Tree Genetics & Genomes (2011) 7:1069–1078
DOI 10.1007/s11295-011-0395-1

ORIGINAL PAPER

Breeding without breeding: minimum fingerprinting effort with respect to the effective population size

Milan Lstibůrek · Kristýna Ivanková · Jan Kadlec · Jaroslav Kobliha · Jaroslav Klápště · Yousry A. El-Kassaby

Received: 30 June 2010 / Revised: 25 March 2011 / Accepted: 19 April 2011 / Published online: 11 June 2011

Abstract We present a probabilistic model to minimize the fingerprinting effort associated with the implementation of the "breeding without breeding" scheme under partial pedigree reconstruction. Our approach is directed at achieving a declared target population's minimum effective population size (N_e), following the pedigree reconstruction and genotypic selection and is based on the graph theory algorithm. The primary advantage of the proposed method is to reduce the cost associated with fingerprinting before the implementation of the pedigree reconstruction for seed parent–offspring derived from breeding arboreta and production or natural populations. Stochastic simulation was conducted to test the method's efficiency assuming a simple polygenic model and a single trait. Hypothetical population consisted of 30 parental trees that were paired at random (selfing excluded), resulting in 600 individuals (potential candidates for forwards selection). The male parentage was assumed initially unknown. The model was used to estimate the minimum genotyping sample size needed to reaching the prescribed N_e. Results were compared with the known pedigree data. The model was successful in revealing the true relationship pattern over the whole range of N_e. Two to three offspring entered genotyping to meet the $N_e = 2$ while 41 to 43 were required to satisfy the $N_e = 14$. Importantly, genetic gain was affected at the lower limits of the genotyping effort. Doubling the number of parents resulted in considerable reduction of the genotyping effort at higher N_e values.

Communicated by R. Burdon

M. Lstibůrek (✉) · K. Ivanková · J. Kobliha · J. Klápště
Department of Dendrology and Forest Tree Breeding, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21 Praha 6, Czech Republic
e-mail: lstiburek@fld.czu.cz

K. Ivanková
Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague, Opletalova 26, 110 00 Praha 1, Czech Republic

J. Kadlec
Faculty of Mathematics and Physics, Charles University in Prague, Ke Karlovu 3, 121 16 Praha 2, Czech Republic

J. Klápště · Y. A. El-Kassaby
Department of Forest Sciences, Faculty of Forestry, University of British Columbia, 2424 Main Mall, V6T 1Z4 Vancouver, BC, Canada

Keywords Breeding without breeding · Pedigree reconstruction · Selection

Introduction

Unraveling the genealogical relationships among members of populations is of great importance to many aspects of biological sciences. The availability of numerous highly variable neutral DNA markers such as microsatellites coupled with the development of sophisticated mathematical methods for pedigree reconstruction (see Jones and Ardren 2003 for a review) has made this task manageable. Pedigree reconstruction has proven to be of substantial value to conservation biology, ecological genetics, and, more recently, has

been advanced as a complement to conventional tree breeding methods (Lambeth et al. 2001; Grattapaglia et al. 2004; El-Kassaby and Lstibůrek 2009).

The "breeding without breeding (BWB)" concept was proposed by El-Kassaby and Lstibůrek (2009) as a complement or even alternative to classical tree breeding (see Massah et al. 2009). The method uses either partial or full pedigree reconstruction to assemble unstructured full- (FS) and half-sib (HS) families originating from natural mating among parents in breeding arboreta and their eventual use in quantitative genetics analyses for backwards, forwards, and combined selection. Several factors affect the efficiency of the proposed method; namely, (1) the number of parents within a breeding arboretum, (2) the size of half-sib family selected for fingerprinting and subsequently pedigree reconstruction, (3) the effectiveness of pedigree reconstruction in family assemblage (number of loci used and their informative nature and the declared types I and II error rates), (4) the level of gene flow (the higher the gene flow the greater the number of uninformative genotyping events), (5) female and male fertility variation, and (6) the quality of the open-pollinated (OP) progeny testing. The number of parents in the breeding arboreta, the portion of OP families selected for fingerprinting, and the level of gene flow all directly affect the DNA fingerprinting effort and ultimately the cost associated with assembling half- and full-sib families. To capitalize on the advantage of the proposed BWB method, the size of investment associated with DNA fingerprinting should be optimized. To address the DNA fingerprinting cost, we subdivide the problem into three components: (1) the genetic response to selection, (2) the effectiveness of the pedigree reconstruction model, and (3) the desired minimum effective population size of the selected individuals needed to establish a production population (seed orchard). While these components seem to be independent from each other, their interplay can substantially affect the DNA fingerprinting investment.

One way to optimize the sample size of each OP family for DNA fingerprinting is to restrict the within-family sampling to the top individuals based on their HS-based breeding values. El-Kassaby and Lstibůrek (2009) presented evidence that the genetic response to selection, in BWB, is a function of the proportion of male parents entering the pedigree reconstruction, and furthermore, they demonstrated that with fingerprinting a relatively small proportion of the candidate population, significant genetic response can be achieved in accordance with the theoretical expectations (unpublished results). This conclusion holds true not only for simple truncation selection, but also for constrained selection, where the constraint is the effective population size. This is the selection method used by practitioners, e.g., where the goal is to establish a seed orchard with a prescribed effective population size.

