Department of Data Analysis and Mathematical Modelling

Exploring and exploiting genetic variation for balancing short- and long-term genetic gains in plant breeding

ir. David M. Vanavermaete

Thesis submitted in fulfillment of the requirements for the degree of
Doctor (Ph.D.) of Bioscience Engineering: Mathematical Modelling

Academic year 2021-2022

Supervisors:	Prof. dr. Bernard De Baets Department of Data Analysis and Mathematical Modelling, Ghent University, Belgium
	Prof. dr. ir. Jan Fostier Department of Information Technology, Ghent University - imec, Belgium
	Prof. dr. ing. Steven Maenhout Department of Plants and Crops Ghent University, Belgium
Examination committee:	Prof. dr. ir. Geert Haesaert (Chairman)
	Prof. dr. ir. Dirk Reheul dr. Gregor Gorjanc Prof. dr. Stijn Luca Prof. dr. Isabel Roldán-Ruiz
Dean:	Prof. dr. ir. Marc Van Meirvenne
Rector:	Prof. dr. ir. Rik Van de Walle

David Marcel Vanavermaete

Exploring and exploiting genetic variation for balancing short- and long-term genetic gains in plant breeding

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Bioscience Engineering: Mathematical Modelling

Academic year 2021-2022

Dutch translation of the title:
Nederlandse titel: Exploratie en exploitatie van de genetische variatie voor het balanceren van korte en lange termijn opbrengst in plantenveredeling

Please refer to this work as follows:
D. Vanavermaete (2021). Exploring and exploiting genetic variation for balancing short- and long-term genetic gains in plant breeding, PhD Thesis, Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium.

Dankwoord

Het dankwoord is altijd een speciaal hoofdstuk. Het zijn de eerste woorden die u als lezer leest, terwijl dit voor mij de laatste woorden zijn die ik neerpen. Ik wil van deze woorden graag gebruik maken om iedereen te bedanken die mij heeft gesteund tijdens mijn doctoraat en heeft meegeholpen tot de kwaliteit van dit werk.

Vooreerst wil ik graag mijn promotoren Bernard, Jan en Steven bedanken om mij de opportuniteit te bieden aan dit doctoraat te werken en mij doorheen het hele traject te begeleiden. Tijdens de vele meetings hebben hun suggesties en feedback geleid tot de huidige kwaliteit van dit werk. Vooral tijdens het schrijven van de eerste paper, kon ik op hun expertise rekenen om mijn Engelse schrijfvaardigheid aan te scherpen.

Ik wil ook graag de vakgroep bedanken (IDLab en Kermit), Doctoral school en het administratief personeel voor de technische ondersteuning, en het organiseren van verschillende cursussen en activiteiten doorheen het jaar. "Den bureau" zou nooit geweest zijn wat het was zonder de vele collega's waar ik zoveel uren mee gespendeerd heb. De vele babbels en lunchpauzes zorgden voor een deugddoende ontspanning tijdens het werken.

Gretel, je hebt mij ondertussen gekend in goede en slechtere dagen. Bedankt dat je mij altijd bent blijven steunen. Natuurlijk kan ik ook mijn viervoeter Aico niet vergeten die steeds klaarstond als ik thuis kwam om mij te herinneren aan zijn wandeling. Ik kon ook altijd terecht bij vrienden wanneer ik even nood had aan ontspanning. Jeroen, we kennen elkaar al lang en nog steeds kijk ik uit naar onze wekelijkse game avonden. Benji, bedankt voor de vele bezoekjes aan Gent en

de lange conversaties; Sebastian, Michiel en Wouter bedankt voor de vele DnD sessies waar ik altijd naar uitkeek.

Uiteraard zou ik hier vandaag niet staan zonder de steun en hulp van mijn ouders. Mama en papa, bedankt om mij te blijven steunen en mij de opportuniteiten aan te bieden om mezelf te blijven ontwikkelen.

Tijdens mijn vierde en laatste jaar heb ik ook de kans gehad een thesisstudent te mogen begeleiden. Ik ben Jonas heel dankbaar voor zijn sterke motivatie en enthousiasme. Samen hebben we verschillende nieuwe ideeën kunnen uitwerken binnen mijn onderzoeksveld.

Het is altijd moeilijk om niemand te vergeten, dus wil ik ook iedereen bedanken die ik niet specifiek genoemd heb, maar altijd op kon rekenen.

David Vanavermaete
Gent, Augustus 2021

Contents

Preface **v**

Contents **vii**

Summary **xiii**

Nederlandstalige samenvatting **xv**

List of acronyms **xviii**

List of Figures **xx**

List of Tables **xxxi**

I Introduction and background 1

1 Introduction **3**
1.1 General overview 3
1.2 Research questions 5
1.3 Roadmap to this dissertation 6
1.3.1 Part 1 6
1.3.2 Part 2 8
1.3.3 Part 3 9

2 Biological background **11**
2.1 Historical background 11
2.1.1 The Mendelian era 11
2.1.2 The green revolution 12
2.2 Genetic background 14
2.2.1 Quantitative genetics 14
2.2.2 Dominance effect 17
2.2.3 Sexual reproduction 20
2.2.4 Doubled haploid 20

2.3 Concept of inbreeding 21
2.4 Breeding scheme 24
2.4.1 Pure-line breeding 24
2.4.2 Hybrid breeding 25
2.4.3 Recurrent breeding 27
2.4.4 Pedigree method 28
2.5 Genetic markers 28
2.5.1 The introduction of genetic markers in plant breeding 28
2.5.2 Genetic markers and their applicability 30
2.5.3 Genotyping techniques 31
2.6 Phenotyping 31
2.7 Marker-assisted selection 32
2.8 Genomic selection 33

3 Mathematical background 35
3.1 The infinitesimal model 35
3.2 Linear mixed effects model 37
3.3 The genomic best linear unbiased predictor 38
3.4 Gibbs sampling 44
3.5 Bayesian models 45
3.5.1 Bayes A model 45
3.5.2 Bayes B model 45
3.5.3 Bayes C model 46
3.6 Prediction accuracy 46

4 Simulation background 47
4.1 Introduction 47
4.2 Experimental design 48
4.3 Genome construction 50
4.4 Construction of the breeding population 51
4.5 Simulation of a breeding cycle 52
4.6 Prediction model and training population 54

II Methods to preserve the genetic variation in breeding programs 57

5 The scoping method 59
5.1 Introduction 60
5.2 Materials and methods 62
5.2.1 Breeding scheme 63
5.2.2 The baseline method 64

5.2.3 The backcrossing method . . . 64
5.2.4 The scoping method . . . 65
5.2.5 The combined method . . . 66
5.2.6 The population merit method . . . 66
5.2.7 The maximum variance total method . . . 66
5.3 Results . . . 67
5.3.1 The baseline method . . . 67
5.3.2 The backcrossing method . . . 68
5.3.3 The scoping method . . . 69
5.3.4 The combined method . . . 71
5.3.5 The population merit method . . . 72
5.3.6 The maximum variance total method . . . 73
5.3.7 Robustness of the scoping method . . . 74
5.4 Discussion . . . 74
5.4.1 Risks of truncation selection . . . 74
5.4.2 Preserving genetic variation for long-term benefits . . . 77
5.4.3 Reintroducing genetic material in the breeding population . 79
5.4.4 Comparison of the scoping method with existing methods . 80
5.4.5 Limitations of the Pearson correlation . . . 82
5.4.6 Prediction model . . . 83
5.4.7 Updating the training population . . . 85
5.4.8 Cost analysis . . . 86
5.5 Conclusion . . . 87

6 The adaptive scoping method 89
6.1 Introduction . . . 90
6.2 Materials and methods . . . 91
6.2.1 Breeding scheme . . . 92
6.2.2 The scoping method . . . 93
6.2.3 The adaptive scoping method . . . 94
6.2.4 Prediction model . . . 94
6.2.5 Simulation of the population . . . 95
6.3 Results . . . 96
6.3.1 Performance of the adaptive scoping method . . . 96
6.3.2 Robustness of the adaptive scoping method . . . 97
6.4 Discussion . . . 101
6.4.1 The effect of a variable scoping rate on the genetic gain . . 101
6.4.2 Optimizing the breeding population within a predefined time frame . . . 101
6.4.3 Comparison of the scoping and adaptive scoping methods . 102
6.5 Conclusion . . . 103

7 The deep scoping method 105

7.1 Introduction . . . 106
7.2 Materials and methods . . . 109
7.2.1 Breeding scheme . . . 109
7.2.2 Truncation selection . . . 111
7.2.3 Haploid estimated breeding values . . . 111
7.2.4 The deep scoping method . . . 112
7.2.5 The HUC method with bridging . . . 116
7.2.6 Prediction model . . . 117
7.2.7 Simulation of the population . . . 118
7.3 Results . . . 119
7.3.1 Truncation selection . . . 119
7.3.2 The deep scoping method . . . 119
7.3.3 The HUC method with bridging . . . 121
7.3.4 Robustness of the deep scoping method . . . 123
7.4 Discussion . . . 123
7.4.1 Introducing and preserving the genetic variation in the breeding population . . . 123
7.4.2 Population size and required resources . . . 125
7.4.3 GEBVs versus HEBVs . . . 125
7.4.4 Comparison of the HUC method and the deep scoping method 127
7.4.5 Flow from the pre-breeding population into the elite population . . . 129
7.4.6 Applying the deep scoping method . . . 131
7.4.7 Designing the elite population . . . 133
7.4.8 The size of the gene bank . . . 133
7.4.9 Updating the training population . . . 135
7.4.10 Influence of the prediction model . . . 136
7.4.11 Cost analysis . . . 139
7.5 Conclusion . . . 142

8 Oracle methods 143
8.1 Introduction . . . 143
8.2 Material and methods . . . 144
8.2.1 The true selection method . . . 144
8.2.2 Selection of an optimal training population . . . 145
8.2.3 The forward selection method . . . 146
8.2.4 The backward selection method . . . 146
8.2.5 The stepwise selection method . . . 147
8.2.6 Prediction model . . . 147
8.3 Results . . . 148
8.3.1 The true selection method . . . 148
8.3.2 Optimizing the training population . . . 149
8.4 Discussion . . . 150

8.4.1 The greedy selection of QTL alleles 150
8.4.2 Reaching the theoretical maximum genetic value 151
8.4.3 Genetic values, phenotypic values, and genomic estimated breeding values 152
8.4.4 The size of the training population 152
8.4.5 The genetic relationship between the training population and the breeding population 153
8.4.6 Studying the optimal training population 155
8.5 Conclusion 157

III Conclusions and perspectives 159

9 Discussion and future work 161
9.1 Genomic estimated breeding values and their limitations 161
9.2 Capacity required to preserve the genetic variation 163
9.3 An alternative crossing block design 165
9.4 Using a high marker density in the scoping method 165
9.5 Preserving the genetic variation in the long term 166
9.6 Balancing the short- and long-term genetic gain 167
9.7 Combining truncation selection and the scoping method 168
9.8 Pearson correlation 169
9.9 Limitations of the simulation study 170
9.10 Future work 171

10 Conclusion 173

Bibliography 175

Appendix 189

Curriculum vitae 201

Summary

Since prehistory, when man started to settle, and shifted from a hunter-gatherer to a settled-agricultural lifestyle, plants have played a crucial role in the development and survival of humankind. Over time, plants have been cultivated and selected to improve favorable characteristics. They are usually used as a source of food, feed, construction material (e.g. fibers), aesthetics (e.g. flowers), fuel, or other non-food products (e.g. latex, cotton, etc.). Through plant breeding, the yield of most crops has improved over time, but to meet the required food supplies to support the expected 9.5 billion people in 2050, the yield of these crops should increase up to 70%. Unfortunately, the current genetic progress may not be sufficient. The climate change makes it even more challenging to reach that goal, increasing the pressure on food security. Truncation selection is often used to rapidly increase genetic gain, but this is often associated with the loss in genetic variation. It causes the loss of several favorable QTL alleles from the breeding population, reducing the long-term genetic gain. Therefore, truncation selection can only deliver a temporary solution to increase the genetic gain. To meet the needs of tomorrow, new selection methods are needed to increase the genetic gain in the short as well as in the long term.

With the help of a simulation study, we demonstrated the dangers and limitations of using a greedy parental selection method. The loss of favorable QTL alleles is problematic as it reduces the maximum reachable genetic value of the breeding population. Because breeders are often focused on maximizing the short-term genetic gain, the dangers of greedy selection are often overlooked. To reach the required long-term genetic gains, we propose new parental selection methods that avoid the loss of favorable QTL alleles by preserving the genetic variation of the

breeding population. Each method is simulated in a similar way, allowing for a fair comparison between the different parental selection methods under the same circumstances.

Both the scoping method and the deep scoping method were able to maximize the long-term genetic gain. The scoping method was able to reach up to 15% points higher genetic gains compared to truncation selection, whereas the deep scoping method was even able to reach up to 19% points higher genetic gains compared to truncation selection. Not only did both methods reach the highest long-term genetic gains compared with other existing methods, but the loss in short-term genetic gain was minimized, making both methods a promising alternative for truncation selection.

Nederlandstalige samenvatting

Sinds de prehistorie, toen de mens overstapte van een jager-verzamelaar naar een agrarische levensstijl, hebben planten een belangrijke rol gespeeld in de ontwikkeling van de mens. Doorheen de tijd werden planten gekweekt en geselecteerd om gunstige eigenschappen door te geven aan de volgende generatie. De meeste gecultiveerde gewassen worden gebruikt als voedselbron, constructiemateriaal (bv. vezels), esthetica (bv. bloemen), farma, brandstof, of als ruw materiaal (bv. latex, katoen, etc.). Met behulp van plantenveredeling kon de opbrengst in de meeste gewassen aanzienlijk verhoogd worden, maar dit is voorlopig nog steeds onvoldoende om de voedselvoorziening van de voorspelde 9,5 miljard mensen in 2050 te ondersteunen. Er wordt verwacht dat de opbrengst met meer dan 70% zou moeten toenemen maar helaas is de huidige vooruitgang nog onvoldoende. Door de huidige klimaatsveranderingen wordt het nog uitdagender om het doel van 70% in 2050 te halen waardoor de druk op de voedselvoorraad blijft toenemen. Truncatieselectie wordt vaak gebruikt om snel de genetische opbrengst op korte termijn te maximaliseren, maar leidt ook vaak tot het verlies van genetische variatie. Truncatieselectie is geassocieerd met het verlies van verschillende gunstige QTL allelen uit de veredelingspopulatie, waardoor de genetische opbrengst op lange termijn sterk verlaagd wordt. Hierdoor kunnen methoden zoals truncatieselectie slechts een tijdelijke oplossing bieden waarbij de opbrengst enkel op korte termijn verhoogd wordt. Om een duurzamere oplossing te vinden, zijn er nieuwe selectie methoden nodig die de genetische opbrengst op zowel korte als op lange termijn maximaliseren.

Met een simulatiestudie konden we de gevaren en nadelen van een gulzige selectie aantonen. Het verlies van gunstige QTL allelen is vaak problematisch aangezien dit de maximaal haalbare genetische waarde van de veredelingspopulatie

sterk verlaagt. De meeste veredelaars zijn sterk gefocust op het economische aspect en zijn dus enkel geïnteresseerd in het maximaliseren van de genetische opbrengst op hele korte termijn waardoor de gevaren van een gulzige selectie vaak over het hoofd worden gezien. Daarom stellen wij nieuwe selectiemethoden voor die het verlies van gunstige QTL allelen kan voorkomen door de genetische variatie van de veredelingspopulatie te behouden. Elke methode werd onder dezelfde omstandigheden ontwikkeld en gesimuleerd zodat de verschillende methoden eerlijk met elkaar vergeleken kunnen worden.

Zowel de scopingmethode als de deep scopingmethode resulteerden in een hogere genetische opbrengst vergeleken met truncatieselectie (gulzige selectie). De scopingmethode kon tot 15 %-punt hogere genetische waarden behalen in vergelijking met truncatieselectie, terwijl dit verder opliep naar 19 %-punt voor de deep scopingmethode. Beide methoden slaagden er ook in de genetische waarde op korte termijn hoog te houden, waardoor deze methoden een perfect alternatief vormen voor truncatieselectie.

List of acronyms

AFLP	Amplified fragment length polymorphism
BC	Breeding cycle
bp	Base pair
gBLUP	genomic best linear unbiased predictor
BRR	Bayesian ridge regression
DH	Double haploid
DNA	Deoxyribonucleic acid
EMBV	Expected maximum haploid breeding value
IBD	Identical by descent
GBS	Genotyping by sequencing
GCA	General combining ability
GEBV	Genomic estimated breeding value
GLSE	Generalized least square estimator
GS	Genomic selection
GV	Genomic value
GWAS	Genome-wide association study
HEBV	Haploid estimated breeding value
HYV	High yielding variety

LD	Linkage disequilibrium
MAS	Marker-assisted selection
ML	Maximum likelihood
MME	Mixed model equations
MVT	Maximum variance total method
NGS	Next-generation sequencing
OCS	Optimal cross selection
OHV	Optimal haploid value
OLS	Ordinary least squares
PCR	Polymerase chain reaction
REML	Restricted maximum likelihood
RS	Recurrent selection
rr	Ridge regression
RRS	Reciprocal recurrent selection
SCA	Specific combining ability
SNP	Single nucleotide polymorphism
SR	Scoping rate
SRS	Simple recurrent selection
SSD	Single-seed descent
SSR	simple sequence repeats
TP	Training panel
UC	Usefulness criterion
QTL	Quantitative trait loci

List of Figures

1.1 Roadmap to this dissertation. The arrows indicate the order in which the different chapters should be read. 7

2.1 Overview of important milestones in plant breeding, genetics and quantitative genetics. 13
2.2 Illustration of a DNA string containing two polynucleotide chains that are coiled around each other to form a double helix that is interconnected with the nucleobases adenine, cytosine, guanine, and thymine. The DNA string is further condensed into a chromosome. The chemical structure of the four nucleobases is illustrated. The H-bonds between two nucleobases are illustrated with a red-dotted bond and the hydrogen atom at which the DNA is bonded is marked in magenta. 15
2.3 Left panel: an overview of a Mendelian trait. Right panel: an overview of a quantitative trait. The phenotypic observations of a Mendelian trait can be categorized into different discrete groups. A quantitative trait is often controlled by hundreds of QTLs resulting in unique phenotypes. 16
2.4 Illustration of absolute dominance. Absolute dominance occurs when the expression of the dominant allele (A) suppresses the expression of the recessive allele (a). 18
2.5 Illustration of codominance (top left panel), incomplete dominance (top right panel), overdominance (bottom left panel), and underdominance (bottom right panel). In codominance, the phenotype of the heterozygous individuals is an intermediate blend of the phenotype of both homozygous individuals, whereas in incomplete dominance, the phenotype of the heterozygous individuals closely resembles to one of the homozygous phenotypes. In overdominance, the heterozygous individuals have a superior phenotype, whereas in underdominance, the homozygous individuals have a superior phenotype. 19

2.6 Overview of the meiosis process. The gamete is constructed based on the haploids of the parents. The number of crossing overs is Poisson distributed with λ, the expected rate of occurrence, equal to the length of the chromosome in Morgan. The locations of the crossing overs are uniformly distributed over the chromosome. . . . 21
2.7 Overview of how the ancestral genetic information is passed over different breeding cycles (BC). In this figure, a breeding cycle represents a crossing event between two parents resulting in one offspring. If two individuals contain the same allele that has been copied from a common ancestor, then both alleles will be identical by descent (IBD). 23
2.8 Illustration of pure-line breeding in which two parents are crossed followed by 5 to 10 generations of inbreeding. 25
2.9 Illustration of the hybrid breeding scheme in which individuals from two heterotic groups are crossed to create a heterozygous and superior F1 hybrid. 27
2.10 Illustration of the reciprocal recurrent selection scheme. First, parents are crossed to create F1 hybrids. After two (or three generations) of single-seed descent, the new breeding cycle is obtained from which new parents can be selected. 29

4.1 Mean genetic value and maximum reachable genetic value for different heritabilities (top left), parental selection intensities (PSI) (top right), QTLs (bottom left) and individuals per cross (IPC) (bottom right). 49
4.2 Schematic overview of the simulator based on the UMN and NDSU datasets. Both datasets contain the genotype of each individual and a genetic map with the relative location of each SNP loci in the genome. A genome is constructed by randomly selecting 100 SNP loci as QTL. The remaining 1490 SNPs serve as markers. Based on the genotype of these 100 QTLs, the phenotype of each individual is calculated. From both datasets, 50 individuals with the highest phenotype are selected and randomly crossed with each other. The base population is obtained after two generations of single-seed descent. 51

5.1 Overview of the recurrent selection scheme. First, 50 couples of parents (P_1, P_2) each produce 20 offspring yielding a total of 1000 F1 hybrids. Then, after two generations of single-seed descent, 1000 F3 individuals are obtained. From those F3 individuals, new parental lines are selected. Three different parental selection methods are considered: i) the baseline method selects 100 parents with the highest GEBVs (truncation selection); ii) the scoping method combines the selection of 50 parents (P_1) with the highest GEBVs and 50 parents (P_2) that maximize the genetic variation (see Eq. (5.1)); iii) the backcrossing method selects every tenth breeding cycle the P_2 parents from the base population. After the parental selection, the TP is updated according to the tails method. 64

5.2 Simulation results using the baseline method over 50 breeding cycles. Mean genetic value of the breeding population increases rapidly over the first breeding cycles. The truncation selection, however, causes the loss of several favorable QTL alleles, reducing the maximum reachable genetic value and causing a premature convergence of the genetic value to a local optimum. The top-10 individuals of the population have a higher mean genetic value than the breeding population, but after several breeding cycles, the genetic variation is reduced, closing the gap between the top-10 individuals and the rest of the breeding population. 68

5.3 Simulation results using the backcrossing method over 50 breeding cycles. At every tenth breeding cycle, when backcrossing occurs, the mean genetic value drops suddenly as a result of the (re-)introduction of genetic material from the base population. This also introduces several favorable QTL alleles that might have been eliminated during earlier breeding cycles, causing the maximum reachable genetic value to increase. The drop in the mean genetic value is relatively modest for the top-10 individuals. Only five breeding cycles are required to recover from the backcrossing event and higher genetic values are reached in subsequent cycles. Over 50 breeding cycles, the breeding population attains a higher genetic value compared to the baseline method. 69

5.4 Genetic relationship between the breeding population and the training population and the prediction accuracy of the backcrossing method over 50 breeding cycles. At every tenth breeding cycle, when backcrossing occurs, the genetic relationship between the TP and breeding pool decreases, while the prediction accuracy increases. However, the prediction accuracy will only reach a maximum value over the next breeding cycle. 70

5.5 Simulation results using the scoping method for a scoping rate of 0.1, 0.3 and 0.6, simulated over 50 breeding cycles. Additionally, the results of the baseline method are shown for the sake of comparison. In the top figure, the mean genetic value of the top-10 individuals and the maximum reachable genetic value are shown for different scoping rate values and the baseline method. In the middle figure, the mean genetic value of the breeding population is shown for different scoping rate values and the baseline method. In the bottom figure, the rate of QTL fixation is shown for different scoping rate values and the baseline method. 71
5.6 Simulation results using the scoping method combined with the backcrossing method simulated over 50 breeding cycles. Replacing the nine breeding cycles between every two backcrossings with the scoping method helps to retain the genetic variation in the breeding population. This leads to a better fixation of the favorable QTL alleles and to a higher maximum reachable genetic value. Moreover, higher mean genetic values are reached compared to the backcrossing method. Further increasing the scoping rate (> 0.6) only leads to negligible improvements of the mean genetic value. 72
5.7 Genetic value of the different parental selection methods over 50 breeding cycles. The genetic value in the long term is the lowest when using the baseline method, followed by the maximum value total (MVT) method, population merit method and the scoping method, which delivers the highest genetic values in the long term. 73
5.8 Genetic value of the different parental selection methods for different heritabilities over 50 breeding cycles. The genetic value in the long term is the lowest when using the baseline method, followed by the maximum value total (MVT) method, population merit method and the scoping method, which delivers the highest genetic values in the long term. In other words, each method can maintain its relative position towards the other methods indicating that the heritability does not influence the effectiveness of the different parental selection methods. 75
5.9 Mean genetic value of the top-10 individuals using the baseline, backcrossing, maximum variance total (MVT), population merit, scoping and combined methods simulated over 50 breeding cycles with at the left 50 QTLs and at the right 200 QTLs. Each method can maintain its relative position towards the other methods indicating that the number of QTLs does not influence the effectiveness of the different parental selection methods. 76

5.10 Genetic variance of the different parental selection methods over 50 breeding cycles. The genetic variance drops the fastest when using the baseline method, followed by the maximum variance total (MVT) method, population merit method and the scoping method with a scoping rate of 0.3. Both the backcrossing and combined methods result in a high genetic variation after each backcrossing event, but a drop in the genetic variation is observed in the subsequent breeding cycles. 77
5.11 Prediction accuracy of the different parental selection methods over 50 breeding cycles. The prediction accuracy drops the fastest when using the baseline method, followed by the maximum variance total (MVT) method, population merit method and the scoping method with a scoping rate of 0.3. 78
5.12 Mean genetic value of the top-10 individuals, Mean genetic value of the breeding population and the prediction accuracy (Pearson correlation) are shown for a breeding population using truncation selection and the scoping method simulated over 50 breeding cycles. The Pearson correlation degrades over each breeding cycle while the genetic gain is increased. 82
5.13 Simulation results of the baseline method (top line) and scoping method (bottom line) using a training population in which low-frequency markers are removed (left) or using a training population in which all markers are used (right). Each prediction model results in the same long-term genetic gain. When low-frequency markers are removed from the training population, a lower genetic gain is observed. . . . 84
5.14 Effect of the TP update method on the genetic value using the scoping method over fifty breeding cycles. Only a small difference is observed between the *top*, *bottom*, *tails* and *random* update methods. The *no change* update method, which does not update the TP, rapidly converges to a low mean genetic value and several favorable QTL alleles are eliminated from the breeding population, thus reducing the maximum reachable genetic value. 86

6.1 Overview of the recurrent breeding scheme. First, 50 couples of parents (P_1, P_2) each produce 20 offspring, yielding a total of 1000 F1 hybrids. After two generations of single-seed descent, 1000 F3 individuals are obtained. From those F3 individuals, new parental lines are selected. 92

6.2 Top panel: the mean genetic value of the top-10 individuals and the maximum reachable genetic value for the adaptive scoping method with a value for t of respectively 10, 20, 30, 40 and 50 breeding cycles. Middle panels: the mean genetic values for the different parental selection methods at breeding cycles 15, 25, 35, 45 and 55. Bottom left panel: the SR for the different parental selection methods. Bottom right panel: the genetic variation for the different parental selection methods. 98

6.3 Mean genetic values of the top-10 individuals using the scoping method (SR = 0.3) and the adaptive scoping method for $t = 10$, $t = 20$ and $t = 50$. In the short term, the scoping method yields the highest genetic gains. Over time, $t = 10$ will result in higher genetic gains followed by $t = 20$, and $t = 50$. 99

6.4 Simulation results of the original and adaptive scoping methods (using t = 10, 20 and 50) for a heritability of 0.2 and 0.8 using 100 QTLs (top) and for a heritability of 0.5 using 50 and 200 QTLs (bottom). The impact of both methods on the genetic value and on the maximum reachable genetic value is reported. In each case, shortly after t breeding cycles, the adaptive scoping method results in the highest genetic value throughout a certain number of breeding cycles. 100

6.5 Simulation results of the scoping and adaptive scoping methods. At each breeding cycle, the mean genetic value of the top-10 individuals and the maximum reachable genetic value is depicted for the method and/or value of t that yields the highest genetic value. The adaptive scoping method yields superior genetic values during a short time window that follows breeding cycle t. 103

7.1 Mean genetic value (GV) of the top-10 individuals in the breeding population using the population merit method (left) and the scoping method (right) after first applying 0, 5, or 20 breeding cycles (BC) of truncation selection (black line). When the genetic variation of the breeding population is already reduced by means of truncation selection, both the population merit method and the scoping method result in a lower genetic value. 108

7.2 Overview of the recurrent breeding scheme. First, 50 couples of parents (P_1, P_2) each produce 20 offspring yielding a total of 1000 F1-hybrids. Then, after two generations of single-seed descent, 1000 F3-individuals are obtained. From those F3-individuals, new parental lines are selected. Three different parental selection methods are considered: i) Truncation selection selects 100 parents with the highest GEBVs and crosses them randomly; ii) The deep scoping method introduces new genetic information into the breeding population while maximizing the short- and the long-term genetic gain; iii) The HUC method with bridging introduces new genetic information into the breeding population by means of a bridging population. 110
7.3 The haplotype is split into different haplotype segments containing an equal number of markers. Next, the HEBV is calculated per segment by summing up the marker effects of each segment separately. 113
7.4 Left panel: overview of the deep scoping method. First, the individuals with the highest GEBVs are selected as elite parents. Next, for each layer, individuals from the elite population and the previous layer are selected. For Layer 0 elite individuals are combined with individuals of the gene bank. For each of the subsequent layers, the individuals can mature in the breeding population, increasing their genetic value, while the genetic variation is gradually decreased. Right panel: overview of the HUC method with bridging. First, the individuals with the highest GEBVs are selected as elite parents. Next, individuals of the gene bank with the highest H-scores are selected and crossed with individuals of the elite population containing the highest GEBVs. 117
7.5 Simulation results of truncation selection over 20 breeding cycles. When truncation selection is used, the mean genetic gain of the top-10 individuals increases rapidly. Unfortunately, truncation selection also causes fixation of unfavorable QTL alleles, leading to a decrease in maximum reachable genetic value and causing a premature convergence of the mean genetic value of the top-10 individuals. . . . 120

7.6 Simulation results of the deep scoping method and the HUC method with bridging starting at breeding cycles 5, 10, 15, and 20. Prior to the selection methods, truncation selection is used to reduce the genetic variation of the breeding population (black line). By deploying a gene bank, new QTL alleles are introduced into the breeding population, increasing the maximum reachable genetic value and avoiding a premature convergence of the genetic value. Compared to the HUC method with bridging, the deep scoping method can introduce more QTL alleles into the breeding population leading to a higher maximum reachable genetic value. The deep scoping method reaches a higher mean genetic values of the top-10 individuals in the long term compared to the HUC method with bridging. . 121
7.7 Overview of how individuals flow between the different subpopulations when the deep scoping method is applied on a breeding population after five initial breeding cycles of truncation selection. A color scheme is used to indicate the change in genetic value and genetic variation of the different subpopulations. The black and dark red arrows represent the selection of respectively the first and second parent. From the first two layers, very few individuals are selected into the elite population. After progressing over three to four layers, an average of one to two parents per breeding cycle is accepted into the elite population. 122
7.8 Simulation results of the deep scoping method and the HUC method with bridging for a heritability of 0.2 and 0.8 using 100 QTLs (top) and for a heritability of 0.5 using 50 and 200 QTLs (bottom line). The impact of both methods on the genetic value and on the maximum reachable genetic value is shown after 5 (left) and 20 (right) breeding cycles of truncation selection. In all six cases, the deep scoping method resulted in higher genetic values in the long term compared with the HUC method with bridging. 124
7.9 Genetic values of the scoping method (SR=0.3) using the GEBVs and the HEBVs as a selection criterion. Replacing the GEBVs with HEBVs resulted in a lower genetic value in the short as well as in the long term. 127
7.10 Mean genetic value of the top-10 individuals is compared between the deep scoping method and the HUC method with bridging using a paired sampled t-test, for two scenarios with respectively 5 and 20 initial breeding cycles of truncation selection. The difference in mean genetic value with a 95% confidence interval is reported. The color of each dot indicates whether the difference in genetic value between both methods is significant ($p < 0.05$). 129

