Molecular marker applications in oat (*Avena sativa* L.) breeding and germplasm diagnostics

.

Utilisation des marqueurs moléculaires pour l'amélioration et le diagnostic variétal chez l'avoine (*Avena sativa* L.)

Benazir Katarina Marquez

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the MSc degree in Biology

.

Une thèse soumise à la Faculté des études supérieures et postdoctorales
en vue de l'obtention de la maîtrise en biologie

Department of Biology・Département de biologie
Faculty of Science・Faculté des sciences
University of Ottawa・Université d'Ottawa

ABSTRACT

The ability to identify germplasm and select traits accurately is fundamental to successful plant breeding. Pedigrees and molecular markers facilitate these processes; however misleading experimental results can occur when incorrect relationships and/or cultivar names are recorded. Molecular markers can identify these inconsistencies, and with advances in genotyping technology these diagnostics can be done faster and more objectively. This study aimed to develop molecular marker assays and graphical genotyping methodologies for cultivar identification, seed purity assessment and trait selection in oat (*Avena sativa* L.). KBioscience's Allele-Specific PCR (KASP™) and genotyping-by-sequencing (GBS) technologies were applied to a set of current Canadian oat cultivars to evaluate their utility for identifying cultivars and detecting intra-cultivar variation. Both KASP™ and GBS detected different extents of heterogeneity among a set of 160 seeds that originated from four seed sources of four cultivars. In both cases, the detected variation did not appear to be limited to a specific cultivar or seed source, reinforcing that all cultivars are heterogeneous. Graphical genotyping localized heterogeneity to specific chromosome regions, thereby distinguishing physical contamination from true genetic heterogeneity and heterozygosity. Pre-existing genotype data for 700 oat cultivars and breeding lines were also used to construct graphical genotypes for pedigree validation and discovery of potential sources for favourable quantitative trait loci (QTL) alleles. This methodology used historical QTLs and anchoring markers to identify 25 putative "high oil" allele carriers. The results from this study will provide diagnostic tools for cultivar identification and pedigree validation, in addition to meaningful information about existing heterogeneity and possible QTL locations in current cultivars.

RÉSUMÉ

La capacité d'identifier les matériaux génétiques est la base de l'amélioration variétale. La généalogie et les marqueurs moléculaires facilitent l'identification variétale, mais des résultats erronés pourraient se produire lorsque l'apparentement et / ou les noms de cultivars sont enregistrés de façon incorrecte. Les marqueurs moléculaires peuvent identifier ces erreurs. L'avancement des technologies de génotypage permettent que ces diagnostics soient faits plus rapidement et de manière plus objective. Cette étude a pour objectif de développer des essais de marqueurs moléculaires et d'employer le génotypage graphique pour l'identification des cultivars, l'évaluation de la pureté des semences et l'application en amélioration chez l'avoine (*Avena sativa* L.). Le PCR d'allèle-spécifique de KBioscience (KASP™) et le « génotypage-par-séquençage» (GBS) ont été appliqués à un ensemble de cultivars d'avoine canadiens pour évaluer leur utilité dans l'identification variétale et la détection de la variation intra-variétale. KASP™ et GBS ont détectés l'hétérogénéité à différent degrés parmi 160 graines provenant de quatre cultivars et quatre sources. La variation détectée ne semble pas être limitée à un cultivar ou une source de graines spécifique, renforçant le fait que la variation intra-variétale existe au sein des cultivars. Le génotypage graphique a permis de distinguer la contamination physique de la vraie hétérogénéité génétique ou hétérozygotie. En utilisant ces génotypages graphiques, l'hétérogénéité a été aussi localisée à des régions chromosomiques spécifiques. Les données génotypiques pour 700 cultivars d'avoine et lignées en cours de sélection ont été utilisées pour construire des génotypages graphiques afin de valider les généalogies et de découvrir des sources potentielles d'allèles favorables. En utilisant les locus de caractères quantitatifs (LCQ) historique et des marqueurs de référence, 25 sources potentielles d'allèles ont été identifiées. Ces allèles ont le potentiel d'augmenter la teneur en huile dans les graines d'avoine. Les résultats de cette étude fourniront des outils de diagnostic pour l'identification des cultivars et la validation des généalogies. De plus, les informations utiles sur la variation intra-variétale et les positions possibles des LCQ chez les différents cultivars sont également fournies.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Drs. Nicholas Tinker and Douglas Johnson. This work would not have been possible without their guidance, patience, and support.

I would also like to thank Charlene Wight and Yung-Fen Huang for their assistance in the preparation of this thesis. I am also thankful for the camaraderie of the other members of the Tinker lab: Biniam Hizbai, Phil Couroux, and Jiro Hattori.

I am indebted to Drs. Jennifer Mitchell Fetch, Weikai Yan, and Aaron Beattie for contributing plant material to this project; Dr. Jesse Poland for contributing to the genotyping-by-sequencing work presented here; and to the members of the Collaborative Oat Research Enterprise (CORE) for allowing me to use their collective data in my analyses.

Finally, I am grateful for my family and friends for their encouragement in this endeavour.

This thesis is dedicated in loving memory to:

VN (1939-2006), RM (1930-2012), and EN (1928-2013)

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADP	allele discrimination plot
AAFC	Agriculture and Agri-Food Canada
AFLP	amplified fragment length polymorphism
bp	base pair
CDC	Crop Development Centre
CDOC	Canadian Diagnostic Oat Cultivars
CFIA	Canadian Food Inspection Agency
CIMMYT	International Maize and Wheat Improvement Center
CL	confidence level
cM	centimorgan
CORE	Collaborative Oat Research Enterprise
CRC	Cereal Research Centre
CSGA	Canadian Seed Growers Association
CTAB	cetyl trimethylammonium bromide
d	genetic distance
DArT	Diversity Array Technology
DNA	deoxyribonucleic acid
DUS	distinct, uniform, stable
DxE	Dal x Exeter
ECORC	Eastern Cereal and Oilseed Research Centre
EST	expressed sequence tag
FAM	6-carboxyfluorescein
FAOSTAT	Food and Agriculture Organization of the United Nations
FRET	fluorescence resonance energy transfer
Gb	gigabase
GBS	genotyping-by-sequencing
GG	graphical genotype
HEX	6 - carboxy - 2',4,4',5',7,7' - hexachlorofluorescein
hm	haplotype match
IBD	identical by descent
IBS	identical by state
ISHS	International Society for Horticultural Science
KASP™	“Kompetitve Allele-Specific PCR”
KOD	KASP™-on-Demand
KxO	‘Kanota’ x ‘Ogle’
LOD	logarithm of odds
MAB	marker-assisted breeding

MAG/GOD	Market Analysis Group/Grain and Oilseeds Division
MAF	minor allele frequency
NGS	next-generation sequencing
NIL	near-isogenic line
nt	nucleotide
NTC	no-template control
PBR	Plant Breeders' Rights
PCR	polymerase chain reaction
PIC	polymorphic information content
POOL	Pedigrees of Oat Lines
QTL	quantitative trait locus
RAPD	random amplified polymorphic DNA
rf	recombination frequency
RFLP	restriction fragment length polymorphism
RFU	relative fluorescent unit
RIL	recombinant inbred line
r_k	kinship coefficient
r_M	Mantel statistic
ROX	5-(and 6)-carboxy-X-rhodamine
RNA	ribonucleic acid
SCAR	sequenced characterized amplified region
SIU	Seed Increase Unit
SNP	single nucleotide polymorphism
SSR	simple sequence repeat
UNEAK	Universal Network-Enabled Analysis Kit
UPGMA	unweighted pairwise group method with arithmetic mean

Chapter 1
Introduction

1.1 Oat

Oat (*Avena* spp.) is a cereal belonging to the Gramineae (Poaceae) family that is well-adapted to cool climates and grown mainly in temperate regions or in winter seasons. The centre of origin is unknown, but most species are considered to have originated in the Mediterranean basin or the Middle East (Murphy and Hoffman, 1992). Oat occurs in three ploidy levels: diploid, tetraploid, and hexaploid, with a base number of seven chromosomes.

1.1.1 The oat genome

Cultivated oat (*Avena sativa* L.) is an allohexaploid (2n=42) with three diploid progenitor genomes: AA, CC, and DD (Baum, 1977). Diploid oats with AA or CC are reproductively isolated from each other and are distinct in chloroplast DNA and chromosome constitution (Rines *et al.,* 2006). The D genome of the hexaploid is thought to be a recent variant of A due to its high similarity, however no extant DD genome diploids have been identified. The other species of oat are described and reviewed by Baum (1977), Coffman (1977), and Leggett (1992).

The genome size of oat has been estimated to be 11.3-14.0 Gb (Bennett and Leitch, 1976; Luo *et al.,* 2012). A large proportion (83%) of the genome is comprised of repetitive sequences (Flavell *et al.* (1977), some of which are retrotransposons or their remnants. A

set of genomic clones from a chromosome-specific (18D) library were characterized, and about half of the assembled contigs contained known repeats largely belonging to LTR-transposons (81%), transposons (16%), and non-LTR transposons (3%) (Luo *et al.,* 2012). No definitive estimate of the total amount of retrotransposons has been reported in oat, however estimates have been reported for members of the Ty1-*copia* LTR- retrotransposon family: pAs17 and OARE-1. Hexaploid oat contains 24, 000 (Linares *et al.,* 1999) and 10, 000 (Kimura *et al.,* 2001) copies of each, respectively. Repetitive DNA sequences are dispersed or clustered in the oat genome and can be used as cytological markers to distinguish chromosomes from the A/D and C genomes (Ananiev *et al,* 2002). In situ hybridization using these repetitive sequences as probes has identified translocations in oat species (Jellen *et al.,* 1994). The 7C-17A translocation is particularly widespread in *A. sativa* cultivars, as well as other hexaploid species of *Avena* (Jellen *et al.,* 2000). Additionally, many genes appear to be duplicated on multiple chromosomes (Portyanko *et al.,* 2001).

1.1.2 Oat as a commodity

Oat is grown primarily to produce grain for human and animal feed, but is also grown as forage or green manure (Kim *et al.,* 2014), or as a nurse crop to establish legumes and other grasses (Webster, 1986). Oat for grain production is grown mostly in Eurasia and North America, where it is well-adapted to the cool and moist climates. Russia, Canada, and Poland are the top three producers of oat in the world (FAOSTAT, 2012). Among Canada's 15 principal field crops, oat is ranked seventh by production (MAG/GOD, 2014). In 2012, Canada produced approximately 2.7 million metric tonnes of oat (FAOSTAT, 2012). Canada is also the world's largest oat exporter, with most products being sent to the United States

of America (MAG/GOD, 2010). Exports are forecasted to increase by 8% due to an increase in total supply and a return to normal trade patterns with the USA (MAG/GOD, 2014).

In light of its health benefits, oat has become known as a health food. Due to their high levels of β-glucan, consuming oat products can lower cholesterol (Braaten *et al.,* 1994) and improve glucose and insulin metabolism. Oat is also used in cosmetics because of its fine and uniform starch granules, which are considered to relieve itch and provide a non-allergenic replacement for talc. Avenanthramides, polyphenols found only in oat, also appear to give oat an anti-inflammatory effect (Sur *et al.,* 2008). Value-added traits, in addition to agronomic traits such as yield, are continually being investigated to improve oat and expand its market value.

1.2 Plant breeding

Plant breeding is an iterative process that modifies plants by deliberately hybridizing specific parents and selecting progeny with favourable qualities. Developed progeny can then be used as parents in later crosses. Essentially, plant breeding is artificial selection for traits that are agronomically important or add value to a crop.

Oat cultivars are usually developed and released as pure lines; consequently, the overall objective is to recombine as many favourable genes as possible into a single homozygous genotype. Mendelian genetics lies at the heart of plant breeding; however, the traits desired in a plant cultivar are not always simply inherited. Most agronomic traits are controlled quantitatively by multiple genes and environmental effects (Tinker, 2008). Genes can also be inherited in a linked fashion, leading to some allele combinations being

inherited more commonly than others. Selection, random genetic drift, epigenetics, and genetic interaction such as epistasis, also lead to a deviation from expected Mendelian inheritance.

1.2.1 Oat breeding in Canada

The late 1800s marked the beginning of oat breeding in Canada. Oat improvement was necessary at the time to meet the demand of the growing Canadian livestock industry (Welsh *et al.,* 1953; McKenzie and Harder, 1995). Public breeding programs at government experimental farms and several agricultural colleges generated highly productive cultivars from the 1900s-1930s. Backcrossing in the 1960s introduced rust resistance genes into many of the cultivars developed at that time. Similarly, introgression of wild oat (*Avena sterilis* L.) germplasm in the 1970s helped develop cultivars resistant to newer forms of stem and crown rust. Current cultivar development continues in public and private breeding programs.

Despite significant progress in oat cultivar development in Canada, the narrowing of the Canadian oat gene pool is of concern. Pedigrees show that cultivars developed since 1930 have been based mainly on a genetic foundation of fewer than 10 important ancestral varieties and landraces (Kibite *et al.,* 2004). Furthermore, a decrease in average genetic diversity has been detected in cultivars released since 1950 (Fu *et al.,* 2003a). This drop in diversity may be partially explained by the increased efforts to introduce rust resistance by backcrossing and introgression. To avoid further narrowing the Canadian gene pool, it would be advantageous to introduce new, diverse germplasm into breeding programs.

1.2.2 Canadian cultivars

A “cultivar” is defined as a group of plants that: a) has been selected for a specific trait or combination of traits, b) is distinguishable, uniform, and stable (DUS) in these traits, and c) retains these traits when reproduced by appropriate means (ISHS, 2009). The term “variety”, used in Canada’s *Seeds Act* and *Plant Breeders’ Rights Act*, is synonymous with cultivar[1].

Canada’s *Seeds Act and Regulations* oversees the testing, inspection, quality and sale of seeds in the country. According to Section 3(1)(b) of the *Seeds Act*, "no person shall sell or advertise for sale in Canada or import into Canada seed of a variety that is not registered in the prescribed manner." There are currently 123 hulless and spring oat cultivars registered for use in Canada (CFIA, 2013). The majority are Canadian-bred cultivars; however, cultivars originating from the United States, United Kingdom, and Sweden are also registered.

For a cultivar to be eligible for registration in Canada, the representative reference sample must meet specific varietal purity standards set by the Canadian Seed Growers’ Association (CSGA) or the *Seeds Regulations* (67(1)(f)). Accordingly, a representative reference sample of an oat cultivar can only contain one to eight off-types or impurities (depending on crop species) in approximately 10, 000 plants. As the official Canadian pedigreeing agency, the CSGA prescribes “varietal purity standards” and oversees the seed crop certification of most agricultural crops (CSGA, 2014).

[1] The term variety may also apply to a botanical variety (a subtaxon of species). To avoid confusion, “cultivar” is used throughout this thesis except when directly referring to or quoting published documents using the term “variety” or “plant variety”. The term “line” is used in reference to inbred breeding material that is not recognized as a variety or cultivar.

1.2.2.1 Pedigreed seed classes and crop certification

Seed crop certification produces pedigreed seeds through a program of planned production, record keeping, unbiased inspection, and rigid standards (CSGA, 2014). Certification maintains cultivar-specific characteristics by limiting the number of generations or multiplications of pedigreed seed classes.

There are five categories of pedigreed seed: breeder, select, foundation, registered, and certified. Breeder seed is considered as the purest seed and is developed and maintained by plant breeders. The other four categories of pedigreed seed are derived from generations and multiplications from breeder seed. The first generation from breeder seed is referred to as select seed, which is produced by authorized seed growers to maintain its varietal identity and purity (CSGA, 2014). Foundation seed is the second generation produced from breeder seed, which is rogued for off-types in order to meet cultivar descriptions and Foundation purity standards. Registered seed is also produced to maintain a variety's identity and purity, but, unlike select seed, comes from breeder, select, or foundation seed (CSGA, 2014). Under the requirements of the *Seeds Act and Regulations,* registered seed must also be graded and labelled by persons authorized by the CFIA. Certified seed is the second generation from foundation seed, which is produced by CSGA seed growers for sale to farmers for commercial production. There is no generation limit for oat breeder seed; however, select seed is limited to five generations while foundation, registered, and certified seed are limited to only one generation.

1.2.3 Marker-assisted breeding (MAB)

Most agronomically important traits are quantitative; i.e., affected by multiple genes and the environment. Quantitative trait loci (QTL) are areas of a chromosome where an allele substitution in one or more genes is statistically inferred to cause a change in the magnitude of a trait. Molecular markers that are linked to QTL can be used to improve the speed, accuracy, or efficiency of plant breeding either through the identification of better parents, or through the selection of progeny with desirable traits. Any breeding strategy that employs molecular markers is referred to as marker-assisted breeding (MAB) or as marker-assisted selection. In contrast to phenotypic selection, where it is often necessary to observe a whole row of plants or a whole bin of seed, MAB can be applied to select a single seed or seedling that carries a desired trait (Tinker, 2008; Gnanesh *et al.,* 2013). Overall, MAB would have the advantage of freeing up resources that would otherwise be used to grow and measure plants that have less overall value in the program.

1.3 Molecular markers

A molecular marker is a specific DNA sequence associated with a specific region in the genome (i.e., a locus). The locus may be a functional gene or an arbitrary "landmark" that can be used for genome analysis (Varshney *et al.,* 2009). There are now many types of molecular marker; for example, restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), Diversity Array Technology (DArT) markers, and single nucleotide polymorphisms (SNPs).

Molecular markers can be used to "fingerprint" plants so that cultivars can be distinguished from each other. Polymorphic information content (PIC, Botstein *et al.,* 1980) estimates the discriminatory power of a marker at a locus as follows:

$$PIC = 1 - \sum_{i=1}^{n} {f_i}^2 \qquad [1.1]$$

where f_i is relative allele frequency. PIC values of 0 represent markers that cannot discriminate any alleles at the locus (i.e., the locus is monomorphic). A PIC value of 1 suggests that a marker is highly discriminative, with many alleles at the locus observed in equal frequencies. Bi-allelic markers have a maximum PIC value of 0.5, which occurs when half of a group of taxa carries one allele and the rest carry the other allele. Larger PIC values can only occur when there are more than two alleles. PIC is often used to evaluate the utility of markers in plants (Smith *et al.,* 1997).

1.3.1 Restriction fragment length polymorphisms (RFLPs)

RFLPs refer to restriction fragments whose length is variable because of a polymorphic restriction enzyme recognition sequence or a physical insertion or deletion between recognition sequences. A polymorphism within a recognition site effectively removes an enzyme cut site, thus producing longer restriction fragments. On the other hand, an insertion or deletion of DNA between two functioning cut sites produces restriction fragments that vary in length. Digestion with a restriction enzyme results in too many DNA fragments to visualize; therefore, Southern hybridization with specific probes is used to visualize and assay specific RFLP loci. Alternatively, PCR using primers designed to anneal to

either side of the site can also be used to score RFLPs (Brown, 2002). RFLPs produce characteristic patterns that can be used to discriminate between cultivars.

RFLP analysis revealed a relatively high level of polymorphism among 83 North American oat cultivars (O'Donoughue *et al.,* 1994). Forty-eight cDNA clones were used as probes, of which one could distinguish unambiguously among 33 cultivars. A set of 205 polymorphic fragments revealed by 11 probes fully diagnosed the unique identity of all 83 cultivars studied. Thus, RFLPs provide a very effective assessment of DNA variation. RFLPs were also used to develop the early linkage maps in oat (O'Donoughue *et al.,* 1992; Rooney *et al.,* 1994; Wight *et al.,* 2003) because of their abundance and frequent co-dominant nature. However, the analysis of RFLPs can be time- and resource-consuming. Automation is not possible and a single assay provides limited information. RFLP probes are also difficult to transfer between laboratories since the probes must be physically maintained. To overcome these limitations, PCR-based markers were developed.

1.3.2 Random amplified polymorphic DNA (RAPDs)

RAPD analysis is based on the use of a single random primer that is short enough to anneal to both template strands to produce one or more PCR-amplified fragments. These fragments are sometimes polymorphic so by testing numerous primers, polymorphic amplicons can be discovered throughout the genome of a species. PCR fragments are visualized on agarose gels and polymorphic amplicons are scored for presence or absence in individual samples. An advantage of this method is that it can be used in the absence of genomic sequence data (Chial, 2008). RFLPs and amplified fragment length polymorphisms

(AFLPs) can also be applied without prior sequence knowledge, but when available it improves probe or primer design and optimizes analysis.

Guillin *et al.* (1998) developed an identification scheme for 53 Canadian oat cultivars based on RAPDs. Multiple primers were used to create fingerprints and only 13 were necessary to identify all 53 cultivars. A dichotomous key was created to identify the cultivars based on the presence or absence of diagnostic bands. This identification key was tested on breeder seed of five cultivars and was found to be repeatable and reliable. However, to be used routinely for cultivar identification additional investigations were needed but were not completed. Guillin *et al.* suggested that repeatability in other laboratories and DNA band homology needed to be further investigated before RAPDs could be put to practical use. They also suggested that diagnostic bands be converted to locus-specific sequenced characterized amplified region markers (SCARs) to allow for more efficient cultivar identification. Use of the developed key is hindered by RAPD analysis since each RAPD has to be applied one at a time and then diagnostic bands must be scored amid other bands. Furthermore, the number of oat cultivars currently registered in Canada has more than doubled since this identification key was reported.

1.3.3 Amplified fragment length polymorphisms (AFLPs)

AFLPs are similar to RFLPs in that they both involve variable lengths of restriction fragments. However, AFLPs are generated by digestion with two different enzymes followed by selective amplification. Choice of enzymes is dependent on the organism being studied, but must consist of a rare cutter and a frequent cutter (Vos *et al.,* 1995). Following double digestion, all fragments are left with an enzyme-specific 5' overhang. Adapters with

complementary overhangs are ligated to the ends of the fragments to act as primer binding sites (Chial, 2008). AFLP primers are designed with a selective 3' extension of one or more nucleotides. This allows the primers to hybridize only to fragments that have sequences, adjacent to the adapter and restriction sites, which complement the selective extension. Fingerprinting of complex genomes generally require two amplification steps. The first amplification, referred to as pre-amplification, usually involves primers with only one selective nucleotide, whereas the second amplification uses radioactively labelled primers with three selective nucleotides (Vos *et al.,* 1995). AFLPs are visualised and scored on a polyacrylamide gel and can create unique banding patterns like RFLPs and RAPDs. Alternatively, AFLPs can be separated and visualized using capillary electrophoresis.

Canadian oat cultivars have been subjected to AFLP analysis to determine patterns of genetic diversity, but the analysis has not been used specifically for cultivar identification. On the other hand, Polish oat cultivars have been fingerprinted using AFLPs and RAPDs. Paczos-Grzeda (2004) reported that two primers from the nine used in a simplified AFLP method produced enough polymorphism to distinguish all 19 oat cultivars studied.

RFLP and AFLP markers are advantageous for detecting polymorphism because of their abundance. However, since RFLP and AFLP marker systems cannot be automated they are not well suited for the high-throughput demands of MAB (Gustafson *et al.,* 2010). PCR-based marker systems that can be applied quickly, without the need for gel electrophoresis would be more appropriate for breeding applications.

1.3.4 Simple sequence repeats (SSRs)

SSRs are an array of repeat sequences that vary in length in the genome and are often referred to as microsatellites. Sequences generally consist of a 1-, 2-, 3-, or 4-bp unit repeated 10-20 times (Brown, 2002). SSRs are ubiquitous and evenly distributed in eukaryotic genomes, and are often highly polymorphic and reproducible (Li *et al.,* 2000). SSRs are assayed using PCR and visualized on an agarose or polyacrylamide gel. Although SSRs were used in oat as early as 2000, only 348 markers have been published largely because they are difficult and costly to develop (Wight *et al.,* 2010). The majority of these SSRs were developed from ESTs (Becher, 2007), RFLPs (Pal *et al.,* 2002), or SSRs from other species (Gustafson *et al.,* 2010). Oat SSRs have been applied to investigate relationships among *Avena* species and cultivars (Li *et al.,* 2000; Montilla-Bascón *et al.,* 2013), as well as region-specific polymorphism (Holland *et al.,* 2001). However, only one study has used SSRs explicitly for fingerprinting. Wight *et al.* (2010) described a set of SSRs capable of distinguishing 35 North American cultivars. One of these markers, referred to as MAMA5, was recently reported to be capable of discriminating between 36 Spanish oat cultivars (Montilla-Bascón *et al.,* 2013).

1.3.5 Diversity Array Technology (DArTs)

Diversity array technology (DArT) is a microarray-based technique used to develop molecular markers using genomic representations and diversity panels. A genomic representation is generated from genomic DNA that has been complexity reduced by digestion with a pair of restriction enzymes, then ligated to adapters, cloned, amplified and purified (Jaccoud *et al.,* 2001). A diversity panel is composed of pooled genomic

representations, from an organism or population of organisms, which are arrayed onto a solid support. DArT detects the presence or absence of polymorphic fragments by hybridizing single genomic representations to the array. These representations are labelled with a green or red fluorescent dye depending on the analysis chosen for DArT (Jaccoud *et al.,* 2001).

DArT analysis can be performed to contrast samples by mixing and hybridizing two representations to a diversity panel. Using this approach, each representation is labelled with either green or red fluorescent dye. Significant differences in the fluorescent dye signal ratio reveal which fragments contain a polymorphism between the two samples. A DNA sample, converted to a representation, can also be genetically fingerprinted using DArT when it is hybridized to a diversity panel with a representation comprised of fragments common to all array elements (Jaccoud *et al.,* 2001). Like the first approach, each representation is labelled with either dye; however, polymorphisms are identified by binary distribution of the signal ratios. The detected sequence variation, in either analysis, can be a SNP or an insertion-deletion mutation.

The DArT platform is well suited to genotyping little-characterized genomes, because it provides comprehensive genome coverage without prior sequence information (Wenzl *et al.,* 2004). The first DArT markers developed in oat had full genome coverage and were used to improve an existing genetic map (Tinker *et al.,* 2009). DArT markers have already facilitated research in oat cultivar improvement by helping to identify QTLs for oil (Hizbai *et al.,* 2012), vernalization response (Nava *et al.,* 2012) and *Fusarium* resistance (He *et al.,* 2013). In addition, DArTs have allowed linkage disequilibrium (Newell *et al.,* 2010) and

genome-wide association for β-glucan (Newell *et al.,* 2012; Asoro *et al.,* 2013) to be explored in oat.

1.3.6 Single nucleotide polymorphisms (SNPs)

A SNP generally describes a locus in a genome where two alleles are observed within a population. SNPs represent the most common type of sequence polymorphism in plant and animal genomes (Waters *et al.,* 2008)**.** SNPs arise from mutations, but most are removed by mismatch repair. Thus, SNPs that escape this process become fixed (Kunkel and Erie, 2005). Since all SNPs arise from mutations, the number of SNPs within a population reflects forward and reverse mutation rates. In turn, these rates provide insight on the number of generations that separate the population from its closest relative and other ancestors (Edward *et al.,* 2008)**.** SNPs are also the basis of other detectable polymorphisms, such as RFLPs and AFLPs. Due to recent advances in DNA sequencing technology it is now cost-effective to directly detect SNPs (Waters *et al.,* 2008).