While the El-Kassaby and Lstibůrek (2009) study was based on a retrospective approach, in the current study, we are investigating the minimum DNA fingerprinting efforts with respect to the effective population size (N_e) with the primary question: how many individuals should be genotyped/fingerprinted in order to meet the declared minimum N_e? This assumes no a priori knowledge of the relatedness patterns due to sharing of common male parents. We developed a probabilistic model to assist in calculating the appropriate sample size needed to plan the DNA fingerprinting activities, and consequently the focus is directed to determining the minimum numbers of individuals required, rather than the actual costs. A cost function could be developed specifically, taking our model into consideration.

Model

General description

Our main goal was to select a set of unrelated individuals from a large candidate population originated from open pollination among members of a breeding arboretum. The selected individuals will be used to establish a production population (seed orchard) through what is known as forwards selection. The genealogical restriction among the selected individuals, in this case, requires selecting only one single individual from every HS family. Assuming that the original founders (parents within the breeding arboretum) were all unrelated (total of n_s), then it is advantageous to select the single offspring with the highest predicted breeding value within each HS family, where the resulting effective population size (N_e) of the orchard is simply determined by the number of selected families (i.e., complete avoidance of full- or half-siblings). It should be pointed out that this breeding value is based on HS performance evaluation, thus, it is conceivable that some of the selected individuals could be sharing common male parents (which will be revealed during the pedigree reconstruction). This inevitably reduces the N_e, and this reduction will be proportional to the number of male parents in the breeding arboretum and could be further reduced by the magnitude of male gametic contribution

variance (i.e., male fertility variation). Conversely, if many individuals per family were included in the genotyping phase prior to pedigree construction, this may increase the genotyping cost without significant benefit to the selection outcome. We, therefore, searched for a set of n offspring whose effective population size is greater than N_e with probability higher than a prescribed probability P. In other words, the breeder is choosing how many unrelated individuals should be put in an orchard. Thus, the census number n will be equivalent to the number of genotyped samples and subsequently those used in the pedigree reconstruction. This basic system is depicted in Fig. 1.

Prerequisites and key assumptions

Let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with corresponding mean μ and variance σ^2. In this study, we took advantage of the well-known property of location-scale families, i.e., that two parameters are sufficient to describe both location and shape of the distribution. Second, the probability content within two and three standard deviations of the mean is 0.9544 and 0.9974, respectively. Thus, occurrence of values outside such an interval is rare.

We assume a simple polygenic trait in a founder population consisting of unrelated and non-inbred trees. Additive genetic effects (a) are normally distributed $\mathcal{N}(\mu_S, \sigma_a^2)$, where μ_S is an arbitrary mean and σ_a^2 is a variance of additive effects. Random mating is assumed, with no selfing or contamination by alien pollen donors (thus, all potential parents could be identified). Thus, n_d offspring are distributed randomly across a notional complete diallel scheme (including reciprocal crosses, but excluding diagonal elements). Considering a cross between female x and male y, the distribution of additive genetic effects in their offspring is

$$\mathcal{N}\left(\frac{a_x + a_y}{2}, \frac{\sigma_a^2}{2}\right), \tag{1}$$

where a_x and a_y are additive genetic effects of female and male parents, respectively.

Hypothetically, seed is collected from each female parent (e.g., plus-tree selections in a breeding orchard) and the OP progeny trial is established. The orchard has been in a reproductive age long enough that the OP seed could be collected from each clone. For each individual offspring, the mother tree is assumed known and can be checked with 100% accuracy and the father tree is initially assumed unknown.

Next, offspring are hypothetically measured at a suitable selection age and sorted by the predicted additive genetic values within each family. Based on our model, the best set of n offspring is determined, and

