping population being selected from opposite ends of the phenotypic spectrum and showing the least genetic relatedness for RNASeq-derived markers, the number of loci with SNPs in the mapping population was also disappointingly low. Of 9,689 loci, less than 6 % (561 loci) contained a SNP between the two parents and could be used for linkage mapping and QTL analysis. A number of other studies in crops, which applied GBS or the alternative Restriction-site Associated DNA (RAD) Sequencing, have produced greater number of markers possibly because of the use of a smaller population size and greater depth of sequencing, populations with greater levels of polymorphism, or more successful DNA extraction protocols, and have consequently used more markers for downstream applications (See Table 1.2). Specifically, for these studies in Table 1.2, a mean of 27% of markers available were used for linkage mapping in comparison to the 6% used here. The percentage of usable markers varies greatly from study to study and the results presented in Chapter 3 are comparable to some, such as the Lambel et al. (2014) study which applied GBS in watermelon. In this case GBS produced 527,844 loci, 379,125 were successfully aligned to the reference genome but only 266 were used for linkage mapping; yet this did not inhibit the identification of a significant QTL for resistance (Lambel et al. 2014). In this thesis, the production of viable resource for watercress were also not inhibited by the small number of usable markers and the lack of a reference genome. However, they are a strong reminder that expectations and targets must still be moderated by the genetic limitations of each particular project and species; and won't necessarily be resolved by large scale sequencing.

Historically, crops with a larger market have been studied for decades through pre-NGS approaches and improved through traditional breeding. Large databases and long-term projects have been required to target traits of interest. In this work, we have presented a case of study where little data predates the work and a combination of NGS-facilitated techniques were applied to create data and resources from scratch. This modern approach, although not comprehensive or final, has been very successful at producing reusable and valuable resources for any future work on the crop and has taken a much smaller investment in both time and money by stakeholders than what might have been required historically. Whilst other groups take similar approaches in the study of other crops, the next few decades are likely to be an informatic revolution in plant sciences and crop breeding.

In recent times, plant breeding objectives have shifted from securing or increasing the value of the farmer's investment to enhancing the nutritional quality, appeal or functionality of a crop in the eyes of the consumer. This shift is particularly important for society in general because a large range of health problems are rooted in malnutrition and the consumption of plant-based diets has not been successfully increased by campaigning and public education; breeding healthier crops is a viable route to improved human health (Patil et al. 2014). A few successful examples of breeding for enhanced nutritional value in crops include Beneforte broccoli (Traka et al. 2013), Vitamin A-enhanced Golden rice (Ye et al. 2000; Beyer et al. 2002), and orange-skinned sweet potatoes which are rich in β-carotene (Laurie et al. 2015), with further examples are compiled in Newell-McGloughlin (2008). Despite the need to improve the process of breeding for nutritionally superior crops, by defining optimal nutrient levels, optimized post-harvest practices and improved quantification of the true effect, if any, of these crops in the population (Patil et al. 2014); these examples show that it is possible to nutritionally enhance staple crops by breeding for select biosynthetic pathways linked to a target trait. Modern phenotyping capabilities and NGS technologies, coupled with computational power and bioinformatics hold the additional promise that once nutritional candidate genes are identified, they can be quickly found and evaluated across species and phyla, further increasing the breeding value of that gene (Newell-McGloughlin 2008). For example, all the transcripts identified in the glucosinolate biosynthetic pathway of radish had a similar transcript in watercress (Chapter 2), highlighting the possible parallel targets in breeding.

The findings of the genomic studies described in this thesis are a strong example of the ability to apply modern technologies to great progress in understudied or orphaned crops and provide a direction for breeding for each defined target trait, as well as insight into modification of current cultivation practices that may enhance crop nutritional quality. The phytonutrient profile of watercress was quantified here in three ways: 1) total antioxidant (AO) capacity of watercress sap, 2) glucosinolate (GLS) concentrations of watercress tissue, and 3) the potency of watercress sap as an inhibitor of human cancer cells (IC_{50}). Traits pertaining to the health benefit of the consumer are often directly related to secondary plant metabolites, which protect the plant from environmental threats or enhance its interaction with the external environment. For example, GLS play a role in the deterrence of herbivory (Newman et al. 1992). The role of these compounds inherently suggests a strong environmental component in their synthesis and accumulation, making such traits particularly complex to predict or explain. The ability to explain complex traits depends on a number of factors, including the genetic architecture of the trait: its heritability, the number of loci and the size of each effect, and the approach to its characterization: the population size, number of markers and their density (Hayes et al. 2010). It is often the case that complex traits, such as the phytonutrient profile of a crop, will be controlled by a large number of small effect polymorphisms making it difficult to confidently address the traits determinants. Understanding the architecture of these traits is critical to our ability to select components to breed for.

The AO capacity of food crops appears to be an example of such a trait. Due to the lack of compound specificity of AO assays in this work, this trait characterization cast a broad net across many potential secondary plant metabolites, or phytonutrients, that have AO properties and might benefit the consumer. Therefore, it is probably not surprising that the 145 DE genes detected in Chapter 2 fell within 23 Gene Ontology (GO) categories relating to plant immunity, stress and stimuli response. These included genes directly involved in known AO biosynthesis pathways such as phenylpropanoid, carotenoid and tocopherol biosynthesis. One example is the enzyme ferulate 5-hydroxylase (*AT4G36220*), which is involved in lignin, sinapic acid and anthocyanin biosynthesis (Ruegger et al. 1999; Humphreys & Chapple 2002; Maruta et al. 2014) and was also recently identified as a candidate gene for breeding improved AO in lettuce (Damerum et al. 2015). A minor QTL for AO capacity was also identified in Chapter 3, however only one marker underlying the QTL showed similarity to known genes; a cyclin (*AT5G11300*). Interestingly, a link of AO compounds to cell cycling and cell wall functions has been made through the wounding response in plants, involving cyclin-dependent kinases (Neubauer et al. 2012). Overall, the AO trait appeared to have low heritability across the mapping population and was found to be strongly altered by the environment of cultivation. Payne et al. (2015) found AO capacity of the UoS watercress collection to differ significantly between field and controlled environments and the same was true for the AO capacity of the mapping population in Chapter 3. Studies of AO characters in oilseed rape, tomato and lettuce have also identified a low trait heritability and a strong G x E effect (Marwede et al. 2005; Rousseaux et al. 2005; Hayashi et al. 2012). Rousseaux et al. (2005) found that only 35% of antioxidant QTL were consistent in 2 out of 3 field trials. This trait seems particularly susceptible to or stimulated by environmental variation and here watercress plants grown under variable, natural conditions consistently are characterized by much higher AO capacity, and thus a higher health benefit for the consumer. These findings highlight two important broader considerations: firstly, the significance of completing field-based versus controlled-condition trials in plant sciences and, secondly, the role of environmental stimulation as a component to production of nutrient-dense crops.

Recently, awareness has grown that traditional studies investigating plant responses to single stressors under controlled conditions are not as useful as intended, because they cannot be combined to predict plant responses or phenotypes under multiple stresses or natural conditions (Plessis et al. 2015; Prasch & Sonnewald 2015; Thoen et al. 2016). Gene expression under natural conditions has been limited by analytical and computational capabilities, yet the majority *Arabidopsis thaliana* (L.) Heynh genes are expressed differently under natural conditions than in the laboratory (Richards et al. 2012). This suggests an overarching need for molecular plant sciences to work under ‘real-world’ settings in order to produce meaningful knowledge that is informative for applied branches, such as crop improvement or environmental conservation. In Chapter 2, whole-transcriptome sequencing in watercress was carried out on plants grown under commercial conditions so that any results were closely linked to what the consumer was exposed to. In addition, progress is being made to increase computational power and analytical capabilities so as to handle greatly complex, multi-trait and multi-stress datasets. Plessis et al. (2015) were able to identify multiple groups of co-expressed genes from RNASeq in field experiments, relate them to environmental variables, and then use this information to model plant responses in a different dataset. Thoen et al. (2016) combined results from multiple stress or multi stress experiments and developed a complex, mixed model, genome-wide approach toward identifying QTL, candidate genes and their interactions across experiments. The capability to take molecular plant science to the natural world through genomics and bioinformatics is particularly exciting for applied crop sciences and can produce valuable insight into intricate pathways affecting breeding targets.

The effect of various environmental variables on the phytonutrient profile of a crop was discussed in Chapter 1.3.6 and is acknowledged as a key component in producing nutrient-dense food stuff. Based on the environmentally-driven differences encountered throughout this work (Chapter 3 and 4) and although it would require adding levels of complexity to the conceptual model of how to deliver and classify nutrient-rich crops, it seems important to at least consider the issue holistically. Hennig et al. (2014) similarly suggested a joint approach of plant and food scientist in delivering a healthy diet. In the case of watercress, a raw leafy salad, where the environment seems to play a major role in the output of phytonutrient traits; is a stimulated or a crop genetically predisposed with greater 'sensitivity' on the farm correlated to a healthier crop on the plate? Answering these types of questions would need to engage the entire production chain: from crop sciences, to encompassing the effect of processing, through to nutrigenomics - studying the relationship between human nutrition and the human genome (Mutch et al. 2005). Cohort studies are the primary methodology of identifying the health benefit associated with vegetable consumption, specifically looking into the occurrence of chronic disease across a sample population. More direct examples include the quantification of antioxidant compounds in the blood of study participants (Gill et al. 2007), protective effects of a dietary treatment on exercise-induced DNA damage (Fogarty et al. 2013), the effect of dietary treatment on the gut microbiome (F. Li et al. 2009) and the quantification of toxins or other compounds of interest in the urine of study participants (Hecht et al. 1995). However, none of these approaches have considered the genetic makeup of the dietary treatment itself. Particularly in light of new information regarding the differences in how individuals respond to diet or medical treatment, called nutrigenetics (Mutch et al. 2005; Newell-McGloughlin 2008), this component of the equation may be particularly important. Overall, each of component of food production has been studied independently but ultimately contribute and interact as a whole to produce the phytonutrient phenotype consumed and the benefit gained (Figure 5.1).

In this work, an attempt was made to bridge some of the gap between the crop and consumer health benefit by quantifying the effect of the watercress mapping population, grown in two contrasting environments, directly on human cancer cells as a phenotype of the plant (Chapter 3). To our knowledge, this is first attempt to identify QTL and genetic variation for such a trait. We identified three QTL for the chemopreventive potency of watercress (IC_{50}), explaining between 5- 7.8 % (total of approximately 20 %) of phenotypic variation, and found sequence similarity of two markers underlying one of the QTLs with plant defence genes, suggesting the possibility of a defence gene cluster (Chapter 3). It would be particularly interesting to apply these types of approaches to the broader questions and targets of nutritious crop breeding and begin to elucidate the complex relationship between genotype, phenotype, environment and dietary-gained human health benefits.

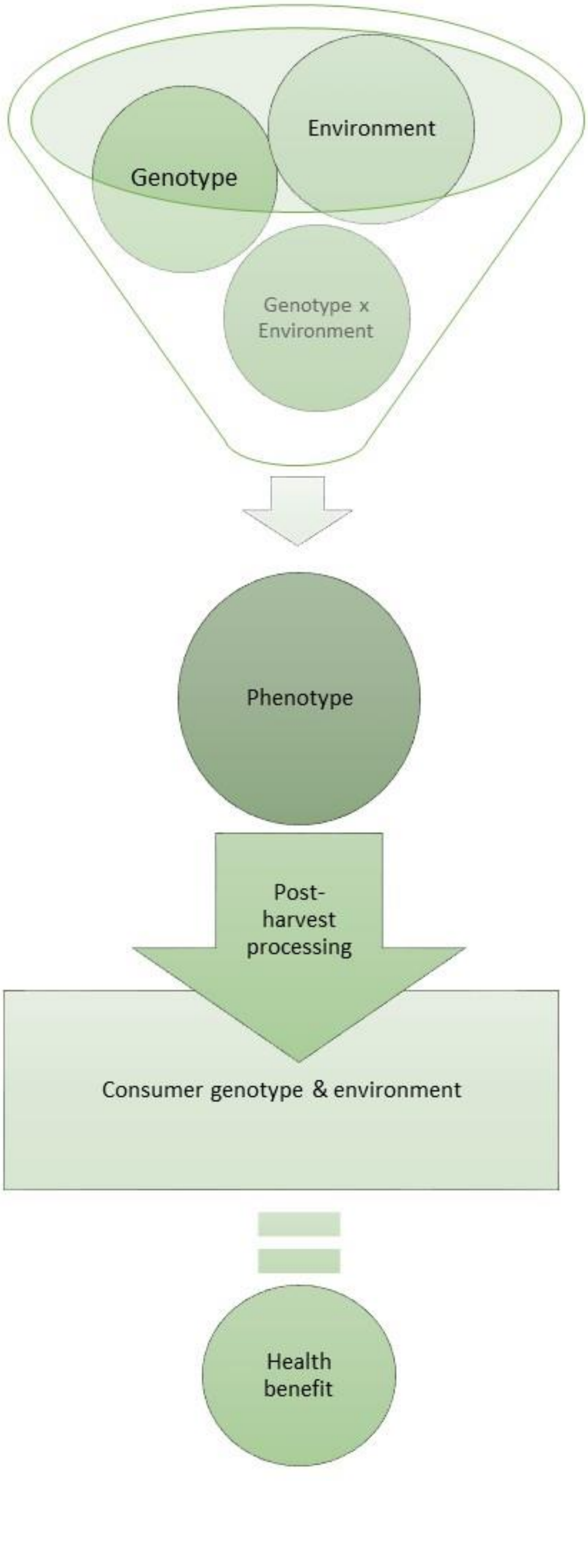


Figure 5.1 A conceptual schematic of the variables and processes affecting the ultimate target of breeding crops that deliver human health

5.3 Future Directions for Watercress Breeding

The level of consensus DNA and RNA sequences encountered in Chapter 2 & 3 would suggest that there is limited genetic variation within the current watercress stock available for study. Therefore, the significant amount of phenotypic variation also present between accessions and parents must be a results of phenotypic plasticity – the ability of a genotype to produce differing phenotypes under alternative environmental conditions (Bradshaw 1965) – or possibly epigenetic differences (Wong et al. 2005). In fact, Howard & Lyon (1952) described watercress as a very plastic species long ago, differing greatly in height based on the amount of water available. What considerations and approaches would be suitable for breeding within this context of low genetic diversity and high phenotypic variability between accessions?

Small gains can be made when breeding is targeted to particular environments with environment-specific QTLs. El-Soda et al. (2014) review the effect of phenotypic plasticity on breeding and QTL mapping, referred to as QTL x E. The authors point out that when there is a strong environmental effect on complex traits, cross-environmental or ‘constitutive’ QTL are unlikely and that the best targets will be ‘neutral’ QTL, that secure a gain in the priority environment of cultivation and do not have negative effects in other environments (El-Soda et al. 2014). Therefore, dissecting the data in an environment by environment fashion is appropriate and requires little layering on of unnecessary analytical complexity. If a reaction norm, or a pattern of how the phenotype changes with across environments for a particular genotype, can be defined then that could be used in the place of discrete traits in QTL analysis as well (Stratton 1998; El-Soda et al. 2014).

In addition, the use of techniques that incorporate gene expression may prove highly suitable for watercress. RNASeq was applied in Chapter 2 to account for phenotypic differences and efficiently produced promising candidate genes. Here, expression QTL (eQTL) analysis may also be fruitful. This approach treats gene expression as the phenotype for QTL analysis and is focused on identifying loci that regulate gene expression of complex traits (Gilad et al. 2008; Kliebenstein 2009). This approach is feasible because gene expression has proven to be surprisingly heritable (Gilad et al. 2008; West et al. 2007). A good example of this application – and in fact relevant to breeding target of resource-use efficiency in watercress – utilized eQTL mapping against phosphorus (P) availability in *Brassica rapa* L. (Hammond et al. 2011). The study identified a genomic hotspot enriched for genes linked to P metabolism, collocating QTL and eQTL, and concluded that the expression differences associated with soil P are highly heritable and would make good breeding targets (Hammond et al. 2011).

Low genetic diversity between parents did not hinder efforts to explain phenotypic variation in this work. In Chapter 2, genes with known functions in related biosynthetic pathways were mined for, based on sequence similarity to *Arabidopsis,* and these candidates are immediately useful in hunting for allelic variation across the watercress material. The QTL identified in Chapter 3 offer a first look at the genetic architecture of important traits in watercress and account for decent percentages of the phenotypic variation in the traits (Chapter 3). These results are sufficient for breeders to make improvements to the crop through basic MAS and for scouting for new allelic variation across the germplasm collections. However, further studies in this population to test the robustness of the QTL, the identification of candidate genes and their validation could be completed.

Several approaches are possible to examining the stability of these QTL and these are explored here. Examining the ability of the markers to predict phenotype in the F_3 progeny or other material from the collection would be one option (Collard et al. 2005). Taking the mapping population forward to later generations and repeating QTL analysis would also improve resolution, as higher QTL resolution is derived from a greater numbers of recombination break points (The Complex Trait Consortium 2003). The primary QTL identified could also be fine mapped to further resolve the underlying regions and discover causal candidate genes. As watercress does not have a genome sequence or physical map available, similarities with an annotated close relative would be required to putatively describe any gene function. *Arabidopsis* is the best described relative but recently the genome of close relative *Barbarea vulgaris* was also drafted (Byrne et al. 2017). *B. vulgaris* is increasingly used in studies of metabolomics and pest-resistance (Kuzina et al. 2011) and resources developed may become useful in watercress based on close genome synteny. Therefore, a suitable approach to fine mapping these QTL might be to clone and sequence the current LOD-support region for target QTLs, annotating the resulting sequence based on similarity to *Arabidopsis*; and candidate gene mining for causal candidates (Salvi & Tuberosa 2005). However, as a 10 cM region in *Arabidopsis* can include 440 genes, the size of each primary QTL will influence whether this can be done directly of whether further work, such as the development of Near Isogenic Lines (NILs), which only vary for the target genomic region, is needed to define a manageable cloning target (Salvi & Tuberosa 2005). A different approach is to carry out differential expression analysis for individuals with differing genotypes for the particular QTL of interest, thus using gene expression analysis to detect the most likely candidate genes (Gelli et al. 2014; Jian et al. 2017). This might be particularly suited to watercress for which gene expression appears to play a significant role in the phenotype.

Although the validation of candidate markers is not essential for breeding as a robust marker is sufficient for MAS; it would contribute significantly to resolving the genetic architecture and contributing factors behind a trait. When a reference genome exists and candidate genes can be visually selected, they are commonly validated at first instance with quantitate PCR (qPCR). RNAi can also be used to silence the expression of a candidate gene and examine its role (Waterhouse & Helliwell 2003). Transgenic approaches are typically be used to resolve the function of a gene (Moose & Mumm 2008) and watercress is malleable to transformation, as protocol are already established (Jin et al. 1999; Park et al. 2011). The current development of the highly efficient and accurate 'genome editing' technologies may prove to be the most efficient methodology to prove function of the candidates genes, or for the development of new cultivars - pending classification of their products as Genetically Modified (GMOs) or not (Schaart et al. 2016; Scheben et al. 2017).

It is anticipated that the mapping population developed for this work (Chapter 3) will be carried forward to produce a permanent resource of Recombinant Inbred Lines (RILs) for watercress. Once RILs are established, they will hopefully enable many more genomic studies in this species. Along the way to RILs, several techniques could be applied to deliver further information on these or other traits that were not mapped in this thesis, i.e. glucosinolate content and P-use efficiency. These include repeating QTL analysis using the current map and genotype information, but with phenotypic means collected from the $F_{2:3}$ families (Austin & Lee 1996), offering the opportunity to expand QTL detection to further environments or multiple years. The limitation of this application is in the availability of seed at these early stages, as adequate replication is needed for $F_{2:3}$ families while in parallel enough seed must be guaranteed to take RILs forward. On the other hand, watercress benefits from the ability to grown clonally and tested in multiple environments, as we have shown in Chapter 3, potentially addressing this limitation.

Bulked Segregant Analysis (BSA) and selective genotyping are two other approaches that could be taken to utilize this existing watercress mapping population in new traits and generations. Both techniques require a mapping population and use smaller pools of individuals with strong phenotypic differentiation to pin down polymorphic markers that are associated to the trait of interest; reducing costs at a loss of descriptive information regarding the QTL (Collard et al. 2005). BSA has now been adapted further with the use of RNASeq (BSRSeq) in discovering markers (Ramirez-Gonzalez et al. 2015), as was done recently to discover critical genes to waterlogging in maize (Du et al. 2017).

An alternative to QTL mapping for candidate gene discovery is association mapping. Association mapping detects the QTL associated with a trait based on linkage disequilibrium, or the non-random relationship between markers, across a collection of naturally-occurring genotypes, instead of within a family (Langridge & Fleury 2011). Any naturally evolved variation in the watercress germplasm collection that contributed to the overlaying phenotypic diversity could be highlighted through a Genome-Wide Association Study (GWAS). However, given that GWAS requires a large number of diverse genotypes for sufficient resolution (Varshney et al. 2009; El-Soda et al. 2014), it is not convincing that GWAS would produce better results for the current watercress germplasm collection that the mapping population approach utilized here.

Finally, expanding the genetic diversity available in the UoS germplasm collection should be a priority, as it has the potential to quickly increase the genetic diversity available to study and breed with. The collection could be improved by the addition of wider-reaching global accessions and through the exploration of the natural origins of watercress, thought to be Eastern Europe and the Middle-East (Bleeker et al. 1999). Wild-collected material may have undergone more outcrossing events, in comparison with cultivated material, and could contain more varied polymorphisms inducted by adaptation to local habitats and stressors. The addition of closely-related species should also be considered to introduce greater genetic diversity. Interspecific crosses were used in combination with molecular technologies in *B. napus* (Obermeier & Friedt 2015). Mutagenesis and transgenic work is thought to be the fastest way to introduce new genetic variation in a crop (Moose & Mumm 2008), however material targeted toward commercialization needs to be sensitive to regulations regarding GMOs or the adverse reaction of potential consumers to the breeding procedures used. Transgressive segregation, following a cross is critical, to mixing up alleles and generating new combinations in natural conditions and breeding projects (Rieseberg et al. 1999). As was observed in Chapter 3, it is a feasible option for reaching new phenotypic ranges in watercress breeding, despite apparent limitations in genetic diversity across the cultivated crop.

5.4 Conclusions

In conclusion, a valuable new set of genomic resources has been developed for watercress. This work has elucidated the genetic structure of important breeding characters, including the complex trait of chemopreventive potency of watercress in human cells. It has also uncovered important information regarding the effect of environmental variables on watercress characters and the genetic structure of the crop itself, through examining accessions and a segregating family in the commercial setting. In this final discussion, outstanding themes have been considered and tools for future work have been presented. This information can now inform and equip the next generation of watercress breeding and research.

5.5 References

Austin, D. F., & Lee, M. (1996). Comparative mapping in F2:3 and F6:7 generations of quantitative trait loci for grain yield and yield components in maize. *Theoretical and Applied Genetics*, *92*(7), 817–826. https://doi.org/10.1007/BF00221893

Beyer, P., Al-Babili, S., Ye, X., Lucca, P., Schaub, P., Welsch, R., & Potrykus, I. (2002). Golden Rice: introducing the beta-carotene biosynthesis pathway into rice endosperm by genetic engineering to defeat vitamin A deficiency. *The Journal of Nutrition*. https://doi.org/10.1093/jn/132.3.506S

Bleeker, W., Huthmann, M., & Hurka, H. (1999). Evolution of the hybrid taxa in Nasturtium R.BR. (Brassicaceae). *Folia Geobotanica*, *34*, 421–433.

Bradshaw, A. D. (1965). Evolutionary significance of phenotypic plasticity in plants. *Advances in Genetics*, *13*, 115–155. https://doi.org/10.1016/S0065-2660(08)60048-6

Byrne, S. L., Erthmann, P. Ø., Agerbirk, N., Bak, S., Hauser, T. P., Nagy, I., … Asp, T. (2017). The genome sequence of *Barbarea vulgaris* facilitates the study of ecological biochemistry. *Scientific Reports*, *7*, 40728. https://doi.org/10.1038/srep40728

Cavell, B. E., Syed Alwi, S. S., Donlevy, A. M., Proud, C. G., & Packham, G. (2012). Natural product-derived antitumor compound phenethyl isothiocyanate inhibits mTORC1 activity via TSC2. *Journal of Natural Products*, *75*(6), 1051–1057.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, *142*(1–2), 169–196. https://doi.org/10.1007/s10681-005-1681-5

Damerum, A., Selmes, S. L., Biggi, G. F., Clarkson, G. J., Rothwell, S. D., Truco, M. J., … Taylor, G. (2015). Elucidating the genetic basis of antioxidant status in lettuce (*Lactuca sativa*). *Horticulture Research*, *2*, 15055. https://doi.org/10.1038/hortres.2015.55

Di Noia, J. (2014). Defining powerhouse fruits and vegetables: a nutrient density approach. *Prev Chronic Dis*, *11*, 130390.

Du, H., Zhu, J., Su, H., Huang, M., Wang, H., Ding, S., … Xu, Y. (2017). Bulked segregant RNA-seq reveals differential expression and SNPs of candidate genes associated with waterlogging tolerance in maize. *Frontiers in Plant Science*, *8*, 1022. https://doi.org/10.3389/fpls.2017.01022

El-Soda, M., Malosetti, M., Zwaan, B., Koornneef, M., & Aarts, M. (2014). Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends in Plant Science*, *19*(6), 390–398. https://doi.org/10.1016/J.TPLANTS.2014.01.001

Fogarty, M. C., Hughes, C. M., Burke, G., Brown, J. C., & Davison, G. W. (2013). Acute and chronic watercress supplementation attenuates exercise-induced peripheral mononuclear cell DNA damage and lipid peroxidation. *The British Journal of Nutrition*, *109*(2), 293–301.

Gelli, M., Duo, Y., Konda, A. R., Zhang, C., Holding, D., & Dweikat, I. (2014). Identification of differentially expressed genes between sorghum genotypes with contrasting nitrogen stress tolerance by genome-wide transcriptional profiling. *BMC Genomics*, *15*(1), 179. https://doi.org/10.1186/1471-2164-15-179

Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, *24*(8), 408–415. https://doi.org/10.1016/j.tig.2008.06.001

Gill, C. I. R., Haldar, S., Boyd, L. a, Bennett, R., Whiteford, J., Butler, M., … Rowland, I. R. (2007). Watercress supplementation in diet reduces lymphocyte DNA damage and alters blood antioxidant status in healthy adults. *The American Journal of Clinical Nutrition*, *85*(2), 504–510.

Hammond, J. P., Mayes, S., Bowen, H. C., Graham, N. S., Hayden, R. M., Love, C. G., … Broadley, M. R. (2011). Regulatory hotspots are associated with plant gene expression under varying soil phosphorus supply in *Brassica rapa*. *PLANT PHYSIOLOGY*, *156*(3), 1230–1241. https://doi.org/10.1104/pp.111.175612

Hayashi, E., You, Y., Lewis, R., Calderon, M. C., Wan, G., & Still, D. W. (2012). Mapping QTL, epistasis and genotype × environment interaction of antioxidant activity, chlorophyll content and head formation in domesticated lettuce (*Lactuca sativa*). *Theoretical and Applied Genetics*, *124*(8), 1487–1502. https://doi.org/10.1007/s00122-012-1803-0

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., & Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics*, *6*(9), e1001139. https://doi.org/10.1371/journal.pgen.1001139

Hecht, S., Chung, F., Richie, J. J., Akerkar, S., Borukhova, A., Skowronski, L., & Carmella, S. (1995). Effects of watercress consumption on metabolism of a tobacco-specific lung carcinogen in smokers. *Cancer Epidemiol. Biomarkers Prev.*, *4*(8), 877–884.

Hennig, K., Verkerk, R., van Boekel, M., Dekker, M., & Bonnema, G. (2014). Food science meets plant science: A case study on improved nutritional quality by breeding for glucosinolate retention during food processing. *Trends in Food Science & Technology*, *35*(1), 61–68. https://doi.org/10.1016/J.TIFS.2013.10.006

Howard, A. H. W., & Lyon, A. G. (1952). *Nasturtium officinale* R. Br. (*Rorippa Nasturtium-Aquaticum* (L.) Hayek). *Journal of Ecology*, *40*(1), 228–245.

Humphreys, J., & Chapple, C. (2002). Rewriting the lignin roadmap. *Current Opinion in Plant Biology*, *5*(3), 224–229. https://doi.org/10.1016/S1369-5266(02)00257-1

Jian, H., Yang, B., Zhang, A., Zhang, L., Xu, X., Li, J., & Liu, L. (2017). Screening of candidate leaf morphology genes by integration of QTL mapping and RNA sequencing technologies in oilseed rape (*Brassica napus* L.). *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0169641

Jin, R., Liu, Y., Tabashnik, B. E., & Borthakur, D. (1999). Tissue culture and Agrobacterium - mediated transformation of watercress, *58*, 171–176.

Kliebenstein, D. (2009). Quantitative genomics: Analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology*, *60*(1), 93–114. https://doi.org/10.1146/annurev.arplant.043008.092114

Kuzina, V., Nielsen, J. K., Augustin, J. M., Torp, A. M., Bak, S., & Andersen, S. B. (2011). *Barbarea vulgaris* linkage map and quantitative trait loci for saponins, glucosinolates, hairiness and resistance to the herbivore *Phyllotreta nemorum*. *Phytochemistry*, *72*(2–3), 188–198. https://doi.org/10.1016/j.phytochem.2010.11.007

Lambel, S., Lanini, B., Vivoda, E., Fauve, J., Patrick Wechter, W., Harris-Shultz, K. R., … Levi, A. (2014). A major QTL associated with *Fusarium oxysporum* race 1 resistance identified in genetic populations derived from closely related watermelon lines using selective genotyping and genotyping-by-sequencing for SNP discovery. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *127*(10), 2105–2115. https://doi.org/10.1007/s00122-014-2363-2

Langridge, P., & Fleury, D. (2011). Making the most of ‘omics’ for crop breeding. *Trends in Biotechnology*, *29*(1), 33–40.

Laurie, S., Faber, M., Adebola, P., & Belete, A. (2015). Biofortification of sweet potato for food and nutrition security in South Africa. *Food Research International*. https://doi.org/10.1016/j.foodres.2015.06.001

Li, F., Hullar, M. A. J., Schwarz, Y., & Lampe, J. W. (2009). Human gut bacterial communities are altered by addition of cruciferous vegetables to a controlled fruit- and vegetable-free diet. *The Journal of Nutrition*, *139*(9), 1685–91. https://doi.org/10.3945/jn.109.108191

Manton, I. (1935). The cytological history of Watercress (*Nasturtium officinale* R. Br.). *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, *69*(1), 132–157.

Maruta, T., Noshi, M., Nakamura, M., Matsuda, S., Tamoi, M., Ishikawa, T., & Shigeoka, S. (2014). Ferulic acid 5-hydroxylase 1 is essential for expression of anthocyanin biosynthesis-associated genes and anthocyanin accumulation under photooxidative stress in Arabidopsis. *Plant Science*, *219–220*, 61–68. https://doi.org/10.1016/J.PLANTSCI.2014.01.003

Marwede, V., Gul, M. K., Becker, H. C., & Ecke, W. (2005). Mapping of QTL controlling tocopherol content in winter oilseed rape. *Plant Breeding*, *124*(1), 20–26. https://doi.org/10.1111/j.1439-0523.2004.01050.x

Moose, S. P., & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology*, *147*(3), 969–77. https://doi.org/10.1104/pp.108.118232

Morozowska, M., Czarna, A., & Jędrzejczyk, I. (2010). Estimation of nuclear DNA content in Nasturtium R. Br. by flow cytometry. *Aquatic Botany*, *93*(4), 250–253.

Mutch, D. M., Wahli, W., & Williamson, G. (2005). Nutrigenomics and nutrigenetics: the emerging faces of nutrition. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *19*(12), 1602–16. https://doi.org/10.1096/fj.05-3911rev

Neubauer, J. D., Lulai, E. C., Thompson, A. L., Suttle, J. C., & Bolton, M. D. (2012). Wounding coordinately induces cell wall protein, cell cycle and pectin methyl esterase genes involved in tuber closing layer and wound periderm development. *Journal of Plant Physiology*, *169*(6), 586–595. https://doi.org/10.1016/j.jplph.2011.12.010

Newell-McGloughlin, M. (2008). Nutritionally improved agricultural crops. *Plant Physiology*, *147*(3), 939–53. https://doi.org/10.1104/pp.108.121947

Newman, R. M., Hanscom, Z., & Kerfoot, W. C. (1992). The watercress glucosinolate-myrosinase system: a feeding deterrent to caddisflies, snails and amphipods. *Oecologia*, *92*, 1–7.

Obermeier, C., & Friedt, W. (2015). Applied oilseed rape marker technology and genomics. In *Applied Plant Genomics and Biotechnology* (pp. 253–295). Elsevier. https://doi.org/10.1016/B978-0-08-100068-7.00016-1

Palaniswamy, U. R., & McAvoy, R. J. (2001). Watercress: A salad crop with chemopreventive potential. *HortTechnology*, *11*(4), 622–626.A

Park, N. Il, Kim, J. K., Park, W. T., Cho, J. W., Lim, Y. P., & Park, S. U. (2011). An efficient protocol for genetic transformation of watercress (*Nasturtium officinale*) using Agrobacterium rhizogenes. *Molecular Biology Reports*, *38*(8), 4947–4953. https://doi.org/10.1007/s11033-010-0638-5

Patil, B. S., Crosby, K., Byrne, D., & Hirschi, K. (2014). The intersection of plant breeding, human health, and nutritional security: Lessons learned and future perspectives. *HortScience*, *49*(2), 116–127.

Payne, A. C. (2011). *Harnessing the Genetic Diversity of Watercess (Rorippa nasturtium - aquaticum) for Improved Morphology and Anti- cancer Benefits: Underpinning Data for Molecular Breeding [doctoral thesis]*. Southampton: University of Southampton.