1.3.6.1 Genotyping-by-sequencing (GBS)

Next generation sequencing (NGS) has become so inexpensive, that raw sequencing outputs are doubling approximately every six months (Poland and Rife, 2012). Sequence-based genotyping has been developed to apply NGS to plant breeding and genomics. GBS is a popular approach that has already been used in important crops for genetic mapping (Elshire *et al.,* 2011; Poland *et al.,* 2012a) and diversity studies (Lu *et al.,* 2013). Traditionally, genotyping germplasm requires a two-step process: marker discovery followed by assay design and genotyping. However, GBS completes marker discovery and genotyping together. The GBS method (Figure 1.1) employs one (Elshire *et al.,* 2011) or two

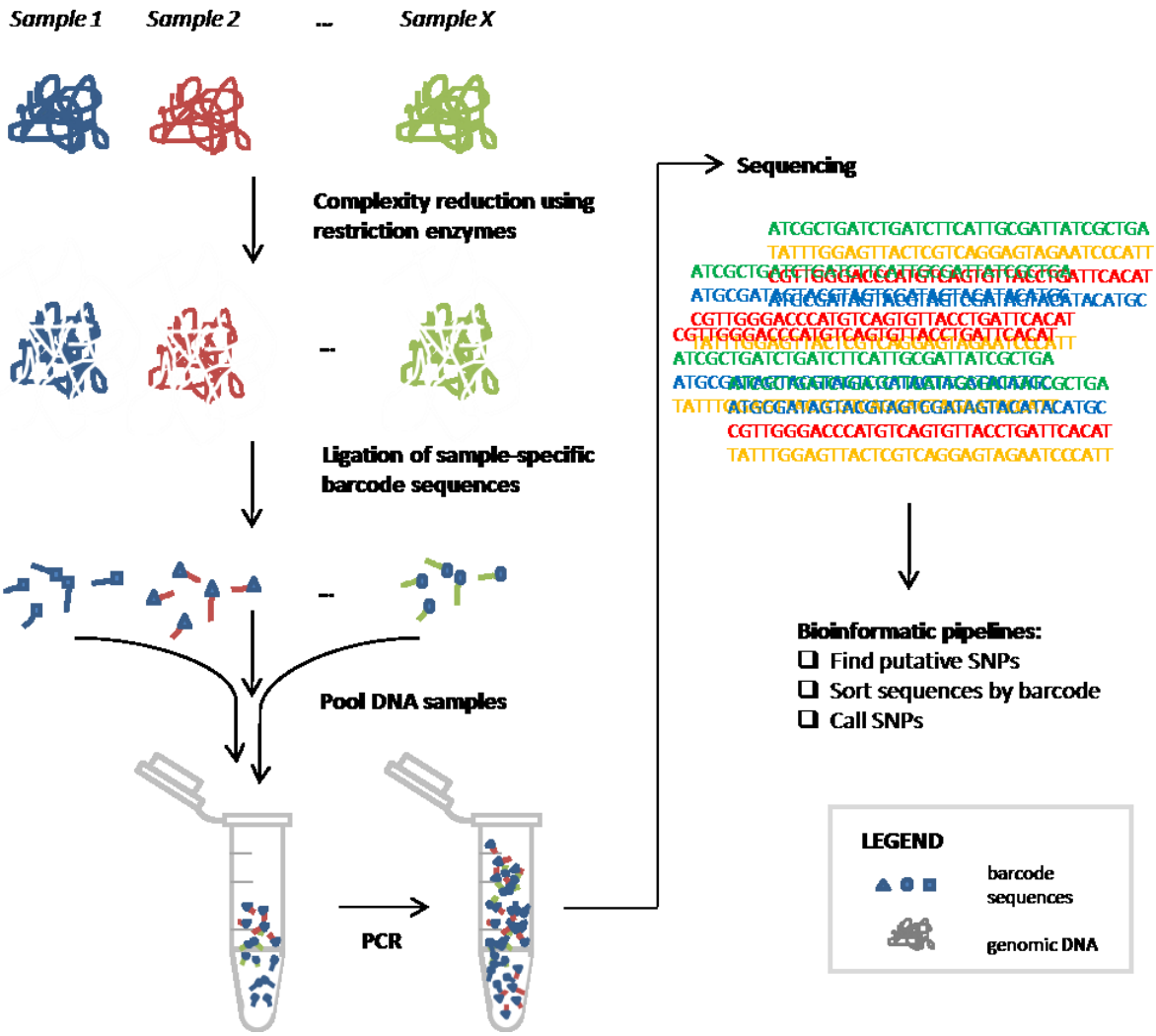

Figure 1.1 Genotyping-by-sequencing (GBS) approach.

(Poland *et al.,* 2012b) restriction enzymes to reduce genome complexity and then ligates short "barcode" sequences to produce multiplex libraries for sequencing. Barcodes are later used to identify which fragments belong to each sample. SNP discovery and genotyping is accomplished using bioinformatic pipelines such as the Universal Network-Enabled Analysis Kit (UNEAK) reported in Lu *et al*. (2013).

UNEAK enables GBS to be applied to any species, regardless of the availability of a reference genome. Consequently, GBS can be applied to non-model species such as switchgrass (Lu *et al.,* 2013) and oat (Huang *et al., submitted*). Modifications to the original GBS method (Elshire *et al.,* 2011) has also enabled GBS to be adapted for highly heterozygous crops like potato (Uitdewilligen *et al.,* 2013). The adaptability of GBS in addition to its low cost and high-throughput manner, make it suitable for breeding applications and cultivar diagnostics. GBS has already proven to be useful for genomic selection in wheat (Poland *et al.,* 2012a). GBS, like any SNP discovery platform, provides sequence information that can be used to develop high-throughput, fluorescence-based genotyping assays like Illumina's GoldenGate and Infinium, Life Technologies' Taqman®, or KBioscience's KASP™.

1.3.6.2 Kompetitive Allele Specific PCR (KASP™)

The KASP™ genotyping system uses a combination of two molecular techniques: allele-specific polymerase chain reaction (PCR) and fluorescence resonance energy transfer (FRET). A KASP™ assay (Figure 1.2) includes: a) two allele-specific primers; b) one common primer; and c) two FRET cassettes, each including a quenched 6-FAM- or 6-HEX-labelled oligonucleotide. Each allele-specific primer is designed with a complementary SNP allele at

its 3’ end to ensure specificity. Additionally, the tail sequence of each allele-specific primer is complementary to one of the fluorescent-dyed oligonucleotides. The common primer introduces the complementary sequence of the tail during the second cycle of PCR. FRET cassettes become unquenched during denaturation steps, allowing for the fluorescent-dyed oligonucleotides to hybridize with amplified DNA. Accumulation of labelled DNA generates a fluorescent signal that can be detected and used for allele discrimination.

KASP™ assays were traditionally performed using a classic thermocycler and fluorescent plate reader. KASP™ genotyping is now often done on a quantitative PCR instrument, so that fluorescence readings can be recorded immediately after the reaction is complete. These results can be interpreted using a plot of normalized relative fluorescence units (nRFUs) for each dye. Depending on the software being used, relative fluorescence is normalized to the fluorescence given off by no-template controls (NTCs) or a passive dye (i.e. ROX). Homozygous genotypes are expected to be plotted at the maximum end of either the x- or y-axis, since each allele-specific primer is associated to one of the two dyes. Heterozygous genotypes, in theory, should have an equal fluorescence signal from each dye, therefore a cluster roughly in between the other two homozygous groups should be observed.

The automation that comes with using KASP™ technology makes it ideal for plant breeding applications where there are often hundreds of samples to be analyzed. KASP™ markers have been successfully used for marker-assisted selection in soybean (Rosso *et al.,* 2011), faba bean (Cottage *et al.,* 2012), oat (Gnanesh *et al.,* 2013), and wheat (Liu *et al.*, 2014). KASP™ markers in soybean and faba bean were developed to detect the presence of

beneficial mutations that improve nutritional value and distinguish determinate growth habit, respectively. The KASP™ markers developed in oat detected the presence or absence of the crown rust resistance gene *Pc91*. KASP™ has also been used to perform quality control genotyping in maize inbred lines (Semagn *et al.,* 2012). Used in this way, KASP™ provides a time- and resource-efficient method of assessing genetic identity and purity within a breeding program. KASP™ assays have also provided enough information in allotetraploid cotton to generate the first genetic linkage map composed entirely of SNP markers (Byers *et al.,* 2012).

KASP™ has been reported to have low error rates, low cost and high efficiency per reaction (Cortés *et al.,* 2011; Rosso *et al.,* 2011; Semagn *et al.,* 2013). For these reasons, KASP™ markers have been heavily used for crop improvement purposes especially at the International Maize and Wheat Improvement Center (CIMMYT). KASP™ is routinely used at CIMMYT for quality control analysis, QTL mapping, and recurrent selection (Semagn *et al.,* 2013). KASP™ has yet to be developed solely for cultivar diagnostics, but since these assays are scalable and do not require gel-based scoring it appears to be a good prospect for fingerprinting.

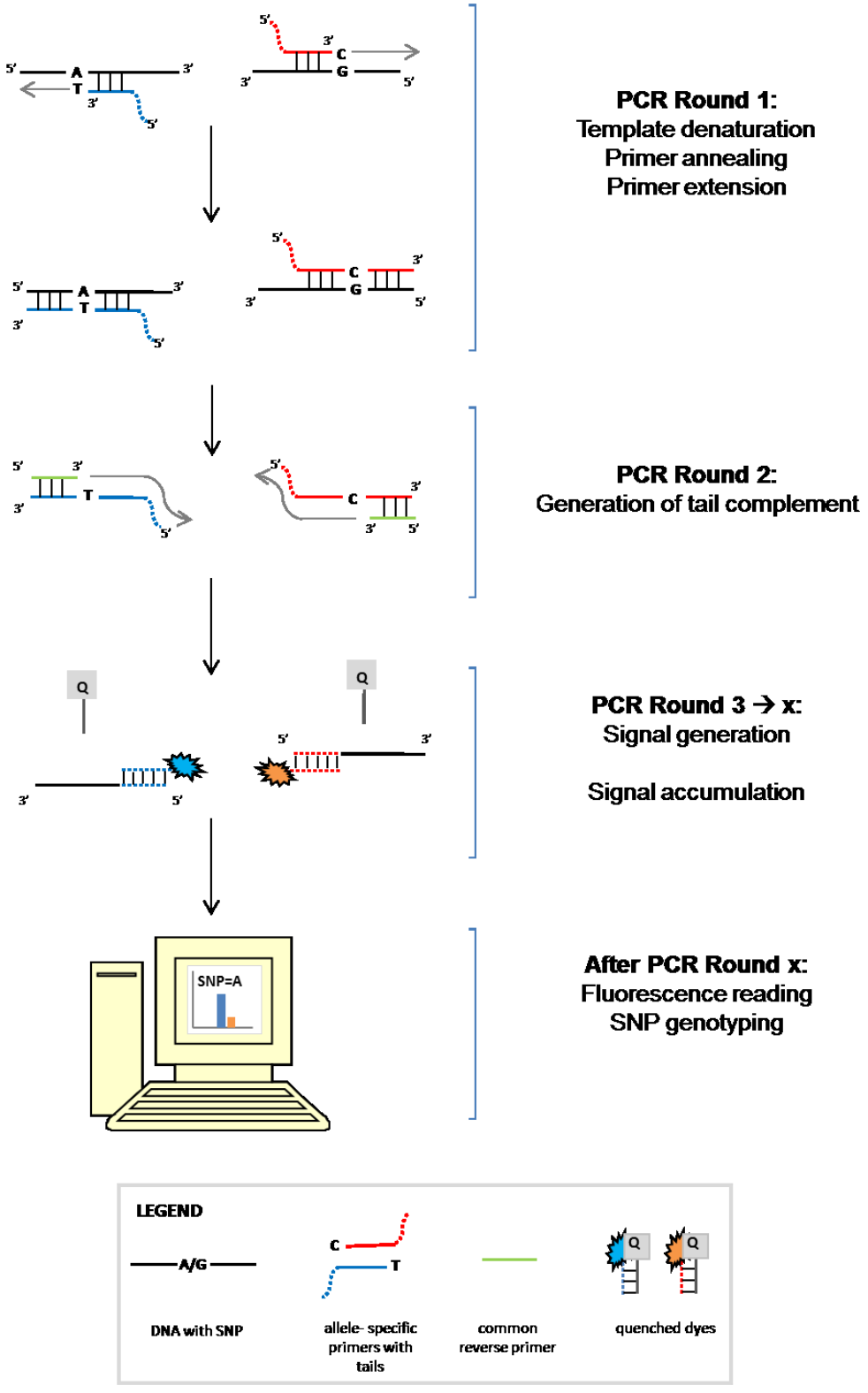

Figure 1.2 Genotyping using Kompetitive Allele-Specific PCR (KASP™). Adapted from: http://www.lgcgenomics.com/genotyping/kasp-genotyping-reagents/how-does-kasp-work/

1.4 Measures of genetic similarity

1.4.1 Marker-based: Genetic distance

Genetic distance gives an indication of the divergence between two populations (i.e. species, individuals, cultivars). This value is proportional to genetic relationships and diversity, in that a small genetic distance between two populations suggests that they are more closely related than two populations with a larger value. Cavalli-Sforza-Edwards (1967), Rogers (1972), and Nei and Li (1979) all developed methods to calculate genetic distance based on molecular markers. Markers common to two populations or samples could have arisen from separate mutations and are considered as "identical by state" (IBS). Alternatively, common markers can be derived from a common ancestor.

Nei and Li (1979) estimated genetic distance using electrophoretic DNA patterns generated from restriction endonuclease digests of mitochondrial DNA. They proposed genetic distance could be estimated using this equation:

$$\hat{F} = \frac{2n_{XY}}{(n_X + n_Y)} \qquad [1.2]$$

where n_{XY} is the number of common bands between populations X and Y; and n_X and n_Y are the number of bands in each respective population. This method is very popular due to its simplicity and has been used in crop species using molecular markers (Messmer *et al.,* 2004). The simple matching coefficient (Equation 3.1) is analogous to Nei and Li's distance.

1.4.2 Pedigree-based: Coefficient of parentage

Coefficients of parentage, also known as coefficients of co-ancestry (*f*), estimate the probability that two alleles at a homologous locus are "identical by descent" (IBD). This coefficient is calculated from pedigree information. Malécot (1948) was the first to describe *f* as:

$$f_{IL} = \frac{\sum \left(\frac{1}{2}\right)^{n_i+p_i} (1 + fJ_i)}{2} \quad [1.3]$$

where *n* and *p* are the number of preceding generations being considered of population I and L, respectively. If kinship is being calculated using information from only immediate parents, *n*=1, but if information for parents and grandparents are available, *n*=2, etc. The coefficient of inbreeding (fJ_i) is the probability that two homologous loci are identical. Programs have been written to compute *f*, such as KIN (Tinker and Mather, 1993) which refers to its output as kinship coefficients (r_k). KIN is used commonly for computing the genetic relationships between oat accessions (De Koeyer *et al.,* 1999, Hizbai *et al.,* 2012, Asoro *et al.,* 2013).

KIN is capable of calculating r_k from pedigree information stored as a text file. This is done under a number of assumptions:

a) when parents of an individual are unknown, r_k is assumed to be 0
b) r_k for an individual and itself is calculated from the inbreeding coefficient ($r_k=1/2(1+r_{pq})$)

c) $r_k = 0$ between ancestors with unknown pedigrees

d) alleles segregate randomly by diploid inheritance

1.4.3 Pedigree errors

Comparisons between genetic distance and coefficients of parentage have been conducted in maize (Messmer *et al.,* 1993), grape (Riaz *et al.,* 2008), barley (Tinker *et al.,* 1993), soybean (Cox et al, 1985), and oat (Paczos-Grzeda, 2004). Under ideal conditions, a comparison between these two measures should yield a straight-line correlation since they both correspond to genetic similarity between two individuals. However, strong deviations have been reported in barley, maize (Melchinger *et al.,* 1991) and oat. Outliers to this correlation could represent individuals that have undergone substantial selection, or they could represent errors in pedigree records. Cultivars once thought to be one cultivar can be shown to be another through molecular marker analysis. Using SSR markers, Bautista *et al.* (2008) were able to identify a mistake in the pedigree of 'Glenora' grape. Its previously thought parent, 'Black Monukka', was actually 'Black Kishmish' based on the obtained SSR profiles. The authors postulated that this mistake was likely due to a recording error (Bautista *et al.,* 2008; Riaz *et al.,* 2008).

1.5 Research Objectives

Plant breeding requires that cultivars are identified correctly and that traits are selected accurately. Misleading experimental results can occur when incorrect relationships and/or cultivar names are recorded in pedigrees. Molecular markers can identify these inconsistencies, and with advances in genotyping technology these diagnostics can be done faster and more objectively. We hypothesize that molecular markers can be developed for germplasm diagnostics and other potential breeding applications. Therefore, the objectives of this research project were to:

i. Determine if a small, manageable set of markers can be used to identify cultivars of interest that are registered and commonly grown in Canada;

ii. Use these markers to investigate variation among and within multiple seed sources from four cultivars; and

iii. Determine if a combination of marker and pedigree analysis can identify errors in pedigree records and putative carriers of favourable QTL alleles.

Statement of Contribution I

The following chapter is written as a manuscript in preparation for submission by Benazir K. Marquez, Charlene P. Wight, Jennifer Mitchell Fetch, Aaron D. Beattie, Douglas A. Johnson, Jesse Poland, and Nicholas A. Tinker titled “Diagnostic KASP™ and GBS SNP assays facilitate identification and purity assessment of oat cultivars”.

I, Benazir K. Marquez, assembled all experimental materials and performed all of the experiments and data analyses in the following manuscript with the exception of the following:

- Initial collection of seed (J. Mitchell Fetch, A.D. Beattie, W. Yan)
- Preliminary design and testing of 10 KASP™ markers (C.P. Wight)
- GBS library preparation (J. Poland)
- Contract-based high-throughput sequencing (A. Sharpe, National Research Council)

I, Benazir K. Marquez, wrote the original draft of this paper and all revised versions thereafter according to comments received from co-authors.

Benazir Katarina Marquez

Chapter 2

Diagnostic KASP™ and GBS SNP assays facilitate identification and purity assessment of oat cultivars

Benazir K. Marquez[1,2], Charlene P. Wight[1], Jennifer Mitchell Fetch[3], Aaron D. Beattie[4], Douglas A. Johnson[2], Jesse Poland[5], Nicholas A. Tinker[1*]

[1] Eastern Cereal and Oilseed Research Centre, 960 Carling Avenue, Ottawa, ON K1A 0C6, CANADA

[2] Department of Biology, University of Ottawa, 30 Marie Curie Private, Ottawa, ON K1N 6N5, CANADA

[3] Cereal Research Centre, 195 Dafoe Road, Winnipeg, MB R3T 2M9, CANADA

[4] Crop Development Centre, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, CANADA

[5] Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA

*corresponding author

2.1 Abstract

Identification and purity assessment of crop cultivars using quick, simple assays would benefit breeders and industry. Single SNP assays, such as “Kompetitive Allele-Specific PCR” (KASP™) are potentially the fastest and least expensive methods to distinguish cultivars, while genotyping-by-sequencing (GBS) shows potential to provide a parallel whole-genome assay at a low cost. We have applied both the KASP™ and GBS technologies to a set of current Canadian oat cultivars to evaluate the utility of these assays for identifying cultivars and detecting intra-cultivar variation. A set of 11 KASP™ markers (PIC≥0.28) was found to uniquely identify 22 of the cultivars studied. However, KASP™ detected a considerable

amount of heterogeneity in a set of 160 seeds originating from four cultivars and seed sources. GBS validated some of the KASP™ results, but revealed less intra-cultivar variation than was suggested by the KASP™ assays. The variation detected by both methods does not appear to be limited to a specific cultivar or seed source, reinforcing the concept that all cultivars are heterogeneous. Principal component analysis and graphical genotyping helped to distinguish physical contamination from true genetic heterogeneity/heterozygosity and to localize heterogeneity to specific chromosome regions. Localized heterogeneity in cultivars is expected in all inbreeding crop species. Here we have shown that this expectation can be calibrated within a genome-based purity assessment leading to more precise diagnostics. Based on these results, we recommend GBS for accurate cultivar identification and purity assessments and KASP™ for rapid or preliminary cultivar identification.

KEYWORDS

KASP™; genotyping-by-sequencing; *Avena sativa*; seed purity; cultivars

---

2.2 Introduction

Proper identification of plant cultivars and assessment of purity have always been important components of the agricultural value chain. Regulators use the identity of cultivars to facilitate seed certification and to enable the tracking and tracing of cultivars in commercial channels and international trade (Canadian Food Inspection Agency (CFIA), 2013). For a cultivar to be registered in Canada, an applicant must provide evidence of

distinguishing phenotypic characteristics that are “distinct, uniform, and stable” (DUS; CFIA, 2013). These characteristics are used to ensure that health and safety requirements are met and that Plant Breeders’ Rights (PBR) are respected. Plant breeders can apply for PBR in order to protect new cultivars as intellectual property. However, many phenotypic traits can be affected by the environment, and the range of traits that can be used in cultivar identification is limited. DNA fingerprints can serve as a complementary DUS characteristic, but are currently not required and cannot be used to replace morphological identity. Applications for cultivar registration or PBR would benefit from the inclusion of molecular data, especially in cases when the number of traits is insufficient to distinguish cultivars.

Many Canadian oat (*Avena sativa* L.) cultivars have been experimentally fingerprinted using different molecular marker systems including: restriction fragment length polymorphisms (O’Donoughue *et al.,* 1994), randomly amplified polymorphic DNA (Guillin *et al.,* 1998), amplified fragment length polymorphisms (AFLP; Fu *et al.,* 2004) and simple sequence repeats (SSRs). Currently, 123 cultivars are registered in Canada of which 40 have been experimentally fingerprinted using SSRs (Fu *et al.,* 2003a; Wight *et al.,* 2010). SSRs have often been favoured over single nucleotide polymorphisms (SNPs) for fingerprinting purposes because they are multi-allelic. However, SSRs are difficult to automate on a large scale, and the number of SSR assays available in oat is limited relative to other crops (Wight *et al.,* 2010). New marker technologies may provide alternatives to SSRs via low cost assays with better genome-wide coverage.

Considerable progress has been made in developing parallel genome assays such as Diversity Arrays Technology (DArT; Tinker *et al.,* 2009) and SNP-based genotyping arrays.

DArT technology has been used to identify quantitative trait loci (QTL) for important nutritional quality traits such as oil (Hizbai *et al.,* 2012), and agronomic traits such as vernalization response (Nava *et al.,* 2012) and *Fusarium* resistance (He *et al.,* 2013). SNP-based assays were used to develop the first physically-anchored map of oat (Oliver *et al.,* 2013). These new SNP assays are expected to be useful for multiple applications, however, they have yet to be explored as a resource for oat cultivar identification or seed purity assessment. A large-scale SNP array is being developed from the work in Oliver *et al.* (2013), but it may not be sufficiently rapid or cost-effective for such an application. Such assays require more time and effort for DNA preparation since samples are typically submitted in batches of 96 or more. KBioscience's Kompetitive Allele-Specific PCR (KASP™) would be more suitable for cultivar diagnostics because its single-plex nature allows for experiment size flexibility and time-efficiency. The KASP™ assay (Figure 1.2) combines allele-specific PCR with fluorescence resonance energy transfer (FRET) to quickly genotype samples at a specific locus. Each assay consists of two allele-specific primers, each designed with a 5' tail complementary to one of two FRET cassettes, enabling a particular fluorescence to be associated with each SNP allele. If multiple assays are done in parallel, this would provide a cost-effective, objective and practical tool for cultivar identification and seed purity evaluation of samples in small numbers.

KASP™ has already proven to be useful for other oat breeding applications such as marker-assisted selection (Gnanesh *et al.,* 2013). In spite of this, before KASP™ assays can be seriously considered for routine use as a source of DUS data, the method must be shown to be reliable at every stage of seed multiplication. However, neutral genetic variation can

remain as residual heterogeneity within a cultivar. Multiple molecular marker technologies have detected such variation within accessions of rice (Olufowote *et al.,* 1997), wheat (Soleimani *et al.,* 2002), oat (Wight *et al.,* 2010), and cowpea (Lucas and Huynh 2013). Variation within a cultivar may also indicate an error in seed handling. For example, SSRs applied to five random samples from cultivars grown in the 2007 Ontario Oat Performance Trial detected five instances of potential contamination (Wight *et al.,* 2010). This result was not surprising since seed harvested from performance trials is not intended for pure-line maintenance, however it does bring attention to the importance of considering seed sources when crossing or designing genetic studies.

Comprehensive assessment of seed purity requires the testing of large samples. According to the Canadian Seed Growers Association (CSGA), inspectors make six counts of 10, 000 plants when inspecting crops for impurities. In order to detect variability that may exist for only one trait in one plant, numerous characteristics such as flowering time, height, and inflorescence type must be observed in each plant. Therefore, it is likely that both a large sample as well as a large number of genetic markers would be required to assess varietal purity. Because of this scale and its corresponding cost, KASP™ technology may not be appropriate for the assessment of seed purity. The GBS method (Figure 1.1) employs one (Elshire *et al.,* 2011) or two (Poland *et al.,* 2012) restriction enzymes to digest the genome into fragments and then ligates a short "barcode" sequence to each prior to sequencing. These barcode sequences allow for several samples to be multiplexed into one lane for sequencing. They are later used to identify which fragment belongs to each sample. Following successful reports of GBS in other crop species such as barley and wheat (Poland

et al., 2012), and switchgrass (Lu *et al.,* 2013), GBS has been applied to oat (Huang *et al., submitted*). Since GBS can be applied at low cost and provides significant genome coverage, this method provides an opportunity for evaluation of seed purity.

The primary objectives of this study were to (1) develop and evaluate a set of KASP™ assays for rapid identification of key Canadian oat cultivars, and (2) assess genetic variability among and within four distinct seed sources using the developed assays and GBS.

2.3 Material and Methods

2.3.1 Plant material

A set of 24 current and representative Canadian oat cultivars were selected after consultation with breeders and available crop insurance reports (Table 2.1). This set is referred to hereafter as the Canadian Diagnostic Oat Cultivars (CDOC-24). Seed used in this study was obtained directly from plant breeders and/or staff at the Eastern Cereal Oilseed Research Centre (ECORC, Ottawa, ON), the Cereal Research Centre (CRC, Winnipeg, MB), the Agriculture and Agri-Food Canada Seed Increase Unit (SIU, Indian Head, SK), and the Crop Development Centre (CDC, Saskatoon, SK).

Eight seeds of each cultivar, from the purest source available, were grown in cyg seed-pack growth pouches (Mega International, St. Paul MN) for at least ten days. Young leaf tissue was harvested, pooled and dried for a minimum of ten days in paper envelopes containing Dry-Packs (Silica Gel Packets.ca, Harrisburg NC).

Table 2.1 A set of 24 oat cultivars selected as Canadian Diagnostic Oat Cultivars (CDOCs). Pedigrees were acquired from the Pedigree of Oat Lines (POOL) database (http://avena.agr.gc.ca/OGIS/, verified 30 Sept. 2013).