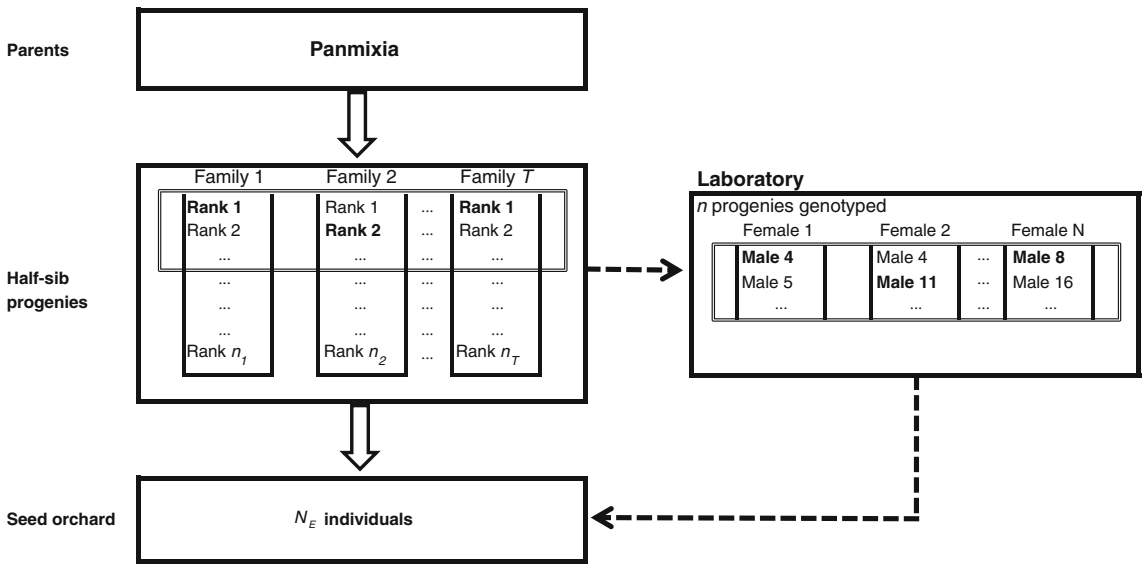

Fig. 1 Schematic model description showing: (1) random mating among the parental population (assuming panmixia and no gene flow), (2) establishment of open-pollinated progeny test, (3) ranking of offspring within each OP family based on their maternal designation (half-sib), (4) selection of the top individuals within each open-pollinated family for genotyping (n), (5) offspring paternal assignment through partial pedigree reconstruction, and (6) the establishment of a seed orchard with the desired N_e (forwards selection). The goal of the model is to estimate n to satisfy N_e with a probability P prior to genotyping

these offspring are subjected to genotyping. New seed orchard is then established out of the best unrelated offspring.

The algorithm is based on the assumption that there exists variation among additive genetic values within a family, as expressed in Eq. 1. Each hypothetical parental combination (i.e., family) in the random mating scheme (diallel excluding selfing) is, therefore, characterized by a specific distribution. Since all families share the same expected variance (polygenic Mendelian segregation), the only difference among respective distributions is due to the different expected mean value (average parental additive value). Due to this variability among families, a given distribution (family) will be fully or partially merged with other distributions while some would be separated further apart, potentially sharing only extreme observations. We can then calculate a conditional probability of a specific value a belonging to a collection of distributions.

For a given offspring i, one parent (female, x) is known prior to the pedigree reconstruction. Let $f_{x,y}$ be a probability density function for a given family (expressing the distribution of the additive genetic values for a specific combination of female (x) and male (y) parents, respectively). The conditional probability $p_{i,y}$ of the ith offspring having a sire y is then given by

$$p_{i,y} = \frac{f_{x,y}(a_i)}{\sum_{k=1}^{n_s} f_{x,k}(a_i)}, k \neq x. \quad (2)$$

Simulation scenarios

We simulated a population consisting of $n_S = 30$ parental trees (contributing as both males and females) and $n_d = 600$ offspring in total (distributed randomly in a complete diallel excluding selfing). This experiment was repeated 100 times for every value of the parameter N_e. Mother trees were assumed known and father trees initially unknown. Parents were assigned randomly to each offspring (satisfying constant family sizes) using the polygenic model (Eq. 1) with the mean additive effect set to 100 and the additive variance set to 100.

An alternative scenario consisted of $n_S = 60$ and $n_d = 2{,}400$ with all additional input parameters being equal to the first scenario.

Algorithm L

The algorithm works by choosing sets of offspring. It estimates the probability p_V that a set V contains N_e independent offspring (i.e., no siblings) and uses this information to evaluate the set. The best up-to-date set is randomly modified to produce new sets, approaching the most diverse ones using a simulated annealing strategy.

We define a score function $\textsc{Score}(V)$ which is used to evaluate a given set of offspring V and to compare it to alternative other ones. It is necessary to quantify the proximity of the estimated $\hat{p}_V$ to the target probability value P. Increasing the p_V value over the P value is not considered ($\hat{p}_V$ beyond P provides no benefit). The score function is then

$$\textsc{Score}_1(V) = \min(\hat{p}_V, P). \quad (3)$$

Aside from the primary consideration (probability), the set should contain trees with high additive genetic (breeding) values—a_V should be considered while the set is formed. This inevitably compromises the maximization of $\hat{p}_V$ (since a set with high a_V is likely derived out of a small number of elite families). From the optimization perspective, a_V should be considered only as $\hat{p}_V$ starts to approach P, (i.e., the algorithm is approaching the space of valid solutions). Therefore, a parameter ϵ has been introduced to regulate the proximity (along the optimum) at which the a_V is considered:

$$\textsc{Score}_2(V) = a_V \cdot (\min(\hat{p}_V, P)/P)^{\epsilon}. \quad (4)$$

Proper weighting of both scores (Eqs. 3 and 4 considered jointly) is controlled by the parameter κ:

$$\textsc{Score}_3(V) = \textsc{Score}_1(V) + \kappa \cdot \textsc{Score}_2(V). \quad (5)$$

There is still a problem associated with this score function. When the parameter N_e is high, it becomes very difficult to identify solutions with $\hat{p}_V > 0$ at the beginning of the algorithm. We tackle this problem by enhancing the score with the estimate of the number of independent trees ($\hat{I}_V$) in the set. The final score is then:

$$\textsc{Score}(V) = \textsc{Score}_3(V) + \frac{\hat{I}_V \cdot \max(0, 0.5 - \hat{p}_V)}{1{,}000}. \quad (6)$$