Payne, A. C., Clarkson, G. J. J., Rothwell, S., & Taylor, G. (2015). Diversity in global gene expression and morphology across a watercress (*Nasturtium officinale* R. Br.) germplasm collection: first steps to breeding. *Horticulture Research*, *2*, 15029. https://doi.org/10.1038/hortres.2015.29

Plessis, A., Hafemeister, C., Wilkins, O., Gonzaga, Z. J., Meyer, R. S., Pires, I., … Purugganan, M. (2015). Multiple abiotic stimuli are integrated in the regulation of rice gene expression under field conditions. *ELife*, *4*(NOVEMBER2015). https://doi.org/10.7554/eLife.08411

Prasch, C., & Sonnewald, U. (2015). Signaling events in plants: Stress factors in combination change the picture. *Environmental and Experimental Botany*, *114*, 4–14. https://doi.org/10.1016/J.ENVEXPBOT.2014.06.020

Ramirez-Gonzalez, R. H., Segovia, V., Bird, N., Fenwick, P., Holdgate, S., Berry, S., … Uauy, C. (2015). RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnology Journal*, *13*(5), 613–624. https://doi.org/10.1111/pbi.12281

Richards, C. L., Rosas, U., Banta, J., Bhambhra, N., & Purugganan, M. D. (2012). Genome-wide patterns of Arabidopsis gene expression in nature. *PLoS Genetics*, *8*(4), e1002662. https://doi.org/10.1371/journal.pgen.1002662

Rieseberg, L. H., Archer, M. A., & Wayne, R. K. (1999). Transgressive segregation, adaptation and speciation. *Heredity*, *83*(4), 363–372. https://doi.org/10.1038/sj.hdy.6886170

Rousseaux, M. C., Jones, C. M., Adams, D., Chetelat, R., Bennett, A., & Powell, A. (2005). QTL analysis of fruit antioxidants in tomato using *Lycopersicon pennellii* introgression lines. *Theoretical and Applied Genetics*, *111*(7), 1396–1408. https://doi.org/10.1007/s00122-005-0071-7

Ruegger, M., Meyer, K., Cusumano, J. C., & Chapple, C. (1999). Regulation of ferulate-5-hydroxylase expression in Arabidopsis in the context of sinapate ester biosynthesis 1. *Plant Physiology*, *119*, 101–110.

Sadeghi, H., Mostafazadeh, M., Sadeghi, H., Naderian, M., Barmak, M. J., Talebianpoor, M. S., & Mehraban, F. (2013). In vivo anti-inflammatory properties of aerial parts of *Nasturtium officinale*. *Pharmaceutical Biology*, (2008), 1–6. https://doi.org/10.3109/13880209.2013.821138

Salvi, S., & Tuberosa, R. (2005). To clone or not to clone plant QTLs: Present and future challenges. *Trends in Plant Science*. https://doi.org/10.1016/j.tplants.2005.04.008

Schaart, J. G., Van De Wiel, C. C. M., Lotz, L. A. P., & Smulders, M. J. M. (2016). Opportunities for products of new plant breeding techniques. *Trends in Plant Science*, *21*(5), 438–449. https://doi.org/10.1016/j.tplants.2015.11.006

Scheben, A., Wolter, F., Batley, J., Puchta, H., & Edwards, D. (2017). Towards CRISPR/Cas crops - bringing together genomics and genome editing. *New Phytologist*. https://doi.org/10.1111/nph.14702

Sheridan, G. E. C., Claxton, J. R., Clarkson, J. M., & Blakesley, D. (2001). Genetic diversity within commercial populations of watercress (*Rorippa nasturtium-aquaticum*), and between allied Brassicaceae inferred from RAPD-PCR. *Euphytica*, *122*, 319–325.

Stratton, D. A. (1998). Reaction norm functions and QTL-environment interactions for flowering time in Arabidopsis thaliana. *Heredity*, *81*(2), 144–155. https://doi.org/10.1046/j.1365-2540.1998.00369.x

The Complex Trait Consortium. (2003). Guidelines: The nature and identification of quantitative trait loci: a community's view. *Nature Reviews Genetics*, *4*(11), 911–916. https://doi.org/10.1038/nrg1206

Thoen, M. P. M., Davila Olivas, N. H., Kloth, K. J., Coolen, S., Huang, P. P., Aarts, M. G.

M., … Dicke, M. (2016). Genetic architecture of plant stress resistance: Multi-trait genome-wide association mapping. *New Phytologist*, 1346–1362. https://doi.org/10.1111/nph.14220

Traka, M. H., Saha, S., Huseby, S., Kopriva, S., Walley, P. G., Barker, G. C., … Mithen, R. F. (2013). Genetic regulation of glucoraphanin accumulation in Beneforté broccoli. *The New Phytologist*, *198*(4), 1085–1095.

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. a. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, *27*(9), 522–530.

Waterhouse, P. M., & Helliwell, C. A. (2003). Exploring plant genomes by RNA-induced gene silencing. *Nature Reviews Genetics*, *4*(1), 29–38. https://doi.org/10.1038/nrg982

West, M. A. L., Kim, K., Kliebenstein, D. J., van Leeuwen, H., Michelmore, R. W., Doerge, R. W., & St Clair, D. A. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics*, *175*(3), 1441–50. https://doi.org/10.1534/genetics.106.064972

Wong, A. H. C., Gottesman, I. I., & Petronis, A. (2005). Phenotypic differences in genetically identical organisms: The epigenetic perspective. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddi116

Ye, X., Al-Babili, S., Klöti, A., Zhang, J., Lucca, P., Beyer, P., & Potrykus, I. (2000). Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science (New York, N.Y.)*, *287*(5451), 303–5.

Appendices

Appendix A Chapter 2 Supplementary Material

A.1 Table S2.1: Complete annotation and differential expression data

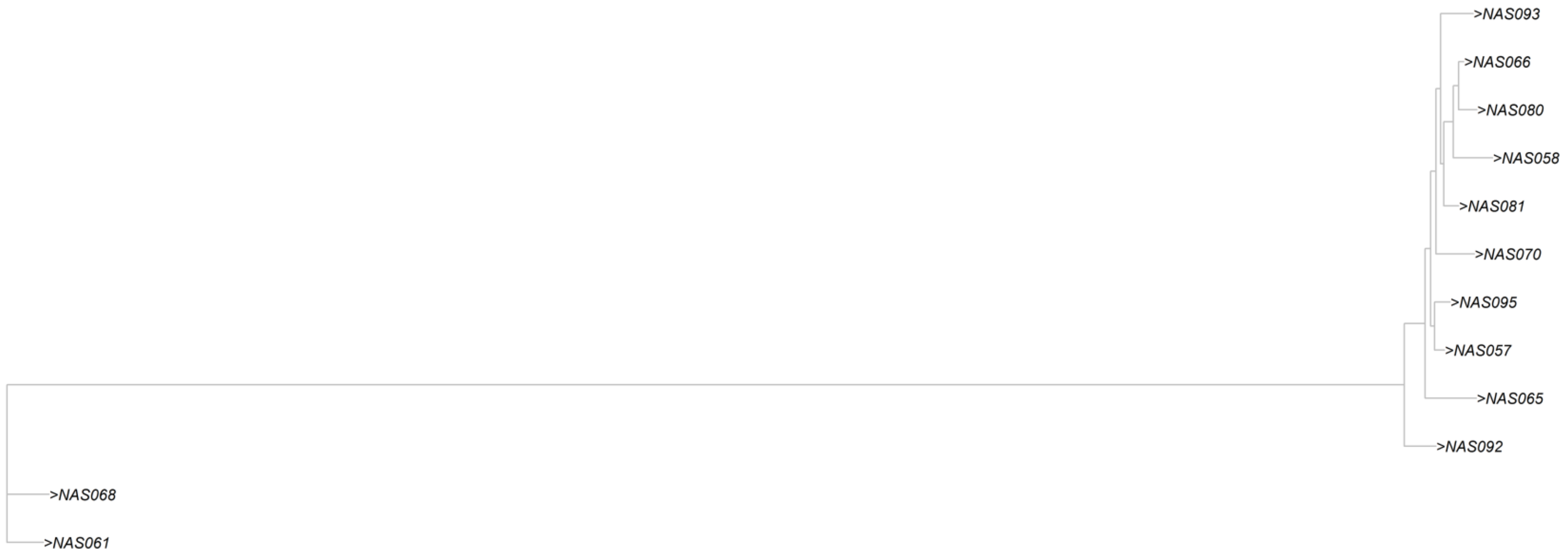
Complete annotation and differential expression data for watercress transcripts assembled and analysed in this study

Available as Additional file 1: Table S1 at:

https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-2704-4/MediaObjects/12864_2016_2704_MOESM1_ESM.xlsx

A.2 Figure S2.1: Genetic distances amongst watercress accessions

Figure S2.1 Genetic variation amongst the watercress accessions used in this study



Available as Additional file 2: Figure S1:

https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-2704-4/MediaObjects/12864_2016_2704_MOESM2_ESM.pdf

A.3 Table S2.2: Raw abundance estimates in the antioxidant differential expression comparison

Raw abundance estimates for each locus and sample in the antioxidant differential expression comparison

Available as Additional file 3: Table S2 at:

https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-2704-4/MediaObjects/12864_2016_2704_MOESM3_ESM.xlsx

A.4 Table S2.3: Raw abundance estimates in the glucosinolate differential expression comparison

Raw abundance estimates for each locus and sample in the glucosinolate differential expression comparison

Available as Additional file 4: Table S3 at:

https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-2704-4/MediaObjects/12864_2016_2704_MOESM4_ESM.xlsx

A.5 Extreme sample selection from archive data

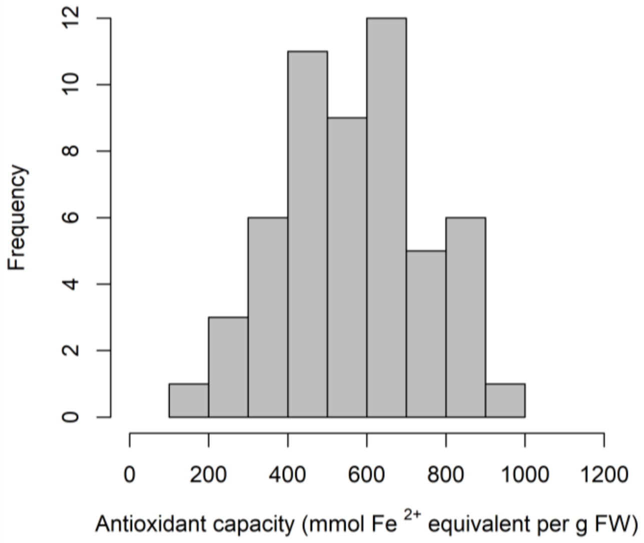
For this study, five high and five low antioxidant capacity samples needed to be identified from previously collected material. As the number of samples that could be sequenced was limited due to the cost of the method, samples were selected with as strict an approach as possible. Therefore, instead of simply selecting and sequencing the highest or lowest samples, the accession’s overall mean ranking was also considered in order to avoid false readings or inconsistent samples. The samples that were ultimately used were to top or bottom-most samples within the most consistently top or bottom-most accessions grown in this particular trial and set of conditions. This selection was informed by, and thus limited by, the small number of replicates and samples overall but also by the amount of material that was available for each desirable sample.

Figure S2.1: Antioxidant capacity data and sample selection

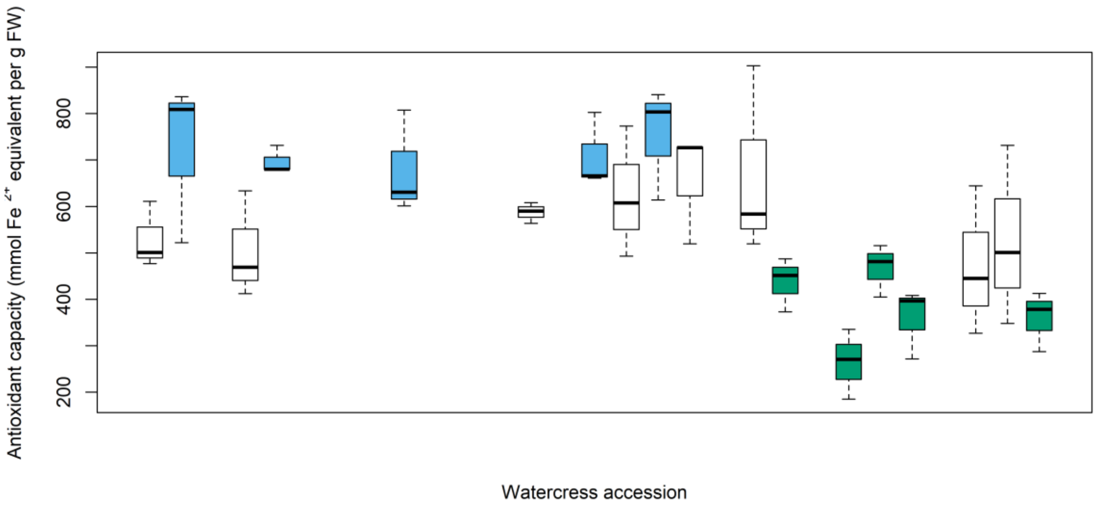
Figure S2.1 A) A histogram exhibiting the antioxidant capacity data from all plants grown in the Spetsbuty, Dorset trial by Dr Adrienne C. Payne.

B) Boxplot showing the antioxidant capacity results by watercress accession, where each box represents an accession. The accessions to be sequenced for this work (blue for high antioxidant and green for low antioxidant capacity) were determined based on consistently high or low sample readings and means and the top or bottom-most sample from that accession was sequenced. Information in the accession used is protected through data confidentiality agreements and further information can be requested from Professor Gail Taylor.

A)



B)



A.6 Table S2.4: Differentially expressed transcripts in high and low antioxidant watercress samples

Comprehensive list of the 145 differentially expressed watercress transcripts in the high and low antioxidant analysis (α ≤ 0.05). For each differentially expressed transcript isoform, both the uncorrected and corrected (for multiple hypothesis testing – false discovery rate) p value are given, alongside the best match locus ID and gene description from annotation

Wx Isoform	P value	FDR	Loci ID	Gene description
c38486_g1_i2	3.53E-12	1.60E-07	AT1G74710	Encodes a protein with isochorismate synthase activity. Mutants fail to accumulate salicylic acid. Its function may be redundant with that of ICS2 (AT1G18870).
c38746_g1_i3	4.30E-12	1.60E-07	AT5G15270	RNA-binding KH domain-containing protein; FUNCTIONS IN: RNA binding, nucleic acid binding; INVOLVED IN: biological_process unknown; LOCATED IN: cellular_component unknown; CONTAINS InterPro DOMAIN/s: K Homology, type 1, subgroup (InterPro:IPR018111), K Homology (InterPro:IPR004087), K Homology, type 1 (InterPro:IPR004088); BEST Arabidopsis thaliana protein match is: RNA-binding KH domain-containing protein (TAIR:AT1G14170.3); Has 5625 Blast hits to 2559 proteins in 215 species: Archae - 0; Bacteria - 48; Metazoa - 3662; Fungi - 737; Plants - 967; Viruses - 0; Other Eukaryotes - 211 (source: NCBI BLink).
c36660_g1_i5	9.08E-12	2.25E-07	AT5G58270	Encodes a mitochondrial half-molecule ABC transporter, a member of ATM subfamily. Mutants are dwarfed, chlorotic plants with altered leaf morphology. ATM3 transcription is

				induced by Cd(II) or Pb(II). Involved in heavy metal resistance. Arabidopsis thaliana has three ATM genes, namely ATM1, ATM2 and ATM3. Only ATM3 has an important function for plant growth. Role in Moco biosynthesis.
c38266_g1_i3	7.49E-11	1.39E-06	AT1G31410	putrescine-binding periplasmic protein-related; FUNCTIONS IN: transporter activity; INVOLVED IN: transport; LOCATED IN: chloroplast; EXPRESSED IN: 20 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Bacterial periplasmic spermidine/putrescine-binding protein (InterPro:IPR001188), Bacterial extracellular solute-binding, family 1 (InterPro:IPR006059); Has 4685 Blast hits to 4685 proteins in 1552 species: Archae - 5; Bacteria - 4396; Metazoa - 0; Fungi - 0; Plants - 37; Viruses - 0; Other Eukaryotes - 247 (source: NCBI BLink).
c23930_g1_i1	1.55E-10	2.30E-06	AT5G13340	unknown protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: chloroplast; BEST Arabidopsis thaliana protein match is: unknown protein (TAIR:AT1G10890.1); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c34668_g1_i2	6.58E-10	8.14E-06	AT5G12420	O-acyltransferase (WSD1-like) family protein; CONTAINS InterPro DOMAIN/s: O-acyltransferase, WSD1, C-terminal (InterPro:IPR009721), O-acyltransferase, WSD1, N-terminal (InterPro:IPR004255); BEST Arabidopsis thaliana protein match is: O-acyltransferase (WSD1-like) family protein (TAIR:AT5G16350.1); Has 30201 Blast hits to 17322 proteins in

				780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c37169_g1_i2	1.74E-09	1.85E-05	AT2G34750	RNA polymerase I specific transcription initiation factor RRN3 protein; FUNCTIONS IN: RNA polymerase I transcription factor activity; INVOLVED IN: biological_process unknown; LOCATED IN: cellular_component unknown; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: RNA polymerase I specific transcription initiation factor RRN3 (InterPro:IPR007991); BEST Arabidopsis thaliana protein match is: RNA polymerase I specific transcription initiation factor RRN3 protein (TAIR:AT1G30590.1); Has 368 Blast hits to 356 proteins in 164 species: Archae - 0; Bacteria - 0; Metazoa - 124; Fungi - 131; Plants - 69; Viruses - 0; Other Eukaryotes - 44 (source: NCBI BLink).
c29280_g1_i1	2.67E-09	2.28E-05	AT1G77420	alpha/beta-Hydrolases superfamily protein; BEST Arabidopsis thaliana protein match is: alpha/beta-Hydrolases superfamily protein (TAIR:AT5G16120.1); Has 4552 Blast hits to 4550 proteins in 1360 species: Archae - 49; Bacteria - 3106; Metazoa - 121; Fungi - 218; Plants - 474; Viruses - 41; Other Eukaryotes - 543 (source: NCBI BLink).
c25889_g1_i3	2.77E-09	2.28E-05	AT5G11350	DNAse I-like superfamily protein; CONTAINS InterPro DOMAIN/s: Endonuclease/exonuclease/phosphatase (InterPro:IPR005135); BEST Arabidopsis thaliana protein match is: DNAse I-like superfamily protein (TAIR:AT1G73875.1); Has 30201 Blast hits

				to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c36445_g1_i10	6.33E-09	4.32E-05	AT5G64120	encodes a cell wall bound peroxidase that is induced by hypo-osmolarity
c27597_g1_i1	6.44E-09	4.32E-05	AT4G15110	member of CYP97B
c35817_g2_i3	6.99E-09	4.32E-05	AT5G52310	cold regulated gene, the 5' region of cor78 has cis-acting regulatory elements that can impart cold-regulated gene expression
c30486_g1_i3	9.18E-09	5.24E-05	AT2G18660	Encodes PNP-A (Plant Natriuretic Peptide A). PNPs are a class of systemically mobile molecules distantly related to expansins; their biological role has remained elusive. PNP-A contains a signal peptide domain and is secreted into the extracellular space. Co-expression analyses using microarray data suggest that PNP-A may function as a component of plant defence response and SAR in particular, and could be classified as a newly identified PR protein. It is stress responsive and can enhance its own expression.
c34644_g1_i7	1.25E-08	6.01E-05	AT2G38240	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein; FUNCTIONS IN: oxidoreductase activity; INVOLVED IN: response to salt stress; EXPRESSED IN: 11 plant structures; EXPRESSED DURING: 6 growth stages; CONTAINS InterPro DOMAIN/s: Oxoglutarate/iron-dependent oxygenase (InterPro:IPR005123); BEST Arabidopsis thaliana protein match is: 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein (TAIR:AT5G05600.1); Has 8819 Blast hits to 8740 proteins in 1012 species: Archae - 0;

				Bacteria - 1137; Metazoa - 112; Fungi - 1067; Plants - 5036; Viruses - 0; Other Eukaryotes - 1467 (source: NCBI BLink).
c38867_g1_i6	1.27E-08	6.01E-05	AT3G52850	Encodes the Vacuolar Sorting Receptor-1 (VSR-1)/Epidermal Growth Factor Receptor-like protein1(VSR-1/ATELP1). Binds vacuolar targeting signals. Involved in sorting seed storage proteins into vacuoles.
c38401_g1_i1	1.30E-08	6.01E-05	NA	NA
c28989_g1_i4	1.95E-08	7.78E-05	AT1G07080	Thioredoxin superfamily protein; FUNCTIONS IN: catalytic activity; INVOLVED IN: biological_process unknown; LOCATED IN: vacuole; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 15 growth stages; CONTAINS InterPro DOMAIN/s: Gamma interferon inducible lysosomal thiol reductase GILT (InterPro:IPR004911), Thioredoxin-like fold (InterPro:IPR012336); BEST Arabidopsis thaliana protein match is: gamma interferon responsive lysosomal thiol (GILT) reductase family protein (TAIR:AT5G01580.1); Has 470 Blast hits to 463 proteins in 91 species: Archae - 0; Bacteria - 0; Metazoa - 320; Fungi - 0; Plants - 101; Viruses - 0; Other Eukaryotes - 49 (source: NCBI BLink).
c33478_g1_i1	1.97E-08	7.78E-05	AT5G07010	Encodes a sulfotransferase that acts specifically on 11- and 12-hydroxyjasmonic acid. Transcript levels for this enzyme are increased by treatments with jasmonic acid (JA), 12-hydroxyJA, JA-isoleucine, and 12-oxyphytodienoic acid (a JA precursor).

c29691_g1_i2	1.99E-08	7.78E-05	AT2G29110	member of Putative ligand-gated ion channel subunit family
c26980_g1_i3	2.80E-08	0.0001038	AT2G41700	ATP-binding cassette A1 (ABCA1); FUNCTIONS IN: ATPase activity, coupled to transmembrane movement of substances, amino acid transmembrane transporter activity; LOCATED IN: plasma membrane; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 15 growth stages; CONTAINS InterPro DOMAIN/s: ATPase, AAA+ type, core (InterPro:IPR003593), ABC transporter-like (InterPro:IPR003439), ABC transporter, conserved site (InterPro:IPR017871); BEST Arabidopsis thaliana protein match is: ATP-binding cassette A2 (TAIR:AT3G47730.1); Has 809471 Blast hits to 378826 proteins in 4091 species: Archae - 14583; Bacteria - 641508; Metazoa - 16186; Fungi - 10844; Plants - 9650; Viruses - 55; Other Eukaryotes - 116645 (source: NCBI BLink).
c35817_g1_i1	3.15E-08	0.0001114	AT5G52310	cold regulated gene, the 5' region of cor78 has cis-acting regulatory elements that can impart cold-regulated gene expression
c36980_g5_i2	3.30E-08	0.0001114	AT1G21310	Encodes extensin 3.
c32073_g2_i1	4.86E-08	0.0001569	AT3G48080	alpha/beta-Hydrolases superfamily protein; FUNCTIONS IN: lipase activity, triglyceride lipase activity, signal transducer activity; INVOLVED IN: lipid metabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: 19 plant structures; EXPRESSED DURING: 12 growth stages; CONTAINS InterPro DOMAIN/s: Lipase, class 3 (InterPro:IPR002921); BEST Arabidopsis thaliana protein match is: alpha/beta-Hydrolases superfamily protein (TAIR:AT3G48090.1); Has 522 Blast hits to 472 proteins in 44 species: Archae - 0; Bacteria -

				4; Metazoa - 0; Fungi - 2; Plants - 484; Viruses - 0; Other Eukaryotes - 32 (source: NCBI BLink).
c36445_g1_i11	5.61E-08	0.0001735	AT5G64120	encodes a cell wall bound peroxidase that is induced by hypo-osmolarity
c31704_g1_i4	1.07E-07	0.0003171	AT1G29100	Heavy metal transport/detoxification superfamily protein ; FUNCTIONS IN: copper ion binding, metal ion binding; INVOLVED IN: copper ion transport, metal ion transport; CONTAINS InterPro DOMAIN/s: Heavy metal transport/detoxification protein (InterPro:IPR006121); BEST Arabidopsis thaliana protein match is: Heavy metal transport/detoxification superfamily protein (TAIR:AT1G06330.1); Has 899 Blast hits to 889 proteins in 44 species: Archae - 0; Bacteria - 0; Metazoa - 3; Fungi - 15; Plants - 881; Viruses - 0; Other Eukaryotes - 0 (source: NCBI BLink).
c34103_g1_i1	1.72E-07	0.0004905	AT2G24850	Encodes a tyrosine aminotransferase that is responsive to treatment with jasmonic acid.
c23835_g2_i1	1.87E-07	0.0005142	AT3G60120	beta glucosidase 27 (BGLU27); FUNCTIONS IN: cation binding, hydrolase activity, hydrolyzing O-glycosyl compounds, catalytic activity; INVOLVED IN: carbohydrate metabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: stem, hypocotyl, sepal, stamen; EXPRESSED DURING: 4 anthesis; CONTAINS InterPro DOMAIN/s: Glycoside hydrolase, family 1 (InterPro:IPR001360), Glycoside hydrolase, family 1, active site (InterPro:IPR018120), Glycoside hydrolase, catalytic core (InterPro:IPR017853), Glycoside

				hydrolase, subgroup, catalytic core (InterPro:IPR013781); BEST Arabidopsis thaliana protein match is: Glycosyl hydrolase superfamily protein (TAIR:AT2G44490.1); Has 11429 Blast hits to 11081 proteins in 1472 species: Archae - 140; Bacteria - 7815; Metazoa - 722; Fungi - 203; Plants - 1548; Viruses - 0; Other Eukaryotes - 1001 (source: NCBI BLink).
c35108_g1_i2	2.54E-07	0.0006745	AT5G48880	Encodes a peroxisomal 3-keto-acyl-CoA thiolase 2 precursor. EC2.3.1.16 thiolases. AT5G48880.1 is named PKT1 and AT5G48880.2 is named PKT2.
c36787_g1_i1	2.98E-07	0.0007637	AT4G22690	member of CYP706A
c35006_g1_i2	3.34E-07	0.0008028	AT3G21340	Leucine-rich repeat protein kinase family protein; FUNCTIONS IN: kinase activity; INVOLVED IN: protein amino acid phosphorylation; LOCATED IN: endomembrane system; EXPRESSED IN: root; CONTAINS InterPro DOMAIN/s: Protein kinase, ATP binding site (InterPro:IPR017441), Protein kinase, catalytic domain (InterPro:IPR000719), Leucine-rich repeat (InterPro:IPR001611), Serine/threonine-protein kinase-like domain (InterPro:IPR017442), Protein kinase-like domain (InterPro:IPR011009), Serine/threonine-protein kinase, active site (InterPro:IPR008271); BEST Arabidopsis thaliana protein match is: Leucine-rich repeat protein kinase family protein (TAIR:AT1G51850.1); Has 160620 Blast hits to 122303 proteins in 4467 species: Archae - 99; Bacteria - 13621; Metazoa - 44084; Fungi - 10001; Plants - 73653; Viruses - 414; Other Eukaryotes - 18748 (source: NCBI BLink).

c9463_g1_i1	3.46E-07	0.0008028	NA	NA
c38401_g2_i1	3.46E-07	0.0008028	NA	NA
c36896_g1_i1	4.38E-07	0.0009862	AT5G13320	Encodes an enzyme capable of conjugating amino acids to 4-substituted benzoates. 4-HBA (4-hydroxybenzoic acid) and pABA (4-aminobenzoate) may be targets of the enzyme in Arabidopsis, leading to the production of pABA-Glu, 4HBA-Glu, or other related compounds. This enzyme is involved in disease-resistance signaling. It is required for the accumulation of salicylic acid, activation of defense responses, and resistance to Pseudomonas syringae. Salicylic acid can decrease this enzyme's activity in vitro and may act as a competitive inhibitor. Expression of PBS3/GH3.12 can be detected in cotyledons, true leaves, hypocotyls, and occasionally in some parts of roots from 10-day-old seedlings. No expression has been detected in root, stem, rosette or cauline leaves of mature 4- to 5-week-old plants.
c37084_g1_i4	8.04E-07	0.0017265	AT2G45550	member of CYP76C
c39789_g1_i5	8.14E-07	0.0017265	AT5G65290	LMBR1-like membrane protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: endomembrane system, integral to membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 15 growth stages; CONTAINS InterPro DOMAIN/s: LMBR1-like membrane protein, conserved region (InterPro:IPR006876); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).

c31742_g1_i2	8.56E-07	0.0017642	AT5G52540	Protein of unknown function (DUF819); LOCATED IN: chloroplast envelope; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Protein of unknown function DUF819 (InterPro:IPR008537); BEST Arabidopsis thaliana protein match is: Protein of unknown function (DUF819) (TAIR:AT5G24000.1); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c36505_g1_i1	1.18E-06	0.0023646	AT4G36220	encodes ferulate 5-hydroxylase (F5H). Involved in lignin biosynthesis.
c23835_g1_i1	1.24E-06	0.0024216	AT3G60120	beta glucosidase 27 (BGLU27); FUNCTIONS IN: cation binding, hydrolase activity, hydrolyzing O-glycosyl compounds, catalytic activity; INVOLVED IN: carbohydrate metabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: stem, hypocotyl, sepal, stamen; EXPRESSED DURING: 4 anthesis; CONTAINS InterPro DOMAIN/s: Glycoside hydrolase, family 1 (InterPro:IPR001360), Glycoside hydrolase, family 1, active site (InterPro:IPR018120), Glycoside hydrolase, catalytic core (InterPro:IPR017853), Glycoside hydrolase, subgroup, catalytic core (InterPro:IPR013781); BEST Arabidopsis thaliana protein match is: Glycosyl hydrolase superfamily protein (TAIR:AT2G44490.1); Has 11429 Blast hits to 11081 proteins in 1472 species: Archae - 140; Bacteria - 7815; Metazoa - 722; Fungi - 203; Plants - 1548; Viruses - 0; Other Eukaryotes - 1001 (source: NCBI BLink).

c35416_g1_i4	1.34E-06	0.0025527	AT4G37550	Acetamidase/Formamidase family protein; FUNCTIONS IN: formamidase activity, hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides; INVOLVED IN: metabolic process; LOCATED IN: vacuole; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Acetamidase/Formamidase (InterPro:IPR004304); BEST Arabidopsis thaliana protein match is: Acetamidase/Formamidase family protein (TAIR:AT4G37560.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c34509_g1_i10	1.47E-06	0.0026403	NA	NA
c34389_g2_i6	1.48E-06	0.0026403	AT2G21630	Sec23/Sec24 protein transport family protein; FUNCTIONS IN: transporter activity, zinc ion binding; INVOLVED IN: intracellular protein transport, transport, ER to Golgi vesicle-mediated transport; LOCATED IN: COPII vesicle coat; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Sec23/Sec24, helical domain (InterPro:IPR006900), Sec23/Sec24 beta-sandwich (InterPro:IPR012990), Sec23/Sec24, trunk domain (InterPro:IPR006896), Zinc finger, Sec23/Sec24-type (InterPro:IPR006895), Gelsolin domain (InterPro:IPR007123); BEST Arabidopsis thaliana protein match is: Sec23/Sec24 protein transport family protein (TAIR:AT4G14160.2); Has 1585 Blast hits to 1570 proteins in 246 species: Archae - 0; Bacteria - 0; Metazoa - 525; Fungi - 514; Plants - 334; Viruses - 0; Other Eukaryotes - 212 (source: NCBI BLink).

c33478_g1_i2	1.49E-06	0.0026403	AT5G07010	Encodes a sulfotransferase that acts specifically on 11- and 12-hydroxyjasmonic acid. Transcript levels for this enzyme are increased by treatments with jasmonic acid (JA), 12-hydroxyJA, JA-isoleucine, and 12-oxyphytodienoic acid (a JA precursor).
c25158_g1_i2	1.74E-06	0.0029687	AT1G23730	beta carbonic anhydrase 3 (BCA3); FUNCTIONS IN: carbonate dehydratase activity, zinc ion binding; INVOLVED IN: carbon utilization; LOCATED IN: cytosol, plasma membrane, membrane; EXPRESSED IN: 14 plant structures; EXPRESSED DURING: 6 growth stages; CONTAINS InterPro DOMAIN/s: Carbonic anhydrase, prokaryotic-like, conserved site (InterPro:IPR015892), Carbonic anhydrase (InterPro:IPR001765); BEST Arabidopsis thaliana protein match is: beta carbonic anhydrase 4 (TAIR:AT1G70410.1); Has 5075 Blast hits to 5060 proteins in 1504 species: Archae - 26; Bacteria - 3914; Metazoa - 59; Fungi - 205; Plants - 361; Viruses - 0; Other Eukaryotes - 510 (source: NCBI BLink).
c36353_g1_i3	1.76E-06	0.0029687	AT2G22400	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein; FUNCTIONS IN: methyltransferase activity, RNA binding; LOCATED IN: cellular_component unknown; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Bacterial Fmu (Sun)/eukaryotic nucleolar NOL1/Nop2p (InterPro:IPR001678), Bacterial Fmu (Sun)/eukaryotic nucleolar NOL1/Nop2p, conserved site (InterPro:IPR018314); BEST Arabidopsis thaliana protein match is: S-adenosyl-L-methionine-dependent methyltransferases superfamily protein (TAIR:AT4G40000.1); Has 8976 Blast hits to 8942 proteins in 2412 species: Archae - 298; Bacteria - 6215; Metazoa - 583; Fungi - 307; Plants - 251; Viruses - 0; Other Eukaryotes - 1322 (source: NCBI BLink).