7.11 Left: an overview of the mean number of individuals that are selected in the elite population for each subpopulation. Right: the maximum reachable genetic value of the elite population for the first ten breeding cycles of the deep scoping method after 5, 10, 15 and 20 breeding cycles of truncation selection. 131
7.12 Simulation results of the deep scoping method using 1 to 10 different layers for a population size of 500, 1000, and 2000 individuals. Increasing the number of layers often results in higher long-term genetic gains. However, if the number of individuals per layer is too low, the long-term genetic gain is reduced. 132
7.13 Simulation results of the deep scoping method and the scoping method over 50 breeding cycles. When the scoping method is used from the first breeding cycle, an overall higher genetic gain is observed in the short and long term. 134
7.14 Difference in genetic value between the deep scoping method using two different breeding populations. The sooner truncation selection is replaced by the deep scoping method, the sooner the genetic gain of the top-10 individuals will increase until the breeding value converges to the same value in the long term. 134
7.15 Simulation results of truncation selection using different crossing block designs. Three crossing block designs were considered: i) crossing the individuals with the highest GEBVs, ii) crossing the individuals with the highest GEBV with an individual that minimizes the genetic relationship between both parents and iii) random crossing. 135
7.16 Simulation results of truncation selection using the top, bottom, random, tails, CDmean and PEVmean methods to update the training panel. Each update method results in a similar long-term genetic gain with the exception of the top update method that results in a lower long-term genetic gain. 136
7.17 Overview of the prediction accuracy of the elite population, the different layers (Layer 1–4) and the pre-breeding population (Layer 0) using truncation selection. The training panel is updated according to the top, bottom, tails, random, CDmean or PEVmean method. . . 137
7.18 Effect of rr-gBLUP and Bayesian models on the genetic gain of a breeding population using the deep scoping method. A slightly lower, but non significant long-term genetic gain is observed for the rr-gBLUP model. 138
7.19 Simulation results of the deep scoping method using rr-gBLUP and BRR. Top left, the predicted variance component of the residuals errors. Right top, the predicted variance component of the additive marker effects, and left bottom, the prediction accuracy. The model is fitted using the available markers or after first removing the low-frequency marker alleles. 140

7.20 Simulation results of truncation selection using rr-gBLUP and BRR. Top left, the predicted variance component of the residuals errors. Right top, the predicted variance component of the additive marker effects, and left bottom, the prediction accuracy. The model is fitted using the available markers or after first removing the low-frequency marker alleles. 141

8.1 Mean genetic value of the top-10 individuals and maximum reachable genetic value of a breeding population using the true selection method, scoping method (SR = 0.3), deep scoping method (BC05), and truncation selection over 50 breeding cycles. The true selection method leads to a high increase of the mean genetic value over the first breeding cycles while the maximum reachable genetic value remains constant, indicating that no favorable QTL alleles are lost. The difference in genetic value between the true selection method and the other methods indicates that further improvements of the parental selection methods could increase the genetic value up to 14 percentage points. 148

8.2 Simulation results of the forward and stepwise selection methods compared to truncation selection. Both the forward and stepwise selection methods have a higher genetic value in the long term compared to truncation selection using the top, bottom, tails, random, PEVmean or CDmean update methods. 150

8.3 Simulation results using the true selection method over 50 breeding cycles. The true selection method leads to a high increase of the mean genetic value over the first breeding cycles. The maximum reachable genetic value remains constant, indicating that no favorable QTL alleles are lost. Due to selection, the frequency of the favorable QTL alleles increases, finally leading to the loss of unfavorable QTL alleles. This eventually results in high genetic values. . 151

8.4 Mean genetic value of a breeding population using truncation selection. The parents are selected based on the GEBV, GV, or phenotypic values. Selecting parents based on the GV will result in high genetic values, followed by phenotypic values and GEBVs (using the tails method). 153

8.5 Overview of the number of individuals that are added to and removed from the training population using the stepwise selection method. Over the first breeding cycles, a lower number of individuals are removed from the breeding population, allowing for an increase in the size of the training population. 154

8.6 Mean genetic relationship between individuals of the training population and the individuals of the breeding population. The top and tails update methods result in a training population with a high mean genetic relationship, whereas the forward selection, stepwise selection, and bottom update methods result in a lower mean genetic relationship. 155
8.7 Top left, the mean genetic value of the training population, top right, the mean genetic value of the breeding population, bottom left, the mean marker variance, and bottom right the absolute residual error. Both the forward and stepwise selection methods select individuals that minimize the residual error, preserves the genetic variation in the training population, while preserving the same mean genetic value as observed in the breeding population. 156

9.1 Simulation overview of truncation selection and the scoping method for different parental selection intensities (PSI). 164
9.2 Mean genetic values of the top-10 individuals and maximum reachable genetic values for the scoping method and the adaptive scoping method using an alternative crossing block design (Maxvar Scoping). 166
9.3 Simulation results of the deep scoping method and the scoping method starting after 1 (left), and 3 (right) breeding cycles of truncation selection. The scoping method can gain the highest long-term genetic values. However, if truncation selection is used prior to the scoping method, the long-term genetic gain decreases, and the deep scoping method can gain higher genetic gains. 168
9.4 Mean genetic value of the top-10 individuals and Pearson correlation of a breeding population using the scoping and deep scoping methods. Both methods are used to select the parents over 50 breeding cycles. 171

List of Tables

2.1 An estimation of the narrow-sense heritability (h^2) of different traits for winter wheat (Teich, 1984), safflower (Zhao et al., 2020), and barley (Bargougui, 2016). 17

4.1 Conversion of the haplotype coding to the genotype coding of an individual. 52
4.2 Overview of the parameters used the simulate a breeding population of Barley over 50 breeding cycles. 56

10.1 The mean genetic value and the standard deviation of the top-10 individuals for each parental selection method over 250 experiments. 190
10.2 The mean genetic value and the standard deviation of the breeding population for each parental selection method over 250 experiments. 191
10.3 The maximum reachable genetic value and the standard deviation of the breeding population for each parental selection method over 250 experiments. 192
10.4 The mean genetic value and the standard deviation of the top-10 individuals for the backcrossing and combined methods over 250 experiments. 193
10.5 The maximum reachable genetic value and the standard deviation of the breeding population for the backcrossing and combined methods over 250 experiments. 194
10.6 The mean genetic value and the standard deviation of the top-10 individuals using the scoping and adaptive scoping methods, averaged over 100 runs. 195
10.7 The maximum reachable genetic value and the standard deviation of a breeding population using the scoping and the adaptive scoping methods, averaged over 100 runs. 196
10.8 The mean genetic value and the standard deviation of the top-10 individuals for the deep scoping method and the HUC method with bridging over 100 experiments. 197

10.9 The maximum reachable genetic value, the genetic value of the fixed QTL alleles, the mean genetic value of the top-10 individuals and the standard deviation using the true selection method. The genetic values are averaged over 100 different experiments. 198
10.10 The maximum reachable genetic value and standard deviation of the breeding population using different training panel update methods. The maximum reachable genetic value is averaged over 100 experiments. 199
10.11 The mean genetic value and standard deviation of the top-10 individuals using different training panel update methods. The genetic value is averaged over 100 experiments. 200

PART I

INTRODUCTION AND BACKGROUND

1

Introduction

1.1 General overview

Imagine a breeder that is responsible for the development of new breeding lines. How will (s)he proceed? The easiest strategy dates back to the beginning of agriculture and consists of cultivating the available breeding lines, and selecting superior lines as parents based on visual morphological characteristics (or phenotype). This method is often referred to as truncation selection, has been used for ages, and coincides with the experience and insight of breeders. Despite the high short-term gains that are often associated with truncation selection, this strategy has certain disadvantages and limitations. First of all, to select individuals based on morphological characteristics, each individual needs to be cultivated, which requires time and resources. Second, different (random) effects (e.g. environmental effects) can also influence the morphological characteristics of individuals resulting in a completely different parental selection compared to the same breeding population under different environmental conditions. A breeding line containing individuals with the same genetic constitution (clones) can still lead to a different phenotype due to differences in the environment they were cultivated in. It is,

therefore, possible that a superior individual in an unfavorable environment has a lower phenotypic value and is thus not selected as a parent, reducing the genetic value of the offspring. There is also a human factor in which certain characteristics will be weighted more severely depending on the preferences of the breeder. Each measurement of the phenotype will also be dependent on the time of sampling and the accuracy of the measurement itself.

The phenotypic expression of certain genes results in a specific morphological characteristic. Therefore, the genetic information of an individual could be used to detect the presence of favorable genes, guiding the parental selection. It was only in 1866, that Mendel was able to classify the phenotypic expression of different genes into discrete classes (Mendel, 1866). However, certain traits could not be classified in discrete classes, and a new theory was needed to predict the phenotypic expression of these genes. That theory, coined quantitative genetics, was developed by Fisher, Wright, and Haldane (Fisher, 1918; Wright, 1931; Haldane et al., 1918). In quantitative genetics, traits like grain yield are controlled by hundreds of genes (or quantitative trait loci (QTL)) that all have an additive effect. Moreover, the trait phenotype increases or decreases depending on the allelic composition of each gene.

Over time, breeders started to collect and store the information of parental lines. Especially in animal breeding, pedigree information was used to guide the parental selection. Based on the idea that the quality of parents should be determined based on the genetic values of their offspring, and not their genetic values, superior parental lines could be selected, resulting in major advancements in animal and plant breeding. By combining the phenotypic information with pedigree information, the impact of the environment and measurement errors is reduced. The cultivation of each (promising) breeding line is still required before it is possible to identify and select the next parents.

With the discovery of genetic markers, genotyping became more accessible, making it possible to apply the concept of quantitative genetics in breeding programs. Marker-assisted selection (MAS) uses the genotype of each individual to predict its estimated breeding value (EBV). Because the locations of most QTLs are unknown, a genetic map is generally constructed requiring a large dataset containing genotypic and phenotypic data. Markers that are located close to a gene of interest are selected and used to guide the parental selection. One way to achieve this is by fitting a linear mixed effects model using the genotypic and phenotypic information of a training population. Next, the phenotypic values of the offspring can be predicted without cultivating or phenotyping the whole breeding population, reducing the time and cost of each consecutive breeding cycle. MAS has been used successfully to improve single-major gene resistance (Miedaner and Korzun, 2012), but despite the expectations, it was unable to predict phenotypic traits that were controlled by many small-effect QTLs, reducing its applicability in quantitative genetics. Genomic selection (GS) was proposed as an alternative method, using all

the available markers over the whole genome. Based on the idea that each QTL allele is in linkage disequilibrium (LD) with at least one or more markers, these markers can be used to grasp and predict each QTL effect. In contrast to MAS, GS does not require genetic mapping and only requires training data to estimate the marker effects. To fit the linear mixed effects model, a training panel containing genotypic and phenotypic data is still required. GS predicts the genomic estimated breeding values (GEBVs) based on the genotype. This reduces the need to cultivate and phenotype whole breeding populations. Based on the prediction model, the genetic contribution of each individual is calculated and used to guide the parental selection. As each parent is selected based on its genotype, the impact of the environment on the prediction criterion will be reduced compared to phenotypic selection. The prediction accuracy of the GEBVs will depend upon the quality of the training dataset.

Genomic selection was able to improve both plant and animal breeding. However, it also introduced new risks. Due to the high efficiency of genomic selection, methods like truncation selection that select individuals with the highest GEBVs result in the loss of genetic variation and the fixation of several QTL alleles. Although this may lead to high short-term genetic gains, truncation selection also leads to the loss of favorable QTL alleles in a breeding population, reducing the maximum reachable genetic value. Eventually, this could lead to a premature convergence of the genetic value to a local optimum. To avoid this, new parental selection methods are needed that preserve the genetic variation and lead to a slower but more accurate fixation of the QTL alleles. These methods should aim to maximize the genetic gain in the long term while preserving the short-term genetic gain.

1.2 Research questions

In genomic selection, different design choices have become so mainstream that they are often used without questioning. In most research studies one predicts the GEBV by using a linear mixed effects model, selects superior lines using truncation selection, and evaluates the performance using the Pearson correlation between the true and predicted selection criterion values.

In this dissertation, the advantages, risks, and limitations of these design choices are assessed. The risks associated with truncation selection are studied and discussed. A hypothesis and several research questions are proposed and are answered in the next chapters.

Hypothesis
The long-term genetic gain of a breeding population can be maximized by preserving or reintroducing the genetic variation in the population.

Research questions

1. Which variables can be used to guide the parental selection?
2. How can the performance of the different parental selection methods be compared?
3. Can the long-term genetic gain be increased by preserving the genetic variation?
4. Is the reintroduction of genetic variation required to maximize the long-term genetic gain?
5. Can a balance be found between the short- and long-term genetic gain?
6. Can the genetic values of a breeding population be maximized after a specific number of breeding cycles?

In this thesis, with the help of a simulation study and based on the proposed researched questions, new parental methods have been developed. Each method is compared with other existing methods under similar conditions and is thoroughly discussed in Part 2. In Part 3, the insights using these methods are used to answer the different research questions.

1.3 Roadmap to this dissertation

This dissertation is divided into three parts: introduction and background, preservation of the genetic variation in breeding programs, and conclusions and perspectives. In Part 1, the biological and mathematical background are thoroughly discussed giving the reader the required insights to understand the experimental work that will be discussed in Part 2. Finally, in Part 3, we summarize and highlight different aspects of the research that was presented in Part 2. We recommend to read this dissertation in a linear way, as each chapter is a continuation of the preceding chapter. Only Chapter 6 is not required to understand Chapters 7 and 8. The roadmap is shown in Figure 1.1.

1.3.1 Part 1

The first part consists (aside from this introduction) of three chapters: the biological background, the modelling background, and the simulation background. In

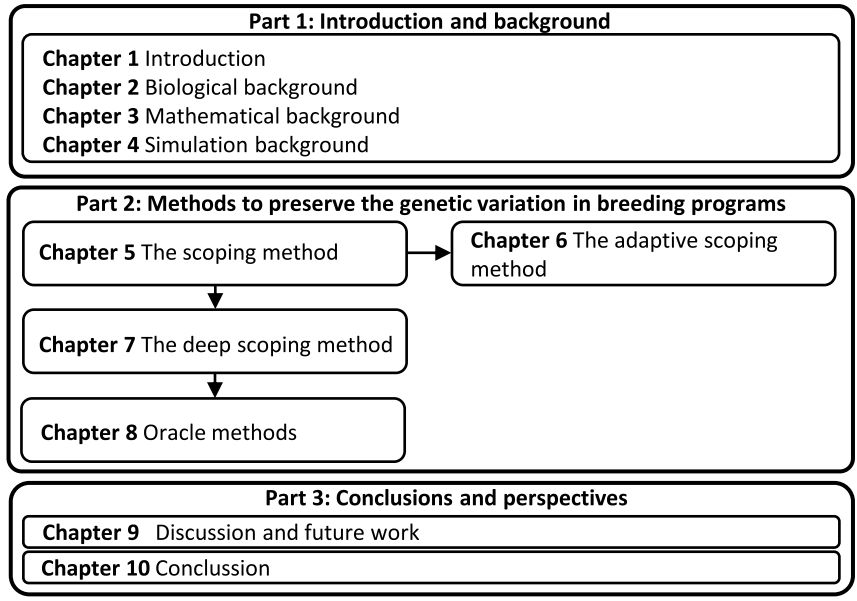

Figure 1.1: Roadmap to this dissertation. The arrows indicate the order in which the different chapters should be read.

each chapter, a theoretical background is given that will allow the reader to navigate and understand the experimental part of this dissertation. Each subject that is discussed in the introduction, has carefully been scaled down to avoid overloading the reader with (unnecessary) information.

Chapter 2 starts with a brief history of plant breeding in which the most relevant milestones are listed. Next, the reader is introduced to quantitative genetics and how the genetic information and the physical appearance of an individual can be linked to each other. Quantitative genetics is the backbone of genomic selection and will play a central role in this dissertation. Next, sexual reproduction on a genetic level will be discussed, explaining how genetic information can be passed on to the next generation. Once the genetic background has been outlined, plant breeding and the different breeding schemes are introduced. Although recurrent breeding is consistently used in this research, other breeding schemes are also discussed. Next, the introduction and application of genetic markers in plant breeding are discussed. Marker-assisted selection was the first strategy to use these markers to guide the parental selection followed by genome-wide association study and genomic selection. Finally, a small overview of the top breeding companies is given.

Chapter 3 will go deeper into the mathematical background of genomic selection. First, the infinitesimal model is discussed, translating the biological background into mathematical equations. Next, the mathematical background of rr-gBLUP and

Bayes A–C models are discussed.

In Chapter 4, the knowledge of the previous two chapters is combined to construct a simulation study that can imitate the behavior of a realistic breeding population. The construction of the simulator, the different design choices, and parameter values will be discussed step by step.

1.3.2 Part 2

In the second part, the author's contributions are listed in four different chapters. In each chapter, a new parental method that preserves the genetic variation and maximizes the long-term genetic value is presented and discussed in detail. The performance of each method is compared with another similar method under different circumstances. Different genome constructions are used to study the robustness of each method against variation in the plant's genome. The different models that were discussed in Part 1 are each used to fit the linear mixed effects model, and different methods to update the training population are considered.

In Chapter 5, the scoping method is proposed as a new parental selection method that combines a two-part selection strategy to preserve the genetic variation while maximizing the genetic progress of the breeding population. This is done by combining two strategies, maximizing the genetic variation of the offspring and selecting parents with a high genetic value. In Chapter 6, the scoping method is further improved by using a variable scoping rate to guide the pre-selection of the parental candidates.

In certain breeding programs, the genetic variation has already decreased due to years of consecutive breeding. Therefore, methods like the scoping method are insufficient, as there is barely any genetic variation left to be preserved. In Chapter 7, the deep scoping method is then proposed as a remedy in which genetic material is reintroduced in the breeding population resulting in a high long-term genetic gain.

In Chapter 8, we propose and discuss different oracle methods to maximize the genetic gain of the breeding population. An oracle method is a theoretical concept in which a value of interest is maximized using the ground truth. Because the oracle method uses data that is normally unavailable in a breeding program, these methods cannot be used in a breeding application but can give a better understanding of how an idealistic selection could influence the long-term genetic gain.

1.3.3 Part 3

In the third and final part, the results obtained in the previous chapters are discussed. Important conclusions and design choices are highlighted and future prospects are discussed. An overall conclusion is given, finalizing a four-year research program.

2

Biological background

2.1 Historical background

2.1.1 The Mendelian era

Plant breeding dates back to the prehistoric period when mankind started to settle, shifting from a hunter-gatherer to a settled-agricultural lifestyle. To understand the principles of plant breeding, we need to go back to the very beginning, more than 10,000 years ago, when the first signs of agriculture and domestication were found in Southwestern Asia near the site of Abu Hureyra (Stuiver et al., 1998). The increasing aridity in the region caused a declining growth of important wild plants, decreasing the available food stock. This led to the cultivation and domestication of different cereals like wheat and rye (Moore et al., 2000; Hillman et al., 2001; Garrard, 1999). Without any specific knowledge of genetics, by selecting plants with favorable properties, unfavorable characteristics like seed shattering and dormancy were eliminated from the population, while favorable traits were

accumulated over each generation. In 3000 BC, most of the major crops in the old world (Africa, Europe, and Asia) were domesticated. This was followed by the new world (America) only 2000 years later. Until the 17th century, crops were selected based on visual observations. It was only in 1694 when Camerarius demonstrated the mechanisms behind the sexual reproduction of plants in his book *De sexu plantarum epistola* that more profound techniques could be developed (Camerarius, 1694). Kölreuter was the first to exploit this knowledge to develop the first hybrid offspring of the tobacco plant (Nicotiana *tabacum*). In the meantime, the first breeding company was established in France by the family of de Vilmorin. Later on, this company developed progeny testing, studying the traits of the offspring to evaluate and guide the selection of the parental lines. Today, de Vilmorin is still an important player in the seed market (Vilmorin, 2021). In the 19th century, Knight, a prominent British horticulturalist and botanist studied the adaptive response of plants and was the first to perform artificial hybridization. Later, Knight also observed dominance, recessive, and segregation properties in peas, but was unable to explain this phenomenon. It was only in 1866 that Mendel published the laws of inheritance, explaining how traits can be inherited (Mendel, 1866). Unfortunately, Mendel's paper was not recognized by the scientific community until De Vries, Correns, and Von Tschermak each rediscovered Mendel's laws of inheritance. Together with the work of Charles Darwin, the foundation of plant breeding was set, starting a new era, the Mendelian era.

Although the works of Darwin and Mendel are fundamental, in the 20th century, breeders were in disagreement whether evolution was caused by continuous variation or by discontinuous variation (Mendel) (Hallauer, 2007). Different genetic studies that were conducted using the principles of Mendel, were not able to classify all the traits in discrete classes. Mendel's theory failed to explain the inheritance of certain traits, opening the search for new theories. Yule was one of the first to recognize that the inheritance of these traits could be explained by a combination of Mendelian laws and the inheritance of quantitative traits (Yule, 1907). After one decade, this theory was developed by R.A. Fisher, S. Wright, and J.B.S. Haldane (Fisher, 1918; Wright, 1931; Haldane et al., 1918), finally recognizing the complementarity between Mendel and Darwin (Hallauer, 2007). Unfortunately, the mathematical tools needed to understand this theory were at that point rather limited. Most plant breeders were already able to improve different crops using pedigree information and progeny testing, and could at that moment not grasp the implications of such a theory on current breeding populations.

2.1.2 The green revolution

The green revolution was marked by different significant advancements in agriculture between 1940 and 1960. In the 1940s, due to an increasing food shortage,

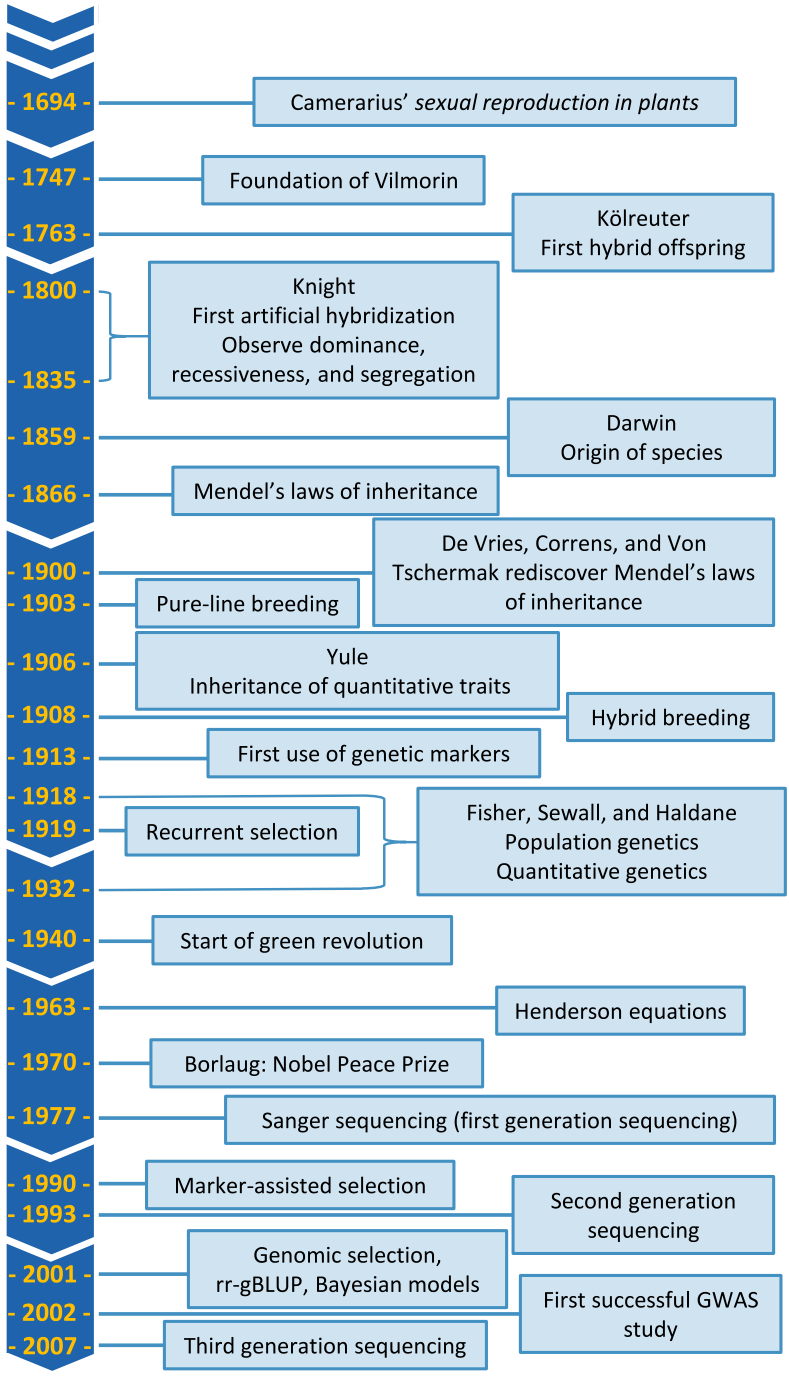

Figure 2.1: Overview of important milestones in plant breeding, genetics and quantitative genetics.

the Rockefeller Foundation released funds to rally researchers around the globe to tackle that problem, starting the green revolution. Soon after, other (inter)national institutions joined. The green revolution was initiated in Mexico and spread to India and Africa. In 1970, the Nobel Peace Prize was awarded to Norman Borlaug, the *father of the green revolution*. During the green revolution, new breeding methodologies were developed resulting in high-yield varieties (HYVs) of different cereals. Those crops were developed using conventional breeding methods, hybrid breeding, and shuttle breeding. The latter consists of cultivating crops at different locations and crossbreeding these varieties to obtain widely adaptable lines. The green revolution also coincides with technological advancements, the use of artificial fertilizers, agrochemicals, and irrigation systems. The rapid advancements in agriculture and the development of HYVs, led to several concerns about the loss of biodiversity. Certain favorable traits that had been selected over the last decades were lost, and the reduced genetic variability could make the HYVs crops susceptible to (a)biotic stress (Kumar, 2007). As a response to this, massive seed banks such as the Svalbard Global Seed Vault were built, preserving the genetic variation of different crop varieties over time.

2.2 Genetic background

2.2.1 Quantitative genetics

In plant breeding, plants are crossed in the hope to pass favorable characteristics from the parents to the next generation, improving a trait of interest which can vary from resistance against pathogens to morphological properties (like strength, flower color, etc.). Although plants are often selected based on morphological properties like yield and length, by understanding how traits are passed to the next generations, a more accurate selection of the parents becomes possible.

Plants are constructed out of different cells containing membrane-bound organelles like a nucleus, endoplasmic reticulum, and mitochondria. The nucleus has a double membrane that encapsulates the nuclear deoxyribonucleic acid (DNA) of that individual. The mitochondria and chloroplasts also contain DNA, but only nuclear DNA will be considered in this dissertation. DNA is composed out of two polynucleotide chains that are coiled around each other to form a double helix that is interconnected with two nucleobases (see Figure 2.2). There are four different nucleobases: adenine (A), cytosine (C), guanine (G), and thymine (T). The nucleobases form a linear connection between the two polynucleotide chains by forming hydrogen bonds. Because of their chemical structure, adenine will only bond with thymine and guanine will only bond with cytosine and vice versa. The sequence in which the base pairs (bp) are present in DNA will determine the genetic

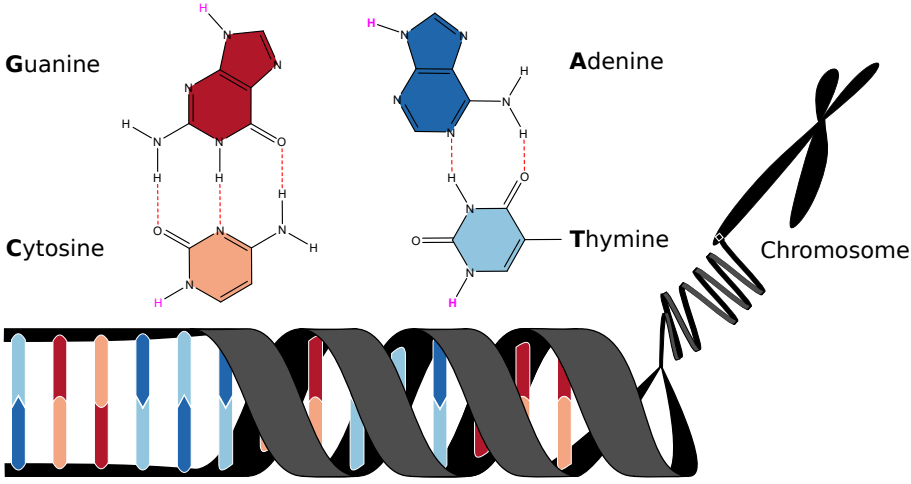

Figure 2.2: Illustration of a DNA string containing two polynucleotide chains that are coiled around each other to form a double helix that is interconnected with the nucleobases adenine, cytosine, guanine, and thymine. The DNA string is further condensed into a chromosome. The chemical structure of the four nucleobases is illustrated. The H-bonds between two nucleobases are illustrated with a red-dotted bond and the hydrogen atom at which the DNA is bonded is marked in magenta.

properties of that individual. To a large extent, there are other properties of DNA such as methylation that also affect the genetic properties of each individual. The sequence of the nucleobases represents the genetic information that after transcription and translation will result in the synthesis of different cell metabolites, affecting the characteristics and functionalities of that individual. The DNA is organised into chromosomes. Each genome is constituted out of one or more chromosomes. Diploid individuals have two copies of each chromosome referred to as homologous chromosomes (Xu, 2010). Several important crops like wheat and potato are polyploid, meaning that their genome contains more than two copies of each chromosome. For example, hexaploid wheat contains six complete versions of each chromosome. The length of the genome of plants can range between 82 Mbp (floating bladderwort) and 149 Gbp (*Paris japonica*), but only a fraction of the genome will result in the transcription and translation of metabolites (Ibarra-Laclette et al., 2013; Pellicer et al., 2018). The regions of DNA that code for the translation of metabolites are called genes. Each gene has different alleles that can influence the phenotype. Homologous chromosomes often contain the same genes, but a different allele can be present at each chromosome. Let us assume a gene with two different alleles: A and a. A diploid plant will have one of the two available alleles on both homologous chromosomes. The individual can be homozygous, meaning that the individual carries the same allele on both homologous chromosomes or heterozygous if a different allele is present on both homologous chromosomes.