Let $T_{\max}$ be the number of tested samples (containing the desired set of n offspring) without any score improvement, α is the probability of a type I error associated with the p_V estimate (defined later) and $o_{\max}$ is the maximum number of random arrangements for estimating the p_V. The algorithm is defined as follows (please, note that n^* denotes a provisional value):

(i) $n^* \leftarrow \max(2, N_e - 1)$ {number of trees in the set}
$n_{\max} \leftarrow \min(n_s/2, n_d)$

(ii) $W \leftarrow$ random sample of n^* offspring {the best set so far}
$T \leftarrow 0$ {number of unsuccessful trials}

(iii) $V \leftarrow$ modified W: each offspring in the set is substituted (with a probability $1/n^*$) by a random candidate. If n^* was increased, extra offspring are randomly added to the set.

$$\text{Score}(V) \leftarrow \begin{cases} \hat{p}_V + \kappa a_V (\hat{p}_V / P)^{\varepsilon} + \dfrac{\hat{I}_V (0.5 - \hat{p}_V)}{1{,}000} & \text{if } \hat{p}_V < 0.5 \\ \hat{p}_V + \kappa a_V (\hat{p}_V / P)^{\varepsilon} & \text{if } 0.5 \leqq \hat{p}_V < P \\ P + \kappa a_V & \text{if } \hat{p}_V \geqq P \end{cases} \tag{7}$$

where

$\hat{p}_V$ is the estimated probability of finding N_e individuals within the set V, as computed by algorithm M in the Appendix

$\hat{I}_V$ is the estimated number of independent individuals in the set,
$a_V = \frac{1}{n^*} \sum_{i \in V} a_i$

(iv) if V has a higher score then W, set $W \leftarrow V$ and $T \leftarrow 0$, {a better set was identified}
otherwise $T \leftarrow T + 1$

(v) if $T = T_{\max}$, then: {W has not been improved over many iterations}

if $\hat{p}_W \geqq P$, set $n_{\max} \leftarrow \min(n_{\max}, n^*+3)$ {success: try higher n^*}
if $n^* \geqq n_{\max}$, finish
$n^* \leftarrow n^*+1$, $T \leftarrow 0$ {next iteration}

(vi) go to (ii)

Individual steps of the algorithm are described in the following part.

ad(i) Variables are initialized. The initial set size n^* is $N_e - 1$, despite it is not satisfying the requirements. The justification is that the algorithm can find sets with $\hat{p}_V > 0$ more easily (reminiscent of the introduction of $\hat{I}_V$ into the score).

ad(iii) A new set is created by modifying the best existing set (the one associated with the best