c12750_g1_i1	2.26E-06	0.003726	AT4G00700	C2 calcium/lipid-binding plant phosphoribosyltransferase family protein; CONTAINS InterPro DOMAIN/s: C2 membrane targeting protein (InterPro:IPR018029), C2 calcium/lipid-binding domain, CaLB (InterPro:IPR008973), Phosphoribosyltransferase C-terminal (InterPro:IPR013583), C2 calcium-dependent membrane targeting (InterPro:IPR000008); BEST Arabidopsis thaliana protein match is: C2 calcium/lipid-binding plant phosphoribosyltransferase family protein (TAIR:AT4G11610.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c34656_g1_i2	2.49E-06	0.0040214	AT1G70780	unknown protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: mitochondrion; EXPRESSED IN: sperm cell, male gametophyte, pollen tube; EXPRESSED DURING: L mature pollen stage, M germinated pollen stage; BEST Arabidopsis thaliana protein match is: unknown protein (TAIR:AT1G23150.1); Has 143 Blast hits to 143 proteins in 17 species: Archae - 0; Bacteria - 0; Metazoa - 0; Fungi - 0; Plants - 143; Viruses - 0; Other Eukaryotes - 0 (source: NCBI BLink).

c16918_g2_i1	2.88E-06	0.0045474	AT4G29050	Concanavalin A-like lectin protein kinase family protein; FUNCTIONS IN: carbohydrate binding, kinase activity; INVOLVED IN: protein amino acid phosphorylation; LOCATED IN: endomembrane system; EXPRESSED IN: 7 plant structures; EXPRESSED DURING: 8 growth stages; CONTAINS InterPro DOMAIN/s: Legume lectin, beta chain (InterPro:IPR001220), Protein kinase, ATP binding site (InterPro:IPR017441), Serine/threonine-protein kinase-like domain (InterPro:IPR017442), Concanavalin A-like lectin/glucanase, subgroup (InterPro:IPR013320), Protein kinase-like domain (InterPro:IPR011009), Serine/threonine-protein kinase, active site (InterPro:IPR008271), Protein kinase, catalytic domain (InterPro:IPR000719), Concanavalin A-like lectin/glucanase (InterPro:IPR008985), Legume lectin, beta chain, Mn/Ca-binding site (InterPro:IPR019825); BEST Arabidopsis thaliana protein match is: Concanavalin A-like lectin protein kinase family protein (TAIR:AT1G70110.1); Has 123408 Blast hits to 121938 proteins in 4814 species: Archae - 120; Bacteria - 14220; Metazoa - 44809; Fungi - 10855; Plants - 34836; Viruses - 425; Other Eukaryotes - 18143 (source: NCBI BLink).
c34986_g2_i1	3.04E-06	0.0046943	AT1G14790	Encodes RNA-dependent RNA polymerase. While not required for virus-induced post-transcriptional gene silencing (PTGS), it can promote turnover of viral RNAs in infected plants. Nomenclature according to Xie, et al. (2004). Involved in the production of Cucumber Mosaic Virus siRNAs.
c29957_g2_i4	3.13E-06	0.0047374	NA	NA

c28825_g1_i5	3.80E-06	0.0056451	AT2G32880	TRAF-like family protein; CONTAINS InterPro DOMAIN/s: TRAF-like (InterPro:IPR008974), MATH (InterPro:IPR002083), TRAF-type (InterPro:IPR013322); BEST Arabidopsis thaliana protein match is: TRAF-like family protein (TAIR:AT2G32870.1); Has 574 Blast hits to 518 proteins in 28 species: Archae - 0; Bacteria - 0; Metazoa - 4; Fungi - 2; Plants - 559; Viruses - 0; Other Eukaryotes - 9 (source: NCBI BLink).
c35428_g1_i3	3.89E-06	0.0056664	AT1G08010	Encodes a member of the GATA factor family of zinc finger transcription factors.
c37066_g1_i2	4.00E-06	0.0057166	AT2G30840	encodes a protein whose sequence is similar to 2-oxoglutarate-dependent dioxygenase
c36086_g1_i11	4.22E-06	0.005915	AT5G42900	cold regulated gene 27; BEST Arabidopsis thaliana protein match is: unknown protein (TAIR:AT4G33980.1); Has 74 Blast hits to 74 proteins in 12 species: Archae - 0; Bacteria - 0; Metazoa - 0; Fungi - 0; Plants - 74; Viruses - 0; Other Eukaryotes - 0 (source: NCBI BLink).
c40478_g1_i5	4.84E-06	0.0065809	AT3G11960	Cleavage and polyadenylation specificity factor (CPSF) A subunit protein; FUNCTIONS IN: nucleic acid binding; INVOLVED IN: biological_process unknown; LOCATED IN: nucleus, chloroplast; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Cleavage/polyadenylation specificity factor, A subunit, C-terminal (InterPro:IPR004871); BEST Arabidopsis thaliana protein match is: damaged DNA binding protein 1A (TAIR:AT4G05420.1); Has 1073 Blast hits to 789 proteins in 185 species: Archae - 0; Bacteria - 0; Metazoa - 332; Fungi - 287; Plants - 277; Viruses - 0; Other Eukaryotes - 177 (source: NCBI BLink).

c34965_g1_i3	4.88E-06	0.0065809	AT3G14990	Encodes a homolog of animal DJ-1 superfamily protein. In the A. thaliana genome, three genes encoding close homologs of human DJ-1 were identified AT3G14990 (DJ1A), AT1G53280 (DJ1B) and AT4G34020 (DJ1C). Among the three homologs, DJ1C is essential for chloroplast development and viability.
c9512_g1_i1	5.16E-06	0.0068378	AT5G18470	Curculin-like (mannose-binding) lectin family protein; FUNCTIONS IN: sugar binding; INVOLVED IN: response to karrikin; LOCATED IN: plant-type cell wall; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 12 growth stages; CONTAINS InterPro DOMAIN/s: Curculin-like (mannose-binding) lectin (InterPro:IPR001480); BEST Arabidopsis thaliana protein match is: lectin protein kinase family protein (TAIR:AT1G67520.1); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c34721_g1_i2	5.53E-06	0.007207	AT1G62975	basic helix-loop-helix (bHLH) DNA-binding superfamily protein; FUNCTIONS IN: DNA binding, sequence-specific DNA binding transcription factor activity; INVOLVED IN: regulation of transcription; LOCATED IN: nucleus; CONTAINS InterPro DOMAIN/s: Helix-loop-helix DNA-binding domain (InterPro:IPR001092), Helix-loop-helix DNA-binding (InterPro:IPR011598); BEST Arabidopsis thaliana protein match is: basic helix-loop-helix (bHLH) DNA-binding superfamily protein (TAIR:AT1G12540.1); Has 533 Blast hits to 533 proteins in 49 species: Archae - 0; Bacteria - 0; Metazoa - 85; Fungi - 0; Plants - 445; Viruses - 0; Other Eukaryotes - 3 (source: NCBI BLink).

c29360_g1_i5	6.43E-06	0.0081564	AT2G22870	embryo defective 2001 (EMB2001); FUNCTIONS IN: GTP binding; INVOLVED IN: embryo development ending in seed dormancy; LOCATED IN: intracellular; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: GTP-binding protein, HSR1-related (InterPro:IPR002917), GTP-binding protein, ribosome biogenesis, YsxC (InterPro:IPR019987); BEST Arabidopsis thaliana protein match is: P-loop containing nucleoside triphosphate hydrolases superfamily protein (TAIR:AT5G11480.1); Has 7972 Blast hits to 7922 proteins in 2512 species: Archae - 103; Bacteria - 5806; Metazoa - 98; Fungi - 230; Plants - 215; Viruses - 0; Other Eukaryotes - 1520 (source: NCBI BLink).
c35531_g1_i7	6.48E-06	0.0081564	AT4G11890	Encodes a receptor-like cytosolic kinase ARCK1. Negatively controls abscisic acid and osmotic stress signal transduction.
c40351_g2_i6	6.84E-06	0.0084624	AT5G24710	Transducin/WD40 repeat-like superfamily protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: plasma membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: WD40 repeat 2 (InterPro:IPR019782), WD40 repeat, conserved site (InterPro:IPR019775), WD40 repeat (InterPro:IPR001680), WD40 repeat-like-containing domain (InterPro:IPR011046), WD40-repeat-containing domain (InterPro:IPR017986), WD40/YVTN repeat-like-containing domain (InterPro:IPR015943), WD40 repeat, subgroup (InterPro:IPR019781); Has 53337 Blast hits to 28879 proteins in 1972 species: Archae - 196; Bacteria - 12524; Metazoa - 15998; Fungi - 8175; Plants - 2336; Viruses - 1195; Other Eukaryotes - 12913 (source: NCBI BLink).

c40487_g1_i5	9.05E-06	0.0105601	AT4G25960	P-glycoprotein 2 (PGP2); FUNCTIONS IN: ATPase activity, coupled to transmembrane movement of substances; INVOLVED IN: transport, transmembrane transport; LOCATED IN: membrane; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: ATPase, AAA+ type, core (InterPro:IPR003593), ABC transporter-like (InterPro:IPR003439), ABC transporter, transmembrane domain, type 1 (InterPro:IPR011527), ABC transporter integral membrane type 1 (InterPro:IPR017940), ABC transporter, transmembrane domain (InterPro:IPR001140), ABC transporter, conserved site (InterPro:IPR017871); BEST Arabidopsis thaliana protein match is: P-glycoprotein 10 (TAIR:AT1G10680.1); Has 857619 Blast hits to 396855 proteins in 4207 species: Archae - 14871; Bacteria - 670687; Metazoa - 17931; Fungi - 12963; Plants - 9451; Viruses - 35; Other Eukaryotes - 131681 (source: NCBI BLink).
c38486_g1_i1	9.12E-06	0.0105601	AT1G74710	Encodes a protein with isochorismate synthase activity. Mutants fail to accumulate salicylic acid. Its function may be redundant with that of ICS2 (AT1G18870).
c25597_g1_i2	9.14E-06	0.0105601	AT4G37110	Zinc-finger domain of monoamine-oxidase A repressor R1; FUNCTIONS IN: zinc ion binding; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Zinc finger, RING-type (InterPro:IPR001841), Cell division cycle-associated protein (InterPro:IPR018866); BEST Arabidopsis thaliana protein match is: Zinc-finger domain of monoamine-oxidase A repressor R1 (TAIR:AT2G23530.1); Has 452 Blast hits to 447 proteins in 93 species: Archae - 0; Bacteria - 0; Metazoa - 137; Fungi - 54; Plants - 229; Viruses - 0; Other Eukaryotes - 32 (source: NCBI BLink).

c21598_g2_i1	9.18E-06	0.0105601	NA	NA
c29280_g1_i3	9.25E-06	0.0105601	AT1G77420	alpha/beta-Hydrolases superfamily protein; BEST Arabidopsis thaliana protein match is: alpha/beta-Hydrolases superfamily protein (TAIR:AT5G16120.1); Has 4552 Blast hits to 4550 proteins in 1360 species: Archae - 49; Bacteria - 3106; Metazoa - 121; Fungi - 218; Plants - 474; Viruses - 41; Other Eukaryotes - 543 (source: NCBI BLink).
c37054_g1_i6	9.57E-06	0.0107658	AT5G15650	RGP2 is a UDP-arabinose mutase that catalyzes the interconversion between the pyranose and furanose forms of UDP-L-arabinose. It appears to be required for proper cell wall formation. rgp1(at3g02230)/rgp2 double mutants have a male gametophyte lethal phenotype. RGP2 fusion proteins can be found in the cytosol and peripherally associated with the Golgi apparatus. RGP2 was originally identified as Reversibly Glycosylated Polypeptide-2. Constitutive expression in tobacco impairs plant development and virus spread.
c38457_g1_i1	9.83E-06	0.010885	AT2G03730	Member of a small family of ACT domain containing proteins. ACT domains are thought to be involved in amino acid binding.
c14719_g1_i1	1.00E-05	0.010885	AT1G33960	Identified as a gene that is induced by avirulence gene avrRpt2 and RPS2 after infection with Pseudomonas syringae pv maculicola strain ES4326 carrying avrRpt2

c37520_g2_i1	1.02E-05	0.010885	AT1G61120	Encodes a geranyllinalool synthase that produces a precursor to TMTT, a volatile plant defense C16-homoterpene. GES transcript levels rise in response to alamethicin, a fungal peptide mixture that damages membranes. This transcriptional response is blocked in JA biosynthetic and JA signaling mutants, but GES transcript levels still rise in response to alamethicin in mutants with salicylic acid and ethylene biosynthetic and/or signaling defects. GES transcripts also accumulate in response to a larval infestation. This enzyme does not localize to the plastids, and it may be present in the cytosol or endoplasmic reticulum.
c34965_g1_i4	1.03E-05	0.010885	AT3G14990	Encodes a homolog of animal DJ-1 superfamily protein. In the A. thaliana genome, three genes encoding close homologs of human DJ-1 were identified AT3G14990 (DJ1A), AT1G53280 (DJ1B) and AT4G34020 (DJ1C). Among the three homologs, DJ1C is essential for chloroplast development and viability.
c22841_g1_i1	1.05E-05	0.010885	AT2G24180	cytochrome P450 monooxygenase
c35531_g1_i4	1.06E-05	0.010885	AT4G11890	Encodes a receptor-like cytosolic kinase ARCK1. Negatively controls abscisic acid and osmotic stress signal transduction.
c34637_g1_i2	1.07E-05	0.010885	AT2G28670	esb1 mutants have increased levels of suberin and altered levels of several ions in their leaves.
c37657_g2_i2	1.13E-05	0.0113209	AT2G19190	Receptor-like protein kinase. Involved in early defense signaling.
c18040_g1_i1	1.21E-05	0.0119397	AT5G46330	Encodes a leucine-rich repeat serine/threonine protein kinase that is expressed ubiquitously. FLS2 is involved in MAP kinase signalling relay involved in innate immunity. Essential in the

				perception of flagellin, a potent elicitor of the defense response. FLS2 is directed for degradation by the bacterial ubiquitin ligase AvrPtoB.
c38624_g2_i6	1.23E-05	0.0120129	AT5G54710	Ankyrin repeat family protein; INVOLVED IN: biological_process unknown; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Ankyrin repeat-containing domain (InterPro:IPR020683), Ankyrin repeat (InterPro:IPR002110); BEST Arabidopsis thaliana protein match is: Ankyrin repeat family protein (TAIR:AT1G34050.1); Has 20807 Blast hits to 11728 proteins in 551 species: Archae - 16; Bacteria - 1568; Metazoa - 11429; Fungi - 1235; Plants - 2476; Viruses - 139; Other Eukaryotes - 3944 (source: NCBI BLink).
c36332_g1_i1	1.25E-05	0.0120425	AT3G11660	encodes a protein whose sequence is similar to tobacco hairpin-induced gene (HIN1) and Arabidopsis non-race specific disease resistance gene (NDR1). Expression of this gene is induced by cucumber mosaic virus. Localization of the gene product is similar to that of NHL3 (plasma membrane) but it is yet inconclusive.
c31344_g1_i1	1.31E-05	0.0124724	AT2G36100	Uncharacterised protein family (UPF0497); CONTAINS InterPro DOMAIN/s: Uncharacterised protein family UPF0497, trans-membrane plant (InterPro:IPR006702), Uncharacterised protein family UPF0497, trans-membrane plant subgroup (InterPro:IPR006459); BEST Arabidopsis thaliana protein match is: Uncharacterised protein family (UPF0497) (TAIR:AT3G11550.1); Has 599 Blast hits to 599 proteins in 21 species: Archae - 0; Bacteria - 0; Metazoa - 0; Fungi - 0; Plants - 599; Viruses - 0; Other Eukaryotes - 0 (source: NCBI BLink).

c37637_g1_i3	1.35E-05	0.012691	AT5G62710	Leucine-rich repeat protein kinase family protein; FUNCTIONS IN: protein serine/threonine kinase activity, protein kinase activity, ATP binding; INVOLVED IN: protein amino acid phosphorylation; LOCATED IN: endomembrane system; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Protein kinase, ATP binding site (InterPro:IPR017441), Protein kinase, catalytic domain (InterPro:IPR000719), Leucine-rich repeat-containing N-terminal domain, type 2 (InterPro:IPR013210), Leucine-rich repeat (InterPro:IPR001611), Serine/threonine-protein kinase-like domain (InterPro:IPR017442), Protein kinase-like domain (InterPro:IPR011009), Serine/threonine-protein kinase, active site (InterPro:IPR008271); BEST Arabidopsis thaliana protein match is: Leucine-rich repeat protein kinase family protein (TAIR:AT1G31420.2); Has 175257 Blast hits to 128076 proteins in 4599 species: Archae - 150; Bacteria - 15408; Metazoa - 46877; Fungi - 10024; Plants - 81619; Viruses - 466; Other Eukaryotes - 20713 (source: NCBI BLink).
c37931_g3_i2	1.37E-05	0.012691	AT1G03590	Protein phosphatase 2C family protein; FUNCTIONS IN: protein serine/threonine phosphatase activity, catalytic activity; INVOLVED IN: N-terminal protein myristoylation; LOCATED IN: plasma membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Protein phosphatase 2C-related (InterPro:IPR001932), Protein phosphatase 2C (InterPro:IPR015655), Protein phosphatase 2C, N-terminal (InterPro:IPR014045); BEST Arabidopsis thaliana protein match is: Protein phosphatase 2C family protein (TAIR:AT4G03415.2); Has 5799 Blast hits to 5797 proteins in 294 species: Archae - 0; Bacteria - 4; Metazoa - 1418; Fungi - 691; Plants - 2505; Viruses - 5; Other Eukaryotes - 1176 (source: NCBI BLink).

c40245_g1_i11	1.38E-05	0.012691	AT1G05570	Encodes a callose synthase 1 catalytic subunit . Member of Glycosyltransferase Family- 48.
c34568_g1_i4	1.41E-05	0.0127887	AT4G26110	Encodes a member of a small gene family of proteins with similarity to nucleosome assembly proteins.May function in nucleotide excision repair. Loss of function mutations have no obvious visible phenotypes but do seem to affect transcription of NER related genes.
c25925_g1_i2	1.45E-05	0.0129925	AT2G43180	Phosphoenolpyruvate carboxylase family protein; FUNCTIONS IN: catalytic activity; INVOLVED IN: metabolic process; LOCATED IN: chloroplast; EXPRESSED IN: 8 plant structures; EXPRESSED DURING: LP.04 four leaves visible, 4 anthesis, petal differentiation and expansion stage; CONTAINS InterPro DOMAIN/s: Pyruvate/Phosphoenolpyruvate kinase, catalytic core (InterPro:IPR015813), Isocitrate lyase/phosphorylmutase (InterPro:IPR000918); BEST Arabidopsis thaliana protein match is: Phosphoenolpyruvate carboxylase family protein (TAIR:AT1G77060.1); Has 5459 Blast hits to 5459 proteins in 1102 species: Archae - 78; Bacteria - 2773; Metazoa - 20; Fungi - 198; Plants - 102; Viruses - 0; Other Eukaryotes - 2288 (source: NCBI BLink).
c38651_g1_i1	1.62E-05	0.0143095	AT5G26740	Protein of unknown function (DUF300); LOCATED IN: endomembrane system; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Protein of unknown function DUF300 (InterPro:IPR005178); BEST Arabidopsis thaliana protein match is: Protein of unknown function (DUF300) (TAIR:AT3G05940.1); Has 921 Blast hits to 913 proteins in 195 species: Archae - 0; Bacteria - 0; Metazoa - 348; Fungi - 193; Plants - 245; Viruses - 0; Other Eukaryotes - 135 (source: NCBI BLink).

c36370_g1_i1	1.67E-05	0.0144944	NA	NA
c37286_g1_i1	1.68E-05	0.0144944	AT2G28630	Encodes KCS12, a member of the 3-ketoacyl-CoA synthase family involved in the biosynthesis of VLCFA (very long chain fatty acids).
c48428_g1_i1	1.78E-05	0.0152007	NA	NA
c37713_g1_i2	1.90E-05	0.0159914	AT2G38940	Encodes Pht1;4, a member of the Pht1 family of phosphate transporters which include: Pht1;1/At5g43350, Pht1;2/At5g43370, Pht1;3/At5g43360, Pht1;4/At2g38940, Pht1;5/At2g32830, Pht1;6/At5g43340, Pht1;7/At3g54700, Pht1;8/At1g20860, Pht1;9/At1g76430 (Plant Journal 2002, 31:341). Expression is upregulated in the shoot of cax1/cax3 mutant.
c19198_g1_i1	1.96E-05	0.0163453	AT4G29700	Alkaline-phosphatase-like family protein; FUNCTIONS IN: hydrolase activity, catalytic activity; INVOLVED IN: metabolic process, nucleotide metabolic process; LOCATED IN: vacuole; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 12 growth stages; CONTAINS InterPro DOMAIN/s: Alkaline phosphatase-like, alpha/beta/alpha (InterPro:IPR017849), Type I phosphodiesterase/nucleotide pyrophosphatase/phosphate transferase (InterPro:IPR002591), Alkaline-phosphatase-like, core domain (InterPro:IPR017850); BEST Arabidopsis thaliana protein match is: Alkaline-phosphatase-like family protein (TAIR:AT4G29690.1); Has 2237 Blast hits to 2218 proteins in 571 species: Archae - 24; Bacteria - 916; Metazoa - 704; Fungi - 177; Plants - 108; Viruses - 6; Other Eukaryotes - 302 (source: NCBI BLink).

c37360_g1_i2	2.00E-05	0.0165305	AT5G63450	member of CYP94B
c31791_g1_i5	2.11E-05	0.0171761	AT5G11410	Protein kinase superfamily protein; FUNCTIONS IN: protein kinase activity, kinase activity, ATP binding; INVOLVED IN: protein amino acid phosphorylation; EXPRESSED IN: 12 plant structures; EXPRESSED DURING: 8 growth stages; CONTAINS InterPro DOMAIN/s: Protein kinase, catalytic domain (InterPro:IPR000719), Serine/threonine-protein kinase-like domain (InterPro:IPR017442), Protein kinase-like domain (InterPro:IPR011009); BEST Arabidopsis thaliana protein match is: Protein kinase superfamily protein (TAIR:AT5G11400.2); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c40328_g2_i2	2.13E-05	0.0171761	AT1G06410	Encodes an enzyme putatively involved in trehalose biosynthesis. Though the protein has both trehalose-6-phosphate synthase (TPS)-like and trehalose-6-phosphate phosphatase (TPP)-like domains, neither activity has been detected in enzymatic assays nor has the protein been able to complement yeast TPS or TPP mutants.
c32173_g1_i2	2.20E-05	0.0175262	AT5G21930	P-Type ATPase, mediates copper transport to chloroplast thylakoid lumen. Required for accumulation of copper-containing plastocyanin in the thylakoid lumen and for effective photosynthetic electron transport
c39223_g1_i11	2.22E-05	0.0175262	AT4G33300	Encodes a member of the ADR1 family nucleotide-binding leucine-rich repeat (NB-LRR) immune receptors.

c29800_g2_i1	2.31E-05	0.0180767	AT4G37980	elicitor-activated gene 3-1 (ELI3-1); FUNCTIONS IN: oxidoreductase activity, zinc ion binding; INVOLVED IN: response to bacterium, plant-type hypersensitive response; EXPRESSED IN: 20 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: GroES-like (InterPro:IPR011032), Polyketide synthase, enoylreductase (InterPro:IPR020843), Alcohol dehydrogenase GroES-like (InterPro:IPR013154), Alcohol dehydrogenase, zinc-containing, conserved site (InterPro:IPR002328), Alcohol dehydrogenase, C-terminal (InterPro:IPR013149), Alcohol dehydrogenase superfamily, zinc-containing (InterPro:IPR002085); BEST Arabidopsis thaliana protein match is: elicitor-activated gene 3-2 (TAIR:AT4G37990.1); Has 39128 Blast hits to 39104 proteins in 3053 species: Archae - 813; Bacteria - 26065; Metazoa - 1241; Fungi - 2911; Plants - 3099; Viruses - 3; Other Eukaryotes - 4996 (source: NCBI BLink).
c39718_g1_i6	2.40E-05	0.0185356	AT5G28510	beta glucosidase 24 (BGLU24); FUNCTIONS IN: cation binding, hydrolase activity, hydrolyzing O-glycosyl compounds, catalytic activity; INVOLVED IN: carbohydrate metabolic process; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Glycoside hydrolase, family 1 (InterPro:IPR001360), Glycoside hydrolase, family 1, active site (InterPro:IPR018120), Glycoside hydrolase, catalytic core (InterPro:IPR017853), Glycoside hydrolase, subgroup, catalytic core (InterPro:IPR013781); BEST Arabidopsis thaliana protein match is: Glycosyl hydrolase superfamily protein (TAIR:AT3G09260.1); Has 11366 Blast hits to 11018 proteins in 1472 species: Archae - 140; Bacteria - 7863; Metazoa - 718; Fungi - 202; Plants - 1442; Viruses - 0; Other Eukaryotes - 1001 (source: NCBI BLink).

c38580_g1_i3	2.53E-05	0.0193164	AT4G35250	NAD(P)-binding Rossmann-fold superfamily protein; FUNCTIONS IN: binding, catalytic activity; INVOLVED IN: metabolic process; LOCATED IN: chloroplast; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: NAD(P)-binding domain (InterPro:IPR016040), NmrA-like (InterPro:IPR008030); BEST Arabidopsis thaliana protein match is: NAD(P)-binding Rossmann-fold superfamily protein (TAIR:AT2G34460.1); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c35872_g1_i1	2.62E-05	0.0193164	NA	NA
c41172_g1_i1	2.62E-05	0.0193164	AT5G11920	Encodes a protein with fructan exohydrolase (FEH) activity acting on both inulin and levan-type fructans (1- and 6-FEH). The enzyme does not have invertase activity.
c34583_g1_i2	2.63E-05	0.0193164	NA	NA
c38717_g1_i2	2.64E-05	0.0193164	AT1G78000	Encodes a sulfate transporter that can restore sulfate uptake capacity of a yeast mutant lacking sulfate transporter genes.
c35426_g1_i2	2.65E-05	0.0193164	AT5G16360	NC domain-containing protein-related; CONTAINS InterPro DOMAIN/s: NC (InterPro:IPR007053); BEST Arabidopsis thaliana protein match is: NC domain-containing protein-related (TAIR:AT3G02700.1); Has 1807 Blast hits to 1807 proteins in 277 species:

				Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c34371_g1_i4	2.76E-05	0.0198906	AT3G05165	Major facilitator superfamily protein; FUNCTIONS IN: substrate-specific transmembrane transporter activity, carbohydrate transmembrane transporter activity, transporter activity, sugar:hydrogen symporter activity; INVOLVED IN: transport, transmembrane transport; LOCATED IN: integral to membrane, membrane; EXPRESSED IN: male gametophyte, pollen tube; EXPRESSED DURING: L mature pollen stage, M germinated pollen stage; CONTAINS InterPro DOMAIN/s: Sugar transporter, conserved site (InterPro:IPR005829), Major facilitator superfamily (InterPro:IPR020846), General substrate transporter (InterPro:IPR005828), Sugar/inositol transporter (InterPro:IPR003663), Major facilitator superfamily, general substrate transporter (InterPro:IPR016196); BEST Arabidopsis thaliana protein match is: Major facilitator superfamily protein (TAIR:AT3G05160.1).
c36892_g1_i3	2.83E-05	0.0202106	AT5G19485	transferases;nucleotidyltransferases; FUNCTIONS IN: transferase activity, nucleotidyltransferase activity; INVOLVED IN: biosynthetic process; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Trimeric LpxA-like (InterPro:IPR011004), Nucleotidyl transferase (InterPro:IPR005835); BEST Arabidopsis thaliana protein match is: Trimeric LpxA-like enzyme (TAIR:AT2G34970.1); Has 6119 Blast hits to 5902 proteins in 1501 species: Archae - 491; Bacteria - 3115; Metazoa - 435; Fungi - 482; Plants - 325; Viruses - 0; Other Eukaryotes - 1271 (source: NCBI BLink).

c38090_g1_i4	2.87E-05	0.0202149	AT4G39030	Encodes an orphan multidrug and toxin extrusion transporter. Essential component of salicylic acid-dependent signaling for disease resistance. Member of the MATE-transporter family. Expression induced by salicylic acid. Mutants are salicylic acid-deficient.
c25449_g1_i1	2.89E-05	0.0202149	AT1G31280	Encodes Argonaute gene that binds viral siRNAs and is involved in antiviral defense response. Regulates innate immunity.
c34637_g1_i1	3.02E-05	0.0209729	AT2G28670	esb1 mutants have increased levels of suberin and altered levels of several ions in their leaves.
c33029_g1_i5	3.15E-05	0.0216788	AT1G22400	UGT85A1; FUNCTIONS IN: in 6 functions; INVOLVED IN: metabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 10 growth stages; CONTAINS InterPro DOMAIN/s: UDP-glucuronosyl/UDP-glucosyltransferase (InterPro:IPR002213); BEST Arabidopsis thaliana protein match is: UDP-glucosyl transferase 85A3 (TAIR:AT1G22380.1); Has 7940 Blast hits to 7832 proteins in 421 species: Archae - 0; Bacteria - 227; Metazoa - 2330; Fungi - 36; Plants - 5216; Viruses - 60; Other Eukaryotes - 71 (source: NCBI BLink).
c31261_g1_i8	3.19E-05	0.0217074	AT1G08680	A member of ARF GAP domain (AGD), A thaliana has 15 members, grouped into four classes. AGD14 belongs to the class 4, together with AGD15.
c32950_g1_i3	3.47E-05	0.0234359	AT4G37400	member of CYP81F
c30375_g1_i2	3.58E-05	0.023895	AT3G04310	unknown protein; Has 44 Blast hits to 44 proteins in 12 species: Archae - 0; Bacteria - 0; Metazoa - 0; Fungi - 0; Plants - 44; Viruses - 0; Other Eukaryotes - 0 (source: NCBI BLink).

c32583_g1_i1	3.61E-05	0.023895	AT5G65280	Encodes a protein with reported similarity to GCR2 a putative G protein coupled receptor thought to be an ABA receptor. Loss of function mutations in GCL1 show no ABA response defects based on assays of seed germination and seedling development.GCL1 also has similarity to LANCL1 and LANCL2, human homologs of bacterial lanthionine synthetase.
c38749_g4_i3	3.69E-05	0.0242198	AT4G10530	Subtilase family protein; FUNCTIONS IN: identical protein binding, serine-type endopeptidase activity; INVOLVED IN: proteolysis, negative regulation of catalytic activity; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Protease-associated PA (InterPro:IPR003137), Proteinase inhibitor, propeptide (InterPro:IPR009020), Peptidase S8/S53, subtilisin/kexin/sedolisin (InterPro:IPR000209), Peptidase S8, subtilisin-related (InterPro:IPR015500), Proteinase inhibitor I9, subtilisin propeptide (InterPro:IPR010259), Peptidase S8/S53, subtilisin, active site (InterPro:IPR022398); BEST Arabidopsis thaliana protein match is: Subtilase family protein (TAIR:AT4G10520.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c35006_g3_i1	3.99E-05	0.025951	AT3G21340	Leucine-rich repeat protein kinase family protein; FUNCTIONS IN: kinase activity; INVOLVED IN: protein amino acid phosphorylation; LOCATED IN: endomembrane system; EXPRESSED IN: root; CONTAINS InterPro DOMAIN/s: Protein kinase, ATP binding site (InterPro:IPR017441), Protein kinase, catalytic domain (InterPro:IPR000719), Leucine-rich repeat (InterPro:IPR001611), Serine/threonine-protein kinase-like domain (InterPro:IPR017442), Protein kinase-like domain (InterPro:IPR011009), Serine/threonine-protein kinase, active site (InterPro:IPR008271); BEST Arabidopsis thaliana protein match is:

				Leucine-rich repeat protein kinase family protein (TAIR:AT1G51850.1); Has 160620 Blast hits to 122303 proteins in 4467 species: Archae - 99; Bacteria - 13621; Metazoa - 44084; Fungi - 10001; Plants - 73653; Viruses - 414; Other Eukaryotes - 18748 (source: NCBI BLink).
c40398_g1_i1	4.04E-05	0.026052	AT1G05380	Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger protein; CONTAINS InterPro DOMAIN/s: Zinc finger, PHD-type (InterPro:IPR001965), Zinc finger, FYVE/PHD-type (InterPro:IPR011011), Acyl-CoA N-acyltransferase (InterPro:IPR016181), Zinc finger, PHD-finger (InterPro:IPR019787); BEST Arabidopsis thaliana protein match is: Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger protein (TAIR:AT4G14920.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c38624_g2_i3	4.26E-05	0.0272438	AT5G54710	Ankyrin repeat family protein; INVOLVED IN: biological_process unknown; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Ankyrin repeat-containing domain (InterPro:IPR020683), Ankyrin repeat (InterPro:IPR002110); BEST Arabidopsis thaliana protein match is: Ankyrin repeat family protein (TAIR:AT1G34050.1); Has 20807 Blast hits to 11728 proteins in 551 species: Archae - 16; Bacteria - 1568; Metazoa - 11429; Fungi - 1235; Plants - 2476; Viruses - 139; Other Eukaryotes - 3944 (source: NCBI BLink).

