Lincoln University
Te Whare Wānaka o Aoraki
CHRISTCHURCH•NEW ZEALAND

Lincoln University Digital Dissertation

Copyright Statement

Development of a SNP Validation Toolset for Wheat

A Dissertation

submitted in partial fulfilment

of the requirements for the Degree of

Bachelor of Science with Honours

at

Lincoln University

by

Hoi Yee Kong

Lincoln University

2015

Abstract of a Dissertation submitted in partial fulfilment of the requirements for the Degree of Bachelor of Science with Honours.

Development of a SNP Validation Toolset for Wheat

by

Hoi Yee Kong

Recent advances in high-throughput technologies and the corresponding growth of the bioinformatics databases, a lot of bioinformatics tools have been constantly developed and published in journals for data-intensive genetic analysis. However, repoducibility has been one of the biggest challenges in delivering bioinformatics. Virtualisation is one of the most commonly used solutions by creating virtualised and isolated environments for running different bioinformatics tools. Containerisation, however, is a more recent, promising, scalable and reproducible approach for delivering bioinformatics. It allows running bioinformatics tools on preconfigured containerised environments, which is lightweight and can be easily updated and shared between researchers. This dissertation has evaluated a bioinformatics toolkit previously developed for genetic marker design in preconfigured containers. The objectives for this dissertation were (1) to improve marker design software to correctly handle melt prediction of amplicons with multiple SNPs, (2) using containerised environments to design and screen PCR assay to (a) validate candidate SNPs detected by GBS of barley *H. bulbosum* introgression line, (b) validate QTL markers of wheat identified by GWAS on chromosome 1A, and (c) convert SNPs from 90K SNP chip for the bread wheat A genome to HRM markers.

Keywords: Bioinformatics tools, High throughput sequencing analysis, Reproducibility, Virtualisation, Containerisation, SNP, marker, SNP validation, CAPS, HRM

Acknowledgements

I wish to acknowledge my principal supervisors John McCallum. I am sincerely thankful to his guidance, patience and wisdom throughout my study. This dissertation would not have been possible without the unconditional support from him. A big thank you to my supervisor Hamish Brown and Plant and Food Research for the research funds to undertake my dissertation this year. I would also like to thank my on campus supervisor Chris Winefield, who provided guidance and comments for this dissertation. A huge thank you go to my unofficial supervisor Paul Johnston for providing education on the cereal plants and laboratory skills, who in addition to proofreading and providing comments on the chapter of the laboratory experiments. All these help was much appreciated. Many thanks to Viji and Mei for their expertise and support which helped me in my laboratory experiments. This dissertation would have been a lot more challenging without their technical assistance and friendly support.

Most importantly, I would like to thank my friends and my sister Heidi. I would not have had the perseverance to get through this year and get this dissertation completed without their support and encouragement. A special thank you to my best friend in New Zealand, Jasmine. I truly appreciated for your support, understanding and many many meals provided. I am very grateful to have undergone the most important and meaningful year in my entire life with this little angel sent from god ☺.

A large thank is given to my mum and dad for their unconditional love and support from overseas. I truly appreciated for sacrificing things in their life to financially support my university studies. It must be hard to not have both daughters around in the past four years. I miss them. Thank you very much.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Literature Review

1.1 Recent molecular marker techniques for genotyping

1.1.1 Background and types of molecular marker

Advances in molecular biology have led to the introduction of many different types of molecular markers. A molecular marker, also known as genetic marker, is an identifiable polymorphic fragment of genetic material, usually DNA, which is associated with a certain location within the genome (Griffiths et al., 2000). It provides information about allelic variation at a given locus, thus it is widely used for detection of specific sequence differences between individuals, populations or species. Allozymes were the first genetic markers established which are based on the variants of protein structure in enzymes that can be separated and distinguished by gel electrophoresis (Schlotterer, 2004). Other DNA-based molecular markers which were previously widely-used for DNA profiling are restriction fragment length polymorphisms (RFLP) and minisatellites, both depend on restriction enzymes which recognise and cut the specific cleavage sequences to produce restriction digest fragments (Langridge and Chalmers, 2004). These markers can be used for detection of DNA variation, for example, a SNP or indel, which causes the formation or removal of a restriction enzyme recognition site resulting in digest fragments with different sizes. The invention and improvement of the polymerase chain reaction (PCR) technology has driven the development of numerous different PCR-based molecular markers including amplified fragment length polymorphisms (AFLP), randomly amplified polymorphic DNAs (RAPDs) , microsatellites and a variety of SNP assay systems (Schlotterer, 2004).

Another technique that is widely used for studying genetic variation is DNA sequencing of complete or partial genomes. Sanger sequencing was previously the most common sequencing method since it was developed in 1977 by Sanger and colleagues (Sanger, Nicklen and Coulson, 1997). However, Sanger sequencing is a slow process and can only sequence a few thousand base pair in a week. In addition, the traditional genotyping method using molecular markers is limiting in the short length of the DNA that can be characterised in each assay. The limitations of these technologies and the high demand of a rapid and low-cost genome sequencing method have driven the development of next generation sequencing technologies which is rapid and high-throughput, cost-effective and less labour required.

1.1.2 Single nucleotide polymorphism (SNP)

Single nucleotide polymorphisms (SNP) are the single nucleotide variations in the DNA sequence of individuals among a species or population. They are the most abundant genetic variations and evenly distributed in high frequencies throughout the genome of most animals and plant species (Wang et al., 2015). For examples, maize has 1 SNP per 60-120 bp, while humans have an estimated 1 SNP per 1000 bp. (Soleimani et al., 2003). SNPs can be mainly found in the non-coding regions of the genome in most organisms studied to date. There are two types of SNPs that are found in coding regions: synonymous which do not change the amino acid sequence, or non-synonymous which cause changes of amino acid sequence (Solemani et al., 2003). Unlike most of the genotyping approaches using molecular markers, SNP analysis does not require DNA separation and thus, can be performed using high-throughput automated technologies.

Since the technology of high-throughput automated sequencing has become more mature and reliable, large amount of SNPs can be easily identified at a low cost by sequencing whole genome or transcriptome of individuals from a wide range of population (Houston et al., 2014).

1.1.3 Use of molecular marker in plants genetics

SNP markers have been widely used for several purposes including crop cultivar identification and construction of high-density genetic maps (Wu et al., 2014; Baldwin et al., 2012; Raman et al., 2014). Due to the marker abundance and the biallelic polymorphism of SNP compared to microsatellite, using SNP markers for detection of marker-trait associations in quantitative trait locus (QTL) and genome-wide association studies (GWAS) has been a recent strategy used in plant genetics.

QTL analysis is used for identifying the genomic regions associated with the quantitative traits using both genetic (linkage maps) and phenotypic data (trait measurements). Comparing with the previous approach for QTL detection based on linkage maps that were constructed using low-throughput molecular markers, the recent approach using high-density genetic maps constructed through NGS for performing QTL analysis provides more complete and precise information about the size, location and estimated effects of detected QTL (Yu et al., 2011; Stange et al., 2013).

GWAS is a genome scanning approach that has become increasingly popular over the last few years since the advanced high-throughput technologies has made large-scale discovery of genome-wide SNP possible. GWAS uses genome-wide SNP data for identifying the associations between SNPs and complex traits. It has been widely performed on large-scale SNP data of many plant, such as rice (Huang et al., 2012; Yang et al., 2014), sorghum (Morris et al., 2012), barley (Pasam et al., 2012), tomato (Shirasawa et al., 2013), wheat (Sukumaran et al., 2015) and apple (Kumar et al., 2013), for trait improvement.

The use of SNPs in genetic analysis has given us a better understanding of the genetic basis of some complex biological systems in plants, such as development and growth, reproduction and adaptation to biotic and abiotic stresses (Raman et al., 2014; Rookiwal et al., 2014).

1.2 Development of SNP marker

1.2.1 SNP discovery

The discovery of SNPs can be performed experimentally by DNA sequencing or searching in silico (Chagne et al., 2012). Before the development of new-generation sequencing technologies, different experimental strategies, including fingerprinting, amplicon resequencing using Sanger's method and in silico alignment of sequence from DNA libraries, had been used for SNP detection. However, these SNP calling strategies were low-throughput, labour-intensive and high in cost. The development and advances in NGS technologies has made these old strategies obsolete. Rapid and cost-effective SNP discovery within genes can be done by transcriptome resequencing using NGS technologies. This methodology was also used coupling with other genome complexity reduction techniques, such as Complexity Reduction of Polymorphic Sequences (CRoPS) (van Orsouw et al., 2007) ,Restriction Site Associated DNA (RAD) (Davey and Blaxter, 2011) and genotyping-by-sequencing (GBS) (He et al., 2014), for discovering SNP in a genome-wide fashion.

1.2.2 SNP validation

A large pool of SNPs identified using NGS technologies needs further processing depending on the specific use case. SNP validation is usually conducted after a large-scale genome-wide SNP discovery, in order to confirm the detection of true SNPs and maximise the number of functional polymorphic markers for the following genetic analysis. This validation step is also required for distinguishing and discarding SNPs from paralogous or homologous genes, and technical errors.

Due to the high-price and intensive labour for validating every single SNP detected during the large-scale SNP discovery, a subset of SNPs is often chosen and screened over an informative set of samples for validation. This subset can be selected randomly or filtered out under specific constraints that might benefit the final genetic analysis.

There are plenty of techniques used for performing SNP validation. These include PCR-based assays, restriction enzyme digestion, resequencing, high resolution melting, and primer extension, among which resequencing using the Sanger method was previously the most commonly used method to validate SNPs. SNPs that pass validation can be converted into a single use marker for marker-assisted breeding or genetic analysis of targets of interest, such as disease resistance genes.

1.3 SNP validation methods

The two methods for validating putative SNPs focused in this study are cleaved amplified polymorphic sequence (CAPS) and high resolution melting (HRM).

1.3.1 Cleaved amplified polymorphic sequence (CAPS)

One of the common methods for SNP validation is through cleaved amplified polymorphic sequence (CAPS). Many reported studies utilised CAPS for validating identified SNPs, such as wheat (Iehisa et al., 2012), melon (Blanca et al., 2011), chickpea (Varshney et al., 2007) and Capsicum (Garces-Claver et al., 2007). Some papers have further converted validated SNPs into CAPS markers (e.g., tomato (Kim et al., 2012), rice (Lee et al., 2009) and soybean (Shu et al., 2011).

CAPS is a technique combining both PCR and restriction enzyme digestion, also known as PCR-RFLP marker, which can be used to differentiate between homozygous and heterozygous alleles. The principle is based on the sequence polymorphism between individuals which is amplified by PCR and digested with restriction enzyme producing digest fragments with different length. (Figure 1.1) The genetic variation can be then visualised and identified by gel electrophoresis of the digested products.

CAPS markers has been widely used in plant genetics for, such as, molecular identification, cloning, genotyping and mutation detection due to the co-dominance and high locus-specificity (Shu et al., 2011). Unlike RFLP, PCR-based CAPS assay requires very small amount of DNA template, and does not require the time-consuming process of Southern blot hybridisation and the use of radioactive isotopes. CAPS markers are also reproducible which can be easily shared between laboratories. However, a CAPS assay involves many steps, including verification of PCR amplification, digestion and incubation of PCR products with restriction enzymes, and separation of digest fragments on a high percentage gel, which could be time-consuming and labour-intensive (Ramkumar et al., 2015).

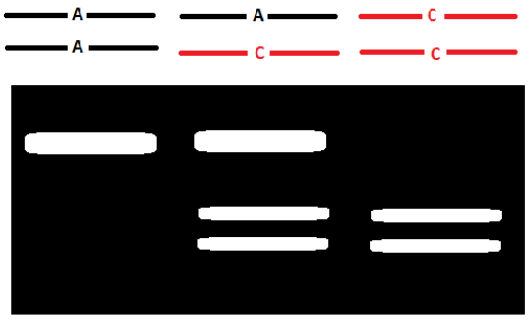


Figure 1.1 Digestion profile of CAPS assays for identifying three possible SNP genotypes – reference homozygote, variant homozygote and heterozygote. This example shows that the allele containing restriction site with C is recognised and cleaved by restriction enzyme, whereas the allele with A is not cleaved. Heterozygous and homozygous of both possible genotypes can be distinguished from the separation of digests on the gel.

1.3.2 High Resolution Melting (HRM)

Another frequently used method for SNP validation is high resolution melting (HRM) which was developed in 2002 by Idaho Technology and the University of Utah (Reed, Kent and Wittwer, 2007). HRM is a powerful technique that is able to detect not only sequence polymorphisms including genetic variations and mutations, but also epigenetic differences by methylation profiling (Muleo et al., 2009; Migheli et al., 2013).

HRM is based on detecting the thermodynamic differences between DNA strands with different sequence. Prior to the HRM analysis, a short SNP target (40-100 bp) is amplified with a saturating dsDNA binding dye, such as EvaGreen, which fluorescents when it binds to dsDNA. As the concentration of PCR amplicon increases, the fluorescent level also rises. At the end of the amplification, amplicons are denatured at 94°C and quickly re-annealed for the formation of homoduplex and heteroduplex, which are the perfect complementary dsDNA and non-perfectly re-annealed dsDNA, respectively (Chagne, 2015). During HRM analysis, PCR amplicons are gradually heated from 50°C or 65°C to 95°C. Once the melting temperature (Tm) is reached, double-stranded amplicons denature and melt apart coupling with a sudden drop of fluorescent level due to the release of intercalating dye from dsDNA. (Figure 1.2)

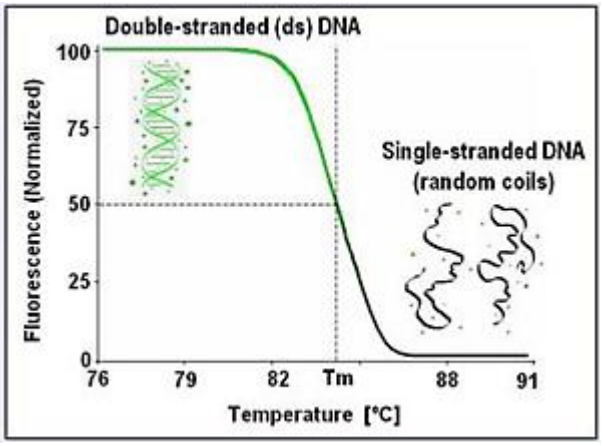


Figure 1.2 Diagram illustrating the denaturation of DNA during HRM analysis. (High Resolution Melting., n.d.).

The level of fluorescence is monitored during the whole process displayed on a melting curve. This melting profile of the amplicon is dependent on its GC content, length, sequence and heterozygosity due to the higher thermal energy required to destabilise a GC pair bound by three hydrogen bonds than an AT pair which has only two hydrogen bonds (Stoep et al., 2009). Variations of Tm and the pattern of melting profile allow the identification of the presence of homozygous and heterozygous variants. Heterozygous samples usually can be distinguished by a different pattern of melting curves caused by heteroduplexes. Whereas, different genotypes of homozygous samples usually have melting curves with the same pattern which can still be identified by the Tm shift (Figure 1.3) (Lehmensiek et al., 2008).

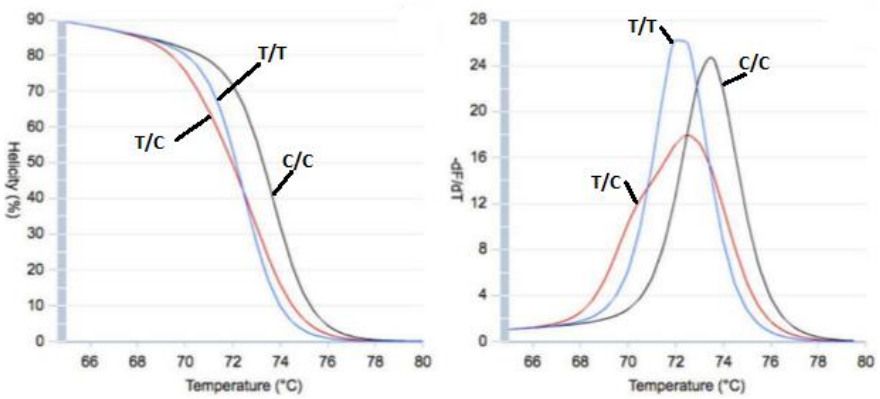


Figure 1.3 HRM profiles of the three possible SNP genotypes. (edited from Dwight et al., 2014)

HRM is a cost-effective method compared to other methods for SNP validation and genotyping, such as sequencing. It is a fast and possibly the simplest closed-tube screening method because there is no additional post-PCR processing, such as separation of products by gel electrophoresis, required after HRM analysis which can be performed in less than 10 minutes (Norambuena et al., 2009).

1.4 Tools for designing PCR based marker assay

There are several rules for design and selection of ideal primer sets for PCR-based marker assays. Melting and annealing temperature, GC content, primer length and product size are some of the parameters that could influence the specificity and efficiency of PCR amplification. Calculating these parameters for every primer design manually could be frustrating and time-consuming. Therefore, primer design for PCR assays has been relying on algorithms that are well developed for primer design and selection.

There are a number of public tools and software available for designing diagnostic PCR assays for validation. A variety of parameters for primer selection can be set prior to performing primer design. Primer3 (Rozen and Skaletsky, 2000) is one of the most widely used primer design programs and consists of a command line program and a HTML web interface called Primer3Plus (Untergasser et al., 2007). There are other common web-based primer design tools such as Primer-BLAST (Ye et al., 2012), OligoCalc (Kibbe, 2007), etc. These tools provide user-friendly interface for biologists to perform primer design for individual markers. However, it could be time-consuming using these tools for bulk design of primer sets to SNPs from NGS data.

1.4.1 Tool for bulk marker design – Galaxy-pcr-marker

Galaxy-pcr-marker (Baldwin et al., 2012) is a generic toolset developed for automating design of PCR assays to validate bulk selection of variants discovered from NGS. It includes two main tools: “find_CAPS.py”, which identifies restriction polymorphisms that might be recognised by and performed CAPS assay with one of the economical restriction enzymes using BioPython, and

“design_primer.py”, which designs flanking primer pairs to a list of variant using BioPython and Primer3 executable, and optionally predicts Tm of reference and variant using the uMelt web service provided by the University of Utah. This toolset provides a few helper scripts which can be used to produce one of the input files, a GFF feature file, by converting read mapper output into GFF3 format. Electronic PCR can be performed following primer design using an additional tool included in this toolset for annotation or redundancy checking.

Comparing with other software or tools developed for primer design which are normally for designing primer sets for a single marker, galaxy-pcr-marker toolset is able to facilitate identification of CAPS and design flanking PCR primer sets on large data sets, such primer sets can be used in different PCR-based assays including HRM.

Galaxy-pcr-marker toolset was developed in BioPython and adapted for use in the Galaxy workflow environment. Galaxy is a web-based bioinformatics workflow framework which provides biologist-friendly graphical user interface (GUI) for tools and complex workflow for bioinformatics tasks. This toolset can be delivered at Galaxy Tool Shed which is a platform for tool developers to share, update and manage their tools across Galaxy. However, due to major changes in the Galaxy environment and the complexity of managing software pre-requisites, the installation of these tools is very complex and in need of an update.