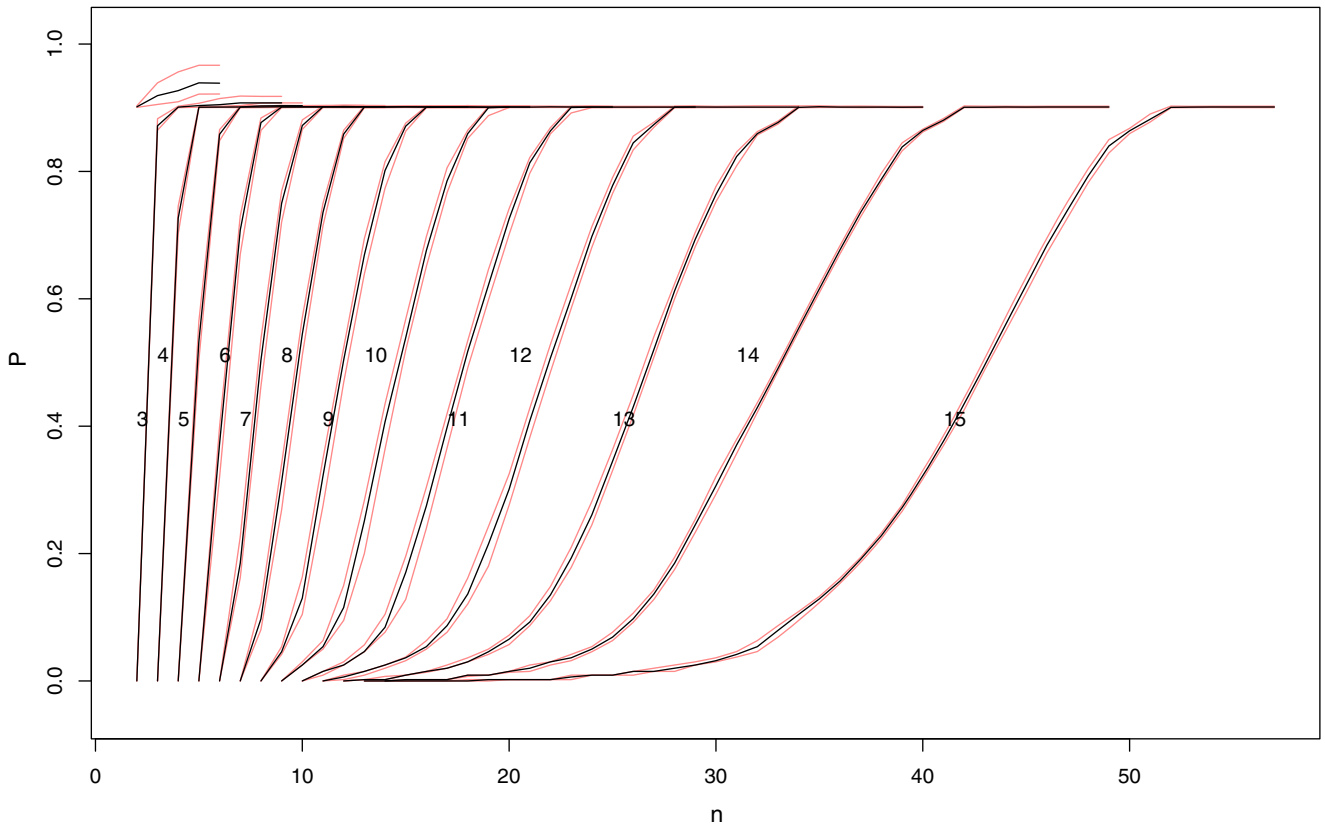

Fig. 2 Probability estimate (y-axis) plotted against n, x-axis. Different *curves* correspond to different prescribed N_e. *Bolded curves* depict medians of 100 iterations; *remaining curves* are 25% and 75% quantiles

score). If the new set should contain additional offspring, random ones are added after this modification. Finally, the score of the new set is computed.

ad(iv) The score of the new set is compared with the score of the best existing set. If the new one is better, it is considered the best set further on. The best set found in iteration n will be denoted Best_n.

ad(v) The algorithm determines whether a sufficient number of trials has passed without a change in the score. If so, the number of individuals n^* in future sets is increased. If the last iteration has produced a feasible solution, the algorithm performs a few additional iterations to find solutions with higher genetic response.

Results

The model was successful in finding the declared set of offspring with sufficient accuracy in both scenarios. Collective results of all runs for $n_S = 30$ are depicted in Figs. 2 and 3. Every set of parallel lines corresponds to one desired effective population size N_e. The x-axis represents the genotyping effort (number of offspring n entering the lab), the y-axis shows collective statistics computed from the best sets of size n found in each run r [$\text{Best}_n(r, N_e)$].

Medians, 25% and 75% quantiles of the estimated probabilities $\hat{p}$ computed over all runs R with the same parameter N_e (scenario $n_S = 30$) are provided in Fig. 2. The bolded lines denote

$$\text{median}(\hat{p}_{\text{Best}_n(r,N_e)}) \quad \text{for } N_e \in \left[2, \frac{n_S}{2}\right], r \in R, \tag{8}$$

while the remaining lines denote 25% and 75% quantiles.

The same statistics for the average additive effect $a_{\text{Best}_n(r,N_e)}$ are plotted in Fig. 3.

In both figures, we can study the progress towards the desired values with increasing sample size n. Both graphs show a pattern of sigmoid curves.

As the prescribed N_e increases to $n_S/2$, high P value becomes more important and the resulting ratio n/N_e grows disproportionately (Fig. 2). Therefore,