c40298_g2_i1	4.39E-05	0.0276613	AT3G01310	Phosphoglycerate mutase-like family protein; FUNCTIONS IN: oxidoreductase activity, transition metal ion binding, acid phosphatase activity; INVOLVED IN: oxidation reduction; LOCATED IN: plasma membrane; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: ATP-grasp fold, RimK-type (InterPro:IPR013651), Histidine phosphatase superfamily, clade-2 (InterPro:IPR000560), Ferritin/ribonucleotide reductase-like (InterPro:IPR009078); BEST Arabidopsis thaliana protein match is: Phosphoglycerate mutase-like family protein (TAIR:AT5G15070.1); Has 35333 Blast hits to 34131 proteins in 2444 species: Archae - 798; Bacteria - 22429; Metazoa - 974; Fungi - 991; Plants - 531; Viruses - 0; Other Eukaryotes - 9610 (source: NCBI BLink).
c26576_g1_i3	4.40E-05	0.0276613	AT1G01680	plant U-box 54 (PUB54); FUNCTIONS IN: ubiquitin-protein ligase activity; INVOLVED IN: response to stress, protein ubiquitination; LOCATED IN: ubiquitin ligase complex; CONTAINS InterPro DOMAIN/s: UspA (InterPro:IPR006016), U box domain (InterPro:IPR003613); BEST Arabidopsis thaliana protein match is: RING/U-box superfamily protein (TAIR:AT1G01660.1); Has 2251 Blast hits to 2122 proteins in 140 species: Archae - 0; Bacteria - 22; Metazoa - 203; Fungi - 25; Plants - 1809; Viruses - 3; Other Eukaryotes - 189 (source: NCBI BLink).
c34972_g1_i2	4.52E-05	0.0281733	AT4G11120	translation elongation factor Ts (EF-Ts), putative; FUNCTIONS IN: translation elongation factor activity; INVOLVED IN: translational elongation; LOCATED IN: mitochondrion; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Translation elongation factor Ts, conserved site (InterPro:IPR018101), Translation elongation factor EFTs/EF1B (InterPro:IPR001816), UBA-like (InterPro:IPR009060), Translation elongation factor EFTs/EF1B, dimerisation

				(InterPro:IPR014039); BEST Arabidopsis thaliana protein match is: elongation factor Ts family protein (TAIR:AT4G29060.1); Has 9471 Blast hits to 8571 proteins in 2664 species: Archae - 0; Bacteria - 5908; Metazoa - 120; Fungi - 25; Plants - 215; Viruses - 0; Other Eukaryotes - 3203 (source: NCBI BLink).
c38167_g2_i5	4.56E-05	0.0282084	AT1G47530	MATE efflux family protein; FUNCTIONS IN: antiporter activity, drug transmembrane transporter activity, transporter activity; INVOLVED IN: drug transmembrane transport, ripening, transmembrane transport; LOCATED IN: plasma membrane, membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Multi antimicrobial extrusion protein MatE (InterPro:IPR002528); BEST Arabidopsis thaliana protein match is: root hair specific 2 (TAIR:AT1G12950.1); Has 9964 Blast hits to 9893 proteins in 2013 species: Archae - 182; Bacteria - 7111; Metazoa - 140; Fungi - 326; Plants - 1355; Viruses - 0; Other Eukaryotes - 850 (source: NCBI BLink).
c33496_g1_i4	4.92E-05	0.0301514	AT3G55960	Haloacid dehalogenase-like hydrolase (HAD) superfamily protein; FUNCTIONS IN: phosphatase activity; INVOLVED IN: biological_process unknown; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: Dullard-like phosphatase domain (InterPro:IPR011948), NLI interacting factor (InterPro:IPR004274); BEST Arabidopsis thaliana protein match is: Haloacid dehalogenase-like hydrolase (HAD) superfamily protein (TAIR:AT1G29780.1); Has 2169 Blast hits to 2162

				proteins in 238 species: Archae - 0; Bacteria - 0; Metazoa - 723; Fungi - 425; Plants - 384; Viruses - 0; Other Eukaryotes - 637 (source: NCBI BLink).
c37140_g1_i2	5.17E-05	0.0314804	AT5G41040	Encodes a feruloyl-CoA transferase required for suberin synthesis. Has feruloyl-CoA-dependent feruloyl transferase activity towards substrates with a primary alcohol.
c36779_g1_i1	5.60E-05	0.0337797	AT4G21570	Protein of unknown function (DUF300); INVOLVED IN: biological_process unknown; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Protein of unknown function DUF300 (InterPro:IPR005178); BEST Arabidopsis thaliana protein match is: Protein of unknown function (DUF300) (TAIR:AT1G11200.1); Has 836 Blast hits to 835 proteins in 194 species: Archae - 0; Bacteria - 0; Metazoa - 284; Fungi - 192; Plants - 240; Viruses - 0; Other Eukaryotes - 120 (source: NCBI BLink).
c32426_g1_i4	5.74E-05	0.0343751	AT3G33530	Transducin family protein / WD-40 repeat family protein; FUNCTIONS IN: nucleotide binding; INVOLVED IN: biological_process unknown; LOCATED IN: chloroplast; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: WD40 repeat-like-containing domain (InterPro:IPR011046), WD40/YVTN repeat-like-containing domain (InterPro:IPR015943); BEST Arabidopsis thaliana protein match is: Transducin family protein / WD-40 repeat family protein (TAIR:AT2G26610.1); Has 166 Blast

				hits to 141 proteins in 59 species: Archae - 0; Bacteria - 0; Metazoa - 92; Fungi - 4; Plants - 46; Viruses - 0; Other Eukaryotes - 24 (source: NCBI BLink).
c37366_g1_i1	5.80E-05	0.0344271	AT2G46430	cyclic nucleotide gated channel (CNGC4), downstream component of the signaling pathways leading to HR/resistance
c26859_g1_i1	6.01E-05	0.0354057	AT1G12080	Vacuolar calcium-binding protein-related; BEST Arabidopsis thaliana protein match is: Vacuolar calcium-binding protein-related (TAIR:AT1G62480.1); Has 21609 Blast hits to 9169 proteins in 1089 species: Archae - 145; Bacteria - 3596; Metazoa - 6157; Fungi - 1827; Plants - 760; Viruses - 333; Other Eukaryotes - 8791 (source: NCBI BLink).
c34475_g3_i1	6.12E-05	0.0357904	NA	NA
c39640_g2_i2	6.48E-05	0.0376026	AT1G21980	Type I phosphatidylinositol-4-phosphate 5-kinase. Preferentially phosphorylates PtdIns4P. Induced by water stress and abscisic acid in Arabidopsis thaliana. Expressed in procambial cells of leaves, flowers and roots. A N-terminal Membrane Occupation and Recognition Nexus (MORN)affects enzyme activity and distribution.
c30065_g1_i1	6.64E-05	0.0381889	AT4G33000	Encodes a member of the calcineurin B-like calcium sensor gene family. Mediates salt tolerance by regulating ion homeostasis in Arabidopsis. In the shoot SCABP recruits the protein kinase

				SOS2 to the plasma membrane. SCABP is partially redundant with SOS3 but neither protein can fully complement the loss of the other protein.
c35911_g1_i2	6.85E-05	0.039104	AT3G57030	Calcium-dependent phosphotriesterase superfamily protein; FUNCTIONS IN: strictosidine synthase activity; INVOLVED IN: alkaloid biosynthetic process, biosynthetic process; LOCATED IN: endoplasmic reticulum, plasma membrane, plant-type cell wall; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Strictosidine synthase, conserved region (InterPro:IPR018119), Strictosidine synthase (InterPro:IPR004141), Six-bladed beta-propeller, TolB-like (InterPro:IPR011042); BEST Arabidopsis thaliana protein match is: Calcium-dependent phosphotriesterase superfamily protein (TAIR:AT5G22020.1); Has 1145 Blast hits to 1130 proteins in 241 species: Archae - 1; Bacteria - 292; Metazoa - 224; Fungi - 14; Plants - 486; Viruses - 0; Other Eukaryotes - 128 (source: NCBI BLink).
c36248_g2_i1	7.76E-05	0.0436887	AT5G20960	Encodes aldehyde oxidase AA01.

c31746_g1_i2	7.77E-05	0.0436887	AT5G18900	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein; FUNCTIONS IN: oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors, oxidoreductase activity, oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, L-ascorbic acid binding, iron ion binding; INVOLVED IN: oxidation reduction, peptidyl-proline hydroxylation to 4-hydroxy-L-proline; LOCATED IN: nucleus, cytoplasm; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 15 growth stages; CONTAINS InterPro DOMAIN/s: Prolyl 4-hydroxylase, alpha subunit (InterPro:IPR006620), Oxoglutarate/iron-dependent oxygenase (InterPro:IPR005123), Metridin-like ShK toxin (InterPro:IPR003582); BEST Arabidopsis thaliana protein match is: P4H isoform 2 (TAIR:AT3G06300.1); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c41293_g1_i1	7.84E-05	0.0437802	AT1G53625	unknown protein; Has 29996 Blast hits to 6987 proteins in 655 species: Archae - 23; Bacteria - 6686; Metazoa - 10521; Fungi - 1178; Plants - 7439; Viruses - 681; Other Eukaryotes - 3468 (source: NCBI BLink).
c38929_g1_i4	8.13E-05	0.0442276	AT1G36370	Encodes a putative serine hydroxymethyltransferase.
c27794_g1_i1	8.14E-05	0.0442276	NA	NA

c31561_g1_i1	8.16E-05	0.0442276	AT5G57510	unknown protein; Has 27 Blast hits to 27 proteins in 9 species: Archae - 0; Bacteria - 0; Metazoa - 0; Fungi - 0; Plants - 27; Viruses - 0; Other Eukaryotes - 0 (source: NCBI BLink).
c39816_g1_i3	8.17E-05	0.0442276	AT3G23410	Encodes a fatty alcohol oxidase.
c24178_g1_i1	8.22E-05	0.0442276	NA	NA
c40454_g3_i4	8.42E-05	0.0449704	AT1G72140	Major facilitator superfamily protein; FUNCTIONS IN: transporter activity; INVOLVED IN: oligopeptide transport, response to nematode; LOCATED IN: plasma membrane, membrane; EXPRESSED IN: 18 plant structures; EXPRESSED DURING: 8 growth stages; CONTAINS InterPro DOMAIN/s: Oligopeptide transporter (InterPro:IPR000109), Major facilitator superfamily, general substrate transporter (InterPro:IPR016196); BEST Arabidopsis thaliana protein match is: Major facilitator superfamily protein (TAIR:AT1G22540.1); Has 7231 Blast hits to 7080 proteins in 1330 species: Archae - 0; Bacteria - 3541; Metazoa - 585; Fungi - 456; Plants - 2171; Viruses - 0; Other Eukaryotes - 478 (source: NCBI BLink).
c14973_g1_i1	8.81E-05	0.0458757	AT3G47930	L-Galactono-1,4-lactone dehydrogenase, catalyzes the final step of ascorbate biosynthesis
c28918_g2_i2	8.81E-05	0.0458757	NA	NA
c22782_g1_i2	8.83E-05	0.0458757	AT3G22060	contains Pfam profile: PF01657 Domain of unknown function that is usually associated with protein kinase domain Pfam:PF00069, however this protein does not have the protein kinase domain

c38932_g1_i3	8.84E-05	0.0458757	AT5G54170	Polyketide cyclase/dehydrase and lipid transport superfamily protein; CONTAINS InterPro DOMAIN/s: Lipid-binding START (InterPro:IPR002913); BEST Arabidopsis thaliana protein match is: Polyketide cyclase/dehydrase and lipid transport superfamily protein (TAIR:AT1G64720.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c34036_g1_i3	9.00E-05	0.0463838	AT5G05670	signal recognition particle binding; FUNCTIONS IN: signal recognition particle binding; LOCATED IN: endoplasmic reticulum, plasma membrane; EXPRESSED IN: 25 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: Signal recognition particle receptor, beta subunit (InterPro:IPR019009); BEST Arabidopsis thaliana protein match is: P-loop containing nucleoside triphosphate hydrolases superfamily protein (TAIR:AT2G18770.1); Has 1807 Blast hits to 1807 proteins in 277 species: Archae - 0; Bacteria - 0; Metazoa - 736; Fungi - 347; Plants - 385; Viruses - 0; Other Eukaryotes - 339 (source: NCBI BLink).
c38341_g2_i2	9.17E-05	0.0469445	AT3G29320	Encodes a plastidic alpha-glucan phosphorylase. In vitro, the enzyme has a preference for maltooligosaccharides, such as maltoheptaose.

A.7 Table S2.5: Differentially expressed transcripts in high and low glucosinolate watercress samples

Comprehensive list of the 94 differentially expressed watercress transcripts in the high and low glucosinolate analysis (α ≤ 0.05). For each differentially expressed transcript isoform, both the uncorrected and corrected (for multiple hypothesis testing – false discovery rate) p value are given, alongside the best match locus ID and gene description from annotation

Wx isoform	P value	FDR	Loci ID	Gene description
c40368_g2_i4	1.94E-20	1.37E-15	AT1G14040	EXS (ERD1/XPR1/SYG1) family protein; LOCATED IN: integral to membrane; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 12 growth stages; CONTAINS InterPro DOMAIN/s: SPX, N-terminal (InterPro:IPR004331), EXS, C-terminal (InterPro:IPR004342); BEST Arabidopsis thaliana protein match is: EXS (ERD1/XPR1/SYG1) family protein (TAIR:AT2G03240.1); Has 1168 Blast hits to 1094 proteins in 207 species: Archae - 0; Bacteria - 0; Metazoa - 260; Fungi - 396; Plants - 362; Viruses - 0; Other Eukaryotes - 150 (source: NCBI BLink).
c40593_g2_i4	3.33E-10	1.13E-05	AT5G15680	ARM repeat superfamily protein; FUNCTIONS IN: binding; INVOLVED IN: biological_process unknown; LOCATED IN: plasma membrane, membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 12 growth stages; CONTAINS InterPro DOMAIN/s: Armadillo-type fold (InterPro:IPR016024); Has 30201 Blast hits to 17322

				proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c23161_g1_i1	5.50E-10	1.13E-05	NA	NA
c37556_g1_i9	6.41E-10	1.13E-05	AT1G32850	ubiquitin-specific protease 11 (UBP11); FUNCTIONS IN: cysteine-type endopeptidase activity, ubiquitin thiolesterase activity; INVOLVED IN: ubiquitin-dependent protein catabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: sperm cell, male gametophyte, pollen tube; EXPRESSED DURING: M germinated pollen stage; CONTAINS InterPro DOMAIN/s: Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2, conserved site (InterPro:IPR018200), Peptidase C19, ubiquitin-specific peptidase, DUSP domain (InterPro:IPR006615), Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2 (InterPro:IPR001394); BEST Arabidopsis thaliana protein match is: ubiquitin-specific protease 10 (TAIR:AT4G10590.1); Has 11311 Blast hits to 7509 proteins in 254 species: Archae - 0; Bacteria - 6; Metazoa - 5825; Fungi - 2000; Plants - 1417; Viruses - 8; Other Eukaryotes - 2055 (source: NCBI BLink).
c29502_g1_i3	8.38E-10	1.18E-05	AT1G01060	LHY encodes a myb-related putative transcription factor involved in circadian rhythm along with another myb transcription factor CCA1

c39754_g1_i5	1.22E-09	1.43E-05	AT1G27752	Ubiquitin system component Cue protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: chloroplast; CONTAINS InterPro DOMAIN/s: Ubiquitin system component Cue (InterPro:IPR003892); Has 423 Blast hits to 397 proteins in 140 species: Archae - 0; Bacteria - 35; Metazoa - 153; Fungi - 103; Plants - 62; Viruses - 0; Other Eukaryotes - 70 (source: NCBI BLink).
c40489_g1_i2	6.71E-09	6.75E-05	AT3G19960	member of Myosin-like proteins
c30515_g1_i2	1.21E-08	0.0001065	NA	NA
c33663_g1_i2	2.35E-08	0.0001836	AT2G35500	Encodes a protein with some sequence similarity to shikimate kinases, but a truncated form of this protein (lacking a putative N-terminal chloroplast transit peptide) does not have shikimate kinase activity in vitro.
c36520_g2_i1	3.99E-08	0.0002809	AT1G28350	Nucleotidylyl transferase superfamily protein; FUNCTIONS IN: tyrosine-tRNA ligase activity, nucleotide binding, aminoacyl-tRNA ligase activity, ATP binding; INVOLVED IN: translation, tyrosyl-tRNA aminoacylation, tRNA aminoacylation for protein translation; LOCATED IN: cytoplasm; CONTAINS InterPro DOMAIN/s: Rossmann-like alpha/beta/alpha sandwich fold (InterPro:IPR014729), Tyrosyl-tRNA synthetase, class Ib, archaeal/eukaryotic cytosolic (InterPro:IPR015624), Tyrosyl-tRNA synthetase, class Ib, bacterial/mitochondrial (InterPro:IPR002307), Aminoacyl-tRNA synthetase, class Ib (InterPro:IPR002305); BEST Arabidopsis thaliana protein match is: Tyrosyl-tRNA synthetase, class Ib, bacterial/mitochondrial (TAIR:AT2G33840.1); Has 6761 Blast hits to 4745 proteins in 1539

				species: Archae - 718; Bacteria - 3084; Metazoa - 577; Fungi - 521; Plants - 241; Viruses - 10; Other Eukaryotes - 1610 (source: NCBI BLink).
c35180_g1_i3	4.57E-08	0.0002927	AT1G08260	Similar to POL2A, DNA polymerase epsilon catalytic subunit. Essential for Arabidopsis growth. Null homozygotes are embryo lethal, partial loss of function alleles show embryo patterning defects such as root pole displacement. Delayed progression through cell cycle results in embryos with smaller numbers of larger cells.
c30367_g1_i3	8.70E-08	0.0004596	AT3G44820	Phototropic-responsive NPH3 family protein; FUNCTIONS IN: signal transducer activity; INVOLVED IN: response to light stimulus; LOCATED IN: cellular_component unknown; EXPRESSED IN: 8 plant structures; EXPRESSED DURING: LP.06 six leaves visible, LP.04 four leaves visible, 4 anthesis, 4 leaf senescence stage, petal differentiation and expansion stage; CONTAINS InterPro DOMAIN/s: NPH3 (InterPro:IPR004249), BTB/POZ (InterPro:IPR013069), BTB/POZ fold (InterPro:IPR011333), BTB/POZ-like (InterPro:IPR000210); BEST Arabidopsis thaliana protein match is: Phototropic-responsive NPH3 family protein (TAIR:AT1G30440.1); Has 907 Blast hits to 880 proteins in 40 species:

				Archae - 0; Bacteria - 0; Metazoa - 32; Fungi - 0; Plants - 873; Viruses - 0; Other Eukaryotes - 2 (source: NCBI BLink).
c40656_g1_i2	9.01E-08	0.0004596	NA	NA
c40035_g1_i2	9.49E-08	0.0004596	AT4G22505	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein; INVOLVED IN: lipid transport; CONTAINS InterPro DOMAIN/s: Bifunctional inhibitor/plant lipid transfer protein/seed storage (InterPro:IPR016140), Plant lipid transfer protein/hydrophobic protein, helical domain (InterPro:IPR013770); BEST Arabidopsis thaliana protein match is: protease inhibitor/seed storage/lipid transfer protein (LTP) family protein (TAIR:AT4G22470.1); Has 2123 Blast hits to 2070 proteins in 299 species: Archae - 1; Bacteria - 212; Metazoa - 674; Fungi - 154; Plants - 889; Viruses - 39; Other Eukaryotes - 154 (source: NCBI BLink).
c38877_g2_i6	9.79E-08	0.0004596	AT5G41100	FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: plasma membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; BEST Arabidopsis thaliana protein match is: hydroxyproline-rich glycoprotein family protein (TAIR:AT3G26910.2); Has 1497 Blast hits to

				1191 proteins in 214 species: Archae - 4; Bacteria - 102; Metazoa - 485; Fungi - 316; Plants - 187; Viruses - 37; Other Eukaryotes - 366 (source: NCBI BLink).
c37926_g1_i6	1.13E-07	0.0004953	AT3G06350	Encodes a bi-functional dehydroquinate-shikimate dehydrogenase enzyme that catalyzes two steps in the chorismate biosynthesis pathway.
c30062_g2_i2	1.56E-07	0.0006467	NA	NA
c30746_g1_i1	1.86E-07	0.000727	AT1G56070	encodes a translation elongation factor 2-like protein that is involved in cold-induced translation. Mutations in this gene specifically blocks low temperature-induced transcription of cold-responsive genes but induces the expression of CBF genes and mutants carrying the recessive mutations fail to acclimate to cold and is freezing sensitive.
c51516_g1_i1	2.66E-07	0.0009496	NA	NA

c29605_g1_i1	2.83E-07	0.0009496	AT5G60390	GTP binding Elongation factor Tu family protein; FUNCTIONS IN: calmodulin binding, translation elongation factor activity; INVOLVED IN: translational elongation; LOCATED IN: mitochondrion, nucleus, cytoplasm; EXPRESSED IN: cotyledon, male gametophyte, guard cell, pollen tube, seed; EXPRESSED DURING: L mature pollen stage, M germinated pollen stage, seed development stages; CONTAINS InterPro DOMAIN/s: Translation elongation factor EFTu/EF1A, C-terminal (InterPro:IPR004160), Translation elongation factor EFTu/EF1A, domain 2 (InterPro:IPR004161), Translation elongation factor EF1A/initiation factor IF2gamma, C-terminal (InterPro:IPR009001), Protein synthesis factor, GTP-binding (InterPro:IPR000795), Translation elongation/initiation factor/Ribosomal, beta-barrel (InterPro:IPR009000), Translation elongation factor EF1A, eukaryotic/archaeal (InterPro:IPR004539); BEST Arabidopsis thaliana protein match is: GTP binding Elongation factor Tu family protein (TAIR:AT1G07940.2); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c39727_g3_i2	2.83E-07	0.0009496	AT4G25700	Converts beta-carotene to zeaxanthin via cryptoxanthin.
c39217_g2_i1	3.20E-07	0.0010238	AT1G50010	Encodes alpha-2,4 tubulin. TUA2 and TUA4 encode identical proteins.
c36454_g1_i1	4.94E-07	0.0015138	NA	NA
c38907_g1_i1	5.64E-07	0.0016559	AT3G52890	KCBP-interacting protein kinase interacts specifically with the tail region of KCBP

c40126_g1_i2	1.10E-06	0.0029596	AT1G35220	unknown protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: cellular_component unknown; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; Has 313 Blast hits to 185 proteins in 75 species: Archae - 0; Bacteria - 0; Metazoa - 200; Fungi - 0; Plants - 67; Viruses - 0; Other Eukaryotes - 46 (source: NCBI BLink).
c24927_g1_i2	1.13E-06	0.0029596	AT3G09770	Encodes a ubiquitin E3 ligase LOG2 (LOSS OF GDU2). Required for GLUTAMINE DUMPER1(GDU1)-induced amino secretion.
c36393_g1_i2	1.13E-06	0.0029596	NA	NA
c26443_g1_i3	1.28E-06	0.0030045	AT5G04070	NAD(P)-binding Rossmann-fold superfamily protein; FUNCTIONS IN: oxidoreductase activity, binding, catalytic activity; INVOLVED IN: oxidation reduction, metabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: NAD(P)-binding domain (InterPro:IPR016040), Glucose/ribitol dehydrogenase (InterPro:IPR002347), Short-chain dehydrogenase/reductase SDR (InterPro:IPR002198); BEST Arabidopsis thaliana protein match is: NAD(P)-binding Rossmann-fold superfamily protein (TAIR:AT5G15940.1); Has 38503 Blast hits to 38471 proteins in 2813 species: Archae - 360; Bacteria - 23739; Metazoa - 3799; Fungi - 2489; Plants - 1519; Viruses - 0; Other Eukaryotes - 6597 (source: NCBI BLink).

c38590_g1_i1	1.28E-06	0.0030045	AT1G09160	Protein phosphatase 2C family protein; FUNCTIONS IN: protein serine/threonine phosphatase activity, catalytic activity; LOCATED IN: plasma membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Protein phosphatase 2C-related (InterPro:IPR001932), Protein phosphatase 2C (InterPro:IPR015655), Protein phosphatase 2C, N-terminal (InterPro:IPR014045); BEST Arabidopsis thaliana protein match is: Protein phosphatase 2C family protein (TAIR:AT1G68410.2); Has 5570 Blast hits to 5569 proteins in 486 species: Archae - 4; Bacteria - 416; Metazoa - 1194; Fungi - 467; Plants - 2390; Viruses - 7; Other Eukaryotes - 1092 (source: NCBI BLink).
c35497_g1_i2	1.28E-06	0.0030045	AT3G58830	haloacid dehalogenase (HAD) superfamily protein; FUNCTIONS IN: catalytic activity; LOCATED IN: chloroplast; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 15 growth stages; CONTAINS InterPro DOMAIN/s: HAD-superfamily hydrolase, subfamily IIIA (InterPro:IPR006549), Protein of unknown function DUF2010 (InterPro:IPR019001), HAD-superfamily phosphatase, subfamily IIIA (InterPro:IPR010021); Has 1169 Blast hits to 1165 proteins in 593 species: Archae - 0; Bacteria - 976; Metazoa - 0; Fungi - 84; Plants - 42; Viruses - 0; Other Eukaryotes - 67 (source: NCBI BLink).
c38005_g3_i1	1.45E-06	0.0032945	AT3G53750	Member of the Actin gene family. Expressed in mature pollen.
c37295_g1_i8	1.52E-06	0.0033472	AT3G27560	encodes a protein with kinase domains, including catalytic domains for serine/threonine as well as tyrosine kinases. a member of multi-gene family and is expressed in all tissues examined.

c27828_g2_i1	1.61E-06	0.0034287	AT1G49760	polyadenylate-binding protein, putative / PABP, putative, similar to poly(A)-binding protein GB:AAF66825 GI:7673359 from (Nicotiana tabacum). Highly and ubiquitously expressed. Member of the class II PABP family.
c38857_g8_i9	1.80E-06	0.0037179	AT1G11950	Transcription factor jumonji (jmjC) domain-containing protein; CONTAINS InterPro DOMAIN/s: Transcription factor jumonji/aspartyl beta-hydroxylase (InterPro:IPR003347), Transcription factor jumonji (InterPro:IPR013129); BEST Arabidopsis thaliana protein match is: transcription factor jumonji (jmjC) domain-containing protein (TAIR:AT1G62310.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c34012_g2_i2	1.95E-06	0.003923	AT1G19170	Pectin lyase-like superfamily protein; FUNCTIONS IN: polygalacturonase activity; INVOLVED IN: carbohydrate metabolic process; LOCATED IN: mitochondrion; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Pectin lyase fold/virulence factor (InterPro:IPR011050), Parallel beta-helix repeat (InterPro:IPR006626), Pectin lyase fold (InterPro:IPR012334), Glycoside hydrolase, family 28 (InterPro:IPR000743); BEST Arabidopsis thaliana protein match is: Pectin lyase-like superfamily protein (TAIR:AT3G42950.1); Has 4384 Blast hits to 4371 proteins in 509 species: Archae - 6; Bacteria - 1552; Metazoa - 14; Fungi - 1224; Plants - 1428; Viruses - 0; Other Eukaryotes - 160 (source: NCBI BLink).

c29073_g1_i1	2.04E-06	0.0039808	NA	NA
c31735_g1_i1	2.33E-06	0.0044293	NA	NA
c40124_g1_i6	3.32E-06	0.0061074	AT5G12430	Encodes one of the 36 carboxylate clamp (CC)-tetratricopeptide repeat (TPR) proteins (Prasad 2010, Pubmed ID: 20856808) with potential to interact with Hsp90/Hsp70 as co-chaperones.
c39412_g2_i4	3.38E-06	0.0061074	AT5G60600	Encodes a chloroplast-localized hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate (HMBPP) synthase (HDS), catalyzes the formation of HMBPP from 2-C-methyl-D-erythrytol 2,4-cyclodiphosphate (MEcPP). The HDS enzyme controls the penultimate steps of the biosynthesis of IPP and dimethylallyl diphosphate (DMAPP) via the MEP pathway and may serve as a metabolic control point for SA-mediated disease resistance. In the light, the electrons required for the reaction catalyzed by HDS are directly provided by the electron flow from photosynthesis via ferredoxin. In the dark however, the enzyme requires an electron shuttle: ferredoxin-NADP<sup>+</sup> reductase.
c36893_g1_i1	3.74E-06	0.0065565	NA	NA
c40271_g1_i3	3.82E-06	0.0065565	AT1G67110	member of CYP709A
c31679_g1_i1	4.01E-06	0.0067185	NA	NA

c29079_g1_i2	4.89E-06	0.0080002	AT1G53170	encodes a member of the ERF (ethylene response factor) subfamily B-1 of ERF/AP2 transcription factor family (ATERF-8). The protein contains one AP2 domain. There are 15 members in this subfamily including ATERF-3, ATERF-4, ATERF-7, and leafy petiole.
c23602_g1_i1	5.18E-06	0.0082837	AT5G58480	O-Glycosyl hydrolases family 17 protein; FUNCTIONS IN: cation binding, hydrolase activity, hydrolyzing O-glycosyl compounds, catalytic activity; INVOLVED IN: carbohydrate metabolic process; LOCATED IN: anchored to plasma membrane, plasma membrane, anchored to membrane; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: X8 (InterPro:IPR012946), Glycoside hydrolase, catalytic core (InterPro:IPR017853), Glycoside hydrolase, family 17 (InterPro:IPR000490), Glycoside hydrolase, subgroup, catalytic core (InterPro:IPR013781); BEST Arabidopsis thaliana protein match is: O-Glycosyl hydrolases family 17 protein (TAIR:AT4G17180.1); Has 2632 Blast hits to 2560 proteins in 135 species: Archae - 0; Bacteria - 0; Metazoa - 4; Fungi - 17; Plants - 2601; Viruses - 0; Other Eukaryotes - 10 (source: NCBI BLink).
c28698_g1_i1	5.56E-06	0.0086994	AT5G02500	encodes a member of heat shock protein 70 family.
c38958_g1_i1	5.85E-06	0.0089589	AT3G58560	DNAse I-like superfamily protein; CONTAINS InterPro DOMAIN/s: Endonuclease/exonuclease/phosphatase (InterPro:IPR005135); BEST Arabidopsis thaliana protein match is: DNAse I-like superfamily protein (TAIR:AT3G58580.1); Has 1372 Blast hits

				to 1328 proteins in 220 species: Archae - 0; Bacteria - 20; Metazoa - 540; Fungi - 247; Plants - 315; Viruses - 0; Other Eukaryotes - 250 (source: NCBI BLink).
c35692_g1_i4	6.47E-06	0.0096663	AT4G35730	Regulator of Vps4 activity in the MVB pathway protein; CONTAINS InterPro DOMAIN/s: Protein of unknown function DUF292, eukaryotic (InterPro:IPR005061); BEST Arabidopsis thaliana protein match is: Regulator of Vps4 activity in the MVB pathway protein (TAIR:AT1G34220.2); Has 794 Blast hits to 754 proteins in 186 species: Archae - 0; Bacteria - 4; Metazoa - 200; Fungi - 205; Plants - 312; Viruses - 0; Other Eukaryotes - 73 (source: NCBI BLink).
c37852_g1_i1	6.59E-06	0.0096663	NA	NA
c29032_g1_i4	6.93E-06	0.0098144	AT4G13030	P-loop containing nucleoside triphosphate hydrolases superfamily protein; Has 35333 Blast hits to 34131 proteins in 2444 species: Archae - 798; Bacteria - 22429; Metazoa - 974; Fungi - 991; Plants - 531; Viruses - 0; Other Eukaryotes - 9610 (source: NCBI BLink).
c36962_g1_i1	6.97E-06	0.0098144	NA	NA
c29057_g1_i2	7.44E-06	0.0102718	NA	NA
c35669_g1_i5	7.61E-06	0.0102997	AT3G58490	Encodes a long-chain base 1-phosphate (LCBP) phosphatase that is expressed in the endoplasmic reticulum.

c38259_g1_i9	8.04E-06	0.010687	AT1G70610	member of TAP subfamily
c33345_g1_i3	8.92E-06	0.011593	AT5G24270	encodes a calcium sensor that is essential for K+ nutrition, K+/Na+ selectivity, and salt tolerance. The protein is similar to calcineurin B. Lines carrying recessive mutations are hypersensitive to Na+ and Li+ stresses and is unable to grow in low K+. The growth defect is rescued by extracellular calcium.
c39490_g1_i4	9.05E-06	0.011593	AT5G18830	Encodes a member of the Squamosa Binding Protein family of transcriptional regulators. SPL7 is expressed highly in roots and appears to play a role in copper homeostasis. Mutants are hypersensitive to copper deficient conditions and display a retarded growth phenotype. SPL7 binds to the promoter of the copper responsive miRNAs miR398b and miR389c.
c33774_g1_i1	9.28E-06	0.0116748	NA	NA
c28379_g1_i3	9.60E-06	0.0118565	AT5G45020	Glutathione S-transferase family protein; LOCATED IN: cellular_component unknown; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Glutathione S-transferase, predicted (InterPro:IPR016639), Glutathione S-transferase, C-terminal (InterPro:IPR004046), Glutathione S-transferase, C-terminal-like (InterPro:IPR010987), Glutathione S-transferase/chloride channel, C-terminal (InterPro:IPR017933), Thioredoxin-like fold (InterPro:IPR012336); BEST Arabidopsis thaliana protein match is: Glutathione S-transferase family protein (TAIR:AT4G19880.1).

c40351_g2_i1	9.83E-06	0.0119354	AT5G24710	Transducin/WD40 repeat-like superfamily protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: plasma membrane; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: WD40 repeat 2 (InterPro:IPR019782), WD40 repeat, conserved site (InterPro:IPR019775), WD40 repeat (InterPro:IPR001680), WD40 repeat-like-containing domain (InterPro:IPR011046), WD40-repeat-containing domain (InterPro:IPR017986), WD40/YVTN repeat-like-containing domain (InterPro:IPR015943), WD40 repeat, subgroup (InterPro:IPR019781); Has 53337 Blast hits to 28879 proteins in 1972 species: Archae - 196; Bacteria - 12524; Metazoa - 15998; Fungi - 8175; Plants - 2336; Viruses - 1195; Other Eukaryotes - 12913 (source: NCBI BLink).
c17382_g1_i1	1.12E-05	0.0134165	NA	NA
c32392_g1_i3	1.21E-05	0.014246	AT4G08790	Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase family protein; FUNCTIONS IN: hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, nitrilase activity; INVOLVED IN: response to cadmium ion, nitrogen compound metabolic process; EXPRESSED IN: 21 plant structures; EXPRESSED DURING: 11 growth stages; CONTAINS InterPro DOMAIN/s: Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase (InterPro:IPR003010); BEST Arabidopsis thaliana protein match is: Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase family protein (TAIR:AT5G12040.1); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).