The toolset can also be accessed through GitHub (https://github.com/cfljam/galaxy-pcr-markers) and used directly from the command line or in authoring tools, such as Jupyter notebook (http://jupyter.org/).

1.5 Aims and objectives

Large-scale sequence variants can be identified by whole genome sequencing. Following the discovery of sequence polymorphism, bioinformatics tools are required for designing diagnostic assays for variant validation.

The aim of this dissertation is to evaluate the potential of a generic toolkit previously developed for large-scale PCR-based variant validation (Baldwin et al., 2012).

The main objectives to achieve this goal were

1 Improve assign design software to correctly handle melt prediction of multi-SNP amplicons

2 Use containerised software environments to design and screen PCR assays to:
   a. validate candidate SNPs detected by GBS of barley *H. bulbosum* introgression lines
   b. convert 90k SNP Chip SNPs for the bread wheat A genome to HRM markers
   c. validate SNP markers across a putative flowering time/earliness *per se* QTL identified by GWAS on wheat Chromosome 1A

Chapter 2

Bioinformatics and dry lab experiments

2.1 Brief introduction of bioinformatics

High-throughput technologies with increasing data volume has turned molecular biology into a data-intensive discipline (Spjuth et al., 2015). Bioinformatics, which combines computer science, statistics, mathematics and engineering, has become an essential field in science for storage, analysis and sharing of biological data using high-performance computing resources. Facilitating reproducibility of bioinformatics projects has been one of the main goal nowadays. Adopting a reproducible research by sharing code and data enables others to verify the findings, apply it to their own data or extend the old approach to new applications (Buffalo, 2014).

Bioinformatics resources

There are plenty of bioinformatics resources developed, mostly notably bindings for major language including R/Bioconductor (Gentleman et al., 2004), Biopython (Cock et al., 2009) and BioPerl (Stajich et al., 2002). These bioinformatics tools are mainly used for sequence alignment, sequence analysis and format conversion. Most of the tools were designed and developed for the Unix command line which could be convenience and easy to facilitate and parallelise in Unix when dealing with large amount of data (Buffalo, 2014). However, new users might find operating a command line interface difficult due to a high degree of memorisation and familiarity required for operation. Thus, there were some applications developed with GUI which can be easily operated by new users although users have less control over the files and operating system (OS). Traditional bioinformatics analysis is usually carried out with multiple steps using different these tools on a server or local computer. The integration of these tools allowing automated data analysis is commonly referred to pipelines or workflows (Spjuth et al., 2015). There are several scientific workflow systems available that allows streamlining the construction, execution and sharing of workflows for conducting efficient scientific analysis. For examples, Taverna (Oinn et al., 2004), Galaxy (Blankenberg et al., 2010), Kelper (Altintas et al., 2004) and Chipster (Kallio et al., 2011) are the common scientific workflow management systems with GUI that can cater to users without extensive bioinformatics background. There are also some lightweight workflow systems, such as Bpipe (Sadedin et al., 2012) and BcBio (https://github.com/chapmanb/bcbio-nextgen), which allow construction of workflows as custom scripts by experienced bioinformaticians using a programming language such as Bash, Python or Perl.

Packaging these bioinformatics resources allows users to download and install multiple pieces of software in the required version at once without undergoing individual download and installation

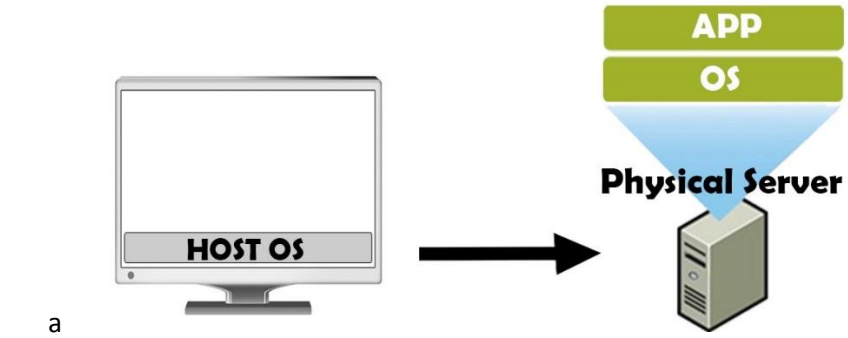
steps which are often frustrating and time-consuming (Field et al., 2006). However, he installation, distribution, and maintenance of these bioinformatics resources for scientific analysis has been the difficulties in delivery of bioinformatics.

2.2 Delivery of bioinformatics

The approaches for delivering bioinformatics tools has been evolving rapidly over the last decade. Ten years ago, these tools were only delivered to the users by installing directly on physical servers or local computers (Figure 2.1).

Bioinformatics tools are mostly developed based on other packages of software in a specific version, therefore, users using this approach often suffer from dealing with complicated dependency which is one of the most frustrating issues in distribution of bioinformatics (Boettiger, 2014). Installing tools on a physical server could not only cause configuration issues, but also require physical hardware systems and computational knowledge for setup and maintenance. Although a local desktop can be easily set up and operated by user with basic computer knowledge, different tools that have different OS requirement and other dependencies cannot be installed on the same local machine. Some users had multiple computers set up running under different environments with different OS and other dependencies in order to solve this problem. However, this solution was expensive and unreproducible.

In the past five years, delivery of bioinformatics has evolved from physical server-based to virtualisation or cloud computing. Using virtualisation to create isolated environments for running different bioinformatics tools is one of the most common solutions for the dependency problem. Many tools can also be reached on the cloud through cloud service providers, such as Amazon Web Services and NZGL. There were several cloud computing platforms developed providing user-friendly interface for biologists to perform data analysis using applications and tools available on the cloud. BaseSpace (https://basespace.illumina.com/) is a genomic cloud computing environment developed by Illumina which offers a wide variety applications and workflows for automated NGS data analysis and management and storage hosted on Amazon Web Services. It provides a non-specialist-friendly GUI which allows a wide range of audiences to perform analysis and share data with other researchers. Google Genomics (https://cloud.google.com/genomics/) is another genomic cloud environment providing an application program interface for storage, processing and sharing of large-scale sequencing data.



a

b

Figure 2.1 Diagram illustrating the previous approaches for accessing bioinformatics tools and applications on (a) physical server or (b) multiple local computers.

2.3 Challenges

2.3.1 Diverse audience

One of the challenges that bioinformaticians have been facing in the delivery of these packages and tools is the diversity of audience. These bioinformatics tools have users from different areas, such as bioinformaticians, statisticians, mathematicians and biologists, who have different biological and computational knowledge and different levels of familiarity with programming languages. Many bioinformatics tools were developed and designed to be used on a command line in a Unix environment which does not have user-friendly interfaces and require basic knowledge in programming to operate. This might be a problem for researchers, who perform lab experiments and generate large-scale data, as they often only have basic computational knowledge and they are usually familiar with Microsoft Windows (Carvalho and Rustici, 2013). This is also the reason why there are many different tools and workflow systems developed for working under different environment and programming level.

2.3.2 Reproducibility

Furthermore, a reproducible research requires bioinformatics tools that can be easily shared and distributed to the diverse audience.

Installation and configuration

However, the installation of tools and configuration of a complex and suitable set of dependencies has been one of the biggest problems in distribution of bioinformatics. Installation could be an issue when user is working under unsupported environment. For example, Primer3-py (Untergasser et al., 2012) is a python application program interface that includes Primer3 library providing a simple interface for automated primer design. However, this package was built and tested on Linux and Mac OS X systems, and it does not provide official Windows support. Thus, Windows users might have trouble in installation of this package under Windows environment. Missing libraries, packages, compilers, and search paths could also lead to installation issues (Collberg et al., 2014). The configuration is another issue causing problems when accessing or employing these bioinformatics tools (Boettiger, 2014). This usually due to the tools having shared packages or libraries with other tools but where they depend on different and incompatible versions of the shared packages or libraries.

Maintenance

Following the development stage, maintenance of these tools are needed, including bug fixing, upgrade for the version of programming language and evolution of OS, and improvement of algorithms and features. However, these changes could potentially alter the results generated by the code, or in worse scenarios, break the whole program.

Benchmarking

Newly developed and published bioinformatics software can be evaluated against the current state of software using benchmarking. This can be used to assure high quality, identify benefits and limitations of the tools, evaluate the improvements from the new methods and allow user to choose the suitable tool for the specific aims and purposes of research (Aniba et al., 2010). With the increasingly use of high throughput technologies and new bioinformatics tools constantly created and published in journals, a number of benchmarks have been designed and running for a regular basis for comparative evaluation of latest developed tool against the current state of tools. However, the installation and configuration of a large amount of tools for evaluation could be problematic.

2.4 Solutions

2.4.1 Virtual machine

Virtualisation would be one of the solutions for the installation and configuration problems of bioinformatics tools that operate under different environment. Setting and starting up virtual machines on a local computer or physical server through a specialised software called hypervisor has been one of the most common approaches for virtualisation. The hypervisor, such as VirtualBox (Oracle Corporation), can emulate hardware, including CPU, memory, hard disk and other hardware resources, completely providing a virtual platform for virtual machines to run different guest OS on the same physical machine (Dash, 2013). Vagrant (HashiCorp) is another software for creating virtual machine that also allows user to package and provision the configuration and setup of a virtual machine into a script called Vagrantfile (Peacock, 2013). It provides a reproducible method to regenerate pre-build virtual environments.

Bioinformatics tools, therefore, can be installed and used on an isolated virtual system containing a complex set of preconfigured bioinformatics software (Nocq et al., 2013) (Figure 2.2). Furthermore, setting up multiple virtual machines with different on-demand execution environments on the same host reduces the need for physical hardware systems. Also, virtual environments on virtual machines can be easily backed up, recovered from disaster and shared between researchers.

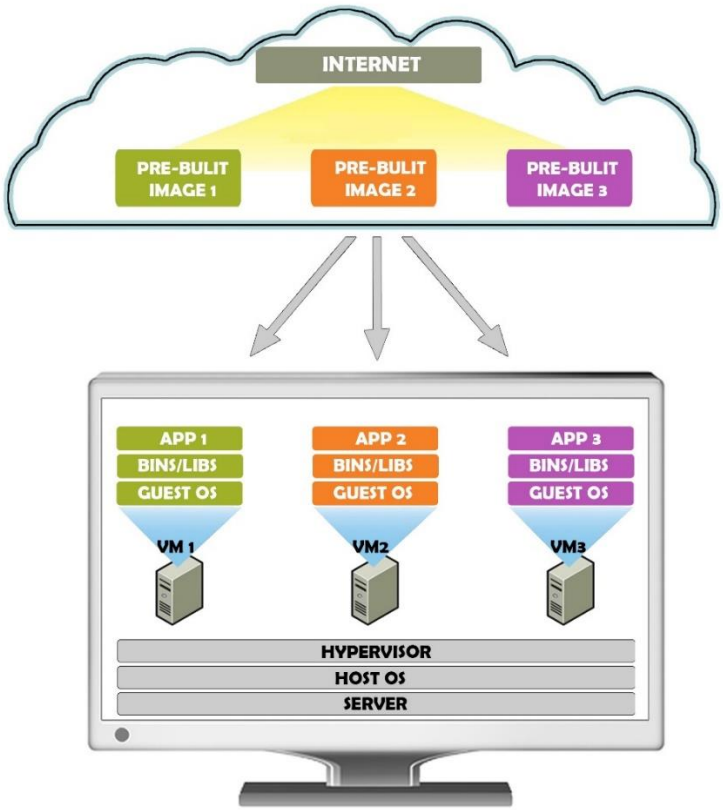


Figure 2.2 A virtualization architecture for development of isolated environments using virtual machine.

2.4.2 Virtual environment

Dependency problem in bioinformatics delivery could be solved by starting up a virtual machine for working under an isolated environment in a separated guest OS. However, having too many virtual machines installed and run on a single local computer host when multiple environments are required could reach the limitations of memory, CPU and scheduling which leads to low performance or crashes and poor user experience (Boettiger et al., 2014).

An alternative of virtualisation is creating isolated virtual environments that have their own directories and paths. The packages in specific version required for different projects and their dependencies can be installed, run and maintained in separate environments on the same local computer host. Switching between environments can be done by simply executing two lines of command to deactivate the current environment and activate another one. Using virtual environments, therefore, allows users to work with specific version of Python or libraries without creating problems with different dependencies required between projects.

There are different tools available for creating virtual environments. For example, virtualenv (https://virtualenv.pypa.io/), is an environment manager which creates isolated working environment that has its own installation directories so that libraries are not shared between virtual workspaces.

Another tool widely used for creating virtual environments is conda. It works similar to virtualenv, but it is further developed as an environment manager as well as a package manager. Conda allows installation of packages with both Python and non-Python installation tasks. Multiple versions of packages and their dependencies can be installed in independent environments which can be switched between easily. It works on all the three most commonly used OS, which are Linux, Apple Mac OS X and Microsoft Windows.

Conda can be installed through Anaconda (Continuum Analytics) which is a scientific Python distribution that includes different versions of Python, Jupyter notebook and more than 300 Python packages commonly used for science computing, engineering and data analysis, such as Numpy, SciPy, Pandas and matplotlib. However, installing Anaconda with all the packages might be time wasting and storage consuming. A smaller alternative of Anaconda is Miniconda (Continuum Analytics) which includes only Python, conda and its dependencies. Other software and packages that are required for a particular project can be installed manually or automatically using a pre-build requirement list.

2.4.3 Containerised environments

Compared to traditional virtualisation using heavyweight virtual machines, a more promising approach for delivering bioinformatics tools is through containerised environments. Traditional virtualisation that has individual guest OS installed on each of the multiple virtual machines could be wasteful in terms of memory, bandwidth and storage. Whereas, containerisation is a lightweight virtualisation based on building containers on a containerisation engine rather than the hypervisor. The best known containerisation engine is Docker (Docker, Inc.). Software or applications can be wrapped up in a Docker container containing code, tools and libraries required for running the applications. Unlike virtual machines, each Docker container runs on a single Docker engine sharing the same Linux kernel with the host machine. Lightweight containers, therefore, are much smaller than virtual machines in size conducting better performance (Boettiger et al., 2014). A Docker container can be easily created by running a Docker image built by following the exact instructions from Dockerfile which is a script containing all the commands stored on the registry hub, such as Docker Hub or GitHub (Figure 2.3).

This approach can be used to create the containerised working environment that has all the required dependencies configured and perhaps some other useful tools installed for conducting scientific analysis. This Docker-based approach resolves the dependency issue and avoid the tedious installation of several pieces of software by automatically building the ideal working environment using the pre-built Dockerfile (Tommaso et al., 2015). Moreover, containers can adapt very easily to the changes of the dependencies during maintenance of the bioinformatics tools by simply updating the Dockerfile manually or committing changes to the local Docker image and push it to the Docker registry.

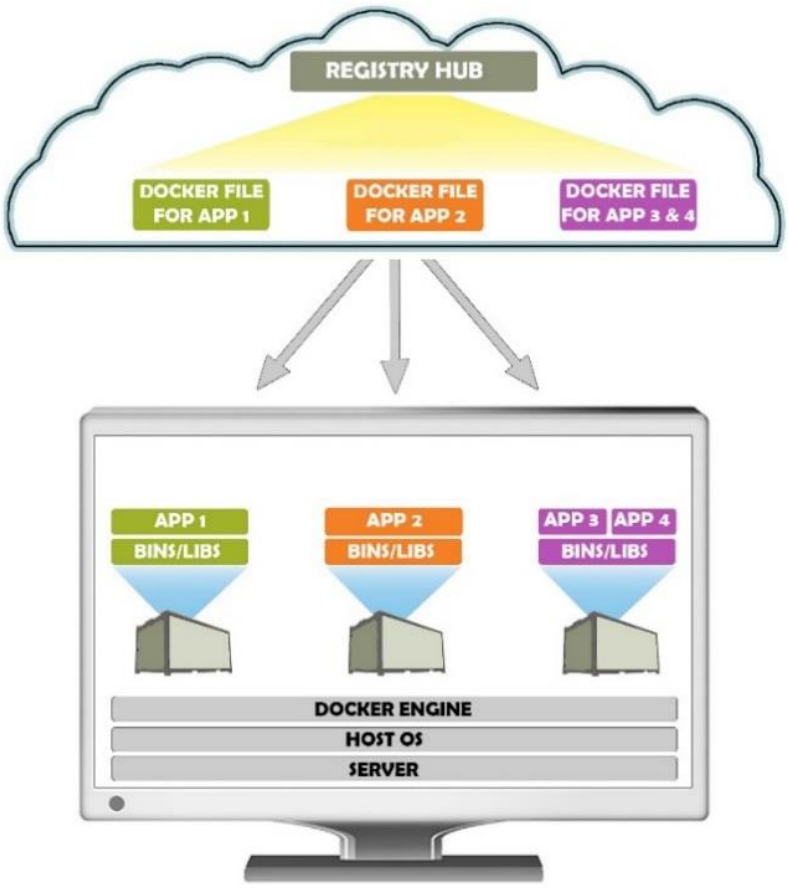


Figure 2.3 A containerisation architecture for building containerised environments in a local server or computer.

2.5 Dry lab experiments in this study

2.5.1 IPython notebook as electronic notebook

In this dissertation, all the bioinformatics tasks were conducted in Jupyter notebook which is one of the main contributions to reproducible research.

Jupyter notebook, previously called IPython notebook, is a web-based interactive authoring tool which allows creating, editing and sharing of dynamic documents that contain live code with rich media output, mathematics, computations and explanatory text (Shen, 2014). Jupyter notebook comes with IPython (stands for interactive Python) kernel which allows using Python within notebooks. Many other additional kernels can be installed for running notebooks in other programming languages, such as R, Julia and Perl. Also, many packages and interactive widgets can be used within Jupyter notebook for data manipulation and visualisation producing rich output in real-time. Moreover, notebooks can be easily shared with others through some basic platforms, such as Dropbox, Google Drive or even emails. Storing notebooks as GitHub Gists (https://gist.github.com/) for rendering using nbviewer (http://nbviewer.ipython.org/) , which is a Jupyter notebook viewer, or directly on GitHub is another alternative for sharing notebooks. Another recently developed notebook viewer, Gistexec (https://github.com/rgbkrk/gistexec) , is a combination of Gist and notebook providing an interactive interface with executable cells which allows viewer to see the code as well as the actual output instantly.