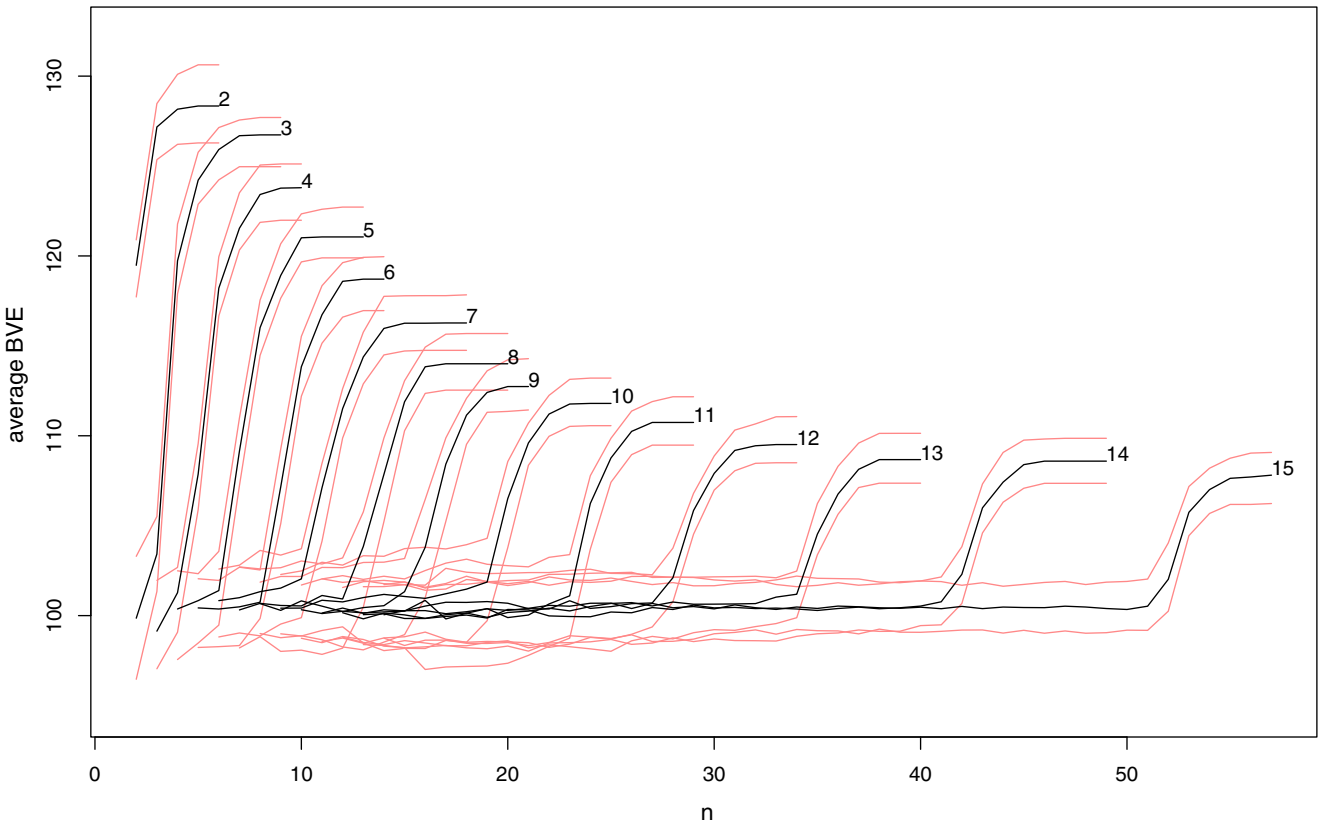

Fig. 3 Average additive genetic value (genetic response, y-axis) plotted against n, x-axis. Different curves correspond to different prescribed N_e. *Bolded curves* depict medians of 100 iterations; *remaining curves* are 25% and 75% quantiles

Table 1 Resultant census size (genotyping effort, n) is provided at each declared N_e value

Results of individual iterations were compared with true values (number of unrelated individuals). *Example:* Let N_e be two. In 90 (out of 100) iterations, the algorithm returned a group of two unrelated individuals, where in 89 iterations the census size was equal to two, and in the remaining one iteration, it corresponded to three. No success in declaring two unrelated individuals was reported in ten iterations (these values are given in parentheses)

N_e	n	truly matched:														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	2	(10)	89													
	3		1													
3	3		(1)	11												
	4		(8)	55	25											
4	5			(11)	48	40										
	6					1										
5	6						3									
	7			(1)	(15)	34	39	8								
6	8					(2)	3	5								
	9				(1)	(11)	31	31	15	1						
7	10						(1)	2	2							
	11						(9)	40	33	13						
8	13							(6)	27	38	14	3				
	14							(1)	2	6	3					
9	15									1	1	1				
	16								(5)	23	34	24	4		1	
	17									2	3	1				
10	19								(2)	(2)	17	18	11	3		
	20								(2)	(3)	21	13	8			
11	23									(1)	(5)	25	20	10		
	24									(1)	(5)	9	16	8		
12	27												1			
	28										(1)	(11)	34	32	11	1
	29											(2)	1	6		
13	33												(1)	1		1
	34											(2)	(7)	34	26	3
	35											(1)	(1)	9	13	1
14	41													(1)	4	3
	42													(11)	51	23
	43												(1)	(1)	4	1

much higher genotyping effort is required to achieve the desired N_e with a reasonable probability.

A similar trend could be observed in Fig. 3. As expected, the additive effect decreases as the sample size increases due to the inclusion of individuals to satisfy the desired N_e. Maximization of the genetic response is performed only when a suitable probability level has been attained. It should be emphasized that higher N_e leads in general to a serious reduction in expected gain, which is something the breeder should carefully consider.

Final estimates are compared with ground-truth evaluation of the resultant sets at the same scenario ($n_S = 30$, Table 1). An agreement to the true values was reported in 89 iterations out of 100, which corresponds to the desired probability value ($P = 90\%$). It is expected that a higher number of runs would converge to 90% success. As indicated here, given a sufficient number of runs, the algorithm is successful in estimating the mean n. At higher N_e values, the solution was nearby the upper limit $n_S/2$. In many applications, a higher number of parents is found and the Table could be easily expanded to higher N_e values.

The effect of increasing the population size to 60 parents is presented in Table 2. While in lower N_e values, the effect is only negligible, with larger (and more

Table 2 Comparison of the genotyping effort in scenarios with 60 and 30 parents under constant prescribed $P = 90\%$

	N_e												
	2	3	4	5	6	7	8	9	10	11	12	13	14
n_{60}/n_{30}	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.86	0.77	0.70	0.65	0.60	0.56

Example: At $N_e = 10$, the genotyping effort in the scenario with 60 parents is 77% of the scenario with 30 parents

realistic) N_e values, the genotyping effort is reduced substantially.