Cultivar	Seed		Breeding Program	Locations Grown[b]	Pedigree
	Type	Source[a]			
AC Assiniboia	increase	CRC	Winnipeg	MB	*Pc68/*7*Robert
AC Morgan	certified	CRC	Lacombe	AB,SK,MB	OT523/OT764
AC Mustang	trial	ECORC	Lacombe	AB,SK,MB	Cascade/Fraser
AC Rigodon	increase	ECORC	Sainte-Foy	ON,PQ	Ogle // RL2897*2 / Kent /3/ Q.O.174.19
Bradley	breeder	ECORC	Ottawa	ON,PQ	Ida//AC Aylmer/Goslin
Calibre	breeder	CDC	Saskatoon	AB,SK	Gemini/Clintford
Cantal	certified	ECORC	Sainte-Foy	ON,PQ	QO190.2/QO189.5
CDC Baler	breeder	CDC	Saskatoon	AB,SK,MB	Av2401/2/SO86044
CDC Dancer	breeder	CDC	Saskatoon	SK,MB	OT344/W90279
CDC Orrin	breeder	CDC	Saskatoon	AB,SK,ON,PQ	OT349/J775-1
Derby	breeder	CDC	Saskatoon	AB,SK,MB	Calibre/Cascade
Dieter	increase	ECORC	Ottawa	ON/PQ	*Pc*68/Donegal//Capital
Furlong	certified	CRC	Winnipeg	SK,MB	w93069(*Pg*16)/AC Assiniboia
Jordan	breeder	SIU	Winnipeg	AB,SK,MB	OT377/Ronald
Leggett	certified	CRC	Winnipeg	SK,MB	OT294/Pc94
Manotick	increase	ECORC	Ottawa	ON,PQ	AC Stewart/06075
Nice	certified	ECORC	McGill	ON,PQ	QO617.16/Donegal
Pinnacle	certified	CRC	Winnipeg	SK,MB	OT248/OT237//AC Medallion
Prescott	increase	ECORC	Ottawa	ON,PQ	OA973-1/AC Aylmer
Ronald	breeder	CRC	Winnipeg	SK,MB	W89329 (dwarf)/AC Medallion
Souris	increase	CRC	Fargo, ND,USA	MB	ND90141/ND900118
Summit	trial	CRC	Winnipeg	MB	Ronald/OT299
Triple Crown	certified	CRC	Svalöv, Sweden	AB,SK,MB	NOSN/WW17187//Sirene/Sang///WW17734
Waldern	trial	CRC	Lacombe	AB	Gemini/Cascade

[a] Source: CDC: Crop Development Centre (Saskatoon, SK), CRC: Cereal Research Centre (Winnipeg, MB), ECORC : Eastern Cereal Oilseed Research Centre (Ottawa, ON), SIU: Agriculture and Agri-Food Canada Seed Increase Unit (Indian Head, SK) [b] Locations are Canadian provinces where cultivars are recommended or commonly grown: AB: Alberta, MB: Manitoba, ON: Ontario, PQ: Quebec, SK: Saskatchewan

In addition to the representative sample above, three additional samples of seed from each of 'AC Morgan', 'Jordan', 'Leggett', and 'Ronald' were obtained so that the following sources were represented: breeder seed, certified seed, seed harvested from an internal cultivar increase, and seed harvested from a registration or PBR trial (Table 2.2). Ten seeds from each of these sources (a total of 160 seeds, referred to hereafter as the CDOC-160) were grown as previously described, but harvested and dried as single plants.

Table 2.2 Detailed information for the Canadian Diagnostic Oat Cultivars (CDOC-160) subset. Unknown information is indicated by (-).

Cultivar	Seed origin	Year harvested	Harvest Location	Trial
AC Morgan	breeder	2010	Indian Head, SK	
	certified	2005	-	
	increase	2010	Glenlea, MB	
	trial	2010	Portage la Prairie, MB	Plant Breeder's Rights
Jordan	breeder	-	Indian Head, SK	
	certified	2007	-	
	increase	2005	Indian Head, SK	Western Cooperative Oat Registration
	trial	2011	Glenlea, MB	Total Avenanthramide Trial
Leggett	breeder	2005	Indian Head, SK	
	certified	2009	-	
	increase	2004	Portage la Prairie, MB	Plant Breeder's Rights
	trial	2010	Portage la Prairie, MB	Plant Breeder's Rights
Ronald	breeder	2001-02	Lincoln, New Zealand	
	certified	2006	-	
	increase	2004	Glenlea, MB	
	trial	2007	Portage la Prairie, MB	Plant Breeder's Rights

2.3.2 DNA preparation

DNA for each CDOC-24 was extracted from 20 mg of dried bulked leaf tissue using a DNeasy Plant Mini kit[2] (Qiagen Inc., Mississauga ON) according to the manufacturer's protocol. DNA from a single leaf of each CDOC-160 plant was extracted using the same kit. Extracts were quantified using Quant-iT Picogreen dsDNA assays (Invitrogen, Carlsbad CA) according to the protocol reported by Blotta *et al.* (2005). CDOC-24 and CDOC-160 extracts were diluted to 2 ng/μL and 1 ng/μL working volumes, respectively, with sterile distilled water.

2.3.3 Kompetitive Allele-Specific PCR assay (KASP™) design

Fifty-nine KASP™ assays were designed by KBioscience's KASP™-on-Demand (KOD) service from SNPs previously discovered by sequencing of genomic DArT fragments (Tinker *et al.,* 2009), cDNA fragments (Oliver *et al.,* 2012), or from GBS assays (Huang *et al., submitted*). As part of the KOD service, KBioscience validated these assays using a set of 19 progeny from the 'Kanota' x 'Ogle' mapping population (Wight *et al.,* 2003) and a panel of 25 cultivars. Based on the clustering patterns of this data (not shown), 30 KASP™ assays were chosen to genotype the CDOC-24.

2.3.4 KASP™ genotyping

Assays were prepared with 4 μL genomic DNA, 4 μL 2x KASP™ Master Mix and 0.11 μL KASP™ Primer Mix. KASP™ 2x Master Mix contained the FRET reporting system (FAM, HEX) and PCR reagents. KASP™ Primer Mixes, unique for each SNP, contained two allele-specific primers and one common primer. Both mixes were obtained from LGC

[2] The effect of DNA extraction on KASP was evaluated using four additional methods and is further described in Appendix 1.

Genomics/KBioscience (Beverly, MA). Triplicate assays using CDOC-24 DNA were arranged on 96-well plates (Fischer Scientific Company, Ottawa ON) with a minimum of three no-template controls and a positive control sample ('Tardis'). Assays were carried out using a CFX96 Real-Time System with C1000 Thermal Cycler (BioRad, Hercules CA). Amplification and fluorescence reading were conducted using the following protocol: 94°C 15', 10 touchdown cycles of 94°C 20" then 65°C 1' (-0.8°C/cycle), 30 cycles of 94°C 20" then 57°C for 1', 5 cycles of 25°C 1' then plate read. Assays on the CDOC-160 samples were carried out once. Alleles were discriminated based on normalized fluorescence readings taken at the final protocol step using CFX Manager's (BioRad) automatic calling option. Allele discrimination plots (ADPs) were confirmed and adjusted by visual inspection of fluorescence data. Alleles were recorded as 1 (FAM) or 3 (HEX).

2.3.5 Genotyping-by-sequencing (GBS)

Libraries containing CDOC-160 DNA (one full and one partial 95-plex) were constructed using the P384 adapter set (Poland *et al.,* 2012b). A single randomly-placed blank well was included on each plate for quality control. Genomic DNA was digested simultaneously with *Pst*I and *Msp*I. Barcoded adapters were ligated to each individual sample following digestion. Samples from each plate were pooled into a single library and PCR-amplified. Detailed protocols can be found in Poland *et al.* (2012). Each 95-plex library was sequenced on a single lane of Illumina HiSeq 2000 at the National Research Council (Saskatoon, SK) by the DNA Technologies core facility.

2.3.6 Genotyping using GBS and the UNEAK pipeline

The universal network-enabled analysis kit (UNEAK) non-reference pipeline (Lu *et al.,* 2013, http://www.maizegenetics.net/gbs-bioinformatics), which is part of the TASSEL 3.0 bioinformatics analysis package (Bradbury *et al.,* 2007) was used for SNP discovery and genotyping. The primary GBS analysis was done using the full CDOC-160 set as the reference sample. Reads containing a barcode sequence followed by the expected *Pst*I (CTGCAG)-*Msp*I (CCGG) cut sites were retained and trimmed to 64 bp. Identical reads present in more than five samples were retained and collapsed into tags. Pairwise alignment was implemented with the retained tags to find tag pairs with a 1 bp mismatch, considered afterward as candidate SNPs. In order to reduce false positive SNP calls, a network filter was employed to identify only reciprocal tag pairs. An error tolerance rate of 0.03 was applied to filter out possible paralogs and sequencing errors. Initial quality and genetic filtering in the UNEAK pipeline was performed with default parameters for minor allele frequency (MAF, 0.05-0.5) and call rate (0-1). Additional genetic filtering was performed using an in-house program to retain only those SNPs showing maximum 5% heterozygosity, minimum MAF of 10%, and maximum 5% missing scores with CDOC-160 as the reference sample. Alleles were recorded for further genetic analysis as '1' (reference allele) or '3' (alternate allele) with '2' indicating a heterozygote. The reference and alternate alleles for each GBS SNP can be found in Table A.2.

For comparison, the above GBS pipeline was repeated using a reference sample that consisted of the CDOC-160 set plus an additional 418 mapping progeny and parents from six mapping populations (Oliver *et al.,* 2013). For this analysis, secondary filtering was

applied to retain SNPs with minimum MAF of 1% and maximum 50% missing scores to ensure that loci segregating at very low frequencies were not omitted.

2.3.7 Data analysis

Kinship coefficients (r_k, Equation 1.2) between pairwise combinations of the CDOC-24 entries were calculated using pedigree information stored in the Pedigrees of Oat Lines database (http://avena.agr.gc.ca/OGIS/) and the KIN program (Tinker and Mather, 1993). All cultivars, landraces, and breeding lines in the pedigrees were assumed to be 100% homozygous and homogeneous for the purpose of computing r_k. Polymorphic information content (PIC) was calculated for each KASP™ marker as:

$$PIC = 1 - (f_1^2 + f_2^2) \quad [2.1]$$

where f_1 is the frequency of allele 1 and f_2 is the frequency of allele 2 (Botstein *et al.,* 1980). Miscalls were recorded where a genotype was different from the results of two other identical assays, and a "miscall rate" was computed as:

$$\left(\frac{number\ of\ miscalled\ assays}{total\ number\ of\ assays\ completed}\right) \cdot 100\% \quad [2.2]$$

Heterozygous calls that differed from a homozygous call were considered as 0.5 of a miscalled assay. Diagnostic KASP™ assays were selected based on PIC > 0.18 (MAF > 10%) and a miscall rate of < 10%.

Dissimilarity was computed with GBS data (2,212 SNPs) using the Euclidean distance metric. Cluster analyses were performed using the *hclust* function included in the R statistical package. Principal component analysis (PCA) was performed with the *prcomp* function.

Graphical genotypes (GGs) were generated using GGT 2.0 (van Berloo, 2008) with 1,344 GBS SNPs using map locations determined by Huang *et al.* (*submitted*). A crossover event was graphically represented in the most likely position: the exact middle of two opposite genotypes. Level of polymorphism was calculated with GBS data (4,307 SNPs) as:

$$\left(\frac{number\ of\ polymorphic\ loci}{number\ of\ polymorphic\ loci + number\ of\ monomorphic\ loci}\right) \cdot 100\% \qquad [2.3]$$

Markers with PIC and heterozygous allele frequencies ≤ 0.09 were defined as monomorphic.

2.4 Results

2.4.1 Pedigree analysis

Pedigree analysis was conducted to observe the extent of kinship among the CDOCs. Kinship coefficients were calculated for all pairwise combinations and displayed as a matrix in Figure A.1. The average r_k within the CDOCs is 11.4% (range: 0.38%-85.02%). Pairs of cultivars with the greatest degree of co-ancestry generally originated from the same breeding programs, as expected. An UPGMA dendrogram based on r_k grouped CDOCs by breeding location (Figure 2.1). In some cases, immediate ancestors are included within this set. For example, since Jordan was bred from a cross involving Ronald (Table 2.1), these two CDOCs share an r_k of 66%. This value is higher than the expected value of 50% for a parent and its progeny, because Ronald and Jordan's other parent 'OT377' are also related. 'Triple Crown', the one cultivar originating outside of North America, is the least related to the other 23 cultivars (r_k=0.38%- 2.31%) based on known pedigree data.

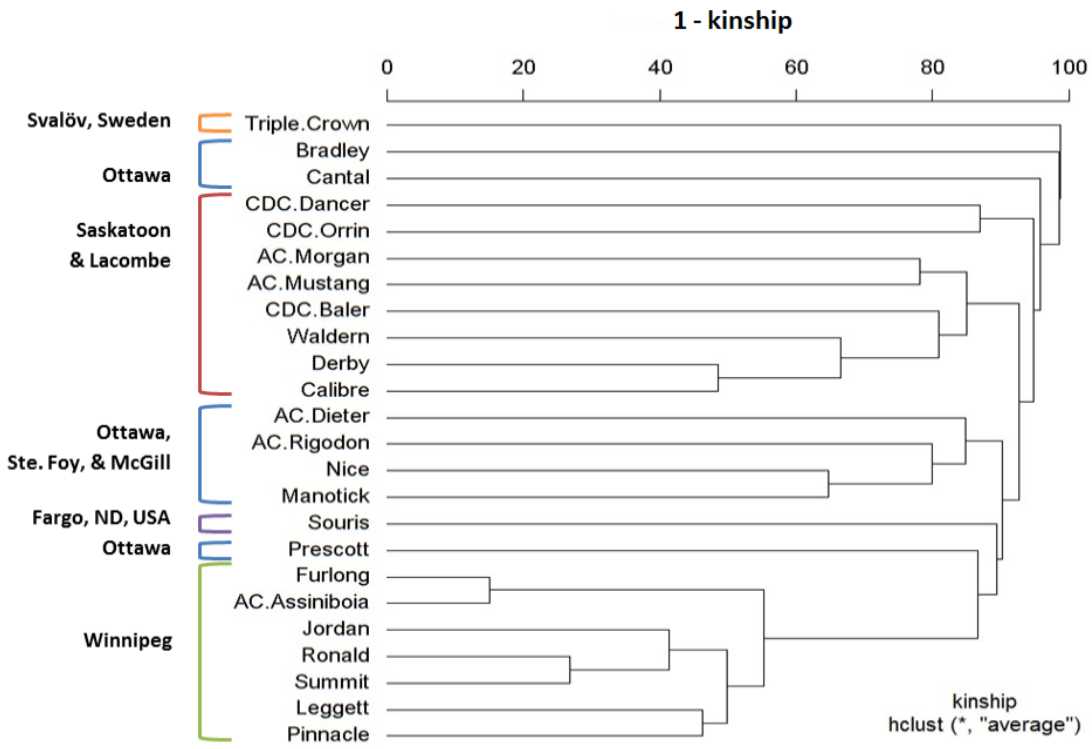

Figure 2.1 Dendrogram illustrating the kinship between the CDOC-24. Each node represents one cultivar. Kinship (r_k) was calculated using information acquired from the Pedigree of Oat Lines (POOL) database (http://avena.agr.gc.ca/OGIS/; verified 30 September 2013). The height scale represents 1- r_k.

2.4.2 KASP™ fingerprinting of 24 Canadian cultivars

A selected set of 11 KASP™ assays uniquely identified 22 of the CDOC-24 (Figure 2.2). The average length of allele-specific primer sequences, excluding KBioscience's proprietary tails, was 24 nt. The average length of common primers was 25 nt (Table A.1). 'CDC Baler' and 'Waldern' remained indistinguishable using these markers. One assay (K_ES01_c13432_100) revealed a heterozygote in the cultivar 'Summit', thus distinguishing it from all other CDOCs. These KASP™ markers have an average PIC of 0.4 (0.28-0.50) and miscall rate of 3% (0%-6%) (Table 2.3). Based on linkage analysis of SNPs, from which

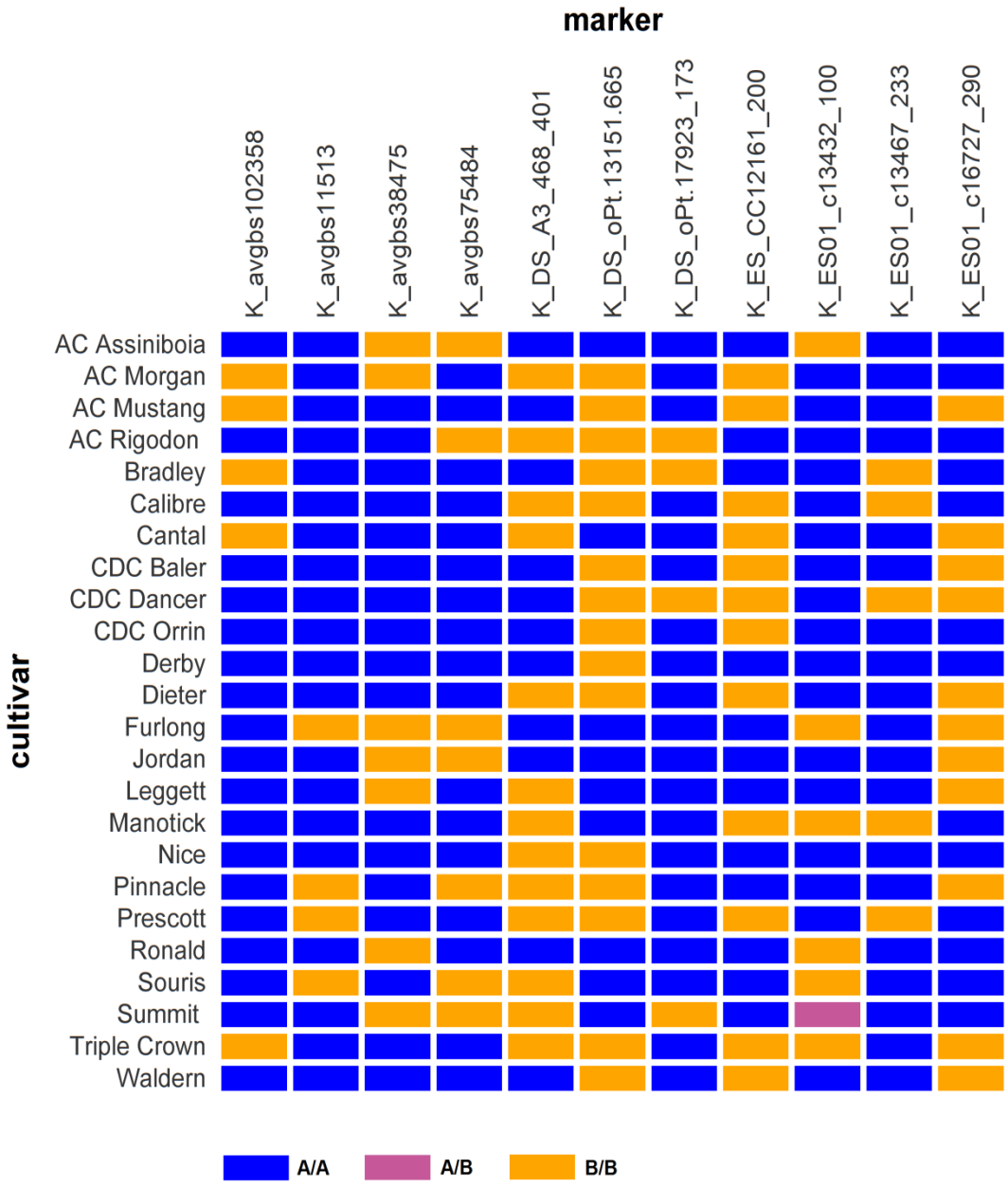

Figure 2.2 Bitmap of genotype profiles of the Canadian Diagnostic Oat Cultivars. Alleles A (blue) and B (orange) correspond to the alleles associated with FAM and HEX in each KASP assay, respectively (Table 2.3). Heterozygotes are presented in purple.

the KASP™ markers were developed (reported elsewhere), this set of KASP™ markers is expected to represent seven of the chromosomes of oat (Table 2.3). Close proximity of two pairs of diagnostic markers is coincidental, as these map positions were determined afterward.

Table 2.3 Assay characteristics and map location of 11 KASP™ markers selected as a minimal set to diagnose identity of Canadian oat cultivars. Unknown information is indicated by (-).

MARKER	SNP Source	Allele[a]		Consensus Map (Oliver *et al.*, 2013)		Statistics	
		A	B	Chromosome	Pos (cM)	PIC[b]	Miscall%
K_ES01_c16727_290	EST	C	G	3C	75	0.50	3%
K_ES01_c13467_233	EST	A	G	1C	37	0.33	4%
K_ES01_c13432_100	EST	A	G	14D	97	0.41	2%
K_ES_CC12161_200	EST	T	C	-	-	0.50	6%
K_DS_oPt.17923_173	DArT	T	G	20D	0	0.28	3%
K_DS_oPt.13151.665	DArT	T	G	19A	0	0.47	0%
K_DS_A3_468_401	DArT	T	G	21D	51	0.50	6%
K_avgbs75484	GBS	T	G	9D	23	0.41	1%
K_avgbs38475	GBS	T	C	19A	41	0.41	0%
K_avgbs11513	GBS	T	C	3C	78	0.28	0%
K_avgbs102358	GBS	A	G	9D	21	0.33	4%

[a] FRET cassette oligos are labelled with FAM (6-carboxyfluorescein) or HEX (6-carboxy - 2',4,4',5',7,7') in each KASP™ assay. Each oligo is complementary to the tail sequence of one allele-specific primer, allowing for FAM and HEX to discriminate alleles A and B, respectively.

[b] Polymorphic information content (PIC) was calculated from CDOC-24 material.

2.4.3 Variation among and within seed sources detected using KASP™

To investigate variation within cultivars, subsets of breeder, certified, increase, and trial seed from Jordan, AC Morgan, Leggett, and Ronald (CDOC-160) were genotyped with the KASP™ diagnostic set (Figure 2.3). K_ES01_c13467_233 and K_DS_oPt.17923_173 were monomorphic among the four cultivars studied as expected (Figure 2.2). Variation among and within seed sources was detected for all four cultivars. Four markers were polymorphic among the set of 40 Jordan seeds, five markers among Leggett, five markers among Ronald, and six among AC Morgan. Genotype profiles for each seed were compared to those established earlier in the bulked reference sample. AC Morgan seeds were the most consistent, with 21 of 40 (52.5%) being identical to the reference profile (Figure 2.2) at all loci. Ronald, Jordan, and Leggett followed with 15 (37.5%), 7 (17.5 %), and 4 (10 %) consistent seeds, respectively. The inconsistencies did not appear limited to a particular seed source. Most of the detected variation fell outside of the estimated miscall rate, with the exception of K_DS_A3_468_401 in Ronald genotypes. KASP™ was observed to always detect substantial variation at loci expected to possess allele B (i.e. orange loci in Figure 2.3) while those expected to have an A allele (blue) were usually consistent. Triplicate KASP™ assays performed with the CDOC-24 reference samples of Jordan, AC Morgan, Leggett, and Ronald were consistent, with the exception of K_avgbs102358 in AC Morgan and K_avgbs75484 in Ronald. Both assays produced one miscall out of three replicates.

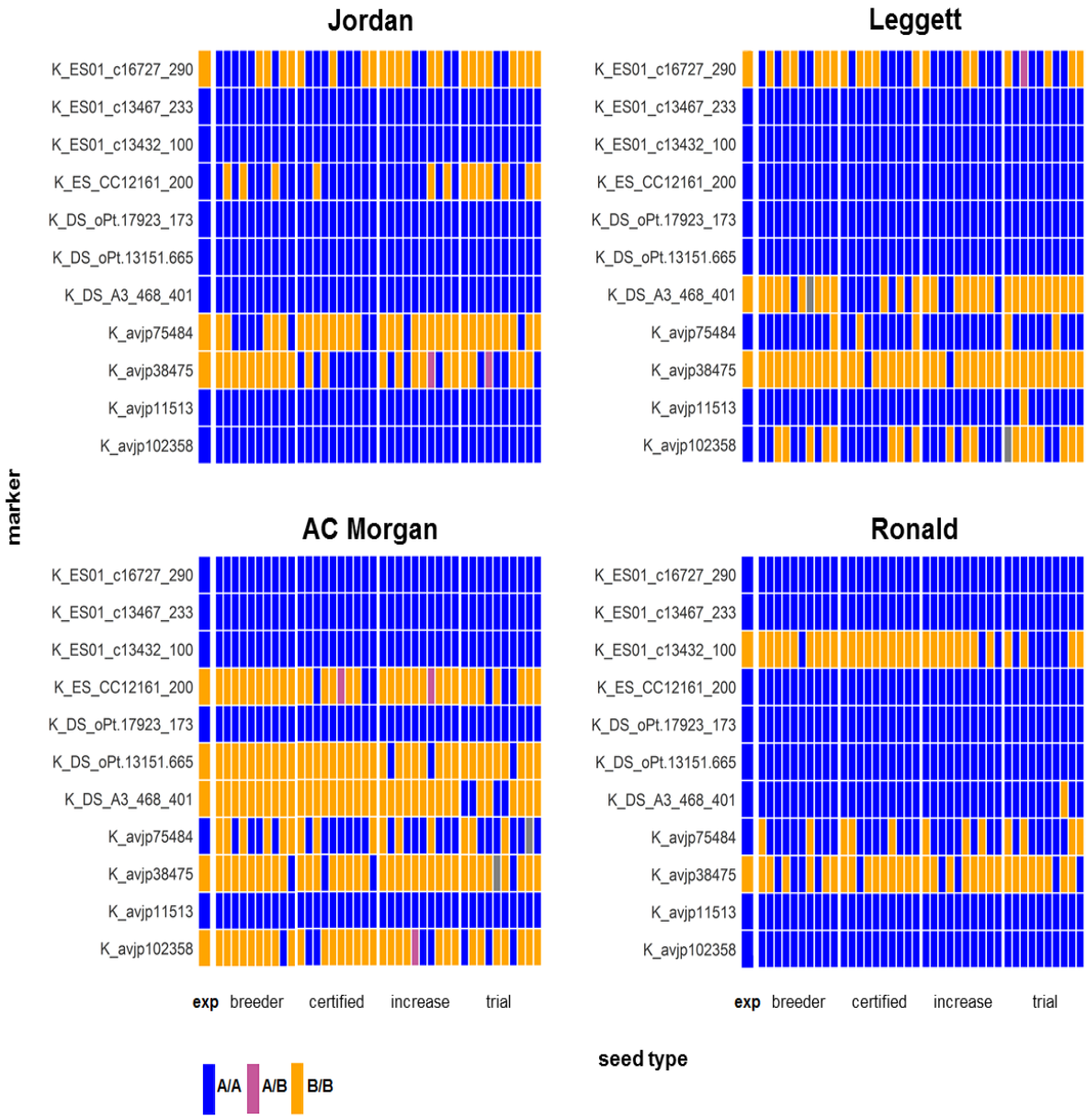

Figure 2.3 Bitmaps illustrating the heterogeneity detected by KASP™ within the CDOC-160. Alleles A (blue) and B (orange) correspond to the alleles associated with FAM and HEX in each KASP™ assay, respectively (Table 2.3), with heterozygotes shown in purple. Missing data is indicated in grey. Expected (exp) genotype profiles based on reference samples (Figure 2.2) are shown to the left of each cultivar panel. Samples from each cultivar are organized by seed source and separated by vertical white spacing.

2.4.4 Variation among and within seed sources detected using GBS

GBS was performed on the same CDOC-160 DNA to validate results of KASP™ assays and to explore intra-cultivar variation using a high density whole-genome assay. A total of 2,212 SNPs were called with stringent filtering parameters. Of these 1,344 (60.8%) were also mapped in an oat consensus framework (Huang *et al., submitted*). Only 21.8% of all GBS loci (20.5% of mapped loci) showed substantial variation (defined as MAF > 0.18) within one or more oat cultivars. Thus, while GBS provided a much larger set of diagnostic loci, each GBS locus was less likely to be polymorphic within a cultivar than one of the selected KASP™ assays. Most of the variation observed using GBS fell within distinct map regions and showed allele frequencies between 40% and 60% in the cultivar in which polymorphism was observed (Table A.3).