According to Mendel's laws, the phenotypic observation of a trait can be classified

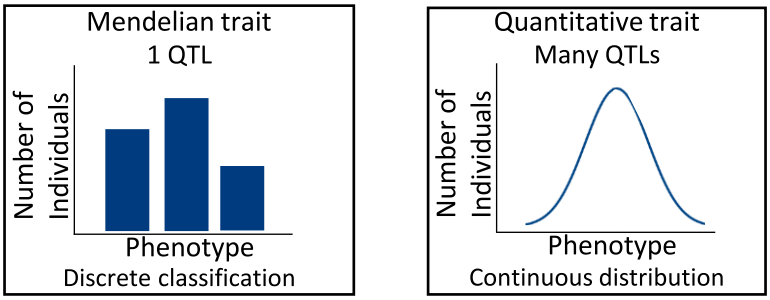

Figure 2.3: Left panel: an overview of a Mendelian trait. Right panel: an overview of a quantitative trait. The phenotypic observations of a Mendelian trait can be categorized into different discrete groups. A quantitative trait is often controlled by hundreds of QTLs resulting in unique phenotypes.

into discrete classes (see left panel of Figure 2.3). However, different traits like grain yield cannot be predicted by using a discrete classification. These traits are controlled by hundreds of genes or quantitative trait loci (QTLs), each having an additive effect on the total grain yield (see right panel of Figure 2.3). Each QTL has a positive or negative effect on the phenotype. The total effect of the different QTLs is known as the genetic value (GV). The breeding value is the total effect of the different QTLs that an individual can pass to its offspring. Because, in this dissertation dominance effects and epistatic effects are assumed absent (see later), the genetic value and the breeding value will be the same. The genetic value will often be different from the phenotypic value. This is because the phenotypic value is also influenced by environmental effects. Two genetically identical plants (clones) can have a different morphology when grown in a different environment. Mathematically, we can express the phenotype of individual i (y_i) as:

$$y_i = g_i + E_i, \tag{2.1}$$

with g_i the genetic value (or genetic effect) of the i-th individual and E_i the effect of the environment on the phenotypic value of the i-th individual. The relationship between the genetic value and the environmental effect is expressed by the heritability. Based on Eq. (2.1), the phenotypic variance (σ_y^2) can be written as:

$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2 + \text{cov}(\mathbf{g}, \mathbf{E}), \tag{2.2}$$

with σ_g^2 the genotypic variance, σ_e^2 the environmental variance and cov($\mathbf{g}$, $\mathbf{E}$) the covariance between the genetic and environmental effects. Assuming that the covariance between the genetic and environmental effects is zero, the broad-sense

heritability (H^2) can be expressed as:

$$H^2 = \frac{\sigma_g^2}{\sigma_y^2}. \tag{2.3}$$

The broad-sense heritability represents the fraction of phenotypic variance that is due to the variation in genetic values. In quantitative genetics, each allele is considered to have an additive effect on the phenotype, therefore, the narrow-sense heritability (h^2) can be used instead:

$$h^2 = \frac{\sigma_u^2}{\sigma_y^2}, \tag{2.4}$$

with σ_u^2 the variance component of the additive values. The genetic variance can also be influenced by other components such as dominance, epistatic, and maternal-paternal effects. The heritability gives an indication of how the genotypic and environmental effects influence a trait. A heritability close to 0 indicates that a trait is mainly controlled by environmental effects (e.g. language spoken) whereas a heritability of almost 1 indicates that a trait is mainly controlled by the genotype (e.g. blood group). In this dissertation, the term heritability will always be used to refer to the narrow-sense heritability. In Table 2.1 an estimation of the heritability for a specific breeding population of winter wheat, barley and safflower is given. Note that the heritability is population and trait specific.

Crop	Trait	Heritability (h^2)
Winter wheat (*Triticum aestivum*)	Grain yield	0.30
	Plant height	0.77
Barley (*Hordeum vulgare*)	Spike number	0.49
	Grain number per spike	0.24
	Thousand kernel weight	0.25
Safflower (*Carthamus tinctorius*)	500 seed weight	0.66
	Grain yield	0.31
	Plant height	0.50

Table 2.1: An estimation of the narrow-sense heritability (h^2) of different traits for winter wheat (Teich, 1984), safflower (Zhao et al., 2020), and barley (Bargougui, 2016).

2.2.2 Dominance effect

In a diploid individual, the effect of the genotype on the phenotype is defined by both alleles and the interaction between them. Five different effects can be distinguished: dominance, codominance, incomplete dominance, overdominance, and underdominance (see Figures 2.4 and 2.5). The dominance effect was demon-

strated by Mendel in which one of the two alleles is dominant (A) whereas the other allele is recessive (a). If the dominant allele is present on at least one of the homologous chromosomes (AA/Aa/aA), only the dominant allele will determine the expression of the phenotype. The recessive phenotype can only be observed when an individual contains both recessive alleles (aa).

In codominance both alleles are expressed. For example, a homozygous phenotype can be expressed as a flower with either white (*AA*) or red (*aa*) petals. The heterozygous phenotype (*Aa*) will be expressed as an intermediate blend of both homozygous phenotypes resulting in pink petals. In incomplete dominance, the dominant allele cannot completely mask the phenotypic expression of the recessive allele. The heterozygous phenotype will closely resemble to one of the homozygous phenotypes. Incomplete dominance can be observed in flowers, in which a homozygous gene will result in e.g. either a red or a white flower whereas a heterozygous gene will result in flower with a lighter shade of red or a darker shade of white.

Overdominance occurs when a heterozygous genotype (Aa) results in a superior phenotype compared to the homozygous genotypes (aa/AA). In other words, the heterozygous genotype has a higher fitness. The human sickle cell anemia is an example of overdominance. Hemoglobin is a four-part molecule built from two α hemoglobin chains and two β hemoglobin chains. Sickle cell anemia is controlled by a recessive gene that interferes in the synthesis of the β hemoglobin chain resulting in malformed red blood cells. Sickle cell anemia is associated with different health problems and reduced life expectancy. However, the abnormal form of the blood cells gives that individual an advantage against malaria, a tropical disease caused by *Plasmodium* that uses the red blood cells to reproduce. A homozygous genotype at that gene will either lead to sickle cell anemia or an increased sensitivity for malaria. A heterozygous genotype, however, will reduce the negative effects of sickle cell anemia and will still offer a partial resistance against malaria.

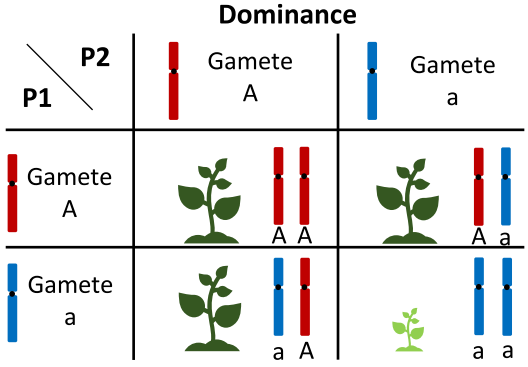

Figure 2.4: Illustration of absolute dominance. Absolute dominance occurs when the expression of the dominant allele (*A*) suppresses the expression of the recessive allele (*a*).

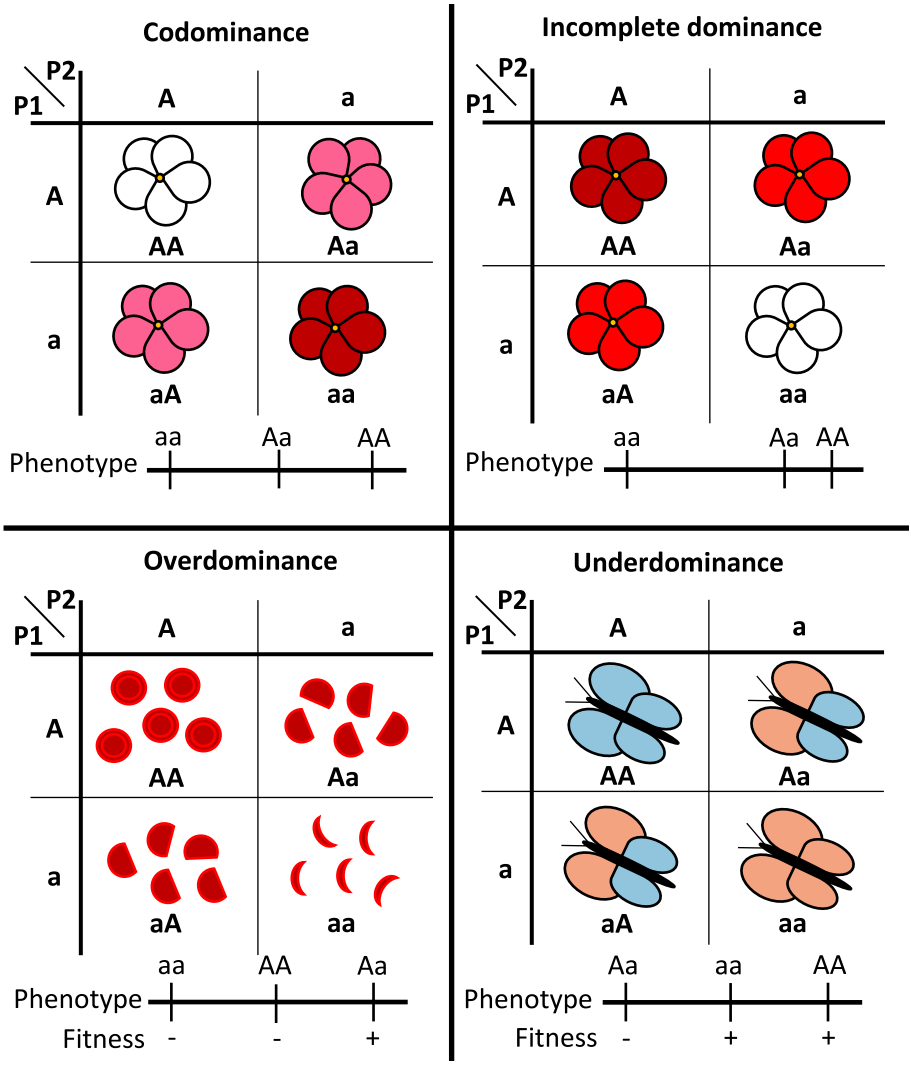

Figure 2.5: Illustration of codominance (top left panel), incomplete dominance (top right panel), overdominance (bottom left panel), and underdominance (bottom right panel). In codominance, the phenotype of the heterozygous individuals is an intermediate blend of the phenotype of both homozygous individuals, whereas in incomplete dominance, the phenotype of the heterozygous individuals closely resembles to one of the homozygous phenotypes. In overdominance, the heterozygous individuals have a superior phenotype, whereas in underdominance, the homozygous individuals have a superior phenotype.

Underdominance is the opposite of overdominance, in which a heterozygous genotype has a reduced fitness compared to a homozygous genotype. Because heterozygous individuals have an inferior phenotype, recessive genotypes are unstable and tend to fixate one of the two alleles in the population. Underdominance can be found in Pseudacraea *eurytus*, an African butterfly that contains two alleles that each imitates the appearance of a toxic butterfly species. An individual with a homozygous genotype at that gene will be able to imitate the appearance of a toxic butterfly, decreasing the likelihood to be killed by his predator. Individuals that are heterozygous at that gene will have an intermediate appearance combining morphological characteristics from both alleles, making them distinguishable from both toxic butterfly species and thus increasing their likelihood to be eaten by a predator.

Until now, only the interactions between different alleles at one locus or gene have been considered. Epistasis is the interaction between different loci or genes and was first described by Bateson (1909). For example, a gene controlling hair color will result in a different phenotype if at another locus, the allele for baldness is present. The phenotypic expression of the gene that controls baldness will mask the hair color of that individual. In diploid individuals, epistatic effects can be observed between genes or between different alleles of both genes.

2.2.3 Sexual reproduction

Sexual reproduction allows for the merging of two gametes, combining genetic information of both parents (see Figure 2.6). Gametes are haploid cells that are obtained after meiosis. During meiosis, the genetic information of both homologous chromosomes is crossed-over, creating two haploid gametes containing one recombined version of each pair of homologous chromosomes. Because crossing overs occur at random places, each gamete will contain a unique set of chromosomes.

2.2.4 Doubled haploid

Double-haploidization (DH) is a technique in which haploid cells undergo a chromosome doubling to create a homozygous diploid individual. The haploids can be obtained via parthenogenesis (asexual reproduction) after which chemical agents like colchicine and acenaphthene are applied to induce the chromosome doubling (Rédei, 2008). The double haploid technique reduces the required time for the development of homozygous lines, increasing the efficiency of a breeding scheme.

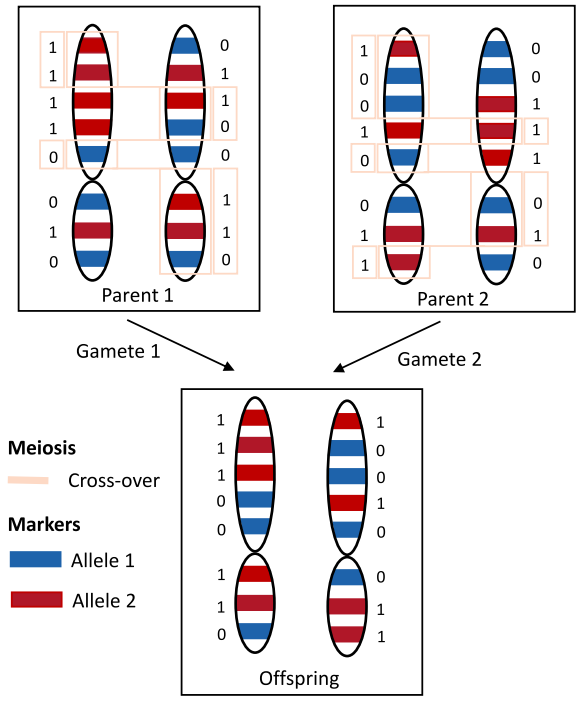

Figure 2.6: Overview of the meiosis process. The gamete is constructed based on the haploids of the parents. The number of crossing overs is Poisson distributed with λ, the expected rate of occurrence, equal to the length of the chromosome in Morgan. The locations of the crossing overs are uniformly distributed over the chromosome.

2.3 Concept of inbreeding

Inbreeding is a natural phenomenon that occurs in a finite population, where two individuals with a common ancestor are crossed. Inbreeding is often used to create superior homozygous lines. However, by consistently crossing closely related individuals in a breeding population, certain deleterious recessive alleles could be inherited from each parent and determine the expression of the phenotype, decreasing the fitness of the offspring.

If both parents have the same chromosomal region containing alleles of a common ancestor, then there is a possibility that both parents pass that chromosomal region to their offspring. In other words, the offspring inherit on both homologous chromosomes the same ancestral genetic information, which is identical by descent (IBD) (see Figure 2.7). At that point, both copies of a chromosomal region have been passed in two separate loops from the common ancestor to an offspring via at least two parents (Falconer and Mackay, 2000). The probability that an individual inherits two alleles that are IBD is better known as the inbreeding

coefficient F_t. Assuming that the ancestor is not inbred, the inbreeding coefficient can be calculated as:

$$F_t = (0.5)^{n_1+n_2+1}, \tag{2.5}$$

with n_1 and n_2 the number of generations between the parent and the common ancestor (Pyeritz et al., 2018). For example, if two cousins are crossed, both cousins are separated from their ancestor (grandparents) by two generations and thus $n_1 = n_2 = 2$, resulting in an inbreeding coefficient of 0.03125. The inbreeding coefficient can easily be calculated using pedigree information. Since the introduction of genetic markers (see Section 2.5), the inbreeding coefficient can also be calculated using genotypic marker scores, allowing for the development of different parental selection methods that try to control the inbreeding coefficient in the next generation (Brisbane and Gibson, 1995; Akdemir and Sánchez, 2016). Because the inbreeding coefficient is calculated with respect to the base population, a different inbreeding coefficient will be obtained if the base population is chosen at an earlier or later generation. Therefore, controlling the inbreeding coefficient is often mismanaged as there is no safe or dangerous value for F_t (Woolliams et al., 2015). It is more important to study and control the rate of inbreeding (ΔF) as it determines the effective population size (N_e) such that $N_e = (2\Delta F)^{-1}$. Several negative effects are associated with higher values of ΔF such as loss of heterozygosity, fixation of unfavorable alleles, and the loss of favorable QTL alleles in the breeding population.

The coancestry coefficient is the probability that two individuals have the same allele by descent. This probability can be calculated based on pedigree information. When genetic markers are available, the genetic relationship can be used to obtain the coancestry coefficient. The coancestry coefficient between two individuals coincides with the expected inbreeding coefficient of their offspring. A high coancestry coefficient indicates that individuals are closely related, resulting in a low genetic variation of the breeding population. To obtain high long-term genetic gains, the coancestry coefficient should be controlled such that a good trade-off is obtained between genetic progress and the preservation of genetic variation (Lindgren and Mullin, 1997). In literature, different strategies have been proposed to combine genetic progress and minimize the coancestry coefficient for long-term profits (Brisbane and Gibson, 1995; Lindgren and Mullin, 1997).

In a breeding program, the goal is to maximize the genetic gain (ΔG). According to Rendel and Robertson (1950) the genetic gain can be written as:

$$\Delta G = i\rho\sigma_g/T, \tag{2.6}$$

with i the selection intensity, ρ the accuracy of selection calculated as the Pearson correlation between the true and predicted criterion values, σ_g the genetic variation, and T the time. From this equation, the genetic gain can be maxi-

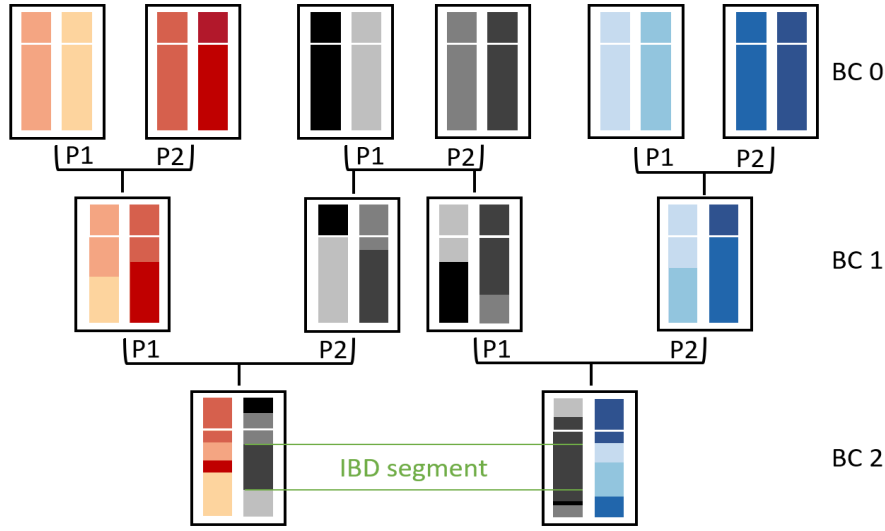

Figure 2.7: Overview of how the ancestral genetic information is passed over different breeding cycles (BC). In this figure, a breeding cycle represents a crossing event between two parents resulting in one offspring. If two individuals contain the same allele that has been copied from a common ancestor, then both alleles will be identical by descent (IBD).

mized by increasing the selection intensity (i), but unfortunately, increasing the selection intensity will also increase ΔF quadratically, leading to high coancestry coefficients (Woolliams et al., 1993). A second approach to maximize the genetic gain is by increasing the prediction accuracy. This can be achieved in different ways.In summary, a good balance should be found between achieving a high genetic gain and preserving the genetic variation. A low value of ΔF indicates that the selection is far from optimal resulting in low genetic progress, whereas a high value of ΔF indicates an intensive selection resulting in high short-term, but low long-term genetic gain. Ideally, ΔF is kept at a constant pre-defined value while the genetic gain is maximized. Unfortunately, it is difficult to predict the F_{t+1}. Assuming that a parent contains a chromosomal region that is IBD, the offspring can only inherit that region on one homologous chromosome. Therefore, at least two breeding cycles are required before a region is IBD on both homologous chromosomes, making it difficult to calculate the impact of each parent on the inbreeding coefficient. However, it is possible to calculate ΔF based on the genetic long-term contribution of each individual (Wray and Thompson, 1990). The long-term genetic contribution of the i-th individual (r_i) is defined as the proportion of the genes of that individual that will be present in all the individuals of the breeding population in the long term (Woolliams et al., 1993). The genetic relationship between r_i and ΔF was further corrected by Woolliams et al. (2000) for non-random mating and

can be mathematically expressed as:

$$\Delta F = 1/4(1-\alpha)\sum_{i=1}^{n} r_i^2\,, \tag{2.7}$$

with α a factor derived from the Hardy-Weinberg equilibrium and n the number of individuals. Based on this equation, methods to keep the ΔF at a constant value should control the long-term genetic contribution of the parental population by preserving the genetic variation in the breeding population (Woolliams et al., 2015). Another strategy is to control the rate of coancestry.

2.4 Breeding scheme

2.4.1 Pure-line breeding

The pure-line method is one of the first breeding schemes that was adopted by breeders. It was designed to create new homozygous superior lines, starting from existing cultivars. Pure-line breeding can be seen as a strategic approach to inbreeding in which inbreeding can be used to pass favorable characteristics to the offspring while minimizing the risks. The pure-line method is illustrated in Figure 2.8 and is mainly used for self-pollinating crops like barley and soybean. Self-pollinating crops are fertilized by their pollen, and can pass only their genetic material to the next generations and will therefore result in homozygous lines. Assuming a heterozygous individual, then the degree of homozygosity after t breeding cycles can be calculated as $1-(0.5)^t$.

The pure-line method starts with the selection of superior lines from a genetically broad population (Allard, 2021). Both parents should contain desirable characteristics that are preferably complementary to each other. In the first breeding cycle, both parents are crossed with each other, creating heterozygous F1 hybrids that contain the genetic material of both parents. Over the next breeding cycles, the superior lines are selected and the next generation is obtained via self-pollination. The required time to create a homozygous line will depend upon the crop and often varies between five and ten breeding cycles. Over the first breeding cycles, the selection is mainly used to remove unfavorable traits from the breeding population. In the subsequent breeding cycles, field tests are used to test the performance of the progeny under different circumstances and environments (local and regional tests) and only the most superior lines are selected. Finally, the high-performance homozygous lines can be used for commercialization purposes.

Single-seed descent (SSD) is a breeding technique that is often used to rapidly fixate genes in a breeding line. Plants are grown in conditions that do not allow

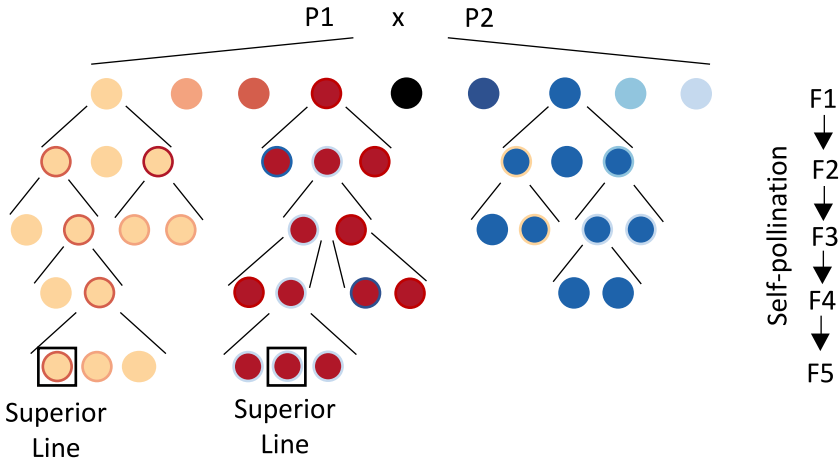

Figure 2.8: Illustration of pure-line breeding in which two parents are crossed followed by 5 to 10 generations of inbreeding.

for an accurate evaluation of the yield potential. In other words, seeds can be collected in a short time frame, but their characteristics cannot be evaluated to guide the selection. This means that single-seed descent can result in a superior homozygous breeding line, but it can also result in an inferior breeding line. In other words, single-seed descent cannot guarantee that the best offspring will be obtained but allows for a rapid evaluation of homozygous lines after hybridization (Gajghate et al., 2018). The number of required generations will depend upon the desired degree of homozygosity. SSD can thus be used to efficiently produce a large number of homozygous lines in artificial conditions (Boerma and Cooper, 1975).

2.4.2 Hybrid breeding

Hybrid breeding was developed in the 20th century and has been able to increase the yield of major crops like maize, and rapeseed (Duvick, 2005; Whitford et al., 2013; Melchinger and Gumber, 1998b). In hybrid breeding, favorable characteristics from two inbred lines are combined to produce a genetically uniform progeny that is superior to both parents. This process is illustrated in Figure 2.9. Compared to pure-line breeding, in hybrid breeding each parent has to come from a different heterotic group and is therefore unrelated to each other. According to Melchinger and Gumber (1998a) a heterotic group consists of individuals that display a simi-

lar combining ability. These individuals can be related or unrelated to each other and come from the same or a different population. To construct a heterotic group, hybrid tests are required to acquire knowledge about the combining ability, hybrid performance, and genetic diversity (Boeven et al., 2016). Individuals can be classified into heterotic groups based on pedigree information. Nowadays, genetic markers like single nucleotide polymorphisms (SNPs) are used, resulting in a more reliable classification of the inbred lines into the different heterotic groups (Qi-Lun et al., 2008; Pejic et al., 1998).

When two homozygous parents from a different heterotic group are crossed, the heterozygous F1 hybrids can be superior to both parents. This effect is known as heterosis and has been described by Shull and Gowen (1952). The positive effect of heterosis is a result of the heterozygosity of the F1 hybrids in which recessive genes that are often associated with unfavorable traits are masked. In other words, dominance, and epistatic effects play an import role in hybrid breeding. Assuming that both parents are homozygous, all the F1 hybrids will have the same genetic information. The heterosis effect has contributed to the success in hybrid breeding and was used to improve different crop characteristics like increased vigor, fruitfulness, speed of development, and pest control (Beckett et al., 2017; Shull and Gowen, 1952).

Over time, heterotic groups are improved by inter-population selection using recurrent breeding in which the mean performance and genetic variance are used as selection criteria. There should be no exchange of genetic material between different heterotic groups. Based on the available parents of both heterotic groups, the genetic values of the hybrids can be predicted based on the general combining ability (GCA) and specific combining ability (SCA) (Sprague and Tatum, 1942). The GCA is related to the genetic value of each parent when crossed with members of the complementary heterotic group and the influence of the additive marker effects on the genetic value of the offspring, whereas the SCA represents the non-additive effects of both parents on the genetic value of the offspring. Based on both heterotic groups, the number of possible genotypes after hybridization rapidly increases up to millions of possibilities. Therefore, breeders often rely on prediction models to guide the parental selection such that the performance of the hybrids is maximized. The factorial design crosses all the available inbred lines to create a broad population of different F1 hybrids. This design would be able to predict the effect of future parental crosses on the genetic value of the offspring, but would also require large datasets that are often unavailable due to limited resources. The incomplete factorial design has been proposed as a solution, only evaluating a subset of the hybrids and thus only requiring a smaller dataset. The top-cross design is a third popular method to construct a training population in which individuals of the first heterotic group with a varying genotype are selected and crossed with one or more test individuals of the second heterotic group. The F1 hybrids are evaluated and used to predict the outcome of future crosses. Once a training pop-

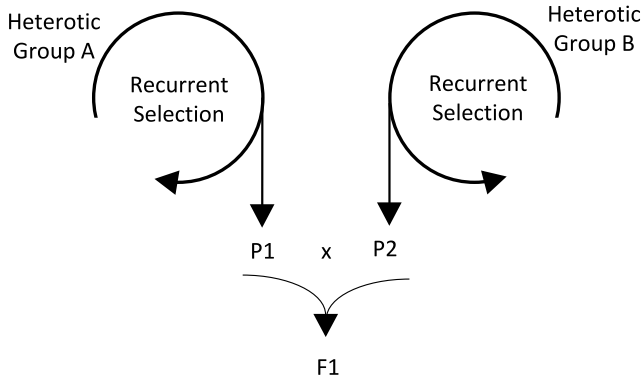

Figure 2.9: Illustration of the hybrid breeding scheme in which individuals from two heterotic groups are crossed to create a heterozygous and superior F1 hybrid.

ulation is constructed, parents from both heterotic groups are selected to create F1 hybrids that can be commercialized.

2.4.3 Recurrent breeding

A recurrent breeding scheme is a closed cycle to increase the frequency of favorable QTL alleles in the breeding population (Orf, 2008). Over each cycle, the favorable QTL alleles are accumulated in the population, continuously increasing the genetic gain in the breeding population. In the case of polygenic traits, the trait is controlled by many genes and thus when recurrent breeding is applied, a frequency shift is observed in which the genetic variance is rearranged in the offspring, increasing the genetic gain of the population (White, 2004). In other words, a recurrent breeding cycle is used to improve a breeding population while preserving the genetic variation to the extent possible. Different types of recurrent breeding have been developed over time: simple recurrent selection, recurrent selection on GCA, recurrent selection on SCA, reciprocal recurrent selection, and marker-assisted recurrent selection.

In simple recurrent selection (SRS), parents are selected from a breeding population based on their phenotype after which a new breeding population is obtained by intermating the selected parents. This method was already used more than 10,000 years ago to improve the first crops, but over time, the simple recurrent selection was replaced by other more efficient breeding schemes (White, 2004). The recurrent selection on GCA (RSGCA) or the half-sib recurrent selection uses

heterozygous testers to increase the genetic gain of a population and is most effective in a breeding population with incomplete dominance. The recurrent selection on SCA (RSSCA) uses homozygous testers to select and combine individuals of a breeding population according to the SCA and will thus be more efficient for traits that are controlled by genes with overdominance or when the breeding value is also governed by non-additive effects like epistasis. The reciprocal recurrent selection (RRS) combines both GCA and SCA to increase the genetic gain of two source populations simultaneously (Comstock et al., 1949). The RRS uses heterozygous testers and is efficient with incomplete dominance, complete dominance, and overdominance. The marker-assisted recurrent selection uses e.g. SNP markers to guide the selection in a recurrent breeding cycle. First, parents are intermated to create F1 hybrids. After two or three generations of single-seed descent, the new breeding population is obtained from which new parents can be selected (see Figure 2.10).

2.4.4 Pedigree method

The pedigree method is often used by breeders to improve self-pollinating and cross-pollinating plants. The pedigree method is used to develop homozygous lines starting with F2 individuals. The development of those F2 individuals is often done using other methods that are less labor-intensive. Over each breeding cycle, the pedigree method will select the best families (offspring that occur from the same parental cross). From the selected families, the seed of each individual is cultivated in a row and the most prominent rows are selected. Finally, the superior individuals are selected from these prominent rows. The selected individuals are then crossed and the offspring is evaluated (Capettini, 2009). In the last breeding cycles, the breeding lines that survived the different selection rounds are tested in different environments before commercializing the superior inbred lines.