c32349_g1_i2	1.28E-05	0.0145342	AT1G35530	Encodes FANCM, a highly conserved helicase that functions as a major factor limiting meiotic crossover formation.
c30220_g1_i2	1.29E-05	0.0145342	AT5G41790	encodes a protein that physically interacts specifically with the putative coiled-coil region of COP1 in vitro. In hypocotyl and cotyledon protoplasts, it is associated to the cytoskeleton, but not in the root. expression is not regulated by light.
c39588_g1_i7	1.32E-05	0.0145342	AT1G28380	This gene is predicted to encode a protein involved in negatively regulating salicylic acid-related defense responses and cell death programs. nsl1 mutants develop necrotic lesions spontaneously and show other features of a defense response, such as higher levels of SA and disease resistance-related transcripts, in the absence of a biotic stimulus. The NSL1 protein is predicted to have a MACPF domain, found in proteins that form a transmembrane pore in mammalian immune responses. NSL1 transcript levels do not appear to change in response to biotic stresses, but are elevated by cycloheximide in seedlings, and by sodium chloride in roots.

c38069_g1_i6	1.32E-05	0.0145342	AT4G10320	tRNA synthetase class I (I, L, M and V) family protein; FUNCTIONS IN: isoleucine-tRNA ligase activity, nucleotide binding, aminoacyl-tRNA ligase activity, zinc ion binding, ATP binding; INVOLVED IN: response to cadmium ion, tRNA aminoacylation for protein translation; LOCATED IN: cytosol; EXPRESSED IN: male gametophyte, guard cell, epidermis, cultured cell, pollen tube; EXPRESSED DURING: L mature pollen stage, M germinated pollen stage; CONTAINS InterPro DOMAIN/s: Aminoacyl-tRNA synthetase, class I, conserved site (InterPro:IPR001412), Isoleucyl-tRNA synthetase (InterPro:IPR018353), Isoleucyl-tRNA synthetase, class Ia (InterPro:IPR002301), Aminoacyl-tRNA synthetase, class 1a, anticodon-binding (InterPro:IPR009080), Rossmann-like alpha/beta/alpha sandwich fold (InterPro:IPR014729), Isoleucyl-tRNA synthetase, class Ia, N-terminal (InterPro:IPR015905), Valyl/Leucyl/Isoleucyl-tRNA synthetase, class I, anticodon-binding (InterPro:IPR013155), Valyl/Leucyl/Isoleucyl-tRNA synthetase, class Ia, editing (InterPro:IPR009008), Aminoacyl-tRNA synthetase, class Ia (InterPro:IPR002300); BEST Arabidopsis thaliana protein match is: tRNA synthetase class I (I, L, M and V) family protein (TAIR:AT5G49030.3); Has 38868 Blast hits to 32849 proteins in 3074 species: Archae - 1055; Bacteria - 22228; Metazoa - 780; Fungi - 735; Plants - 304; Viruses - 0; Other Eukaryotes - 13766 (source: NCBI BLink).
c36356_g1_i1	1.47E-05	0.0159136	NA	NA
c38741_g1_i3	1.54E-05	0.0164496	AT5G57380	Encodes a plant homeodomain protein VERNALIZATION INSENSITIVE 3 (VIN3). In planta VIN3 and VRN2, VERNALIZATION 2, are part of a large protein complex that can include the polycomb group (PcG) proteins FERTILIZATION INDEPENDENT ENDOSPERM (FIE),

				CURLY LEAF (CLF), and SWINGER (SWN or EZA1). The complex has a role in establishing FLC (FLOWERING LOCUS C) repression during vernalization.
c7863_g1_i1	1.57E-05	0.0164496	NA	NA
c38130_g1_i1	1.66E-05	0.0167365	AT1G67300	Major facilitator superfamily protein; FUNCTIONS IN: carbohydrate transmembrane transporter activity, sugar:hydrogen symporter activity; INVOLVED IN: transport, transmembrane transport; LOCATED IN: integral to membrane, membrane; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Sugar transporter, conserved site (InterPro:IPR005829), Major facilitator superfamily (InterPro:IPR020846), Sugar/inositol transporter (InterPro:IPR003663), General substrate transporter (InterPro:IPR005828), Major facilitator superfamily, general substrate transporter (InterPro:IPR016196); BEST Arabidopsis thaliana protein match is: Major facilitator superfamily protein (TAIR:AT1G79820.2); Has 41411 Blast hits to 40853 proteins in 2427 species: Archae - 690; Bacteria - 24236; Metazoa - 4821; Fungi - 7159; Plants - 2674; Viruses - 2; Other Eukaryotes - 1829 (source: NCBI BLink).
c40271_g1_i1	1.67E-05	0.0167365	AT1G67110	member of CYP709A
c40392_g1_i1	1.69E-05	0.0167365	AT1G22930	T-complex protein 11; CONTAINS InterPro DOMAIN/s: T-complex 11 (InterPro:IPR008862); BEST Arabidopsis thaliana protein match is: T-complex protein 11 (TAIR:AT4G09150.1); Has

				9929 Blast hits to 7090 proteins in 817 species: Archae - 19; Bacteria - 1454; Metazoa - 4175; Fungi - 699; Plants - 382; Viruses - 14; Other Eukaryotes - 3186 (source: NCBI BLink).
c39269_g1_i5	1.73E-05	0.0167365	AT2G19600	member of Putative potassium proton antiporter family
c37003_g1_i2	1.73E-05	0.0167365	AT1G05350	NAD(P)-binding Rossmann-fold superfamily protein; FUNCTIONS IN: binding, oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor, catalytic activity, cofactor binding; INVOLVED IN: metabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 15 growth stages; CONTAINS InterPro DOMAIN/s: D-isomer specific 2-hydroxyacid dehydrogenase, NAD-binding (InterPro:IPR006140), UBA/THIF-type NAD/FAD binding fold (InterPro:IPR000594), Molybdenum cofactor biosynthesis, MoeB (InterPro:IPR009036), NAD(P)-binding domain (InterPro:IPR016040); BEST Arabidopsis thaliana protein match is: SUMO-activating enzyme 2 (TAIR:AT2G21470.1); Has 12729 Blast hits to 12531 proteins in 2437 species: Archae - 211; Bacteria - 8204; Metazoa - 1010; Fungi - 713; Plants - 367; Viruses - 0; Other Eukaryotes - 2224 (source: NCBI BLink).
c31027_g1_i1	1.74E-05	0.0167365	NA	NA

c25937_g1_i1	1.76E-05	0.0167365	AT3G09630	Ribosomal protein L4/L1 family; FUNCTIONS IN: structural constituent of ribosome; INVOLVED IN: translation; LOCATED IN: in 9 components; EXPRESSED IN: 26 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS InterPro DOMAIN/s: Ribosomal protein L4/L1e (InterPro:IPR002136), Ribosomal protein L4/L1e, eukaryotic/archaeal, conserved site (InterPro:IPR013000); BEST Arabidopsis thaliana protein match is: Ribosomal protein L4/L1 family (TAIR:AT5G02870.1); Has 1142 Blast hits to 1141 proteins in 405 species: Archae - 316; Bacteria - 17; Metazoa - 269; Fungi - 177; Plants - 117; Viruses - 0; Other Eukaryotes - 246 (source: NCBI BLink).
c29857_g1_i1	2.17E-05	0.0203602	NA	NA
c23290_g1_i1	2.31E-05	0.0213876	AT3G57490	Ribosomal protein S5 family protein; FUNCTIONS IN: structural constituent of ribosome; INVOLVED IN: translation; LOCATED IN: cytosolic small ribosomal subunit, ribosome, intracellular, membrane; EXPRESSED IN: 24 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Ribosomal protein S5, eukaryotic/archaeal (InterPro:IPR005711), Ribosomal protein S5, N-terminal (InterPro:IPR013810), Double-stranded RNA-binding-like (InterPro:IPR014720), Ribosomal protein S5, C-terminal (InterPro:IPR005324), Ribosomal protein S5 domain 2-type fold (InterPro:IPR020568), Ribosomal protein S5 (InterPro:IPR000851), Ribosomal protein S5 domain 2-type fold, subgroup (InterPro:IPR014721), Ribosomal protein S5, N-terminal, conserved site (InterPro:IPR018192); BEST Arabidopsis thaliana protein match is: Ribosomal protein S5 family protein (TAIR:AT1G59359.1); Has 8574 Blast hits to 8565 proteins in 2893 species:

				Archae - 263; Bacteria - 5125; Metazoa - 697; Fungi - 227; Plants - 154; Viruses - 0; Other Eukaryotes - 2108 (source: NCBI BLink).
c39643_g1_i4	2.40E-05	0.0219588	AT2G29080	encodes an FtsH protease that is localized to the mitochondrion
c37776_g1_i6	2.46E-05	0.0222285	AT1G30270	Arabidopsis thaliana CBL-interacting protein kinase 23. CIPK23 serves as a positive regulator of the potassium transporter AKT1 by directly phosphorylating AKT1. CIPK23 is activated by the binding of two calcineurin B-like proteins, CBL1 and CBL9.
c33396_g1_i1	2.50E-05	0.0222932	AT1G61340	Encodes a F-box protein induced by various biotic or abiotic stress.
c28553_g3_i1	2.63E-05	0.0231922	ATMG01360	cytochrome c oxidase subunit 1
c37106_g1_i4	3.10E-05	0.026922	NA	NA
c30048_g1_i1	3.75E-05	0.0321695	NA	NA
c26066_g1_i1	3.92E-05	0.0332262	AT4G27970	Encodes a protein with ten predicted transmembrane helices. The SLAH2 protein has similarity to the SLAC1 protein involved in ion homeostasis in guard cells. But, it is not expressed in guard cells and cannot complement a slac1-2 mutant suggesting that it performs a different function. SLAH2:GFP localizes to the plasma membrane.

c40459_g1_i4	4.02E-05	0.0337125	AT3G23530	Cyclopropane-fatty-acyl-phospholipid synthase; FUNCTIONS IN: cyclopropane-fatty-acyl-phospholipid synthase activity; INVOLVED IN: lipid biosynthetic process; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Amine oxidase (InterPro:IPR002937), Cyclopropane-fatty-acyl-phospholipid/mycolic acid synthase (InterPro:IPR003333), Adrenodoxin reductase (InterPro:IPR000759); BEST Arabidopsis thaliana protein match is: Cyclopropane-fatty-acyl-phospholipid synthase (TAIR:AT3G23510.1); Has 14893 Blast hits to 14869 proteins in 1968 species: Archae - 117; Bacteria - 7289; Metazoa - 156; Fungi - 477; Plants - 317; Viruses - 0; Other Eukaryotes - 6537 (source: NCBI BLink).
c29678_g1_i3	4.32E-05	0.0357558	AT2G42780	FUNCTIONS IN: molecular_function unknown; INVOLVED IN: regulation of transcription; LOCATED IN: integral to membrane, nucleus; EXPRESSED IN: 20 plant structures; EXPRESSED DURING: 11 growth stages; CONTAINS InterPro DOMAIN/s: RNA polymerase II transcription factor SIII, subunit A (InterPro:IPR010684).
c39092_g1_i6	4.73E-05	0.0385958	AT5G23050	acyl-activating enzyme 17 (AAE17); FUNCTIONS IN: catalytic activity, ligase activity; INVOLVED IN: metabolic process; LOCATED IN: cellular_component unknown; EXPRESSED IN: 22 plant structures; EXPRESSED DURING: 15 growth stages; CONTAINS InterPro DOMAIN/s: AMP-binding, conserved site (InterPro:IPR020845), AMP-dependent synthetase/ligase (InterPro:IPR000873); BEST Arabidopsis thaliana protein match is: acyl-activating enzyme 18 (TAIR:AT1G55320.1); Has 35333 Blast hits to 34131 proteins in 2444

				species: Archae - 798; Bacteria - 22429; Metazoa - 974; Fungi - 991; Plants - 531; Viruses - 0; Other Eukaryotes - 9610 (source: NCBI BLink).
c37703_g1_i4	4.77E-05	0.0385958	AT5G62650	Tic22-like family protein; CONTAINS InterPro DOMAIN/s: Tic22-like (InterPro:IPR007378); Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLink).
c17535_g1_i1	5.11E-05	0.0408882	AT1G72730	DEA(D/H)-box RNA helicase family protein; FUNCTIONS IN: helicase activity, nucleic acid binding, ATP-dependent helicase activity, ATP binding; INVOLVED IN: biological_process unknown; LOCATED IN: cytosol, plasma membrane, plant-type cell wall; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 11 growth stages; CONTAINS InterPro DOMAIN/s: RNA helicase, DEAD-box type, Q motif (InterPro:IPR014014), DNA/RNA helicase, DEAD/DEAH box type, N-terminal (InterPro:IPR011545), RNA helicase, ATP-dependent, DEAD-box, conserved site (InterPro:IPR000629), DEAD-like helicase, N-terminal (InterPro:IPR014001), DNA/RNA helicase, C-terminal (InterPro:IPR001650), Helicase, superfamily 1/2, ATP-binding domain (InterPro:IPR014021); BEST Arabidopsis thaliana protein match is: eukaryotic translation initiation factor 4A1 (TAIR:AT3G13920.1); Has 48726 Blast hits to 48144 proteins in 3104 species: Archae - 749; Bacteria - 26664; Metazoa - 6044; Fungi - 4769; Plants - 2616; Viruses - 17; Other Eukaryotes - 7867 (source: NCBI BLink).

c39022_g3_i2	5.17E-05	0.0408882	AT5G51040	unknown protein; CONTAINS InterPro DOMAIN/s: Protein of unknown function DUF339 (InterPro:IPR005631); Has 532 Blast hits to 532 proteins in 207 species: Archae - 0; Bacteria - 285; Metazoa - 16; Fungi - 41; Plants - 40; Viruses - 0; Other Eukaryotes - 150 (source: NCBI BLink).
c16125_g1_i1	5.51E-05	0.0430761	NA	NA
c37477_g1_i6	5.75E-05	0.0443326	AT5G37260	Encodes a MYB family transcription factor Circadian 1 (CIR1). Involved in circadian regulation in Arabidopsis.
c39750_g1_i1	5.79E-05	0.0443326	NA	NA
c26798_g1_i3	6.58E-05	0.0495475	AT1G09850	Arabidopsis thaliana papain-like cysteine peptidase
c33946_g1_i5	6.61E-05	0.0495475	AT1G21580	Zinc finger C-x8-C-x5-C-x3-H type family protein; FUNCTIONS IN: zinc ion binding, nucleic acid binding; INVOLVED IN: biological_process unknown; LOCATED IN: cellular_component unknown; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Zinc finger, CCCH-type (InterPro:IPR000571); Has 6412 Blast hits to 4180 proteins in 441 species: Archae - 2; Bacteria - 615; Metazoa - 1993; Fungi - 936; Plants - 1540; Viruses - 149; Other Eukaryotes - 1177 (source: NCBI BLink).

Appendix B Chapter 3 Supplementary Material

B.1 Protocol and Results for Cross (F_1) Genotyping

Methodology

The software ProSeq3 (Filatov 2009) was used to align the parental transcriptome sequence data alongside the k2g15d2 watercress transcriptome; both resources were produced during the work for Voutsina et al. (2016). Transcripts in common between the two parents and the reference transcriptome were isolated and polymorphisms analysed. A further filtering step was applied as possible to identify polymorphic loci (with at least five segregating sites) within genes of the glucosinolate (GLS) biosynthetic pathway. Nine transcripts were assessed for suitable primer and restriction enzymes (RE) combinations. The desirable RE would have a restriction site in one parent but not the other, due to the polymorphism within these transcripts, and offspring treated with the particular restriction enzyme would fragment according to the parent from whom their allele was inherited. A successful cross at the F_1 generation be heterozygote was expected to show both parental signatures. Suitable restriction enzymes were identified using Restriction Mapper Version 3 (www.restritctionmapper.org). The matching sequences were checked for introns against the closest *Arabidopsis* sequences and primers were designed using the Primer3 online tool (Untergrasser et al. 2012). The final sequence and restriction enzyme pairs used to genotype the F_1 plants are described in Table S3.1.

DNA was extracted via CTAB DNA extraction protocol with some modifications for fresh tissue (Doyle & Doyle 1987) from each F_1 plant and the parents. DNA concentrations were assessed on a Thermo Scientific NanoDrop 1000 (Thermo Fisher Scientific, U.K.) to ensure extraction success and acceptable quality. Polymerase chain reaction (PCR) amplification was completed as in Chapman *et al.* (2008) with the following conditions: 3 min denaturation at 95 ºC, then ten cycles of 30 s at 94 °C, 30 s at 65 °C, with a one degree per cycle reduction in annealing temperature, and 45 s at 72 °C. This was followed by thirty 30s cycles at 94 ºC, then 55 ºC and 45 s at 72 °C. The amplification was ended with 20 min at 72 °C. Successful amplification of the presence of desired amplification was confirmed with gel electrophoresis. Amplified DNA was RE digested with the appropriate RE at 37 °C. Gel electrophoresis was used to characterize the digestion signature of each F_1 individual in comparison with the parent lines.

Results

Of three polymorphic fragments/RE pairs tested, two were suitable for genotyping (Table S3.1). Both digests using the RE MslI, distinctly cut one parent but not the other and were employed in F_1 genotyping. Figure S3.1.a shows the results for c39658/MslI digest, where Parent B cuts into two fragments (550 and 143 bp size) with an additional third large fragment (approximately 600 bp size) which probably represents an incomplete digestion by the RE. Nevertheless, it is clearly evident that Parent A does not possess the digestion site and has not been cut. The F_1 samples contain both Parent B fragments as well as a distinct larger fragment similar to Parent A, suggesting that they contain DNA from both parents. As the Parent B digestion was not complete, leaving behind that third undigested band (600bp), we ran the c38685/MslI digest as additional confirmation (Figure S3.1.b). In this case, Parent A was digested but Parent B was not while all F_1 samples are again digested. Together these results indicate that the F_1 plants produce the signatures of both parents in two instances and must therefore contain DNA for both.

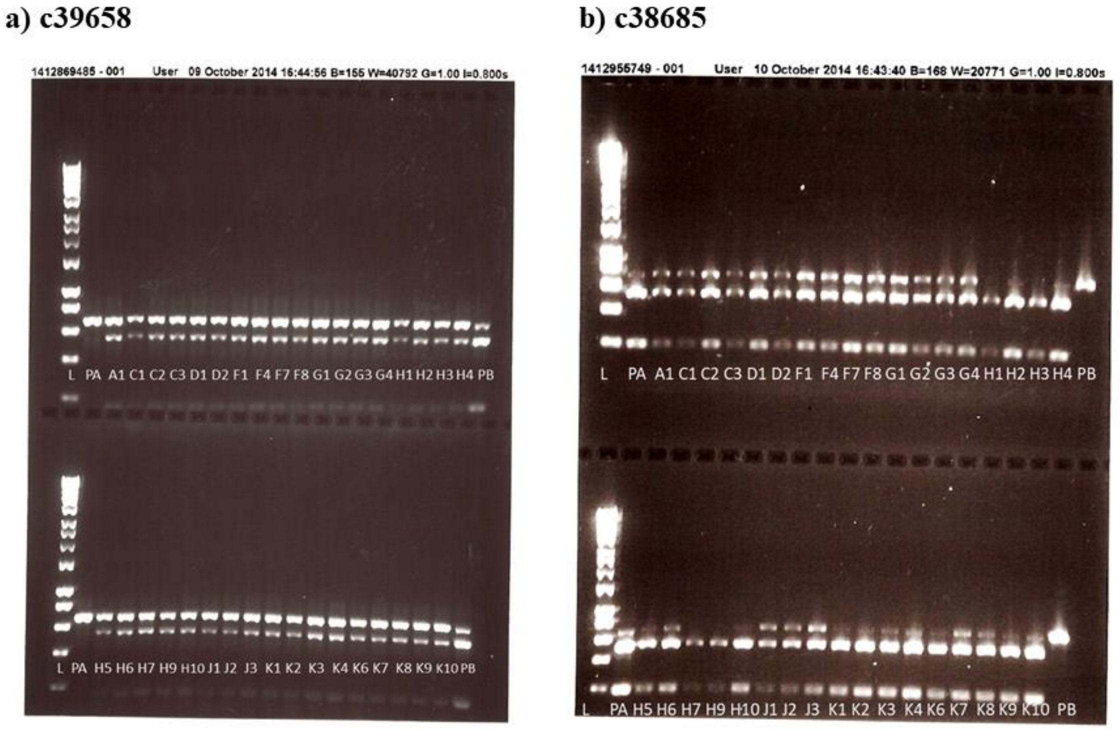
Table S3.1: Molecular tools designed to genotype F_1 offspring.

Target sequence and parent, restriction enzyme (RE), cut site and expected fragment sizes.

Target Parent	Amplified transcript	RE	Cut site	Fragment sizes
Parent B	C38685	HindiIII	562 of 1318	598, 110
Parent B	C39658	MslI	607 of 1009	552, 143
Parent A	C38685	MslI	976 of 1318	524, 184

Figure S3.1: Gel electrophoresis images of genotyping results

Figure S3.1 Genotype images confirming the presence of genetic material from both parents in the F_1 plants. Gel image a) shows locus c39658 digest by the restriction enzyme MslI. This primer/restriction enzyme pair will cut Parent B but not Parent A. Gel image b) shows locus c38685 digest by the restriction enzyme MslI. This primer/restriction enzyme pair will cut Parent A but not Parent B. The first position is annotated L and contains ladder and the F_1 samples are flanked by parents A on the left and B on the right.


a) c39658
1412869485 - 001 User 09 October 2014 16:44:56 B=155 W=40792 G=1.00 I=0.800s
L PA A1 C1 C2 C3 D1 D2 F1 F4 F7 F8 G1 G2 G3 G4 H1 H2 H3 H4 PB
L PA H5 H6 H7 H9 H10 J1 J2 J3 K1 K2 K3 K4 K6 K7 K8 K9 K10 PB
b) c38685
1412955749 - 001 User 10 October 2014 16:43:40 B=168 W=20771 G=1.00 I=0.800s
L PA A1 C1 C2 C3 D1 D2 F1 F4 F7 F8 G1 G2 G3 G4 H1 H2 H3 H4 PB
L PA H5 H6 H7 H9 H10 J1 J2 J3 K1 K2 K3 K4 K6 K7 K8 K9 K10 PB

References

Chapman, M.A. et al., 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *The Plant cell*, 20(11), pp.2931–2945.

Doyle, J.J. & Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, 19, pp.11–15.

Filatov, D.A., 2009. Processing and population genetic analysis of multigenic datasets with ProSeq3 software. *Bioinformatics*, 25(23), pp.3189–3190.

Untergrasser, A. et al., 2012. Primer3 - new capabilities and interfaces. *Nucleic Acids Research*, 40(15), p.e115.

Voutsina, N. et al., 2016. Characterization of the watercress (*Nasturtium officinale* R. Br.; Brassicaceae) transcriptome using RNASeq and identification of candidate genes for important phytonutrient traits linked to human health. BMC Genomics, 17, p.378.

B.2 Mapping Population Ploidy Assessment

Methodology

An important factor to the construction of a genetic linkage map is the definition of ploidy in the species and particular mapping population. We used fluorescent microscopy to address these queries. Chromosome imaging was attempted in collaboration with the Biomedical Imaging Unit at the UoS (David A Johnston). Traditional staining with aceto-orcein and Carmine dye were tested. Followed by fluorescent nucleic stains, DAPI (4’6-diamidino-2-phenylindole) and Sytox Orange (Thermo Fisher Scientific) on a Leica DMi8 confocal microscope at 405 nm wavelength, capturing an image every 0.33 um. Both types of imaging were completed using fresh root tips from both seedlings and cuttings, as well as a control daffodil.

Despite capturing beautiful images of chromocentres, we were unable to identify any dividing cells in watercress. Thus, buds from F2 plants with were fixed, before 11:00 am in 1:3 glacial acetic acid: absolute ethanol and incubated at 4 ºC for 24 hours. The material was washed thoroughly with, then stored and shipped in 70 % ethanol to Dr E. Wijnker (Wageningen University and Research Centre, The Netherlands) for further examination.

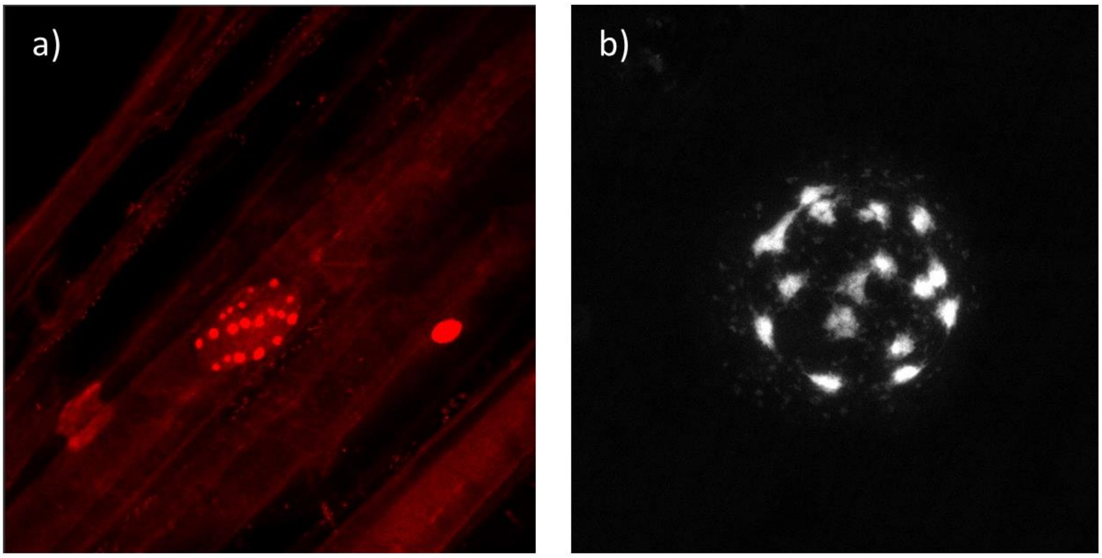
Results

Efforts to answer questions at the UoS regarding chromosome number and ploidy were not particularly conclusive. Traditional stains with aceto-orcein and carmine dye worked on daffodil root tips (control) but not watercress where root caps were not visible at all. Fluorescent imaging showed distinct parcels containing approximately 16 – 20 bright dots (Figure 3.2.a), similar to the chromocentres to chromocentres. Although interesting these images showed no obvious cell division and could not be used to answer our questions. However, we did establish that the watercress nucleus appears to be approximately 2 µm in size.

Microscopy work completed by our collaborator Dr Wijnker (Wageningen University and Research Centre) was more fruitful. Several images were produced, using the fixed flower bud material, of meiosis and the end of prophase in mitosis. Sixteen pairs of homologous chromosomes were consistently seen (Figure 3.2.b). Dr Wijnker concluded that the mapping population has 32 chromosomes and is either a diploid (2n = 32), or an allotetraploid (2n = 4x = 32) which is stable and behaved like a large diploid. With these results, we were able proceed to assemble a genetic linkage map with diploid assumptions.

Figure S3.2: Fluorescent microscopy of watercress nucleus

Figure S3.2 DAPI stained images of watercress obtained using fluorescent microscopy at a) UoS Biomedical Imaging Unit captured condensed chromocentres in root cells and b) produced by Dr E Wijnker (Wageningen University and Research Centre) shows sixteen bivalent pairs of chromosomes at the end of meiotic prophase.



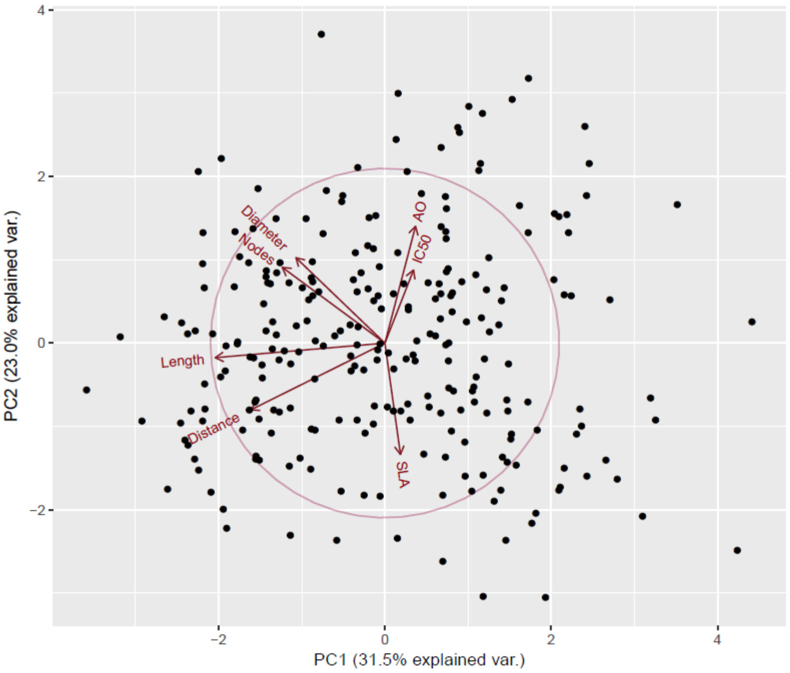
B.3 Supplementary File 3.1: Fitted survival curves for MCF-7 breast cancer cells against F_2 individuals in Control and Field trials

The survival curves fitted to data from MTS cell viability assays conducted for each F_2 sample against MCF-7 human breast cancer cells to calculate the IC_{50} maximal inhibitor concentration for each sample.

Available as Supplementary File 3.1 on the accompanying CD.

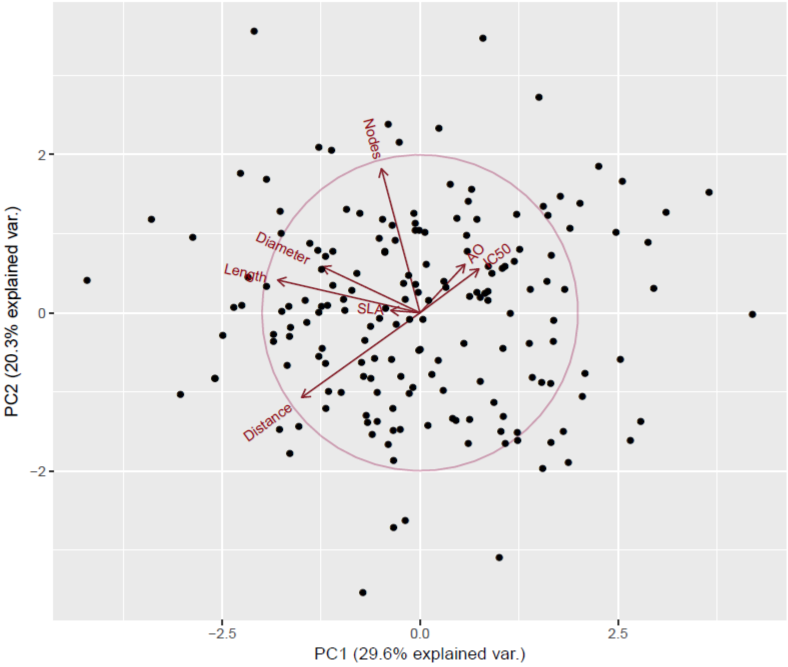
B.4 Figure S3.3: Principal component analysis (PCA) for Control environment

Figure S3.3 Principle component analysis for F2 trait data collected in the Control environment. Principle Component 1 (PC1) is primarily driven by stem length and mean internode distance, whereas PC2 by specific leaf area and antioxidant capacity.



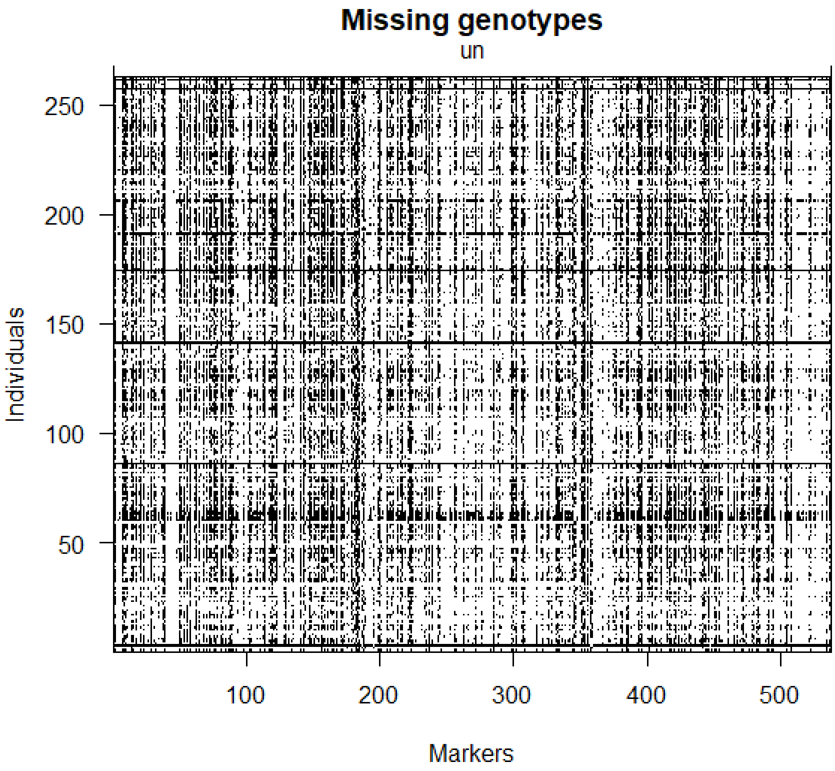
B.5 Figure S3.4: Principle component analysis (PCA) for Field environment

Figure S3.4 Principle component analysis for F_2 trait data collected in the Field environment. Principle Component 1 (PC1) is primarily driven by stem length and mean internode distance, followed by stem diameter, whereas PC2 by the number of nodes.