SNPs were also called from a larger reference sample consisting of the CDOC-160 and six mapping populations to evaluate the influence of using a larger reference set with relaxed filtering parameters. In this analysis, a much larger set of 21,931 SNPs were called, of which 4,307 (19.6%) were mapped to the oat consensus framework. We considered the 4,307 mapped loci to represent a sample of GBS loci that is relatively unbiased by the CDOC-160 samples. Of these, 3,905 were polymorphic in the CDOC-160, representing a polymorphism rate of 90.7%. The majority (87.6%) of SNPs called from only the CDOC-160 were also called in the larger reference sample. An additional 734 markers with PIC ≤ 0.1 were called using the alternate reference sample with relaxed stringency (Table A.4). These markers did not reveal any previously undetected heterogeneity, but variation within AC Morgan samples did become more noticeable.

Principal component analysis (PCA, Figure 2.4A) with the 2,212 stringently filtered SNPs revealed that all seeds clustered by cultivar, but that all cultivars contained some allelic variation. The first two principal components (PCs) accounted for 45% and 26% of the variation, respectively. Jordan, AC Morgan and Ronald formed tight clusters, while Leggett formed two noticeably separate clusters. One Ronald trial seed (Rt7, Figure 2.4A) clustered tightly with seeds of Leggett suggesting it is a contaminant. One Leggett increase seed (Li2, Figure 2.4A) fell farther from the rest of the Leggett sample which suggests that Li2 may not be Leggett nor any of the other three cultivars evaluated. Rt7 and Li2 were both omitted from their respective cultivar datasets in further analyses to avoid bias.

To further investigate the genetic heterogeneity of each cultivar, PCAs were performed for each cultivar using only polymorphic SNPs (Figure 2.4B). A Leggett seed (Li9), two AC Morgan increase seeds (Mi5, Mi6), and a Ronald breeder seed (Rb6) were also omitted from these analyses due to high proportions of missing data. Subsequent PCA of each cultivar revealed the heterogeneity within the sampled seeds in better detail. The first and second principal components explained an average of 11.5% (6.8-11%) and 8.95% (6.6-12.7%), respectively, of the variation in each PCA. Jordan seeds appear to be the most heterogeneous, followed by Ronald, Leggett and AC Morgan based on cluster tightness observed in individual PCA plots (Figure 2.4B). Jordan separated into four clusters, none of which reflected seed source. Leggett seeds segregated into two clusters mirroring the results of the first PCA. Ronald and AC Morgan both appeared as tight clusters with a few seeds distinctly different.

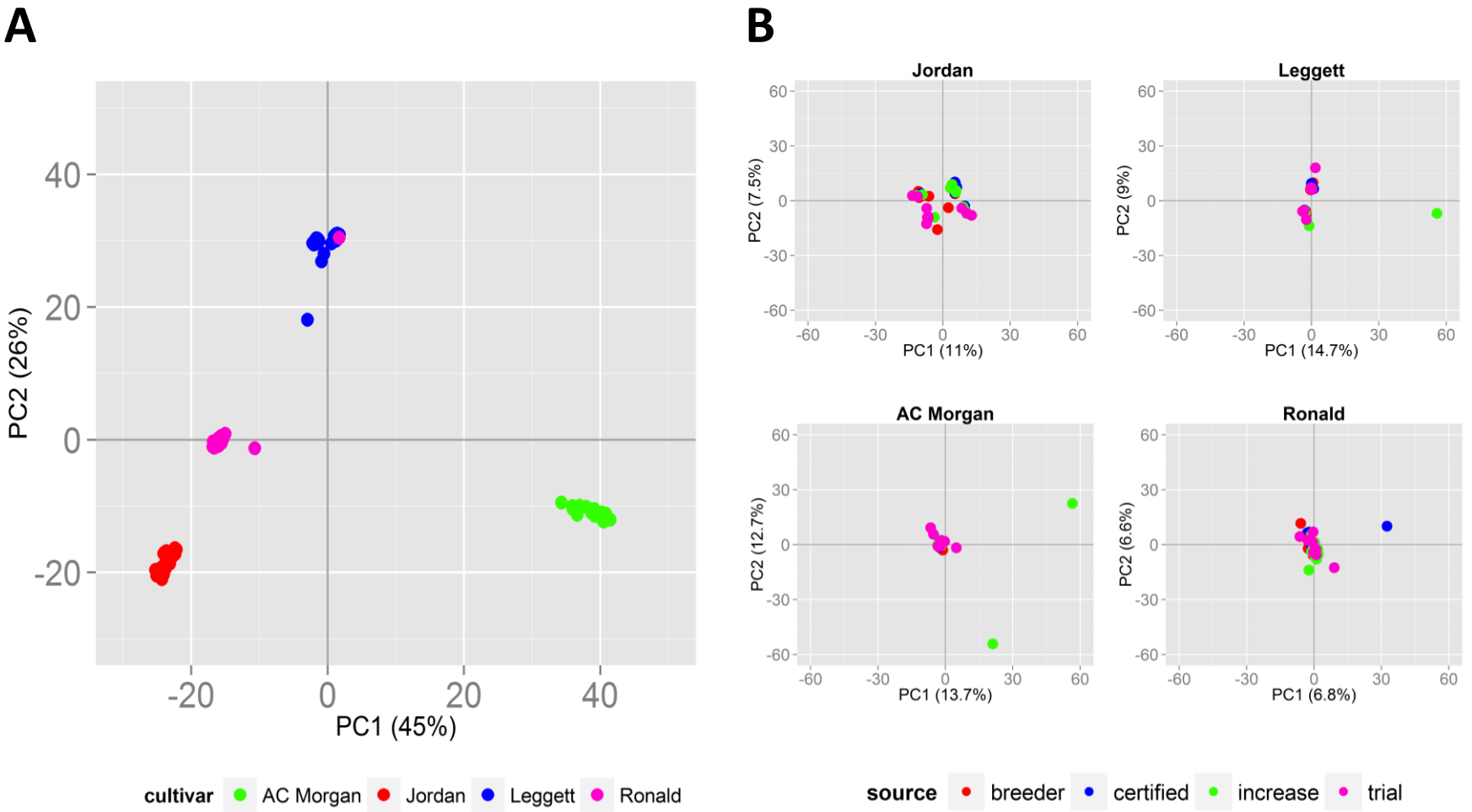

Figure 2.4 Principal component analyses (PCA) of genotyping-by-sequencing SNPs in (A) oat cultivars 'AC Morgan', 'Jordan', 'Leggett', and 'Ronald' and (B) within cultivars among breeder, certified, increase, and trial seed. Each circle represents one seed sample.

Degree of variation did not appear related to seed origin in any of the four cultivars. Average heterozygosity of each seed source was 0.79%, 0.75%, 0.69%, and 0.59%, for breeder, certified, increase, and trial seed, respectively. This pattern reflects the expectation that heterozygosity decreases with successive generations of self-propagation.

2.4.5 Graphical genotyping of GBS SNPs

GGs based on the 1,344 mapped GBS SNPs allowed for localization of heterogeneity and comparison to the approximate regions of KASP™ loci. Regions containing nine KASP™ loci were homogeneous within most cultivars based on GBS loci, while heterogeneity was revealed in two KASP™ regions: at K_avgbs38475 (chromosome 19A, 45 cM) in Leggett and Ronald seeds; and at K_DS_A3_468_401 (chromosome 21D, 51 cM) in AC Morgan trial seeds. The region containing the K_avgbs102358 and K_avgbs75484 loci on chromosome 9D (20-23 cM) was also observed to be extremely variable in the 40 Leggett seed samples (Figure 2.5). These seeds could also be discriminated into two separate groups using the genotype data generated by KASP™ at the K_avgbs102358 locus (Figure 2.3). These groups reflected the two clusters observed in the PCA score plot (Figure 2.4B). When the markers from this region were removed from the Leggett dataset and the remaining markers were used to perform PCA, the Leggett samples formed one distinct cluster (data not shown).

GGs also confirmed that Rt7 and Li2 varied significantly from their respective cultivars. Rt7 differed from Ronald seeds at every chromosome, while Li2 differed from Leggett seeds at chromosomes 1C, 3C, 4C, 5C, 14D, 16A , 17A-7C, and 19A (Figure A.3). Additionally, two AC

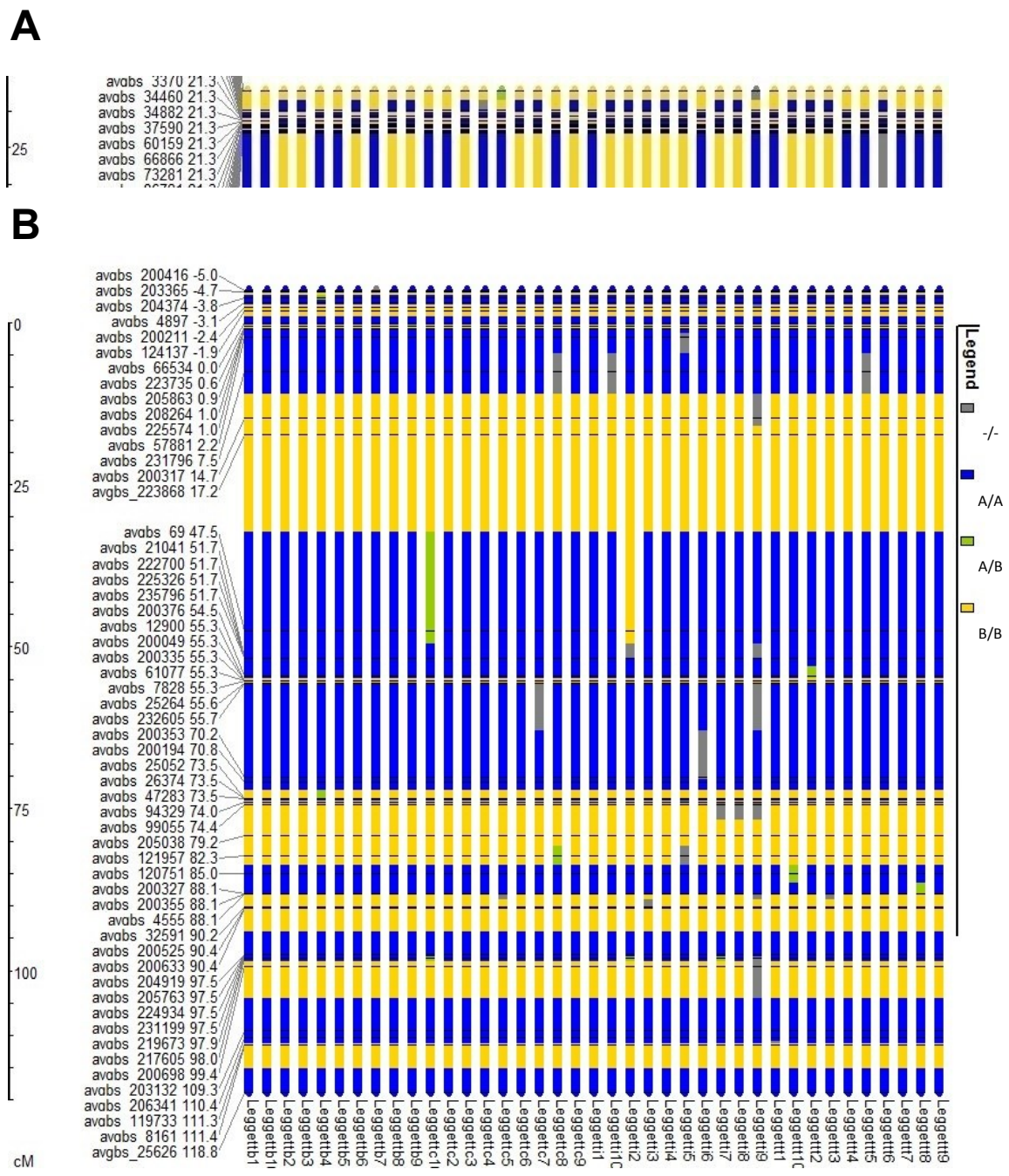

Figure 2.5 Graphical genotypes of 'Leggett' seeds demonstrating (A) localized heterogeneity on chromosome 9D (20-25 cM) and (B) homogeneity on chromosome 19A. Alleles A and B are colour-coded blue and yellow, respectively. Heterozygous alleles are colour-coded in green and missing data in grey.

Morgan seeds (Mt1 and Mt7) were also revealed to vary substantially at chromosomes 1C, 5C, 6C, 13A, 19A, 20D, and 21D (Figure A.3) compared to the other AC Morgan samples.

2.5 Discussion

2.5.1 CDOC-24 Fingerprinting

The original motive for this study was to establish a set of KASP™ assays capable of identifying key Canadian oat cultivars. We determined that a small set of 11 KASP™ assays are sufficient for identifying 22 of the CDOCs studied, leaving only CDC Baler and Waldern indistinguishable from each other. The CDOCs are all registered in Canada and were chosen due to their prevalence across the country. While these criteria allowed for economically-important cultivars to be chosen, it did introduce the difficulty of increased genetic relatedness within the set. Many of the CDOCs were bred in the same breeding program; consequently parents of some CDOCs are also included in the set. CDC Baler and Waldern were bred from different programs: the Crop Development Center and the Lacombe Research Centre, respectively (Table 2.1). However, these cultivars still share two common ancestors within three generations. CDC Baler and Waldern share a kinship coefficient (r_k) of 12.26, meaning the likelihood of inheriting the same allele from the same source is 12.3%. This suggests the possibility that the developed KASP™ assays are detecting SNPs that were inherited by both cultivars from one or both of their common ancestors. Nevertheless, more closely related cultivars can be identified from each other (e.g. 'AC Assiniboia' and 'Furlong', r_k=85.02). These differences are also due to sampling a small number of diagnostic loci, some of which originate from common genetic regions. Four of the 11 assays probe SNPs in similar regions: K_ES01_c16727_290 and K_avgbs11513 map

within 3 cM of each other on chromosome 3C, while K_avgbs102358 and K_avgbs75484 map within 2 cM of each other on chromosome 9D. Map locations of GBS loci were updated after these KASP™ assays were developed, when they were initially thought to be located in distinct regions.

Collectively, the developed KASP™ assays were estimated to have an average miscall rate of 3%, which is comparable to Illumina's GoldenGate assays in wheat and maize (Akhunov *et al.,* 2009; Yan *et al.,* 2009). To the best of our knowledge, only a few studies using KASP™ with polyploid species have reported any measure of genotyping error. However, a scientific review of the KASP™ method in maize and wheat has reported genotype mismatch rates comparable to the miscall rates for each of the 11 diagnostic KASP™ assays (Table 2.3). Mismatches among maize and wheat positive controls have been stated to vary from 0% to 2.9% for samples genotyped on the same plate, and 0% to 3.1% for samples genotyped on separate plates (Semagn *et al.,* 2013). Conversely, a very low genotyping error of 0.4% was reported in replicated KASP™ assays done with autotetraploid potato cultivars (Uitdewilligen *et al.,* 2013), but this error rate was estimated from a single cultivar. The miscall rates reported in this study used results from all 24 cultivars selected, thus it should provide a more conservative estimate of genotyping error.

The presence of multiple paralogous loci, as is the normal case in hexaploid oat, can produce a genotyping artifact described as a cluster shift. This refers to a pattern observed on genotyping plots where the clusters have shifted from their expected position due to detection of a paralogous locus. When loci were being chosen for KASP™ assay design, any which showed a strong cluster shift and/or appeared to interrogate multiple loci (i.e.

displayed more than two or three primary clusters) were discarded. Additionally, KASP™ primers are designed so that the allele-specific nucleotide is at the 3' end, which lowers the probability of extension if primers hybridize to similar but not identical sequences. However, paralogous loci could still have interfered with the target loci to different extents because of small variations in PCR conditions between replicated assays. The interference of highly similar loci would produce fluorescence readings that could change the ratio of FAM:HEX, thereby causing opposite allele calls. The same logic can be applied to repetitive sequences that are similar to the targeted sequence surrounding the SNP. Since identical replicates revealed miscalls, we would recommend that any critical diagnostics done with this set of markers should always be replicated at least three times.

Another complication in genotyping hexaploid oat, or any polyploid, is the presence of identical polymorphisms at loci in homoeologous chromosomes. A hypothetical A/B homozygous SNP present in all three genomes of oat could produce eight possible genotypes: AAAAAA, AAAABB, AABBAA, BBAAAA, BBBBAA, BBAABB, AABBBB, and BBBBBB. A KASP™ assay that targets a locus such as this would generate four clusters on an ADP. Any designed assays with these patterns were not chosen to genotype the CDOCs, and ADPs generated after genotyping did not exhibit a multi-cluster pattern. Therefore this type of polymorphism is not likely to have complicated the genotype results presented here.

The four GBS-derived KASP™ assays produced fewer miscalled genotypes than those derived from DArTs and ESTs. GBS loci mostly occur in non-genic regions of the genome (Elshire *et al.,* 2011), which evolve more rapidly than ESTs and genic DArTs. Therefore the context sequence surrounding a GBS locus, which is used to interrogate the SNP, may be

less prone to duplication and thus more reliable. Although these observations are based on a relatively small sample of loci, they suggest that GBS will likely provide a good source for future KASP™ marker design.

Caution should also be applied when using this set of assays beyond the 24 cultivars studied here. The genotype profiles established in this study are only likely to distinguish a CDOC from the other 21 identifiable CDOCs. The KASP™ diagnostic set was able to detect heterozygous loci in some of the CDOCs, but only reliably in Summit using K_ES01_c13432_100. It is important to remember that the heterozygous genotype is obtained from a bulked sample of seed, and that it likely represents a homozygous but heterogeneous locus. This single assay could be useful as a simple test for that cultivar, provided the samples are based on multiple bulked plants. This small set of KASP™ assays may not have differentiated all 24 CDOCs, but it does provide a foundation and fingerprinting protocol for other KASP™ assays with oat.

2.5.2 Variation detected in the CDOC-160 using KASP™

KASP™ assays detected a surprising amount of variation within the CDOC-160. Only 52.5% of AC Morgan seeds matched the reference profile established with the CDOC-24, while Leggett seeds were the most inconsistent with only 10% of seeds matching the reference profile. This amount of variation is considerably higher than what has been reported using SSRs (Wight *et al.,* 2010) and AFLPs (Fu *et al.,* 2006). If we were to consider that the seeds assayed here were contaminants introduced from seed handling errors, it would imply that all the sources we sampled from were substantially impure. When an oat crop is officially inspected, six counts of 10,000 plants are evaluated visually for key morphological traits.

Within crops grown from certified seed, the CSGA allows only five plants from different oat cultivars to be present (CSGA, 2013). This constitutes a maximum impurity content of 0.05%, which is drastically lower than the variation detected by KASP™. However, this maximum is almost certainly underestimated, because it is based on observable traits that have been deliberately selected for homogeneity. In contrast, the KASP™ markers target allelic variants that are likely to be phenotypically neutral.

The high KASP™-based variation within the CDOC-160 may be explained partially by miscall rates. As seen with the CDOC-24, triplicate assays are recommended when using this diagnostic set as multiple factors could influence the final allele call. However due to practical limitations, the CDOC-160 could only be genotyped once. This has a definite impact on interpretation, exemplified by the substantial variation detected only at loci expected to possess allele B. Two loci: K_avgbs102358 and K_avgbs75484 both produced one miscall among three replicated assays in AC Morgan and Ronald (CDOC-24). This suggests that if the CDOC-160 genotyping had been repeated, KASP™-based variation would have been less extensive.

Furthermore, the CDOC-160 variation is being judged based on the genotype profiles established using the CDOC-24. A comparison of AFLP data collected in bulked and single-plant samples from five oat cultivars suggested that bulking reveals less variation (Fu *et al.,* 2006) and may in fact homogenize any within-cultivar variation (Fu *et al.,* 2003b). Therefore, the bulking of CDOC-24 plants may have introduced a bias toward the alleles detected in the initial seed sources of AC Morgan, Jordan, Leggett, and Ronald. It also cannot be ignored that only 11 markers were used to assess genetic variation in the CDOC-

160. A previous study in rice suggests using at least 24 EST-SSR markers for assessing seed genetic purity (Moorthy and Babu, 2011). It is for this reason that we tested the alternate approach based on GBS.

2.5.3 Variation detected in the CDOC-160 using GBS

GBS revealed significantly less genotype variation than KASP™, however a fifth of the loci genotyped were still variable (MAF≥0.18) in at least one of four cultivars. PCA with this data revealed that the CDOC-160 could be clearly differentiated into their respective cultivars (Figure 2.4A). Thus, within the 2,212 GBS loci presented here, many additional cultivar-specific loci could potentially be converted into more reliable KASP™ assays. However, not all GBS can be converted because the GBS context sequence is only 64-bases long, while the KASP™ assay generally requires at least 20 bases on each side of the SNP. PCA also revealed two plausible contaminant seeds (Li2 and Rt7) in the Leggett and Ronald subsets (Figure 2.4A). GGs of these two seeds clearly demonstrate substantial variation from the remaining Leggett and Ronald seeds (Figure A.3). There is no conclusive way of determining if Rt7 was introduced into the Ronald seed stock by seed handling errors in the field or during sampling for this study. However, it does appear that a seed handling error outside of this study is responsible for Li2. GBS suggests that it does not originate from Jordan, AC Morgan, or Ronald. Additionally, KASP™ data for Li2 does not correspond with any of the genotype profiles for the CDOC-24, therefore it is unlikely to be a seed from any of the cultivars studied. Impurities of one seed in 40 (2.5%) exceed the standards prescribed by the CSGA, however in both cases the potential contaminant seed was observed in a trial (Rt7) or an increase (Li2) which would not have been inspected by CSGA or used within the certified

seed system. Seed impurities detected by molecular means have been reported to be 2-3% higher than those determined by morphological traits (Moorthy and Babu, 2011).

Beyond these two contaminant seeds, most of the variation observed using GBS assays fell within distinct map regions and showed allele frequencies of approximately 50% in the cultivar in which polymorphism was observed (Tables A.3-A.4). This result is consistent with the expectation that most cultivars contain residual heterogeneity in distinct chromosomal regions that were heterozygous in the last generation from which seed was bulked from a single plant (i.e. the founding generation).

Seed is usually bulked from selected plants at a generation such as F_4 (Tinker, 2008) but this practice varies among breeding programs. Both Leggett (Mitchell Fetch *et al.,* 2007) and Ronald (Mitchell Fetch *et al.,* 2003) were bulked at F_6, while Jordan was bulked at F_5 (Mitchell Fetch and Brown, 2009) and AC Morgan at F_{10} (Kibite and Menzies, 2002). Differences between the degree of variation within each cultivar were revealed by PCA (Figure 2.4B), with cluster tightness generally representative of the generation in which seed was bulked. Since Jordan, Leggett, and Ronald were bulked in the earlier generations, they show more diversity than AC Morgan which was bulked latest. AC Morgan shows the least variation and is represented by the tightest cluster. However, this trend can be affected by other factors. Progeny that are segregating in chromosomal regions containing major adaptive loci will often be eliminated during the selection phases, since they will not be phenotypically homogeneous. Variants that slip through this selection may be removed by roguing later in the breeding cycle. The initial number of segregating loci is also affected by the diversity of parents, since related parents will contain identical chromosome regions

that will not segregate. Nevertheless, the amount of neutral variation present in bulked progeny can be substantial, and this is reflected in the results presented here.

Interestingly, Leggett samples formed two distinct clusters in the PCA (Figure 2.4) which is also reflected in the two groups divided by K_avgbs102358 in the KASP™ results. Although all cultivars contained variable regions, this distinct split may represent a region containing a disproportionately large number of heterogeneous loci. Upon inspection of the GGs, the variation in Leggett responsible for dividing the seeds into two groups can be localized to chromosome 9D in the 20-25 cM region (Figure 2.5A). A review of the oat QTL literature did not reveal any significant QTL in this region, supporting the hypothesis that residual variation is most likely to remain where effects are neutral.

A caveat of using the stringently filtered dataset of 2,212 SNPs is that loci with low PIC (nearly monomorphic) may be omitted. Such data is more error prone, but may also be informative for seed purity assessment. Reducing the filtering parameters allowed for 734 additional markers (PIC ≤ 0.1) to be observed. These additional markers were more prevalent within AC Morgan, but were mostly located on the same chromosomes where original GGs revealed variation. However, the addition of these almost monomorphic loci produces a more noticeable difference between Mt1, Mt7, and the other AC Morgan samples (Table A.4). While studying a smaller, cleaner set of data provides more than sufficient material for seed purity assessment, one must be mindful of the potentially informative data being filtered out.

GGs also revealed that GBS loci were heterogeneous in genomic regions where KASP™ markers were polymorphic in AC Morgan, Leggett, and Ronald. This suggests that the variation revealed by KASP™ may not be entirely caused by miscall rate or comparison bias to CDOC-24. As such, KASP™ remains a possible tool in assessing seed purity albeit better suited for smaller sets of samples. GBS is recommended for larger sets of samples as it provides a suitable amount and quality of genotype data without performing replicate assays.

Graphical genotyping using GBS data has shown the potential for visualizing the chromosomal location of heterogeneity within cultivars. This method of data interpretation would be suitable for breeders as an efficient tool for selection, and for inspectors as a method of discovering possible seed contaminants. It may also allow for strategic development of near-isogenic lines or fine mapping strategies to isolate QTL that are hypothesized to lie within a short region of heterogeneity in a cultivar.

2.6 Conclusion

This study evaluated the utility of KASP™ and GBS for identifying cultivars and detecting intra-cultivar variation. KASP™ results revealed that a small set of markers is sufficient for identification among 22 Canadian oat cultivars. KASP™ also revealed a substantial amount of heterogeneity within cultivars, which was only partially validated by GBS. The intra-cultivar variation detected by both technologies did not appear to be limited to a specific cultivar or seed source, supporting the expectation that all cultivars are heterogeneous. Graphical genotyping of GBS data allowed for genomic heterogeneity to be localized to

specific chromosome regions and proved an appropriate method for purity assessment in larger sets of seeds. Therefore, we recommend KASP™ for rapid cultivar identification and GBS for more conclusive identification and/or for purity assessment.

Statement of Contribution II

The following chapter is written in the format of a manuscript.

I, Benazir K. Marquez, performed the analyses presented using data generated by the Collaborative Oat Research Enterprise (CORE). This data has been previously published in:

Tinker NA, Kilian A, Wight CP, *et al.* (2009) New DArT markers for oat provide enhanced map coverage and global germplasm characterization. *BMC Genomics* 10:39.

Oliver RE, Tinker NA, Lazo GR, *et al.* (2013). SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species. *PLoS One*, *8*(3), e58068.

Huang Y, Poland JA, Wight CP, Jackson EW, Tinker NA (2013). Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *Manuscript submitted to PLoS One for publication.*

I, Benazir K. Marquez, wrote the original draft of this paper and all revised versions thereafter.

Benazir Katarina Marquez

Chapter 3
Breeding applications of graphical genotyping

3.1 Abstract

Plant breeders attempt to recombine favourable alleles from different parental sources. To maintain genetic diversity and increase the likelihood of novel variation, they may select parents that are adapted but genetically divergent. A strategy such as this relies on knowledge of genetic diversity, which is traditionally inferred from pedigree analysis. However, such strategies are adversely affected by errors or missing information in pedigrees. Furthermore, progress in molecular marker technology and quantitative trait loci (QTL) discovery have provided large amounts of data that can be a challenge to interpret. Graphical representations of molecular marker data can be employed to evaluate plant material in a simple and visual way. Here we present two cases where a combination of pedigree analysis and graphical genotyping can be applied to common challenges faced by oat breeders.