2.5 Genetic markers

2.5.1 The introduction of genetic markers in plant breeding

The crops available today are the result of decades of breeding. Before biotechnological techniques were available, breeders used morphological characteristics to select the most successful individuals of a breeding population. Therefore, breeders needed to grow and monitor every single individual before making a final selection, making this an expensive and time-consuming process. With the discov-

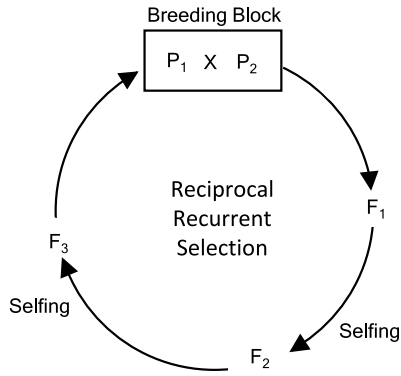

Figure 2.10: Illustration of the reciprocal recurrent selection scheme. First, parents are crossed to create F1 hybrids. After two (or three generations) of single-seed descent, the new breeding cycle is obtained from which new parents can be selected.

ery of genetic markers, genetic information can be mapped and used to guide and improve current breeding programs. A genetic marker can be a chromosomal landmark or allele that traces a specific DNA region, a specific DNA sequence at a known position on the genome, or a gene that can be used as a probe to detect the presence of a specific nucleus, chromosome, or locus (King and Stansfield, 1997). In general, three different classes of genetic markers are distinguished: morphological, biochemical, and molecular markers. The most primitive ones are morphological markers, which are based on morphological characteristics. Biochemical markers are based on the detection of biochemical metabolites like the production of certain proteins or enzymes. Both the morphological and biochemical markers are limited in number and often require a fully developed organism, reducing the usefulness of these markers (Winter and Kahl, 1995). Molecular markers are used to detect variation in certain DNA regions and can, therefore, be applied at the earliest development stage of any organism. Different molecular markers are available like single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs).

When a marker is located close to a gene (or QTL) of interest, that marker and gene could be in linkage disequilibrium (LD), meaning that both the gene and the marker are often inherited together. In that case, the marker acts as a proxy to monitor the presence or absence of that gene in the genome. For biallelic SNP markers, the LD will remain the same for different allelic combinations between two loci, however, when multiallelic markers such as SSRs are used, the LD value between two loci will depend on the allelic combination. Based on the infinitesimal model (see Subsection 3.1), a trait of interest is controlled by many different

QTLs. Because most QTL locations and effects are unknown, markers are used instead to predict the trait of interest via marker-assisted selection or genomic selection. The introduction of molecular markers in breeding programs has different advantages (Roychowdhury et al., 2013):

1. Time saving: with molecular markers, the genetic information of an individual can be collected in every stage of its development making it possible to select and develop individuals in an earlier stage and thus reducing the time per breeding cycle.

2. Consistency: compared to morphological characteristics, molecular markers are not influenced by environmental effects, resulting in a more consistent evaluation of the individuals.

3. Biosafety: molecular markers can work as a proxy to indicate if disease resistance is present or absent without the need to isolate the gene itself or to inoculate the pathogen.

4. Efficiency: markers make it possible to evaluate the different progeny lines in an early stage, resulting in more efficient breeding as unsuccessful breeding lines can be rejected directly.

Molecular markers were able to improve the genetic gain in several breeding programs and helped to introduce several new characteristics like tolerance to abiotic stress, resistance to pathogens, and agronomic characteristics in a breeding population (Tester and Landridge, 2010; Song et al., 2003). Using molecular markers, different variables like the inbreeding coefficient, allele frequency, and genetic relationship of the breeding population can be calculated and used to guide the parental selection, further optimizing current breeding programs (Akdemir and Sánchez, 2016).

2.5.2 Genetic markers and their applicability

Different types of markers are available to translate the genetic information present in the genome of an individual. SSRs, also known as microsatellites, are short strings of DNA containing one to six nucleotides that are repeated several times (Appleby et al., 2009). In plants, SSRs are widely spread throughout genic and intergenic regions of the genome and occur with a frequency of approximately 1 SSR per 50 kbp (Morgante et al., 1994; Cox and Mirkin, 1997). The number of repeats is specific and can be influenced by mutations. These mutations are caused by errors made by the DNA polymerase enzyme during DNA replication, which often results in the addition or subtraction of a repeat unit (Strand et al., 1993). Because SSRs are abundantly present in the genome of plants and are highly polymorphic,

SSRs are often used as markers in MAS, genetic diversity analyses, and genetic mapping (Ashkani et al., 2012; Appleby et al., 2009).

SNPs are the most abundant markers present in a genome. An SNP is a locus at which in a certain fraction of the population the nucleotide at that locus is substituted with another nucleotide. The different nucleotides that can be detected in a SNP are referred to as the marker alleles. Biallelic SNP markers will always have two possible nucleotides at each locus. Because SNPs have a low mutation rate and remain stable during evolution, they are often used as genetic markers in a wide range of applications such as MAS, GS, and LD studies.

2.5.3 Genotyping techniques

The simplest and most popular way to obtain the genotype of SSR markers is by using PCR technology (Weber and May, 1989). First, a specific primer is required. Primers are oligonucleotides and are specific for each SSR marker. For most crops, SSR primers are already available in public libraries, otherwise, they have to be built from scratch. The primers are complementary to a region flanking the SSR. During PCR, the SSR region is amplified (Mason, 2015). Next, the amplified fragments can be visualized using capillary electrophoresis.

GeneChip assays are based on the hybridization of allele-specific oligonucleotide (ASO) probes with a specific DNA fragment to distinguish SNP alleles. To do so, the difference in thermal stability between a perfectly matched and mismatched ASO probe and the target DNA is used. Because it is difficult to achieve an optimal hybridization, 40 different ASOs are used to identify each SNP loci. When hybridization occurs, a fluorescence pattern can be observed and used to distinguish SNP alleles (Chagné et al., 2007).

Illumina Microarray Technology uses silica microbeads that contain thousands (or even millions) of genotypes or oligonucleotides that can be used to identify the genotype of an individual. The DNA fragments are added to the microarray. Each probe on the microbead will bind with a complementary DNA fragment that is flanking a marker of interest. Next, allele specificity is conferred by extending each hybridized probe with one labeled nucleotide using DNA polymerase. After exciting the microbeads with a laser, the intensity of that signal can be used to calculate the allelic ratio at each locus (Illumina, 2021).

2.6 Phenotyping

The term phenotyping was coined by Johannson during his research on the inheritance of seed size in beans (Johannsen, 1903). In plant breeding, the phenotype

can refer to a wide range of traits. Those traits can range from grain yield to the presence of certain cell metabolites. In today's breeding programs, elite plants are often selected based on hundreds of measurements. To do so, high-throughput phenotyping can be used, often using non-destructive sensors and imaging technology (Furbank and Tester, 2011). Over time, phenotyping using sensors and cameras has become more affordable due to a decrease in the cost of the required equipment (Deery et al., 2014). Different sensors like thermography point sensors, drones, and phenomobiles have been successfully used in the field (Walter et al., 2015). In greenhouses, robots can be deployed to measure different characteristics in an automated way (Leister et al., 1999).Phenotyping is still an expensive process, but with the development of different automated phenotyping platforms, we can expect that the cost will further decrease over the next years.

2.7 Marker-assisted selection

Marker-assisted selection (MAS) uses molecular markers that are in strong LD with one or more QTL alleles. Although it seemed that MAS could revolutionize animal and plant breeding, it is mainly constrained to monogenic traits that are controlled by major QTL effects (Xu and Crouch, 2008; Dekkers and Hospital, 2002). Because MAS is able to accurately predict these traits, it plays a significant role in the backcrossing of major genes in elite varieties (Holland, 2004; Heffner et al., 2009). MAS makes it possible to evaluate individuals at an early development stage, reducing the overall cost, selecting for low-heritability traits, and improving the selection by eliminating environmental, pleiotropic, or epistasis effects (Roychowdhury et al., 2013). Unfortunately, MAS has also several limitations. First, because MAS uses molecular markers that are in strong LD with QTLs, linkage mapping is required to find such markers. Therefore, huge datasets with different breeding populations are needed. Breeding programs are often limited in funds and often use a biparental population with a limited size. By consequence, such datasets are underpowered causing a poor estimation of the QTL positions and reduce the applicability of MAS in most breeding programs (Dekkers and Hospital, 2002; Schön et al., 2004). Second, MAS fails to grasp the effect of many small-effect QTLs, which is essential for the successful development of several crop varieties containing complex quantitative traits (Lande and Thompson, 1990). Third, the use of a dataset with a biparental population that does not represent the current breeding population could result in a poor estimation of the QTL positions (Jannink et al., 2001; Sneller et al., 2009).

2.8 Genomic selection

Genomic selection (GS) is an improved version of MAS in which the challenges imposed by MAS for quantitative traits are overcome. In contrast to MAS, GS does not reduce the number of markers to those that are associated with a large-effect QTL but uses high-density markers that are equally distributed over the whole genome. When high-density markers are used, each QTL should be in LD with at least one marker. The remaining markers will then be able to capture the remaining genetic variance, leading to an accurate estimation of large- and small-effect QTLs (Heffner et al., 2009, 2010; Beyene et al., 2015). GS still needs a training population to estimate the marker effects. A training population contains genotypic and phenotypic information of individuals and is used to estimate all the marker effects simultaneously. Next, the genomic estimated breeding values (GEBVs) of a breeding population can be calculated. In contrast to the training population, only genotypic information is required (Meuwissen et al., 2001). The GEBVs can then be used as a selection criterion and replace the need for phenotypic data. The performance of the GEBVs can be calculated based on a test population. This population contains the genotype and phenotype of individuals that have not been used in the training population. Based on the predicted marker effects, the GEBVs can be calculated and compared with the phenotype. The prediction accuracy is usually defined as the Pearson correlation between the GEBVs and the phenotypic values. Over time, a breeder should update the training population to keep an accurate estimation of the marker effects (Neyhart et al., 2017). GS was able to revolutionize current breeding programs in both plant and animal breeding (Hayes et al., 2009; Heffner et al., 2009). Not only did GS improve the genetic gain in different major crops, but it also reduced the need for phenotypic data, which is not only costly to collect but also requires fully developed individuals. By using the GEBVs as a selection criterion, an individual can already be evaluated as a seedling, increasing the time efficiency of most breeding programs.

GS uses a large number of markers and, therefore, there are often more marker effects that need to be estimated than there are observations available in the training population. Certain markers can also be correlated to each other resulting in the problem of rank deficiency. One could consider reducing the number of markers, but this would conflict with the main idea of GS in which marker selection is avoided (Meuwissen et al., 2001; Jannink et al., 2010). When predicting the marker effects, overfitting should be avoided at all cost. Overfitting can occur when the estimated marker effects captures the residual noise in the model, leading to an accurate prediction of the individuals present in the training population but fails to predict the GEBVs of individuals in the test population. Different methods like Best Linear Unbiased Prediction (BLUP), ridge regression, Bayes A–C models, etc. have been proposed to predict the marker effects while avoiding overfitting. In the next paragraph, a short overview of the different models is listed. In the next chapter,

each model will be discussed in more depth.

The genomic Best Linear Unbiased Predictor (gBLUP) is commonly used in GS and estimates the marker effects as a random effect. BLUP assumes that the marker effects are normally distributed with a mean 0 and a variance component σ_u^2. Bayes A–C models are based on the assumption that each marker effect has a different variance component, that certain markers do not affect the phenotype, or that certain marker effects follow a Student t-distribution. Different machine learning techniques like support vector machines (SVM) and random forest (RF) have also been used, but their application in GS is limited and will therefore not be considered.

Genomic selection facilitates the rapid selection of superior genotypes and accelerates the progress in plant breeding. By taking into account environmental interactions, dominance, and epistatic effects, GS is widely used to improve phenotypic traits in different major crops (Sweeney et al., 2021). GS is also used for improving multiple traits at once or to select individuals that are resistant to a certain pathogen or (a)biotic stress (Lenz et al., 2020; Pincot et al., 2020). Aside from plant breeding, genomic selection is also used in animal breeding to improve the livestock, but also to reduce their environmental impact (Pryce and Mekonnen, 2020).

3

Mathematical background

Before being able to simulate a breeding population, the biological background needs to be translated into mathematical equations. Therefore, in this chapter, the mathematical background of quantitative genetics and genomic selection is discussed. The infinitesimal model translates the biological background of quantitative genetics into mathematical equations that will be used in the next chapters to simulate a breeding population. Next, the linear mixed effects model is introduced. This model can be used to predict breeding values of the genotyped members of a breeding population. To do so, the marker effects (and variance components) are estimated using the ridge regression genomic Best Linear Unbiased Predictor (rr-gBLUP) or Bayesian models.

3.1 The infinitesimal model

To develop new breeding methods, it is not only important to collect sufficient data of existing breeding populations, but also to implement these methods in field experiments. Taking into account that the number of breeding cycles in plant

breeding is limited to only a few per year, it can take a considerably long time before the effects of these methods can be studied. The infinitesimal model is a simple and robust model that simulates the behavior of a breeding population over different breeding cycles. By simulating a breeding population, the effects of a breeding method can be studied in a shorter time frame and under different conditions. The infinitesimal model or polygenic model assumes that a quantitative trait is controlled by an infinitely large number of independent additive small-effect QTLs (Bulmer, 1971; Fisher, 1918). In a large population, a quantitative trait of the offspring is normally distributed with a mean genetic value situated between the genetic value of both parents. The true breeding value (a_i) of individual i is defined as:

$$a_i = \sum_{k=1}^{L} Z_{ik} q_k \,, \tag{3.1}$$

with L the number of QTLs, Z_{ik} the genotype of the i-th individual at the k-th QTL and q_k the k-th additive QTL effect. This equation can be rewritten as:

$$a_i = \sum_{k=1}^{L} X_{ik} \,, \tag{3.2}$$

with X_{ik} the contribution to the genetic value of the k-th QTL of the i-th individual (Lange, 1997). Assuming that each locus has a similar QTL effect, then, according to the central limit theorem, $\mathbf{a}$ approximately follows a multivariate normal distribution (Lange, 1978; Fisher, 1918).

The phenotypic value of a trait is defined as:

$$y_i = a_i + E_i \,, \tag{3.3}$$

with y_i the phenotypic value of the i-th individual and E_i the environmental effect of the i-th individual. The simplest way to define the environmental contribution is by assuming that $\mathbf{E}$ follows a normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Upsilon})$ with $\boldsymbol{\Upsilon} = \sigma_E^2 \mathbf{I}_n$, σ_E^2 the environmental variance and $\mathbf{I}_n$ the identity matrix of dimension n. Asides the environmental contribution, phenotypic data often contains measurement errors, therefore, a residual error is modeled according to a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$ and σ_e^2 the residual variance. Taking into account the residual error (ϵ), the phenotypic value of an individual i is:

$$y_i = a_i + E_i + \epsilon_i \,. \tag{3.4}$$

3.2 Linear mixed effects model

A linear regression model is a statistical approach to express the mathematical relationship between different variables using the equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \tag{3.5}$$

with $\mathbf{y}$ a vector containing n different observations, $\mathbf{X}$ the incidence matrix with dimension $n \times (p+1)$, $\boldsymbol{\beta}$ a vector containing $p+1$ fixed effects and $\boldsymbol{\epsilon}$ the residual error. The variance of the observations ($\mathbf{y}$) can be calculated as: $\text{var}(\mathbf{y}) = \text{var}(\boldsymbol{\epsilon}) = \mathbf{V}$. For example, let us assume that there is a linear relationship between the milk production and the body mass per cow and between the milk production and the amount of digested feed per cow. The milk production ($\mathbf{y}$) can then be predicted with $\mathbf{X}$, a matrix containing in the first column only ones, in the second column the mass of each cow, and in the third column the amount of digested feed of each cow. $\boldsymbol{\beta}$ will contain three variables, the averaged milk production per cow, the effect of the body mass on the milk production and the effect of the amount of digested feed on the milk production. Normally, fixed effects ($\boldsymbol{\beta}$) are unknown and need to be estimated with the help of a training population containing data about the milk production, body mass, and feed quantity for a population of cows. The fixed effects are estimated using the generalized least square estimator (GLSE):

$$\text{GLSE}(\boldsymbol{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\,, \tag{3.6}$$

with $\mathbf{V}$ the covariance matrix. Because $\mathbf{V}$ is in most cases unknown, the ordinary least squares (OLS) is more commonly used. Assuming that $\mathbf{V} = \mathbf{I}\sigma_u^2$, the covariance matrix is simplified and fixed effects can easily be predicted as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,, \tag{3.7}$$

with $\hat{\boldsymbol{\beta}}$ the estimated fixed effects. In vivo, predicting milk production based on feed quantity and body mass is difficult, but the use of genetic markers has been proven useful to predict complex traits such as milk production. Based on the infinitesimal model, the phenotypic values ($\mathbf{y}$) can be calculated as follows:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Q}\boldsymbol{\alpha} + \boldsymbol{\epsilon}\,, \tag{3.8}$$

with $\mathbf{1}_n$ a vectors of size n containing ones, μ the phenotypic mean, $\mathbf{Q}$ a matrix containing the genetic information of the QTLs (coded as -1, 0, and 1), $\boldsymbol{\alpha}$ the QTL effects and $\boldsymbol{\epsilon}$ the residual error (Goddard, 2009; Goddard et al., 2011; Hayes et al., 2009). The genetic information ($\mathbf{Q}$) of the QTLs is often unknown and difficult

to estimate, therefore, genomic markers covering the whole genome are used instead. Because in genome-wide regression, the numbers of markers are usually much larger than the number of observations, a problem of rank deficiency can occur, where marker effects cannot be estimated at the same time due to low degrees of freedom (Neves et al., 2012). Therefore, marker effects are predicted using a linear mixed effects model, treating the marker effects as random effects and not as fixed effects (Meuwissen et al., 2001). The linear mixed effects model is expressed as:

$$\mathbf{y} = \mathbf{1}_n\beta + \mathbf{Zu} + \boldsymbol{\epsilon}\,, \tag{3.9}$$

with β the fixed effect containing the phenotypic mean, $\mathbf{Z}$ is the incidence matrix containing the genotype of each individual (coded as -1, 0, and 1) and $\mathbf{u}$ the additive marker effects with length k. The additive marker effects follow a normal distribution with a mean zero and a covariance matrix $\mathbf{G}$. The residual errors $\boldsymbol{\epsilon}$ also follow a normal distribution with a mean zero and a covariance matrix $\mathbf{R}$. Assuming that the random effects and the residual error are independent from each other, the variance of y is simplified to:

$$\text{var}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}\,, \tag{3.10}$$

with $\mathbf{G} = \text{var}(\mathbf{u})$ and $\mathbf{R} = \text{var}(\boldsymbol{\epsilon})$. The linear mixed effects model can also be expressed in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta + \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1k} \\ Z_{21} & Z_{22} & \dots & Z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{nk} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{3.11}$$

The additive marker effects and mean phenotypic value need to be estimated, therefore, a handful of methods are available. Although rr-gBlup is the most commonly used method, an overview of the different methods is listed in detail in the next paragraphs.

3.3 The genomic best linear unbiased predictor

The Best Linear Unbiased Predictor (BLUP) was introduced by Henderson (1963) predicting phenotypic values of a breeding population using pedigree information. Over time, pedigree information was replaced by molecular markers leading to the development of the genomic Best Linear Unbiased Predictor (gBLUP). The introduc-

tion of molecular markers combined with gBLUP has improved livestock breeding programs, leading to a 20-50% increase of the prediction accuracy (VanRaden et al., 2009). While gBLUP uses a linear model to predict the phenotype, non-linear models like Bayes A, Bayes B and Bayes C models have also been proposed. The use of both linear and non-linear models have resulted in similar prediction accuracies (Moser et al., 2009), but gBLUP has a better performance compared with the Bayesian models when many small-effect QTLs are present. By simultaneously predicting all the marker effects at once, higher prediction accuracies are observed compared with single marker regression (Yang et al., 2010). To simultaneously predict all the marker effects, a linear mixed effects model (see Eq. (3.9)) is used under the following assumptions:

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right), \tag{3.12}$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{V}),$$

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R},$$

with $\mathbf{X}$ a vector of size n containing ones, $\mathbf{G} = \sigma_u^2 \mathbf{I}_k$ and $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. The mean squared error of the linear mixed effects model is given as:

$$\mathrm{E}[(\mathbf{u} - \hat{\mathbf{u}})^2] = \iint (\hat{\mathbf{u}} - \mathbf{u})^2 f(\mathbf{u}, \mathbf{y}) d\mathbf{y} d\mathbf{u}, \tag{3.13}$$

with $f(\mathbf{u}, \mathbf{y})$ the joint probability density function of the additive marker effects $\mathbf{u}$ and the phenotypic values $\mathbf{y}$, which is given as:

$$f(\mathbf{u}, \mathbf{y}) = \frac{\exp\left(-\frac{1}{2} \begin{bmatrix} \mathbf{u}' & (\mathbf{y} - \mathbf{X}\beta)' \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \mathbf{u} \\ \mathbf{y} - \mathbf{X}\beta \end{bmatrix} \right)}{(2\pi)^{(n+k)/2} |\boldsymbol{\Sigma}|^{1/2}}, \tag{3.14}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{G} & \mathbf{G}\mathbf{Z}' \\ \mathbf{Z}\mathbf{G} & \mathbf{V} \end{bmatrix},$$

and k the number of elements in vector $\mathbf{u}$. Assuming that $\mathbf{V}$ is invertible, the determinant of $\boldsymbol{\Sigma}$ can be calculated as:

$$|\boldsymbol{\Sigma}| = |\mathbf{V}||\mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}|, \tag{3.15}$$

and therefore, $\boldsymbol{\Sigma}^{-1}$ can be calculated using the blockwise inversion:

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \mathbf{W} & -\mathbf{W}\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} \\ -\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{W} & \mathbf{V}^{-1} + \mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} \end{bmatrix}, \tag{3.16}$$

with

$$\mathbf{W} = (\mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G})^{-1}. \tag{3.17}$$

The joint density function can then be rewritten as:

$$f(\mathbf{y}, \mathbf{u}) = \frac{\exp\left(-\frac{1}{2}\left((\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u})'\,\mathbf{R}^{-1}\,(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u}) + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u}\right)\right)}{(2\pi)^{(n+k)/2}|\mathbf{R}|^{1/2}|\mathbf{G}|^{1/2}}. \tag{3.18}$$

To minimize the mean squared error, Eq. (3.18) is maximized by taking the partial derivatives with respect to β and $\mathbf{u}$:

$$\frac{\partial f(\mathbf{y}, \mathbf{u})}{\partial \beta} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\beta - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}) f(\mathbf{y}, \mathbf{u}), \tag{3.19}$$

$$\frac{\partial f(\mathbf{y}, \mathbf{u})}{\partial \mathbf{u}} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\beta - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u} + \mathbf{G}^{-1}\mathbf{u}) f(\mathbf{y}, \mathbf{u}). \tag{3.20}$$

An optimum can be found by setting both derivatives to zero, which results in the following equations:

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\beta} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{u}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}, \tag{3.21}$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\beta} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}, \tag{3.22}$$

with $\hat{\beta}$ and $\hat{\mathbf{u}}$ the estimated values for β and $\mathbf{u}$, respectively. These equations are known as the mixed model equations (MME) and can be written in a matrix form:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{3.23}$$

The MME are commonly simplified by assuming that both the molecular marker effects and the residual effects are independent and that both variance components σ_u^2 and σ_e^2 are homoscedastic:

$$\text{var}(\mathbf{u}) = \mathbf{G} = \sigma_u^2 \mathbf{I}_k, \quad \text{var}(\boldsymbol{\epsilon}) = \mathbf{R} = \sigma_e^2 \mathbf{I}_n. \tag{3.24}$$

By taking into account these assumptions, the MME are further simplified to:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \delta\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \tag{3.25}$$

with $\delta = \sigma_e^2/\sigma_u^2$. If initial values for both variance components are known, both β and $\mathbf{u}$ can be calculated by solving Eq. (3.25). Because δ can be seen as a penalization factor that controls the sum of squares of the marker effects ($\mathbf{u}$), this approach is often referred to as ridge regression gBLUP (rr-gBLUP). The values for

both variance components can be estimated iteratively via the Restricted Maximum Likelihood method (REML). By rewriting Eq. (3.12), **G** and **R** can be written in function of $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, two vectors containing the variance components of respectively **G** and **R**:

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \begin{bmatrix} \mathbf{G}(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\boldsymbol{\phi}) \end{bmatrix}\right). \tag{3.26}$$

Based on the probability density function of the phenotypic values $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{V})$ the log-likelihood function is given as:

$$l(\beta, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\phi}|\mathbf{y}) = -\frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta) + n\log(2\pi\sigma^2) + \log(|\mathbf{V}|)\right). \tag{3.27}$$

The log-likelihood function can be differentiated with respect to β and the variance components. Differentiating the log-likelihood with respect to β and equating it to zero will result in the GLSE (see Eq. (3.6)). Compared to the maximum likelihood estimator, REML only maximizes the part of the log-likelihood that is invariant to the fixed effects. The derivation of REML is beyond the scope of this thesis, but can be found in De Coninck et al. (2014) and Patterson and Thompson (1971). The REML log-likelihood is given as:

$$l_{\text{REML}}(\sigma^2, \boldsymbol{\gamma}, \boldsymbol{\phi}) = -\left((n-1)\log(\sigma^2) + \log|\mathbf{G}| + \log(|\mathbf{R}|) + \log(|\mathbf{C}|) + \frac{\mathbf{y'Py}}{\sigma^2}\right), \tag{3.28}$$

with n the number of observations, k the number of markers, **C** the coefficient matrix of the MME (see Eq. (3.25)) and:

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}. \tag{3.29}$$

By making the following assumptions: $\sigma^2 = \sigma_e^2$, $\mathbf{R} = \mathbf{I}_n$, $\mathbf{G} = \gamma\mathbf{I}_k$ and $\gamma = \sigma_u^2/\sigma_e^2$, the REML log-likelihood can be written in function of the two variables σ_e^2 and γ:

$$l_{\text{REML}}(\sigma_e^2, \gamma) = -\left((n-1)\log(\sigma_e^2) + k\log(\gamma) + \log(|\mathbf{C}|) + \frac{\mathbf{y'Py}}{\sigma_e^2}\right). \tag{3.30}$$

Both **C** and **P** only depend on γ and, therefore, based on the REML log-likelihood, if γ is known, a solution for σ_e^2 exists that maximizes the log-likelihood:

$$\sigma_e^2 = \frac{\mathbf{y'Py}}{n-1}. \tag{3.31}$$

Differentiating the REML log-likelihood with respect to γ results in the score function that will be used in a next step to find both variance components iteratively.

The score function is mathematically expressed as:

$$\frac{l_{\text{REML}}}{\partial\gamma} = -\left(\frac{k}{\gamma} - \frac{\text{tr}(\mathbf{C}^{ZZ})}{\gamma^2} - \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sigma_e^2\gamma^2}\right), \tag{3.32}$$

with $\mathbf{C}^{ZZ}$ the lower right block of the inverse of $\mathbf{C}$. Because the score function cannot be used to directly calculate γ, an iterative scheme is used in which both the value for γ and σ_e^2 are approximated. The Newton–Raphson method is a well-known iterative method that can maximize the log-likelihood relying on the first differentiation of the log–likelihood with respect to the different parameters and the Hessian matrix. The gradient is given as:

$$\nabla l_{\text{REML}(\sigma_e^2,\gamma)} = \begin{bmatrix} \frac{\partial l_{REML}(\sigma_e^2,\gamma)}{\partial\sigma_e^2} \\ \frac{\partial l_{REML}(\sigma_e^2,\gamma)}{\partial\gamma} \end{bmatrix}, \tag{3.33}$$

and the Hessian matrix is given as:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2}{(\partial\sigma_e^2)^2} & \frac{\partial^2}{\partial\sigma_e^2\partial\gamma'} \\ \frac{\partial^2}{\partial\sigma_e^2\partial\gamma'} & \frac{\partial^2}{\partial\gamma\partial\gamma'} \end{bmatrix} l_{\text{REML}}(\sigma_e^2,\gamma). \tag{3.34}$$

Because the construction of the Hessian matrix may be tedious, it can be replaced by the Fisher information matrix which is easier to construct (Patterson and Thompson, 1971).

The iterative scheme of Newton–Raphson can be written as:

$$\boldsymbol{\kappa}_{i+1} = \boldsymbol{\kappa}_i - \mathbf{H}_i^{-1}\nabla l_{\text{REML}}(\boldsymbol{\kappa}_i), \tag{3.35}$$

with $\boldsymbol{\kappa}_i$ a vector containing γ and σ_e^2 at the i-th iteration, $\mathbf{H}_i$ the Hessian matrix of l_{REML} at the i-th iteration and $\nabla l_{\text{REML}}(\boldsymbol{\kappa}_i)$ the gradient of the REML log-likelihood with respect to $\boldsymbol{\kappa}_i$.

When the Hessian matrix is replaced by the Fisher information matrix, the calculation of the trace of a large-sized matrix is still required. To avoid this, the average between the observed and expected Hessian can be computed, resulting in the averaged information matrix (AI-REML) (Thompson et al., 2003):

$$\mathbf{H}_{\text{AI}} = \frac{1}{\sigma_e^2}\mathbf{Q}'\mathbf{P}\mathbf{Q}, \tag{3.36}$$

with

$$\mathbf{Q} = \begin{bmatrix} \frac{y}{\sigma_e^2} & \frac{\mathbf{Z}\hat{\mathbf{u}}}{\gamma} \end{bmatrix}. \tag{3.37}$$

Substituting $\mathbf{H}_i^{-1}$ with $\mathbf{H}_{\text{AI},i}^{-1}$ in Eq. (3.35) results in the AI-REML. Based on a dataset of a breeding population containing phenotypic values ($\mathbf{y}$) and genotypic information ($\mathbf{Z}$) coded as -1, 0 or 1, u, β and both variance components can be estimated by following the next steps:

1. Set an initial value for γ. Next, ξ is defined, such that when the relative update of γ becomes smaller than ξ, convergence is reached and a solution for β, $\mathbf{u}$, σ_u^2 and σ_e^2 is found.

2. The MME (see Eq. (3.25)) are constructed. Using the Cholesky decomposition, a solution for β and $\mathbf{u}$ is found.

3. Based on the coefficient matrix ($\mathbf{C}$) of the linear mixed effects equations, an analytical solution for σ_e^2 can be calculated using Eq. (3.31).

4. Based on $\hat{\mathbf{u}}$ and the analytical solution of σ_e^2, $\mathbf{Q}$ can be calculated according to Eq. (3.37).

5. The matrix $\mathbf{P}$ is calculated according to Eq. (3.29) assuming that $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$ with $\mathbf{G} = \gamma\ \mathbf{I}_k$ and $\mathbf{R} = \mathbf{I}_n$. Next, $\mathbf{H}_{\text{AI}}$ is calculated according to Eq. (3.36).

