Accuracy of genomic selection in biparental populations of flax (*Linum usitatissimum* L.)

CrossMark

Frank M. You[a,*], Helen M. Booker[b], Scott D. Duguid[a], Gaofeng Jia[a,b], Sylvie Cloutier[c]

[a]Morden Research and Development Centre, Agriculture and Agri-Food Canada, Morden, MB R6M 1Y5, Canada
[b]Crop Development Centre, Department of Plant Sciences, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada
[c]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON K1A 0C6, Canada

ARTICLE INFO

ABSTRACT

Flax is an important economic crop for seed oil and stem fiber. Phenotyping of traits such as seed yield, seed quality, stem fiber yield, and quality characteristics is expensive and time consuming. Genomic selection (GS) refers to a breeding approach aimed at selecting preferred individuals based on genomic estimated breeding values predicted by a statistical model based on the relationship between phenotypes and genome-wide genetic markers. We evaluated the prediction accuracy of GS (r_{MP}) and the efficiency of GS relative to phenotypic selection (*RE*) for three GS models: ridge regression best linear unbiased prediction (RR-BLUP), Bayesian LASSO (BL), and Bayesian ridge regression (BRR), for seed yield, oil content, iodine value, linoleic, and linolenic acid content with a full and a common set of genome-wide simple sequence repeat markers in each of three biparental populations. The three GS models generated similar r_{MP} and *RE*, while BRR displayed a higher coefficient of determination (R^2) of the fitted models than did RR-BLUP or BL. The mean r_{MP} and *RE* varied for traits with different heritabilities and was affected by the genetic variation of the traits in the populations. GS for seed yield generated a mean *RE* of 1.52 across populations and marker sets, a value significantly superior to that for direct phenotypic selection. Our empirical results provide the first validation of GS in flax and demonstrate that GS could increase genetic gain per unit time for linseed breeding. Further studies for selection of training populations and markers are warranted.

Crown Copyright © 2016 Production and hosting by Elsevier B.V. on behalf of Crop Science Society of China and Institute of Crop Science, CAAS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Genomic or genome-wide selection (GS) is a breeding method based on the relationship between phenotype and a genome-wide set of genetic markers. A practical GS approach in breeding includes several steps [1–3]: (1) construction of an optimal training population that is genetically diverse and large; (2) phenotyping individuals of the training population in multiple environments; (3) genotyping individuals of the training population with a genome-wide set of genetic markers; (4) fitting an optimal statistical model based on the phenotypic and genotypic data, and estimating marker effects in the model; (5) genotyping test individuals with the markers used in GS model fitting; and (6) applying the GS model to

---

estimate genomic estimated breeding values (GEBVs) of test individuals for selection. GS has been proposed to be superior to conventional phenotypic selection and maker assisted selection (MAS) in time and money savings, and thereby to increase the efficiency of plant breeding [4].

GS is commonly applied in animal breeding [2], and extensive studies of GS in plants have been performed since 2007. Two research approaches have been employed to evaluate the efficiency of GS in plant breeding. The first is based on simulated [5–10] or real [11–13] marker data with simulated population data. Computer simulation is advantageous for generating data based on strict assumptions and for investigating the relationship of GS accuracy with different levels of factors influencing GS such as population type, marker density, linkage disequilibrium of populations, QTL number, and population size. The second approach is to use empirical data, a practical approach to demonstrating actual GS efficiency in plant breeding. To date, results of evaluation of GS accuracy have been reported for several annual crops such as maize [14–23], barley and *Arabidopsis* [14], wheat [4,16], rice [24–26], sugar beet [27], and sugarcane [28], and perennial trees such as loblolly pine, eucalyptus [29–31], and apple [32].