Discussion

Pedigree reconstruction is a fundamental component to the application of BWB for two essential reasons: (1) changing the original pedigree matrix from being entirely HS to a mix of HS and FS (this is needed for the attainment of greater individual-tree breeding value accuracy (Lynch and Walsh 1998)) and (2) coancestry control during the selection phase of production populations. Thus, the amount of experimental effort (i.e., number of DNA fingerprinted individuals) will greatly affect the cost and feasibility of BWB implementation. The development of mathematical algorithms that effectively reduce the genotyping efforts will be beneficial and directly support the application of BWB.

In the present study, while we were primarily interested in presenting the model's applicability, we used simple scenarios to investigate the interplay among N_e, n, and P. Our model can be used as a tool to estimate the minimum fingerprinting effort with respect to the declared N_e. As indicated above, final fingerprinting effort is a function of additional variables, such as the power of the pedigree reconstruction protocol, etc. Therefore, efforts should be taken with the selection of loci used (high polymorphism information content value (Botstein et al. 1980)), minimal genotyping error and the use of appropriate pedigree reconstruction models. It should be stated also that the level of gene flow to the breeding arboreta is expected to increase the genotyping efforts and this increase is proportionate to pollen contamination (Friedman and Adams 1985; El-Kassaby et al. 1989). However, if the realized selection differential between the breeding arboreta and outside pollen source is far from zero, it is expected that most of the extreme observations (top-ranking individuals) within a family are sired by selected parents (those with high general combining ability). Such parents are easily identified during the pedigree reconstruction. At the same time, most contaminating parents' offspring would be ranked low within each family and their respective contribution to the selected set is therefore negligible.

An interesting result from this work is the observed substantial reduction in genotyping effort when larger parental population was simulated. With larger number of male parents, individual's gametic contribution is reduced, resulting in lower likelihood of sharing an identical male parent. Thus, further reduction in the genotyping effort is expected, even though this would likely result in substantial reduction in genetic gain due to lower selection differential of the parental population.

An additional factor that might limit the number of male parents captured from pedigree reconstruction is the male fertility variation commonly observed in tree breeding arboreta and seed orchards (Schoen and Stewart 1986, 1987). If male gametic contributions should deviate from H-W expectations, conditional probabilities should be calculated correspondingly, assuming that an estimate of actual gametic contribution is available (see Funda et al. 2011). In our current model, such estimates of male gametic contributions could be used as an input. There are no restrictions on the distribution of such contributions.

Conclusions

In this study, we presented a novel probabilistic model, based on graph theory, to minimize the fingerprinting effort needed for pedigree reconstruction for assembling half- and full-sib families from open-pollination in breeding arboreta or seed orchards. This model will reduce cost associated with the implementation of the "breeding without breeding" strategies. Computer program is available from the corresponding author upon request. The work also revealed a counter-intuitive feature associated with reduced fingerprinting efforts needed when larger parental populations were used.

Acknowledgements We are grateful to Rowland Burdon and two anonymous reviewers for their critical review and many helpful comments on this article. The access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is highly acknowledged. Support from the Czech Science Foundation (GAČR; grant 521/07/P337; M. Lstibůrek) and the National Agency for Agricultural Research (NAZV; grant QH81172; M. Lstibůrek) and (NAZV; grant QH81160; Jaroslav Kobliha) and the Natural Sciences and Engineering Research Council of Canada (Discovery and IRC Grants) and the Johnson's Family Forest Biotechnology Endowment to Y. A. El-Kassaby are highly appreciated.

Appendix

ad(iii) Probabilities are estimated by the Monte Carlo method:

Algorithm M

Let the number of all outcomes be denoted as A and that of all successful outcomes as S.

(i_M) Set $A \leftarrow 0$, $S \leftarrow 0$.

(ii_M) $A \leftarrow A + 1$
Assign randomly male parents to offspring, so that i^{th} offspring is sired by a male parent y with a probability $p_{i,y}$.

(iii_M) Find the biggest subset of offspring M with all parents different. {algorithm MinRedl2}
If $|M| \geqq N$, increase $S \leftarrow S + 1$.

(iv_M) If $A \geqq o_{\max}$ or
if P is outside the confidence interval logit(S, A, α), {we reject $H_0 : p = P$}
return result $\hat{p} = S/A$,
else go to (ii_M).

ad(iii_M) The problem of finding the largest subset of offspring with different parents was converted to the problem of finding a maximum matching in a general graph, a well-developed subject in graph theory (Agnarsson and Greenlaw 2008).
As a graph, we define a pair (V, E) containing a set of vertices V and edges E connecting pairs of vertices. In our settings, the vertices are declared as parents while the edges are respective offspring. A pairing M is a subset of E where no two edges share common vertex. Vertices sharing an edge with a vertex V will be denoted as its neighbors.
For finding maximum matchings in graphs, we utilized the fast approximation algorithm MinRed12 (MR) by Magun (1998).

(i_{MR}) $|M| \leftarrow 0$

(ii_{MR}) Remove all vertices without edges and all duplicate edges;
consider vertex v with the lowest number of neighbors Δ_v.