6. The score function is calculated using Eq. (3.32) and multiplied with the inverse of $\mathbf{H}_{\text{AI}}$ resulting in a 2×2 update matrix. Next, the sum of the lower right element of the updated matrix and γ_i is taken yielding the value for γ_{i+1} which will be used in the next iteration. If γ_{i+1} has a negative value, the relative update (the lower right element of the updated matrix) is divided by 2 until the new value for γ_{i+1} has a value greater than zero.

7. Repeat step 2 to step 6 until convergence, which is reached when the relative update becomes smaller than ξ.

In this dissertation, the additive marker effects are estimated using the `rrBLUP` package in R (Endelman, 2011). The `rrBLUP` package uses spectral decomposition to maximize the log-likelihood function. When $\beta = (\mathbf{X'V}^{-1}\mathbf{X})\mathbf{X'V}^{-1}\mathbf{y}$ and $\hat{\sigma}_e = \frac{\mathbf{y'Py}}{n-1}$, the log-likelihood function is maximized. Using spectral decomposition, the eigenvalues $\boldsymbol{\lambda}$ can be obtained from the following equation:

$$\begin{aligned}\sigma_e^2 &= (\mathbf{U'y})'\text{diag}[(\lambda_1 + \gamma, \ldots, \lambda_{n-1} + \gamma)]^{-1}\mathbf{U'y} \\ &= \sum_{s=1}^{n-1} \frac{\eta_s^2}{\lambda_s + \gamma},\end{aligned} \tag{3.38}$$

with $\mathbf{U}$ a eigenvector matrix of size $n \times (n-1)$, λ_s the s-th eigenvalue and $\boldsymbol{\eta}$ is defined as $\mathbf{U'y} = [\eta_1, \eta_2, \ldots, \eta_{n-1}]$ (Kang et al., 2008). The restricted maximum likelihood can be rewritten such that:

$$f_{\text{REML}}(\gamma) = \frac{1}{2}\left[(n-1)\log\left(\frac{n-1}{2\pi}\right) - (n-1) \right.$$
$$\left. -(n-1)\log\left(\sum_{s=1}^{n-1}\frac{\eta_s^2}{\lambda_s+\gamma}\right) - \sum_{s=1}^{n-1}\log(\lambda_s+\gamma)\right], \quad (3.39)$$

with $f_{\text{REML}}(\gamma)$ the rewritten log–likelihood function. The differentiation of the log–likelihood function is given as:

$$f'_{\text{REML}}(\gamma) = \frac{n-1}{2}\frac{\sum_s \eta_s^2/(\lambda_s+\gamma)^2}{\sum_s \eta_s^2/(\lambda_s+\gamma)} - \frac{1}{2}\sum_s \frac{1}{\lambda_s+\gamma}. \quad (3.40)$$

The parameter γ can be obtained using the Newton–Raphson method. Because the spectral decomposition needs to be calculated only once, this method has a lower time complexity compared to the classical REML using the Newton–Raphson method. The time complexity of REML is $O(rn^3)$ while using spectral decomposition the time complexity becomes $O(n^3 + rn)$ with r the number of iterations (Kang et al., 2008).

3.4 Gibbs sampling

The Gibbs sampling is a special case of the Metropolis-Hastings algorithm to obtain a sequence of observations from a multivariate distribution (Geman and Geman, 1984). Gibbs sampling is used when it is difficult to sample from the multivariate distribution, but not from the conditional distributions. The sampling procedure can be divided into three steps:

1. Initialize X_1^1
2. Sample $X_2^1 \sim P(X_2|X_1^1)$
3. Sample $X_1^2 \sim P(X_1|X_2^1)$
4. Repeat Steps 2 and 3 k times

The BGLR package in R is used in the next part to estimate the additive marker effects ($\hat{\mathbf{u}}$) using Bayesian models. In the case of Bayes Ridge-Regression (BRR), the Gibbs sampler is used to estimate $\hat{\mathbf{u}}$, using the posterior probability distributions and initial values for β, σ^2, γ, and ϕ. Next, the different variables are repeatedly sampled from their conditional distributions:

Sample β^{i+1} from $p(\beta|{\sigma^2}^i, \gamma^i, \phi^i, y)$;

Sample ${\sigma^2}^{i+1}$ from $p(\sigma^2|\beta^{i+1}, \gamma^i, \phi^i, y)$;

Sample γ^{i+1} from $p(\gamma|\beta^{i+1}, {\sigma^2}^{i+1}, \phi^i, y)$;

Sample ϕ^{i+1} from $p(\phi|\beta^{i+1}, {\sigma^2}^{i+1}, \gamma^{i+1}, y)$;

with $i \in \{1, 2, ..., k\}$ (Adeniyi and Yahya, 2020). The Gibbs sampler requires an initial value for each variable. This could interfere with the sampling over the first iterations resulting in an unreliable approximation of the multivariate distribution. To resolve this, the first b iterations referred to as burn-in are not considered. Only after the burn-in, k new iterations are generated to approximate the multivariate distribution. The BRR should result in a similar estimate of the variance components as rr-gBLUP.

3.5 Bayesian models

In this dissertation, the additive marker effects and variance components of the linear mixed effects model are estimated using rr-gBLUP. Other models like Bayes A, Bayes B, and Bayes C have also been studied (Meuwissen et al., 2001). The three different Bayesian models are implemented using the BGLR package in R (Pérez and de los Campos, 2014).

3.5.1 Bayes A model

The Bayes A model estimates the variance component of each marker (Meuwissen et al., 2001). Compared to rr-gBLUP, the additive marker effects do not have a common variance. Each marker u_i has a variance $\text{var}(u_i) = \sigma^2_{u_i}$. The variance component $\sigma^2_{u_i}$ is sampled from an inverted chi-square distribution $\chi^{-2}(\nu, S)$ with ν the number of degrees of freedom and S a scaling parameter (Wang et al., 1993). The Bayes A model cannot be estimated directly, however, Gibbs sampling can be used.

3.5.2 Bayes B model

In vivo, not all marker loci affect the trait of interest. Therefore, in the Bayes B model, the Bayes A model is extended to account for the possibility that a genetic

marker has a zero effect (Meuwissen et al., 2001). The marker variance is then defined as:

$$\begin{cases} \sigma^2_{u_i} = 0, & \text{, with probability } \pi \\ \sigma^2_{u_i} \sim \chi^{-2}(\nu, S) & \text{, with probability } (1-\pi) \end{cases}. \tag{3.41}$$

Similar as the Bayes A model, the Bayes B model uses a Gibbs sampling algorithm to estimate model parameters, including the additive marker effects.

3.5.3 Bayes C model

The Bayes C model assumes that a fraction (π) of the marker effects has a zero effect while the other fraction ($1-\pi$) follows a multivariate normal distribution with a common marker effects variance $\mathcal{N}(\mathbf{0}, \sigma^2_u \mathbf{I}_k)$ (Lorenz et al., 2010; Habier et al., 2011). The linear mixed effects model is expressed as:

$$y = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}\,, \tag{3.42}$$

The parameter π needs to be estimated. If π is zero, the Bayes C model reduces to rr-gBLUP.

3.6 Prediction accuracy

In genomic selection, the prediction accuracy is often measured as correlation between the GEBVs and the true breeding values (Calus et al., 2008; Rabier et al., 2016). The correlation between the GEBVs ($\hat{\mathbf{g}}$) and the genetic values ($\mathbf{g}$) is calculated by means of the Pearson correlation (ρ):

$$\rho(\hat{\mathbf{g}}, \mathbf{g}) = \frac{\text{cov}(\hat{\mathbf{g}}, \mathbf{g})}{\sigma_{\hat{g}}\sigma_g}\,, \tag{3.43}$$

with $\sigma_{\hat{g}}$ the standard deviation of the GEBVs and σ_g the standard deviation of the genetic values. Because in field experiments, the true breeding values cannot be measured, the prediction accuracy can be estimated as the correlation between the GEBVs and the phenotypic values. In the next chapters, the term prediction accuracy will always refer to the Pearson correlation between the GEBVs and the genetic values.

4

Simulation background

4.1 Introduction

To study the long-term effects of a design choice in a breeding population, years of consecutive breeding are required. By simulating a breeding population, the long-term effects can be estimated and studied. The infinitesimal model is used to simulate a breeding population, combining both the biological and mathematical backgrounds. The backbone structure of the simulator was developed by Neyhart et al. (2017), using the packages `GSSimTPUpdate` and `hypred` in R (version 3.6.3). Two datasets of North American barley (Hordeum *vulgare*) from the University of Minnesota (UMN) and the University of North Dakota (NDSU) are used to construct a base population. The UMN and NDSU dataset contain the genotype of 1590 SNP markers coded as -1, 0, and 1 for 384 and 380 individuals, respectively. For each marker, the marker position, as well as the genotype is available.

4.2 Experimental design

Two types of user inputs can be distinguished: inputs that control the crop characteristics (e.g. heritability, number of QTLs, etc.) and inputs that influence the design of the breeding population (e.g. parental selection intensity, individuals per cross, etc.). The heritability (h^2) and the number of QTLs are normally not within the control of the breeder but can be used to simulate a specific breeding population. In most crops, the exact number of QTLs is still unknown. GWAS is used to detect and identify these QTLs, but depending on the quality and quantity of the available data, the estimation of QTLs could be misleading. In spring and winter barley, 217 and 143 QTLs are detected in the genome, respectively (Xu et al., 2018). The number of QTLs controlling a single trait ranges between 29 QTLs for leaf length (Du et al., 2019) and 66 QTLs for grain length (Xu et al., 2018). The selection of barley often relies on different traits like grain mass, grain length, number of spikes, etc. The heritability of different traits for barley is listed in Table 2.1. Normally, 100 QTLs and a heritability of 0.5 are used to simulate a breeding population of barley. According to the literature, both values are a realistic representation for the genomic characteristics of barley and have already been used in a similar simulation study (Neyhart et al., 2017). However, the heritability and the number of QTLs can be underestimated and thus, other values for the heritability and the number of QTLs will also be used in this dissertation.

To understand how the number of QTLs and the heritability can influence the genetic value of the breeding population, a simulation was performed using truncation selection, selecting the individuals with the highest GEBVs and crossing them randomly. The impact of the number of QTLs on the genetic value is shown in the bottom left panel of Figure 4.1. When a lower number of QTLs are present, higher genetic gains are observed in the long term. Assuming that the number of markers remains the same, when the number of QTLs decreases, more markers are available to grasp the genetic variation, allowing for a better estimation of the QTL effects. When truncation selection is used, selecting individuals with the highest GEBV (or phenotype) as parents, often results in the loss of favorable QTL alleles (see Chapter 5). Regardless if the parental selection is based on the GEBVs or the phenotype, parents are chosen based on a single value. Therefore, a small-effect QTL can be masked by the presence of other QTLs, resulting in the loss of that small-effect QTL. This is often the case when a high number of QTLs are present, and results in a lower long-term genetic gain. Moreover, a higher number of QTLs increase the probability that a favorable QTL allele is in LD with an unfavorable QTL allele, reducing the maximum reachable genetic value.

The impact of the heritability on the genetic gain is shown in the top left panel of Figure 4.1. The heritability represents the fraction of the phenotypic variation that can be explained by the genotype. A trait with a high heritability will hardly

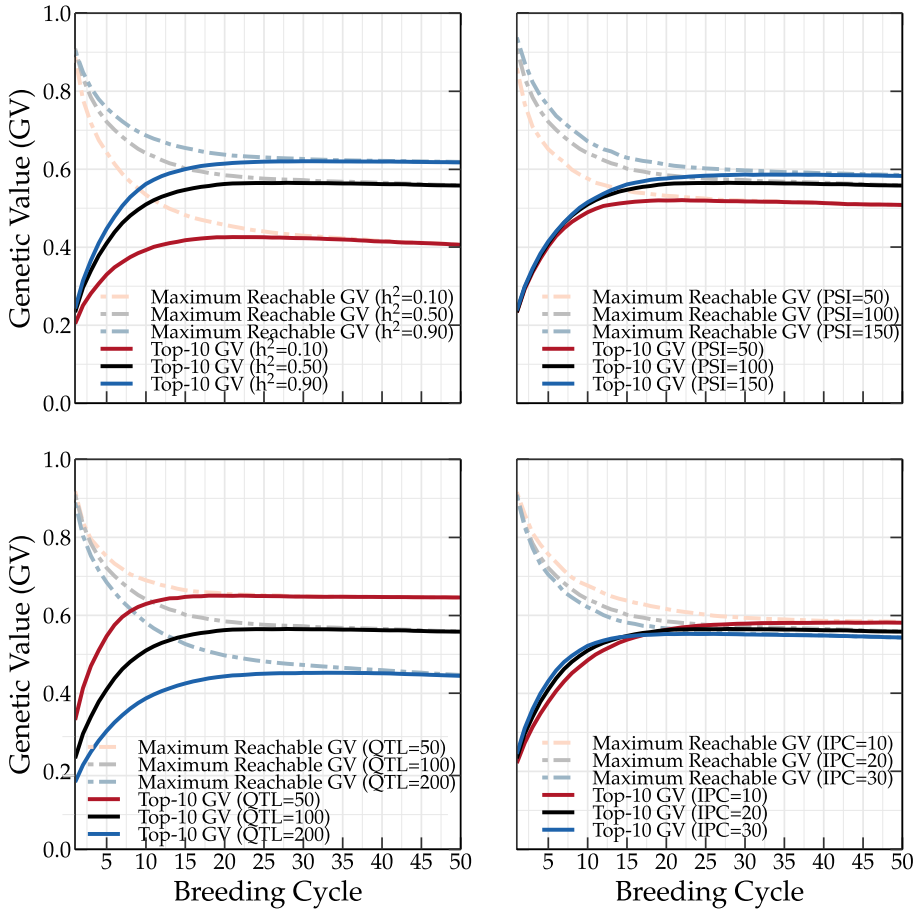

Figure 4.1: Mean genetic value and maximum reachable genetic value for different heritabilities (top left), parental selection intensities (PSI) (top right), QTLs (bottom left) and individuals per cross (IPC) (bottom right).

be influenced by the environment and the phenotype or GEBVs can easily be used to select individuals containing favorable QTL alleles. Therefore, when truncation selection is used, high genetic gains will rapidly be obtained. However, when a trait has a low heritability, the selection of individuals containing favorable QTL alleles becomes more tedious. An individual with a low genetic value may have a high phenotypic value due to the impact of the environment, decreasing the genetic progress in the next generations. In other words, a low heritability will result in a lower genetic gain, both in the short and the long term.

The parental selection intensity (PSI) and the number of individuals per cross (IPC) define the size of the breeding population. A default value of 100 and 20 is set for PSI and IPC, respectively, resulting in a breeding population of 1000 individuals. The impact of the PSI and IPC on the mean genetic value of the top-10 individuals is shown in respectively the top right panel and bottom right panel of Figure 4.1. The PSI represents the number of parents that are selected in each breeding cycle, controlling the amount of genetic variation that can be passed from the breeding population to the next generation. By selecting a low number of parents, only a small amount of genetic variation can be passed to the next generation resulting in a rapid fixation of the QTL alleles and a premature convergence of the genetic value. This can be avoided by increased the number of parents, but it will also increase the financial cost of the breeding population. The IPC represents the number of offspring that is generated between two parents. Increasing the IPC results in different recombinations between the genetic information of both parents. Compared to PSI, increasing the IPC decreases the genetic gain in the long term. When each pair of parents produces more progeny, the probability that closely related individuals are selected as a parent in the next breeding cycle increases. This leads to the loss of genetic variation causing a premature convergence of the genetic value.

The choice of the prediction model, the size of the training population, and the training population update method will also have major repercussions on the prediction accuracy. Increasing the size of the training population will increase the prediction accuracy to a certain extent, but it will also increase the financial cost for phenotyping and genotyping such a population. Therefore, the training population size is kept constant in the remainder of this work, but different methods to update the training population and different prediction models will be used in Part 2.

4.3 Genome construction

A genome of barley is constructed based on two datasets containing marker information and haplotype information of 1590 biallelic SNP loci coded as 0 and 1.

Based on the marker information, a genetic map is constructed containing the marker position for each chromosome. Next, 100 QTLs ($L = 100$) are selected randomly from the available 1590 biallelic SNP loci. The remaining 1490 biallelic SNP loci are available as markers for prediction and selection purposes. The QTL effects are calculated according to a geometric series. At the k-th QTL, the favorable homozygote has a value of a^k, the heterozygote a zero value and the unfavorable homozygote a value of $-a^k$ with $a = (L-1)/(L+1)$ (Bernardo, 2009). The favorable QTL alleles are randomly divided between the first (haplotype = 0) and second (haplotype = 1) allele. Dominance and epistatic effects are assumed to be absent, but could be incorporated if required. In total, 100 different genomes are constructed and are used for 100 simulation runs for each parental selection method. Each genome has the same number of QTLs and the same QTL effects, but because the QTL locations are randomly sampled, each genome will result in a unique scenario. An overview of the simulator is given in Figure 4.2.

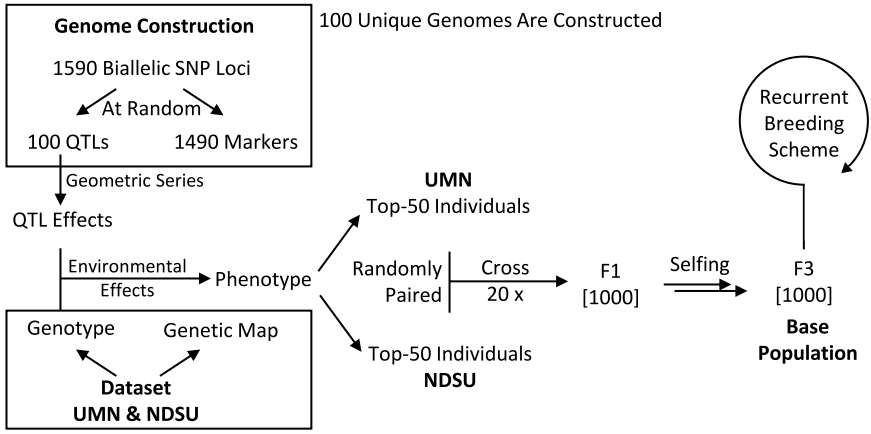

Figure 4.2: Schematic overview of the simulator based on the UMN and NDSU datasets. Both datasets contain the genotype of each individual and a genetic map with the relative location of each SNP loci in the genome. A genome is constructed by randomly selecting 100 SNP loci as QTL. The remaining 1490 SNPs serve as markers. Based on the genotype of these 100 QTLs, the phenotype of each individual is calculated. From both datasets, 50 individuals with the highest phenotype are selected and randomly crossed with each other. The base population is obtained after two generations of single-seed descent.

4.4 Construction of the breeding population

The base population is constructed based on the haplotype information of both datasets. The haplotype is represented in a matrix of size $2n \times k$ with n the number of individuals and k the number of markers. The haplotype of an individual i is represented at rows $2i-1$ and $2i$ and is coded as 0 and 1, representing which SNP allele is present at each marker location. The two haplotypes of an individual

i can be converted into its genotype $\mathbf{Z}_i$ following the conversion rules shown in Table 4.1. Simulation studies that use a different coding for the genotype (like -0.5, 0, 0.5 and 0, 1, 2) will yield the same results. However, if the linear mixed effects model is extended with a polynomial of the genotype (e.g. $\mathbf{Z}^2$), then the coding will influence the simulation results and methods using a different coding should not be compared with each other (Martini et al., 2019). Both the genetic and

Haplotype 1	0	1	0	1
Haplotype 2	0	0	1	1
Genotype	-1	0	0	1

Table 4.1: Conversion of the haplotype coding to the genotype coding of an individual.

phenotypic values are calculated based on the genotypic information. The genetic value of the i-th individual (g_i) is obtained by summing the QTL effects present in the genome. The phenotypic value of the i-th individual (y_i) is calculated as follows:

$$y_i = \frac{1}{3}\sum_{j=1}^{3}(g_i + e_j + \epsilon_{ij}), \tag{4.1}$$

with e_j the j-th environmental effect and ϵ_{ij} the residual effect of the i-th individual and the j-th environment. The environmental effect is averaged over three different environments ($j = 3$) drawn from a normal distribution with mean 0 and a variance component σ_E^2 that is set to eight times the genetic variance (Bernardo, 2014). The residual effect is drawn from a normal distribution with mean 0 and a variance component σ_e^2:

$$\sigma_e^2 = n_e(\sigma_E^2/h^2) - \sigma_E^2, \tag{4.2}$$

with n_e the number of environments. The residual variance component is scaled to simulate a population with a heritability of 0.5.

To create the base population, the top-50 individuals of the UMN dataset are randomly paired with the top-50 individuals of the NDSU dataset. Only at this stage of the simulation, the parental selection uses phenotypic values to guide the parental selection.

4.5 Simulation of a breeding cycle

A new breeding cycle starts with the selection of 100 parents according to one of the parental selection methods that will be proposed in Part 2. The selected parents are paired and each couple produces 20 offspring resulting in 1000 F1 hybrids. After two generations of single-seed descent 1000 F3 individuals are obtained. These individuals form the new breeding population.

Different methods to select and cross parents will be proposed in Part 2 and form the backbone of this dissertation. Each of our parental selection methods tries to preserve the genetic variation and maximize the genetic gain in the long term while preserving the short-term genetic gain. Our methods will be compared with other approaches such as truncation selection in which 100 individuals with the highest GEBVs are selected and crossed randomly.

To simulate a parental cross, a gamete is constructed for each parent by recombining both haplotypes. At each marker, the parent has two alleles that together define the genotype of that parent at that marker. To simulate the recombination, the haplotype information of one of the two homologous chromosomes is copied. Then, at a certain location, a crossing over is simulated, copying the haplotype information of the other homologous chromosome until the next crossing over. The number of crossing overs is Poisson distributed with $\lambda = l$ and l the length of the chromosome in Morgan. The locations where the crossing overs will occur are sampled uniformly at random. An overview of the gamete construction is given in Figure 2.6. The gametes of both parents will become the new genotype of that offspring. In case of single-seed descent, the same mechanism is used to create the offspring, but now, both gametes will originate from the same parent.

Several variables are calculated to track the progress of the breeding population during simulation. The genetic relationship matrix $\mathbf{G}$ is calculated as follows (VanRaden, 2008):

$$\mathbf{G} = \frac{\mathbf{MM'}}{2\sum_{i=1}^{k} P_i(1-P_i)}, \tag{4.3}$$

with $\mathbf{M}$ a matrix with n rows and k columns of which each column is calculated as $\mathbf{Z}_i - \mathbf{1}_n[2(P_i - 0.5)]$, n the number of individuals in the breeding population, $\mathbf{Z}_i$ the genotype of n individuals at the i-th marker, $\mathbf{1}_n$ a vector of size n containing ones, k the number of markers, and P_i the frequency of the second allele at the i-th marker. The averaged inbreeding coefficient F is calculated as:

$$F = \frac{1}{n}\sum_{i=1}^{n} G_{ii} - 1, \tag{4.4}$$

with G_{ii} the diagonal elements of the genetic relationship matrix. To track the fixation of the different QTL alleles, new variables are introduced. The maximum genetic value is the sum of the favorable QTL effects. The fixed genetic value is the sum of the QTL effects that are fixed. The maximum reachable genetic value is the sum of the QTL effects that are fixed (both favorable and unfavorable) and the sum of the favorable QTL effects that are not yet fixed. It represents the maximum genetic value that could still be reached, taking into account the fixation of unfavorable QTL alleles. All these variables are converted into percentages, where the maximum genetic value of 1 can only be achieved if all favorable QTL

alleles are present.

Before selecting the next parents, GEBVs are predicted using the `rrBLUP` (Endelman, 2011) package or `BGLR` package (Pérez and de los Campos, 2014) in R. To do so, a phenotypic and genotypic information is required. In the first breeding cycle, the base population is used as training population. Once the GEBVs are predicted, the parents are selected according to one of the described parental selection methods and paired to construct the crossing block. Both the phenotypic and genotypic values are calculated and used to track the progress of the genetic gain throughout the different breeding cycles. Based on the GEBVs, in each breeding cycle, the training population (TP) is updated according to the *tails* method, selecting the bottom-75 and top-75 individuals while the oldest 150 lines are eliminated from the training population (Neyhart et al., 2017).

4.6 Prediction model and training population

The parental selection schemes are based on GEBVs that are obtained by fitting a linear mixed effects model:

$$\mathbf{y} = \mathbf{1}_n\beta + \mathbf{Z}_{\text{TP}}\mathbf{u} + \boldsymbol{\epsilon}\,, \tag{4.5}$$

with $\mathbf{y}$ a vector with phenotypic values, n the number of individuals in the training population, $\mathbf{1}_n$ a vector of size n containing ones, β the fixed effect (phenotypic mean), $\mathbf{Z}_{\text{TP}}$ the incidence matrix of the training population with marker information (coded as -1, 0 and 1), $\mathbf{u}$ the marker effects following a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{G})$ with $\mathbf{G} = \sigma_u^2\mathbf{I}_k$ (with $\mathbf{I}_k$ the identity matrix of dimension k) and $\boldsymbol{\epsilon}$ the residual effects following a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \sigma_e^2\mathbf{I}_n$. Both variance components σ_u^2 and σ_e^2 are estimated by means of restricted maximum likelihood (REML). The GEBVs of the individuals are calculated as:

$$\hat{\mathbf{g}} = \mathbf{Z}\hat{\mathbf{u}}\,, \tag{4.6}$$

with $\hat{\mathbf{g}}$ the GEBVs, $\mathbf{Z}$ the marker information (coded as -1, 0, and 1) and $\hat{\mathbf{u}}$ the predicted marker effects. Assuming that the phenotypic data of the entire base population is available, it can be used to construct the initial TP with a total size of 764 individuals. During subsequent breeding cycles, the TP is updated with 150 new individuals selected from the breeding population, limiting the required phenotyping effort per cycle to only 150 individuals. The 150 individuals that have been the longest in the TP are removed keeping the size of the TP constant. The removal of old lines in the TP does not affect the prediction accuracy significantly, but reduces the required computation time. The selection of 150 new individuals during the TP update is done using the tails method (Neyhart et al., 2017). In

Chapter 5, markers with a minor allele frequency smaller than 0.03 are removed. After reevaluation, we demonstrated that removing low-frequency markers from the training population resulted in a lower genetic gain. Therefore, in the subsequent chapters, all markers are used to fit the linear mixed effects model.

Although the training population is updated by using the tails method, other methods to update the training population will also be studied. The training population can also be updated selecting 150 individuals with the highest GEBVs (*top* method), with the lowest GEBVs (*bottom* method) or the individuals can be selected at random (*random* method) (Neyhart et al., 2017). Individuals can also be selected to minimize the prediction error variance (*PEVmean* method) or by maximizing the expected reliability of the predictions (*CDmean* method) (Rincent et al., 2012). The PEVmean method selects individuals in the training population that minimize the prediction error variance (PEV):

$$\text{PEV} = \text{var}(\hat{\mathbf{u}} - \mathbf{u}). \tag{4.7}$$

The PEV can be calculated using the MME equation Eq. (3.25) in combination with a contrast matrix $\mathbf{c}$ and a design matrix $\mathbf{M}$:

$$\text{PEV} = \text{diag}\left[\frac{\mathbf{c}'(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \delta\mathbf{G})\mathbf{c}}{\mathbf{c}'\mathbf{c}}\right] \times \sigma_e^2, \tag{4.8}$$

with $\mathbf{M}$ defined as $\mathbf{M} = \mathbf{I}_t - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, $(\mathbf{X}'\mathbf{X})^{-}$ the generalized inverse of $\mathbf{X}'\mathbf{X}$, t the number of individuals that are added to the training population, $\delta = \sigma_e^2/\sigma_u^2$, $\mathbf{G}$ the genetic relationship matrix, and $\mathbf{I}_t$ the identity matrix of size t (Rincent et al., 2012; Maenhout et al., 2010; Laloë, 1993; Neyhart et al., 2017). Complementary to PEV, the generalized coefficient of determination (CD) can be used to construct a training population that maximizes the expected reliability of the contrast. The CD is defined as:

$$\text{CD} = \text{diag}\left[\frac{\mathbf{c}'(\mathbf{G} - \delta(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \delta\mathbf{G}^{-1})^{-1})\mathbf{c}}{\mathbf{c}'\mathbf{G}\mathbf{c}}\right], \tag{4.9}$$

and ranges between 0 and 1, indicating whether the predictions are unreliable (CD $\sim$ 0) or reliable (CD $\sim$ 1).

Systematically updating the training population will increase the financial cost, but if the training population remains unchanged, the prediction accuracy will degrade over time causing low prediction accuracies and ultimately resulting in low genetic gains (Neyhart et al., 2017). Therefore, updating the training population should always be considered.

An overview of the parameter values used in the simulation study is listed in Table 4.2. In Part 2, the simulator is used in each chapter to simulate a breeding

Parameter	Value
Genome	Barley
Number of chromosomes	7
Number of QTLs	100
Number of markers	1490
Heritability	0.5
Population size	1000
Number of breeding cycles	50
Number of iterations	100
Parental selection	100
Individuals per cross	20
Number of environments	3
Prediction model	rr-gBLUP
Update training population	150
training population update method	Tails

Table 4.2: Overview of the parameters used the simulate a breeding population of Barley over 50 breeding cycles.

population over 50 breeding cycles. Specific modification of the simulator will be reported per chapter if required.

PART II

METHODS TO PRESERVE THE GENETIC VARIATION IN BREEDING PROGRAMS

5

The scoping method

Genomic selection has been successfully implemented in plant and animal breeding. The transition of parental selection based on phenotypic characteristics to genomic selection (GS) has reduced breeding time and cost while accelerating the rate of genetic progress. Although breeding methods have been adapted to include genomic selection, parental selection often involves *truncation* selection, selecting the individuals with the highest genomic estimated breeding values (GEBVs) in the hope that favorable properties will be passed to their offspring. This ensures genetic progress and delivers offspring with high genetic values. However, several favorable quantitative trait loci (QTLs) alleles risk being eliminated from the breeding population during breeding. We show that this could reduce the mean genetic value that the breeding population could reach in the long term with up to 40%. In this chapter, by means of a simulation study, we propose a new method for parental mating that is able to preserve the genetic variation in the breeding population, preventing premature convergence of the genetic values to a local optimum, thus maximizing the genetic values in the long term. We do not only prevent the fixation of several unfavorable QTL alleles, but also demonstrate that the genetic values can be increased by up to 15 percentage points compared with truncation selection.

The material of this chapter is based on the following publication:
Vanavermaete, D., Fostier, J., Maenhout, S., De Baets, B., 2020. Preservation of genetic variation in a breeding population for long-term genetic gain. G3, 10(8), 2753–2762.