For millennia, flax has been used as a food source and to produce durable fibers and linen. More recently, flax has become an important multi-purpose crop, owing to an increasing demand for both oil and fiber [33]. Flax seeds typically contain 35–50% oil composed of five main fatty acids: palmitic (PAL, 6.0%), stearic (STE, 2.5%), oleic (OLE, 19.0%), linoleic (LIO, 13.0%), and linolenic (LIN, 55.0%) [34,35]. LIN is also referred to as α-linolenic acid (ALA). Recent work has shown that flax's omega-3 fatty acids (LIN) and plant estrogens contribute to reducing blood cholesterol levels and mitigate heart disease and certain cancers in humans [36–38]. The major breeding aims of linseed development are high seed yield (YLD), high oil content (OIL), and high (>65%), or low (2–4%) LIN content. The first registered high-LIN linseed cultivar in Canada is NuLin 50 with 68% LIN (http://www. viterra.ca) [39]. Also, low-LIN (2–4%) and high-LIO (65–70%) cultivars have been obtained by mutation breeding [39–41]. High-LIN flaxseed is one of the richest dietary sources of ALA and is also a good source of soluble fiber mucilage [42], whereas low LIN in seeds will effectively improve the oxidative stability and suitability of linseed oil for food uses [43]. Flax straw and its processed forms are widely used in the manufacturing of fine papers and some industrial fiber products such as the interior paneling of vehicles (http://www.flaxcouncil.ca/english/index.jsp). The major breeding aims of fiber flax are increased straw yield, fiber content in straw, fiber quality, and resistance to disease and abiotic stresses. However, phenotyping of seed yield, seed quality, and fiber traits is time-consuming, labor-intensive, and consequently costly. In addition, most of these traits are quantitative. Conventional breeding using phenotyping continues to predominate in flax breeding programs. Even MAS based on single QTL can be ineffectual because of potential overestimated QTL effects and small proportions of the genetic variation explained by the QTL [3]. By predicting breeding values of these traits for selection without prior phenotyping, GS provides an alternative approach to the quantitative traits in flax breeding.

Single sequence repeats (SSRs) or microsatellites are stretches of DNA containing a variable number of short tandem repeats. They are generally codominant, highly polymorphic, abundant, and reliable, and can be readily developed from existing genomic sequences or expressed sequence tags (ESTs) [44]. Currently more than 1400 SSR markers have been developed in flax from EST libraries [45–47] or genomic sequences [45,48–51]. These SSR markers have been used for the construction of genetic maps [52,53], genetic diversity assessment [54,55], QTL mapping [53], and association studies [56,57]. A total of 770 markers were incorporated into a consensus map from three biparental populations. The map had a total length of 1551 cM with a mean marker density of one marker every 2 cM and covered an estimated 74% of the predicted flax genome size of 370 Mb [52]. Thus, these SSR markers span most regions of the flax genome. Using the same set of SSR markers, Cloutier et al. detected two major QTL each for LIO, LIN, and iodine value (IOD), and one major QTL for PAL, in a doubled-haploid (DH) population of 78 lines generated from a cross between SP2047 and UGG5-5 [53]. Soto-Cerda et al. [56] used the association mapping approach and a flax core collection of 390 accessions with 460 SSR markers to identify QTL for seed quality traits in this germplasm collection. A total of nine QTL were associated with OIL, LIO, and LIN, some of which colocalized with QTL previously identified in the SP2047/ UGG5-5 biparental population [53]. These previous studies provide data useful for validating results from GS in flax.

The objective of this study was to explore the feasibility of GS in flax breeding by comparing accuracies and relative efficiencies of genomic prediction in multiple biparental populations using genome-wide SSR marker sets and several GS statistical models for seed yield and seed quality traits in flax.

2. Materials and methods

2.1. Populations

Three biparental populations were used for evaluation of GS. The first population (BM) was generated by single-seed descent from a cross between CDC Bethune [58] and Macbeth [59], and consisted of 243 F_6-derived recombinant inbred lines (RILs). Its two parents were high-yielding Canadian linseed cultivars containing 55–57% LIN [58,59]. The second population (EV) comprised 90 F_6-derived RILs from a cross between E1747, an ethyl methane sulphonate (EMS)-induced low LIN breeding line [60], and Viking, a French fiber flax cultivar grown widely in 2000 but deregistered in 2012. The third population (SU) was an F_1-derived doubled haploid (DH) population of 78 lines obtained from a cross between the breeding line SP2047, which gave rise to a yellow-seeded Solin variety called Linola 2047 that contains only 2–3% LIN, and breeding line UGG5-5, which is a high-LIN line with 63–66% LIN [53,61]. These three populations have been used for genetic mapping and QTL detection [52,53,62].

2.2. Phenotypic data

Lines from the three biparental populations were evaluated in field tests over 3 or 4 years (2009–2012) at two sites (Morden,