Note: This vertex is always removed. Along with the vertex, potential neighbor is removed as well (this is determined by the number of neighbors):

(iii_{MR}) If $\Delta_v = 1$, remove v and its neighbor w,
if $\Delta_v = 2$, remove v and merge its neighbors w_1 and w_2,
else remove w with the lowest number of neighbors Δ_w.

Note: While the first case ($\Delta_v = 1$) is obvious, the second case ($\Delta_v = 2$) is questionable (which one should be removed). As one vertex must always be removed, the two are simply merged. We are not interested at the exact form of such a pairing. It is sufficient to declare the number of paired vertices. If $\Delta_v > 2$, a heuristic approach is used (this is the only approximation step in the algorithm). Let us choose a particular neighbor w out of all potential neighbors v so that it has the fewest number of neighbors. This one is removed. The general idea is that vertices with the lowest number of neighbors interfere the least to the pairing among other vertices. Details are provided in Magun (1998).

(iv_{MR}) $|M| \leftarrow |M| + 1$ {Add edge vw to pairing}
If $|M| \geqq N$, finish.
If there are any more edges in the graph, go to (ii_{MR}).

Note: We are testing a null hypothesis that $\hat{p} = P$ against the alternative that $\hat{p} \neq P$. If H_0 is rejected, we can stop testing the current set. As a result, an outcome of this algorithm must always satisfy this testing, irrespective of its exact score value.
For hypothesis testing, we will use the logit interval, based on the approximation of binomial distribution of logit functions. Such a function is used in the next step with the exception that the exact binomial interval is used for $X = 0, 1, Z - 1, Z$. If $\hat{p}$ is inside this interval, we are not rejecting the hypothesis.

ad(iv_M) To test, whether the probability is estimated accurately, an interval logit(X, Z, α) is used in the algorithm, where X = number of successful tests, Z = number of all tests, and α is the probability of the type I error. The value $c = -0.5$ is taken from Edwardes (1998).
Set

$$x = X - c, \quad z = Z - 2c,$$

$$\varphi = \exp\left(\Phi(1-\alpha/2)\sqrt{\frac{z}{x(z-x)}}\right),$$

where Φ is the cumulative distribution function for the normal distribution. Then

$$\text{logit}(X, Z, \alpha) = \begin{cases} \left\langle 0, 1 - \sqrt[z]{\alpha/2} \right\rangle & \text{if } X = 0 \\ \left\langle 1 - \sqrt[z]{1-\alpha}, \dfrac{x}{x + (z-x)/\varphi} \right\rangle & \text{if } X = 1 \\ \left\langle \dfrac{x}{x + (z-x)\varphi}, \sqrt[z]{1-\alpha} \right\rangle & \text{if } X = Z-1 \\ \left\langle \sqrt[z]{\alpha/2}, 1 \right\rangle & \text{if } X = Z \\ \left\langle \dfrac{x}{x + (z-x)\varphi}, \dfrac{x}{x + (z-x)/\varphi} \right\rangle & \text{else.} \end{cases}$$

References

Agnarsson G, Greenlaw R (2008) Graph theory: modeling, applications, and algorithms. Prentice-Hall, Englewood Cliffs

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

Edwardes MD deB (1998) The evaluation of confidence sets with application to binomial intervals. Stat Sinica 8:393–409

El-Kassaby YA, Rudin D, Yazdani R (1989) Levels of outcrossing and contamination in two Scots pine seed orchards. Scand J Forest Res 4:41–49

El-Kassaby YA, Lstibůrek M (2009) Breeding without breeding. Genet Res 91:111–120

Friedman ST, Adams WT (1985) Estimation of gene flow into seed orchards of loblolly pine (*Pinus taeda* L.). Theor Appl Genet 69:609–615

Funda T, Liewlaksaneeyanawin C, Fundova I, Lai BSK, Walsh C, Niejenhuis AV, Cook C, Graham H, Woods J, El-Kassaby YA (2011) Congruence between parental reproductive investment and success determined by DNA-based pedigree reconstruction in conifer seed orchards. Can J For Res 41:380–389

Grattapaglia D, Ribeiro VJ, Resende GDSP (2004) Retrospective selection of elite parent trees using paternity testing with microsatellite markers: an alternative short term breeding tactic for *Eucalyptus*. Theor Appl Genet 109:192–199

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. Mol Ecol 12:2511–2523

Lambeth C, Lee B-C, O'Malley D, Wheeler N (2001) Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. Theor Appl Genet 103:930–943

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland

Magun J (1998) Greedy matching algorithms: an experimental study. ACM J Exp Algorithmics 3:22–28

Massah N, Wang J, Russell JH, Niejenhuis AV, El-Kassaby YA (2009) Genealogical relationship among members of selection and production populations of Yellow Cedar (*Callitropsis nootkatensis* [D.Don] Oerst.) in the absence of parental information. J Hered 101:154–163

Schoen DJ, Stewart SC (1986) Variation in male reproductive investment and male reproductive success in white spruce. Evolution 40:1109–1120

Schoen DJ, Stewart SC (1987) Variation in male fertilities and pairwise mating probabilities in *Picea glauca*. Genetics 116:141–152