5.1 Introduction

In times of climate change and rapid population growth, new methods need to be developed to further improve different crop properties like yield and resistance to pathogens and drought (Tester and Landridge, 2010). These properties are controlled by different chromosomal regions or quantitative trait loci (QTLs), making it difficult to improve crop properties by only relying on phenotypic characteristics (Dekkers and Hospital, 2002). Initially, pedigree information was used to guide the selection of parental lines in animal and plant breeding. Nowadays, molecular markers like single nucleotide polymorphisms (SNPs) serve as proxies for QTLs, assuming that markers are in strong linkage disequilibrium with one or more QTLs (de Roos et al., 2008). The linear relationship between the genetic markers (genotype) and the phenotype can then be estimated using a mixed effects model. This concept was first introduced in marker-assisted selection (MAS), but only minor improvements in yield were reported (Goddard and Hayes, 2002). Genomic selection was introduced as an alternative for MAS (Meuwissen et al., 2001). By using markers that cover the complete genome, the fraction of the genetic variance that can be explained by the molecular markers was better captured, leading to an improved estimation of large and small QTL effects (Heffner et al., 2009, 2010; Beyene et al., 2015). Genomic selection improved yield in animal and plant breeding and reduced the time in between breeding cycles (Hayes et al., 2009). For example, crops like oil palm (Elaeis *guineensis Jacq.*) reach sexual maturity after three years but require 13 to 15 years before phenotypic characteristics can be obtained: the transition of phenotypic selection to genomic selection reduced the time of one breeding cycle from 15 to three years (Cros et al., 2018). In the course of time, genomic selection has further evolved and has become a powerful tool in animal and plant breeding (Meuwissen et al., 2001; Bernardo and Yu, 2007; Crossa et al., 2010). Over the last years, several advancements were achieved ranging from yield maximization to the development of new drought/heat-resistant plants (Wang et al., 2019; Sun et al., 2019; Suontama et al., 2019). Nevertheless, the implementation of genomic selection in certain breeding populations with complex traits and environmental interactions is still challenging (Juliana et al., 2018; Voss-Fels et al., 2018).

Several simulation studies on genomic selection have resulted in high prediction accuracies and genetic values in the short term (VanRaden et al., 2009; Hayes

et al., 2009). These studies often rely on *truncation* selection of the parents, leading to a high genetic gain in the short term but the loss of favorable QTL alleles, genetic variation and prediction accuracy over time (Jannink, 2010). Truncation selection selects the top fraction of the individuals based on their genomic estimated breeding values (GEBVs), which serve as estimators for the true breeding values. Because the GEBVs are calculated as the sum of the estimated additive marker effects, the contribution of favorable small-effect QTLs can be concealed leading to their loss in the breeding population, thus reducing long-term genetic gain. The loss of those favorable QTL alleles could be reduced by weighting the marker effects of favorable low-frequency alleles more heavily, thereby safeguarding long-term gain (Jannink, 2010; Liu et al., 2015). In recent years, different parental methods have been developed that aim to reduce the loss in genetic variation. This helps to increase the prediction accuracy and the genetic gain in the long term. To preserve genetic variation, the selection of closely related individuals should be avoided (Lindgren and Mullin, 1997) or the inbreeding coefficient should be minimized (Brisbane and Gibson, 1995). Although genomic selection uses GEBVs for parental selection, alternative score functions to guide the parental selection have been proposed. Daetwyler et al. (2015) proposed a parental selection scheme based on the genomic optimal haploid value, selecting parents that optimize the genetic values of their offspring. This method was further improved by simulating the meiosis between parental haploids, yielding an improved prediction of offspring. This, in turn, leads to a more accurate evaluation of the double haploids, thereby guiding the parental selection to further increase long-term genetic gain (Müller et al., 2018). In an alternative approach, Lehermeier et al. (2017) proposed the *criterion of usefulness*, which takes into account the selection intensity, mean genetic value and genetic variance of the breeding population, improving the long-term genetic gain.

Over the last years, new mating designs have been proposed to further improve the parental selection and maximize the genetic gain in the short or long term. In a new mating design, the genetic variation is preserved by penalizing crosses between two parents with high coancestry (Cervantes et al., 2016). Moreover, long-term gain was further improved by also minimizing the rate of inbreeding and controlling the allele heterozygosity and allele diversity (Akdemir and Sánchez, 2016). The introduction of an optimal mating design using a two-part plant breeding selection with rapid recurrent genomic selection reduced the drop in genetic diversity, thus maximizing the conversion of genetic variance into genetic gain (Gorjanc et al., 2018).

Although parental selection methods play a major role in the realization of long-term genetic gain, as long as those methods are based on GEBVs, the results will be influenced by the choice of the prediction model and the training population design. Several training population designs have been proposed although no significant difference was observed in the long term, as long as the training popu-

lation was systematically updated over time (Akdemir et al., 2015; Rincent et al., 2012; Neyhart et al., 2017).

In this chapter, the *scoping* method is presented as a new parental mating scheme to reduce the loss of favorable QTL alleles by preserving the genetic variation and thus maximizing the genetic value in the long term. The scoping method combines genetic progress (truncation selection) and the preservation of the genetic variation of each marker in the breeding population. Based on the observation that two closely related individuals might contain a different rare marker allele, both individuals should be selected to preserve the genetic variation of both markers in the breeding population. Therefore, in contrast to other methods, the genetic relationship or inbreeding coefficient is not taken into account, but individuals are selected based on their genotype, ensuring the maximal selection of the different marker alleles and thus maximizing the genetic variance of their offspring. By doing so, we reduce the risk of premature convergence of the genetic values to a local optimum. Combined with truncation selection, the genetic progress is ensured in the short as well as in the long term. We benchmark our proposed scoping method against two existing selection strategies: the *population merit* method (Lindgren and Mullin, 1997) and the *maximum variance total* method (Cervantes et al., 2016). Both methods try to maximize long-term genetic gain by preserving the genetic variation of the breeding population. Whereas the scoping method preserves the genetic variation by maximizing the variation of each marker, the population merit method preserves the genetic variation by minimizing the average genetic relationship of the parental population. Both the population merit method method and the maximum variance total method aim to maximize the genetic variation of the parental population, and thus are good candidates against which our proposed scoping method can be benchmarked.

We also propose the backcrossing method. This method combines truncation selection with the reintroduction of new genetic information to help preserve the genetic variation of the breeding population. Reintroducing new genetic information could suddenly reduce the genetic gain of the breeding population. Therefore, the backcrossing method is only used to study to which extent reintroducing genetic variation can maximize the genetic gain in the long term.

5.2 Materials and methods

We adopt the base population and breeding scheme of Neyhart et al. (2017), making it possible to compare our results with truncation selection as reported by Neyhart et al. (2017). The base population consists of two datasets of North American barley (Hordeum *vulgare*) from the University of Minnesota (UMN) and the University of North Dakota (NDSU) counting respectively 384 and 380 six-row spring

inbred lines with 1590 biallelic SNP loci. Recurrent selection is applied to the base population to simulate the later breeding cycles (see Subsection 5.2.1).

The scoping method, in which a parental selection method is combined with a new mating design, is proposed and compared with truncation selection with random mating. The scoping method tries to maximize the genetic values in the long term, while preserving the genetic variation of the breeding population. It aims to avoid the loss of positive-effect QTL alleles, preventing the convergence of the genetic values to a local optimum. Because this method might select extreme GEBVs, the Pearson correlation cannot be used to evaluate the selection method due to its sensitivity to outliers. Instead, the mean genetic value of the breeding population, calculated on the basis of the true breeding values, is used to measure and evaluate the genetic gain for each method. The mean genetic value of the top-10 individuals is also reported. Our method aims to maximize the genetic gain of the top-10 individuals, while the remaining individuals of the breeding population serve to preserve important genetic marker alleles in the breeding population.

5.2.1 Breeding scheme

The recurrent selection scheme is illustrated in Figure 5.1. Starting with 100 individuals, a crossing block is constructed, pairing the selected individuals. Each couple produces 20 offspring resulting in a total of 1000 F1 hybrids. After two generations of single-seed descent, 1000 F3 individuals are obtained. These individuals form the new breeding population from which again 100 parents are selected to start a new breeding cycle. This selection occurs either according to the *baseline*, scoping, backcrossing, population merit or maximum variance total methods. The first breeding block (in breeding cycle zero) pairs 50 individuals of the NDSU dataset with 50 individuals of the UMN dataset with the highest phenotypic value, regardless of the parental selection method. This design choice ensures that each parental selection method has the same number of individuals in the breeding population over each breeding cycle. The subsequent parental selections are fully based on GEBVs, reducing the financial cost of phenotyping. A linear mixed effects model is used to obtain GEBVs from molecular marker scores (see Subsection 4.6). Each simulation consists of 50 breeding cycles. This number was specifically selected as it allowed to compare and visualize the converging behavior of each examined method. All results are averaged over 250 simulation runs.

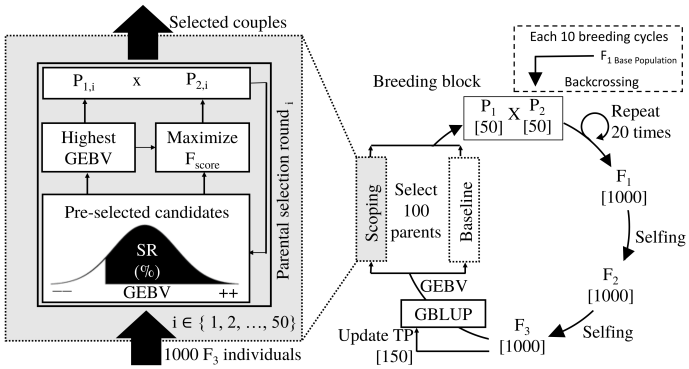

Figure 5.1: Overview of the recurrent selection scheme. First, 50 couples of parents (P_1, P_2) each produce 20 offspring yielding a total of 1000 F1 hybrids. Then, after two generations of single-seed descent, 1000 F3 individuals are obtained. From those F3 individuals, new parental lines are selected. Three different parental selection methods are considered: i) the baseline method selects 100 parents with the highest GEBVs (truncation selection); ii) the scoping method combines the selection of 50 parents (P_1) with the highest GEBVs and 50 parents (P_2) that maximize the genetic variation (see Eq. (5.1)); iii) the backcrossing method selects every tenth breeding cycle the P_2 parents from the base population. After the parental selection, the TP is updated according to the tails method.

5.2.2 The baseline method

The baseline method selects 100 parents with the highest GEBVs (truncation selection) and pairs them randomly. The idea is that favorable properties will be passed on to the next offspring, leading to high short-term gain and rapid fixation of favorable QTL alleles. However, several favorable QTL alleles will be eliminated from the breeding population during breeding, reducing long-term gain and causing the convergence of the genetic values to a local optimum.

5.2.3 The backcrossing method

The backcrossing method introduces new genetic material into the population at fixed points in time. Several favorable QTL alleles that might have been eliminated during earlier breeding cycles are reintroduced in the breeding population, making it possible to escape from a local optimum and improve the genetic values in the long term. The backcrossing method is identical to the baseline method except for the parental selection at every tenth breeding cycle. At this point, 50 parents with the highest GEBVs are randomly crossed with 50 randomly chosen individuals from the F1 hybrids of the original base population. The F1 hybrids of the base population should show the highest degree of heterozygosity and thus reintroduce the highest amount of genetic variation into the next breeding population. Moreover, by randomly selecting F1 hybrids, each hybrid has the same probability to be selected as a parent, making it possible to reintroduce important unknown QTL

alleles in the current breeding population. We expect a drop in the mean genetic value after each backcrossing event, but a higher genetic gain might be obtained in the long term.

5.2.4 The scoping method

The scoping method continuously preserves genetic variation, avoiding premature convergence to a local optimum, while ensuring a gradual increase of genetic values over breeding cycles. The parental selection is split into two parts: the pre-selection and the selection. First, a fraction of the breeding population with the highest GEBVs is pre-selected using truncation selection. This fraction, referred to as the scoping rate (SR), can take a value between 0.1 and 1. A scoping rate of 0.1 pre-selects 100 individuals (10%) of the breeding population, whereas a scoping rate of 1 pre-selects the entire breeding population (100%). During the selection, 100 different parents are chosen from the pre-selected population. In contrast to the baseline method, parents are not paired randomly. From the pre-selected individuals, the one with the highest GEBV is chosen as the first parent. The second parent is chosen from the pre-selected individuals in such a way that the genetic variation of selected parents is maximized over each marker. Mathematically, the following score function is maximized:

$$F_{\text{score}} = \sum_{i=1}^{k} \text{var}(\mathbf{Z}_i) p_i \,, \tag{5.1}$$

with k the number of markers, $\mathbf{Z}_i$ the i-th row of the $k \times n$ matrix $\mathbf{Z}$ containing the genotypes (coded as -1, 0 and 1) of the n already selected individuals and where $\mathbf{p}$ denotes a vector of k Boolean values. Initially, p_i is set to a value of 1 for all marker positions. When both alleles at marker i are present, p_i is set to 0. Thereby, the score function will maximize the variance of the genotype over each marker for which both alleles are not yet present in the selected population, thus avoiding the loss of low-frequency marker alleles. If p_i equals 0 for all markers, the value of each p_i is restored to 1, again maximizing the variance over all the markers. At this moment, all the available marker alleles of the current breeding population are present in the selected parental population.

The scoping method combines truncation selection with a new mating design, pairing individuals with high GEBVs with individuals that maximize the genetic variation of their offspring. The pre-selection process avoids that individuals with lower GEBVs, which might maximize the genetic variation of certain parents, are not available for selection and thus avoids the loss of genetic gain in the short term. We expect that the mating design will reduce the loss of marker alleles while the

pre-selection will eliminate unfavorable QTL alleles over time. This should lead to a slower but more accurate fixation of the favorable QTL alleles.

5.2.5 The combined method

The backcrossing method uses the baseline method to design the crossing block between each backcrossing occurring every tenth breeding cycle. In this combined method, the scoping method is used to select the parents between every backcrossing (breeding cycles 1-9). We expect that the scoping method will reduce the loss in genetic variation caused by the baseline method, while the backcrossing will further increase the genetic variation and thereby increase the long-term genetic gain.

5.2.6 The population merit method

The population merit method was introduced by Lindgren and Mullin (1997) and aims to preserve the genetic variation of the breeding population by taking into account the average coancestry of the parental population. Normally, the average coancestry is calculated based on pedigree information. Unfortunately, this information is not available for both datasets. Therefore, the average genetic relationship will be used instead. The population merit B_ω is calculated as:

$$B_\omega = \hat{g}_m - c\phi_\omega\,, \tag{5.2}$$

with $\hat{g}_m$ the mean genetic value of the parental population, c a penalty weight and ϕ_ω the average genetic relationship of the parental population. At each breeding cycle the population merit is maximized. First, 100 individuals are selected using truncation selection. Second, the mean genetic value of the parental population and the average genetic relationship are calculated. Third, the population merit is maximized iteratively by replacing each parent with another individual of the breeding population that increases the population merit. To do so, the mean genetic value of the parental population and the average genetic relationship have to be recalculated each time. The population merit is maximized when the parental population remains unchanged.

5.2.7 The maximum variance total method

The maximum variance total (MVT) method aims to maximize the genetic variance of the breeding population (Cervantes et al., 2016). The method was developed by

Bennewitz and Meuwissen (2005) and further modified by Cervantes and Meuwissen (2011). The genetic variance criterion $\text{var}(u_w)$ is calculated as:

$$\text{var}(u_w) = \frac{1}{n}\sum_{i=1}^{n}\left((1+F_i) - 2\overline{\mathbf{G}}_p\right), \tag{5.3}$$

with n the number of selected parents, F_i the inbreeding coefficient of parent i and $\overline{\mathbf{G}}_p$ the average genetic relationship of the parents. Originally, the genetic variance criterion was calculated using the average coancestry, but due to the lack of pedigree information, the average coancestry was replaced with the average genetic relationship. Similar to the population merit method, the genetic variance is maximized iteratively. However, the MVT method does not take into account the genetic value. Therefore, it can only be used in a pre-selected population to guide the final parental selection. The MVT method is used to select the P_2 parents from a pre-selected population similar to the scoping method. First, 300 individuals are pre-selected using truncation selection. Second, from the pre-selected individuals, 100 parents are selected using truncation selection. Finally, the P_2 parents are iteratively replaced such that the genetic variance criterion of the parental population is maximized by only using the pre-selected individuals. We expect a higher long-term gain compared with the baseline method, but a lower genetic gain compared with the scoping method.

The mean genetic value of the breeding population, mean genetic value of the top-10 individuals and the maximum reachable genetic value of all the proposed methods are reported in Table 10.1, Table 10.2 and Table 10.3, respectively.

5.3 Results

5.3.1 The baseline method

The baseline method combines truncation selection with random mating. Our results are similar to those reported by Neyhart et al. (2017). During the first 10 to 20 breeding cycles, we observe a steep increase in genetic value and rapid fixation of QTL alleles (see Figure 5.2). The maximum reachable genetic value is reduced by more than 40%, due to the loss of favorable QTL alleles in the breeding population. It is interesting to also consider the mean genetic value of the 10 individuals with the highest genetic values. Those individuals are of particular interest to breeders for commercialization purposes. Therefore, their genetic value is more important than the mean genetic value of the breeding population. In the baseline method, the top-10 individuals have a higher genetic value over the first breeding cycles, but due to strong fixation, the genetic variation is reduced and

the difference between the top individuals and the breeding population average becomes smaller.

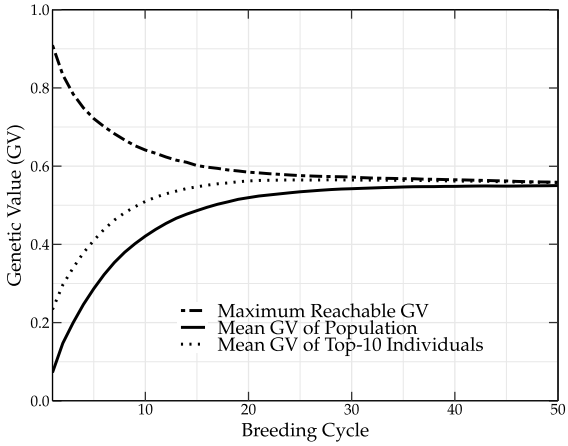

Figure 5.2: Simulation results using the baseline method over 50 breeding cycles. Mean genetic value of the breeding population increases rapidly over the first breeding cycles. The truncation selection, however, causes the loss of several favorable QTL alleles, reducing the maximum reachable genetic value and causing a premature convergence of the genetic value to a local optimum. The top-10 individuals of the population have a higher mean genetic value than the breeding population, but after several breeding cycles, the genetic variation is reduced, closing the gap between the top-10 individuals and the rest of the breeding population.

5.3.2 The backcrossing method

The backcrossing method is organized in periodic blocks of ten breeding cycles (see Figure 5.3). Over the first nine breeding cycles, the results are similar to those of the *baseline* method. In the tenth breeding cycle, the introduction of new individuals causes a drop in the number of fixed QTLs. This indicates that several QTL alleles that had been eliminated from the breeding population over the preceding nine breeding cycles are reintroduced. Obviously, only a fraction of those QTL alleles has a favorable effect on the genetic value. Therefore the inclusion of new individuals causes the mean genetic value to drop. This also causes a decrease in the genetic relationship between the breeding population and the training population (TP), leading to a poor estimation of the GEBVs (see Figure 5.4). Hence, in the next breeding cycle, only a small increase of the genetic value is observed. This problem is resolved after a single TP update, leading to a better estimation of the GEBVs in the subsequent breeding cycles. Similar to the baseline method, the top-10 individuals have a slightly higher genetic value than the average of the breeding population. After each backcrossing, only a small drop in genetic value of the top-10 individuals is observed.

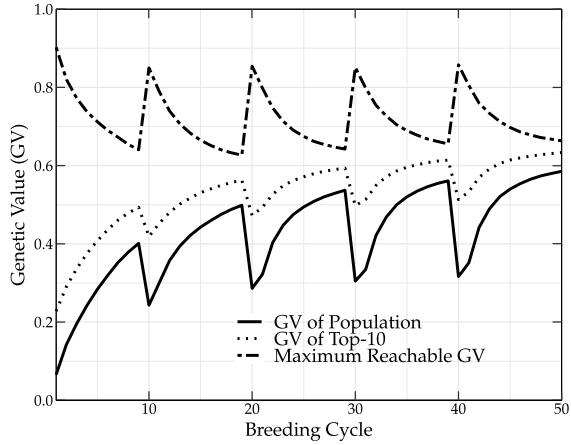

Figure 5.3: Simulation results using the backcrossing method over 50 breeding cycles. At every tenth breeding cycle, when backcrossing occurs, the mean genetic value drops suddenly as a result of the (re-)introduction of genetic material from the base population. This also introduces several favorable QTL alleles that might have been eliminated during earlier breeding cycles, causing the maximum reachable genetic value to increase. The drop in the mean genetic value is relatively modest for the top-10 individuals. Only five breeding cycles are required to recover from the backcrossing event and higher genetic values are reached in subsequent cycles. Over 50 breeding cycles, the breeding population attains a higher genetic value compared to the baseline method.

During the nine breeding cycles that follow each backcrossing event (every tenth cycle), increasingly higher mean genetic values are obtained. This indicates that each time when new individuals are included, favorable QTL alleles are better preserved in the breeding population. This makes it possible to escape from a local optimum and approach the optimal genetic value. After 50 breeding cycles, the backcrossing method yields a genetic value that is 7% points higher than the baseline method. The global optimum is not yet reached and the loss of several favorable QTL alleles causes the genetic value to reach a value of 63% of the global optimum after 50 breeding cycles.

5.3.3 The scoping method

The scoping method introduces the scoping rate (SR) as a new parameter. With the scoping rate, the breeder can control what fraction of the upper tail of the GEBV distribution will be considered for parental selection. Using a small scoping rate, only individuals with high GEBVs will be considered, leading to truncation selection. When a higher scoping rate is used instead, individuals with lower GEBVs will also be considered as candidates, making it possible to preserve the genetic variation of the breeding population. The scoping rate provides the breeder with the option to choose between the maximization of the rate of genetic progress in the short

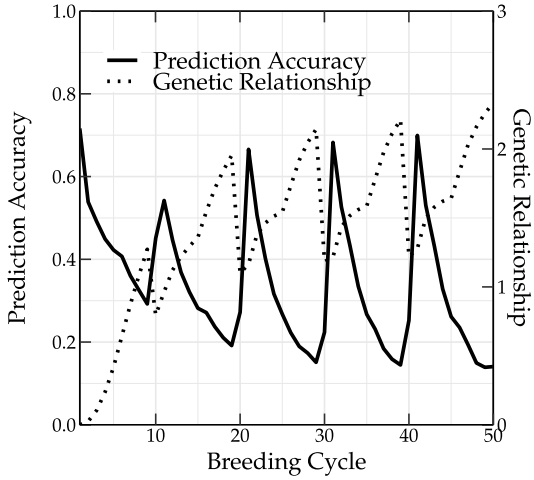

Figure 5.4: Genetic relationship between the breeding population and the training population and the prediction accuracy of the backcrossing method over 50 breeding cycles. At every tenth breeding cycle, when backcrossing occurs, the genetic relationship between the TP and breeding pool decreases, while the prediction accuracy increases. However, the prediction accuracy will only reach a maximum value over the next breeding cycle.

term on the one hand or the maximization of the genetic variation in the long term on the other hand. As expected, the scoping method yields somewhat lower mean genetic values over the first ten breeding cycles (see Figure 5.5). However, the mean genetic value of the top-10 individuals is only slightly lower compared with the baseline method. Certainly, for small scoping rate values (0.1 to 0.3) the difference in genetic value is negligible.

After the tenth breeding cycle, the loss of several favorable QTL alleles causes the baseline method to reach a local optimum, rendering it less efficient than the scoping method. In contrast, by preserving the genetic variation within the breeding population, the scoping method strongly reduces the loss of favorable QTL alleles, thus preserving the potential to reach high genetic values. A higher scoping rate will better prevent the loss of favorable QTL alleles, however, due to a slower increase in genetic value, a high scoping rate will require a longer time before outperforming the baseline method. Therefore, the use of a smaller scoping rate is preferred. It delivers high genetic values in both the short and the long term.

A scoping rate of 0.1 is a special case as it results in the same parental selection as the baseline method, but it uses an alternative mating design to maximize the genetic variation of the offspring. After 50 breeding cycles, this leads to a 4 percentage points higher mean genetic value of the top-10 individuals in favor of the scoping method. This demonstrates that maximizing the genetic variation

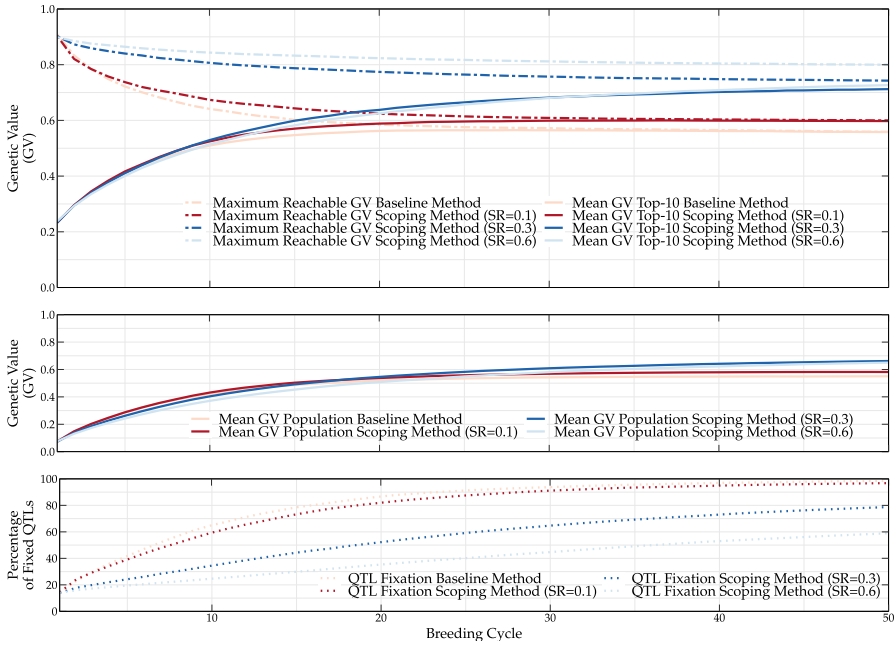

Figure 5.5: Simulation results using the scoping method for a scoping rate of 0.1, 0.3 and 0.6, simulated over 50 breeding cycles. Additionally, the results of the baseline method are shown for the sake of comparison. In the top figure, the mean genetic value of the top-10 individuals and the maximum reachable genetic value are shown for different scoping rate values and the baseline method. In the middle figure, the mean genetic value of the breeding population is shown for different scoping rate values and the baseline method. In the bottom figure, the rate of QTL fixation is shown for different scoping rate values and the baseline method.

increases the genetic value in the long term. A scoping rate of 0.3 yields high genetic values in both the short and the long term. Only eight breeding cycles are needed before the top-10 individuals outperform the baseline method. Over those eight breeding cycles, the difference in genetic value between the baseline method and the scoping method is negligible. After 12 breeding cycles, the mean genetic value of the population surpasses that of the baseline method. Ultimately, after 50 breeding cycles, the scoping method with a scoping rate of 0.3 yields a mean genetic value of 0.71 over the top-10 individuals, a 15 percentage points increase compared with the baseline method.

5.3.4 The combined method

The combination of the scoping and backcrossing methods helps to preserve the genetic variation, leading to a general improvement of the backcrossing method (see Figure 5.6). Although higher genetic values are observed compared to the backcrossing method, the combined method cannot outperform the scoping method.

Moreover, our simulation study indicates that with the proper use of the scoping method, the addition of new genetic information is not required.

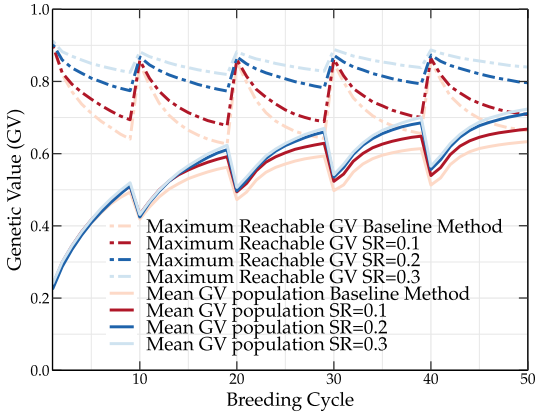

Figure 5.6: Simulation results using the scoping method combined with the backcrossing method simulated over 50 breeding cycles. Replacing the nine breeding cycles between every two backcrossings with the scoping method helps to retain the genetic variation in the breeding population. This leads to a better fixation of the favorable QTL alleles and to a higher maximum reachable genetic value. Moreover, higher mean genetic values are reached compared to the backcrossing method. Further increasing the scoping rate (> 0.6) only leads to negligible improvements of the mean genetic value.

5.3.5 The population merit method

The population merit method preserves the genetic variance by reducing the average genetic relationship of the parental population, leading to a higher genetic gain in the long term compared with the baseline method (see Figure 5.7). Despite the fact that a higher long-term gain is observed, the population merit method only retains a fraction of the genetic variation, still causing the fixation of several unfavorable QTL alleles and a premature convergence of the genetic value to a local optimum. Compared to the scoping method, the same genetic value is observed over the first eight breeding cycles. However, the population merit method causes a strong reduction in the maximum reachable genetic value rendering this method less efficient in the long term than the scoping method. Several values for the penalty weight c were tested and the best results for $c = 20$ are reported. At breeding cycle 50, only an 8 percentage points increase in the genetic value was observed for the population merit method compared with the baseline method, while a 15 percentage points increase in genetic value was observed for the scoping method.

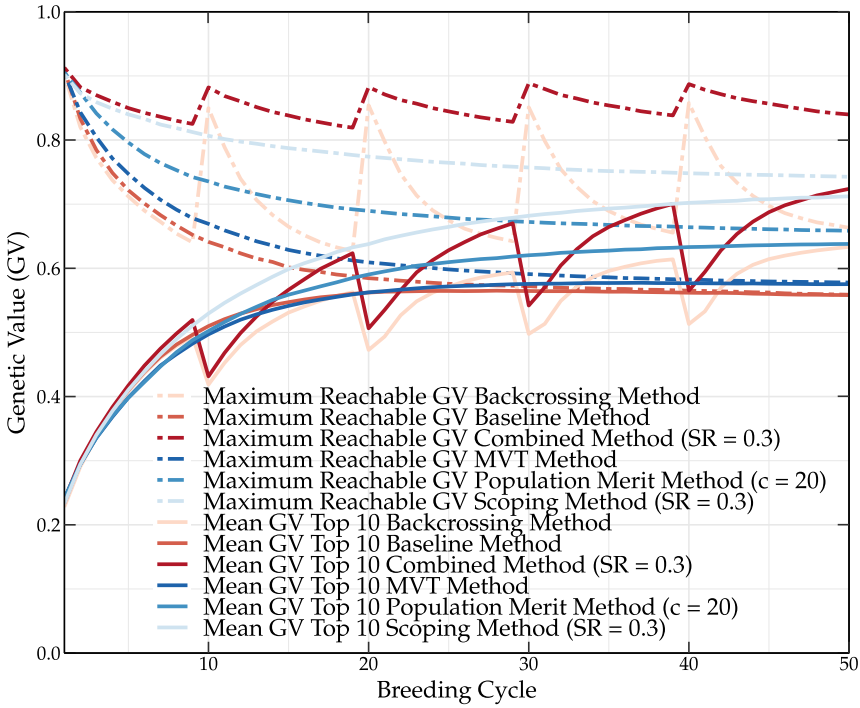

Figure 5.7: Genetic value of the different parental selection methods over 50 breeding cycles. The genetic value in the long term is the lowest when using the baseline method, followed by the maximum value total (MVT) method, population merit method and the scoping method, which delivers the highest genetic values in the long term.