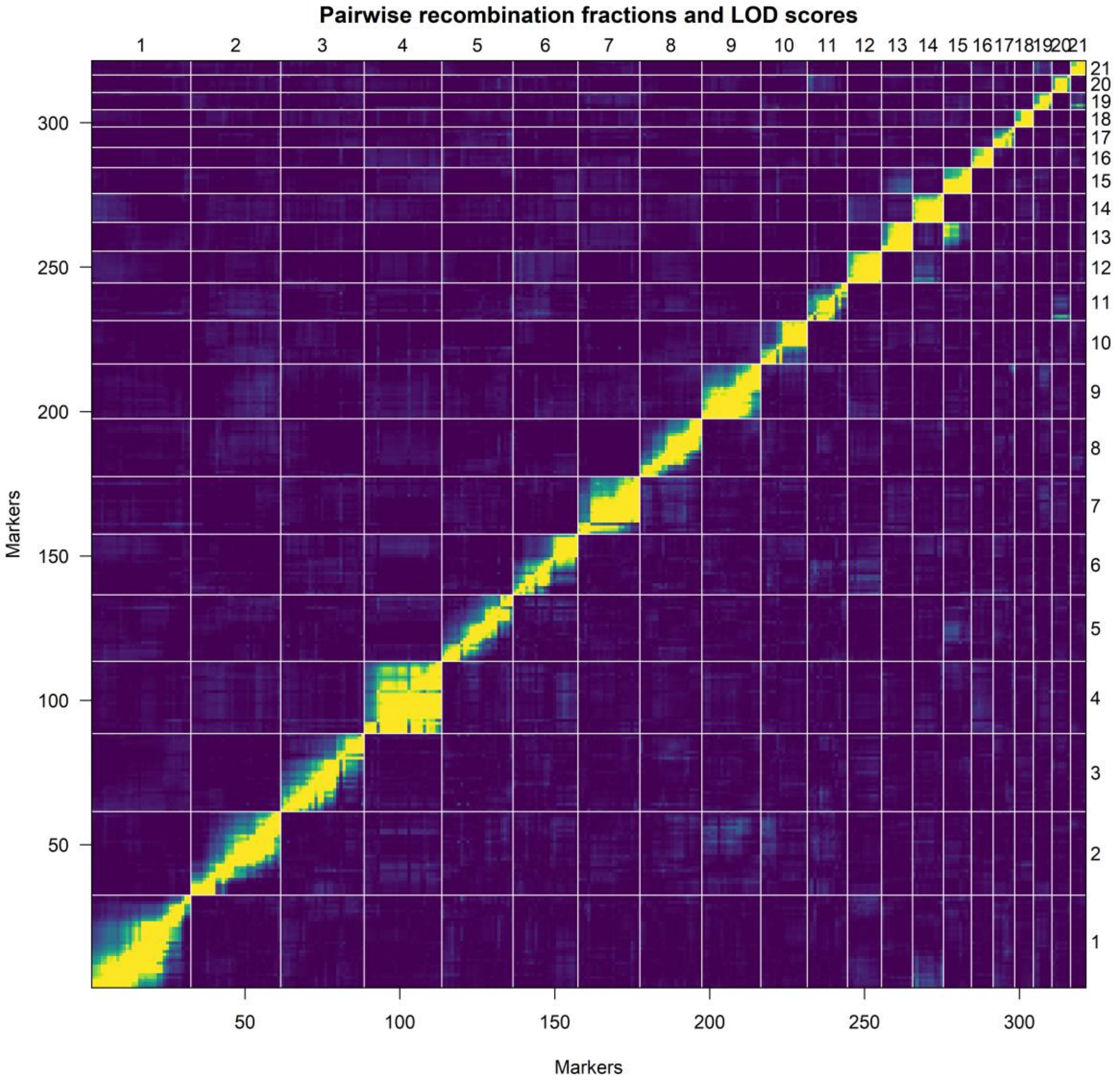
B.6 Figure S3.5: Visual representaion of individual/marker combination genotyped by Genotyping-By-Sequencing

Figure S3.5 Missing genotypes by F_2 individual and by marker in the complete polymorphic marker dataset. Missing data is visualized in black.



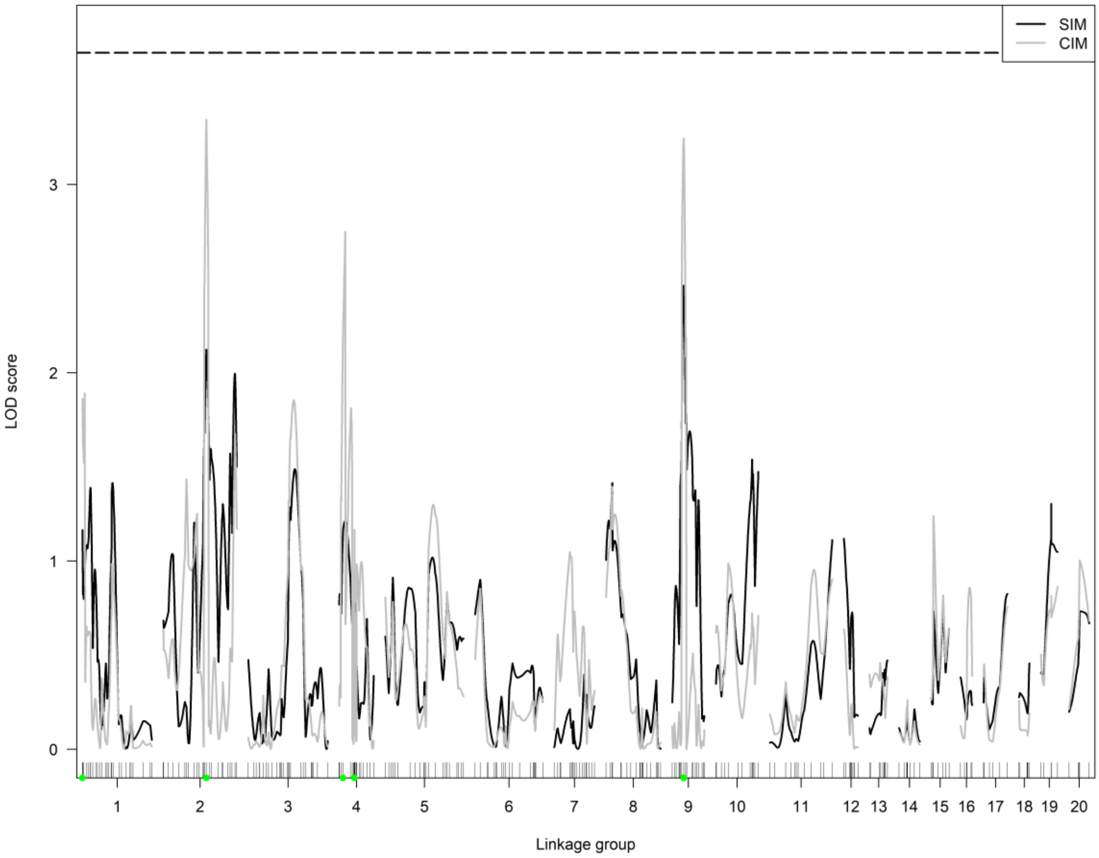
B.7 Figure S3.6: Linkage map pairwise LOD score and recombination fraction

Figure S3.6 Estimated recombination fractions (upper left triangle) and LOD scores (lower right triangle) for each pair of markers in each linkage group. Yellow indicates a high and blue a low store.



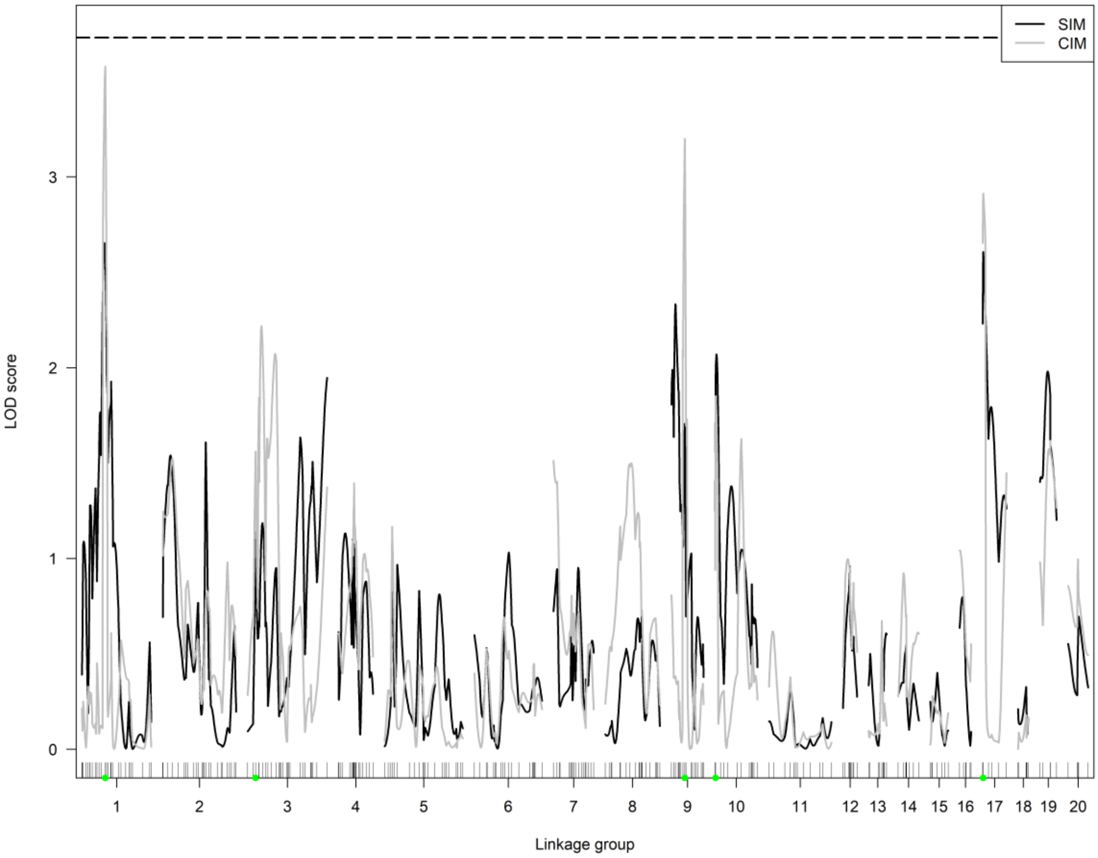
B.8 Figure S3.7: CIM model for antioxidant capacity in the Control trial

Figure S3.7 LOD scores plotted for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for antioxidant capacity (mmol Fe^{2+} equivalent per g fresh weight) in the Control trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.



B.9 Figure S3.8: CIM model for antioxidant capacity in the Field trial

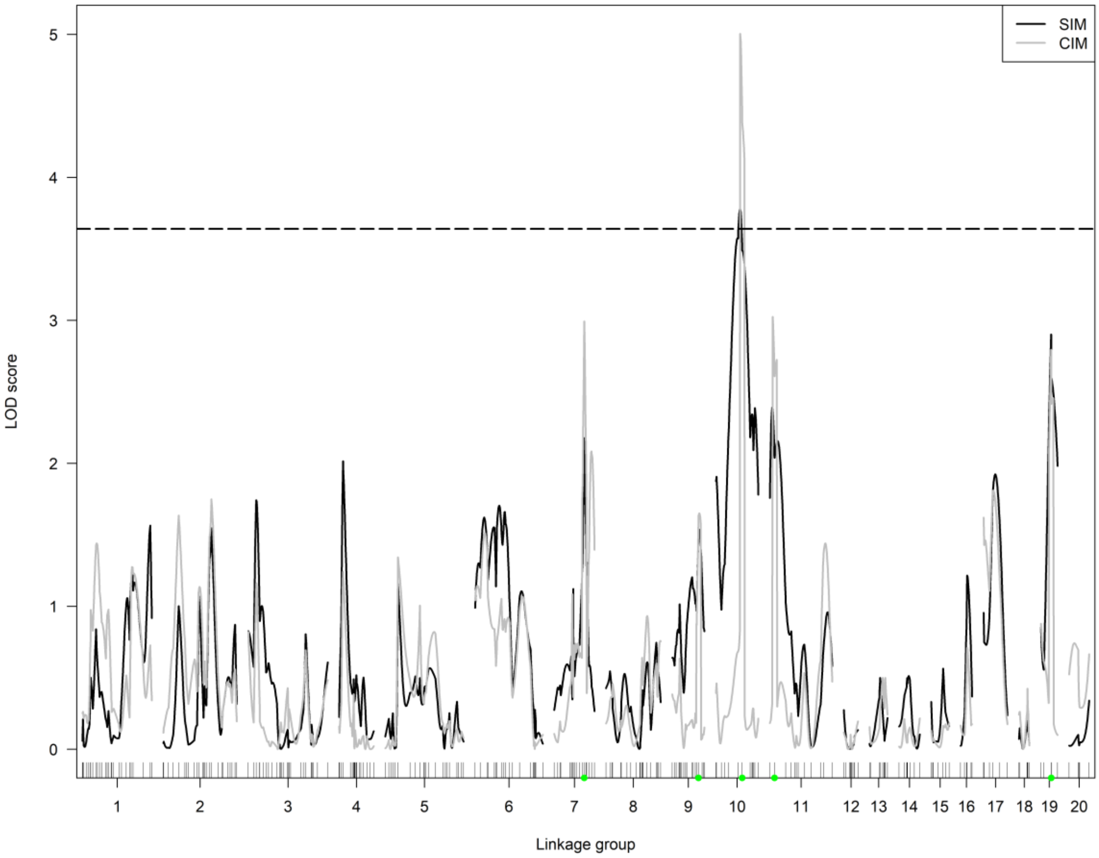
Figure S3.8 LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for antioxidant capacity (mmol Fe^{2+} equivalent/g fresh weight) in the Field trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

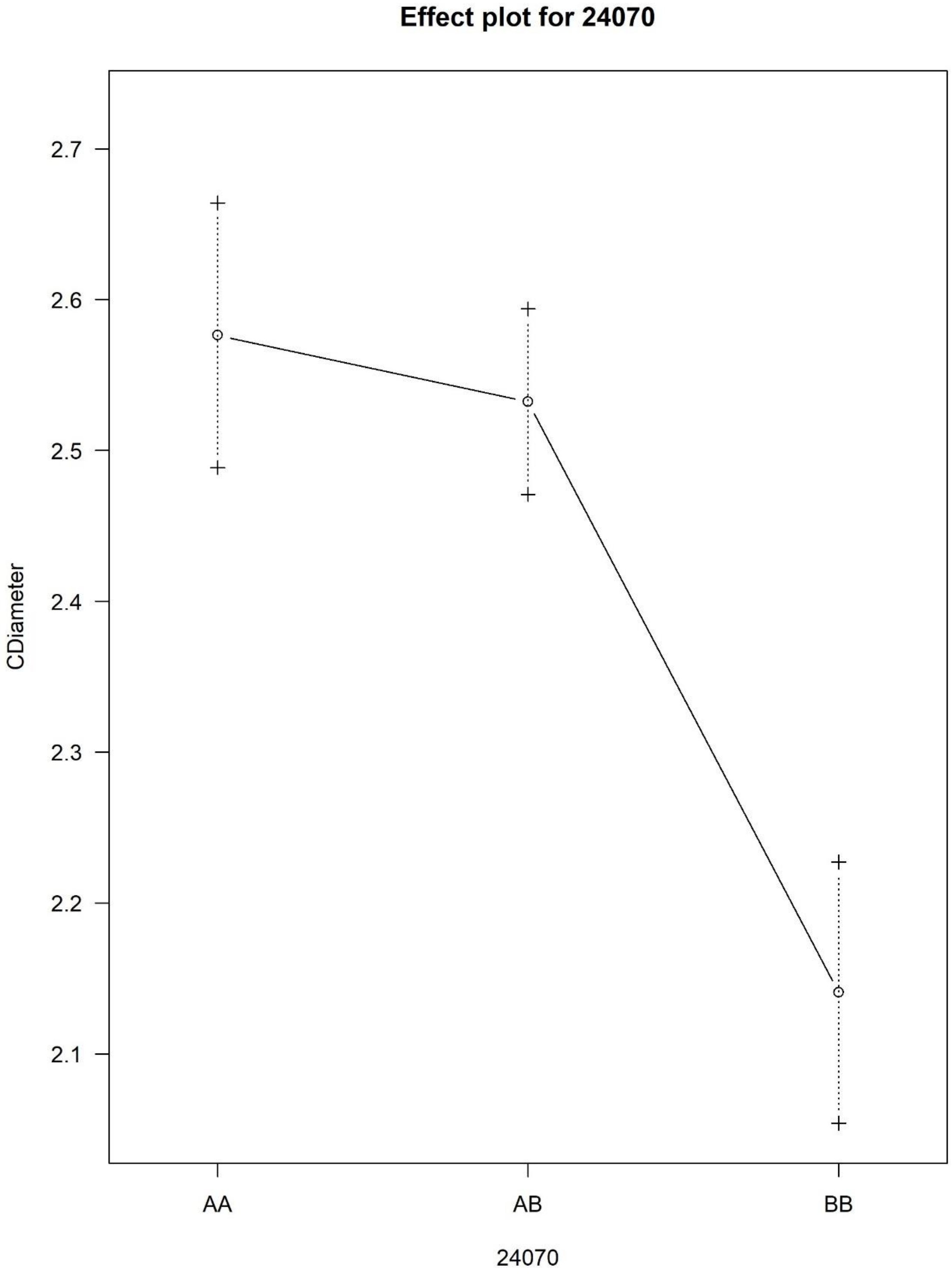


B.10 Figure S3.9: CIM model and QTL effect plot for stem diameter in the Control trial

Figure S3.9 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for stem diameter (mm) in the Control trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of stem diameter (mm) against genotype for the peak QTL on LG 10 (Marker 24070). Error bars indicate ± SE.

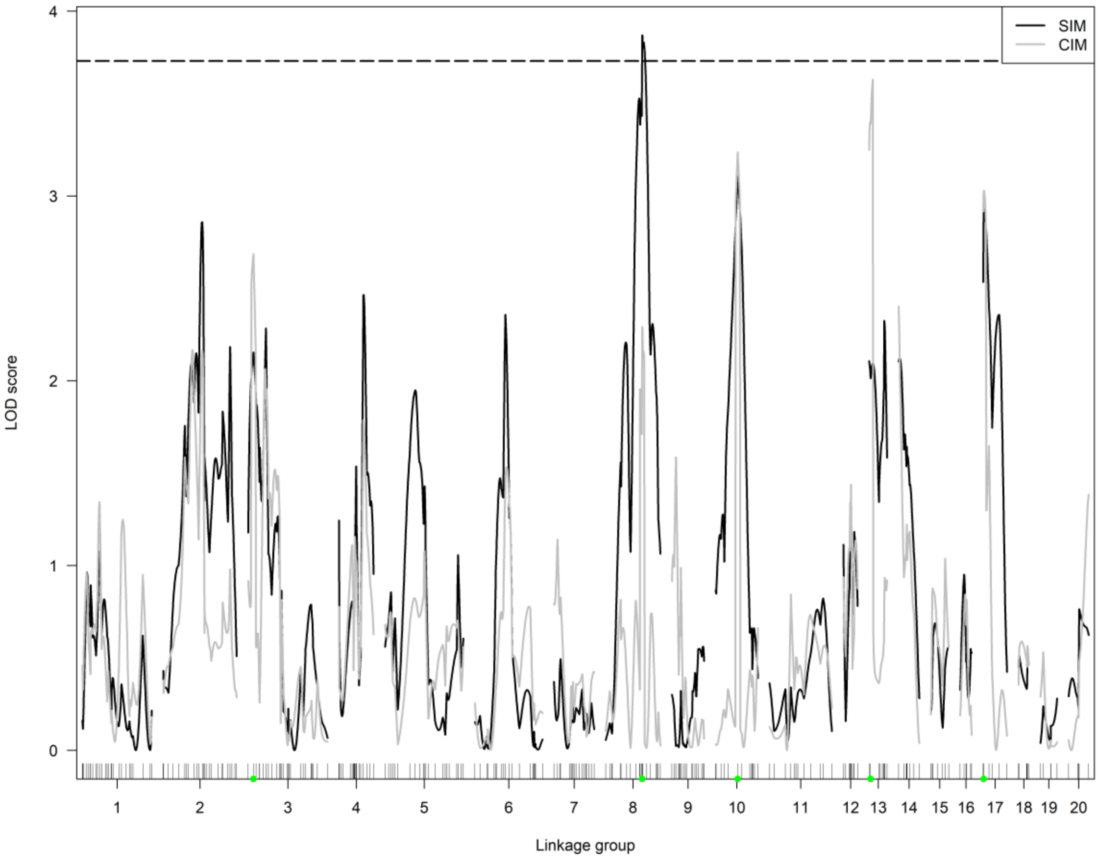


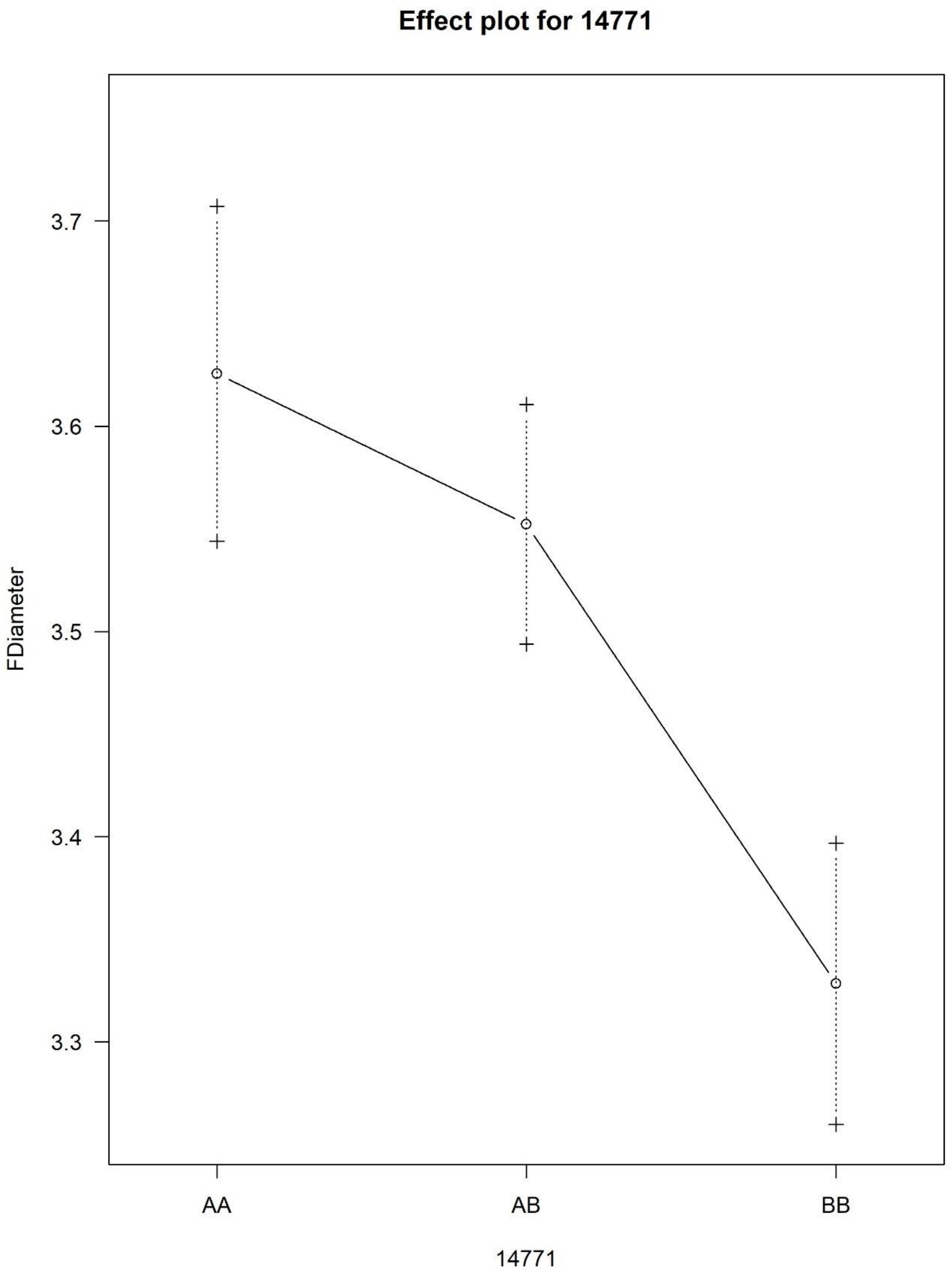

Effect plot for 24070
CDiameter
2.7
2.6
2.5
2.4
2.3
2.2
2.1
AA
AB
BB
24070

B.11 Figure S3.10: CIM model and QTL effect plot for stem diameter in the Field trial

Figure S3.10 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for stem diameter (mm) in the Field trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of stem diameter (mm) against genotype for the peak QTL on LG 13 (Marker 14771). Error bars indicate ± SE.

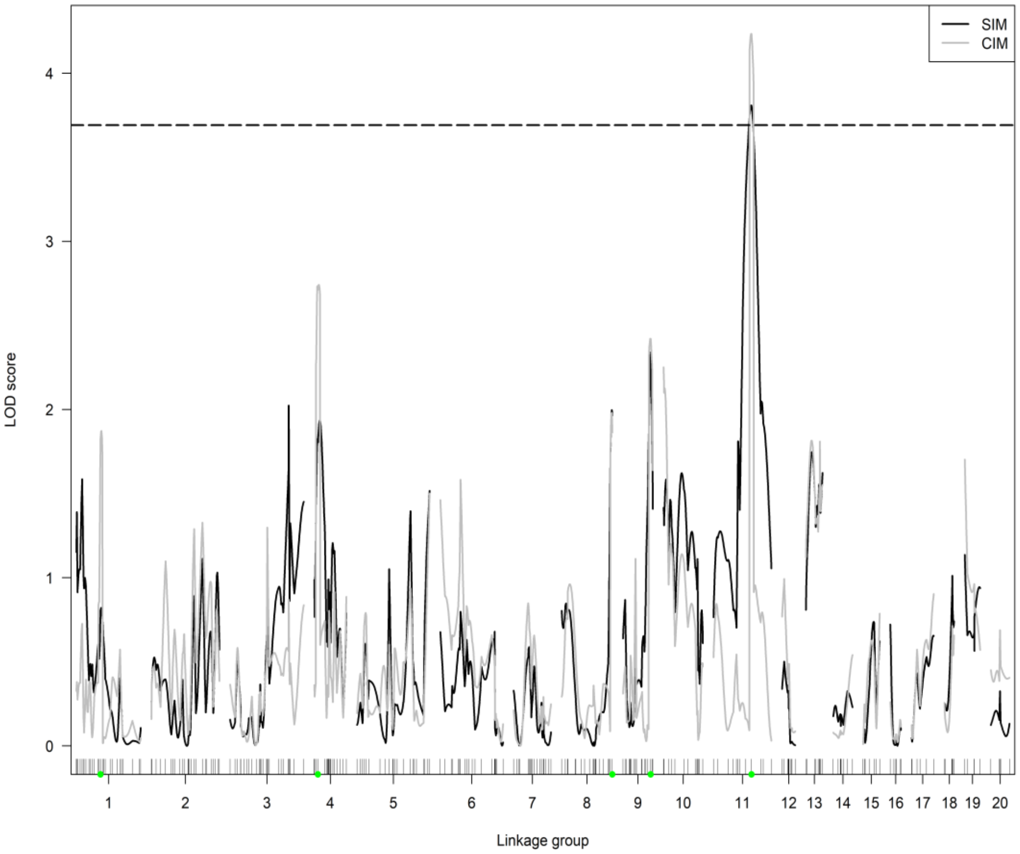


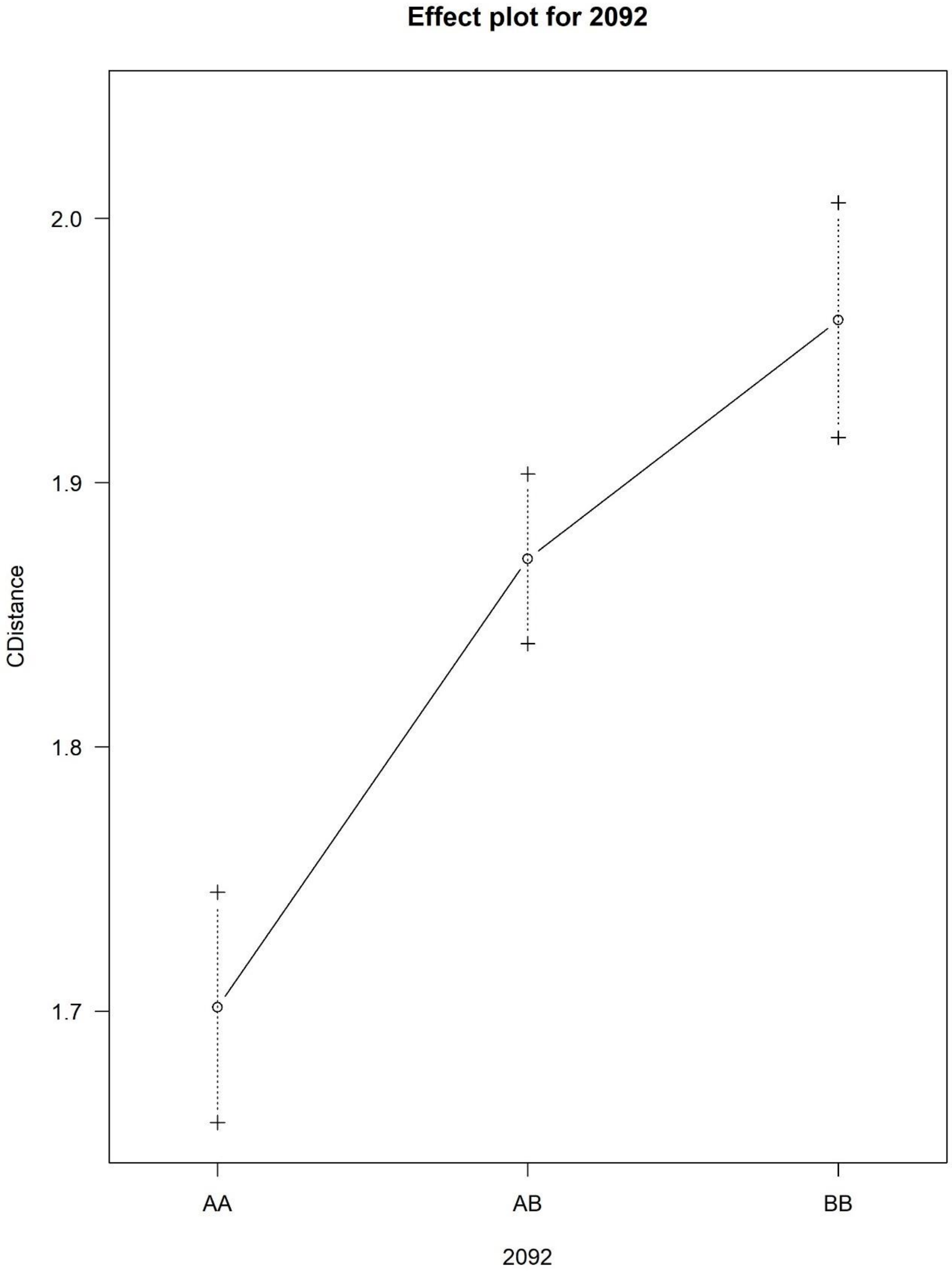

Effect plot for 14771
FDiameter
3.7
3.6
3.5
3.4
3.3
AA
AB
BB
14771

B.12 Figure S3.11: CIM model and QTL effect plot for mean internode distance in the Control trial

Figure S3.11 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for mean internode distance (cm) in the Control trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of mean internode distance (cm) against genotype for the peak QTL on LG 11 (Marker 2092). Error bars indicate ± SE.