There are a number of ways for launching Jupyter notebook. Firstly, some scientific Python distributions, such as Anaconda, include Python, Jupyter notebook and other packages that are commonly used for scientific computing as mentioned above. However, these scientific Python distributions are local applications which require user admin privileges for installation. Secondly, there are browser extensions that allow user to launch Jupyter notebook within a browser without installing any other software. For example, Jupyterdrive is a notebook extension which can be installed through the package called Jupyter-drive from GitHub (https://github.com/jupyter/jupyter-drive). This extension allows Jupyter notebook to be launched within Google Chrome and use Google Drive for file management. Thirdly, there are managed Jupyter notebook servers available in the cloud. Wakari (https://wakari.io/) is one of the examples that is hosted in Amazon Elastic Compute Cloud (EC2). It allows user to conduct data analysis and visualisation in Jupyter notebook launched in a browser-based Python and Linux environment using Anaconda. Data and notebooks can be stored and shared in the cloud. Also, user can launch a notebook on a prebuilt virtual machine in the cloud, such as NotebookCloud (https://notebookcloud.appspot.com/docs) which allows launching and controlling Jupyter notebook server in Amazon EC2 from a browser. Alternatively, notebook server can also be run within a virtual machine on a local computer launching the notebook through a

browser. Furthermore, another way to launch a Jupyter notebook is to pull and run a Docker container within a lightweight virtual machine. There are plenty of prebuilt Docker images available on GitHub and Docker Hub that contain ready-to-run Jupyter notebook, and some useful tools and packages.

In this study, Jupyter notebook was launched on a Docker container called cfljam/socker (https://github.com/cfljam/socker). It was developed and maintained by John McCallum (john.mccallum@plantandfood.co.nz) and was designed for use in statistical genetics and genomics. It contains Jupyter notebook with some important dependencies, such as primer3-py and bcbio-gff. It also provides some genetic tools, including VCFtools, VCFLib, Samtools, BedTools and R genetics tools.

2.5.2 Configuration and enhancement of bioinformatics environment

Improvement to melt

In the previous version of the primer design software, the mutation of the reference to the mutant amplicon only includes a single target SNP. The current version of the software was extended to include all SNPs within the amplicon in order to handle correct melt prediction of multi-SNP amplicons.

Here are the links of the primer design software enhancement committed to GitHub repository: https://github.com/cfljam/galaxy-pcr-markers/commit/675b979d30a0c4433bb123242a47cf75b9612ab8?diff=split

and the secret gist of the notebook for testing improved primer design software with a small subset of SNPs: https://gist.github.com/hymmikong/cd6fac83481aa98d276d

Steps for setting up working environment

1. In order to run a Docker container on a Window machine, Boot2Docker, a lightweight Linux distribution based on Tiny Core Linux that allows to run Docker daemon on Windows, was installed. The installation of Docker on Windows has recently migrated from Boot2Docker to Docker Machine which can be installed using Docker Toolbox. The schematic workflow for working environment setup and primer design is illustrated in Figure 2.4.

2. To run docker container cfljam/socker, Boot2Docker need to be started up by either running the following start commands or using the Boot2Docker start shell script.

```
boot2docker init
# which creates a virtual machine for Boot2Docker if one does not
exist.
```

```
boot2docker up
# which start up the boot2Docker virtual machine and Docker daemon.
```

   When Boot2Docker is up, the docker image of cfljam/socker can be pull down from GitHub and a writable container layer over the docker image is created at the same time by executing the following command. A "-v" flag is used for mounting a local directory from the Docker daemon's host into this new container. '-p' flag is for publishing a port of the container to the host.

```
docker run -rm -p 8888:8888 -v /my_local_dir:/vm_mount_point –it
cfljam/socker
```

3. In order to use toolkit galaxy-pcr-marker, its git repository need to be cloned to create a copy on local directory using following command.

```
git clone https://github.com/cfljam/galaxy-pcr-markers.git
```

4. To run Jupyter notebook, browse URL: http://localhost:8888/. Jupyter notebook is now ready for primer design or other data analysis.

2.5.3 Procedures of primer designing for wheat QTL marker validation as an example:

Link to the secret gist of the primer designing notebook:
https://gist.github.com/hymmikong/bbf2f0dda0daa56bc8fb

1. Both of the find_CAPS.py and design_primer.py tools require a FASTA sequence file and a GFF feature file as input files. A FASTA file containing reference sequence of wheat genome-wide distributed SNPs was converted from the supplementary table (Table S5) containing annotation of SNP loci from Wang et al. (2014) using dataframe and iterator in pandas which is a Python package for data structuring.
2. A GFF file containing features of all SNPs from the SNP array from Wang et al. (2014) was generated using code adapted from one of the helper scripts from galaxy-pcr-marker toolkit, vcr_gff.py.
3. Specific GFF and target files for selected markers were created using several Bash commands. IPython has a cell magic, %%bash, which allows the execution of Bash command lines within IPython or Jupyter notebook. http://nbviewer.ipython.org/github/ipython/ipython/blob/1.x/examples/notebooks/Cell%20Magics.ipynb
4. After preparing all the input files, primer design was conducted using the tool, design_primers.py, from galaxy-pcr-marker with following parameters:
    - maximum number of primer pairs to return = 5
    - minimum product size = 60
    - maximum product size = 120
    - do uMelt prediction = yes

    The tool was used in command line that was executed in Jupyter notebook using a line magic in IPython, Shell capture (%sc) which runs shell command and capture output, shown as below. The output of the primer design was processed and rendered using a Python module, StringIO and Pandas dataframe.

```
%sc HRM_primer_output=python \
../../../galaxy-pcr-markers/design_primers.py \
-i ../Data_files/wheat_snp_fasta.fasta \
-g ../Data_files/Chr1A_DTF_candidates.gff \
-T ../Data_files/Chr1A_DTF_candidates.targets \
-n 1  -p 60 -P 120   -u
```

5. Finally, the primer sets with Tm difference that is larger than 0.3 were written into a csv file. All the details of primer sets designed for both barley and wheat markers validation are provided in Appendices A.1 to A.4.

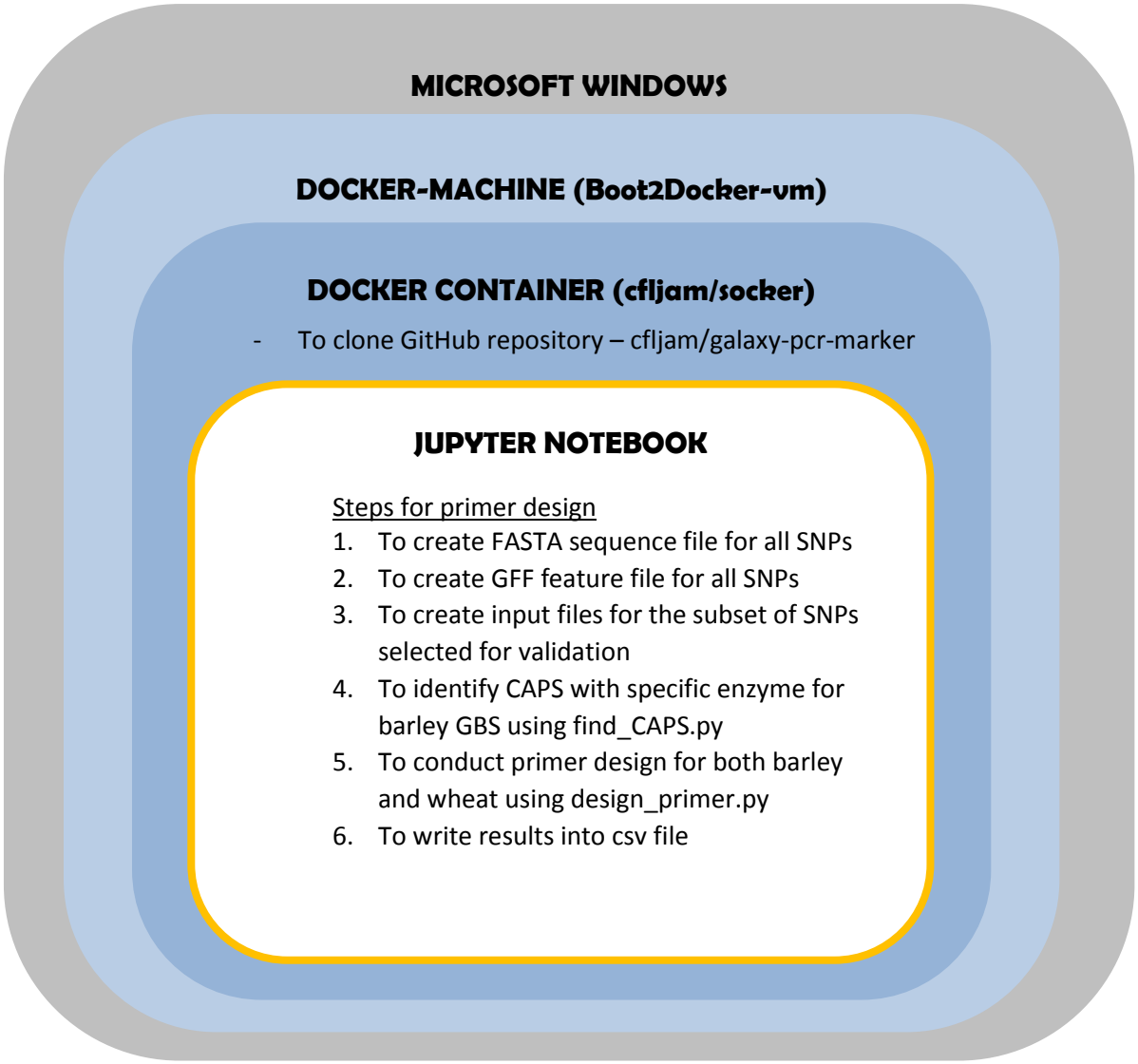


Figure 2.4 A schematic workflow for working environment setup and primer design.
Boot2Docker virtual machine created and started up by initialising Boot2Docker. A Docker container was then pull down from GitHub and ran on the Docker-machine. The toolkit was cloned to the local directory and used for primer design conducted in Jupyter notebook which was launched on a local browser.

2.5.4 Performing In silico PCR after SNP validation

In silico PCR, so called electronic PCR (ePCR), is sometimes performed following the design of primer sets in order to check for redundancy or to annotate reference sequences. Due to the polyploid nature of wheat genome, designed primer sets are more likely to target homologues of the actual gene of interest. In this dissertation, ePCR for all the wheat SNP markers was carried out after validation with regard to confirm that the in vitro PCR was amplifying a single target region from the genome of interest. ePCR was performed in Jupyter notebook using Ipcress (https://github.com/nathanweeks/exonerate), which is an in silico PCR experiment simulation system contained in the Exonerate package (Slater and Birney, 2005), by simply providing a set of wheat reference sequences retrieve from EnsemblPlants (ftp://ftp.ensemblgenomes.org/pub/plants/release-29/fasta/triticum_aestivum/dna/), and a file containing details of primer sets in a particular format. Ipcress then produced the predictions of PCR products for these primer sets. Results of ePCR experiments for the wheat markers in this dissertation are summarised (as number of hits) and provided in the Appendices B.3 and B.4.

Chapter 3

SNP validation and wet lab experiments

3.1 Barley

3.1.1 Background

Leaf rust is a fungal disease that affects stems, leaves and grains of cereal, including barley, wheat and rye, which causes serious seasonal crop loss. *Hordeum bulbosum*, is a secondary gene pool of cultivated barley and has desirable traits of barley (*Hordeum vulgare*), especially for pathogen resistance or tolerance (Wendler et al., 2014). Thus, *H. bulbosum* has been used as a source of genetic introgression for barley improvement by providing access to genetic diversity outside the primary gene pool of cultivated barley.

In previous study, an adult plant resistance APR gene which contributes to partial resistance of barley leaf rust (or slow rusting) was introgressed from *H. bulbosum* 'A17' into a barley cultivar 'Emir' creating introgression line (IL) '200A12'.

The performance of partial resistance to leaf rusting of '200A12' was described by Pickering and colleagues (2004). This paper observed that the latency period, which is the time taken for 50% of the eventual number of pustules to develop, on '200A12' was 16% longer than on 'Emir', indicating that there might be some genetic components in '200A12' that contributes to slow rusting.

A population of F_2 individuals was previously developed by back-crossing the IL '200A12' with its barley parent. The main goal of this experiment is to validate the primer sets of bulk markers designed using Galaxy-pcr-marker toolkit and ideally detect any recombination events in the F_2 population using validated markers.

3.1.2 Materials and methods

Barley plant material

DNA samples for validation of barley CAPS markers were previously extracted and provided by Plant and Food Research Lincoln. The IL '200A12' was produced by crossing the barley (*H. vulgare*) cultivar 'Emir' with the *H. bulbosum* 'A17' for hybrid seed production. Diploid progeny was then produced by backcrossing the hybrid as the pollen parent with its barley parent 'Emir' for elimination of the H. bulbosum chromosomes. IL '200A12' was created when there was a recombination event happened between the barley and *H. bulbosum* chromosomes during gamete formation (pollen, egg). It is followed by crossing barley cultivar 'Emir' and the IL '200A12' for mapping the partial resistance gene

in IL '200A12' for the development of a F_2 population of 183 individuals. A breeding scheme is shown in Figure 3.4 in the discussion section of this chapter.

For barley HRM markers, fresh young leaf tissue from 'Emir', '200A12' and hybrid ('Emir' x '2032') were collected and placed in test tubes which were then stored in the freezer overnight ready for DNA extraction. DNA was extracted from these frozen leaf tissues using MAS DNA extraction method (Chao and Somers, 2012).

CAPS markers

PCR amplification

A total of 21 CAPS markers were tested on 7 samples selected from the F_2 population that contained a mixture of homozygous (VV and BB) and heterozygous (VB) genotypes, where V and B represent a haploid genome equivalent of *H. vulgare* ('Emir') and *H. bulbosum* ('200A12'), respectively. The validation of each marker was performed with a negative control where the DNA template was substituted with sterile water to ensure solutions were free of contamination.

All markers were amplified in a 10 µl reaction volume that contained 1 x PCR buffer (ReddyMix), 0.2U of Thermo-Prime Taq DNA polymerase (Thermo Fisher), 0.2 mM dNTPs, 1.5 mM MgCl2, 0.3 µM each primer and 20 ng DNA template. Amplification of markers was conducted on a Mastercycler Pro S (Eppendorf) using the following thermal cycle conditions: an initial denaturation period of 94°C for 2 minutes, then 40 cycles of 94°C for 30 seconds, 50°C for 30 seconds and 72°C for 30 seconds, followed by a final extension period of 72°C for 5 minutes.

To check whether a single product was amplified in each reaction, 4 µl of the PCR product was separated by electrophoresis using a 3% agarose gel. The length of the PCR product was estimated using the standard 1 Kb plus DNA ladder (Life Technologies) and was compared to the expected PCR product size provided by the primer design tool.

Troubleshooting

In order to obtain the best amplification possible, the markers with the poor PCR amplification were re-amplified under different conditions including an increase in Mg concentration from 1.5mM to 2mM or 2.5mM and changing the annealing temperature from 50°C to 55°C.

Digestion

5 µl of the product of the PCR reactions showing a clear single band on the gel were digested in a 10 µl reaction using 2U of one of the seven restriction enzymes (TaqI, RsaI, HincII, DdeI, AluI, HinfII and DpnII) with the appropriate buffer and double-distilled water. The digests were incubated for 4 hours at 37°C (or 65°C for TaqI digests). Digested products were run on a 3% agarose gel for separation and visualisation under UV after staining with ethidium bromide.

Testing across the F2 population

The markers with the digested amplicons showing clear identification of its genotype were selected and tested across the F2 population with the total of 183 samples and a few positive and negative controls using the same protocol. (Figure 2.5)

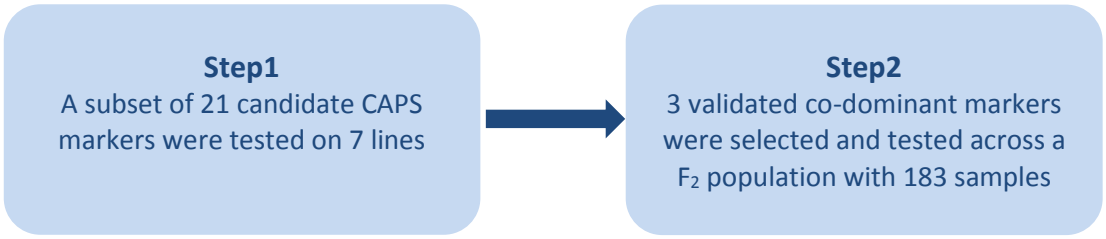


Figure 2.5 Steps of the validation process for barley CAPS markers.

HRM markers

PCR amplification

48 barley HRM markers were tested on 3 DNA samples of the three possible genotypes (VV, VB and BB) and negative control. They were first amplified in a 10 µl reaction using 5 X HOT FIREPol EvaGreen HRM Mix (Solis BioDyne), 10 µM of forward and reverse primer and 20 ng DNA template overlayed with 20 µl PCR grade mineral oil (SIGMA) to prevent evaporative losses during amplification and melting. 5 X HOT FIREPol EvaGreen HRM Mix comprises HOT FIREPol DNA Polymerase, 5x EvaGreen HRM buffer, 12.5 mM MgCl2, dNTPs, EvaGreen dye and BSA. Touchdown PCR amplification was carried out on a Mastercycler Pro S (Eppendorf) with an initial denaturation of 95°C for 15 min, followed by 10 cycles of 95°C for 30 seconds, annealing temperatures started at 60°C for 30 seconds and decreased by 0.5°C per cycle, and 72°C for 30 seconds for elongation. This was followed by 30 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and an extension at 72°C for 30 seconds. To conclude the PCR reaction, amplicons were heated at 95°C for 30 seconds and rapidly cooled to 28°C for 30 seconds to maximise the formation of heteroduplex if the DNA sample was heterozygous.

High resolution melting

After marker amplification, PCR amplicons were then transferred to a LightScanner (Idaho Technology Inc.) for HRM. The amplicons were melted from 50°C to 95°C. Melting curves were analysed using the Lightscanner software version 2.0.0.1331 (Idaho Technology Inc.) using the 'small amplicon' module.

3.1.3 Results

CAPS markers

A total of 21 markers within the introgression region were amplified. Among the 21 markers, 14 of them resulted in excellent PCR amplification indicated by a clear single band on the agarose gel. The rest of the 7 markers were discarded as 4 of them showed multiple products amplified on all 7 lines, 1 failed amplification in BB genotype, 1 failed to amplify target region and another 1 failed amplification on all lines. Out of the 14 successfully amplified markers, 8 of them were digested by the specific restriction enzyme (Figure 3.1). There were 2 markers showed no digestion in VB genotypes and another 4 markers failed to produce any digested products. Among the 8 markers with digested amplicons, 6 of them were shown to be co-dominant which can differentiate between homozygotes and heterozygotes. Digestion results of 4 of the 6 co-dominant markers are provided in the Appendix C.1. Although one of 8 resulted in partial digestions and bands that were very close to each together, it could possibly be a co-dominant marker as well. For the last marker with digested amplicons, VV was indistinguishable from VB as one of the alleles on VV was also digested (Table 3.1).

Number of primer set designed		21
PCR amplification		
Single clear band		14
Multiple bands		4
Only failed in BB line		1
Failed to amplify target region		1
Failed in all lines		1
Digestion of the 15 markers with good amplification		
Digested		8
Co-dominant	6	
partial digestion and unclear results	1	
one of the alleles on VV was cut	1	
Digestion only on BB line		4
No digestion		2

Table 3.1 Summary of barley CAPS markers validation. Appendix B.1 shows a more detail summary of the result.