KEYWORDS

graphical genotyping; haplotype; pedigree; quantitative trait loci; single nucleotide polymorphism

3.2 Introduction

Plant breeding's central goal is to improve upon current germplasm. This is accomplished through an iterative process of crossing and selection (Tinker, 2008). Breeders make crosses

to accumulate favourable genes and traits, and select progeny in between crosses. Pedigree records facilitate crossing by providing estimates of genetic similarity. Knowledge of genetic relationships among individual plants and populations are important for breeders, since larger genetic variance is expected from crossing genetically divergent parents (Messmer *et al.,* 1993). Unfortunately, pedigrees are subject to human error and unnoticed mistakes can be propagated for years. The use of questionable or incorrect pedigrees can reduce the effectiveness of strategic decisions such as parent selection. Using molecular markers, it has been possible to identify and correct mistakes in known plant pedigrees (Bautista *et al.,* 2008; Riaz *et al.,* 2008; Evans *et al.,* 2011).

Monitoring genetic similarity is also a concern for plant breeders, as modern breeding practices can reduce genetic diversity by homogenizing chromosomal segments containing breeding targets. The possible consequences that could result from such reductions could affect crop adaptation to disease or environmental changes, leaving many cultivars vulnerable. SSR (Fu *et al.,* 2003a) and AFLP (Fu *et al.,* 2004) analysis have shown that the Canadian oat (*Avena sativa* L.) gene pool is not as diverse as previously thought, revealing the negative impacts of breeding during specific time periods. Average diversity in cultivars released after 1950 has decreased. The authors suspect this was due to the increased breeding efforts to introduce crown and rust resistance genes using a limited number of resistant parents. Fu *et al.* (2004) have suggested that to avoid reaching a “genetic ceiling” in oat improvement, new sources of favourable alleles must be explored.

Pedigree and molecular marker data eventually accumulate to sizeable amounts that can be difficult to manage and interpret. Furthermore, progress in quantitative trait loci (QTL)

discovery has provided equal amounts of data that can be useful but remains a challenge to use with other molecular marker data. However, the concept of a graphical genotype (GG) can meld all of this data into a useful and graphical form for easy interpretation. A GG is a visual representation of a chromosome and/or genome based on molecular marker data. Ideally, GGs portray parental origin and allelic composition throughout the genome (Young and Tanksley, 1989), so chromosomal recombination and assortment can be demonstrated. When GGs are made for the progeny of a known cross, where parents have been genotyped, it is easy to incorporate parental origin. However, in cases where parents are not available, pedigree records can be used to find the closest ancestor for comparison. When QTL data is available, GGs can be applied to the study of polygenic traits in plant breeding, since they allow for "whole genome selection" (Young and Tanksley, 1989).

The following analyses strive to illustrate how graphical genotyping can be applied to take advantage of all sources of data provided to breeders. The objectives of this study were to determine if a combination of marker and pedigree analysis could identify (1) possible errors in oat diversity panels, and (2) putative carriers of favourable QTL alleles.

3.2.1 Challenge 1: Validating pedigrees

Pedigrees may contain mistakes that can be propagated for years if not immediately noticed. These errors can have consequences for breeding programs and associated research, especially in terms of cultivar release and crop improvement. However, mistakes in grape (Bautista *et al.,* 2008; Riaz *et al.,* 2008) and apple (Evans *et al.,* 2011) pedigrees have been detected using molecular markers. These mistakes were mainly identified by comparing allele frequencies between progeny and their purported parents. An alternative

approach would involve comparing kinship coefficients (r_k) to genetic distance (*d*). Kinship coefficients, otherwise known as the coefficient of co-ancestry (Malécot, 1948), indirectly estimate genetic similarity based on pedigree information. Conversely, genetic distance is a direct estimate obtained from molecular markers. Under ideal conditions, a comparison between these two measures should yield a negative straight-line correlation since r_k and *d* are inversely proportional. However, since selection only affects genetic distance, strong deviations may be expected. Deviations, such as those observed in barley (Tinker *et al.,* 1993) and oat (Paczos-Grzeda, 2004), may be a result of selection bias as well as from recording errors and incomplete pedigree data.

We hypothesized that we can sometimes identify causes of discrepancy between r_k and *d* (i.e. outliers in the relationship) by investigating graphical genotypes of progeny and putative ancestors. When graphically represented, a cultivar should contain ancestral linkage blocks (i.e. rows of markers) where alleles from a recent common ancestor are preserved. If a significant number of linkage blocks are lacking between an outlier and its proposed ancestors, it would suggest the outlier represents an error in recorded pedigrees. Conversely, common linkage blocks may also provide new information to complete unrecorded or incorrect pedigrees.

3.2.2 Challenge 2: Finding diverse sources of favourable alleles

Maintaining a high level of genetic diversity is a major concern for plant breeding. Low genetic diversity among breeding material is undesirable, as it limits the amount of different and favourable alleles that can be introduced into future cultivars. Ultimately, this would limit the extent of trait improvement.

QTLs are areas of a chromosome where certain allele variants are associated with a change in magnitude of a trait. QTLs are discovered by associating phenotypes with presence or absence of a molecular marker using statistical analyses. QTLs can be mapped using a variety of strategies including the use of recombinant inbred lines (RILs) from a bi-parental mapping population, as well as through association mapping in a diverse population. Parents of a bi-parental population are selected to have contrasting phenotypes for a specific trait. This type of population produces an estimate of genetic linkage that is almost linearly related to physical marker distance, since they result from the recombination of only two parents. Mapping populations also represent a limited source of alleles (at most, two alleles per locus, one from each homozygous parent). Conversely, a diversity population consists of cultivars and breeding lines that possess different pedigrees and are acquired from different geographical locations. This type of population represents multiple sources of diverse alleles; however, the associations gathered in this method of QTL mapping are only reliable for very close loci. Ideally, data from both types of populations should be merged in order to benefit from their respective advantages. We have developed methodology that uses such a merged dataset to create haplotypes anchored to a previously published QTL. Over 700 QTLs have been discovered in oat, and a large portion of these have been recorded in online databases. We hypothesized that cultivars and breeding lines that have identical localized haplotypes in the region of a QTL are likely to contain the same QTL. This is, in fact, the same principle that is used to discover QTL through association mapping, but extended to allow discrete identification and selection of specific pre-validated QTLs.

3.3 Materials and Methods

3.3.1 SNP genotype data

Three sets of genotype data were used in this study. A set of 95 oat cultivars (Table A.5) was selected from previous experiments to represent a diverse global sample of cultivated oat. Genetic characterization of these cultivars was based on 1872 SNP markers from a new Illumina-based assay and a new oat consensus map (Oliver *et al.,* 2013). This dataset (D.1) was used to determine if possible errors could be detected in a diversity panel.

A merged dataset (D.2) consisting of 1928 SNPs for 'Dal', 'Exeter', and the 145 Dal x Exeter (DxE) progeny (Hizbai *et al.,* 2013) were specifically used to validate the methodology used to create QTL-anchored haplotypes. Markers with identical genotypic data were combined together to remove redundancy, resulting in a final dataset of 1,653 binned markers.

Multiple sets of mapping and diversity data were merged using a custom program "Haploscope" (Tinker, *unpublished*). The merged dataset consisted of 84,304 SNPs for 681 cultivars and breeding lines, and 515 mapping progeny from seven mapping populations. SNPs were acquired from previous experiments described in Oliver *et al.* (2013) and Huang *et al.* (*submitted*). Redundancy was removed as above, resulting in a final set of 75,710 binned markers. Only cultivars and breeding lines with at least 90% complete genotype data were retained. This dataset (D.3, Table A.6) was used to identify putative carriers of favourable QTL alleles.

In all datasets, SNPs showing no polymorphism within a set of cultivars and/or breeding lines were deemed uninformative and discarded from analysis.

3.3.2 Data analysis

Kinship coefficients (r_k, Equation 1.3) between taxa pairs were computed using KIN software (Tinker and Mather, 1993) and pedigree information stored in the online database: Pedigree of Oat Lines (http://avena.agr.gc.ca/OGIS/, POOL). Genetic distance (*d*) was estimated between taxa pairs from SNP information using the simple matching coefficient:

$$d = 1 - \frac{x}{y} \quad [3.1]$$

where *x* is the number of shared alleles and *y* is the total number of alleles observed. The Mantel test statistic (r_M; Mantel, 1967) was performed using the *vegan* package in R (Oksanen, 2013) with 999 permutations to assess the correlation between matrices of r_k and *d*. Linear regression analysis and unpaired student's *t* test were performed using Microsoft Excel. Genetic linkage was estimated from mapping populations as recombination frequency (*rf*):

$$rf = \frac{number\ of\ recombinant\ progeny}{total\ number\ of\ progeny} \cdot 100\% \quad [3.2]$$

All graphical genotypes (GGs) were generated using GGT 2.0 (van Berloo, 2008) using non-ambiguous SNPs with consensus map locations. Genotypes were recoded to match parental information when applicable (e.g. all alleles matching parent A were recoded as A). A crossover event was graphically represented in the most likely position: the exact middle of two opposite genotypes.

Haplotype match (*hm*) was quantified as:

$$hm = \frac{number\ of\ matching\ alleles}{total\ number\ of\ alleles} \cdot 100\% \qquad [3.3]$$

De novo linkage mapping was performed using JoinMap® (van Ooijen, 2006). Markers were assigned to linkage groups at LOD 7.

3.3.3 Historical quantitative trait loci and anchoring markers

QTL information was acquired from the online database *oatgenes* (http://avena.agr.gc.ca/oatgenes/, accessed: August 2013). *Oatgenes* currently holds information for 710 QTL predictions from the literature, including their location on the 'Kanota' x 'Ogle' (KxO) map. Each record includes a confidence level (CL, 1-3) based on the experiment-wide or comparison-wide error included in the predicting study. A QTL with the greatest confidence was given a CL of 1 when the study predicting the QTL had an experiment-wide error of < 0.05 or comparison-wide error of < 0.00005. QTLs in studies with a comparison-wide error of < 0.001 or < 0.05 were given CLs of 2 or 3, respectively. A CL could be gained if the QTL was present in more than half of the tested environments in the study. Conversely, a CL was dropped if the QTL was present in less than half of the environments and was not significant in the average over all environments. Only QTLs with a CL of 1 were considered. QTLs of interest were located on the oat consensus map using its homologous KxO location (Oliver *et al.,* 2013) and the closest marker was selected as a QTL-anchored marker.

3.3.4 QTL-anchored haplotype construction using Haploscope

To generate short segments of linked markers, markers within 5% *rf* of a QTL-anchored marker were found in the merged datasets from mapping populations (D.2 and D.3). The seven mapping populations in D.3 were then omitted from further haplotype analysis. Segmental haplotypes from a source cultivar, known to contain a target QTL allele, was then used as a seed to find other cultivars with matching haplotypes based on the following process. Haploscope extracted marker data from all diversity lines in the merged dataset and coded alleles in phase with the selected cultivar. Alleles matching the cultivar were coded as 1 and those not matching were coded as 3, while heterozygotes were coded as 2. Haplotypes were filtered to discard those with < 89% complete data. Markers in the resulting haplotypes were ordered using the consensus map (Oliver *et al.,* 2013) or a *de novo* linkage map for input into GGT 2.0. Cultivars and breeding lines were predicted to possess a QTL by comparing GGs to that of a cultivar known to have the favourable trait. This methodology is summarized in Figure 3.1.

3.4 Results and Discussion

3.4.1 Pedigree validation and exploration

Matrices of genetic distance (*d*) and kinship coefficients (r_k) were calculated among 95 cultivars (D.1). A significant correlation between the matrices was established ($r_M = 0.186$, $p<0.01$), which is consistent with reports in barley (Tinker *et al.,* 1993) and maize (Melchinger *et al.,* 1991)**.** The expected negative correlation between *d* and r_k was also observed ($R^2=0.51$) when both metrics were plotted as a linear regression (Figure 3.2). The correlations observed indicate that *d* and r_k are partially predictive of each other. Therefore

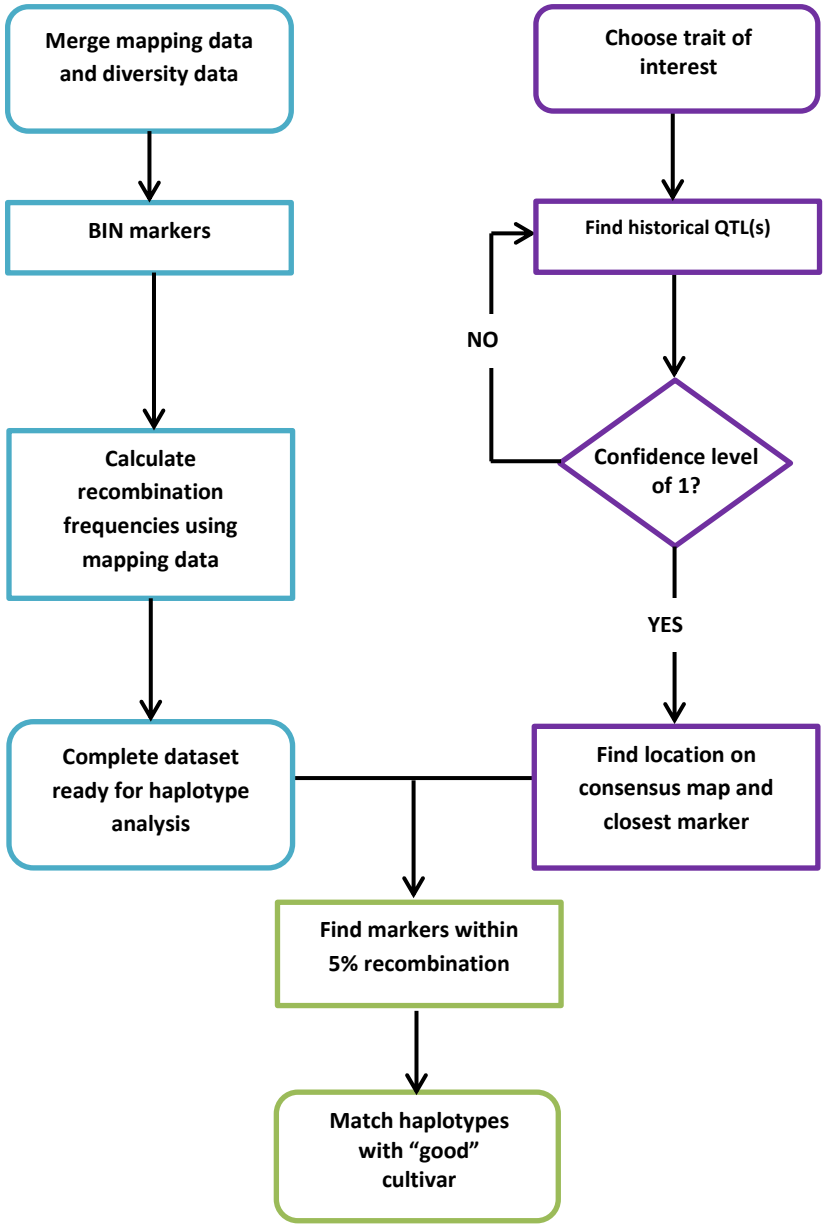

Figure 3.1 Methodology used to generate QTL-anchored haplotypes.

any substantially outlying data point could represent a potential pedigree error, lack of pedigree data, or biased artificial selection toward one parent. Most data points were plotted toward the lower end of the x-axis because pedigree data was not complete for all the cultivars included in D.1.

Pairwise comparisons of cultivars with kinship equalling 0 were filtered out of D.1 to remove some of this bias, since most were due to the complete absence of pedigree information. Nevertheless, some graphical outliers were observed. Two oat families are highlighted as examples in Figure 3.2.

Family 1 (red) contains 'CDC Pro-Fi' and its known parents 'CDC Boyer' and 'Gem'. Pairwise comparisons within Family 1 were found to be very close to the predicted relationship between *d* and r_k. Cultivars chosen as parents are often divergent from each other to improve genetic variance in progeny, and this is shown in Figure 3.2 where CDC Boyer and Gem show a small r_k and greater *d*. A cultivar derived from a single cross is expected to inherit half of its alleles from its female parent, and the other half from its male parent according to Mendelian genetics. This expectation should result in an r_k and *d* of 50%. CDC Pro-Fi has an r_k of 50.8% with each of its parents, which is consistent with the expected value. However, CDC Pro-Fi deviates from the expected *d* with its parents. CDC Pro-Fi shares a genetic distance of 0.32 and 0.19, with CDC Boyer and Gem, respectively. This deviation shows the difference between r_k and *d*, and highlights the limitations of r_k. Kinship is an indirect measure based on probability and Mendelian inheritance and is based on three assumptions: 1) pedigree data is accurate; 2) lines with no known common parentage have

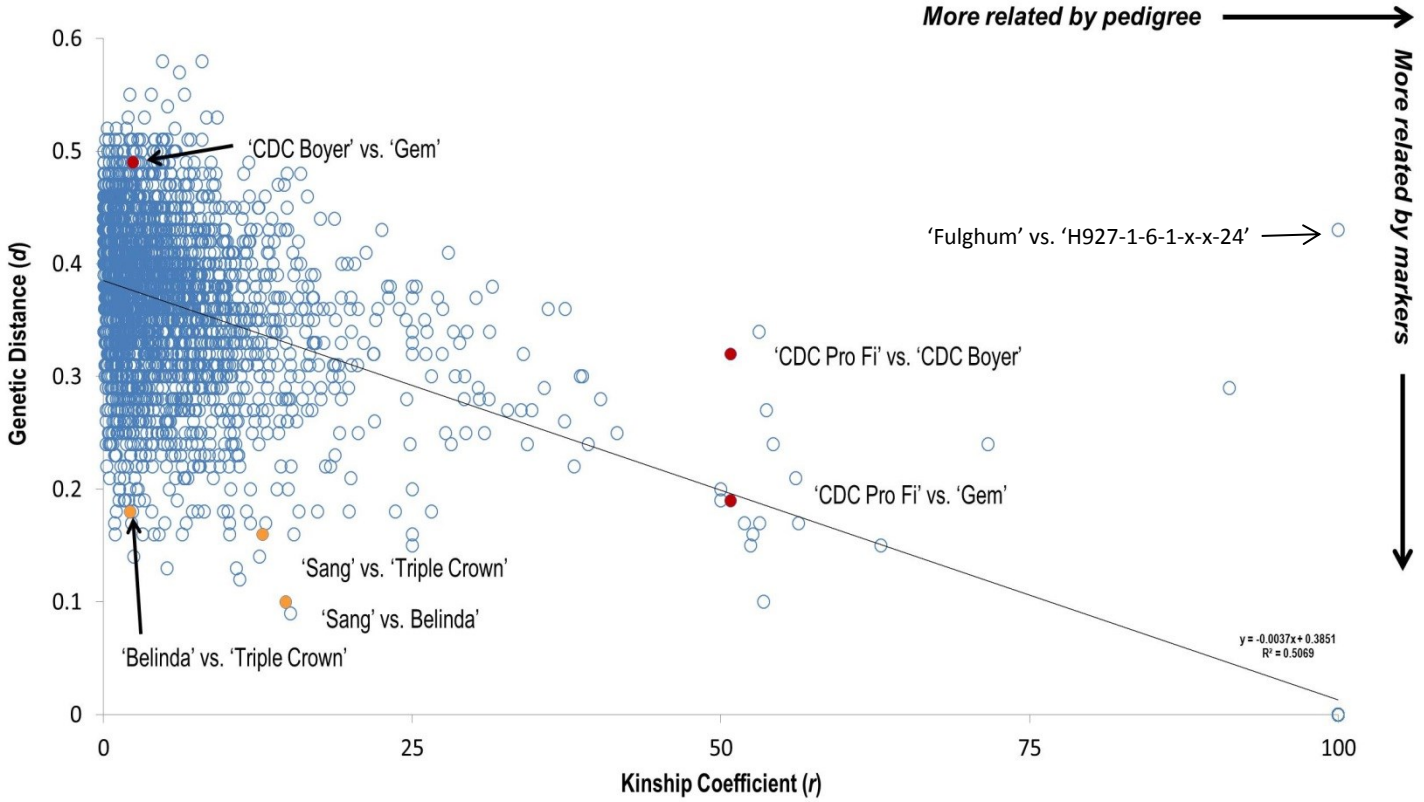

Figure 3.2 Correlation between kinship coefficient (r_k) and genetic distance (d) of pairwise comparisons of 95 oat cultivars. Red-filled circles represent Family 1: 'CDC Pro-Fi', 'CDC Boyer', and 'Gem'. Yellow-filled circles represent Family 2: 'Sang', 'Triple Crown', 'Belinda'. Linear regression was fitted using the Trendline function in Excel, y=-0.0037x + 0.3852, R^2=0.5069.

no shared ancestry; and 3) lines derived from a cross obtained half of their genome from each parent. These assumptions are more than likely to be violated, especially the third, since deviations are expected due to selection, linkage, and random genetic drift. Furthermore, the intensive selection pressure applied during cultivar development may shift genomic distribution toward the parent with a superior or favourable trait. On the other hand, *d* provides a direct measure that makes no assumptions about heritage.

The difference between these two metrics is also exemplified between 'Fulghum' and 'H927-1-6-1-x-x-24' (Figure 3.2). These two cultivars are indicated to be identical in pedigree (r_k=100), both descending completely from the landrace 'Red Rustproof'. Fulghum was originally selected from Red Rustproof because of its early maturity and increased height, while H927-1-6-1-x-x-24 came from a collection of crosses between other selections from Red Rustproof. However, based on SNPs these lines are highly divergent from each other (*d*=0.42). This, in fact, suggests that Red Rustproof should not have been considered as a homozygous cultivar rather as a highly heterogeneous landrace.

Red Rustproof is believed to have originated from Mexico and was introduced into the Southern United States in 1848-49 (Coffman, 1977). Many other cultivars can be traced by selection or reselection back to Red Rustproof, including: 'Appler', 'Aurora', 'Brunker', 'Burt', 'California Red', 'Coast Black', 'Colburt', 'Columbia', 'Culberson', 'Ferguson 560', 'Forkedeer', 'Frazier', Fulghum, 'Fulwin', 'Hairy Culberson', 'Hay', Kanota, 'Navarro', 'New Nortex', 'Nortex', 'Otoe', 'Pentagon', 'Suwannee', 'Tech', 'Tennex', 'Trojan' (Coffman, 1977). This is not an extensive list, as many of these cultivars have one or more synonyms (POOL, accessed: December 2013). Moreover, several of these have become important progenitor

cultivars therefore they can be found in many oat pedigrees. Red Rustproof was originally included in D.1, but was excluded due to missing data. When included in the above analysis, it produced multiple data points close to the one plotted for Fulghum and H927-1-6-1-x-x-24 (data not shown). Those points corresponded to pairwise combinations of Red Rustproof with its selections or derived cultivars. This demonstrates how some outliers from the expected r_k and *d* correlation can be heavily influenced by a widely used ancestor in its breeding history.

Family 2 (orange) contains Swedish cultivars: 'Belinda', 'Triple Crown', and their common ancestor 'Sang'. Contrary to Family 1, members of Family 2 appear to be highly interrelated based on SNPs (*d*=0.1-0.18), but relatively unrelated based on pedigree information (r_k=2.14-14.77). He and Bjørnstad (2012) reported an extensive number of shared loci when they studied Nordic (Norwegian, Swedish, and Finnish) accessions using SSRs, AFLPs, and DArTs. They inferred this to be the product of frequent germplasm exchange and interbreeding among these countries' breeding programs. High genetic diversity was observed in the Swedish cultivars studied, which is in contrast to the results reported in Figure 3.2. However, among the Swedish accessions they studied, they observed frequent use of a number of parental lines. As discussed above, a recurrent parent can influence r_k. In this case, pedigree information is not as complete for this family as the other examples therefore r_k may be underestimated.

To further investigate the relationship between the cultivars represented by these outliers, GGs were inferred using mapped SNPs with non-ambiguous data. As expected, both families share common markers on putative ancestral linkage blocks (Figures 3.3 and 3.4).

The GG for CDC Pro-Fi (Figure 3.3) is composed of 350 SNPs, of which 334 can be traced to either CDC Boyer or Gem. The remaining 16 SNPs could not be traced to either parent due to missing parental genotypes. The number of alleles contributed from parent 1 and parent 2 to its progeny is rarely expected to exceed a ratio of 59:41 based on Chi-square (α=0.05) analysis (Lorenzen *et al.,* 1995). CDC Pro-Fi's GG validates its recorded pedigree with a slightly greater inheritance of linkage blocks from Gem (59%) than from CDC Boyer (41%). The greater inheritance of linkage blocks from Gem may be an indication of selection toward a Gem trait.

CDC Pro-Fi is branded as a high β-glucan cultivar (Izydorczyk *et al.,* 2014), therefore it is reasonable to suspect that Gem also contains high levels of β-glucan. Trait scores measured from nine different environments during a 1998 QUON study[3] shows no significant difference between β-glucan in CDC Boyer and Gem. This suggests that any possible selection toward a favourable Gem trait may not have been linked to β-glucan. However, Gem is resistant to prevalent races of crown rust in addition to having good test weight, groat percentage, groat protein percentage, and high grain yield (Duerst *et al.,* 1999). Selection for any of these traits could be the cause of CDC Pro-Fi inheriting extra Gem linkage blocks. Superimposition of known QTLs onto the CDC Pro-Fi GG could confirm these speculations in the future.