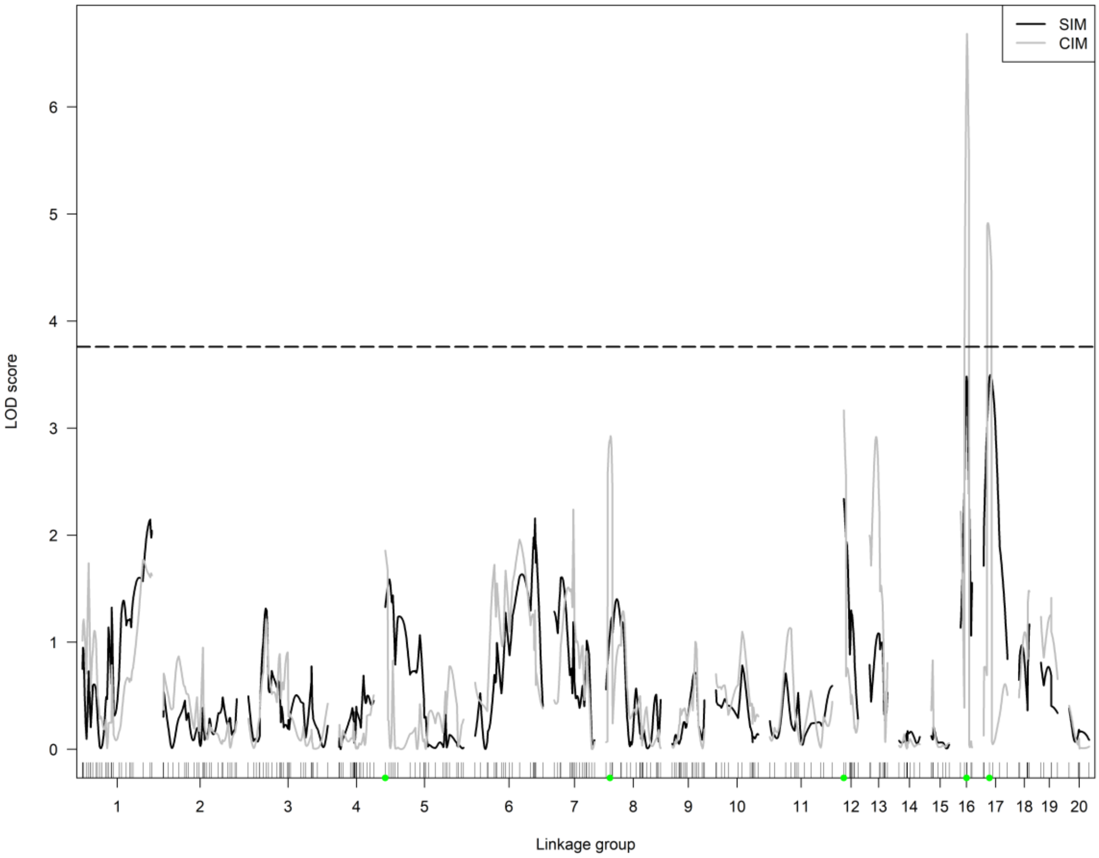


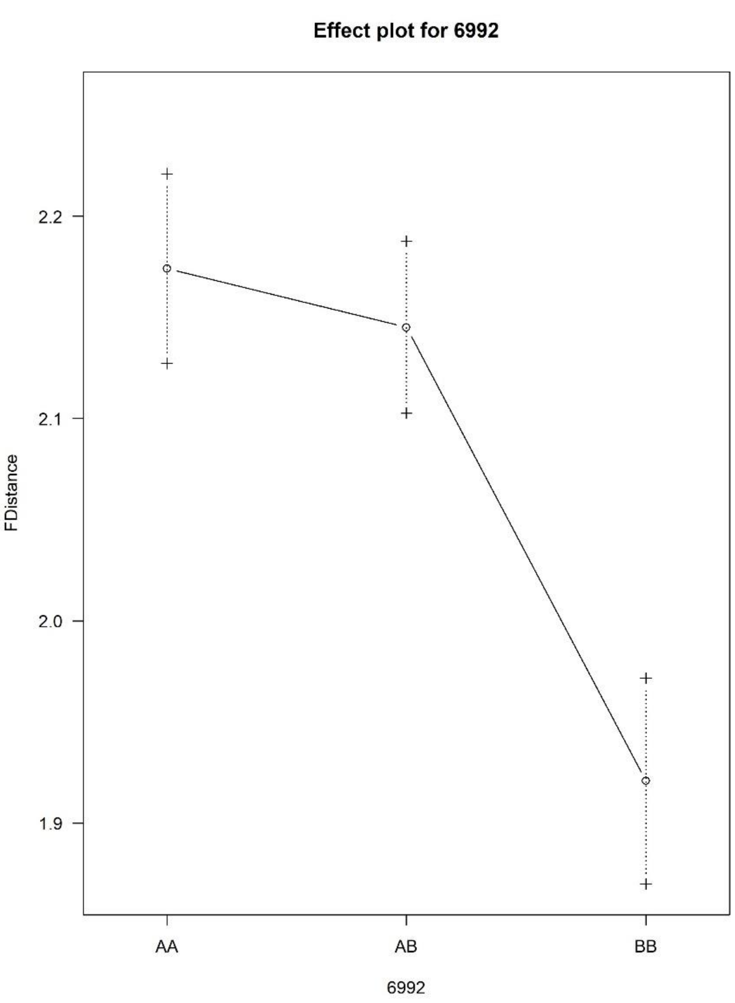

Effect plot for 2092
CDistance
2.0
1.9
1.8
1.7
AA
AB
BB
2092

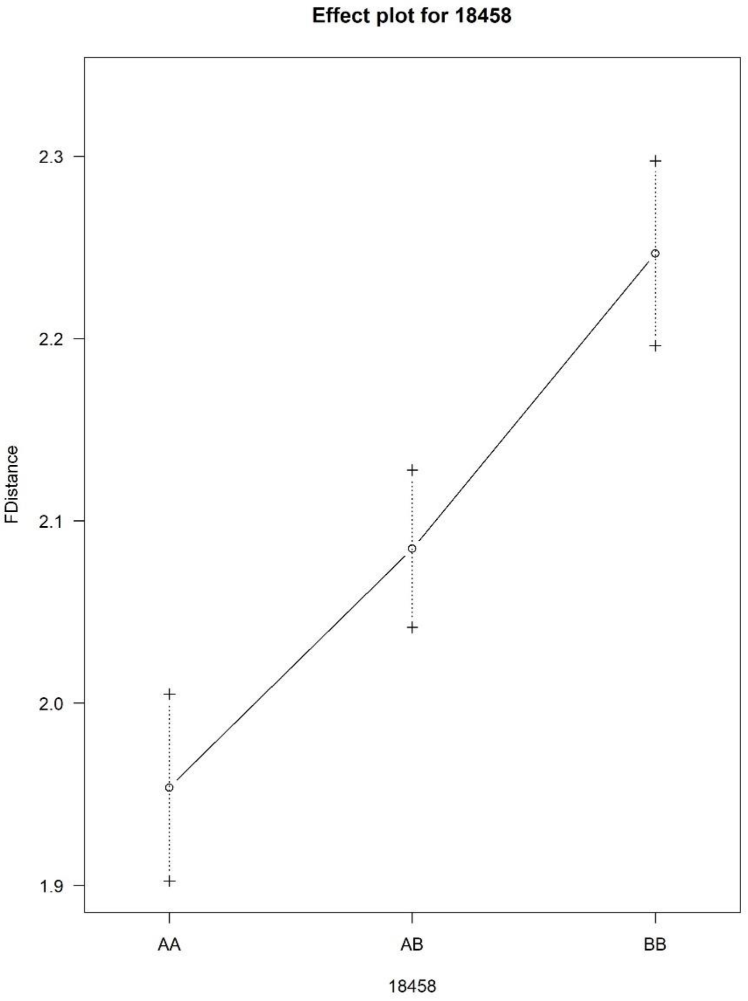
B.13 Figure S3.12: CIM model and QTL effect plots for mean internode distance in the Field trial

Figure S3.12 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for mean internode distance (cm) in the Field trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of mean internode distance (cm) against genotype for the peak QTL on LG 16 (Marker 6992) and LG 17 (Marker 18458). Error bars indicate ± SE.

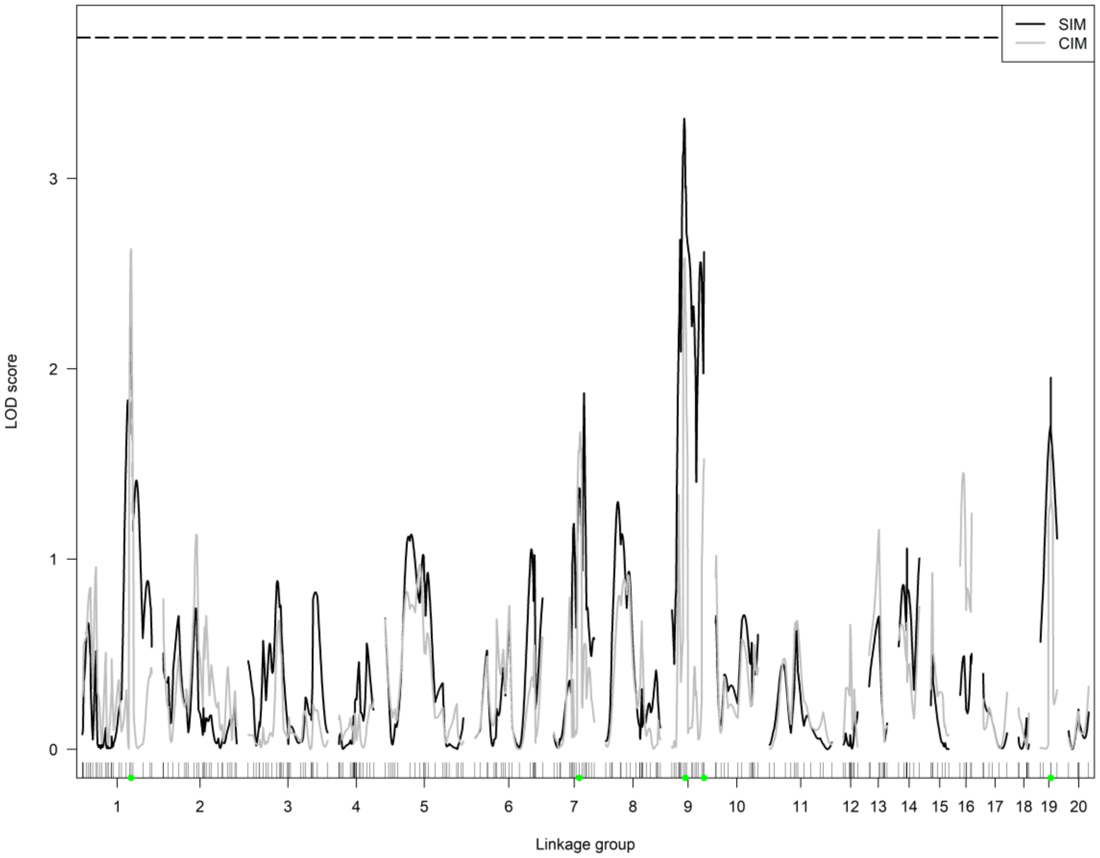



Effect plot for 6992
FDistance
2.2
2.1
2.0
1.9
AA
AB
BB
6992


Effect plot for 18458
FDistance
2.3
2.2
2.1
2.0
1.9
AA
AB
BB
18458

B.14 Figure S3.13: CIM model for IC_{50} in the Control trial

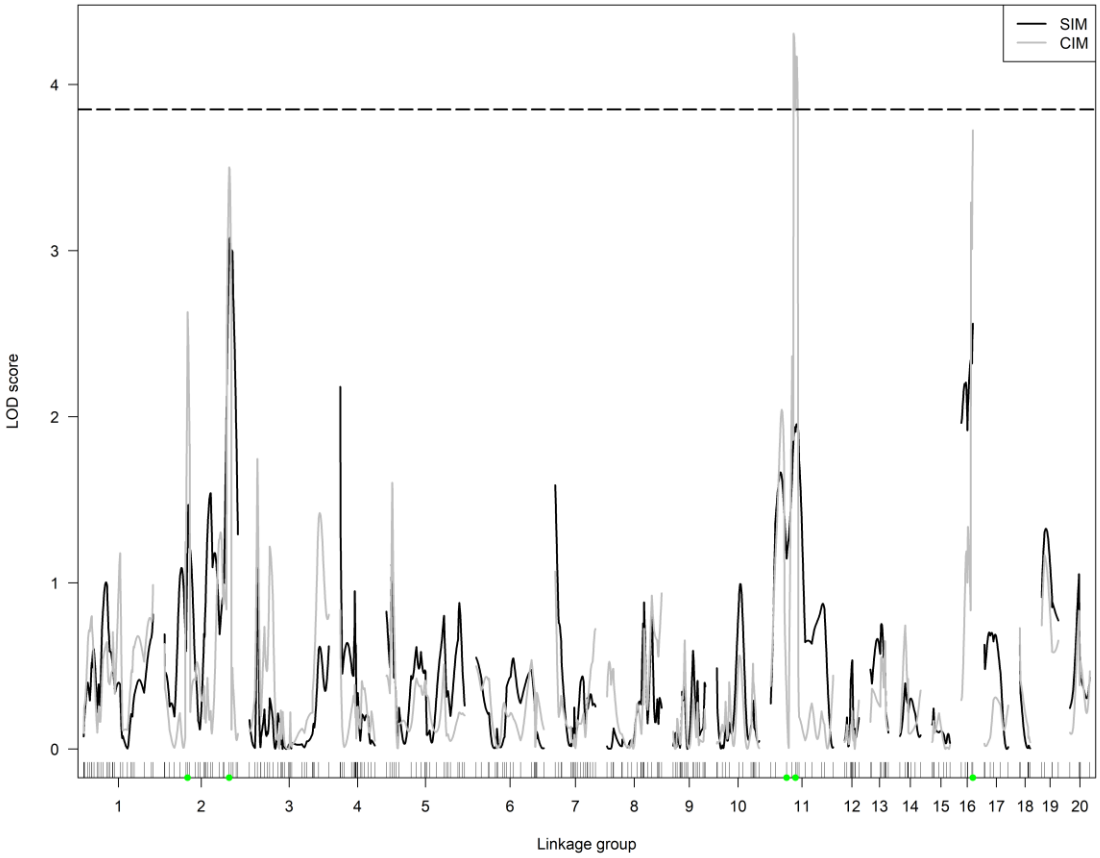
Figure S3.13 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for the square root of IC_{50} cytotoxicity of watercress extract against MCF-7 cancer cells in the Control trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.



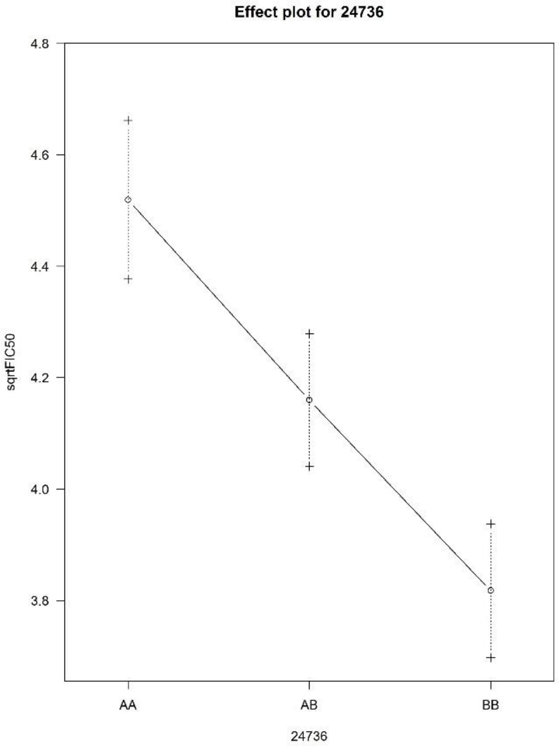
B.15 Figure S3.14: CIM model and QTL effect plots for IC_{50} in the Field trial

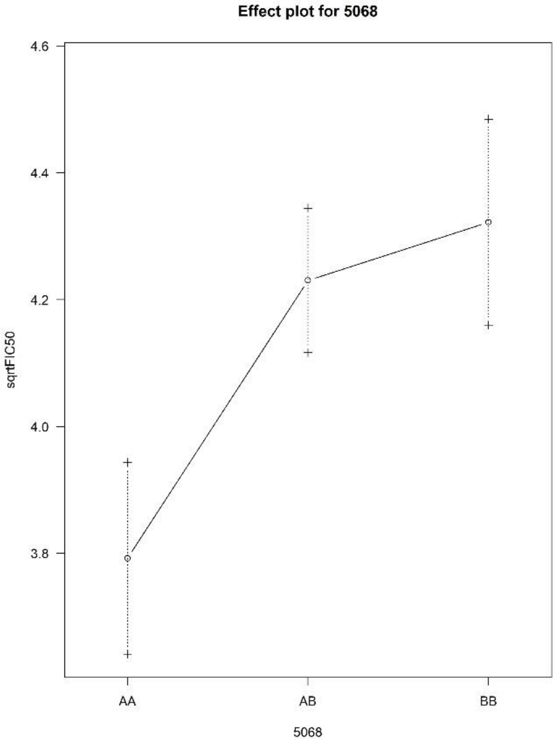
Figure S3.14 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for the square root of IC_{50} cytotoxicity of watercress extract against MCF-7 cancer cells in the Field trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

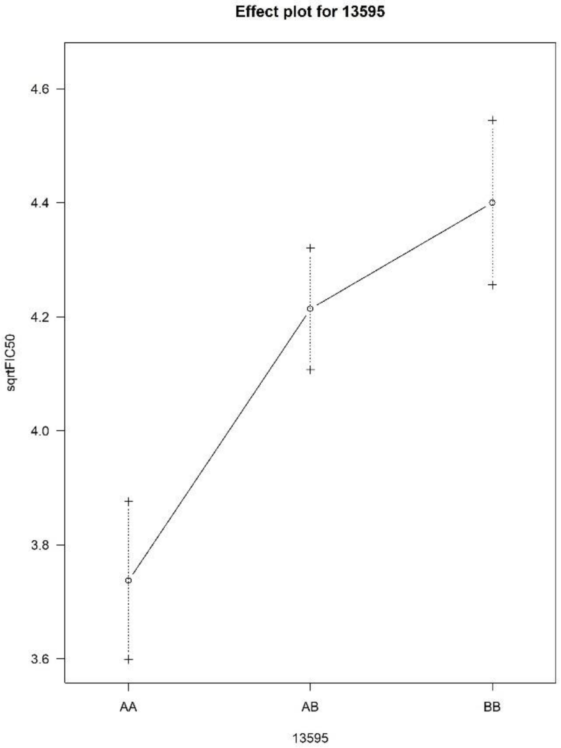
Next page: Plot of IC_{50} (µl/ml) against genotype for the peak QTL on LG 2 (Marker 14736), LG 11 (Marker 5068) and LG 16 (Marker 13595). Error bars indicate ± SE.



Appendix B


Effect plot for 24736

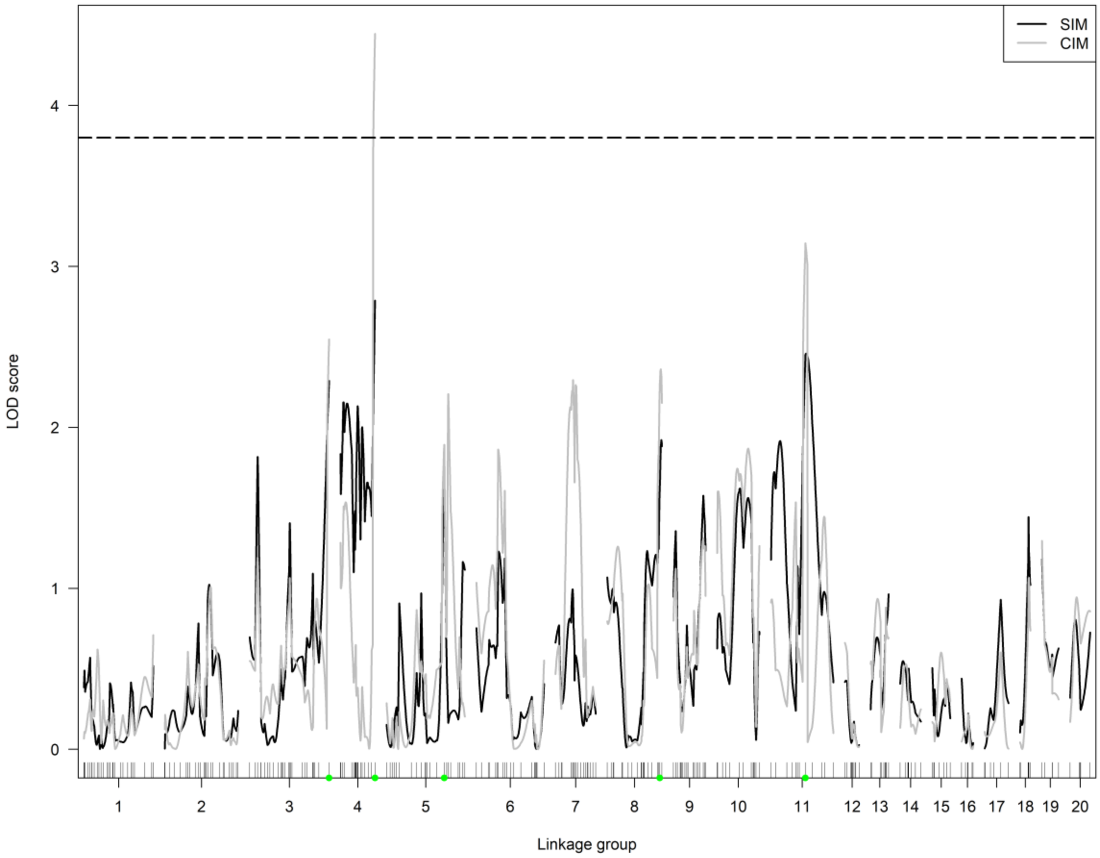

Effect plot for 5068

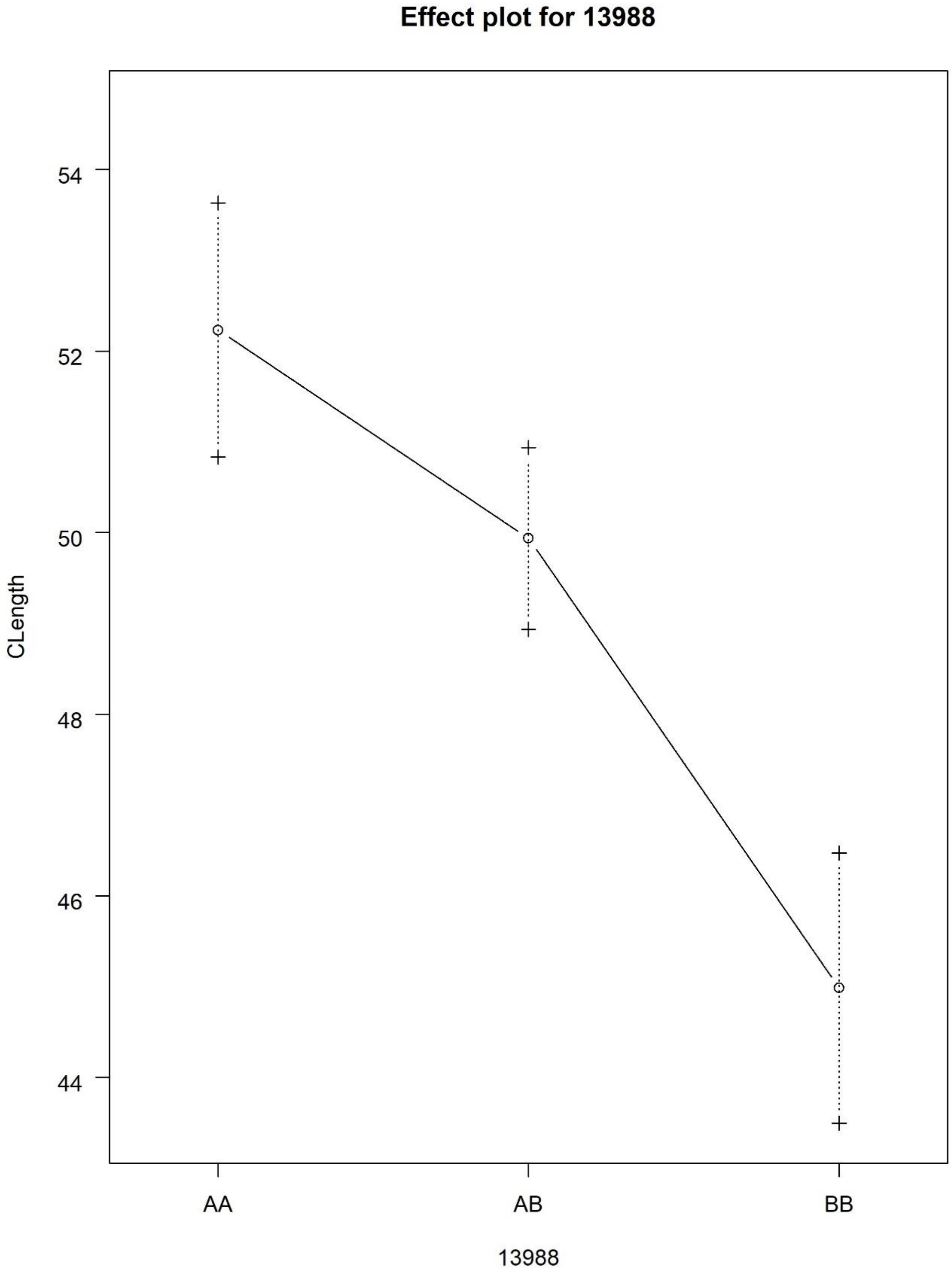

Effect plot for 13595

B.16 Figure S3.15: CIM model and QTL effect plot for stem length in the Control trial

Figure S3.15 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for stem length (cm) in the Control trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of stem length (cm) against genotype for the peak QTL on LG 4 (Marker 13988). Error bars indicate ± SE.

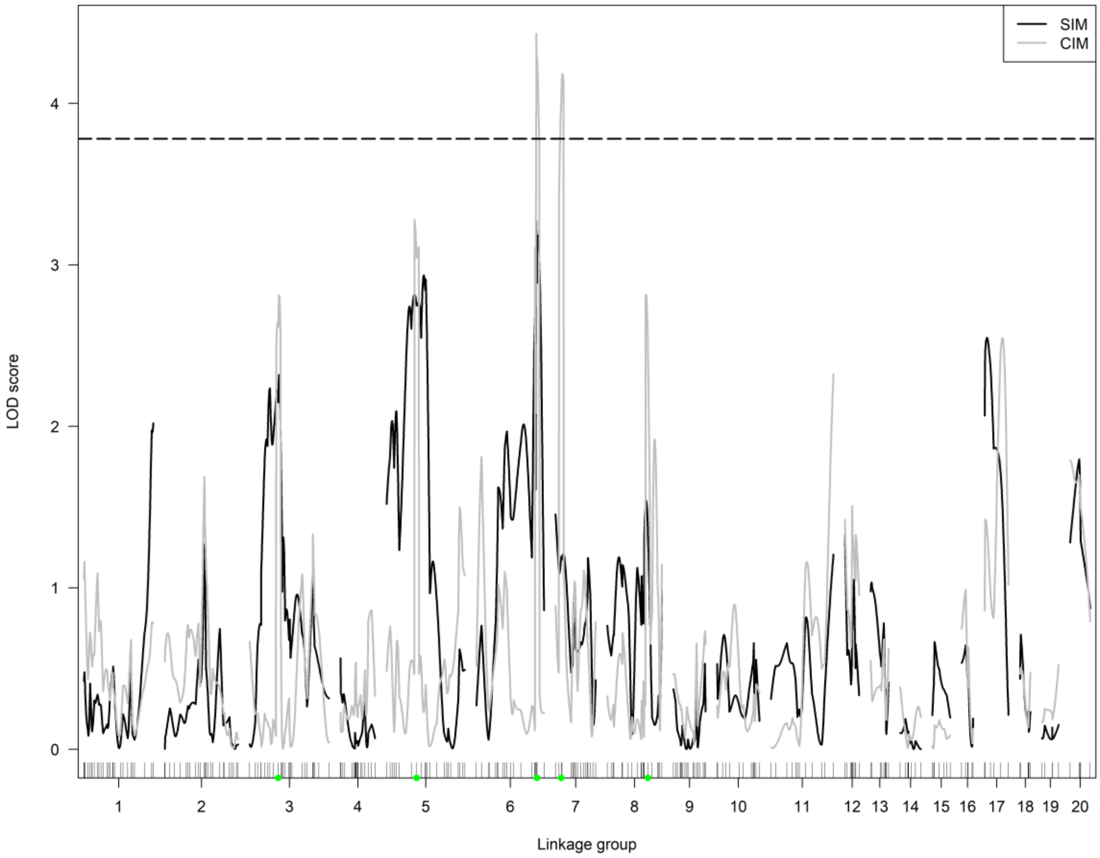



Effect plot for 13988
54
52
50
48
46
44
CLength
AA
AB
BB
13988

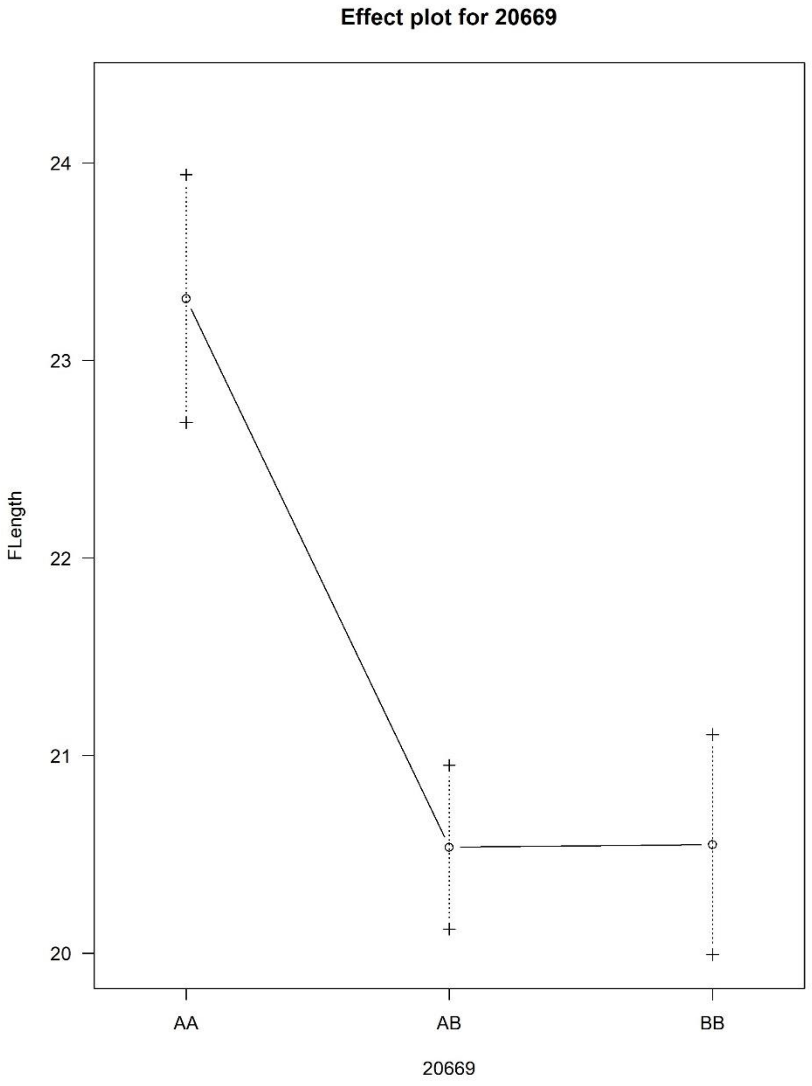
B.17 Figure S3.16: CIM model and QTL effect plots for stem length in the Field trial

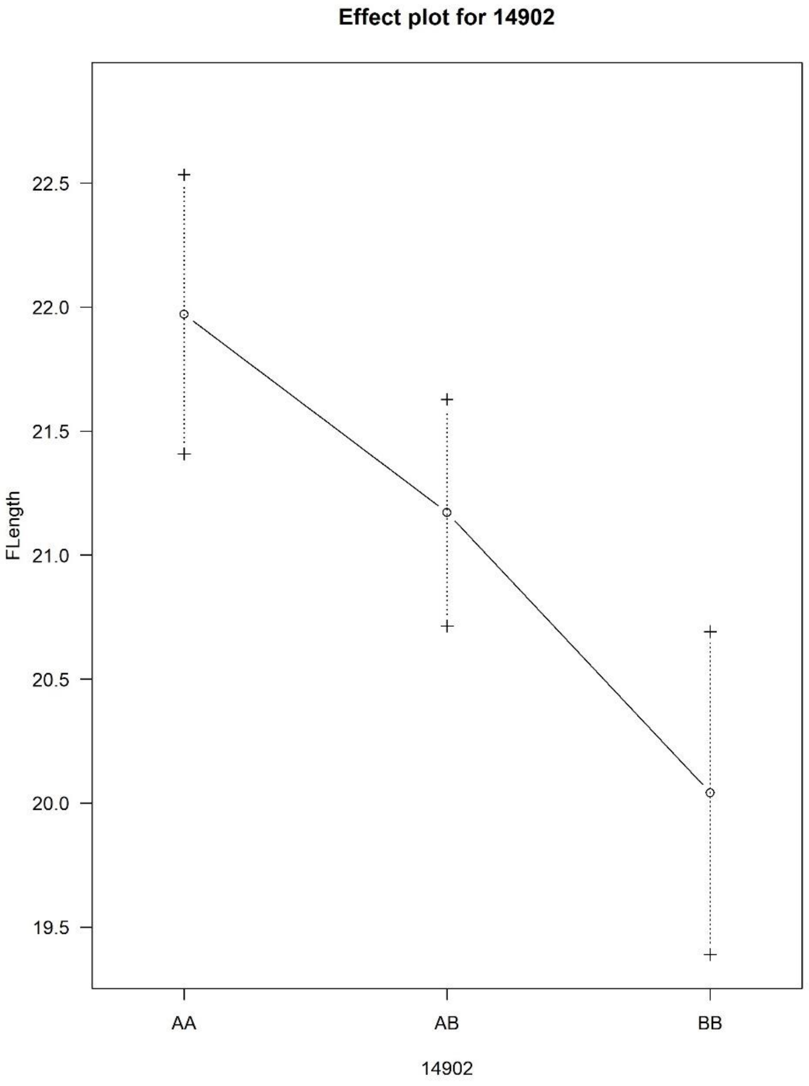
Figure S3.16 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for stem length (cm) in the Field trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of stem length (cm) against genotype for the peak QTL on LG 6 (Marker 20669) and LG 7 (Marker 14902). Error bars indicate ± SE.