5.3.6 The maximum variance total method

The MVT method combines the average genetic relationship and the average inbreeding coefficient to maximize the genetic variation of the breeding population. This method was used to compare the mating design of the scoping method with the MVT method by using the same pre-selected population to select the P_2 parents. Only a small increase of the genetic gain was observed for the MVT method compared with the baseline method (see Figure 5.7). Using a pre-selected population combined with a truncation selection of the P_1 parents, the MVT method only preserved a small part of the genetic variation compared with the scoping method, causing the loss of several favorable QTL alleles and thus reducing the maximum reachable genetic value. At breeding cycle 50, only a 2 percentage points higher genetic value was observed compared with the baseline method, rendering this method less efficient than the scoping method.

The mean genetic value of the breeding population, mean genetic value of the top-

10 individuals and the maximum reachable genetic value of the baseline method, scoping method, population merit method and MVT method are reported in Table 10.1, Table 10.2, and Table 10.3, respectively. The mean genetic value of the top-10 individuals and the maximum reachable genetic value of the backcrossing method, and combined method are reported in Table 10.4, and Table 10.5, respectively.

5.3.7 Robustness of the scoping method

The robustness of the scoping method has been tested and compared with the baseline, backcrossing, combined, population merit and MVT methods using different genome constructions. We have compared the different methods for a heritability of 0.1, 0.3, 0.7 and 0.9 (see Figure 5.8). Increasing heritability will result in a better prediction of the additive marker effects, leading to an improved parental selection and thus guiding the breeding population towards a higher mean genetic value of the top-10 individuals. Decreasing the heritability will have the opposite effect. As the heritability decreases, the effect of the environment becomes more pronounced, making it more challenging to select parents based on the GEBVs. Independent from the value of the heritability, the relative position between each parental selection method remained the same, indicating that the heritability does not influence the effectiveness of the different parental selection methods. However, a higher heritability will reduce the genetic gain between each method. The genetic value was also studied for 50 QTLs and 200 QTLs (see Figure 5.9). Decreasing the number of QTLs means that the same number of markers is now available to grasp the effects of only 50 QTLs. This leads to a higher prediction accuracy and a higher mean genetic value. When the number of QTLs is increased to 200, the opposite effect is observed. Again, each method can maintain its relative position towards the other methods indicating that the number of QTLs does not influence the effectiveness of the different parental selection methods, except for the combined method which resulted in a slightly higher mean genetic value of the top-10 individuals in the 50 QTLs scenario.

5.4 Discussion

5.4.1 Risks of truncation selection

Nowadays, the use of truncation selection is still popular among breeders, despite the fact that fixation of unfavorable QTL alleles associated with this selection method has been reported (Jannink, 2010). By selecting parents based on

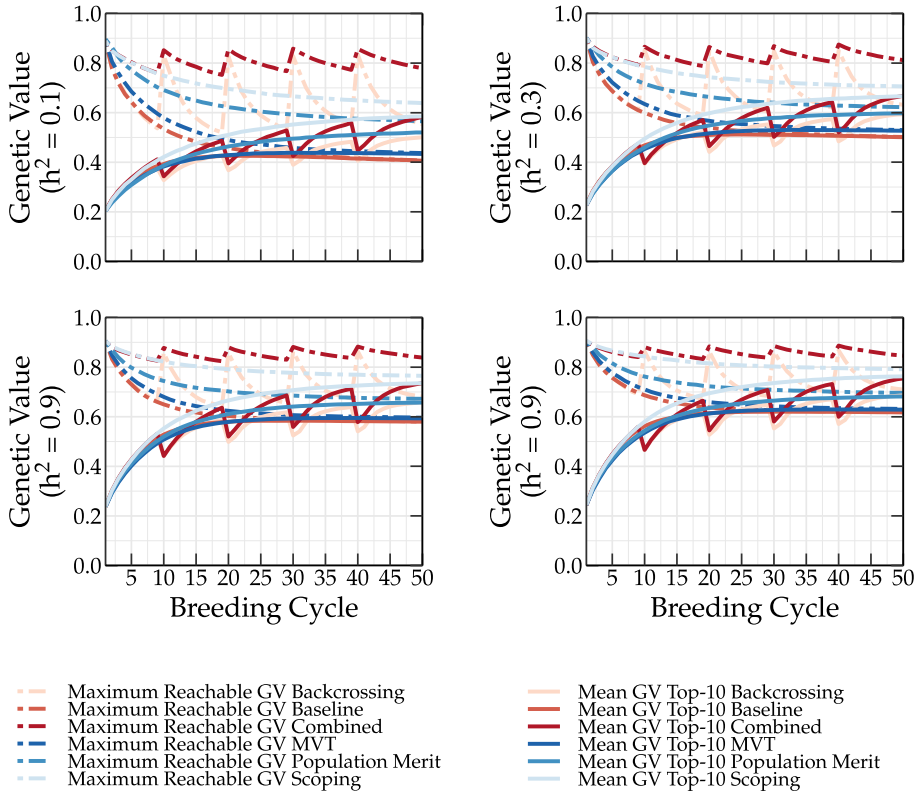

Figure 5.8: Genetic value of the different parental selection methods for different heritabilities over 50 breeding cycles. The genetic value in the long term is the lowest when using the baseline method, followed by the maximum value total (MVT) method, population merit method and the scoping method, which delivers the highest genetic values in the long term. In other words, each method can maintain its relative position towards the other methods indicating that the heritability does not influence the effectiveness of the different parental selection methods.

their GEBV using truncation selection, breeders hope to maximally pass favorable QTL alleles on to the next generation. However, the GEBV represents only a single value per individual that integrates the genetic information of more than 1000 molecular markers (see Eq. (4.6)). In contrast to MAS, in genomic selection, only a fraction of those molecular markers are in strong linkage disequilibrium with QTLs (Meuwissen et al., 2001). By summarizing the information of all those marker effects into a single number, important genetic information is lost, rendering it difficult to detect the presence or absence of favorable QTL alleles. This is especially the case when rare marker effects are masked by the presence of many other marker effects. This was demonstrated in the baseline method, where several negative QTL alleles were fixed in the breeding population. Eynard et al. (2017) simplified the selection of favorable QTL alleles by assigning weights to rare marker alleles. Nevertheless, it is clear that truncation selection does not guarantee the presence of all favorable QTL alleles in the parental population and could

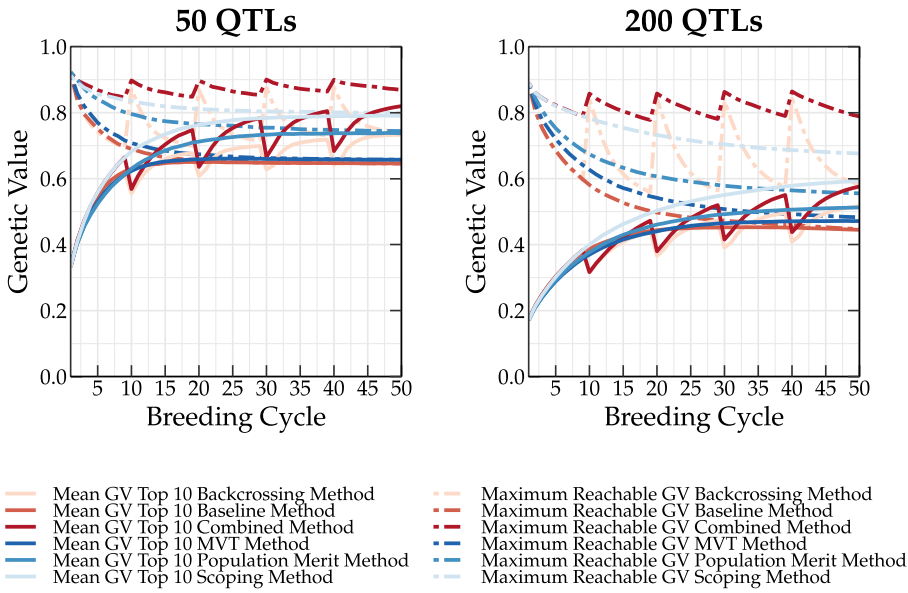

Figure 5.9: Mean genetic value of the top-10 individuals using the baseline, backcrossing, maximum variance total (MVT), population merit, scoping and combined methods simulated over 50 breeding cycles with at the left 50 QTLs and at the right 200 QTLs. Each method can maintain its relative position towards the other methods indicating that the number of QTLs does not influence the effectiveness of the different parental selection methods.

hence result in their loss. However, the baseline method has a positive genetic gain over each breeding cycle, indicating that the fixation of favorable QTL alleles has a higher impact on the genetic value than the fixation of unfavorable QTL alleles. The reduction of the genetic variation of the breeding population, which is often associated with truncation selection, causes a reduction in prediction accuracy (Heffner et al., 2009), which implies poorly estimated marker effects and substandard parental selections (see Figure 5.11). In turn, a poor parental selection in combination with a low genetic variation will further contribute to the loss of favorable QTL alleles as observed in the baseline method. Jannink (2010) tackled this problem by limiting the rate of inbreeding in the TP and thereby reducing the loss of genetic variation. However, methods based on truncation selection still cause the loss of several favorable QTL alleles. The scoping method also tackles this problem by preserving the genetic variation throughout the breeding cycles and increases long-term gain (see Figure 5.10).

Combined with truncation selection, the recurrent selection cycle can also cause the loss of several favorable QTL alleles, further reducing the genetic variation of the breeding population. During selection, 100 parents are selected and divided into 50 couples. Each couple is crossed for twenty times followed by two generations of single-seed descent leading to 1000 F3 individuals. If one couple is able to produce 20 F3 individuals with high GEBVs, they will all be selected in the next

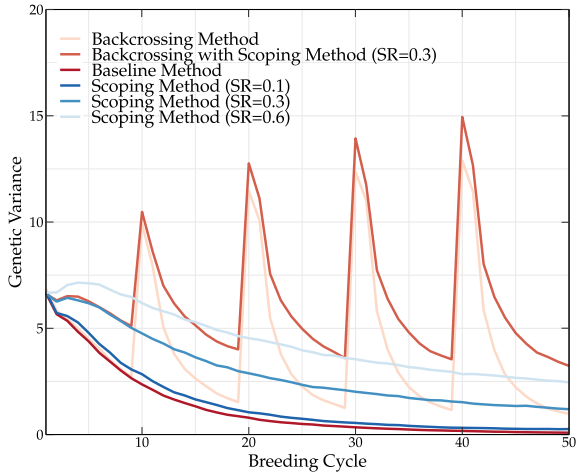

Figure 5.10: Genetic variance of the different parental selection methods over 50 breeding cycles. The genetic variance drops the fastest when using the baseline method, followed by the maximum variance total (MVT) method, population merit method and the scoping method with a scoping rate of 0.3. Both the backcrossing and combined methods result in a high genetic variation after each backcrossing event, but a drop in the genetic variation is observed in the subsequent breeding cycles.

parental population. This means that at least 20% of the parental population will then have the same ancestor, which could cause a strong reduction in genetic variation and the loss of several favorable QTL alleles. Certainly QTL alleles that are masked by many other QTL effects are at risk. Therefore, breeders should always be cautious when applying truncation selection in a breeding program. Using the average genetic relationship or marker information as done by the population merit method and scoping method could prevent the selection of too closely related individuals and thus prevent the loss of genetic variation.

5.4.2 Preserving genetic variation for long-term benefits

Truncation selection causes the loss of several favorable and unfavorable QTL alleles, reducing the genetic variation of the breeding population and causing a premature convergence of the genetic value to a local optimum. Reintroducing new semi-wild species can temporally increase the genetic variation. The backcrossing method uses, therefore, the F1 hybrids of the base population, making it possible to study this method in a predefined simulation setting and compare it with other selection methods. Although the introduction of those F1 hybrids in the parental population helped to increase the long-term genetic gain, the backcrossing method suffered from abrupt changes in mean genetic value after each back-

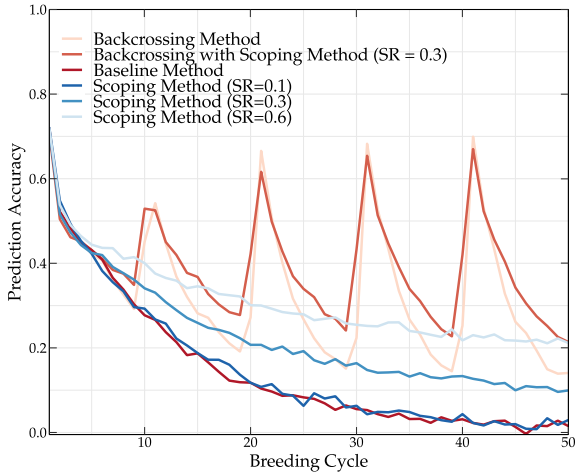

Figure 5.11: Prediction accuracy of the different parental selection methods over 50 breeding cycles. The prediction accuracy drops the fastest when using the baseline method, followed by the maximum variance total (MVT) method, population merit method and the scoping method with a scoping rate of 0.3.

crossing event. Introducing a pre-breeding program, in which semi-wild species are crossed for several generations to increase the reintroduction of favorable QTL alleles could reduce the abrupt changes in the mean genetic value, but it would also reduce the cost and time efficiency of the backcrossing method. The scoping method *consistently* preserves the genetic variation in the breeding population for as long as possible and thus avoids a premature convergence of the genetic value. In less than 20 breeding cycles, both methods yield offspring with higher genetic values compared to the baseline method.

Using the backcrossing method, the global optimum is eventually approached by moving from local optimum to local optimum. However, the relative gain of every consecutive backcrossing period (10 cycles) decreases. The scoping method avoids local optima by consistently preserving the genetic variation in the breeding population for as long as possible and thus approaches the global optimum in a more direct manner. The combination of selecting parents with high GEBVs and preserving both marker alleles in the breeding population has proven to be more efficient than reintroducing genetic information in the breeding population, making the scoping method the most efficient one.

In the backcrossing method, the timing of the backcrossing event is an important design choice. Expediting the backcrossing event to five breeding cycles leads to a decrease in the genetic gain. Each backcrossing event is now only followed by four breeding cycles of truncation selection which is not enough to recover from a sudden drop in the genetic value. Retarding the backcrossing event to 20 breeding cycles will reduce the time efficiency of the backcrossing method. Each backcross-

ing event is now followed by 19 breeding cycles of truncation selection. The truncation selection causes a rapid loss of the genetic variation leading to premature convergence of the genetic value to a local optimum. That optimum is reached well before the next backcrossing event, causing a temporary loss in genetic gain between the period in which the local optimum is reached and the next backcrossing event. In summary, the time between each consecutive backcrossing should be on the one hand large enough to recover from the sudden change in genetic value caused by the reintroduction of several QTL alleles and on the other hand short enough to avoid a premature convergence of the genetic value in a local optimum. Both design criteria were achieved when introducing the backcrossing event every 10 breeding cycles.

Both the scoping method and backcrossing method do not only preserve the genetic variation in the breeding population but also in the TP, leading to an improved prediction accuracy (see Figures 5.10 and 5.11) (Voss-Fels et al., 2018). In the case of the backcrossing method, after nine breeding cycles of truncation selection, most marker frequencies approach either zero or one. Backcrossing the breeding population with the base population yields closely related individuals with alterations on certain markers. This makes it possible to uncover important marker effects that had been masked during previous breeding cycles. In the case of the scoping method, by preserving both marker alleles at each marker, both alleles at each QTL were also preserved in the breeding population. If certain marker effects were masked or poorly predicted, the alternate allele could still be built into the next generation.

The scoping method delivers an important message. Fixation of favorable QTL alleles is not a prerequisite to obtain high genetic values. The scoping method was able to outperform the baseline method and backcrossing method with only 40% of the QTL alleles fixed in the breeding population. Preserving both alleles at each QTL prevents the elimination of poorly predicted QTL alleles. In the combined method, the incorporation of the scoping method not only increased the mean genetic value of the breeding population, but it also increased the maximum reachable genetic value making it easier to converge towards the global optimum.

5.4.3 Reintroducing genetic material in the breeding population

In this chapter, the importance to preserve the genetic variation has already been pointed out several times. The scoping method has been proposed as an alternative method to help reduce the loss in genetic variation. However, even when the scoping method was used, a small fraction of the favorable QTL alleles was still lost during breeding. Therefore, the combined method was proposed, combining the backcrossing method with the scoping method. This method was not only able

to preserve the genetic variation like the scoping method, but it also reintroduced lost QTL alleles during the backcrossing process, leading to the highest maximum reachable genetic values compared to the other methods. Unfortunately, the sudden change in the genetic value after each backcrossing event reduces the usability of this method.

In the long term, the combined method gained the same mean genetic value of the top-10 individuals as the scoping method, but the genetic variation and the maximum reachable genetic value of the combined method are still surprisingly higher than the scoping method, making the combined method potentially more interesting. Combining this method with a pre-breeding program could reduce the drop in the mean genetic value, making it possible to reach the full potential of the combined method. Unfortunately, this will also be accompanied by high costs, reducing its applicability.

Nevertheless, both the backcrossing and combined methods indicate that the reintroduction of genetic variation is important when different favorable and unfavorable QTL alleles have been eliminated from the breeding population. In Chapter 7, a new method is proposed that reintroduces genetic variation in the breeding population while the drop in the genetic value is avoided.

5.4.4 Comparison of the scoping method with existing methods

Two existing methods (the population merit method (Lindgren and Mullin, 1997) and the MVT method (Cervantes et al., 2016)) were compared with the scoping and backcrossing methods. The population merit method calculates a score per parental population and is maximized using an iterative algorithm. By penalizing a high genetic relationship between parents, the loss in genetic variation is minimized. The population merit method delivered a significant improvement compared with the baseline method, but the scoping method was able to outperform the population merit method within the first 10 breeding cycles. The backcrossing method had the same mean genetic gain of the top-10 individuals compared with the population merit method. However, the backcrossing method cannot fully recover after each backcrossing event, making the population merit method more efficient as it delivered a positive genetic gain over each breeding cycle. The combined method also had a sudden drop in the genetic value after each backcrossing event. However, by using the scoping method instead of truncation selection, a higher genetic gain was observed during the next nine breeding cycles. Using the combined method, higher genetic values were observed in the long term, but after each backcrossing event, the combined method needed four breeding cycles before outperforming the population merit method. The population merit method was able to preserve a certain fraction of the genetic variation and thus improve

long-term genetic gain, but only using the genetic relationship matrix was not enough. Over the first breeding cycles, a strong decrease in the maximum genetic value was still observed, indicating the fixation of several unfavorable QTL alleles. This was probably caused by the loss in genetic variation, leading to a lower prediction accuracy and thus a poor estimation of the additive marker effects (see Figures 5.10 and 5.11). The population merit method reduces the information of the genetic relationship matrix into a single averaged value. This certainly helps to preserve the genetic variation, but it does not guarantee that all the marker alleles will be preserved in the breeding population. A decrease of the maximum reachable genetic value is a good indicator to monitor the loss of favorable QTL alleles. A good parental selection method should be able to keep the maximum reachable genetic value fixed. It is clear that the population merit method fails at preventing the loss of those favorable QTL alleles. The scoping method ensures the inclusion of all available marker alleles, reducing the loss of favorable QTL alleles and thus maximizing the genetic variation over each breeding cycle.

Over the first breeding cycles, a lower mean genetic value of the top-10 individuals is observed for the population merit method. Compared with truncation selection, individuals with high GEBVs are exchanged with other individuals to minimize the average genetic relationship of the parental population, affecting the genetic gain in the short term. Compared to the scoping method, each individual can be selected as a parent. This means that an individual with low GEBV and a low genetic relationship could be selected as parent if c is chosen high enough. Although the population merit method had a low genetic gain in the short term, the genetic variation was well preserved, leading to a better prediction of the marker effects and a better parental selection. The MVT method also had a low genetic gain in the short term, despite the fact that a pre-selection was used to avoid the acceptance of low-GEBV individuals. The MVT method did not take into account the genetic relationship per couple or the mean genetic value of the parents, possibly leading to a low genetic gain in the short and the long term. Again, the use of the inbreeding coefficient and relationship matrix only prevents the selection of closely related individuals, but it does not directly prevent the loss of certain QTL alleles. Similar to the GEBVs, the genetic value summarized the genetic information (genotype and allele frequency) into a single value, masking important information during breeding. The MVT method should preserve the genetic information, however, in the current setting, the genetic variation is barely preserved. Testing the MVT method without pre-selection and greedy parental selection (P_1) led to a much higher genetic variation, but because the method does not take into account the genetic value, only low genetic values were observed. It is clear that the combination of the MVT method with a pre-selection and a greedy parental selection was not successful, but it did deliver higher genetic values compared with the baseline method.

5.4.5 Limitations of the Pearson correlation

In genomic selection, the prediction accuracy is defined as the linear correlation between the GEBVs and the true breeding values. This can be measured by taking the Pearson correlation between the GEBVs and the genetic values. Although the Pearson correlation is widely used in GS to evaluate the performance of a method (or prediction model), this performance measure is prone to outliers and is influenced by the amount of genetic variation in the breeding population, resulting in an overestimation of the prediction accuracy (Devlin et al., 1975; Glass and Hopkins, 1996; Goodwin and Leech, 2006).

In Figure 5.12 the mean genetic value of the top-10 individuals, the mean genetic value of the breeding population, and the Pearson correlation are shown for a breeding population using truncation selection and the scoping method. According to the mean genetic value of the top-10 individuals and the Pearson correlation, the scoping method outperforms truncation selection between breeding cycles 7 and 8, while a lower prediction accuracy is observed for truncation selection at breeding cycle 6. According to the mean genetic value of the breeding population, the scoping method will only be able to outperform truncation selection at breeding cycle 15, but at that point, the scoping method already resulted in superior individuals compared to truncation selection. The goal of breeding is to develop new superior breeding lines, and therefore, the mean genetic value of the top-10 individuals is a better choice than the mean genetic value of the breeding population to evaluate a method.

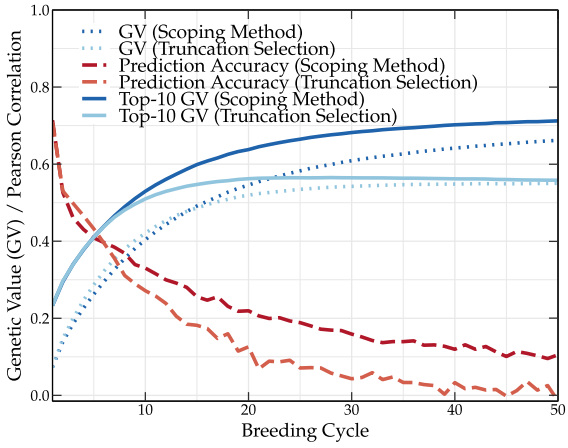

Figure 5.12: Mean genetic value of the top-10 individuals, Mean genetic value of the breeding population and the prediction accuracy (Pearson correlation) are shown for a breeding population using truncation selection and the scoping method simulated over 50 breeding cycles. The Pearson correlation degrades over each breeding cycle while the genetic gain is increased.

At breeding cycle 20, the mean genetic value of the top-10 individuals using truncation selection converges. At that point, the Pearson correlation is still degrading and will only converge around breeding cycle 40. Based on the Pearson correlation, we could expect that truncation selection will have a better performance at breeding cycle 20 than at breeding cycle 50. However, the mean genetic value of the top individuals remains unchanged over these breeding cycles. High long-term genetic gains can be obtained by using a good parental selection strategy and a good estimation of the marker effects. As long as the parental selection strategy requires the estimation of marker effects, if a poor estimation is obtained, a lower long-term genetic gain will be observed. Therefore, we believe that using the mean genetic value of the top-10 individuals is a better and safer choice to correctly evaluate the performance of the different parental selection methods.

Especially for a small scoping rate, the Pearson correlation can be deceiving. Individuals with a low genetic value can be pre-selected if their GEBV is high due to a poor estimation. Because these individuals are often associated with a broad genetic variation, when the scoping method is used, one of these individuals will probably be selected as a P2 parent, leading to F3 individuals (their offspring) with a lower genetic value. Because the genetic value of these F3 individuals will differ from the rest of the population, the Pearson correlation will be overestimated. While the high prediction accuracy could give the breeder the wrong impression, the genetic gain over the subsequent breeding cycles will be lower, indicating that the parental selection was far from optimal.

5.4.6 Prediction model

Different prediction models are available to estimate the marker effects. In genomic selection, rr-gBLUP is often used, but non-linear models like Bayes A, Bayes B and Bayes C have also been used in different research settings. The effect of the different prediction models on the genetic value for the baseline (top line) and scoping (bottom line) methods are evaluated (see Figure 5.13). As expected, the Bayesian models resulted in a similar mean genetic value of the top-10 individuals as rr-gBLUP (Moser et al., 2009).

Normally, in each breeding cycle, low-frequency markers are removed from the training population to increase the prediction accuracy Edriss et al. (2013). To study the effect of low-frequency markers on the prediction accuracy of both rr-gBLUP and the Bayesian models, a breeding population is simulated using all the available markers to fit the mixed effects model. Removing the low-frequency marker alleles had a negative effect on the genetic value. Based on the results as depicted in Figure 5.13, low-frequency markers should not be removed from the training population. Although Bayesian models result in a similar breeding value in the long term, models like the Bayes B and Bayes C models require a parameter

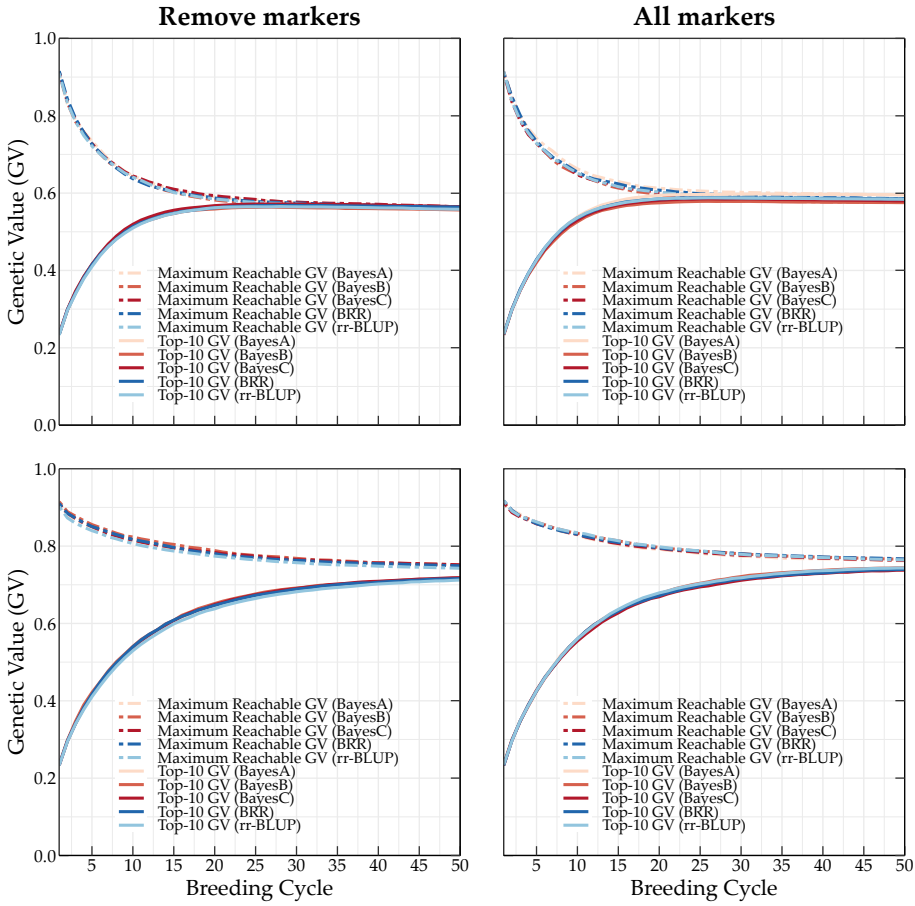

Figure 5.13: Simulation results of the baseline method (top line) and scoping method (bottom line) using a training population in which low-frequency markers are removed (left) or using a training population in which all markers are used (right). Each prediction model results in the same long-term genetic gain. When low-frequency markers are removed from the training population, a lower genetic gain is observed.

π, which needs to be fine-tuned by the breeder. Unfortunately, this fine-tuning process often relies on gut feeling and will thus differ from breeder to breeder. A prediction model like rr-gBLUP can directly be used without the need of fine tuning and is therefore a safer choice.

5.4.7 Updating the training population

Neyhart et al. (2017) reported the importance of systematically updating the TP to counter the decay in prediction accuracy. Several update methods were studied in combination with the scoping method, confirming the importance of updating the training population. Not updating the training population decreases the relationship between the training population and the breeding population (Isidro et al., 2015). In turn, poorly estimated GEBVs will essentially result in a random selection, explaining the premature convergence of the mean genetic value. Over each subsequent breeding cycle, favorable QTL alleles are eliminated from the breeding population, resulting in a much lower genetic gain.

The effect of the different training population update methods is shown in Figure 5.14. The difference in genetic value between the different update methods is almost negligible and is in accordance with Neyhart et al. (2017). In the long term, the *tails* update method results in a slightly higher mean genetic value of the top-10 individuals whereas selecting individuals with the lowest GEBVs (*bottom* update method) results in the lowest mean genetic value of the top-10 individuals. A training population that only contains individuals with the lowest GEBVs will have difficulties in correctly predicting the top individuals. Certainly for the scoping method with a scoping rate of 0.3, it is important that the GEBVs are correctly predicted to avoid the selection of low-GEBV individuals that could decrease the genetic gain over the next breeding cycles. The tails method combines the information of both tails of the population and results in an accurate prediction of the top and bottom individuals. Selecting individuals with the highest GEBVs in the TP (*top* update method) results in an accurate selection of the top individuals but individuals with a lower genetic value are predicted inaccurately. Because each selection method only selects individuals with a high GEBV, selecting individuals with the highest GEBV in the TP had no negative repercussions on the genetic gain. Randomly selecting individuals in the TP results in an overall good prediction accuracy over the whole breeding population and will therefore, also result in a similar long-term genetic value as observed for the top and tails update methods.