[3] Data available at: http://wheat.pw.usda.gov/cgi-bin/graingenes/report.cgi?class=traitstudy;name=Beta%20glucan%2C%201998%20QUON;show=traitscore

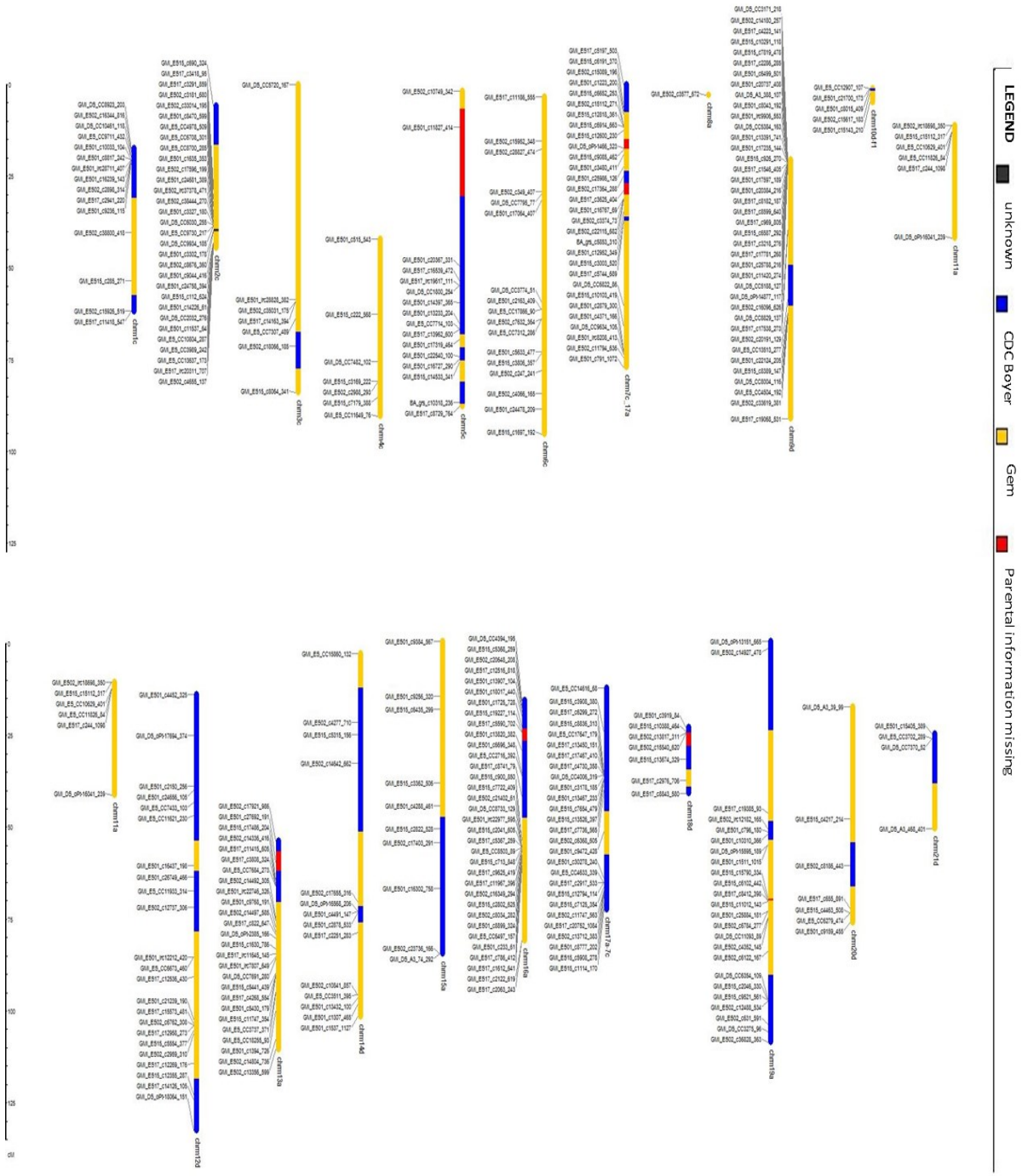

Figure 3.3 Graphical genotype of 'CDC Pro-Fi'. 350 informative markers are shown. Blue regions are inherited from 'CDC Boyer' (39.2%), those inherited from 'Gem' are yellow (56.1%), and red regions represent areas where parental information is missing (4.6%).

The GGs of Belinda and Triple Crown show unexpectedly large portions of chromosomes that could have been co-inherited from an ancestor similar to Sang (Figure 3.4). Sang is presented because it is the only known common ancestor from an earlier generation within the set studied. This strong similarity suggests that there may be additional common ancestors in this family that have not been recorded. He and Bjørnstad (2012) grouped the Nordic oat accessions they studied into three age groups representing separate breeding periods: old (landraces and released before 1950), medium (released between 1960 and 1990), and new (released after 1991). Belinda and Sang were classified as new and medium, respectively. While Triple Crown was not included, it can be assumed that it would have been classified as new based on its release date. Genetic diversity appeared to decrease strongly between the old and medium groups. Therefore, it is plausible that Sang has retained linkage blocks from older cultivars that may be missing in the pedigrees of Belinda and Triple Crown.

None of the outliers found in the linear regression were identified to be errors, however these analyses have further demonstrated the utility of graphical genotyping for pedigree validation and exploration. The GGs presented allow for speculation about selection occurring during each cultivar's breeding history, but none can be validated without any additional trait information. Superimposition of QTL onto GGs is explored in the next section as a method to find putative carriers of favourable alleles.

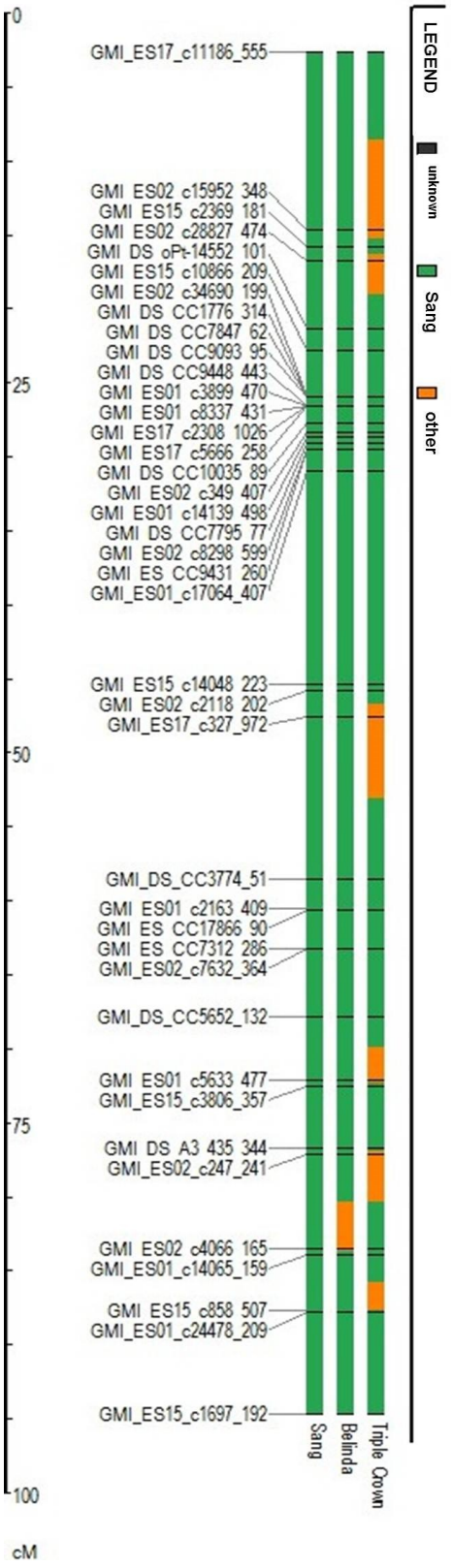

Figure 3.4 Graphical genotype of chromosome 6C in 'Sang', 'Belinda' and 'Triple Crown'. Green regions represent similarities to 'Sang.' Orange regions show locations where 'Belinda' (0.4%, 13% overall) and 'Triple Crown' (25.5%, 21.4% overall) are different from 'Sang'. Orange regions do not necessarily originate from a common ancestor.

3.4.2 Discovery of putative QTL carriers

3.4.2.1 Validation of QTL-anchored haplotypes

To validate the methodology shown in Figure 3.1, GGs of the DxE population were anchored to an oil QTL that was discovered previously in KxO: *cdo665b* (Kianian *et al.,* 1999). GMI_ES_c12621_204 was chosen as the anchoring marker as it maps closest to *cdo665b* on linkage group KO_11_41_20_45. *Cdo665b* maps to position 0, while GMI_ES_c12621_204 maps to position 6.03. Two QTLs (oPt-6135, oPt-17489) that were found to contribute the largest effect on oil in DxE co-locate with *cdo665b* (Hizbai *et al.,* 2012), therefore we hypothesized that a marker in *cdo665b's* proximity would also represent a QTL region in DxE. Based on this hypothesis, we predicted that this population would be separable into two haplotype groups: one group containing alleles for high oil, and the other containing alleles for low oil. Using a merged dataset comprised of DxE mapping data and diversity data from parents Dal and Exeter, Haploscope found 24 markers with ≤ 5% *rf* with GMI_ES_c12621_204. These markers were never mapped in the DxE population, so they were mapped *de novo* to allow markers to be placed in order. Four markers mapped with GMI_ES_c12621_204 (position 1.8) at LOD 7: GMI_ES18_c4764_27 (0), GBS_85336 (2.6), GBS_605858 (12.2), and GMI_ES17_c2122_619 (12.9).

These markers were expressed as GGs for inspection (Figure 3.5). The first three markers appear to be closely linked as no recombination within this region is visible. The 145 DxE lines were separated into two haplotype groups according to their similarity to the "high oil" parent Dal (Figure 3.5). Group 1 consists of progeny that carry only or predominantly Dal alleles at the first three loci. Group 2 consists of progeny that carry only Exeter alleles

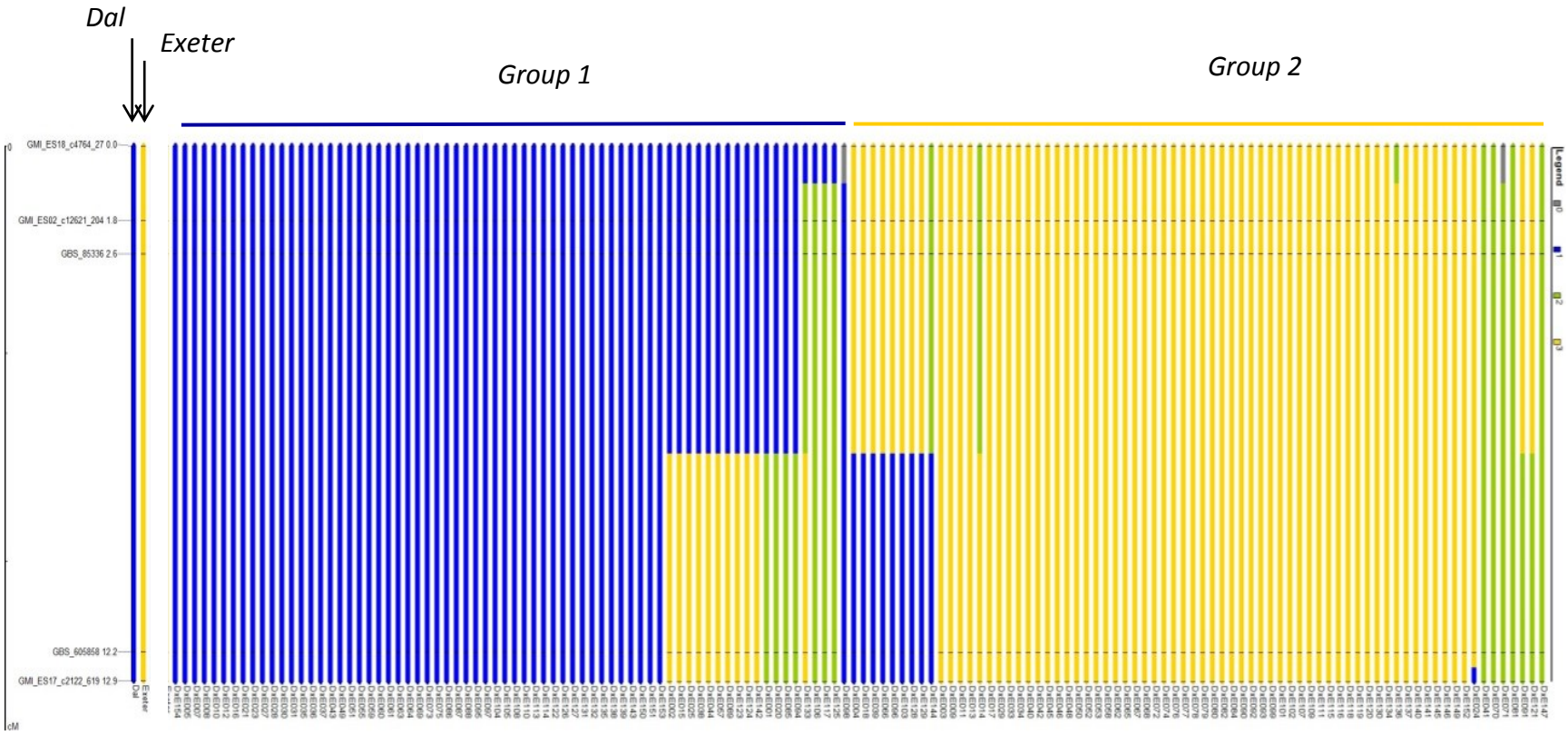

Figure 3.5 Graphical genotypes of the 'Dal' x 'Exeter' population ordered by likeness to 'Dal'. Dal alleles are represented by blue, Exeter alleles by yellow. Heterozygotes are represented by green. Missing data are coloured grey. Graphical genotypes were generated by GGT 2.0 (van Berloo, 2008).

or are heterozygous at these loci. Group 1 and Group 2 can also be divided solely on the presence of the anchoring marker. The average oil content of each haplotype group was calculated using the available phenotype data from the population (Hizbai *et al.,* 2012). Group 1 has statistically significant ($p<0.0001$) higher average oil content (8.38%) than group 2 (7.00%). These observations demonstrate that the developed methodology can produce anchored haplotypes capable of representing historical QTLs.

3.4.2.2 Discovery of putative cdo665b carriers

Haplotypes anchored to *cdo665b* were further pursued with the cultivars and breeding lines contained in D.3. Haplotypes of 114 cultivars representing worldwide diversity (Figure 3.6) and 253 North American breeding lines (Figure 3.7) were studied to find possible carriers. Unlike the validation example, these haplotypes were compared to the "high oil" parent in the original population where *cdo665b* was first mapped: Kanota. GMI_ES_c12621_204 was used again as the anchoring marker and eight additional markers were located within 5% *rf.* These markers were ordered according to the oat consensus map (Oliver *et al.,* 2013) before GGs were generated. Ten cultivars and 15 breeding lines were observed to have identical haplotypes to Kanota and consequently were identified as putative carriers of *cdo665b* (Table 3.1).

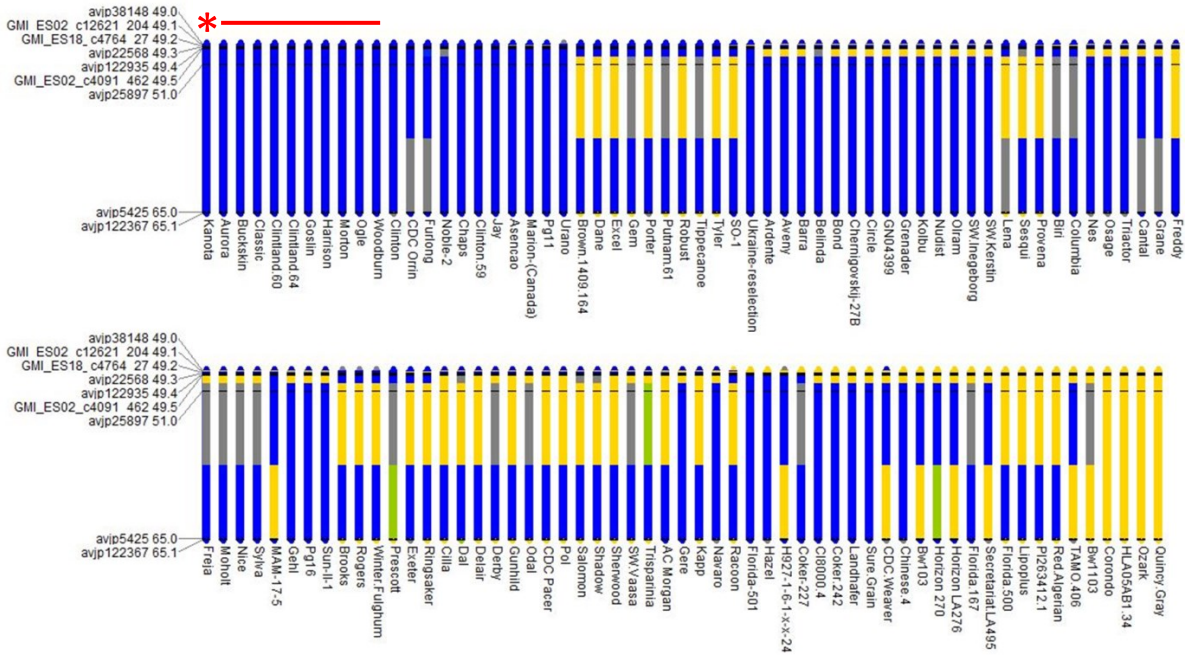

Figure 3.6 Graphical genotypes of diverse cultivars and accessions anchored to *cdo665b* by marker GMI_ES_c12621_204. Alleles matching those in Kanota (*) are represented in blue, non-matching alleles in yellow, and heterozygotes in green. Missing data is represented by grey. Red lines indicate cultivars with identical haplotypes to Kanota. Cultivars and accessions are presented in order by number of alleles identical to Kanota. Graphical genotypes were generated by GGT 2.0 (van Berloo, 2008).

Figure 3.7 Graphical genotypes of North American breeding lines anchored to *cdo665b* by marker GMI_ES_c12621_204. Alleles matching those in Kanota (*) are represented in blue, non-matching alleles in yellow, and heterozygotes in green. Missing data is represented by grey. Red lines indicate breeding lines with identical haplotypes to Kanota. Graphical genotypes were generated by GGT 2.0 (van Berloo, 2008).

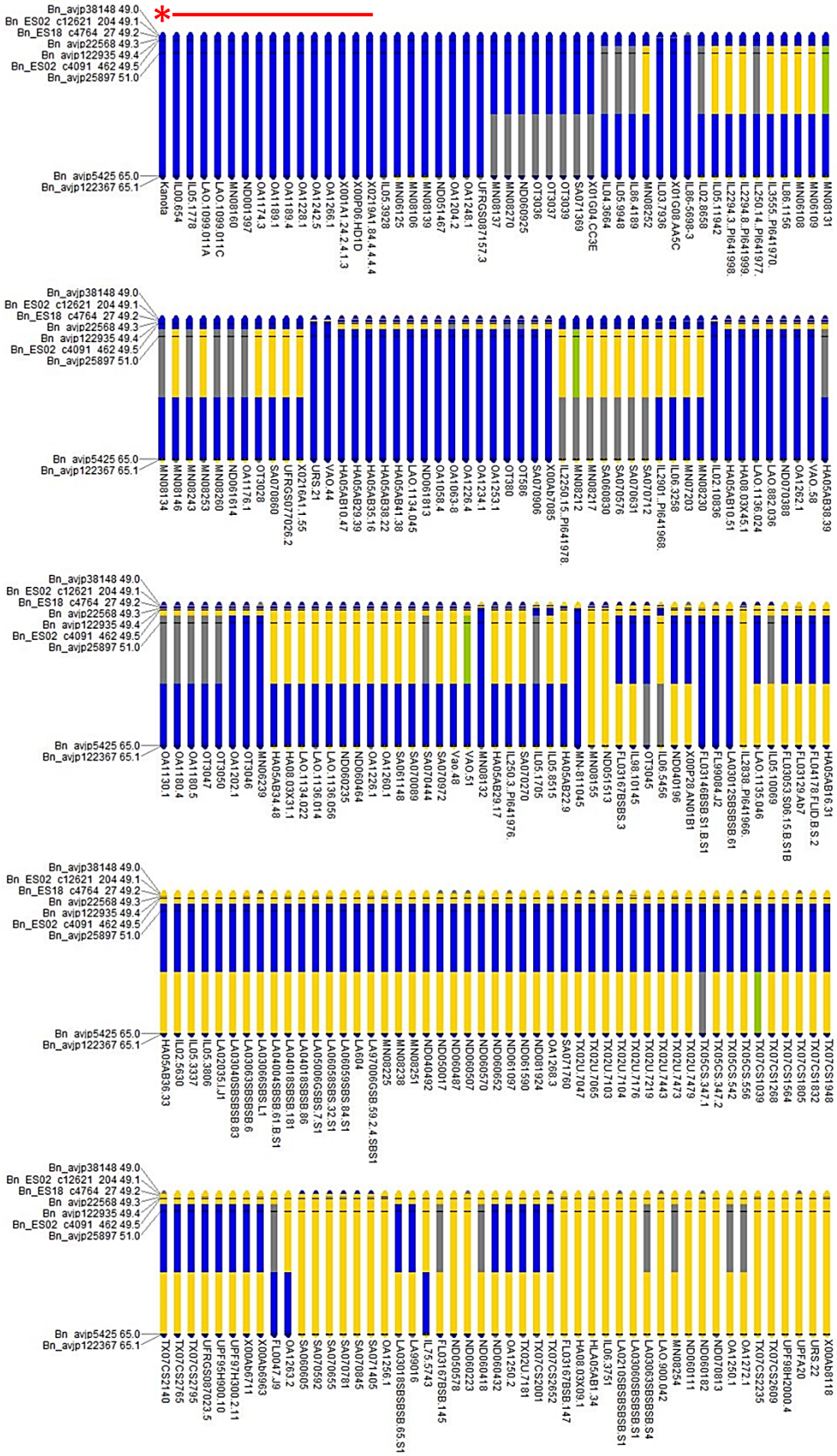
Bn_avjp38148 49.0
Bn_ES02_c12621_204 49.1
Bn_ES18_c4764_27 49.2
Bn_avjp22568 49.3
Bn_avjp122935 49.4
Bn_ES02_c4091_462 49.5
Bn_avjp25897 51.0
Bn_avjp5425 65.0
Bn_avjp122367 65.1
Kanota
IL00.654
IL05.1778
LA0.1099.011A
LA0.1099.011C
MN08160
ND001397
OA1174.3
OA1189.1
OA1189.4
OA1226.1
OA1242.5
OA1266.1
X001A1.24.2.4.1.3
X00P06.HD1D
X02I9A1.84.4.4.4.4
IL05.3928
MN06125
MN08106
MN08139
ND051467
OA1204.2
OA1248.1
UFRGS087157.3
MN08137
MN08270
ND060925
OT3036
OT3037
OT3039
SA071369
X01G04.CC3E
IL04.3664
IL05.9948
IL86.4189
MN08252
IL03.7936
X01G08.AA5C
IL86-5698-3
IL02.8658
IL05.11942
IL2294.3.Pi641998.
IL2294.8.Pi641999.
IL2294.9.Pi641999.
IL2504.Pi641977.
IL3555.Pi641970.
IL86.1156
MN06108
MN06109
MN08131
HA05AB38.39
VAO.58
OA1262.1
ND070388
LA0.892.036
LA0.1136.024
IL05.10069
LA08.03X45.1
HA05AB10.51
IL02.10836
LA0301258SBSBSB.61
MN08230
MN07203
IL06.3258
IL2901.Pi641968.
SA070712
SA070631
SA070576
SA060830
MN08217
MN08212
IL2250.15.Pi641978.
X00Ab7085
SA070906
OT586
OT380
OA1253.1
OA1234.1
OA1226.4
OA1063.8
OA1058.4
ND061813
LA0.1134.045
HA05AB41.38
HA05AB38.22
HA05AB29.39
HA05AB10.47
VAO.44
URS.21
X0216A1.1.55
UFRGS077026.2
SA070860
OT3028
OA1176.1
ND061614
MN08260
MN08253
MN08243
MN08146
MN08134
HA05AB16.31
FL04178.FLIDB.S.2
FL03129.Ab7
FL03053.S06.15.B.S1B
IL05.10069
LA0.1135.046
IL2838.Pi641966.
LA0301258SBSBSB.61
FL99084.J2
FL03146BSB.S1.B.S1
X00P28.AN01B1
ND040196
IL06.5456
OT3045
IL98.10145
FL03167BSBS.3
ND051513
MN08155
MN-811045
HA05AB22.9
IL05.8515
IL05.1705
SA070270
IL250.3.Pi641976.
HA05AB29.17
MN08132
VAO.51
Yao.48
SA070972
SA070444
SA070089
SA06148
OA1260.1
OA1226.1
ND060464
ND060235
LA0.1136.056
LA0.1134.014
LA0.1134.022
HA08.03X31.1
HA05AB34.48
MN06239
OT3046
OA1202.1
OA1180.4
OA1180.5
OT3050
OT3047
OA1180.5
OA1130.1
TX07CS1948
TX07CS1832
TX07CS1805
TX07CS1564
TX07CS1268
TX07CS1039
TX05CS.556
TX05CS.642
TX05CS.347.2
TX05CS.347.1
TX02U7479
TX02U7473
TX02U7443
TX02U7219
TX02U7176
TX02U7104
TX02U7103
TX02U7065
TX02U7047
SA071760
OA1268.3
ND081924
ND061590
ND061097
ND060652
ND060570
ND060507
ND060487
ND050017
ND040492
MN08251
MN08238
MN08225
LA9700.6GSB.59.2.4.SBS1
LA604
LA060585BS.32.S1
LA060585BS.84.S1
LA05006GSBS.7.S1
LA040185BSB.86
LA040185BSB.181
LA040045BSB.61.B.S1
LA030665BSBS.L1
LA030635BSBSB.6
LA030405BSBSBSB.83
LA02035.L31
IL05.3806
IL05.3337
IL02.5630
HA05AB36.33
X00Ab8118
URS.22
UPFA20
UPF98H2000.4
TX07CS2609
TX07CS2235
OA1272.1
OA1250.1
ND070813
ND060182
ND060111
MN08254
LA0.900.042
LA030635BSBSBSB.S4
LA030605BSBSBSB.S1
LA02105BSBSBSBSB.S1
IL06.3751
HLA05AB1.34
HA08.03X09.1
FL03167BSB.147
TX07CS2652
TX07CS2001
TX02U.7181
OA1250.2
ND060432
ND060418
ND060223
ND050578
FL03167BSB.145
IL75.5743
LA99016
LA030185BSBSBSB.65.S1
OA1256.1
SA071405
SA070845
SA070781
SA070655
SA070592
SA060605
OA1263.2
FL0047.J9
X00Ab6963
X00Ab6711
UPF97H300.211
UPF95H900.10
UFRGS087023.5
TX07CS2795
TX07CS2765
TX07CS2140

Table 3.1 Putative carriers of *cdo665b* and their kinship coefficients (r_k) with 'Kanota'. All cultivars and breeding lines have identical localized graphical genotypes (*hm*=100%) to Kanota. Three r_k values are not applicable (n/a) due to unavailable pedigree data.

cultivar	r_k with Kanota (%)	breeding lines	r_k with Kanota (%)
Aurora	100.00	OA1266.1	10.91
Bucksin	4.99	IL00.654	7.90
Classic	4.73	OA1228.1	7.61
Clintland 60	4.71	OA1189.1	5.35
Clintland 64	3.97	OA1189.4	5.35
Goslin	3.57	OA1174.3	5.08
Harrison	2.11	OA1242.5	5.08
Morton	0.10	IL05.1778	3.96
Ogle	0.00	MN08160	3.45
Woodburn	0.00	LAO.1099.011A	2.63
		LAO.1099.011C	2.63
		ND001397	0.00
		X001A1.24.2.4.1.3	n/a
		X00P06.HD1D	n/a
		X0219A1.84.4.4.4.4	n/a

Highly similar haplotypes were expected to occur in cultivars and breeding lines with common ancestry, however comparisons of *hm* and r_k suggest that the majority of putative carriers are not closely related to Kanota (Figure 3.8). Only one of the putative carriers (Figure 3.8A) has an $r_k > 50\%$ to Kanota. 'Aurora' and Kanota have an r_k of 100% suggesting that they have identical pedigrees. According to POOL records (Tinker and Deyl, 2005), Aurora and Kanota are progeny from two different cultivars: Appler and Fulghum, respectively. While these cultivars have different names, they are both pure-line selections from Red Rustproof (Stanton, 1955) and are expected to be very similar in genomic constitution. This suggests that Aurora and Kanota may have acquired *cdo665b* from Red Rustproof or another common ancestor.