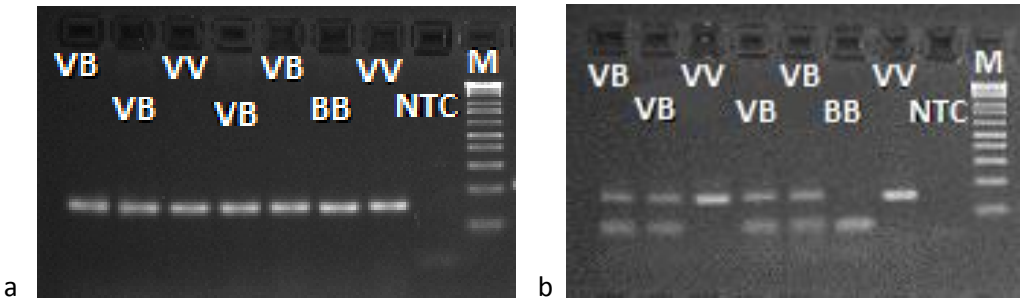


Figure 3.1 Example of the screening result of a co-dominant CAPS marker (CM_1186). (a) Verification of PCR amplification (b) Visualisation of restriction digest result

The 3 co-dominant markers (CM_1202, CM_1186 and CM_1199), which were located at the middle and near the two ends of the introgression region, were further tested across a total of 183 F_2 individuals for additional confirmation or even detection of recombination events within the introgression region. A gel picture of testing CM_1202 across a 96 plate of F_2individuals is shown in Appendix C.2. In most cases, the marker genotypes were in agreement for each DNA sample. However, one recombinant line was successfully identified by a change in the genotype detected for one of the CAPS markers.

HRM markers

A subset of 48 candidate SNPs identified by GBS were selected for SNP validation and tested on 3 samples with 3 different genotypes – VV, BB and VB. High fluorescent level were detected in the early stage of HRM analysis for all markers indicating that DNA was successfully amplified. Among the 48 markers, 14 were co-dominant which 11 displayed simple melting curve patterns (shown in Appendix C.3) and the other 3 markers showed complex profiles. Figure 3.2 shows the HRM profile of one of the co-dominant markers (CM_1158). From the rest of the non-validated markers, 8 of all markers had heterozygotes that were indistinguishable from one of the homozygotes, 6 gave overlapping profiles of the two homozygotes, 12 markers showed overlapping melting curves on all 3 genotypes and 8 markers displayed very complex profiles which were not validated (Table 3.2).

One co-dominant marker was further tested across a total of 96 F_2 individuals (Figure 3.3). Although more than 3 clusters were calculated using the auto-grouping function from the Lightscanner software, 3 main groups could be discriminated by the difference of Tm and profile patterns between curves. The marker genotypes detected from HRM were with agreement for the genotypes of all samples detected from previous validated markers.

Number of primer designed		48
Detected all 3 possible genotypes		14
Good clean HRM profile	11	
Complex HRM profile	3	
Indistinguishable heterozygotes		**8**
Indistinguishable homozygotes		**6**
Overlapping melting curves on all 3 lines		**12**
Messy melting curves		**8**

Table 3.2 Summary of barley HRM markers validation. Appendix B.2 shows a more detail summary of the result.

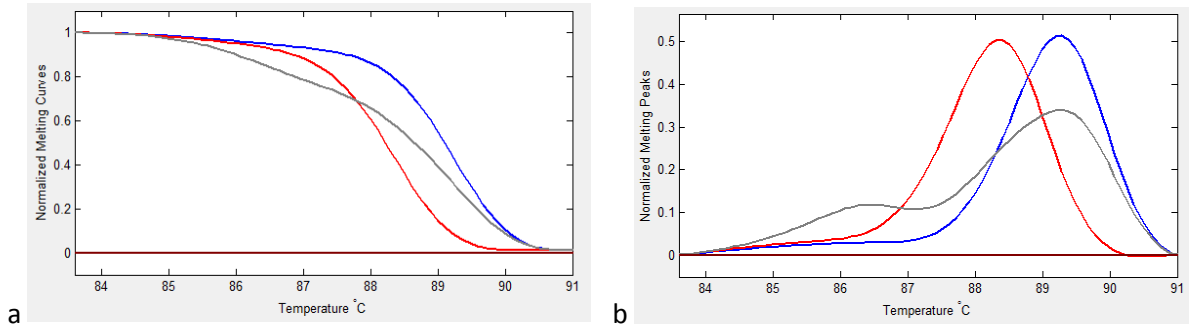


Figure 3.2 (a) Normalised melting curves, and (b) normalised derivative melting curves for the genotyping of CM_1158: G>A variants. It is possible to distinguish the two homozygous samples by their Tm variation and the heterozygous sample by the pattern of the melting curves. Barley homozygote (G/G), *bulbosum* homozygote (A/A) and heterozygote (G/A) are shown in blue, red and grey, respectively.

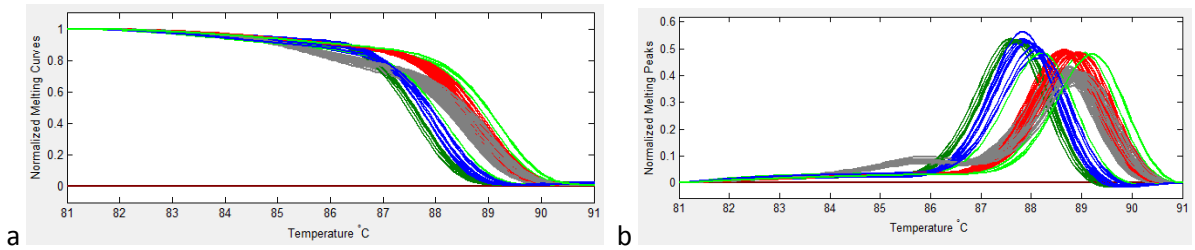


Figure 3.3 (a) Normalised melting curves, and (b) normalised derivative melting curves for CM_1158 tested across 96 F_2 individuals.

Seven of the co-dominant markers were selected and gel electrophoresis was performed to verify the HRM analysis to ensure that only the target sequence was amplified during PCR. All 7 markers displayed a single band with the same product size as the target sequence. Another 16 markers that were not validated were also run on the gel. More than half of the markers gave multiple bands indicating non-specific target DNA amplifications. The results are scored and shown in Appendix B.6 and the gel pictures are shown in Appendix C.4.

3.1.4 Discussion

CAPS markers

In the validation of the 21 candidates of barley CAPS marker, 5 of them produced multiple bands or single band in an unexpected size indicating that the primer sets were targeting to non-specific region. One marker failed to amplify BB line could be due to experimental error.

For those primer sets that were not amplifying a single product in the expected size on all samples, relaxing or tightening PCR conditions could help in improving performance of PCR amplification. Several markers were amplified again with higher concentration of magnesium (from 1.5 mM to 2.5mM Mg) and higher annealing temperature (from 50°C to 55°C) performing better amplification compared to PCR under the original conditions. More PCR troubleshooting could be done to result the best amplification of each marker under its ideal conditions, in order to increase amplification success rate. However, troubleshooting process is worth avoiding as it could be time-consuming and labour-intensive.

The marker that had digested products in BB lines but not in VB lines might be due to preferential allele amplification of *vulgare*. This might be caused by underlying sequence variation in the primer binding region leading to non-preferable amplification of *bulbosum* allele in heterozygotes.

Among the six validated co-dominant markers, 3 of them were selected and further tested across the F_2 population with the total of 183 samples. Most of results agreed with the genotypes of each DNA sample, but except for one which has a VB genotype resulting in a BB genotype detected by one of the CAPS marker located at a far end of the introgression region (CM_1186). This could be a recombinant line that had recombination events occurred between chromosomes during gamete formation (Figure 3.4). Further confirmation, such as resequencing or testing with other markers, are required to ensure that the change of genotype detected was not caused by experimental error.

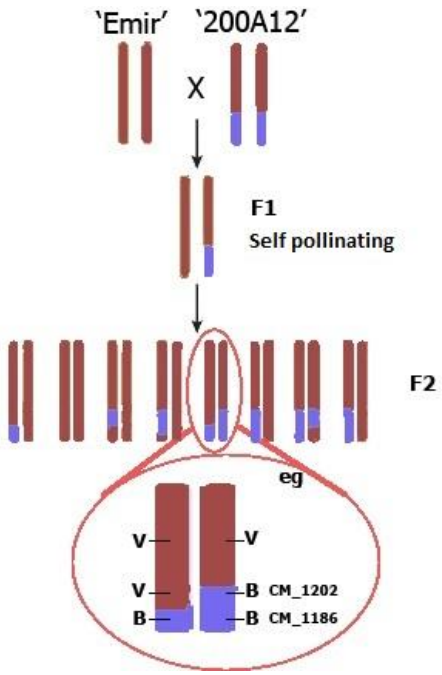


Figure 3.4 A breeding scheme illustrating the development of F_2 population and the possible recombinant chromosomes that could be identified by the validated CAPS markers. Initially the IL '200A12' was crossed to the barley cultivar 'Emir', the F_1 hybrids were self-pollinated and the F_2 progeny were genotyped using the validated CAPS markers for identification of recombinants. 'Emir' chromosome segments are shown in red and IL '200A12' chromosome segments are shown in blue.

HRM markers

There were 14 out of 48 candidates of barley HRM marker validated as co-dominant markers. Among these 14 co-dominant markers, one of them (CM_1158) was selected and tested across the F_2 population with 96 individuals. Although the HRM profile of this marker tested on samples with the 3 possible genotypes gave a clear discrimination between heterozygote and homozygote (Figure 3.2), the HRM profile of the same marker tested across the F_2 population with 96 individuals showed slightly drifted melting curves of samples with the same genotype (Figure 3.3). It could happen when testing HRM marker across a large number of samples with different DNA quality or quantity which could sometimes affect the genotyping results. Temperature calibrator could be used to improve the differentiation of melting curves. Fortunately, in this case, 3 genotypes could still be differenciated by visual inspection of the curve pattern and Tm.

Eight markers were unable to distinguish heterozygotes from one of the homozygotes. They could be similar to the CAPS markers of barley with no digestion on heterozygotes, which was caused by preferential allele amplification. Moreover, the two homozygous variants were indistinguishable in another six markers. Using unlabelled probes might help to improve the indistinguishable melting curves. This strategy is often used for amplicons that contain multiple informative SNPs which could decrease the sensitivity of mutation scanning by producing complex melting curves (Garritano et al., 2009; Smith et al., 2009).

There was high florescent level detected from the negative control of some of the markers which could likely be caused by the formation of primer dimer rather than contamination as it did not happen to the all the samples. In order to confirm what the products amplified in the negative control were, 7 co-dominant markers and 16 complex markers were run on an agarose gel for separation. This can also check whether target region was amplified during PCR producing single product. The gel picture of the 7 good co-dominant markers showed that a single small product was amplified only in all negative controls but not in other samples. It confirmed that there was formation of primer dimer, rather than contamination, causing the detection of high florescent level in negative control. Also, all 7 markers showed a clear single band with the expected size in all samples indicating that specific product was amplified from the specific target region. On the other hand, most of the non-validated markers with messy HRM profile had smear bands or extra small bands suggesting the complex HRM profile was due to the amplification of non-specific target region and formation of primer dimer.

3.2 Wheat

3.2.1 Background and objectives

One of the major challenges in plant genetics recently has been the understanding of the relationship between phenotypic variations and the underlying DNA variation. Recent advances in high-throughput sequencing technologies have allowed the discovery of large numbers of SNPs throughout the genome. These SNPs can be used for studying patterns of genetic diversity and associations between markers and traits by performing genetic analysis such as GWAS. Several high-density wheat SNP genotyping arrays have been developed in recent years (Wang et al., 2014; Cavanagh et al., 2013).

Previously a genome-wide association scan was performed with Illumina SNP chip genotyping on a set of wheat lines sown at different dates for studying genetic basis of the days to flowering in wheat. GWAS results show that there is significant association detected on chromosome 1A region, as shown in Figure 3.5.

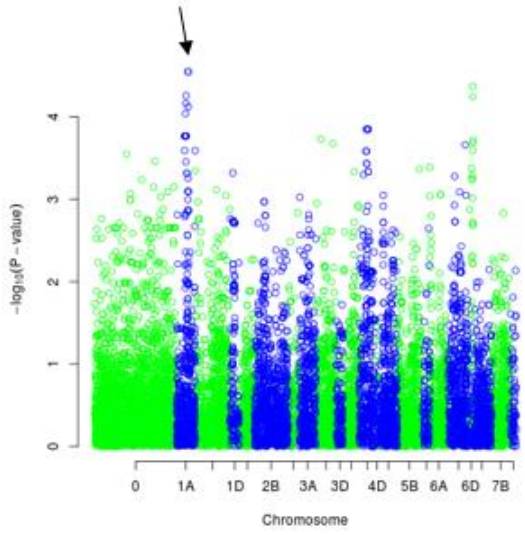


Figure 3.5 SNPs associated with the days to flowering in wheat identified by GWAS. Significant association detected on chromosome 1A region is marker with an arrow.

Two sets of SNPs were validated in the wheat experiment which are the SNPs on chromosome 1A that were suggested to be associated with the days to flowering in wheat, as well as a subset of random SNPs across the A genome. The goal was to determine the effectiveness of HRM as a SNP validation tool in this experiment.

3.2.2 Materials and methods

A total of 48 primer sets were designed including 12 QTL markers from chromosome 1A picked from 17885 SNPs and 36 random SNPs across the A genome selected from 81587 SNPs from Wheat Illumina SNP chip for SNP validation using HRM. All 48 markers were tested on 7 lines with one negative control. The 7 DNA samples provided by Plant and Food Research Lincoln were selected from the wheat flowering lines previously extracted and genotyped using the 90k Wheat Illumina SNP chip. PCR amplification and HRM analysis were performed on all the markers using the same methods and protocols as in the barley experiment mentioned above.

3.2.3 Results

Twelve putative SNPs of QTL markers from chromosome 1A and 36 random SNPs across the A genome were selected for validation. The results of the HRM analysis were then compared with the SNP chip clustering results obtained using the polyploid version of GenomeStudio (GS) software (Illumina). HRM profiles of all 48 SNP markers are given in Appendices C.5 and C.6.

QTL markers

Out of the 12 QTL markers, one of them showed a poor amplification (Figure 3.6a) and another 4 were not validated due to the overlapping melting peaks between the two genotypes (Figure 3.6b). The rest of the 7 markers displayed an HRM profile that matched the SNP genotypes from the SNP chip clustering with clearly different Tm or pattern of melting peak between the two genotypes (Figure 3.6c) (Table 3.3). Among the 7 best markers, several of them showed more complex HRM profiles with multiple peaks or bumps which might be due to the presence of additional SNPs that were not included during the primer design (Figure 3.6d).

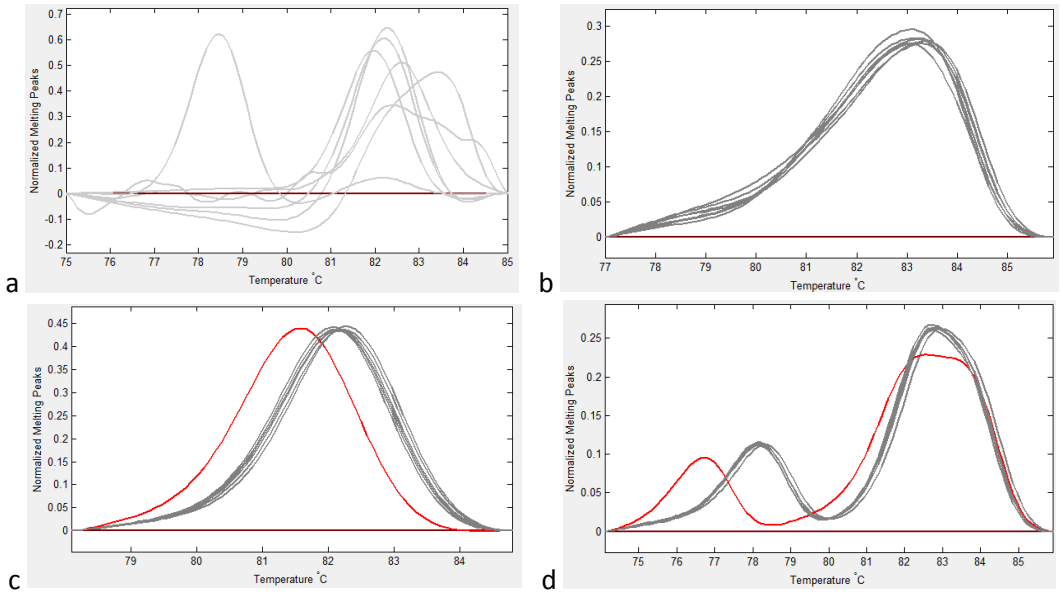


Figure 3.6 Examples of HRM profile of four markers: (a) CM_01210, (b) CM_01214, (c) CM_01215 and (d) CM_01217.

For further evaluation of selected SNP markers, 5 out of the 7 best looking markers were tested across a wider set of 93 genotypes (used in the original GWAS experiment). Four of them displayed clear separation of the melting peaks between the two genotypes (Figure 3.7a). However, one SNP marker (CM_01215) showed poor differetiation due to the drifting of two sets of curves to each other (Figure 3.7b). The normalised melting curves and normalised derivative melting curves of these 5 markers tested across a set of 93 samples are provided in Appendix C.7.

Number of SNP selected	12
Amplification	
Success	11
Fail	1
HRM results of the 11 amplified markers	
Validated - Matching with the SNP chip	7
Not Validated - overlapping melting curves	4

Table 3.3 Summary of wheat HRM QTL markers validation. Appendix B.3 shows a more detail summary of the result.

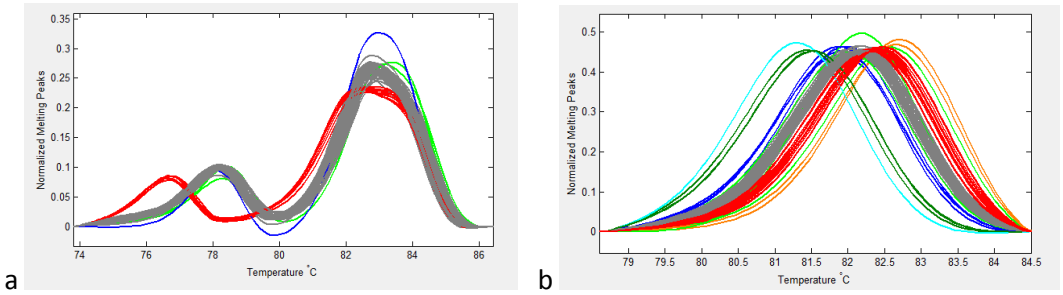


Figure 3.7 Examples of HRM profile of two best looking markers: (a) CM_01217 with clear indication of two genotypes, and (b) CM_01215 with drifted curves

Random markers across the A genome

Fourteen of the 36 candidate SNPs randomly selected across the A genome were positively validated. Among these validated SNPs, genotyping result of 11 markers matched the results from the SNP chip with clear differentiation between genotypes. An example of the normalised derivative melting curves and the SNP chip clustering of one of these 11 markers (CM_01230) are shown in Figure 3.8a and Figure 3.9a. It can be seen that the two well-separated clusters of this marker were perfectly matched by the two clearly separated sets of melting curves with a large difference of Tm. The other 3 markers have matching genotyping results but melting curves were complex (Figure 3.8b).