Appendix B

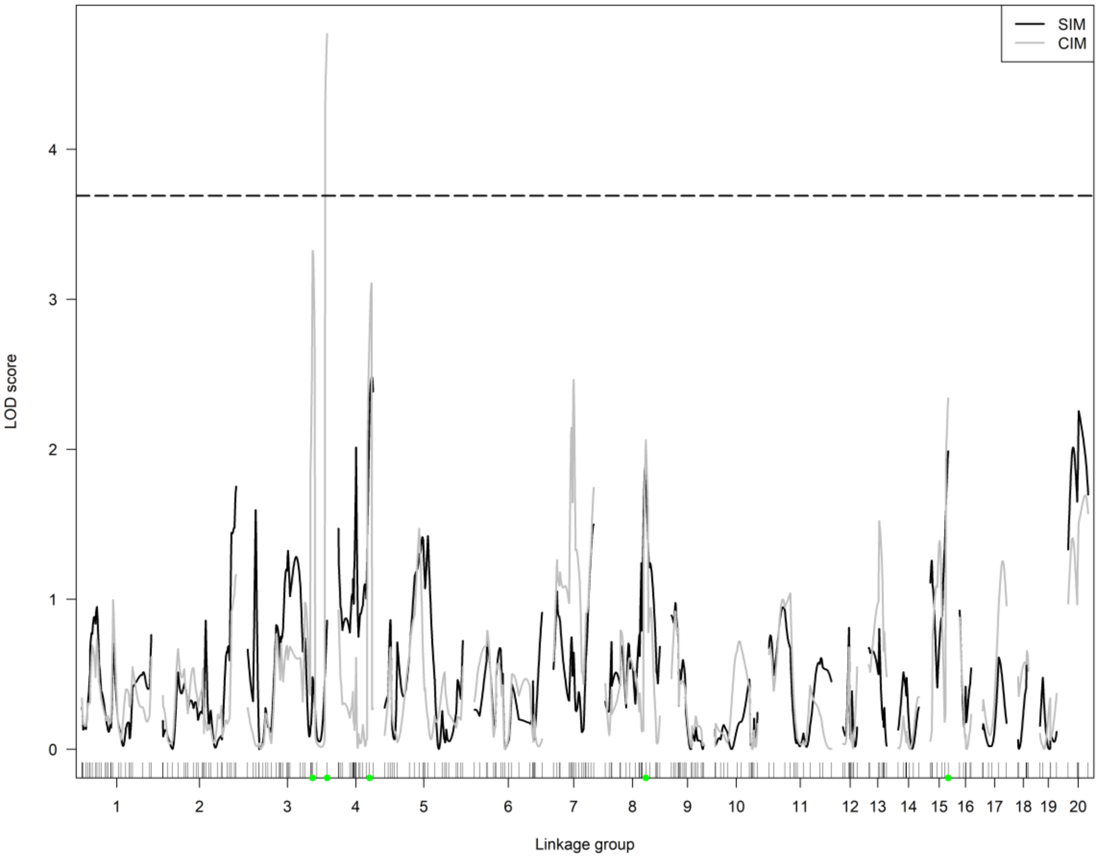

Effect plot for 20669
FLength
24
23
22
21
20
AA
AB
BB
20669

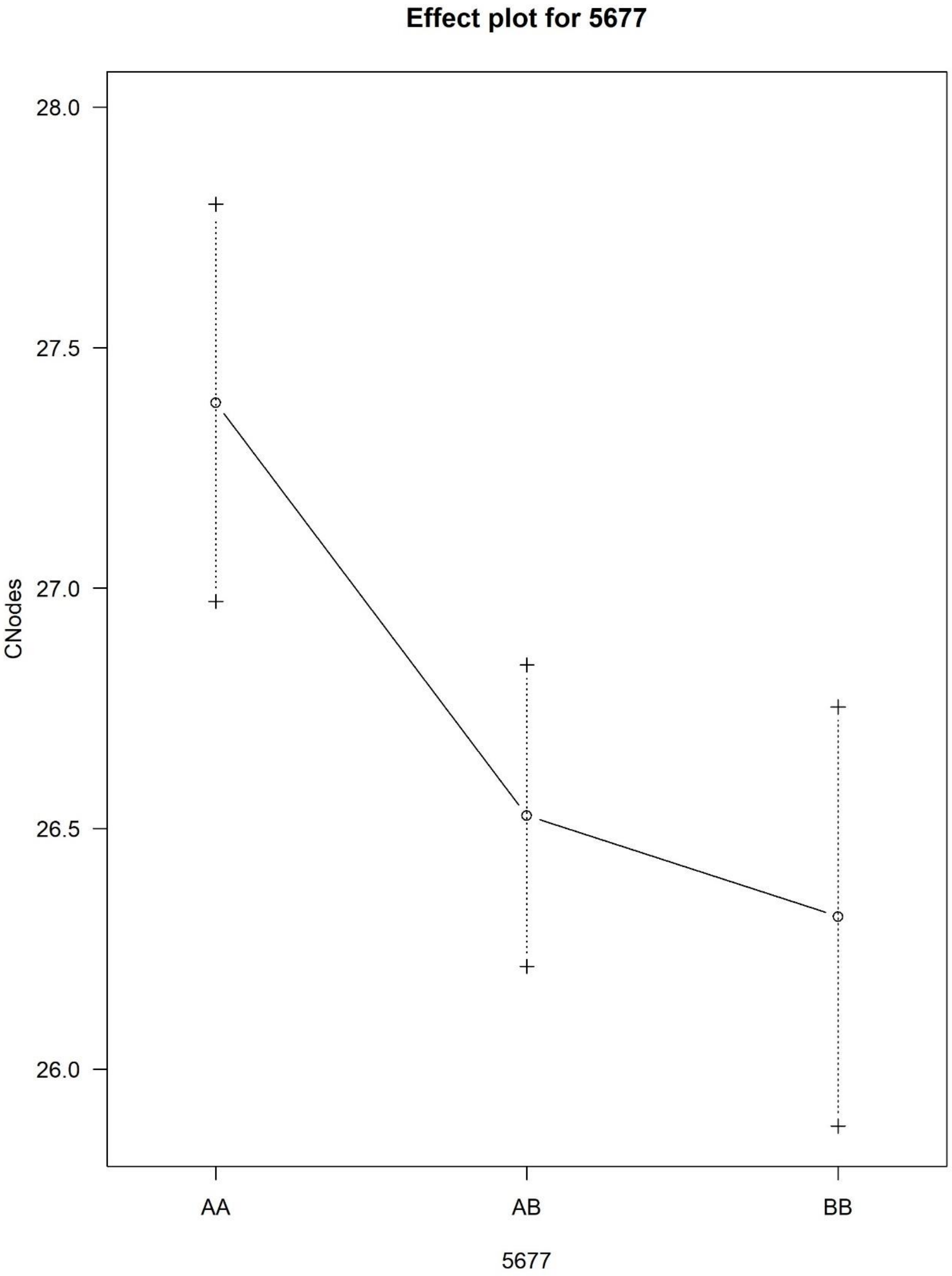

Effect plot for 14902
FLength
22.5
22.0
21.5
21.0
20.5
20.0
19.5
AA
AB
BB
14902

B.18 Figure S3.17: CIM model and QTL effect plot for number of nodes in the Control trial

Figure S3.17 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for number of nodes in the Control trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of number of nodes against genotype for the peak QTL on LG 3 (Marker 5677). Error bars indicate ± SE.

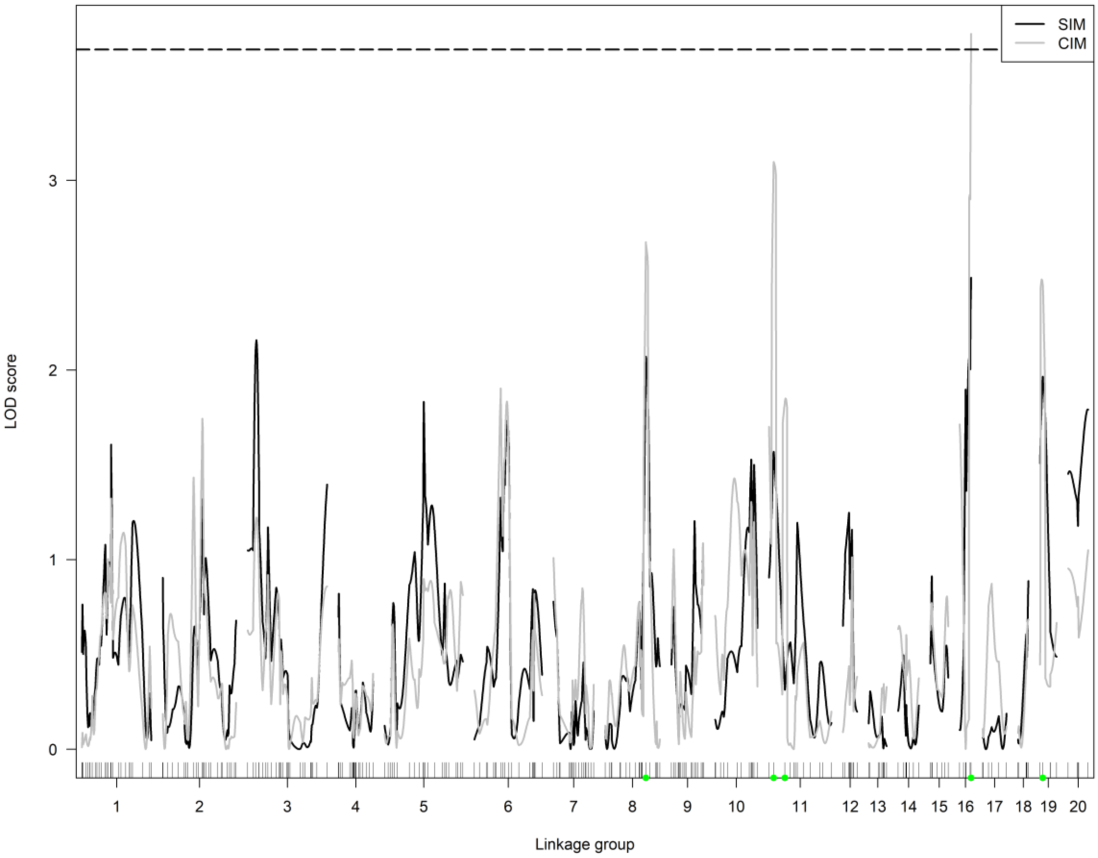


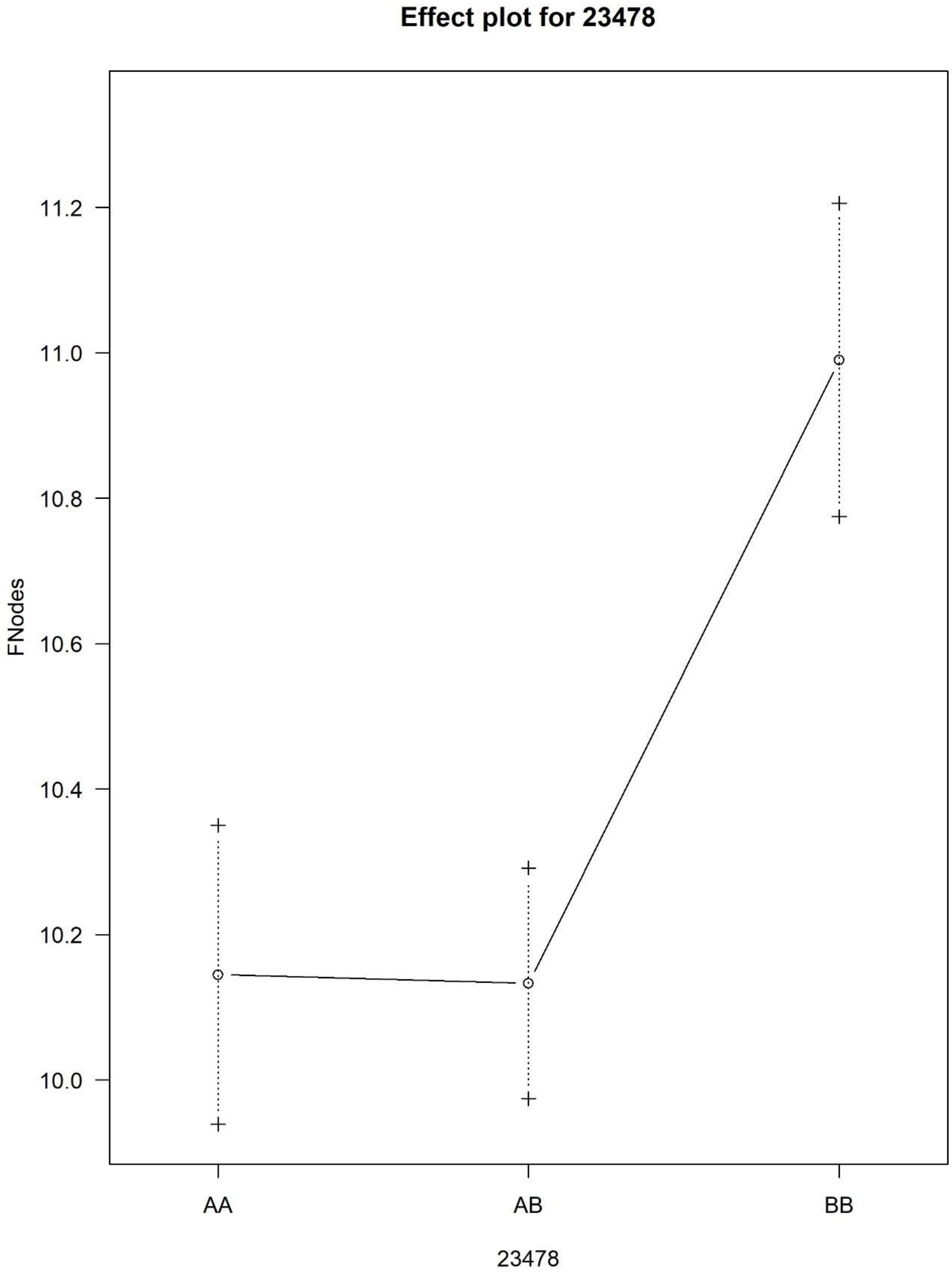

Effect plot for 5677
28.0
27.5
27.0
26.5
26.0
CNodes
AA
AB
BB
5677

B.19 Figure S3.18: CIM model and QTL effect plot for number of nodes in the Field trial

Figure S3.18 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for number of nodes in the Field trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of number of nodes against genotype for the peak QTL on LG 16 (Marker 23478). Error bars indicate ± SE.

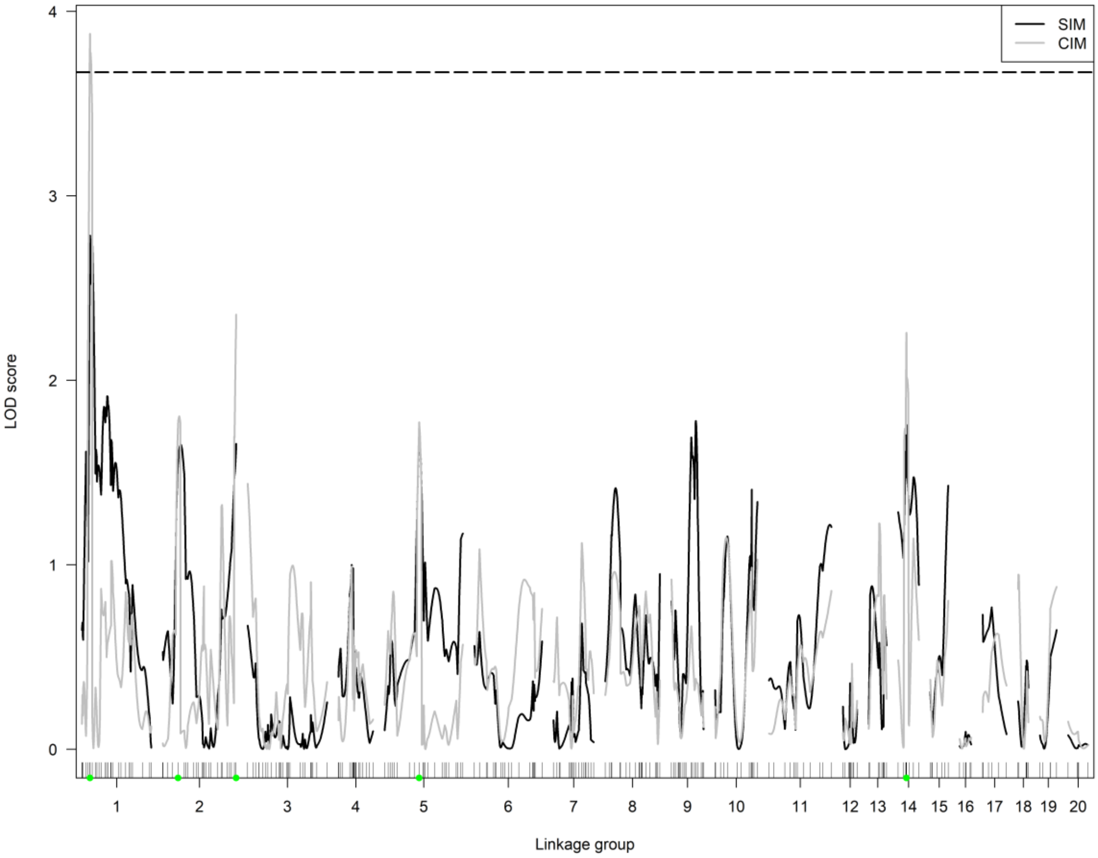


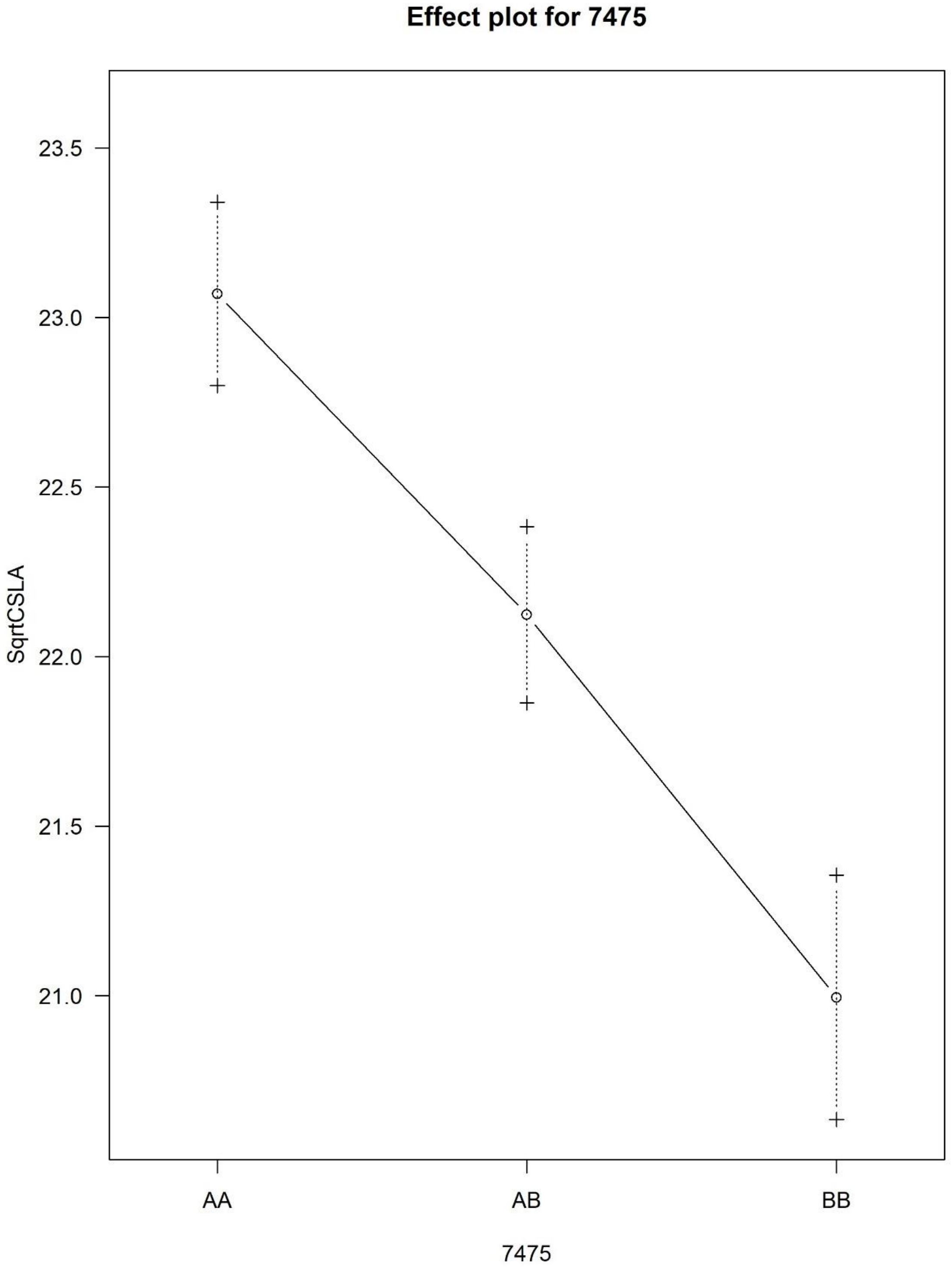

Effect plot for 23478
FNodes
11.2
11.0
10.8
10.6
10.4
10.2
10.0
AA
AB
BB
23478

B.20 Figure S3.19: CIM model and QTL effect plot for specific leaf area in the Control trial

Figure S3.19 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for the square root of specific leaf area (cm^2/g) in the Control trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of the square root of specific leaf area (cm^2/g) against genotype for the peak QTL on LG 5 (Marker 7475). Error bars indicate ± SE.

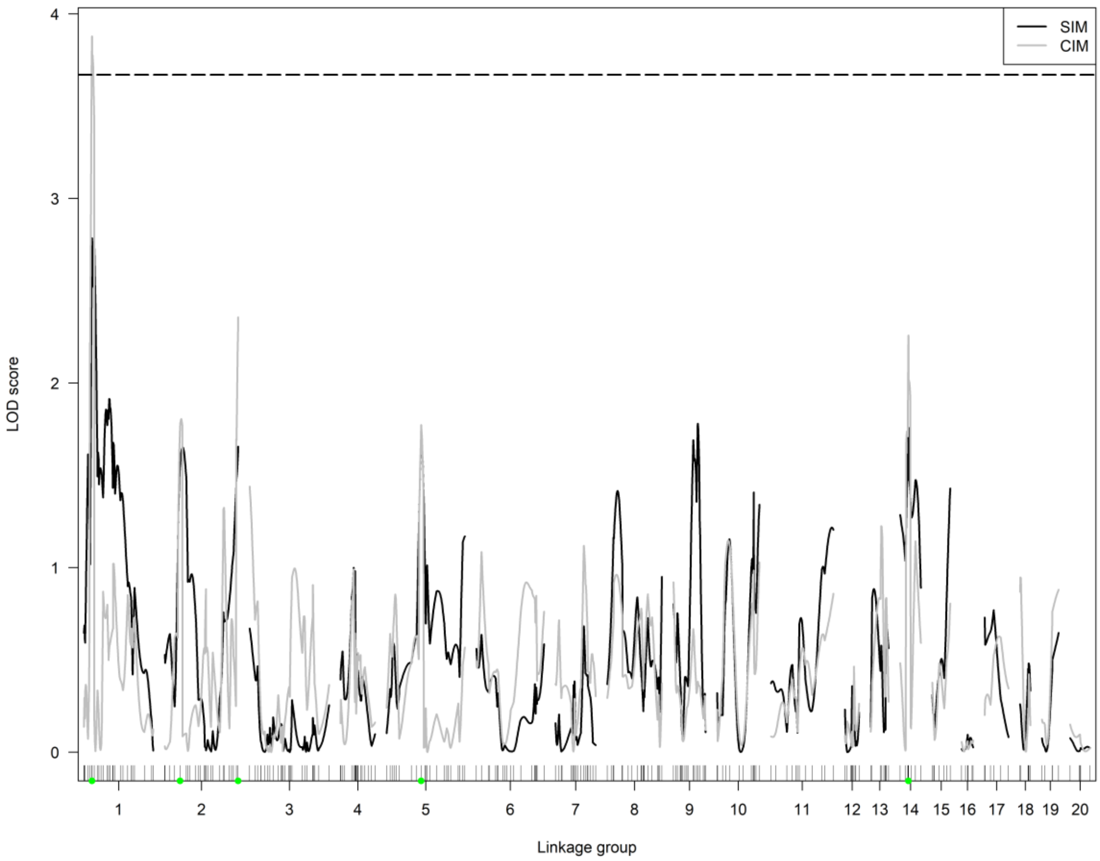


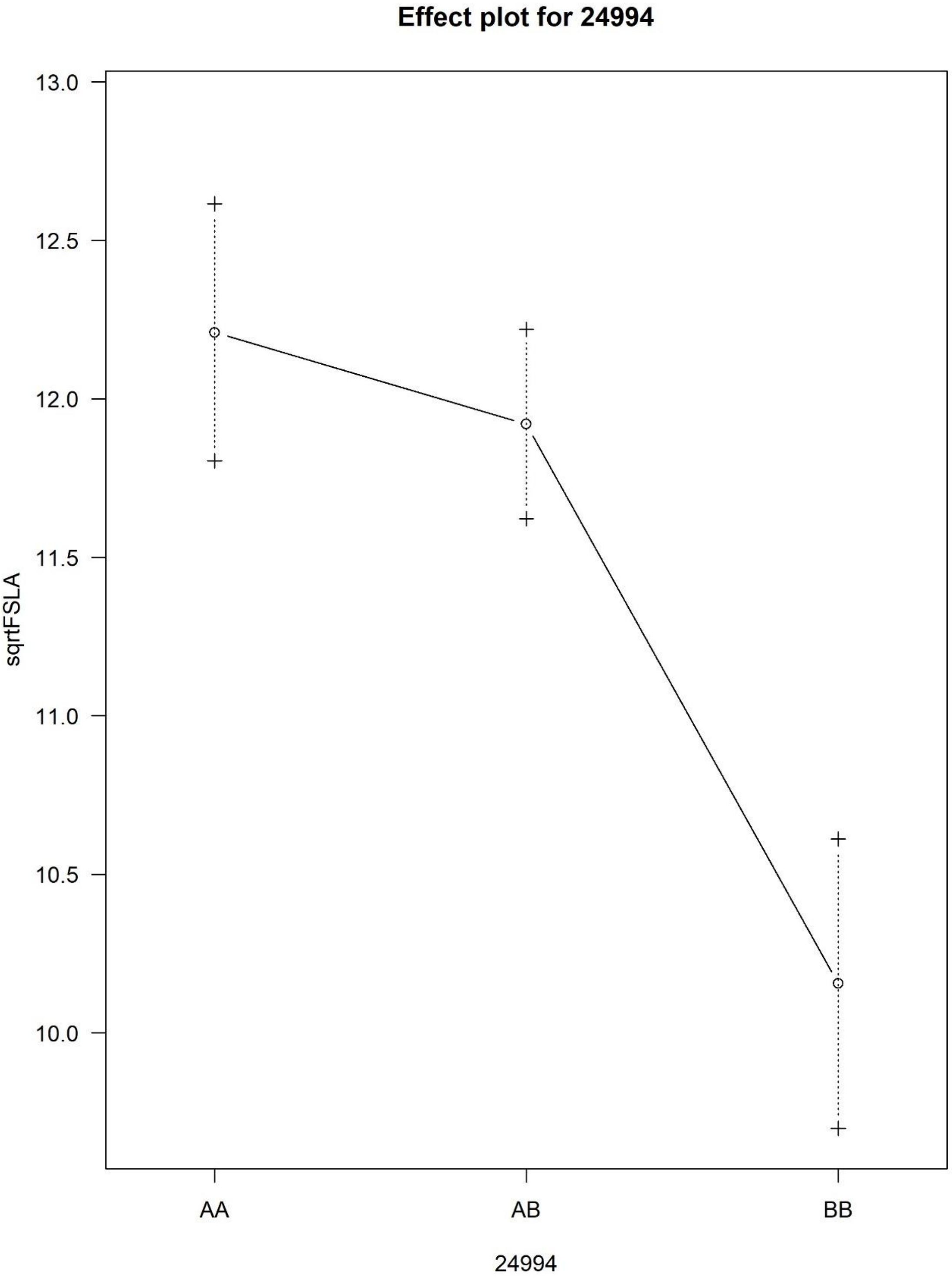

Effect plot for 7475
SqrtCSLA
23.5
23.0
22.5
22.0
21.5
21.0
AA
AB
BB
7475

B.21 Figure S3.20: CIM model and QTL effect plot for specific leaf area in the Field trial

Figure S3.20 This page: LOD scores for composite interval mapping (CIM in grey) and simple interval mapping (SIM in black) for the square root of specific leaf area (cm^2/g) in the Field trial. Markers used as cofactors are indicated in green and the 5 % LOD significance threshold by the dotted line.

Next page: Plot of square root of specific leaf area (cm^2/g) against genotype for the peak QTL on LG 1 (Marker 24994). Error bars indicate ± SE.




Effect plot for 24994
13.0
12.5
12.0
11.5
11.0
10.5
10.0
sqrtFSLA
AA
AB
BB
24994

B.22 Table S3.2: BLASTn annotation of markers underlying QTL detected in this study

Table S3.2 Annotation results for every marker within the 1.5 LOD support interval of a QTL detected in this study based on BLASTn search against the watercress transcriptome (Chapter 2) and the broader NCBI database.

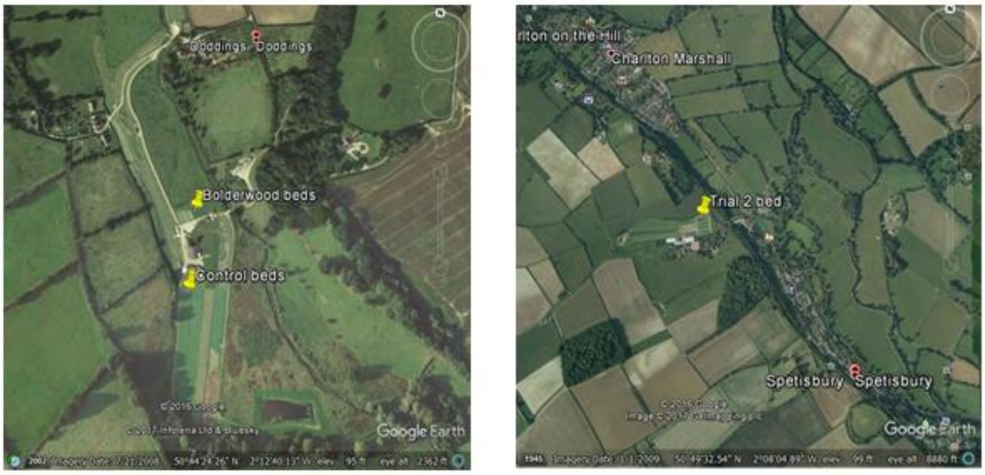
Available as Supplementary Table 3.2 on accompanying CD

Appendix C Chapter 4 Supplementary Material

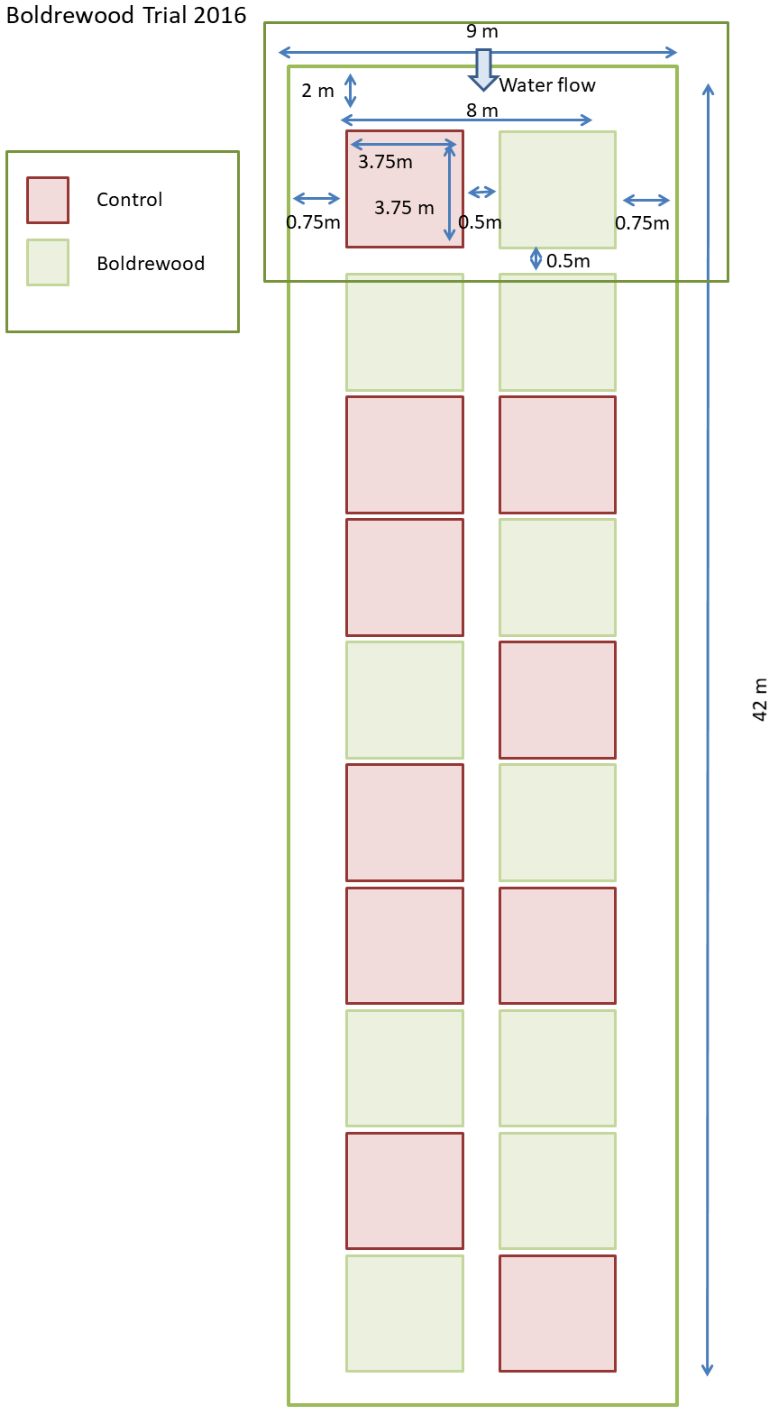
C.1 Figure S4.1: Google Earth visual of trial locations

Figure S4.1 Left: The location of the commercial farm in Trial 1, Dorset, U.K., with close-proximity Boldrewood and Control beds

Right: The location of the commercial farm and watercress bed used in Trial 2, Dorset, U.K.



C.2 Figure S4.2: Randomized design of Trial 2


Boldrewood Trial 2016
Control
Boldrewood
9 m
Water flow
2 m
8 m
3.75m
3.75 m
0.75m
0.5m
0.75m
0.5m
42 m