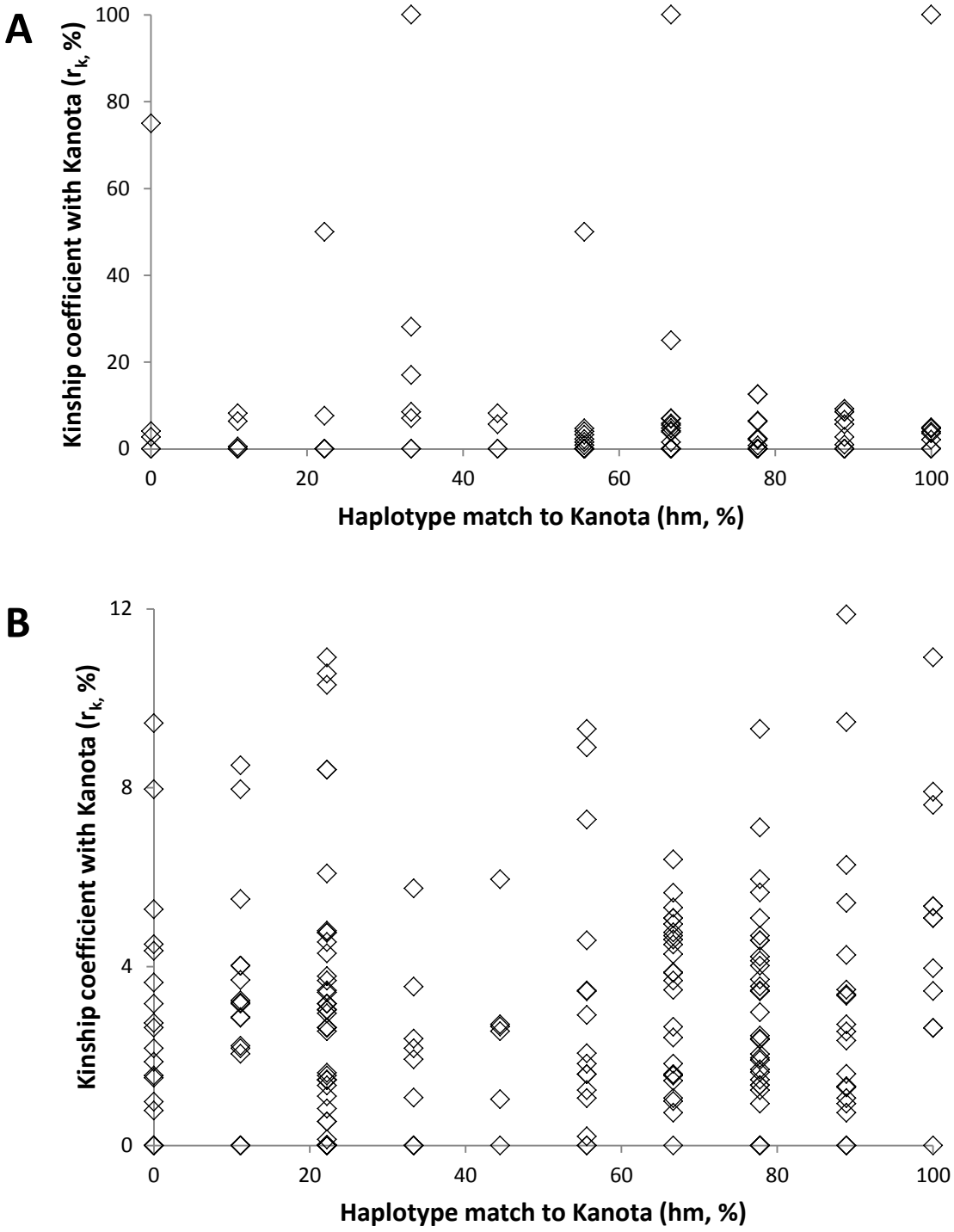

Figure 3.8 Kinship coefficient and haplotype match of (A) 114 cultivars and (B) 228 breeding lines to Kanota. Each (◇) represents one cultivar or breeding line. Comparison between Kanota and itself is omitted. Fifteen breeding lines were omitted from analysis due to unavailable pedigrees.

As seen in section 3.4.1, pure-line selections originating from Red Rustproof are expected to have an r_k of 100% since their pedigrees all trace back to the same cultivar. A caveat of r_k is that it does not take into account the effect of selection that occurs during the breeding process, and therefore does not provide an indication of recombination. This is exemplified by Columbia and H927-1-6-1-x-x-24 that have r_k = 100% with Kanota, but *hm* ≠ 100%. Columbia is an off-type selection of Fulghum (Stadler and Kirkpatrick, 1930), while H927-1-6-1-x-x-24 is the result of Hairy Culberson//Fulghum/Red Rustproof[4]. Neither Columbia (*hm*=66%) nor H927-1-6-1-x-x-24 (*hm*=33%) are known to be high oil cultivars, and this is substantiated by their GGs (Figure 3.6). Neither possesses the anchoring marker: GMI_ES_c12621_204 and consequently both GGs show yellow segments at that locus.

The remaining 24 putative carriers have r_k values < 11%. This suggests that if they do carry *cdo665b*, it is unlikely to have been acquired from Kanota. Therefore, these 24 oat accessions may be diverse sources for the QTL. Five of the 15 breeding lines that may possess the favourable *cdo665b* allele are related (r_k > 20%) to five of the cultivars also identified as putative carriers (Figure A.2). This implies that another source for the QTL may exist in oat families other than Red Rustproof.

The cultivars and breeding lines were predicted to possess *cdo665b* if they had an identical GG to that of Kanota. This criterion is straightforward and was the major reason for being applied in this analysis, however the location of the haplotype match in the "ideal" GG must also be considered. As seen in the DxE example, Group 1 (high oil) and Group 2 (low oil)

[4] This is the pedigree written in Purdy notation. The first cross occurred between Fulghum and Red Rustproof. The resulting progeny was then crossed with Hairy Culberson to produce H927-1-6-1-x-x-24. (GRIN:http://www.ars-grin.gov/cgi-bin/npgs/acc/search.pl?accid=H927-1-6-1-x-x-24&inactive=Yes)

could be divided using the entire GG or only the anchoring marker. There are 11 cultivars and 20 breeding lines that have an *hm* = 88.89% to Kanota (Figure 3.8) and all possess the anchoring marker. Similar to the majority of putative carriers, none of these accessions appear to be related to Kanota (r_k < 10%) and therefore should also be considered as possible diverse allele sources for *cdo665b*.

Interestingly, Ogle has an identical GG to Kanota. Ogle was the other parent of the mapping population where *cdo665b* was discovered, and in theory should not carry the same alleles as Kanota at these loci. The data presented here was obtained from genotyping experiments on Kanota and Ogle DNA that were not from the same plants used in the Kianian *et al.* (1999) study. The difference in plant material could be part of the explanation behind this oddity. However, it should also be noted that the nine markers found by Haploscope are concentrated into two groups on the GGs with a 14 cM gap in between. To strengthen predictions, discovery of additional high density markers that map within this gap will be needed. Misleading predictions, such as Ogle, may be removed once GGs are populated with more markers.

Unlike the DxE example described in section 3.4.2.1, phenotype data for these carriers is not readily available. As such, the cultivars and breeding lines identified to potentially carry *cdo665b* cannot be confirmed quantitatively to have high oil content. Remarks in literature, registration notices, and available breeders' notes about the putative carriers do not include any comments about oil. Therefore, further investigation is required to determine the extent to which the generated GGs predict oil content and presence of *cdo665b.*

The methodology presented here (Figure 3.1) shows promise for future breeding applications. Firstly, the identified cultivars and breeding lines could be used as allele sources for breeding high oil cultivars. This would be helpful in cases where Kanota is not advantageous or desired. Kanota may not be preferred if it has been used frequently and there is a desire to increase diversity in breeding material. Secondly, using GGs to display the anchored haplotypes provides an easy approach for making selections in a large number of plants. If further pursued, this methodology could be applied systematically to all historical QTLs recorded in *oatgenes*. Cultivars and breeding lines identified to potentially carry the favourable allele at each QTL could be collected in another database and used as a resource for oat breeders.

3.5 Conclusion

Advances in molecular marker technology have led to a wealth of data for breeders and scientists. We have presented two cases where graphical genotyping can be used in conjunction with pedigree and QTL information to overcome common challenges faced in plant breeding. GGs of oat families were shown to a) validate pedigrees and b) speculate on missing ancestors. QTL-anchored haplotypes of cultivars and breeding lines revealed putative diverse carriers of the "high oil" allele of *cdo665b.* Using GGs for pedigree validation is straight forward and is expected to be easily repeatable, however further validation of the latter GG application is recommended.

Chapter 4
Discussion

Plant breeding is increasingly powered by pedigrees and molecular markers. These tools facilitate cultivar identification and efficient trait selection. Consequently, when pedigrees contain inaccurate relationships and/or cultivar names, breeding results can vary from expectations. For example, traits expected to be incorporated into progeny may be missing or parents of crosses may not incorporate as much diversity as expected. This can hinder or limit research into crop improvement and delay the breeding process. Fingerprinting cultivars can identify pedigree inconsistencies, and advances in genotyping technology now allow these diagnostics to be done faster and more objectively. As such, this study aimed to develop molecular marker-based methodologies based on KASP™, GBS, and graphical genotyping for cultivar identification, seed purity assessment, and trait selection in oat.

RFLPs, AFLPs, RAPDs, and SSRs have been used to experimentally fingerprint oat. The latter three marker methods are PCR-based yet still rely on gel electrophoresis for scoring. This can slow down diagnostics since automation is not possible. Automated fingerprinting could be possible using a PCR-based method with fluorescence detection. This study has developed an alternative KASP™ protocol for oat, suitable for fingerprinting purposes rather than marker-assisted selection (Gnanesh *et al.,* 2013). This study has also identified a small set of KASP™ markers capable of identifying 22 current Canadian cultivars (Figure 2.2). GBS was not used explicitly for fingerprinting in this study, but the whole-genome data generated for four cultivars distinguished clearly among them, thus showing how GBS could

be used for cultivar identification. However, the sample preparation and data processing necessary for GBS would slow down any routine diagnostics. In spite of this, GBS provides a large amount of data that can be reaped of SNPs for future KASP™ marker design.

The set of diagnostic KASP™ markers detected a considerable amount of heterogeneity in a set of 160 seeds representing four seed sources of four cultivars (Figure 2.3). GBS performed on the same set of samples revealed less intra-cultivar variation and partially validated the KASP™-detected heterogeneity (Figure 2.4). The variation detected by both methods does not appear to be limited to a specific cultivar or seed source. A miscall rate was established for KASP™ in this study, and it is highly suggested that replicate assays be performed when KASP™ is used for diagnostics. However, KASP™ was only performed once on the CDOC-160 and the amount of detected heterogeneity may be over-estimated. Therefore, KASP™ cannot be disregarded as a potential method for seed purity assessments. Additional assays would have to be performed to appropriately evaluate its utility for that breeding application. Based on these results, GBS should be used for more conclusive identification and/or purity assessment.

Graphical genotyping of GBS data demonstrated the localized genomic heterogeneity expected in cultivars of all inbreeding crop species (Figure 2.5, A.3). These GGs allowed for this expected variation to be discriminated from physical contamination. The CSGA allows only five plants from different oat cultivars to be present within 10,000 plants grown from certified seed (CSGA, 2014). This maximum impurity content of 0.05% is almost certainly underestimated, because it is based on observable traits that have been deliberately selected for homogeneity. In contrast, KASP™ and GBS markers target allelic variants that

are likely to be phenotypically neutral. Therefore detecting physical contamination by molecular means could be more accurate than observation, especially in cases where a contaminant is phenotypically similar to the cultivar being inspected. In spite of this, DNA fingerprints are not required as a DUS characteristic and only serve as a complement to traditional characteristics. As such, there are no molecular guidelines concerning the acceptable amount of genomic heterogeneity in registered Canadian cultivars. The expected allele frequencies (Table A.3-.4) caused by localized heterogeneity are presented in this study and could be used as a starting point for such diagnostic guidelines.

This study also demonstrated the utility of graphical genotyping for other breeding applications. Breeders select parents that are genetically divergent to maintain genetic diversity and increase the likelihood of novel variation in progeny. At times it is also desired to use novel parents, outside of current breeding material, to introduce more genetic diversity into a program. Graphical genotyping of SNPs provided a method of visualisation for pedigree validation (Figure 3.3) and speculation about missing ancestors (Figure 3.4). Furthermore, GGs displayed the genomic effect of selection. Graphical genotyping for pedigree validation has already been shown in soybean (Lorenzen *et al.,* 1995) and maize (Smith *et al.,* 1997), therefore the repeatability and utility of this application is expected.

GGs of cultivars and breeding lines were superimposed with historical QTL to identify putative diverse carriers of a QTL for oil. This methodology (Figure 3.1) has potential for future breeding applications, especially for finding diverse parents with favourable traits. However, it is clear that this methodology needs to be further developed. *Cdo665b*, a QTL for oil discovered in KxO, was predicted to be carried by both parents. If this was true, the

QTL would not have been present in the cross where it was discovered. Even if Ogle had not been predicted to have this QTL, additional phenotyping would still be necessary to corroborate that the generated haplotypes truly reflect the QTL in question before being applied by a breeder. Additionally, the effect of environmental conditions on phenotype must be considered when determining the extent to which the generated GGs predict oil content. If further pursued, this methodology could be applied systematically to more historical oat QTLs and collected in a database as a breeding resource.

The present work can also be considered as a resource for breeders. Firstly, KASP™ and GBS have been developed for rapid cultivar identification and purity assessments in oat, respectively. These assays can be applied in the event that pedigrees are questionable or seed contamination is suspected. Secondly, GBS combined with graphical genotyping allowed for heterogeneity within cultivars to be localized to chromosomal regions, and in turn provided a way to distinguish that from physical contamination. These results could provide a baseline for guidelines prescribing the acceptable amounts of heterogeneity in registered cultivars. Lastly, we have evaluated the utility of graphical genotyping for pedigree validation and discovery of putative carriers of favourable QTL alleles. Breeders will find GGs useful for viewing their cultivars and breeding lines as a whole, especially since GGs can demonstrate the effect of selection on the genome. The latter application, which involves anchoring and superimposition of QTL onto GGs, may prove to be useful for breeding but still requires investigation. If further pursued, this methodology would provide a simple way to select allele sources for increasing diversity in a breeding program. In summary, the presented work has provided tools and insights that will assist oat breeding.

REFERENCES

Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor Appl Genet 119:507–17.

Ananiev EV, Vales MI, Phillips RL (2002) Isolation of A/D and C genome specific dispersed anad clustered repetitive DNA sequences from *Avena sativa*. Genome 45:431-441.

Asoro FG, Newell MA, Scott MP, Beavis WD, Jannink J-L (2013) Genome-wide association study for beta-glucan concentration in elite North American oat. Crop Sci 53:542-553.

Baum BR (1977) Oats: wild and cultivated. A monograph of the genus *Avena* L. (Poaceae). Monograph No. 14. . Biosystematics Research Institute (currently ECORC), Ottawa, Canada.

Bautista J, Dangl GS, Yang J, Reisch B, Stover E (2008) Use of genetic markers to assess pedigrees of grape cultivars and breeding program selections. Proceedings of the 2nd Annual National Viticulture Research Conference. Am Soc Enol Viticulture, pp 8–9.

Becher R (2007) EST-derived microsatellites as a rich source of molecular markers for oats. Plant Breeding 126:274–278.

Bennett MD, Leitch I (1976) Nuclear DNA amounts in angiosperms. Phil Trans R Soc Lond B 274: 227-274.

Van Berloo R (2008) GGT 2.0: Versatile software for visualization and analysis of genetic data. J Hered 99:232–236.

Blotta I, Prestinaci F, Mirante S, Cantafora A (2005) Quantitative assay of total dsDNA with PicoGreen reagent and real-time fluorescent detection. Ann dellIstituto Super di sanita 41:119–123.

Botstein D, White R, Skolnick M, Davis R (1980) Construction of a genetic-linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES(2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–5.

Braaten JT, Wood PJ, Scott FW, Wolynetz MS, Lowe MK, Bradleywhite P, Collins MW (1994) Oat beta-glucan reduces blood cholesterol concentration in hypercholesterolemic subjects. Eur J Clin Nutr 48:465-474.

Brown TA (2002) Genome, 2nd ed. John Wiley & Sons, Inc., Rexdale, ON.

Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA (2012) Development and mapping of SNP assays in allotetraploid cotton. Theor Appl Genet 124:1201-1214.

Canadian Food Inspection Agency (2013) List of registered varieties in Canada.

Canadian Seed Growers Association (2014) Canadian regulations and procedures: Canadian Seed Growers ' Association Circular 6.

Cavalli-Sforza L, Edwards A (1967) Phylogenetic analysis: Models and estimation procedures. Evolution (NY) 21:550–570.

Chial H (2008) DNA fingerprinting using amplified fragment length polymorphisms (AFLP): No genome sequence required. Nat Educ 1:176.

Coffman FA (1977) Oat history, identification and classification. 356.

Cortés AJ, Chavarro MC, Blair MW (2011) SNP marker diversity in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 123:827–845.

Cottage A, Gostkiewicz K, Thomas JE, Borrows R, Torres A-M, O'Sullivan DM. (2012) Heterozygosity and diversity analysis using mapped single nucleotide polymorphisms in a faba bean inbreeding programme. Mol Breed 30:1799–1809.

Cox TS, Kiang YT, Gorman MB, Rodgers DM (1985) Relationship between coefficient of parentage and genetic similarity indices in the soybean. Crop Sci 25:529-532.

De Koeyer DL, Phillips RL, Stuthman DD (1999) Changes in genetic diversity during seven cycles of recurrent selection for grain yield in oat, *Avena sativa* L. Plant Breeding 118:37-43.

Duerst R, Kaeppler H, Forseberg R (1999) Registration of 'Gem' oat. Crop Sci Sci 39:869–880.

Edward K, Poole R, Barker G (2008) SNP discovery in plants. In: Henry R (ed) Plant Genotyping II: SNP Technology. CABI, Wallingford, UK, pp 1–29.

Edwards K, Johnstone C, Thompson C (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. Nucleic Acids Res 19:1349.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379.

Evans KM, Patocchi A, Rezzonico F, *et al.* (2011) Genotyping of pedigreed apple breeding material with a genome-covering set of SSRs: trueness-to-type of cultivars and their parentages. Mol Breed 28:535–547.

FAOSTAT (2012) Food and agricultural commodities production. Available at: http://faostat.fao.org/site/339/default.aspx Accessed: March 17, 2014.

Flavell RB, Rimpau J, Smith DB (1977) Repeated sequence DNA relationships in four cereal genomes. Chromosoma 63:205-222.

Fu Y-B, Legge WG, Clarke JM, Richards KW (2006) Genetic variation and relationships of pedigree-known oat, wheat, and barley cultivars releaved by bulking and single-plant sampling. Genet Resour Crop Evol 53:1153–1164.

Fu Y-B, Kibite S, Richards KW (2004) Amplified fragment length polymorphism analysis of 96 Canadian oat cultivars released between 1886 and 2001. Can J Plant Sci 84:23–30.

Fu Y-B, Peterson GW, Scoles G, Rossnagel B, Schoen DJ, Richards KW. (2003a) Allelic diversity changes in 96 Canadian oat cultivars released from 1886 to 2001. Crop Sci 43:1989.

Fu Y-B, Whitwill S, Ferdinandez Y, Coulman BE, Richards KW (2003b) Bulking generated considerable bias in detection of AFLP variations in oat, fringed brome and smooth bromegrass. Mol Ecol Notes 3:476–478.

Gnanesh BN, Mitchell Fetch J, Menzies JG, Beattie AD, Eckstein PE, McCartney CA (2013) Chromosome location and allele-specific PCR markers for marker-assisted selection of the oat crown rust resistance gene *Pc91*. Mol Breed 32:679–686.

Guillin E, Baum BR, Mechanda S (1998) Development of an identification scheme for Canadian registered oat cultivars using RAPDs. Can J Plant Sci 78:605–610.

Gustafson P, Oliver R, Obert D (2010) Development of oat-based markers from barley and wheat microsatellites. Genome 53:458–471.

Haque KA, Pfeiffer RM, Beerman MB, Struewing JP, Chanock SJ, Bergen AW (2003) Performance of high-throughput DNA quantification methods. BMC Biotechnol 3:20.

He X, Bjørnstad Å (2012) Diversity of North European oat analyzed by SSR, AFLP and DArT markers. Theor Appl Genet 125:57–70.

He X, Skinnes H, Oliver RE, Jackson EW, Bjørnstad Å (2013) Linkage mapping and identification of QTL affecting deoxynivalenol (DON) content (*Fusarium* resistance) in oats (*Avena sativa* L.). Theor Appl Genet 126:2655–70.

Hizbai BT, Gardner KM, Wight CP, *et al.* (2012) Quantitative trait loci affecting oil content, oil composition, and other agronomically important traits in oat. Plant Genome J 5:164.

Holland JB, Helland SJ, Sharopova N, Rhyne DC (2001) Polymorphism of PCR-based markers targeting exons, introns, promoter regions, and SSRs in maize and introns and repeat sequences in oat. Genome 44:1065–1076.

Huang Y, Poland J, Wight C, Jackson EW, Tinker NA (2013) Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. Manuscript submitted for publication.

International Society for Horticultural Science (2009) International code of nomenclature for cultivated plants, 8th ed.

Izydorczyk MS, McMillan T, Bazin S, *et al.* (2014) Milling of Canadian oats and barley for functional food ingredients: Oat bran and barley fibre-rich fractions. Can J Plant Sci 94:573–586.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:E25.

Jellen EN, Gill BS, Cox TS (1994) Genomic in-situ hybridization differentiates between A/D-genome and C-genome chromatin and detects intergenomic translocations in polyploid oat species (Genus *Avena*). Genome 37:613-618.

Jellen EN, Beard J (2000) Geographical distribution of a chromosome 7C and 17 intergenomic translocation in cultivated oat. Crop Sci 40:256-263.

Kianian SF, Egli MA, Phillips RL, *et al.* (1999) Association of a major groat oil content QTL and an acetyl-CoA carboxylase gene in oat. Theor Appl Genet 98:884–894.

Kibite S, Menzies J (2002) Registration of 'AC Morgan'oat. Crop Sci 42:2210–2211.

Kim K-S, Tinker N, Newell M (2014) Improvement of oat as a winter forage crop in the Southern United States. Crop Sci (in press)

Kimura Y, Tosa Y, Shimada S, Sogo R, Kusaba M, Sunaga T.Betsuyaka S, Eto Y, Nakayashi H, Mayama S (2001) OARE-1, a Ty1-*copia* retrotransposon in oat activated by abiotic and biotic stresses. Plant Cell Physiol 42:1345-1354.

Kunkel T, Erie D (2005) DNA mismatch repair. Annu Rev Biochem 74:681–710.

Leggett J (1992) Classification and speciation in *Avena*. Oat Science and Technology — Agronomy Monograph no. 33

Li CD, Rossnagel BG, Scoles GJ (2000) The development of oat microsatellite markers and their use in identifying relationships among *Avena* species and oat cultivars. Theor Appl Genet 101:1259–1268.

Linares C, Serna A, Fominaya A (1999) Chromosomal organization of a sequence related to LTR-like elements of Ty-1 *copia* retrotransposons in Avena species. Genome 42: 706-713.

Liu S, Yang X, Zhang D, Bal G, Chao S, Bockus W (2014) Genome-wide association analysis identified SNPs closely linked to a gene resistant to soil-borne wheat mosaic virus. Theor Appl Genet 127:1039-1047.

Lorenzen LL, Boutin S, Young N, Specht JE, Shoemaker RC (1995) Soybean pedigree analysis using map-based molecular markers: I. Tracking RFLP markers in cultivars. Crop Sci 35:1326.

Lu F, Lipka AE, Glaubitz J, *et al.* (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet 9:e1003215.

Lucas M, Huynh B (2013) High-resolution single nucleotide polymorphism genotyping reveals a significant problem among breeder resources. Plant Genome 6:1–5.

Luo X, Wight CP, Zhou Y, Tinker NA (2012) Characterization of chromosome-specific genomic DNA from hexaploid oat. Genome 55:265-268.

Malécot G (1948) Mathématiques de l'hérédité.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220.

Market Analysis Group/Grains and Oilseeds Division (2014) Canada: Outlook for principal field crops.

Market Analysis Group/Grains and Oilseeds Division (2010) Market outlook report volume 2 number 3. 2:1–6.

McKenzie R, Harder D (1995) Oat. In: Slinkard A, Knott D (eds) Harvest of Gold: The History of Field Crop Breeding in Canada. University of Saskatchewan, Saskatchewan, SK, pp 98–112.

Melchinger AE, Messmer MM, Lee M, Woodman WL, Lamkey KR (1991) Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci 31:669–678.

Messmer MM, Melchinger AE, Herrmann RG (1993) Relationships among early European maize inbreds: II. Comparison of pedigree and RFLP data. Crop Sci 33:944–950.

Mitchell Fetch J, Brown PD, Duguid SD, *et al.* (2003) Ronald oat. Can J Plant Sci 83:101–104.

Mitchell Fetch J, Duguid SD, Brown PD, *et al.* (2007) Leggett oat. Can J Plant Sci 87:509–512.

Mitchell Fetch J, Brown PD, Duguid SD, *et al.* (2009) Jordan oat. Can J Plant Sci 89:67–71.

Montilla-Bascón G, Sánchez-Martín J, Rispail N, *et al.* (2013) Genetic diversity and population structure among oat cultivars and landraces. Plant Mol Biol Report 31:3015-1314.

Moorthy K, Babu P (2011) Identification of informative EST-SSR markers capable of distinguishing popular Indian rice varieties and their utilization in seed genetic purity assessments. Seed Sci Technol 39:282–292.

Murphy JP, Hoffman L (1992) The origin, history, and production of oat. Oat Science and Technology — Agronomy Monograph no. 33

Nava IC, Wight CP, Pacheco MT, Federizzi LC, Tinker NA (2012) Tagging and mapping candidate loci for vernalization and flower initiation in hexaploid oat. Mol Breed 30:1295–1312.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273.

Newell MA, Asoro FG, Scott MP, White PJ, Beavis WD, Jannink JL (2012) Genome-wide association study for oat (*Avena sativa* L.) beta-glucan concentration using germplasm of worldwide origin. Theor Appl Genet 125:1687–96.

Newell MA, Cook D, Tinker NA, Jannink J-L (2010) Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies. Theor Appl Genet 122:623–32.

O'Donoughue L, Souza E, Tanksley S (1994) Relationships among North-American oat cultivars based on restriction fragment length polymorphisms. Crop Sci 34:1251–1258.

O'Donoughue LS, Wang Z, Röder M, Kneen M, Leggett M, Sorrells ME, Tanksley SD (1992) An RFLP-based linkage map of oats based on a cross between two diploid taxa (*Avena atlantica* × *A . hirtula*). Genome 35:765–771.

Oksanen J, Guillaume Blanchet F, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2013). vegan: Community ecology package. R package version 2.0-10. Available at: http://CRAN.R-project.org/package=vegan

Oliver RE, Tinker NA, Lazo GR, *et al.* (2013) SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species. PLoS One 8:e58068.

Olufowote J, Xu Y, Chen X, Goto M (1997) Comparative evaluation of within-cultivar variation of rice (*Oryza sativa* L.) using microsatellite and RFLP markers. Genome 40:370–378.

Van Ooijen JW (2006) JoinMap ® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B.V., Wageningen, Netherlands.

Paczos-Grzeda E (2004) Pedigree, RAPD and simplified AFLP-based assessment of genetic relationships among *Avena sativa* L. cultivars. Euphytica 138:13–22.

Pal N, Sandhu JS, Domier LL, Kolb FL (2002) Development and characterization of microsatellite and RFLP-derived PCR markers in oat. Crop Sci 42:912.

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome J 5:92.

Poland JA, Endelman J, Dawson J, *et al.* (2012a) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome J 5:103.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012b) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253.