HRM profile of the 36 random markers, 10 of them showed that more than one genotype from the samples which were grouped in the same cluster on the SNP chip. These usually were wide and spread clusters which could potentially represent multiple variants (Figure 3.9c). Eight showed the same pattern of melting curves of all samples (Figure 3.8c) and matched the genotype called from the SNP chip which gave only a single cluster (Figure 3.9b). The last 4 candidate SNPs were not validated due to either showing messy HRM profiles (Figure 3.8d) or failing to match SNP genotype called from the SNP chip clustering (Table 3.4).

Number of SNP selected		36
Results matching with the SNP chip		14
Good clean HRM profile	11	
Complex HRM profile	3	
More than one genotype detected from one cluster		10
Only one cluster called on the SNP chip		8
Messy HRM profile or fail to match with the SNP chip		4

Table 3.4 Summary of wheat HRM random markers validation. Appendix B.4 shows a more detail summary of the result.

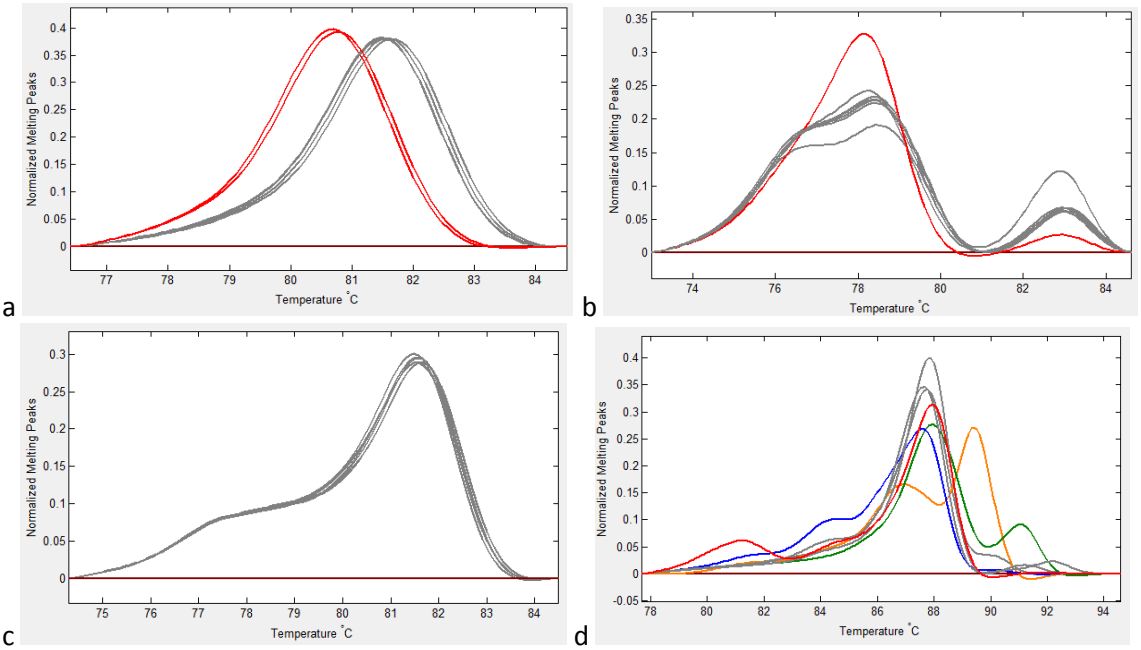


Figure 3.8 Examples of HRM profile of four random markers selected across the A genome: (a) CM_01230, (b) CM_01222, (c) CM_01252 and (d) CM_01247.

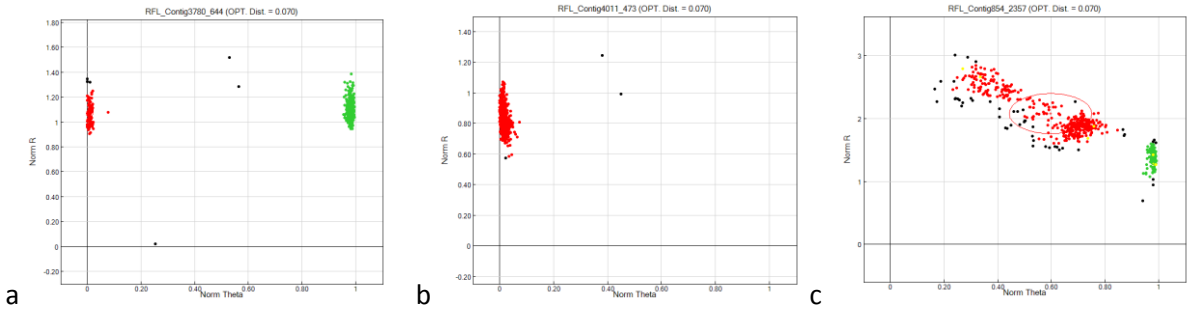


Figure 3.9 Examples of clustering from the SNP chip for three markers: (a) a validated marker (CM_1230) with two well-separated clusters, (b) a non-validated marker (CM_1224) with a single concentrated cluster and (c) a marker (CM_1225) with extra genotype detected from one cluster

3.2.4 Discussion

A 58% conversion rate of the 12 wheat QTL markers was observed in this experiment. The results from ePCR showed that most of the non-validated markers had no hits predicted. This was caused by the poor selection of marker as ePCR was not performed pre-validation due to the missing of wheat reference sequence. Furthermore, there were multiple hits predicted from ePCR for some validated markers with more complex HRM profile which suggests that there might be multiple products amplified from non-specific region. Whereas, two markers with double peaks or bumps on the melting curves resulted in a single hit from ePCR indicating the presence of underlying SNP in the amplicons. Resequencing amplicons of the validated markers might be useful to further confirm that the specific region was amplified from the target chromosome.

On the other hand, the 36 random wheat markers across the A genome had a lower conversion rate (39%) than the 12 wheat QTL markers. This could be due to the choice of SNPs with low minor allele frequency (maf), such as the 8 markers with a single cluster, for validation. Selecting SNPs with high maf could help to result in a higher conversion rate in the next SNP validation experiment.

Moreover, 4 of the 5 QTL markers tested across the 93 samples gave clear separated melting curves between genotypes detected which were with agreement of the genotypes called from SNP chip clustering. However, one of the 5 markers (CM_1215) with drifted melting curves was unable to differentiate between genotypes. Temperature calibrator could be used to improve the differentiation of melting curves. It is based on using a High Sensitivity Genotyping Mastermix (Idaho Technology Inc.) which includes both saturating dye and two sets of internal temperature calibration dsDNA frgments with low and high Tm. This allows genotyping with high sensitivity and accuracy.

Last but not least, a few individuals had slightly different pattern of melting curves and different Tm in all 5 markers tested across population which could be caused by low DNA quality of these samples.

Chapter 4

General discussion

The advances in next generation technologies has allowed the collection of large-scale sequencing data which can be used for advanced genetic analysis in a genome-wide fashion. Bioinformatics has rapidly become an important technology in many areas of biology for data storage and performing data-intensive analysis. However, reproducibility is one of the biggest challenges in the delivery of bioinformatics. One of the approaches recently used in delivery of bioinformatics tools is virtualisation which creates isolated environments for tools with different sets of dependencies. Virtual machine is one the most commonly used methods for creating virtual environments, yet this is a heavyweight virtualisation with OS installation required in each environment which is disk storage consuming. Developing virtual environments using conda or containerised environments using Docker are more promising and scalable approaches for delivering bioinformatics. Conda is an environment and package management system which allows users to create multiple isolated environments with different version of packages and their dependencies installed. Users can also use Docker to create and share containerised environments with required dependencies configured to run bioinformatics tools for performing scientific analysis.

Furthermore, using conda in a containerised environment could be a better approach for bioinformatics distribution. In this approach, conda is contributing as a package manager allowing simple installation of different packages and dependencies for building a containerised environment that can be easily shared on Docker or GitHub registry. The simplicity of conda also allows non-specialist to update dependencies of a Docker image which can be pushed up to registry for sharing.

In this dissertation, diagnostic PCR assays were designed using containerised software environments for bulk SNP validation and conversion in barley and wheat. Bulk marker selection and design of CAPS and HRM PCR assays were conducted in Jupyter notebook launched in a Docker container. Several of these markers of both barley and wheat were validated and converted to CAPS or HRM markers. In traditional marker design practice, markers are designed and validated individually involving intensive troubleshooting steps for ideal PCR amplification which is time and laboratory equipment consuming. Whereas, designing diagnostic PCR assays for bulk markers using automated marker designing tools is a more efficient approach requiring less time, labour and lab equipment to develop validated markers.

Several strategies could be used to improve the outcome of bulk marker designing and screening. Firstly, performing ePCR pre-validation could help to identify and discard the poor choices of marker.

However, ePCR was performed post validation of wheat markers in this experiment, as we did not have reference sequence of wheat until late of the dissertation. Secondly, using other HRM techniques, such as temperature calibrator or unlabelled probes, to conduct highly sensitive and accurate of HRM assays could produce more significant results.

For the improvement of this approach in the bioinformatics aspect, conda packaging of the PCR marker designing tool used in this dissertation could benefit non-specialists to easily install the tool in a running container by executing a single conda command. In addition, although the design software was improved to handle melt prediction of multi-SNP amplicons, breaking down the code of the software into smaller pieces for further testing could be done to ensure correctness.

Both CAPS and HRM assays are suitable methods for validating and genotyping large-scale candidate SNPs. Both methods allow identification of co-dominant markers with distinguishable homozygous and heterozygous genotypes (Ramkumar et al., 2015; Han et al., 2011). CAPS is a robust and simple technique for SNP validation. It is reproducible that can be easily shared between laboratories as it only requires basic laboratory equipment for performing PCR and gel electrophoresis (Baldwin et al., 2012). However, using CAPS markers could be labour-intensive, time-consuming, and possibly costly if unusual digestion enzyme is required (Ramkumar et al., 2015). By contrast, HRM is a simple SNP validation tool which does not require post PCR separation step and restriction enzyme. It had a higher success rate of SNP conversion compared to CAPS observed in this dissertation. On the other hand, HRM is sensitive to DNA quality that partially degraded DNA could produce inconclusive results (Lehmensiek et al., 2008). Samples with non-uniform DNA concentration, pH or salt could affect the HRM performance leading to inaccurate results. HRM could also be unsuitable to genotype SNP with low difference of Tm, such as an A/T polymorphism.

Chapter 5

Conclusion

This dissertation has evaluated the bioinformatics toolkit previously developed for bulk PCR-based marker validation. It was done by using Docker for creating containerised software environments to design and screen CAPS and HRM assays for SNP validation and conversion on barley and wheat.

Containerisation is a scalable approach for delivering reproducible bioinformatics tools as it provides a lightweight platform for creating and sharing containerised environments with configurations set up. Using conda, which is a package and environment manager, in Docker containers allows simple installation of tools and packages for building or updating these containerised environments which can be easily shared to colleagues or public through Docker or GitHub registry. Although this marker designing toolkit is available on GitHub, packaging it into a conda package would make installation easier which could benefit non-specialists.

About 30% of the CAPS and HRM markers of barley were validated as co-dominant markers. One possible recombination event was detected using one of these validated markers, yet further confirmation is required. In wheat, 58% and 39% of the QTL markers on chromosome 1A and random SNP markers across the A genome were validated and converted into HRM markers, respectively. Performing in silico PCR post primer design and using other HRM techniques, such as temperature calibrator or unlabelled probes, could improve the conversion rate and produce more significant HRM results.

Appendix A

Lists of primer sets designed for SNP validation

A.1 List of 21 PCR primer sets designed for barley CAPS markers

Primer Name	Forward Primer	Reverse Primer	Enzyme	Reference Base	Variant Base	Amplicon Size
CM_01186	AACATGGTGCAGGTTTGGTC	TGACTGAAGCCCGACTATCC	AluI	T	G	151
CM_01187	GTACAGGCGGTTGTTGTCTC	TCTTTCATTCCGGGCTACCC	HinfI	G	C	213
CM_01188	GAAAGTGCGGTGCTTTCTTG	CATCCTCCTCCTCGTCGTAG	HindII_TaqI	G	A	240
CM_01189	CTACAGGCTAGGAGACGTGG	TCGACTTCTCCACGTAGCTC	TaqI	T	C	208
CM_01190	ACGCTCTTCATCACGTCCTC	CAGAACGTCACGGGCCTC	DdeI	G	A	187
CM_01191	GTGCTCAAATAACTTGGCCAC	CTATCACCACACGACCAAACC	TaqI_TaqI	A	G	216
CM_01192	TGGTCTGAACAACAATGGCG	AAGGGTTGGGCTGTCAAATC	DpnII_Sau3AI	C	G	228
CM_01193	GATGGCGGCTTGGAAGTAAG	CCTGTCCGATCGATGCAAAG	HinfI	T	C	222
CM_01194	GCGCGCTGTGATGCATATAC	TGTGGAATGGAGTGAGTACCG	HinfI	C	G	227
CM_01195	CATCTGCTTGTAGCTGCACG	GGCAAGACAATGTCCTTGGG	AluI_PvuII	G	C	185
CM_01196	GACTTCGAGGGCCTCTTCTC	CTTCCTCCTTCTCCGGGC	HinfI_TaqI	C	G	183
CM_01197	TGTACGAGTGCGAGAAGGAG	CTTCTGGAGCGAGCATTCAC	DpnII_Sau3AI	C	G	163
CM_01198	TCTTCTTCGTCCCATCTCCG	ACTGTAAAGCTCGTTGGCTG	RsaI	A	G	230
CM_01199	TTGTGCCTTCTCCCTGAGAG	TGCCGAAATCAGCAAGACAC	RsaI	C	G	163
CM_01200	TTTCACGATCGAAGTGGCAC	GAGCTCAACAGTGCAGGTTC	AluI	A	G	238
CM_01201	AAACACAAGCTTGGCGACAG	GCCCTGCTGATATGTTGCTC	RsaI	A	G	224
CM_01202	ACCCAACCCATAGGAAGCTC	AAGCAGAGTACTCCCTTGGG	HindII	C	G	243
CM_01203	NCCACTTAACCATACCTCACC	AAGAATGACGTCCATCCTTGC	DdeI	C	G	249
CM_01204	GTCGGTCGCGAGTTATTCTG	CGATGCCCGTCTTCTCTTC	TaqI	T	C	124
CM_01205	AGTTTGGCTTCGAACCAAGC	TCACACGGTCGGAATGGAC	DdeI	G	A	243
CM_01206	AGCCAGTCACAAATCGCATC	CAAAGCCAGCCTCCAAGC	DpnII_Sau3AI_ClaI	A	G	148

A.2 List of 48 PCR primer sets designed for barley HRM markers

Primer Name	Forward Primer	Reverse Primer	Reference Base	Variant Base	Amplicon Size	Reference Tm	Variant Tm	Tm difference
CM_01138	GGAACTGACTCCAAGCACTG	CCCTAGCGCGACAAAGATG	G	A	86	92.1	91.4	0.7
CM_01139	AACATGGTGCAGGTTTGGTC	CGATATCTGAACCTGCAGGC	T	G	103	91.15	91.8	0.65
CM_01140	GTACAGGCGGTTGTTGTCTC	TCCATTGCTCAATGCTGCAG	C	T	97	88.75	88.25	0.5
CM_01141	GTCAACGCTACAAGGAAGGG	TCAACCCACAGAGCAAATGAC	T	C	116	87.8	88.3	0.5
CM_01142	CTACGACGAGGAGGAGGATG	CCCTGACCAAAGCACAAGAG	C	T	102	93.55	92.75	0.8
CM_01143	CAGACGAGCTCCAGGTAGG	GCACTGCTACCTCCTCGAG	C	A	89	93.75	93.2	0.55
CM_01144	GTGCTCAAATAACTTGGCCAC	CGTTGCTGCAGATTTCGTTG	A	C	82	82.3	82.95	0.65
CM_01145	GCAGGTAAGCCTCTGCAAAC	GGCCACCACCATCATCAAC	G	T	102	88.1	87.6	0.5
CM_01146	CTTGCTCTGGATCTGCAGC	GATCAAACTCCGACACTGCC	G	A	100	86.35	85.85	0.5
CM_01147	GATCTGCAGCATGGCGTC	ATCATGAGCAAGGACCTCGG	T	C	107	94.95	95.45	0.5
CM_01148	GCCCTACTTGTTAGAAAGACACC	CTGCAGCAGAGTCCTTAACG	A	G	115	84.25	84.75	0.5
CM_01149	TGTTCCAGGACTGCCTCATC	GAACCACCGAGATGTGCTTC	T	C	93	93.75	94.3	0.55
CM_01150	CCTTGTCAACGCAGAGGTAG	CAGAAGCTGGGAGTAATGCC	A	C	90	92.5	93.3	0.8
CM_01151	TGATGAATGGAAGGAGCTACTAC	CAGCACTATGAATGTAACAAGTCTG	T	C	89	82.05	82.7	0.65
CM_01152	CTGACTGCAGCTGGTTAACC	ATGTCTGGACACCGGAAGAG	A	G	104	85.7	86.2	0.5
CM_01153	GGATTAGTGTTTGGACTCATCCC	CCCAGCTTCCTATCCTGGTAG	A	G	86	84.75	85.3	0.55
CM_01154	TGCAGAGGCCAGAGTTATCCTC	GTGGCGAGTGGAGGTGGG	G	T	119	90.15	89.25	0.9
CM_01155	ACTTGAGGCCCGTTAAATGG	ACCCAGACTGATGAGAGTCC	T	C	83	86.5	87.05	0.55
CM_01156	GAGCTCATGGAGGAGTCGG	AACTGTCGACGAGGAGGAG	C	T	119	93.95	93.4	0.55
CM_01157	ATCTCCATGATGTCCGGCG	ATGAGCTCCTTCTCGCACTC	A	G	94	93.8	94.35	0.55
CM_01158	CAAACGGATCAGCTAGCCAC	TCAACGAGAGGCTCAAGGAC	G	A	118	92.75	92.25	0.5
CM_01159	GCCTAGCTTGGTGTTGAAGG	AATGGGATCAATGCGTCGTC	A	G	79	87.3	88.1	0.8
CM_01160	TCCCATTCCGGTGTATGTCC	GAGTTGATGAAGCGGTCCTG	A	G	118	90.2	90.7	0.5
CM_01161	TCTTCTTCGTCCCATCTCCG	TAAGGCTGCAGCTAGTCGTG	A	G	96	89.7	90.3	0.6
CM_01162	GATGGTGAAGCCTGGTTTCC	AGATGACGATGATGCACAGC	G	A	80	88.25	87.7	0.55
CM_01163	CACCTGAAATCTGCCTCTGC	TCACATCGACATCAAAGCCG	A	G	110	91.6	92.2	0.6

CM_01164	TGCTGTCACATTGCGAAGAC	ACAAAGGCTCCTCCTACCTG	C	T	81	89.25	88.75	0.5
CM_01165	TCACTCGCCTTGCAAAGATG	ACGGAGGGAGTACATTGCTG	A	G	83	90.2	90.75	0.55
CM_01166	AACACAAGCTTGGCGACAG	TGGCAGGTTCCATTTGCTAC	C	T	96	91.65	90.95	0.7
CM_01167	GCTCAAACATCCACACTTTGC	TCGCCAAGCTTGTGTTTAGC	T	C	117	93.3	93.8	0.5
CM_01168	ATCTACGACCTCGAGCTCAC	GATCGATGCTGCAGTCGATG	A	C	114	93.2	93.8	0.6
CM_01169	CATCCGGCTGCAGATTTCTC	GGAGAACGTGTGGCTGAAAG	C	T	114	93.5	92.8	0.7
CM_01170	TGATTCTCTCCAGCCGACTG	GATCTTGTTGTGCGACTGGG	C	A	114	90.7	90.1	0.6
CM_01171	TGAGCGTTTCATTGAGATGATAAG	TCTTGTACTTGGCATACGGAG	C	A	88	82.6	81.9	0.7
CM_01172	CTGGCACATTGGTCAGACTC	GGTAAACACCTGGAGAAGTGG	C	T	79	86.15	85.45	0.7
CM_01173	GCGTTGTTGGAGATGTCGAG	GCTCTACCTCCAGACCAACC	C	T	92	95.25	94.4	0.85
CM_01174	GGTACACATTGGAAAGTGCAAC	ACGACAGCTTAGGTTTCTTGC	A	G	104	86	86.65	0.65
CM_01175	CTCGCATCGAAAGACCCATC	CAAAGCCAGCCTCCAAGC	G	T	103	95.15	94.5	0.65
CM_01176	CATCCGCTGCAGCCAGTC	TGCTGGATGGGTCTTTCGATG	G	A	79	89.2	88.65	0.55
CM_01177	TGTTGTGGTTGCTAACGTTTC	AAAGGGTCTTCCAAGGATGAC	T	C	93	84.1	84.65	0.55
CM_01178	CGTAGGCGCCACACTACAC	GTCACACGGTCGGAATGGAC	C	T	93	94.65	93.95	0.7
CM_01179	CAGTGCTTGGAGTCAGTTCC	GACTTTCGTGGGAATCAGCC	T	C	93	90.35	91.1	0.75
CM_01180	AGCTTGGTCTACTCCCATGG	TCAGGTAGTGGCTACTGCAC	T	C	83	87.65	88.15	0.5
CM_01181	GTGAGCCACAGGGACTCATC	GCTGCACCTCACCACTTAAC	T	C	90	84.55	85.2	0.65
CM_01182	CCGTTGCAGTTAGTACGTGG	TGTGGTGGTGAGATCGGTAC	T	C	88	88.3	89.05	0.75
CM_01183	GCTGCCATCATAATACCTTGC	GGGAAGCCAAAGAATATTCTCAAG	T	G	110	83.6	84.1	0.5
CM_01184	AGGGTCTTTCGAGGATGACG	NGATGTGCCTCTTGTGGTTG	G	A	100	85.95	85.35	0.6
CM_01185	GGTGACTCACCGTGGCAC	TTCAGAGCGTCCCAAACCAC	C	A	85	94	93.45	0.55

A.3 List of designed 12 PCR primer sets of wheat QTL markers on chromosome 1A

Primer Name	Forward Primer	Reverse Primer	Reference Base	Variant Base	Amplicon Size	Reference Tm	Variant Tm	Tm Difference
CM_01207	GCC AGC ACT TGA ACT TCT CC	ACC AAC ATC ACC ATT CGA CC	T	C	63	84.95	85.7	0.75
CM_01208	TGG GTG TTG ACT GGA ACA AC	TCT ATC GAC GTG TTG ATG GC	T	G	60	84.65	85.65	1
CM_01209	CAT GTT CGT TGC TCA ACA TGC	TTC TAT CGT CAT GTT GCG GC	T	G	62	83.75	84.65	0.9
CM_01210	CAA CAA GTT TAG TTG GAT CAA ATG G	ACC ATC AAC AAC GAT AAG TGT AAC	T	C	67	78.9	79.55	0.65
CM_01211	GCA GTC AAA GGA ATC CAC CC	GGG TCA ACC TTA TCT GCG TC	A	G	95	86.85	87.2	0.35
CM_01212	TTT GCA GCC TCT TCG AAA GG	TGG GCA AGC TGC TGT ATA TTC	T	C	74	86.15	86.5	0.35
CM_01213	ATT TCC CTC CTT GTT GCA GC	GGT AGA GTT ACA TTC TGC TGT GC	A	G	87	89.85	90.4	0.55
CM_01214	GAT ACT CAG CCA CCG GTA GG	TAT TGG CCT ACA GGG TGC TC	A	G	97	87.85	88.2	0.35
CM_01215	TTG TGC ACT TGG TTC TTC GG	CAG CCC AGC TCG AAT GTT TC	A	G	65	84.8	85.3	0.5
CM_01216	ACG CTT GCT TCT GGC TTA TG	ATG CGT TCT CTT CTG GCA TC	T	C	106	86.2	86.6	0.4
CM_01217	CAT GTA GTT GAG GGC ATG GG	ACT TCG ACA TCT CTG TGC TC	T	C	90	85.25	85.8	0.55
CM_01218	TGC TAC TTC CGC AAA CAA CC	AGT GAG GCG AAT GAA CCC TC	T	C	104	91.8	92.2	0.4

A.4 List of designed 36 PCR primer sets of random wheat markers across the A genome

Primer Name	Forward Primer	Reverse Primer	Reference Base	Variant Base	Amplicon Size	Reference Tm	Variant Tm	Tm Difference
CM_01219	TAC AAG AGA CCG AGT GCT GG	GAT GAA GAT GGC GCT GGT G	A	C	73	89.95	90.8	0.85
CM_01220	TCT TTG CTG TGA TGA CCG TG	GCC GAA GAG CGA TAA GAA CG	T	C	77	88.8	89.7	0.9
CM_01221	CTC CCT CAG ATT GAG CTC CG	TAG ACG CGC ACA GAT CGA TG	T	C	66	88.45	89.5	1.05
CM_01222	AAA CAG ATA GTT CAT CAC GTT TGC	TGA CCA TTA TGG ACA AAC CTA ACA G	T	C	60	81.7	82.6	0.9
CM_01223	TGA GAA GTT GTT CCG TAC ATA TCC	CAC AGA TCC GGT ATT TAT CTA CAC C	T	C	67	82.2	82.85	0.65
CM_01224	CCA CAG CAG ATT CCA CTC AAG	TTG TCA GCC TCT GCA GGA TC	T	C	78	85.7	86.55	0.85
CM_01225	CCT TAT CTT CAA CCT AGG AGG C	ATC CAT CAC ACT GCA CAA GG	A	G	74	82.85	83.45	0.6
CM_01226	AAT GGA CTA GAT GCG GGA GG	GGA GGA CAG CAA TCA TTC TGG	T	C	86	87.65	88.25	0.6
CM_01227	CCT ACA AAG AAG TTG CCA CCC	GGT ACC ACC AAT TCT CTG TAC C	T	C	71	83	83.6	0.6
CM_01228	GAG CTG GTC CAC CTC CTC	CCC TGG CAA ACT CAC AGC	T	C	49	84.8	85.7	0.9
CM_01229	CAT GAA CAT CGT GCT CGG G	GAA GGA GAC CGG GAT GTA GC	T	C	51	87.7	88.8	1.1
CM_01230	TGC GTG TGA TGC GTA CTA AG	GAG CCA CCC TTG ATT AAC CC	T	C	58	84.65	85.7	1.05
CM_01231	TTC GTC GTG AAA TAC TCC ACA C	TGC AGT ATT TGA GCG CAC TG	T	C	62	84.65	85.3	0.65
CM_01232	ACA TGA AGC AGA GTG AAG CG	CCT GCT CCT GTC AGA TCC AG	T	C	65	86.4	87.25	0.85
CM_01233	GCC AAA CAA CAA CAT GCT CC	AGA AAC TGA AAT GCC ACG CC	A	G	60	83.75	84.45	0.7
CM_01234	AGG GTT GGT TTA GCA TTG CC	TGA TGT GGT TCA TGG CAT GG	A	G	69	86	86.6	0.6
CM_01235	AGG TAT CGA CTC TAA CGG CG	CTC TGT TCT GCT GCG CTT C	T	G	73	87.95	88.8	0.85
CM_01236	AGC AAC TCC CAC CTC CAC	AGA ATC CTG GAA GCT GGA CG	T	C	93	94	94.6	0.6
CM_01237	AGT GAC GAT GAC CAA AGT GTA AG	CTT CAT CGC TCA CTG ATG CC	T	C	78	84.15	84.95	0.8
CM_01238	TTT GAC TAT TAC CCA ATG ACT CAA C	GGA TGG TGA TAA TGC TGT TTC AC	T	C	73	84.25	84.95	0.7
CM_01239	CTA CAT CGG CTA GGG CTA GG	ACG GAG AGC ATG CAT GTA AC	T	C	81	90.55	91.75	1.2
CM_01240	GCC ATC GAT CAT ACT GCT GC	TAA GGG AGA TAG CAG GCA CG	T	C	66	87.05	87.75	0.7
CM_01241	GGG TTG GGT GTC AGC AAC	GGT GAT GGC TCT TTG AGT GAG	T	G	63	83.8	84.7	0.9
CM_01242	TCT TGA ATT CTT CTT CGA ACT CGG	TTT GTT GAA GAC TGC AGC GG	T	C	67	84.2	85.15	0.95
CM_01243	GAA GGC GTA CAG AAC TGC TC	CTC CCT CTT GGA CGC AAT TG	T	C	69	83.95	84.85	0.9

CM_01244	GGT AGA GGC TGC CGA GAG	TAC CTC TTG GAC ACT GCA CC	T	C	71	90.15	90.85	0.7
CM_01245	TCG TTC TCC TCC TCT TCC AC	AAG CAA CCT TTG GTG CTC TC	T	C	111	88	88.75	0.75
CM_01246	GAA GAT GCT CCT GGA GTG GG	AGC TCC TTG TAC GTC GCC	A	G	75	94.2	94.75	0.55
CM_01247	CAG CCG CCA GAA CTT CAT C	CGA CCG CAA TCC TTA AGG TC	T	C	66	87.3	87.95	0.65
CM_01248	GTA TGC ATG CCT GAG ACC AG	GCA AAG CTG GCT GGA TCT AG	T	G	87	88.25	88.8	0.55
CM_01249	TCC TAC AGG CTC ATC CAT GG	AGG TAT GAA AGC TGC TTC CAT C	T	C	73	83.85	84.5	0.65
CM_01250	TGA ATT CAT GAT CAT GTT CTC TCG	CCC TCT CTT CAA GTG ATT TAT GC	A	C	73	79.65	80.3	0.65
CM_01251	CCC AAG GAG GCA TTA TAG ATT GTA C	ACA TGG CAG CCA TCT GTT TC	A	G	72	82.15	82.7	0.55
CM_01252	TCC ATA TGA TTC ATC ATC TCG CAT C	AAG AAG CGC AGA GGA CCC	T	C	76	85.2	85.95	0.75
CM_01253	GTG GTA TCC TGT TTG CCG TG	TGT CAT CAC ACT TGC CAC AAG	T	G	89	83.8	84.35	0.55
CM_01254	CAT CCT GAA CAA CCT TGG GC	GGT TTC CAG TGT TCT CGA GC	T	C	87	86.65	87.25	0.6

Appendix B

Summary tables for SNP validation

B.1 Summary of 21 barley CAPS markers validation

CAPS_name	Amplification	Digestion	Co-dominance
CM_1186	**√**	**√**	**√**
CM_1187	√	√	√/X (partial digestion)
CM_1188	x failed to amplify target region		
CM_1189	x fail BB	√	
CM_1190	x multiple bands	√	
CM_1191	x multiple bands	x no cut	
CM_1192	**√**	**√**	**√ (but digested Hb allele on VB lines showed very weak band)**
CM_1193	x multiple bands	√	
CM_1194	√	x only cut BB	
CM_1195	√	x only cut BB	
CM_1196	X fail all	/	
CM_1197	**√**	**√**	**√**
CM_1198	√	√	X (one of the alleles on VV was cut)
CM_1199	**√**	**√**	**√**
CM_1200	√	x only cut BB	
CM_1201	**√**	**√**	**√**
CM_1202	**√**	**√**	**√**
CM_1203	x multiple bands	√	
CM_1204	√	x no cut	
CM_1205	√	x no cut	
CM_1206	√	x only cut BB	

B.2 Summary of 48 barley HRM markers validation

HRM_name	Validation	Reason	Clear markers selected for gel eletrophoresis	SB= single band;	Complex markers selected for gel eletrophoresis	SB= single band; MB=multiple bands; PD = primer dimer; SMB=smeared band
CM_1138	x	overlapping			√	SB
CM_1139	x	very complex			√	MB
CM_1140	x	only 2 genotypes detected				
CM_1141	√		√	SB		
CM_1142	x	very complex			√	SMB
CM_1143	√ complex					
CM_1144	√ complex				√	SB
CM_1145	x	only 2 genotypes detected			√	SB
CM_1146	x	very complex			√	SB PD
CM_1147	x	very complex			√	SMB
CM_1148	√					
CM_1149	x	only 2 genotypes detected				
CM_1150	√					
CM_1151	x	only 2 genotypes detected				
CM_1152	√					
CM_1153	x	only 2 genotypes detected				
CM_1154	x	only 2 genotypes detected				
CM_1155	√		√	SB		
CM_1156	x	only 2 genotypes detected			√	SMB
CM_1157	x	only 2 genotypes detected			√	SB
CM_1158	√		√	SB		
CM_1159	x	very complex			√	SB
CM_1160	√		√	SB		
CM_1161	√ complex				√	SB

CM_1162	√		√	SB		
CM_1163	x	only 2 genotypes detected			√√	SMB
CM_1164	x	only 2 genotypes detected				
CM_1165	x	only 2 genotypes detected				
CM_1166	x	only 2 genotypes detected				
CM_1167	√					
CM_1168	x	only 2 genotypes detected				
CM_1169	x	very complex				
CM_1170	x	overlapping				
CM_1171	x	overlapping				
CM_1172	x	overlapping				
CM_1173	√		√	SB		
CM_1174	x	overlapping				
CM_1175	x	very complex			√	SMB
CM_1176	√		√	SB		
CM_1177	x	overlapping				
CM_1178	x	very complex			√	SMB
CM_1179	x	overlapping				
CM_1180	x	overlapping				
CM_1181	x	overlapping				
CM_1182	x	overlapping				
CM_1183	x	overlapping				
CM_1184	x	overlapping				
CM_1185	x	only 2 genotypes detected				

B.3 Summary of 12 wheat HRM markers validation along with ePCR result

HRM_name	SNP_name	Validation	ePCR (hit)
CM_01207	CAP11_c1972_285	√	4
CM_01208	Ex_c21450_396	√	2
CM_01209	IACX219	√	1
CM_01210	Ku_c2898_783	x	0
CM_01211	Kukri_c52952_315	√	3
CM_01212	Ra_c10580_1629	x	0
CM_01213	RAC875_c49760_107	√	1
CM_01214	wsnp_Ex_c10595_17291999	x	0
CM_01215	wsnp_Ex_c1427_2736441	√	1
CM_01216	wsnp_Ex_c42282_48900922	x	0
CM_01217	wsnp_Ku_c2815_5317230	√	3
CM_01218	wsnp_Ku_c30921_40705731	x	1

B.4 Summary of 36 wheat HRM markers validation along with ePCR result

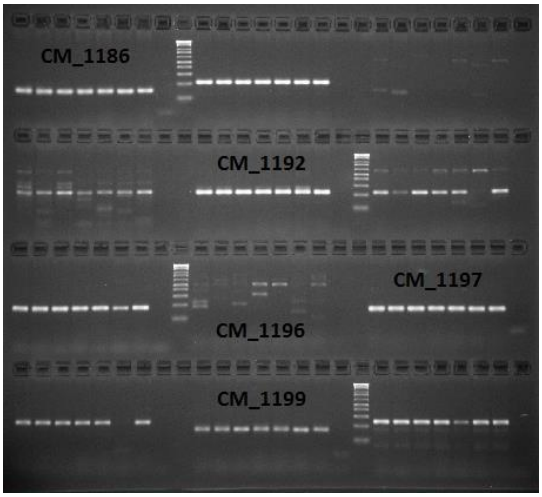
HRM_name	SNP_name	Validation	Reason	ePCR (hit)
CM_01219	BS00062735_51	√		1
CM_01220	BS00096519_51	√		1
CM_01221	Excalibur_c35312_109	√-	good complex	2
CM_01222	Kukri_c27599_1258	√-	good complex	4
CM_01223	RAC875_c754_120	√		2
CM_01224	RFL_Contig4011_473	x	single cluster	0
CM_01225	RFL_Contig854_2357	x	extra group	3
CM_01226	wsnp_Ex_c55245_57821389	√		3
CM_01227	Ku_c17678_1161	x	messy	3
CM_01228	Kukri_rep_c68559_549	x	single cluster	2
CM_01229	RAC875_c68649_457	x	extra group	3
CM_01230	RFL_Contig3780_644	√		1
CM_01231	Excalibur_c26923_569	x	single cluster	3
CM_01232	Excalibur_c63753_211	x	extra group	12
CM_01233	Kukri_rep_c69614_1326	x	extra group	7
CM_01234	RAC875_c63833_145	x	single cluster	3
CM_01235	Tdurum_contig55610_784	x	extra group	2
CM_01236	BobWhite_c11512_157	√		1
CM_01237	Kukri_c44260_577	√		1
CM_01238	Kukri_c46057_646	√		3
CM_01239	Tdurum_contig8404_683	x	extra group	3
CM_01240	BobWhite_c26122_129	√		1