Portyanko VA, Hoffman DL, Lee M, Holland JB (2001) A linkage map of hexaploid oat based on grass anchor DNA clones and its relationship to other oat maps. Genome 44:249-265.

Riaz S, Tenscher AC, Smith BP, Ng DA, Walker MA (2008) Use of SSR markers to assess identity, pedigree, and diversity of cultivated muscadine grapes. J Am Soc Hortic Sci 133:559–568.

Rines HW, Molnar SJ, Tinker NA, Phillips RL (2006) Oat. In: Kole C (ed) Genome Mapping and Molecular Breeding in Plants, Vol. 1 Cereal. Millets. Springer Berlin / Heidelberg, pp 211–242.

Rogers J (1972) Measures of similarity and genetic distance. Studies in genetics. VII. University of Texas Publ, p 145.

Rooney WL, Rines HW, Phillips RL (1994) Identification of RFLP markers linked to crown rust resistance genes *Pc91* and *Pc* in oat. Crop Sci 34:940.

Rosso ML, Burleson S a., Maupin LM, Rainey KM (2011) Development of breeder-friendly markers for selection of *MIPS1* mutations in soybean. Mol Breed 28:127–132.

Seeds Act, Revised Statue of Canada (1985, c. S-8). Retrieved from the Department of Justice Canada website: http://laws-lois.justice.gc.ca/eng/acts/S-8/index.html

Seeds Regulations, Consolidated Regulations of Canada (c. 1400). Retrieved from the Department of Justice Canada website: http://laws-lois.justice.gc.ca/eng/regulations/C.R.C.,_c._1400/index.html

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. Mol Breed 33:1-14.

Semagn K, Beyene Y, Makumbi D, Mugo S, Prasanna BM, Magorokosho C, Atlin G (2012) Quality control genotyping for assessment of genetic identity and purity in diverse tropical maize inbred lines. Theor Appl Genet 125:1487–501.

Smith JSC, Chin ECL, Shu H, *et al.* (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): Comparisons with data from RFLPs and pedigree. Theor Appl Genet 95:163–173.

Soleimani V, Baum B, Johnson D (2002) AFLP and pedigree-based genetic diversity estimates in modern cultivars of durum wheat [*Triticum turgidum* L. subsp durum (Desf.) Husn.]. Theor Appl Genet 104:350–357.

Stadler LJ, Kirkpatrick RT (1930) Columbia oats, a new variety for Missouri, Volumes 270-284. 1–12.

Stanton TR (1955) Oat identification and classification. 206.

Sur R, Nigam A, Grote D, Liebel F, Southall MD (2008) Avenanthramides, polyphenols from oats, exhibit anti-inflammatory and anti-itch activity. Arch Dermatol Res 300:569–574.

Tinker NA, Fortin M, Mather D (1993) Random amplified polymorphic DNA and pedigree relationships in spring barley. Theor Appl Genet 85:976–984.

Tinker NA (2008) Plant breeding. In: Stewart CN (ed) Plant Biotechnology and Genetics: Principles, Techniques and Applications. John Wiley & Sons, Inc., Hoboken, NJ, pp 47–82.

Tinker NA, Deyl JK (2005) A curated internet database of oat pedigrees. Crop Sci 45:2269.

Tinker NA, Kilian A, Wight CP, *et al.* (2009) New DArT markers for oat provide enhanced map coverage and global germplasm characterization. BMC Genomics 10:39.

Tinker NA, Mather D (1993) KIN - Software for computing kinship coefficients. J Hered 84:238.

Uitdewilligen JGAML, Wolters A-MA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One 8:e62355.

Varshney R, Hoisington D, Nayak S, Graner A (2009) Molecular plant breeding: methodology and achievements. In: Gustafson JP, Langridge P, Somers DJ (eds) Plant Genomics: Methods and Protocols. Series: Methods in Molecular Biology. Vol. 513. Humana Press, Totowa, NJ, pp 283–304.

Vos P, Hogers R, Bleeker M, *et al.* (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–14.

Waters D, Bundock P, Henry R (2008) Genotyping by allele-specific PCR. In: Henry R (ed) Plant Genotyping II: SNP Technology. CABI, Wallingford, UK, pp 88–97.

Webster FH (1986) Oats: Chemistry and technology. American Association of Cereal Chemists, St. Paul, MN.

Welsh J, Carson R, Cherewick W, Hagborg WAF, Peturson B, Wallace HAH (1953) Oat varieties- past and present. Publication 891. Canada Department of Agriculture, Ottawa, Canada.

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Killian A (2004) Diversity array technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci USA 101:9915–9920.

Wight CP, Tinker NA, Kianian SF, *et al.* (2003) A molecular marker map in 'Kanota' x 'Ogle' hexaploid oat (*Avena* spp.) enhanced by additional markers and a robust framework. Genome 46:28–47.

Wight CP, Yan W, Mitchell Fetch J, Deyl J, Tinker NA (2010) A set of new simple sequence repeat and avenin DNA markers suitable for mapping and fingerprinting studies in oat (spp.). Crop Sci 50:1207.

Yan J, Yang X, Shah T, *et al.* (2009) High-throughput SNP genotyping with the GoldenGate assay in maize. Mol Breed 25:441–451.

Young N, Tanksley S (1989) Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor Appl Genet 77:95–101.

APPENDIX 1

Determining the optimal DNA extraction method for KASP™ samples

---

Purpose:

Leaf tissue from 'AC Morgan' and 'AC Rigodon' breeder seeds were subjected to five different DNA extraction methods to determine the optimal method for KASP™ sample preparation. Together these methods represented common laboratory techniques (CTAB and rapid) and commercially-available extraction kits optimized for plant DNA (Qiagen DNeasy and ZR Plant/Seed DNA Miniprep) and forensic samples (NucleoMag 96 Trace). Varying DNA purity and quality is expected from these different methods, as they differ in RNA and removal of PCR-inhibitors.

DNA extraction and quantification:

Additional leaf tissue (20 mg) from AC Morgan and AC Rigodon breeder seed was grown and prepared for extraction as described in section 2.3.1. Three commercially available extraction kits were used following respective manufacturers' protocols: NucleoMag 96 Trace (Macherey-Nagel, Germany), DNeasy Plant Mini kit (Qiagen Inc., Mississauga ON), ZR Plant/Seed DNA Miniprep (Zymo Research, Irvine CA). A CTAB method (Tinker *et al.,* 1993) and a rapid DNA extraction method (Edwards *et al.,* 1991) were used to represent other commonly used techniques. DNA concentration, A_{260}/A_{280} and A_{260}/A_{230} purity ratios were measured three times for each sample using a NanoPhotometer P-Class spectrophotometer (Implen, Germany). Agarose gel electrophoresis (1% agarose gel stained with GelRed (Biotium, Hayward CA)) was used primarily to assess DNA degradation, and to confirm

spectrophotometric estimations of DNA concentrations. In the event of a conflict, DNA concentration was approximated by visual comparison with a 50 bp DNA ladder (New England BioLabs Inc., Ipwich MA). All samples were diluted to 2 ng/μL working volumes with sterile distilled water for KASP™.

DNA quality and quantity:

Extraction kits yielded double or more DNA than the CTAB and rapid methods in both cultivars (Table S1). Methods yielded DNA from highest to lowest as follows: Kingfisher (653-1145 ng/mg), Qiagen (511-558 ng/mg), ZR (167-296 ng/mg), rapid (45 ng/mg), and CTAB (2.4 ng/mg). Agarose gel electrophoresis confirmed most sample DNA concentrations except for those extracted using the rapid method. Spectrophotometric estimation of DNA concentration can increase due to contamination of proteins, RNA, and salts (Haque *et al.,* 2003), which appears to be the case with these samples. The purity ratios indicate that DNA samples extracted from commercial kits (A_{260}/A_{280} =1.7-1.9) are less probable to contain ribonucleic acid (RNA) than those extracted using common laboratory methods (A_{260}/A_{280} > 2). However, samples visualized on a 1% agarose gel confirmed the presence of RNA in samples extracted using the Kingfisher kit and the rapid method (Figure S1). CTAB extraction yielded the least amount of DNA and was only faintly seen (Figure S1), thus the presence of RNA could not be confidently assessed. Only the Qiagen method involved digestion with RNase, however the ZR method appears to remove RNA during a wash step while DNA is bound to a filter. A_{260}/A_{230} purity ratios indicate that samples extracted using the two common laboratory methods contain significantly more organic compounds and

salts, which may act as PCR inhibitors, than the commercial kits ($p<0.05$). Electrophoresis also revealed that most samples were degraded to different extents.

Table S1 Quantity and quality of DNA extracted from AC Morgan and AC Rigodon using common laboratory techniques and commercial extraction kits. Averages of three replicate measurements are shown. Missing data is indicated by (-).

		CULTIVAR	
Method		**AC Morgan**	**AC Rigodon**
CTAB (Tinker, 1993)	conc (ng/µL)	2.72	3.33
	yield (ng/mg)	40.78	50.00
	A_{260}/A_{280}	2.42	2.39
	A_{260}/A_{230}	0.44	0.44
rapid (Edwards, 1991)	*conc (ng/µL)	9	< 9
	yield (ng/mg)	45	< 45
	A_{260}/A_{280}	2.09	2.06
	A_{260}/A_{230}	0.95	1.19
NucleoMag 96 Trace	conc (ng/µL)	131	229
	yield (ng/mg)	653.33	1145
	A_{260}/A_{280}	1.75	1.90
	A_{260}/A_{230}	1.25	1.52
Qiagen DNeasy Plant Mini	conc (ng/µL)	102	112
	yield (ng/mg)	511.17	558.33
	A_{260}/A_{280}	1.84	1.85
	A_{260}/A_{230}	1.92	2.28
ZR Plant/Seed DNA Miniprep	conc (ng/µL)	59.1	33.4
	yield (ng/mg)	295.67	167.17
	A_{260}/A_{280}	1.74	1.86
	A_{260}/A_{230}	2.74	-

* Mass of DNA was approximated by comparing intensity of bands to those of the ladder. Concentration was estimated as approximate mass divided by total volume of DNA loaded.

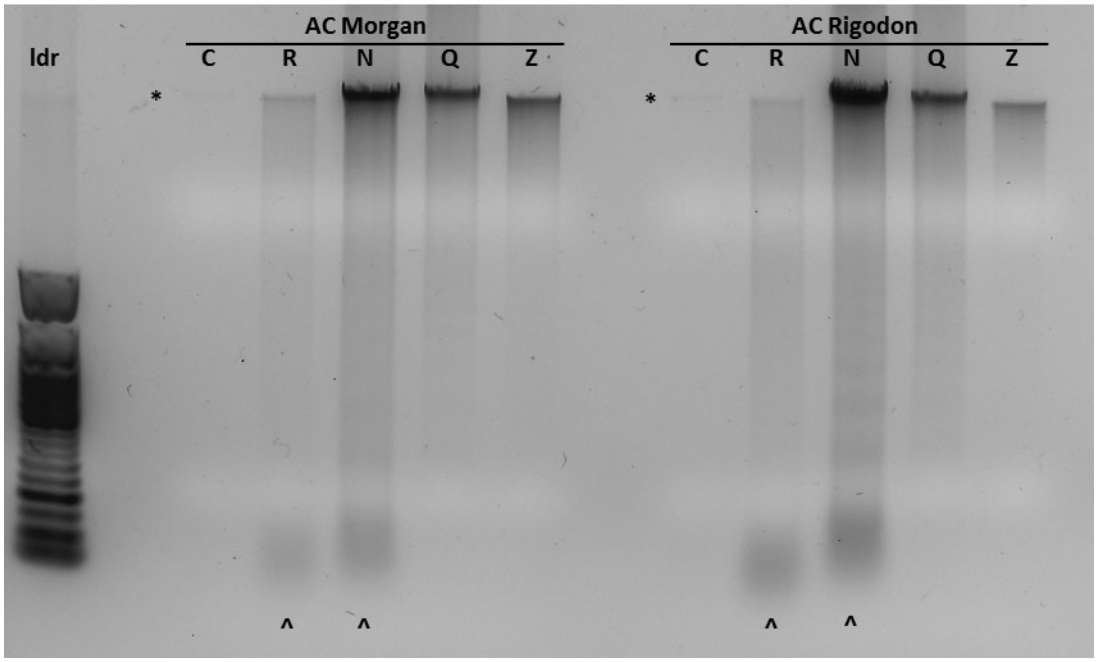

Figure S1 AC Morgan and AC Rigodon DNA extractions visualised on a 1% agarose gel. DNA (3 µL/lane) extracted using the CTAB method (C), rapid method (R), NucleoMag Trace 96 kit (N), Qiagen DNeasy Plant Mini kit (Q), and ZR Plant/Seed DNA Miniprep kit (Z) are presented with a 50 bp ladder (ldr, 5 µL). CTAB extractions (6 µL/lane) can be faintly seen and are indicated with a (*). RNA present in samples are indicated by (^). Gel was stained with GelRed (10 µL/150 mL gel).

Effect of DNA quality on KASP™ genotyping:

The 10 AC Morgan and AC Rigodon samples were assayed with K_avgbs38475 and K_avgbs75484 to assess if and how DNA quality affects the clustering pattern observed on allele discrimination plots (ADPs). Triplicate assays were performed as described in section 2.3.4. The clustering pattern determines how software and its users make allele calls. K_avgbs38475 and K_avgbs75484 were chosen to assay these DNA samples based on their observed cluster tightness and separation after assays with the CDOCs (Figure S2). Assays for K_avgbs38475 produced clusters that were clearly separate but moderately tight, while K_avgbs75484 gave clusters that were moderately separate but clustered tightly. AC Morgan and AC Rigodon were also expected to be distinguished from each other using these assays. Both KASP™ assays were capable of amplifying all the DNA samples from AC Morgan and AC Rigodon. Assays performed with AC Morgan samples produced five unexpected allele calls, while those performed with AC Rigodon produced four (Figure S3). However, when all three assays per sample were considered, the resulting genotype was as expected. The CTAB-extracted DNA from AC Rigodon consistently gave the incorrect allele call when assayed with K_avgbs75484. Samples that were consistent between triplicate assays were found to cluster together in an easily interpreted fashion (Figure S4A). Additionally, the samples that had one inconsistent allele call were also easily interpreted as they were clearly separated in ADPs (Figure S4B). Therefore, extraction method does not appear to have an effect on ADP interpretation.

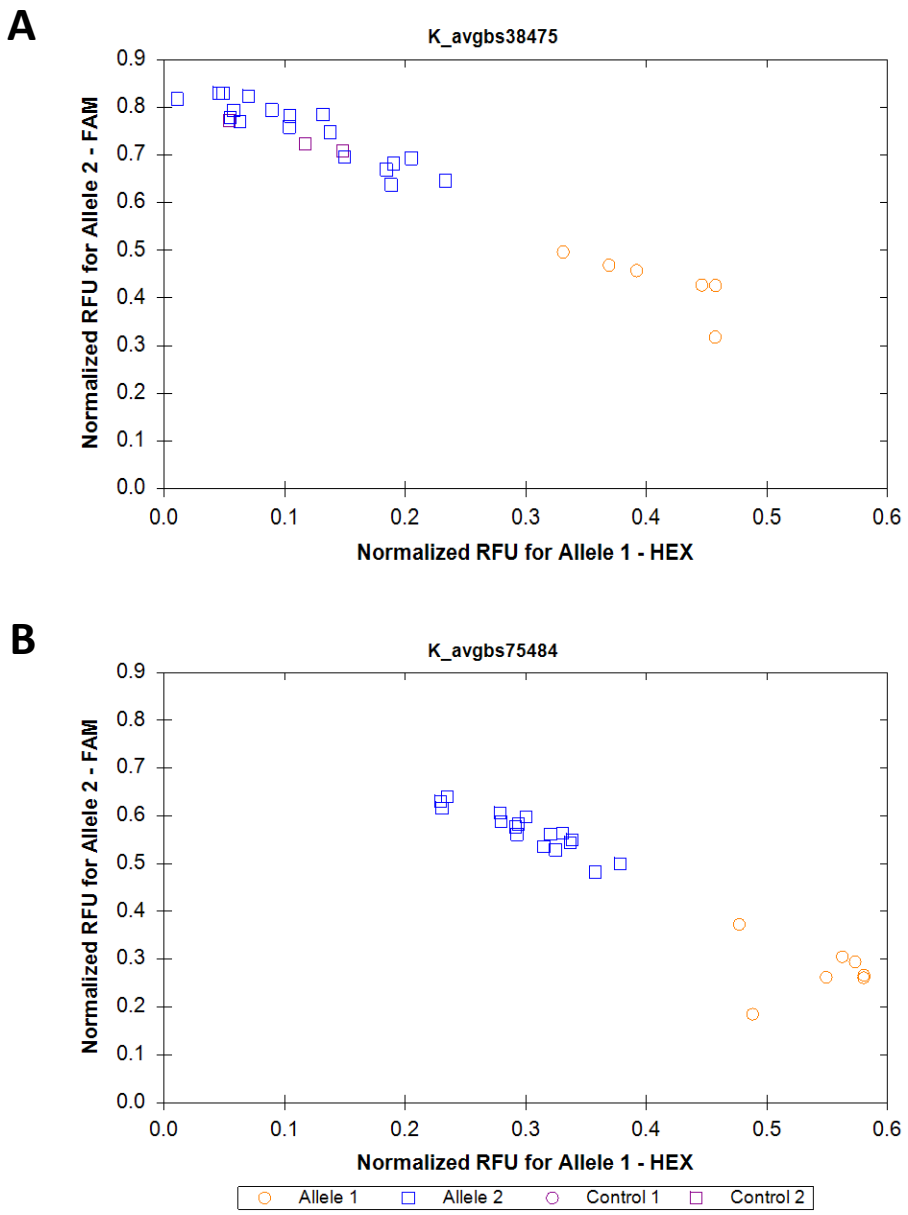

Figure S2 Allele discrimination plots of (A) K_avgbs38475 and (B) K_avgbs75484 triplicate assays with the 24 Canadian Diagnostic Oat Cultivars.

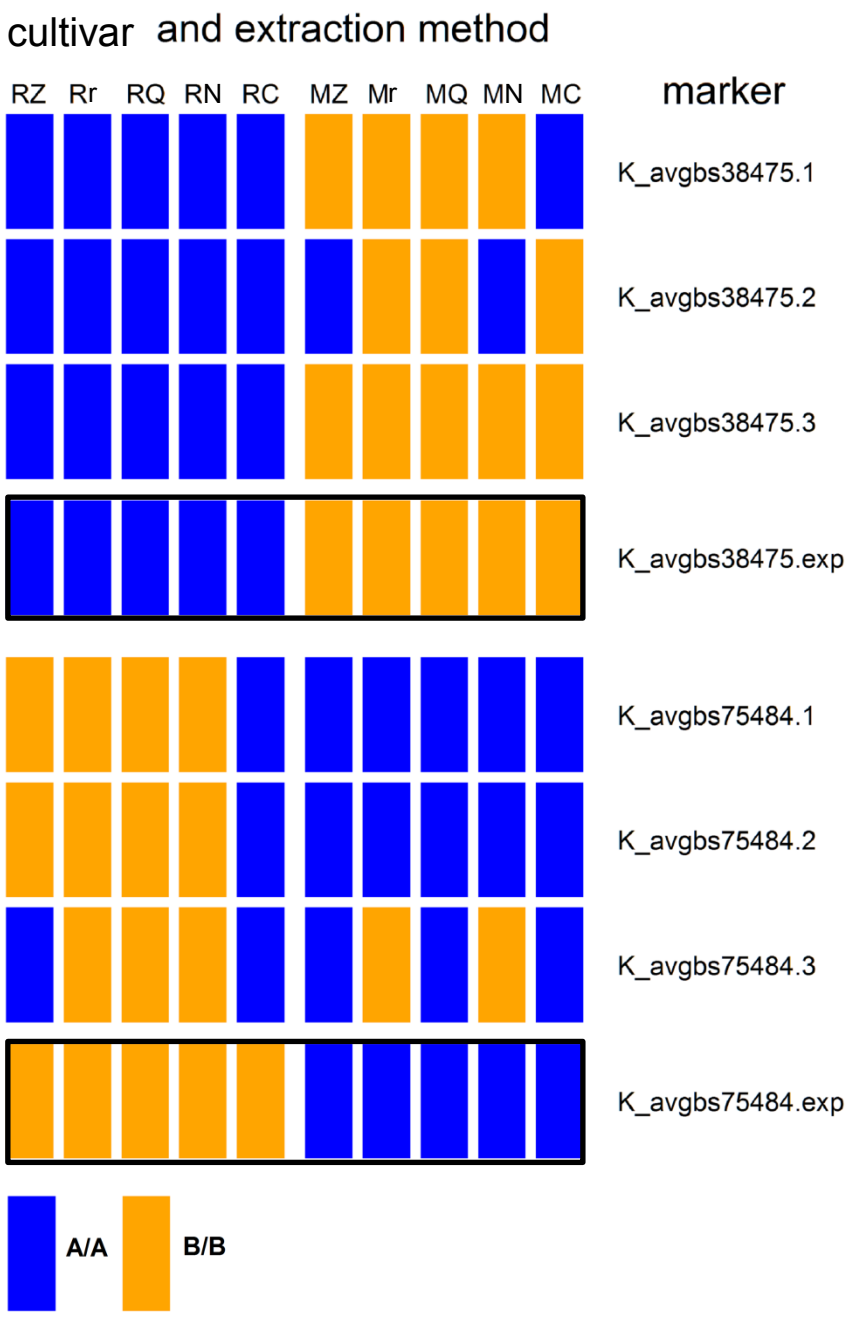

Figure S3 Bitmap of triplicate genotype results for AC Morgan (M) and AC Rigodon (R) samples extracted using five different methods. CTAB (C), NucleoMag 96 Trace (N), Qiagen DNeasy (Q), rapid (r), ZR Plant/Seed MiniPrep (Z). Alleles A and B correspond to the alleles associated with FAM and HEX in each KASP assay, respectively (Table 2.3). Expected (.exp) genotype results for each marker are highlighted in the black squares.

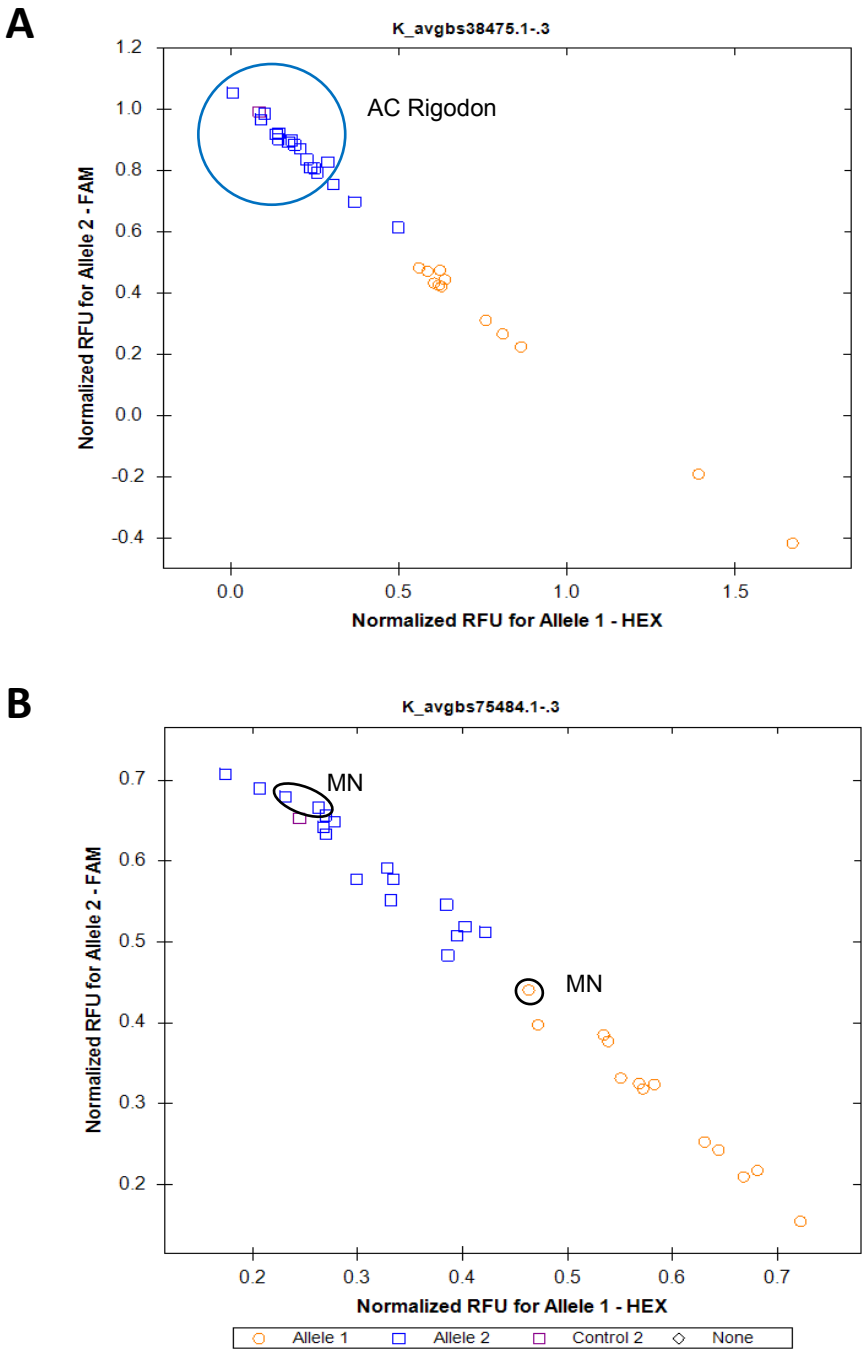

Figure S4 Allele discrimination plots of (A) K_avgbs38475 and (B) K_avgbs75484 assays with AC Morgan (M) and AC Rigodon (R) DNA samples. AC Morgan samples extracted using the NucleoMag Trace 96 kit (MN) are noted to demonstrate clear separation of an inconsistent allele call.

Suitability of various DNA extraction methods for KASP™:

Based on the results obtained (Figures S3-4), all extraction methods yielded DNA that was usable for KASP™. Any organic compounds or salts that were introduced from extraction did not hinder PCR, given that the CTAB (A_{260}/A_{230}= 0.44) and rapid (A_{260}/A_{230}= 0.95, 1.19) samples were able to generate fluorescence in the KASP™ assay. However, since the AC Rigodon (RC) sample consistently produced the opposite genotype than expected, the CTAB extraction method is not suggested for KASP™. AC Morgan and AC Rigodon samples extracted using Qiagen were the only samples to consistently produce the expected genotypes (Figure S3). However, when results from all three replicate assays for each extraction sample were considered, the expected genotype was still produced. This emphasizes the importance of doing replicate assays when using the developed diagnostic set of KASP™ assays.

ADP clustering was not affected by DNA quality, thus there appears to be no added benefit of using a specific method in terms of data interpretation. However, since Qiagen's DNeasy kit produced high quality DNA (RNA and inhibitor-free) and consistent genotypes it warrants recommendation for future KASP™ work. A brief review of studies involving KASP™ shows that the Qiagen method is the standard; however CTAB and KingFisher products have also been reported. In situations where diagnostics must be done on a large number of samples, it may be more efficient to extract DNA using the NucleoMag 96 Trace kit. This commercial kit is used with the automated KingFisher system to extract 96 samples at a time. KASP™ appears to function well with varying quality of DNA, however Qiagen DNeasy kits are recommended for sample preparation.

APPENDIX 2

List of supplemental tables and figures