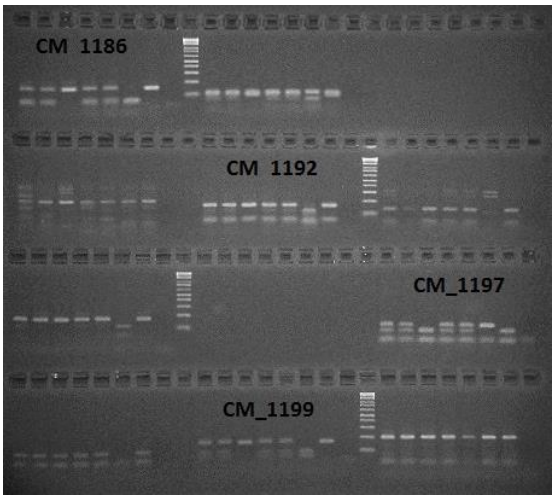
CM_01241	BobWhite_c7114_237	√		1
CM_01242	BS00067456_51	√-	good complex	5
CM_01243	RFL_Contig2531_969	x	extra group	4
CM_01244	tplb0050h15_1007	x	single cluster	8
CM_01245	wsnp_Ex_c7829_13320760	x	not matching	1
CM_01246	Tdurum_contig25539_248	x	extra group	5
CM_01247	BS00010796_51	x	messy	0
CM_01248	BS00022395_51	x	extra group	0
CM_01249	D_contig38730_358	x	single cluster	6
CM_01250	Excalibur_c854_1459	x	not matching	4
CM_01251	RAC875_c1467_1195	x	single cluster	2
CM_01252	RFL_Contig5101_350	x	single cluster	3
CM_01253	wsnp_Ex_c5060_8985678	√		1
CM_01254	BS00066739_51	x	extra group	4

Appendix C

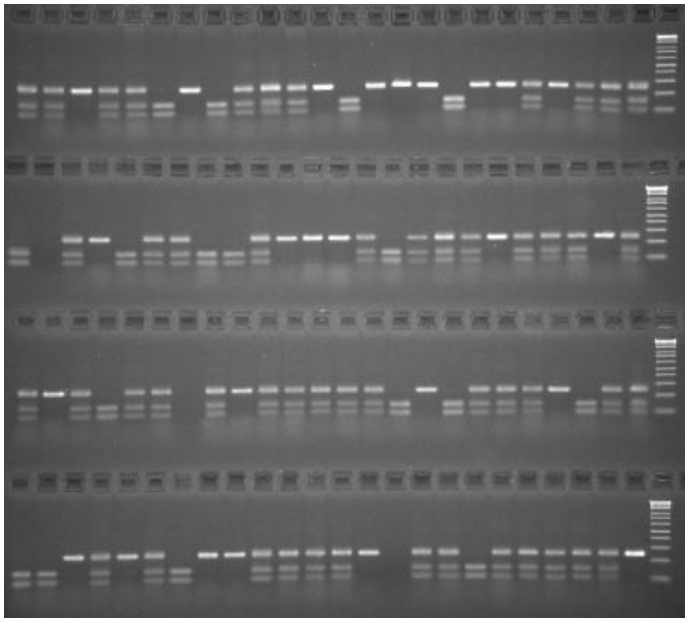
Results of SNP validation

C.1 Gel pictures of (a) PCR amplification and (b) RE digestion under condition 55°C annealing temp and 2.5mM Mg. Four of the six co-dominant CAPS markers, CM_1186, CM_1192, CM_1197 and CM_1199 are shown on both gel pictures. CM_1196 which is the only marker failed to produce clear bands is also shown on the PCR amplification gel picture.

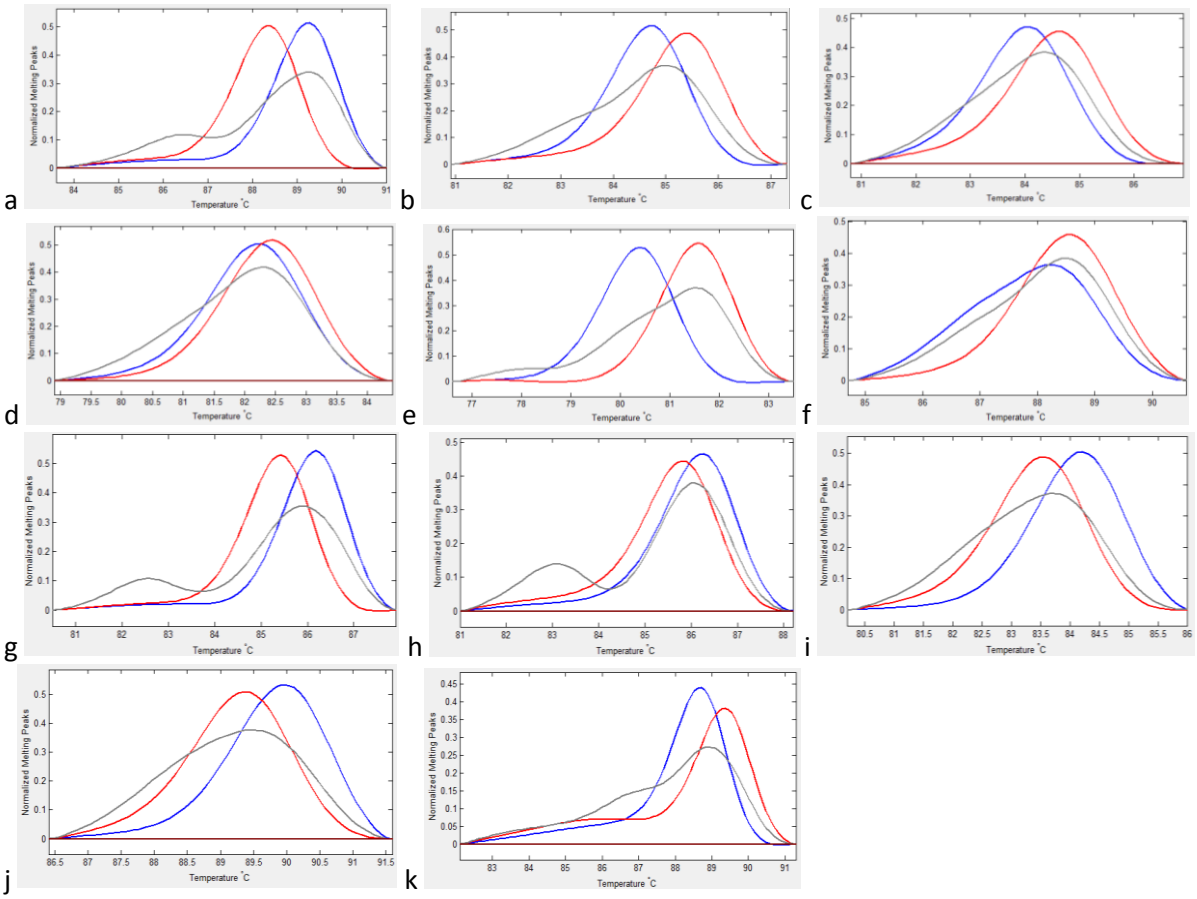
a 

b 

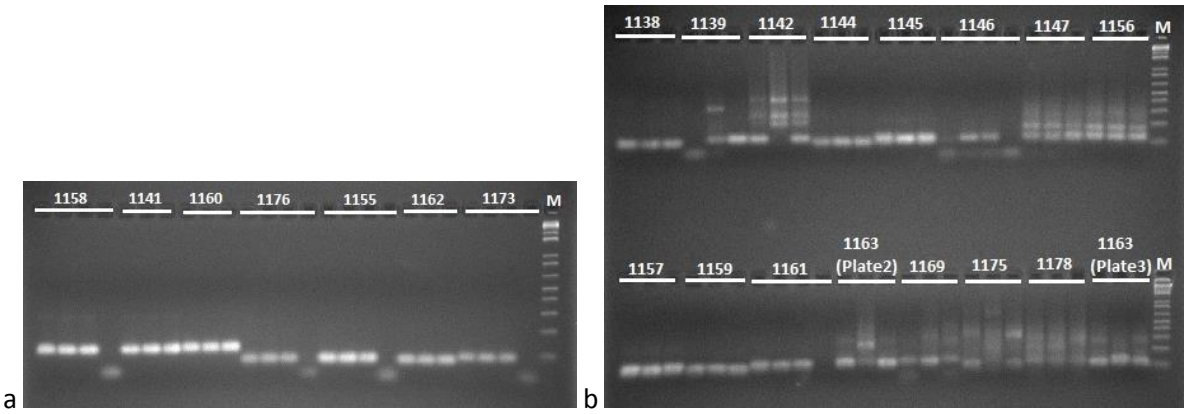
C.2 Example of one of the best barley CAPS markers (CM_1202) tested across the first half of the F2 population



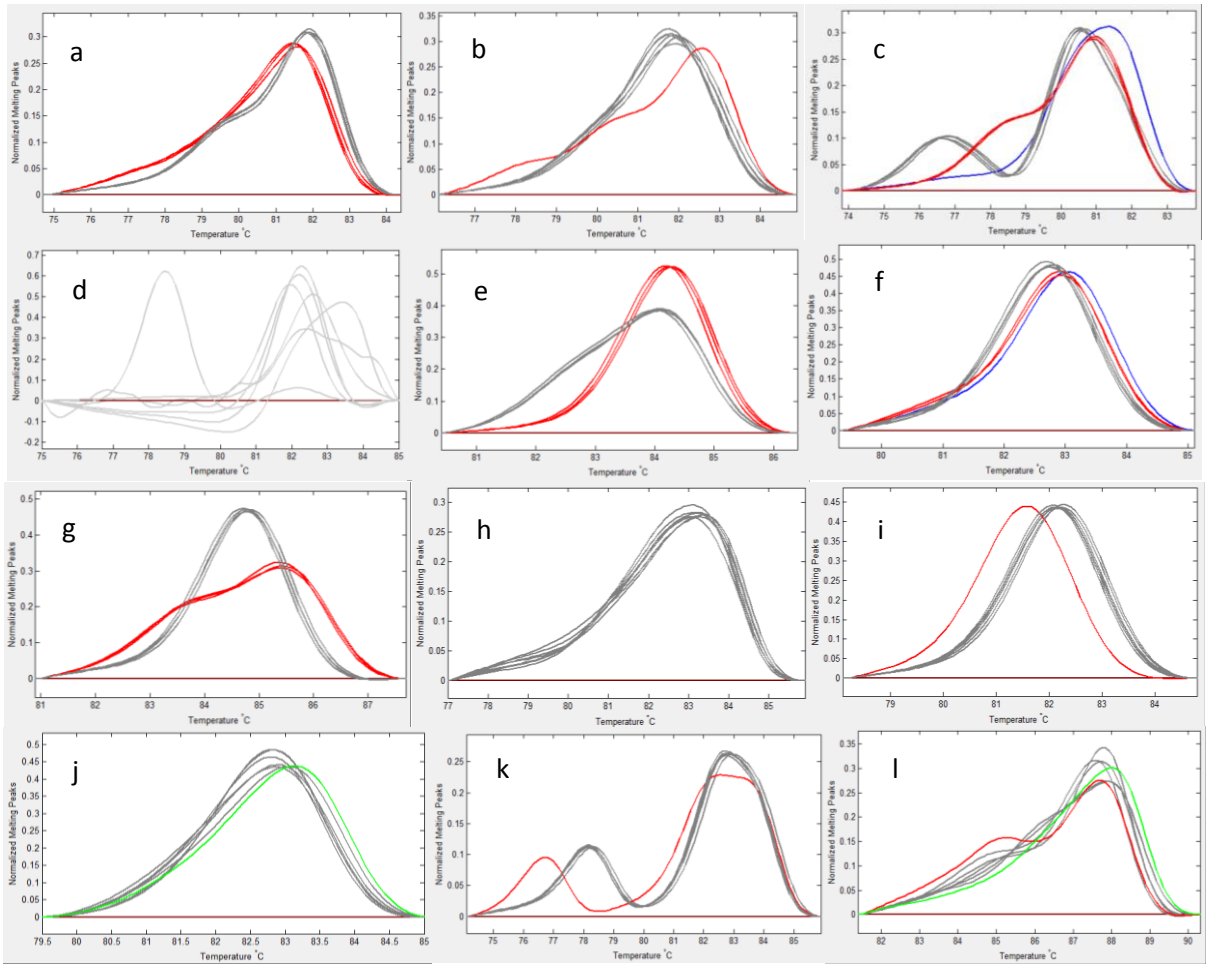
C.3 HRM profile of the 11 co-dominant barley markers tested on 3 DNA samples. (a) CM_1158, (b) CM_1141, (c) CM_1155, (d) CM_1152, (e) CM_1148, (f) CM_1150, (g) CM_1160, (h) CM_1176, (i) CM_1162, (j) CM_1173 and (k) CM_1167



C.4 Gel separation of (a) 7 of the good HRM markers for barley and (b) 16 of the non-validated HRM markers for barley

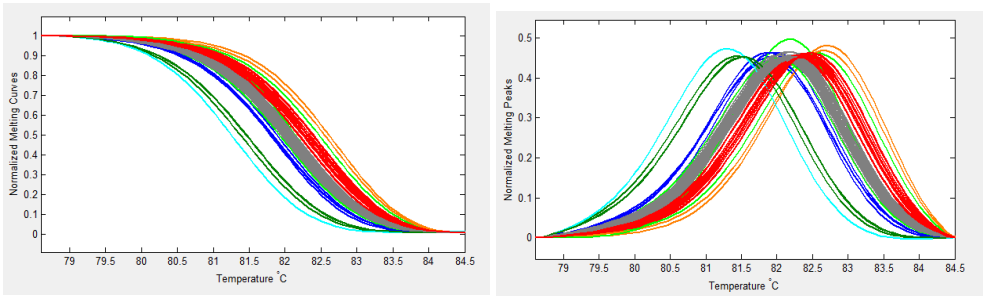


C.5 HRM profile of the 12 wheat QTL markers tested on 7 DNA samples. (a) CM_01207, (b) CM_01208, (c) CM_01209, (d) CM_01210, (e) CM_01211, (f) CM_01212, (g) CM_01213, (h) CM_01214, (i) CM_01215, (j) CM_01216, (k) CM_01217 and (l) CM_01218

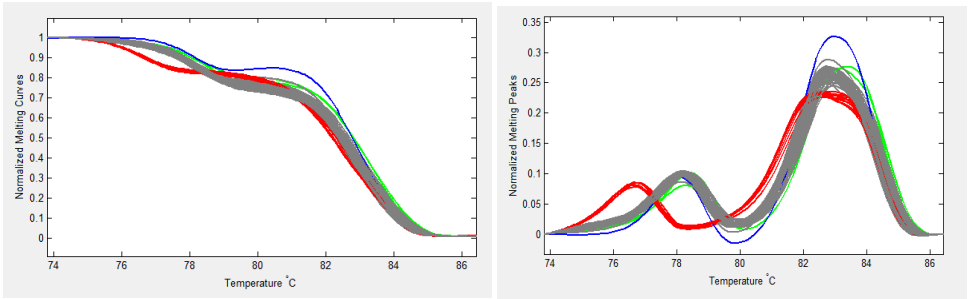


C.6 HRM profile of 5 out of the 7 better looking markers tested across a wide set of 93 samples. (a) CM_01215, (b) CM_01217, (c) CM_01211, (d) CM_01213 and (e) CM_01208

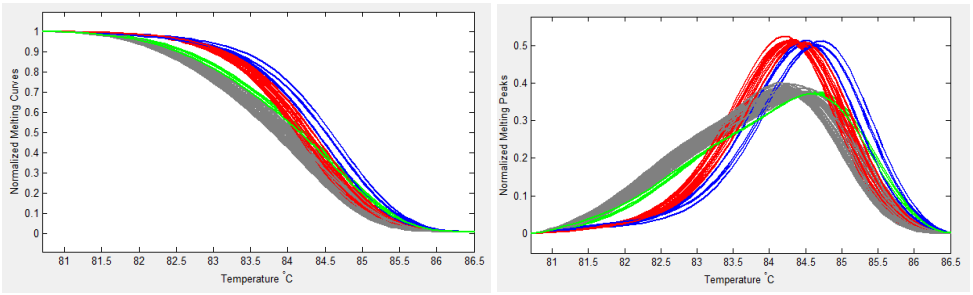
a



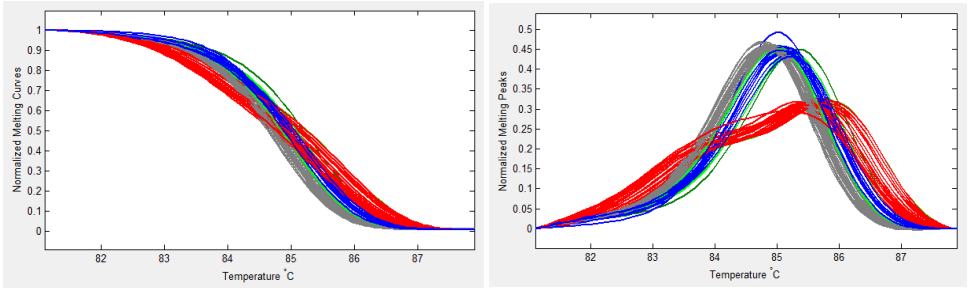
b



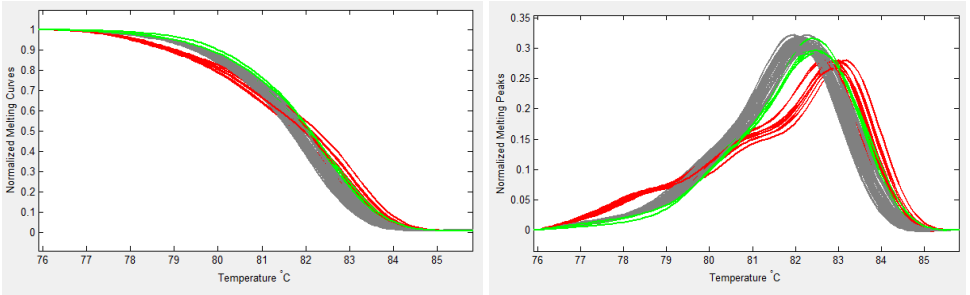
c



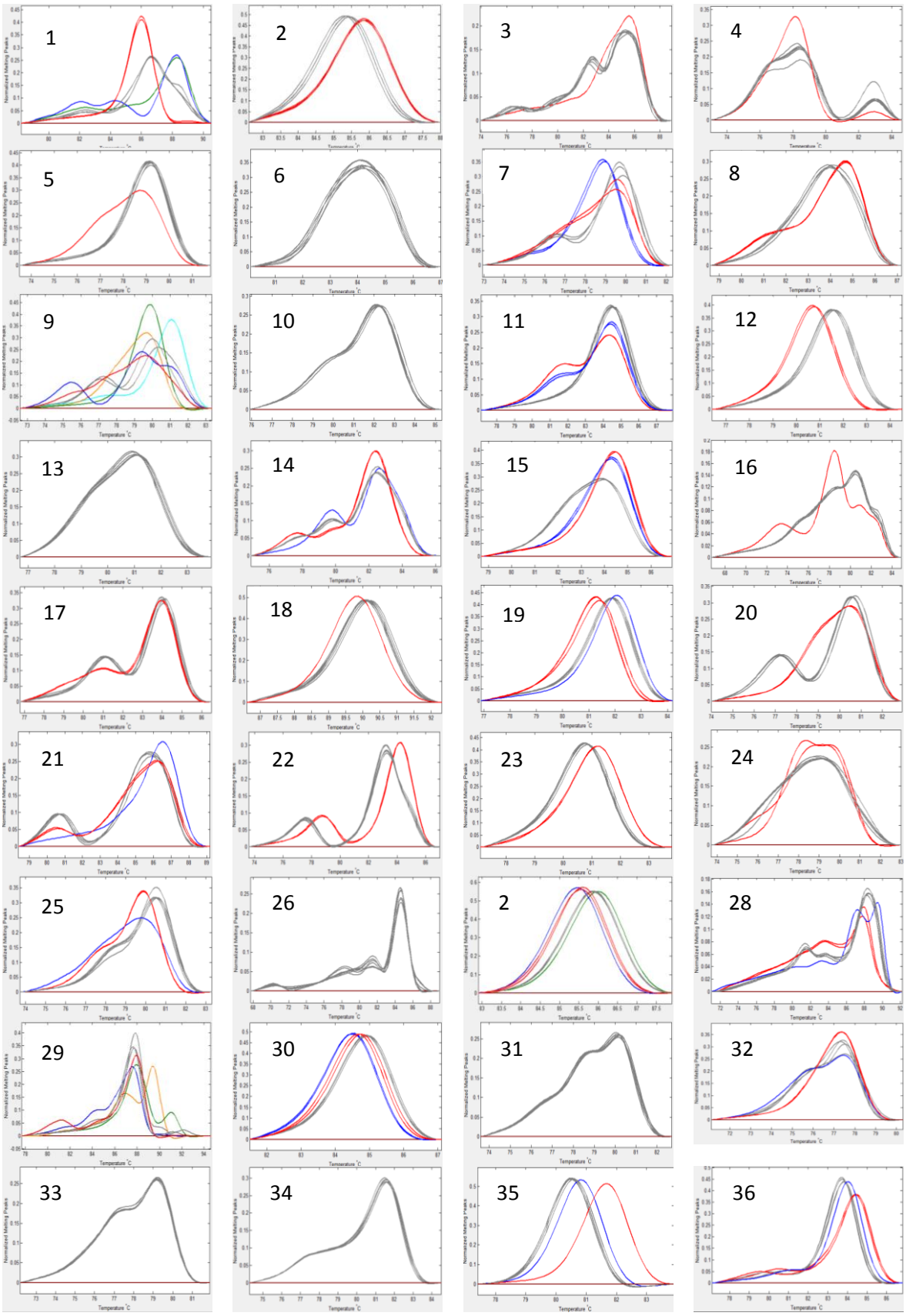
d



e



C.7 HRM profile of the 36 wheat random markers across the A genome tested on 7 DNA samples. (1-36) CM_01219 – CM_01254



References

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., & Mock, S. (2004). Kepler: An Extensible System for Design and Execution of Scientific Workflows. Retrieved from http://users.sdsc.edu/~ludaesch/Paper/ssdbm04-kepler.pdf

Aniba, M. R., Poch, O., & Thompson, J. D. (2010). Issue in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Research, 38(21), 7353-7363.

Baldwin, S., Revanna, R., Thomson, S., Pither-Joyce, M., Wright, K., Crowhust, R., Fiers, M., Chen, L., Machnight, R., & McCallum, J. (2012). A Toolkit fir bulk PCR-based marker design from next-generation sequence data: application for development of a framework linkage map in bulb onion (Allium cepa L.). BMC Genomics, 13(637), 1-9.

Blanca, J. M., Cañizares, J., Ziarsolo, P., Esteras, C., Mir, G., Nuez, F., Garcia-Mas, J., & Picó, M. B. (2011). Melon Transcriptome Characterization: Simple Sequence Repeats and Single Nucleotide Polymorphisms Discovery for High Throughput Genotyping across the Species. The Plant Genome, 4(2), 118-131.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., & Taylor, J. (2010). Galaxy, a web-based genome analysis tool for experimentalists. Current Protocols in Molecular Biology, 19(1021), 1-33.

Boettige, C. (2014). An introduction to Docker for reproducible research, with examples from the R environment. ACM SIGOPS Operating Systems Review. 49(1), 71-79.

Buffalo, V. (2014). Bioinformatics Data Skills : Reproducible and Robust Research with Open Source Tools. Sebastopol, CA, USA: O'Reilly Media. Retrieved from http://www.ebrary.com.ezproxy.lincoln.ac.nz

Carvalho, B. S., & Rustici, G. (2013). The challenges of delivering bioinformatics training in the analysis of high-throughput data. Briefings in Bioinformatics, 14(5), 538-547.

Cavanagha, C. R., Chaob, S., Wangc, S., Huangd, B. E., Stephena, S., Kianic, S., Kianic, S., Forreste, K., Saintenacc, C., Brown-Guediraf, G. L., Akhunovac, A., Seeg, D., Baih, G., Pumphreyi, M., Tomarj, L., Wonge, D., Konge, S., Reynoldsk, M., da Silvak, M. L., Bockelmanl, H., Talbertm, L., Andersonn, J. A., Dreisigackerk, S., Baenzigero, S., Carteri, A., Korzunp, V., Morrelln, P. L., Dubcovskyq, J., Morella, M. K., Sorrellss, M, E., Haydene, M. J., & Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. PNAS, 110(20), 8057-8062.

Chagne, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., Vanderzande, S., Hellens, R. P., Kumar, S., Cestaro, A., Velasco, R., Main, D., Rees, J. D., Iezzoni, A., Mockler, T., Wilhelm, L., de Weg, E. V., Gardiner, S. E., Bassil, N., & Peace, C. (2012). Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple. PloS ONE, 7(2), e31745.

Chagne, D. (2015). Application of the high-resolution melting technique for gene mapping and SNP detection in plants. Methods in Molecular Biology, 1245, 151-159.

Chao, S., & Somers, D. (2012). Wheat and barley DNA extraction in 96-well plates;WheatMAS. http://maswheat.ucdavis.edu/protocols/general_protocols/DNA_extraction_003.htm Accessed 10th July 2015

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoo, M. J. L. (2009). Biopython freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423.

Collberg, C., Proebsting, T., Moraila, G., Shankaran, A., Shi, Z., Warren, A. M. (2014). Measuring Reproducibility in Computer Systems Research. Retrieved from http://reproducibility.cs.arizona.edu/tr.pdf

Dash, P. (2013). Getting Started with Oracle VM VirtualBox. Olton, Birmingham, GBR: Packt Publishing Ltd. Retrieved from http://www.ebrary.com

Davey, J. W., & Blaxter, M. L. (2011). RADSeq- next-generation population genetics. Briefings in Functional Genomics, 9(5), 416-423.

Dwight, Z. L., Palais, R., Kent, J., & Wittwer, C. T. (2014). Heterozygote PCR Product Melting Curve Prediction. Human Mutation, 35(3), 278-282.

Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N., & Thurston, M. (2006). Open software for biologists: from famine to feast. Nature Biotechnology, 24(7), 801-803.

Garces-Claver, A., Fellman, S. M., Gil-Ortega, R., Jahn, M., Arnedo-Andres, M. S. (2007). Identification, validation and survey of a single nucleotide polymorphism (SNP) associated with pungency in Capsicum spp. Theoretical and Applied Genetics, 115, 907-916.

Garritano, S., Gemignani, F., Voegele, C., Nguyen-Dumont, T., le Calvez-Kelm, F., de Silva, D., Lesueur, F., Landi, S., & Tavtigian, S. V. (2009). Determining the effectiveness of High Resolution Melting analysis for SNP genotyping and mutation scanning at the TP53 locus. BMC Genetics, 10(5), 1-12.

Gentleman, R. C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., & Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, 5(10), R80.

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Richard, R. C., & Gelbart, W. M. (2000). An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman.

He, J., Zhao, X., Laroche, A., Lu, Z., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Frontiers in Plant Science 30, 5(484). 1-8.

High Resolution Melting. (n.d.). Retrieved 13 November, 2015, from https://www.dna.utah.edu/Hi-Res/TOP_Hi-Res%20Melting.html

Houston, R. D., Taggart, J. B., Cezard, T., Bekaert, M., Lowe, N. R., Downing, A., Talbot, R., Bishop, S. C., Archibald, A. L., Bron, J. E., Pennan, D. J., Davassi, A., Brew, F., Tinch, A. E., Gharbi, K., & Hamilton, A. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). BMC Genomics, 15(90), 1-13.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C., Fan, D., Lu, Y., Weng, Q., Liu, K., Zhou, T., Jing, Y., Si, L., Dong, G., Huang, T., Lu, T., Feng, Q., Qian, Q., Li, J., & Han, B. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nature Genetics, 44(1), 32-41.

Iehisa, J. C. M., Shimizu, A., Sato, K., Nasuda, S., & Takumi, S. (2012). Discovery of High-Confidence Single Nucleotide Polymorphisms from Large-Scale De Novo Analysis of Leaf Transcripts of Aegilops tauschii, AWild Wheat Progenitor. DNA Research, 19, 487–497.

Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemela, P., Gentile, M., Scheinin, I., Koski, M., Kaki, J., & Korpelainen, E. I. (2011). Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics, 12(507), 1-14.

Kibbe, W. A. (2007). OligoCalc: an online oligonucleotide properties calculator. Nucleic Acids Research, 35, 43-46.

Kim, H., Lee, H., Hyun, J. Y., Won, D., Hong, D. O., & Harn, C. H. (2012). CAPS Marker Linked to Tomato Hypocotyl Pigmentation. Korean Journal of Horticultural Science & Technology, 30(1), 56-63.

Kumar, S., Garrick, D. J., Bink, M. C. A. M., Whitworth, C., Chagne, D., & Volz, R. K. (2013). Novel genomic approaches unravel genetic architecture of complex traits in apple. BMC Genomics, 14(393), 1-13.

Langridge, P. & Chalmers, K. (2004). The Principle: Identification and Application of Molecular Markers. In H. Lörz & G. Wenzel (Eds.), Molecular Marker Systems in Plant Breeding and Crop Improvement (pp. 3-22).

Lee, G., Koh, H., Chung, H., Dixit, A., Chung, J., Ma, K., Lee, S., Lee, J., Lee, G., Gwag, J., Kim, T., & Park, Y. (2009). Development of SNP-based CAPS and dCAPS markers in eight different genes involved in starch biosynthesis in rice. Molecular Breeding, 24, 93-101.

Lehmensiek, A., Sutherland, M. W., & McNamara, R. B. (2008). The use of high resolution melting (HRM) to map single nucleotide polymorphism markers linked to a covered smut resistance gene in barley. Theoretical and Applied Genetics, 117, 721-728.

Migheli, F., Stoccoro, A., Coppede, F., Omar, W. A. W., Failli, A., Consolini, R., Seccia, M., Spisni, R., Miccoli, P., Mathers, J. C., & Migliore, L. (2013). Comparison Study of MS-HRM and Pyrosequencing Techniques for Quantification of APC and CDKN2A Gene Methylation. PloS ONE, 8(1), e52501.

Morris, G. P., Ramub, P., Deshpandeb, S. P., Hashc, C., T., Shahb, T., Upadhyayab, H. D., Riera-Lizarazub, O., Brownd, P. J., Acharyae, C. B., Mitchelle, S. E., Harrimane, J., Glaubitze, J. C., Bucklere, E. S., & Kresovicha, S. (2012). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. PNAS, 110(2), 453-458.

Muleo, R., Colao, M. C., Miano, D., Cirilli, M., Intrieri, M. C., Baldoni, L., & Rugini, E. (2009). Mutation scanning and genotyping by high-resolution DNA melting analysis in olive germplasm. Genome, 52, 252-260.

Nocq, J., Celton, M., Gendron, P., Lemieux, S, & Wilhelm, B. T. (2013). Harnessing virtual machines to simplify next-generation DNA sequencing analysis. Bioinformatics, 29(17), 2075-2083.

Norambuena, P. A., Copeland, J. A., Krenkova, P., Stambergova, A., & Macek, M. Jr. (2009). Diagnostic method validation: High resolution melting (HRM) of small amplicons genotyping for the most common variants in the MTHFR gene. Clinical Bioinformatics, 42(12), 1308-1316.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A., & Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20(17), 3045-3054.

van Orsouw, N. J., Hogers, R. C., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Verstegen, H., & Eijk, M. J. T. (2007). Complexity Reduction of Polymorphic Sequencess (CRoPSTM) A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PloS ONE, 11, e1172.

Pasam, R. K., Shama, R., Malosetti, M., van Eeuwijk, F. A., Haseneyer, G., Kilian, B., & Graner, A. (2012). Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biology, 12(16), 1-22.

Peacock, M. (2013). Creating Development Environments with Vagrant. Olton, Birmingham, GBR: Packt Publishing Ltd. Retrieved from http://www.ebrary.com.ezproxy.lincoln.ac.nz

Raman, H., Dalton-Morgan, J., Diffey, S., Ramen, R., Alamery, S., Edwards, D., & Batley, J. (2014). SNP markers-based map construction and genome-wide linkage analysis in Brassica napus. Plant Biotechnology Journal, 12, 851-860.

Ramkumar, G., Prahalada, G. D., Hechanova, S. L., Vinarao, R., & Jena, K. K. (2015). Development and validation of SNP based functional codominant markers for two major disease resistance genes in rice. Molecular Breeding, 35(129), 1-11.

Reed, G. H., Kent, J. O., Wittwer, C. T. (2007). High-resolution DNA melting analysis for simple and efficient molecular diagnostics. Pharmacogenomics. 8(6), 597-608.

Rookiwal, M., Nayak, S., Thudi, M., Upadhyaya, H. D., Brunel, D., Mournet, P., This, S., Sharma, P. C., & Varshney, R. K. (2014). Allele diversity for abiotic stress responsive candidate genes in chickpea reference set using gene based SNP markers. Frontiers in Plant Science 6, 5(248), 1-11.

Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for General Users and for Biologist Programmers. In S. Misener and S. A. Krawetz (Eds). Methods in Molecular Biology, vol. 132: Bioinformatics Methods and Protocols, Humana Press Inc., Totowa, NJ.

Sadedin, S. P., Pope, B., & Oshlack, A. (2012). Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics, 28(11), 1525-1526.

Sanger, F., Nicklen, S., & Coulson, A. R. (1997). DNA sequencing with chain-terminating inhibitors. PNAS, 74(12), 5463-5467.

Schlotterer, C. (2004). The evolution of molecular markers – just a matter of fashion?, Nature Reviews Genetics, 5, 63-69.

Semagn, K., Babu, R., Hearne, S. & Olsen, M. (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. Molecular Breeding, 33, 1-14.

Shen, H. (2014). Interactive notebooks: Sharing the code. Nature, 515, 151-152.

Shirasawa, K., Fukuoka, H., Matsunaga, H., Kobayashi, Y., Kobayashi, I., Hirakawa, H., Isobe, S., & Tabata, S. (2013). Genome-Wide Association Studies Using Single Nucleotide Polymorphism Markers Developed by Re-Sequencing of the Genomes of Cultivated Tomato. DNA Research, 20(6), 593-603.

Shu, Y., Li, Y., Zhu, Z., Bai, X., Cai, H., Ji, W., Guo, D., & Zhu, Y. (2011). SNPs discovery and CAPS marker conversion in soybean. Molecular Biology Reports, 38, 1841-1846.

Slater, G. S. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 6(31). 1-11.

Smith, B. L., Lu, C. P., & Alvarado Bremer, J. R. (2009). High-resolution melting analysis (HRMA) a highly sensitive inexpensive genotyping alternative for population studies. Molecular Ecology Resources, 10(1), 193-196.

Soleimani, V. D., Baum, B. R., & Johnson, D. A. (2003). Efficient Validation of Single Nucleotide Polymorphisms in Plants by Allele-Specific PCR, With an Example From Barley. Plant Molecular Reporter, 21, 281-288.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., & Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Research, 12(10), 1611-1618.

Stange, M., Utz, H. F., Schrag, T. A., Melchinger, A. E., & Wurschum, T. (2013). High-density genotyping- an overkill for QTL mapping- Lessons learned from a case study in maize and simulations. Theoretical and Applied Genetics, 126, 2563-2574.

van der Stoep, N., van Paridon, C. D., Janssens, T., Krenkova, P., Stambergova, A., Macek, M., Matthijs, G., Bakker, E. (2014). Diagnostic guidelines for high-resolution melting curve (HRM) analysis: an interlaboratory validation of BRCA1 mutation scanning using the 96-well LightScanner. Human Mutation 6, 899-909.

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P., & Reynolds, M. P. (2015). Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. Theoretical and Applied Genetics, Impact Factor,3(51).

Tommaso, P. D., Palumbo, E., Chatzou, M., Prieto, P., Heuer, M. L., & Notredame, C. (2015). The impact of Docker containers on the performance of genomic pipelines. PeerJ, 3, e1273.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G.(2012). Primer3—new capabilities and interfaces. Nucleic acids research, 40(15), e115.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. M. (2007). Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Research, 35, 71-74.

Varshney, R. K., Nayak, S., Jayashree, B., Eshwar, K., Upadhyaya, H. D., & Hoisington, D. A. (2007). Development of cost-effective SNP assays for chickpea genome analysis and breeding. Journal of SAT Agricultural Research, 3(1), 29-31.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., International Wheat Genome Sequencing Consortium, Lillemo, M., Mather13, D., Appels, R., Dolferus, R., Brown-Guedira, G., Korol, A., Akhunova, A. R., Feuillet, C., Salse, J., Morgante, M., Pozniak, C., Luo, M., Dvorak, J., Morell, M., Dubcovsky, J., Ganal, M., Tuberosa, R., Lawley, C., Mikoulitch, I., Cavanagh, C., Edwards, K. J., Hayden, M., & Akhunov, E. (2014). Characterisation of polyploid wheat genomic diversity using a high-density 90000 single mucleotide polymorphism array. Plant Biotechnology Journal, 12, 787-796.

Wang, B., Tan, H., Fang, W., Mehinhardt, L., Mischke, S., Matsumoto, T., & Zhang, D. (2015). Developing single nucleotide polymorphism (SNP) markers from transcriptome sequences for identification of longan (Dimocarpus longan) germplasm. Horticulture Research, 2, 14065.

Wu, B., Zhong, G., Yue, J., Yang, R., Li, C., Li, Y., Zhong, Y., Wang, X., Jiang, B., Zheng, L., Yan, S., Bei, X., & Zhou, D. (2014). Identification of Pummelo Cultivars by Using a Panel of 25 Selected SNPs and 12 DNA Segments. PloS ONE, 9(4), e84506.

Yang, X., Wallom, D., Waddington, S., Wang, J., Sharon, A., Matthews, B., Wilson, M., Guo, Y., Guo, L., Blower, J. D., Vasilakos, A. V., Liu, K., & Kershaw, P. (2014). Cloud computing in e-Science- research challenges and opportunities. The Journal of Supercomputing, 70, 408-464.

Ye, J., Conlouris, G., Zaretsksya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics, 13(134), 1-11.

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., Xiao, J., & Zhang, Q. (2011). Gains in QTL Detection Using an Ultra-High Density SNP Map Based on Population Sequencing Relative to Traditional RFLP-SSR Markers. PloS ONE, 6(3), e17595.