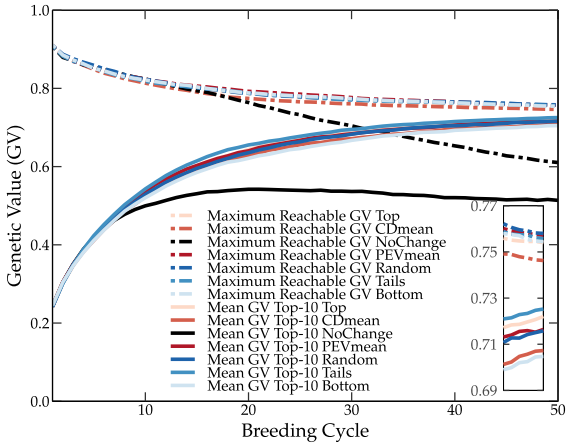

Figure 5.14: Effect of the TP update method on the genetic value using the scoping method over fifty breeding cycles. Only a small difference is observed between the *top*, *bottom*, *tails* and *random* update methods. The *no change* update method, which does not update the TP, rapidly converges to a low mean genetic value and several favorable QTL alleles are eliminated from the breeding population, thus reducing the maximum reachable genetic value.

5.4.8 Cost analysis

The influence of a parental selection method on the profit margin of a breeding program is difficult to assess. Different factors are often in play, influencing the market price, and costs. Therefore, assessing the marginal profit between using the scoping method instead of truncation selection is difficult.

Assuming that each method is performed under the same conditions and with the same resources, then the cost for each method should also be the same. Only the backcrossing method and combined method will deviate, as they require the use of F1-hybrids from the first breeding cycle. The backcrossing method was an interesting case study, but was only a theoretical concept. Therefore, both the backcrossing method and the combined method will not be considered during this analysis.

The scoping, population merit and MVT methods all uses the same settings as truncation selection. Only the parental selection is different. Truncation selection uses the GEBV, whereas the scoping method also uses marker information. The population merit requires the genetic relationship matrix and the MVT requires the degree of inbreeding. All the required information is already available and does not increase the financial cost.

The scoping method gains a 15 percentage points higher genetic value compared to the baseline method. Therefore, we can assume that using the scoping method will deliver a financial advantage after 8 breeding cycles. At that point, a breeding

population using the scoping method will yield improved breeding lines compared to a breeding population using truncation selection. The effect of these breeding lines on the marginal profit are difficult to predict. Probably, due to the superiority of these breeding lines, the market share of the breeding company will grow resulting in a higher profit.

5.5 Conclusion

In our simulation study, we demonstrated the need for an alternative parental selection method to prevent the convergence of the genetic value of the breeding population to a local optimum caused by the loss of favorable QTL alleles. Truncation parental selection leads to a rapid fixation, but also to the loss of several favorable QTL alleles, causing the convergence of the genetic values to a suboptimal value and reducing the possibility to reach the global optimum in the long term. Consistently preserving the genetic variation (scoping method) leads to higher genetic values in the long term and only a slightly lower genetic value in the short term. We conclude that preserving the genetic variation by means of the scoping method is more beneficial than periodically reintroducing new genetic information into the breeding population.

6

The adaptive scoping method

Truncation selection is often used to rapidly achieve short-term genetic gain within a breeding program. Unfortunately, it is also associated with the loss of favorable QTL alleles in the breeding population, causing a premature convergence to sub-optimal genetic values. Parental selection strategies such as the scoping method have been proposed to preserve genetic variation in the breeding population and thus maximize genetic gain in the long term. Nevertheless, for economic reasons, breeders are often interested to maximize the genetic gain in a shorter time frame. We propose a new selection strategy, named the adaptive scoping method, that aims at maximizing the genetic gain within a specific, predefined time frame. Throughout this time frame, the adaptive scoping method progressively changes its selection strategy: during the initial breeding cycles, it attempts to maximally preserve genetic variation, whereas in later breeding cycles, it prioritizes the increase of the genetic value. We demonstrate through simulation studies that the adaptive scoping method is able to maximize the genetic gain for a wide range of time frames and that it outperforms the original scoping method, both in the short and in the long term.

The material of this chapter is based on the following publication:
Vanavermaete, D., Fostier, J., Maenhout, S., De Baets, B., 2021. Adaptive scoping: balancing short- and long-term genetic gain in plant breeding. (under review).

6.1 Introduction

From an economic point of view, breeders aim to maximize the genetic gain as quickly as possible. To this end, they often resort to the use of truncation selection: every generation, individuals that rank highest according to certain traits of interest are selected for breeding. When such parental selection is guided by pedigree information, the reduction in genetic variation is limited (Piepho et al., 2008). However, when truncation selection is based on genotypic data (i.e., genomic selection), a rapid fixation of large-effect quantitative trait loci (QTLs) has been observed (Clark et al., 2011; Pszczola et al., 2012; Jannink, 2010). The loss of favorable QTL alleles in the breeding population reduces the maximum reachable genetic value, ultimately resulting in a premature convergence to a sub-optimal genetic value. Therefore, a successful breeding program should find a balance between genetic gain on the one hand and the preservation of genetic variation in the breeding population on the other hand (Jannink, 2010).

Different methods have been proposed in literature to maximize long-term genetic gain by controlling the average inbreeding coefficient of a population (Wray and Goddard, 1994; Brisbane and Gibson, 1995; Meuwissen, 1997). The inbreeding coefficient of a diploid individual is the probability that, at any given locus, the two alleles are identical by descent (i.e., originate from the same ancestor). It is important to manage the rate at which the average inbreeding coefficient changes between consecutive breeding cycles (Woolliams et al., 2015). A high rate of inbreeding results in a quick, short-term genetic gain but a rapid loss of genetic variation, whereas a low rate of inbreeding yields a better preservation of the genetic variation at the expense of a slower genetic progress. Unfortunately, estimating the rate of inbreeding for the next generation solely based on information of the parents remains difficult. Different methods to predict and control the rate of inbreeding have been proposed. Wray and Thompson (1990) propose the use of the long-term genetic contribution metric, i.e., the proportion of the genes of an individual that will be passed to its descendants in the long term. Meuwissen (1997) proposed to limit the rate of inbreeding by restricting the coancestry between parents using the optimal contribution selection (OCS) method. In similar approaches, the rate of inbreeding was controlled by ensuring a sufficient genetic distance between parents, thus limiting within-family selection (Sonesson et al., 2012; Allier et al., 2020b). Gorjanc et al. (2018) use the OCS method in a two-part breeding program to maximize the long-term genetic gain. Akdemir and Sánchez (2016) propose an optimal mating plan, taking into account the risk of inbreeding,

the allele heterozygosity and allele diversity. Unfortunately, none of these methods allow for the direct construction of the optimal set of parents to be used for breeding and thus, different combinations should be evaluated. Because it is not computationally feasible to enumerate and evaluate all possible combinations of parent individuals, optimization techniques such as genetic algorithms are often used to find a good parental population (Allier et al., 2020b; Gorjanc et al., 2018). However, such techniques tend to converge to local optima, which implies that the optimal parental population may not be found.

In Chapter 5, the scoping method is proposed as an alternative strategy to preserve the genetic variation in a breeding population and thus maximize the long-term genetic gain. The selection of parental individuals is performed in a computationally efficient manner and consists of two steps: pre-selection followed by actual parental selection. First, a certain fraction of the individuals with the highest genomic estimated breeding values (GEBVs) are pre-selected from the breeding population. This fraction is referred to as the scoping rate (SR). A low scoping rate results in the pre-selection of a small set of individuals with only the highest GEBVs, whereas a high scoping rate yields a larger, more diverse set of candidate parents. From this set, parents are selected and coupled aiming for genetic progress as well as the preservation of genetic variation. The scoping method was demonstrated to outperform parental selection methods such as truncation selection, the population merit method (Lindgren and Mullin, 1997) and the maximum variance total method (Cervantes et al., 2016), with a more pronounced superiority in the long term.

For economic reasons, breeders are often interested in the maximization of genetic gain in a shorter time frame. To this end, we propose a modification of the scoping method that aims at optimizing the genetic gain of a breeding population within a predefined number of breeding cycles. This method, referred to as the *adaptive* scoping method, dynamically changes the scoping rate throughout the different breeding cycles: initially, high scoping rates are considered such that the preservation of genetic variation is emphasized, whereas during later breeding cycles, increasing the genetic value is gradually prioritized through lower scoping rates. The adaptive scoping method takes only a single parameter, namely the time frame t (expressed through the number of breeding cycles) during which the genetic gain should be maximized. This unique feature enables breeders to balance exploration and exploitation of their breeding population.

6.2 Materials and methods

The base population and breeding scheme are adopted from Neyhart et al. (2017). The base population is constructed with two datasets of North American barley

(Hordeum *vulgare*) from the University of Minnesota (UMN) and the University of North Dakota (NDSU), counting respectively 384 and 380 six-row spring inbred lines with 1590 biallelic SNP loci. The simulation study was constructed in a similar way as described in Chapter 4, ensuring that the performance of the adaptive scoping method can be compared with that of the original scoping method.

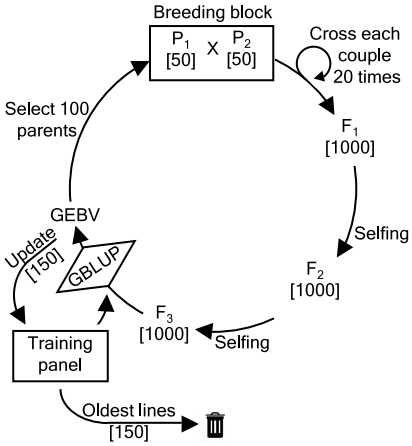

Figure 6.1: Overview of the recurrent breeding scheme. First, 50 couples of parents (P_1, P_2) each produce 20 offspring, yielding a total of 1000 F1 hybrids. After two generations of single-seed descent, 1000 F3 individuals are obtained. From those F3 individuals, new parental lines are selected.

6.2.1 Breeding scheme

The recurrent breeding scheme is depicted in Figure 6.1 and has been described in Chapter 4. In the initial breeding cycle, 50 individuals with the highest phenotypic values of the NDSU dataset are coupled with 50 individuals with the highest phenotypic values of the UMN dataset. Each couple produces 20 offspring and after 2 generations of single-seed descent, the base population is obtained containing 1000 individuals. From this point onward, the parents are selected solely based on the genomic estimated breeding values (GEBVs) to reduce the financial cost of phenotyping. The GEBVs are predicted using a linear mixed effects model that has been fitted using the base population, which contains both phenotypic and genotypic information.

In each subsequent breeding cycle, 100 parents are selected and coupled according to one of the parental selection methods considered to construct a crossing block. Each couple produces 20 offspring resulting in a total of 1000 F1 hybrids. After two generations of single-seed descent, 1000 F3 individuals are obtained. These individuals represent the new breeding population from which parents can

again be selected. Each simulation run consists of 50 breeding cycles and all results are averaged over 100 simulation runs.

6.2.2 The scoping method

In Chapter 5, the scoping method is proposed to preserve genetic variation in a breeding population and thus maximize genetic gain in the long term. The scoping method consists of two steps: pre-selection followed by parental selection. During pre-selection, individuals with the highest GEBVs are chosen. This ensures that genetic progress can be made over the next breeding cycles. The fraction of individuals that is pre-selected is controlled by the scoping rate which ranges between a minimum value $\mathrm{SR}_{\mathrm{min}}$ and a maximum value $\mathrm{SR}_{\mathrm{max}}$. Here, $\mathrm{SR}_{\mathrm{max}} = 1$ (pre-select 100% of the individuals), while $\mathrm{SR}_{\mathrm{min}} = n_s/n_t$, with n_s the number of parent individuals to be selected for breeding and n_t the total population size. For example, if $n_s = 100$ parents are to be selected from a population size $n_t = 1000$, then a fraction of at least 10% ($\mathrm{SR}_{\mathrm{min}} = 0.1$) of the individuals must be pre-selected.

Next, from these pre-selected individuals, $n_s/2$ parental pairs are consecutively chosen as follows: the individual with the highest GEBV is selected as the P1 parent, whereas the P2 parent is taken such that the genetic variation of all selected parents thus far is maximized over all markers. Mathematically speaking, the F_{score} is maximized:

$$F_{\mathrm{score}} = \sum_{i=1}^{k} \mathrm{var}(\mathbf{Z}_i) p_i \,, \tag{6.1}$$

with k the number of genetic markers, $\mathbf{Z}_i$ the i-th column of the $n \times k$ matrix $\mathbf{Z}$ containing the genotypes (coded as -1, 0 and 1) of the n already selected individuals and $\mathbf{p}$ a Boolean vector of length k. Initially, p_i is set to 1 for all marker positions. When both alleles at marker i are present, p_i is set to 0. This way, the F_{score} takes into account only those markers for which both alleles are not yet present in the selected population. Once both alleles are present for all marker positions, all p_i are again set to 1 such that the variance is again maximized over all markers.

The core idea is that the P1 parents drive the genetic progress of the offspring, whereas the P2 parents ensure the preservation of genetic variation. Clearly, the scoping rate controls the trade-off between the degree in which genetic variation can be preserved on the one hand, and the rate at which genetic gain can be made on the other hand.

6.2.3 The adaptive scoping method

The scoping method uses a single, fixed value for the scoping rate across different breeding cycles. In contrast, the adaptive scoping method gradually decreases the scoping rate from its maximum value SR_{max} to its minimum value SR_{min}. The adaptive scoping method takes a single parameter t, expressed through the number of breeding cycles over which the scoping rate is varied. Specifically, at breeding cycle i, the scoping rate takes the value:

$$SR(i) = \begin{cases} \frac{SR_{min} - 1}{t - 1} i + \frac{t - SR_{min}}{t - 1} & \text{, if } 1 \leq i \leq t \\ SR_{min} & \text{, if } i > t \end{cases}. \tag{6.2}$$

In other words, the scoping rate decreases linearly over t breeding cycles from $SR_{max} = 1$ at breeding cycle 1 to SR_{min} at breeding cycle t (and later cycles). As a consequence, during the first breeding cycles, the adaptive scoping method pre-selects a larger number individuals, focusing on the preservation of the genetic variation (exploration). In contrast, at breeding cycles t and later, only elite individuals are pre-selected, maximizing the genetic progress (exploitation). From this set of pre-selected candidate parents, parent pairs are chosen in an identical manner as in the scoping method.

6.2.4 Prediction model

The GEBVs are predicted by fitting a linear mixed effects model:

$$\mathbf{y} = \mathbf{1}_n \beta + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{6.3}$$

with $\mathbf{y}$ a vector of phenotypic values, $\mathbf{1}_n$ a vector of size n containing ones, n the number of individuals in the training panel, β the fixed effect (phenotypic mean), $\mathbf{Z}$ the incidence matrix of the training panel with marker information, $\mathbf{u}$ the marker effects following a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{G})$ with $\mathbf{G} = \sigma_u^2 \mathbf{I}_k$ (with $\mathbf{I}_k$ the identity matrix of dimension k), k the number of markers and $\boldsymbol{\epsilon}$ the residual effects following a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. Both variance components σ_u^2 and σ_e^2 are estimated by means of restricted maximum likelihood (REML). The GEBVs of the individuals are calculated as:

$$\hat{\mathbf{g}} = \mathbf{Z}\hat{\mathbf{u}}, \tag{6.4}$$

with $\hat{\mathbf{g}}$ the GEBVs, $\mathbf{Z}$ the marker information and $\hat{\mathbf{u}}$ the predicted marker effects.

At the start of the simulation study, both the UMN and NDSU datasets are used as training population. In the subsequent breeding cycles, 150 new individuals are phenotyped and added to the training panel according to the 'tails' method, selecting 75 individuals with the highest GEBVs and 75 individuals with the lowest GEBVs (Neyhart et al., 2017). According to Neyhart et al. (2017), this results in a (non-significantly) higher genetic gain compared to other update methods. Before updating the training panel, 150 individuals that have been the longest in the training panel are removed from the training population. This reduces the computational time without reducing the prediction accuracy (Neyhart et al., 2017).

The linear mixed effects model in Eq. (6.3) is fitted using the package `rrBLUP` in R (Endelman, 2011). Even though it has been recommended to remove markers with low levels of polymorphism from the training panel (Chang et al., 2018), we kept all markers as this resulted in a higher prediction accuracy.

6.2.5 Simulation of the population

The simulation study was built upon the work of Neyhart et al. (2017), using the packages `GSSimTPUpdate` and `hypred` in R (version 3.6.3). First, the genome of barley is constructed based on marker position, allele, and chromosomal information. One hundred QTLs ($L = 100$) are selected randomly from the available 1590 biallelic SNP loci. The remaining 1490 biallelic SNP loci are available as markers for prediction and selection purposes. The QTL effects are calculated according to a geometric series. At the k-th QTL, the favorable homozygote will have a value a^k, the heterozygote a value zero, and the unfavorable homozygote a value $-a^k$ with $a = (L-1)/(L+1)$. Dominance and epistatic effects were assumed to be absent. The phenotypic value is calculated over three different environments, each drawn from a normal distribution with mean 0 and a variance component σ^2_E which is defined as eight times the genetic variance (Bernardo, 2014). The phenotypic value of the i-th individual in the j-th environment (y_{ij}) is calculated as follows:

$$y_{ij} = g_i + e_j + \epsilon_{ij}\,, \tag{6.5}$$

with g_i the genetic value of the i-th individual, e_j the j-th environmental effect and ϵ_{ij} the residual effect of the i-th individual and the j-th environment. The residual effect is drawn from a normal distribution with mean 0 and a variance component σ^2_R, with σ^2_R scaled to simulate a population with a heritability (h^2) of 0.5. The phenotypic value of an individual is the averaged value over the three environments. A comprehensive overview of the simulation study has been described in Chapter 4.

To track the fixation of unfavorable QTL alleles, the maximum reachable genetic value is calculating as the sum of the QTL effects that are fixed (both favorable

and unfavorable) and the sum of the favorable QTL effects that are not yet fixed. It represents the maximum genetic value that could still be reached, taking into account the fixation of unfavorable QTL alleles. The maximum reachable genetic value and the mean genetic value are rescaled such that the maximum reachable genetic value has a value of 1. As in Chapter 5, the mean genetic value of the top-10 individuals is reported. These individuals represent the superior lines that are prime candidates for commercialization.

6.3 Results

6.3.1 Performance of the adaptive scoping method

The adaptive scoping method is designed to maximize the preservation of genetic variation over the first breeding cycles and maximize the genetic progress over later cycles. This is achieved by linearly decreasing the scoping rate over the course of t breeding cycles. Figure 6.2 shows the mean genetic value of the top-10 individuals of the population for different values of $t = 10$, 20, 30, 40 and 50 breeding cycles. For a value of $t = 10$, the preservation of variation is quickly traded off for a rapid genetic gain: from breeding cycle 10 onward, only the individuals with the highest GEBVs are considered as parents. At breeding cycle 15, this manifests itself in an at least 4 percentage points higher mean genetic value compared with the other values of t. Nevertheless, due to the rapid reduction of genetic variation, the adaptive scoping method with $t = 10$ quickly loses its ability to drive genetic progress further, yielding the worst genetic values beyond about 30 breeding cycles. The adaptive scoping method with $t = 20$ preserves the genetic variation somewhat longer before focusing on genetic gain. Around breeding cycle 25, this yields the highest mean genetic value compared with other values of t. Again, this advantage quickly degrades throughout later breeding cycles.

In general, the behavior of the adaptive scoping method can be understood as follows: the higher the value of t, the longer genetic variation is preserved before genetic gain is prioritized. As soon as t breeding cycles are completed, the adaptive scoping method resembles the behavior of truncation selection (although the pairing of parents is not random). Therefore, the adaptive scoping method will yield the highest genetic gains shortly after t breeding cycles. This is indicated in the middle panels of Figure 6.2, where the results for different values of t are compared at different breeding cycles: the adaptive scoping method with $t = 10$ yields the highest gain at cycle 15, the adaptive scoping method with $t = 20$ yields the highest gain at cycle 25, etc. The only exception is the adaptive scoping method with $t = 50$. In that case, the adaptive scoping method is not able to outperform the adaptive scoping method with $t = 40$ at breeding cycle 55 but converges to

the same value (see Table 10.6). Clearly, by choosing a particular value for t, a breeder can expect the highest gains during the breeding cycles that immediate follow t, outperforming the adaptive scoping method with different values of t.

We also compare the adaptive scoping method with the original scoping method for a fixed scoping rate across the breeding cycles. In Chapter 5, a scoping rate of 0.3 was suggested to maximize the genetic gain in the short as well as in the long term and is hence also used here. Compared with the scoping method, the adaptive scoping method uses a higher scoping rate and hence pre-selects more individuals during the first breeding cycles, allowing for a better preservation of the available genetic variation. As a consequence, during those initial breeding cycles, the adaptive scoping method suffers less from the loss of favorable QTL alleles and hence preserves a higher maximum reachable genetic value at the expense of a lower mean genetic value of its top-10 individuals (see Figure 6.3).

The short-term sacrifice in genetic gain pays off in the long term. The adaptive scoping method with $t = 10$ outperforms the scoping method after 13 breeding cycles and yields higher genetic values up to breeding cycle 20. At that point, the adaptive scoping method has exploited the remaining genetic variation, quickly leading to the convergence of the genetic value from that point onward. Similarly, the adaptive scoping method with $t = 20$ outperforms the scoping method after 20 breeding cycles. Finally, the adaptive scoping method with $t = 50$ surpasses the scoping method at breeding cycle 30 and yields the highest long-term gain (a 4 percentage point increase compared to the scoping method at breeding cycle 50). At that point, the genetic value of the adaptive scoping method is even higher than the maximum reachable genetic value of the scoping method. This means that the loss of favorable QTL alleles from the population during the initial breeding cycles of the scoping method has caused an insurmountable disadvantage in the long term. We conclude that the adaptive scoping method is able to outperform the original scoping method, both in the short term (when low values of t are used) and in the long term (for high values of t).

The mean genetic value of the top-10 individuals and the maximum reachable genetic value of the adaptive scoping method and the scoping method are reported in Tables 10.6 and 10.7, respectively.

6.3.2 Robustness of the adaptive scoping method

The original scoping method and the adaptive scoping method have been evaluated in different simulation settings. In each experiment, these methods were assessed using 100 different genomes such that the effects of different QTL and marker positions are averaged. The effects of the heritability and the number of QTLs on the genetic gain using both methods have also been tested: simulation

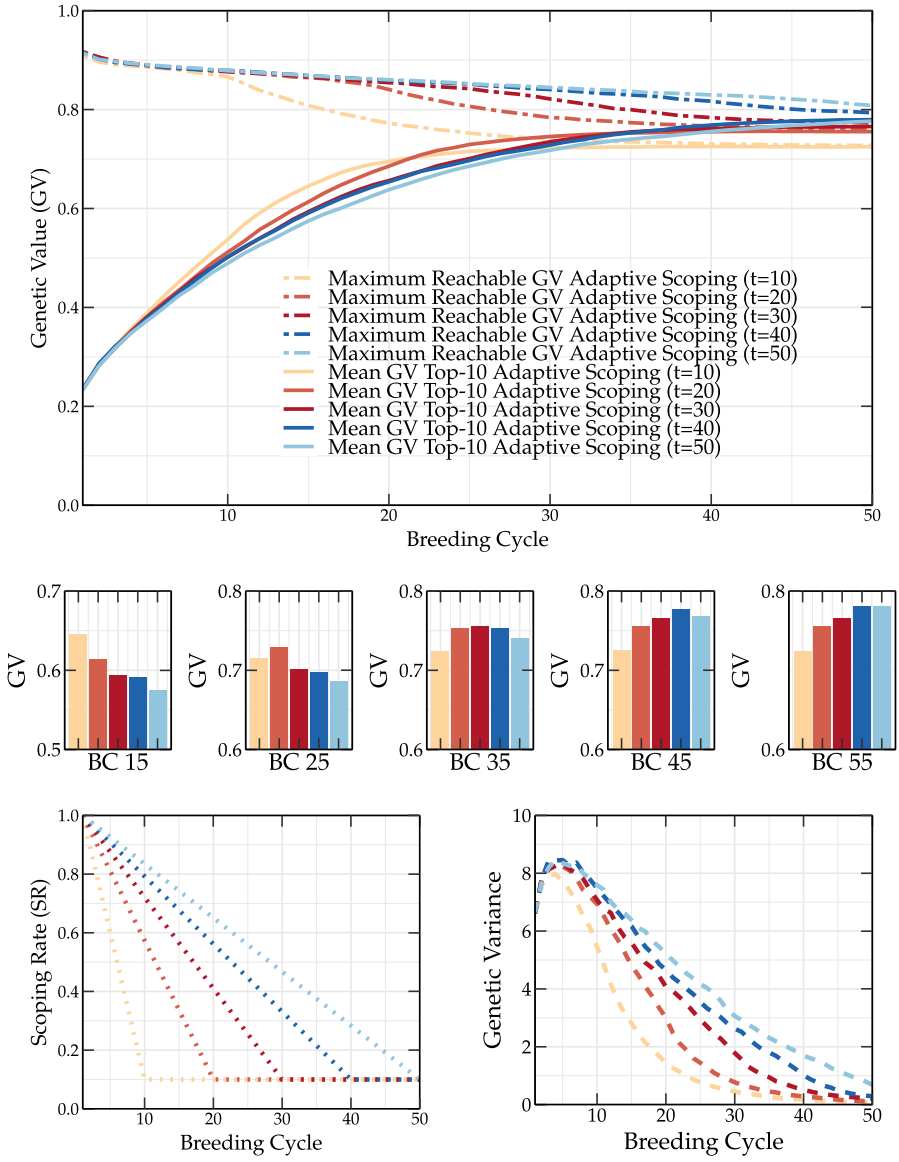

Figure 6.2: Top panel: the mean genetic value of the top-10 individuals and the maximum reachable genetic value for the adaptive scoping method with a value for t of respectively 10, 20, 30, 40 and 50 breeding cycles. Middle panels: the mean genetic values for the different parental selection methods at breeding cycles 15, 25, 35, 45 and 55. Bottom left panel: the SR for the different parental selection methods. Bottom right panel: the genetic variation for the different parental selection methods.

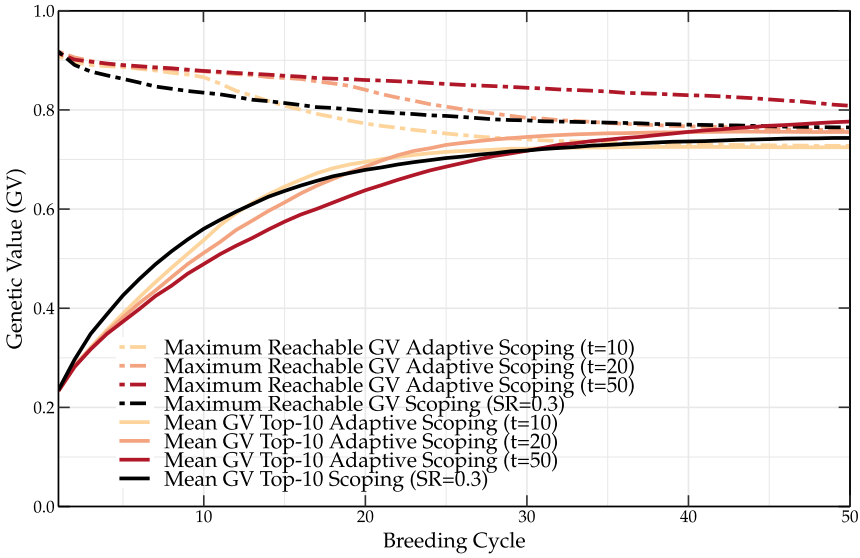

Figure 6.3: Mean genetic values of the top-10 individuals using the scoping method (SR = 0.3) and the adaptive scoping method for $t = 10$, $t = 20$ and $t = 50$. In the short term, the scoping method yields the highest genetic gains. Over time, $t = 10$ will result in higher genetic gains followed by $t = 20$, and $t = 50$.

studies were performed using a heritability of 0.2 and 0.8 using 100 QTLs, and a heritability of 0.5 using 50 and 200 QTLs (see Figure 6.4).

In each case, shortly after t breeding cycles, the adaptive scoping method resulted in the highest genetic value throughout a certain number of breeding cycles. For $t = 50$, the adaptive scoping method always yielded the highest long-term genetic gain. Increasing the heritability improves the prediction accuracy, resulting in higher genetic gains for all methods. Similarly, the GEBVs can be more accurately predicted when fewer QTLs are present. For lower values of the heritability, the effect of the environment becomes more pronounced, making it more challenging to select parents based on the GEBVs. As the adaptive scoping method is better at preserving the genetic variation over the first breeding cycles, a slower but more accurate fixation of the QTL alleles is observed, resulting in higher long-term genetic gains compared to the scoping method.

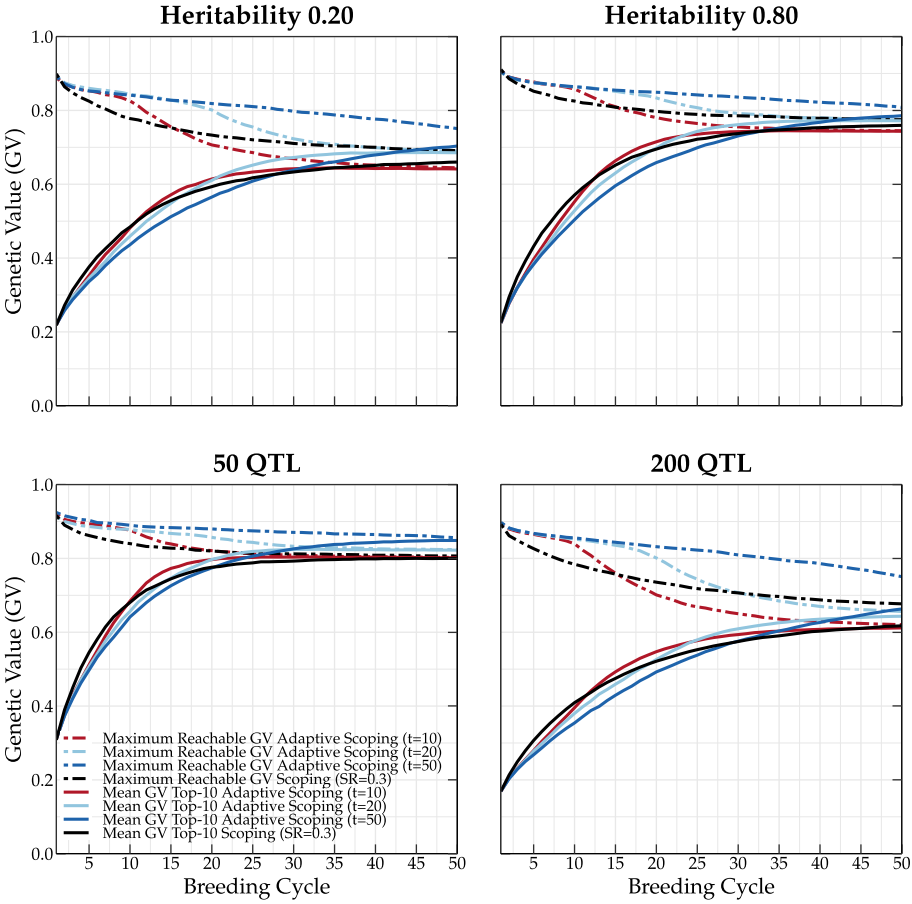

Figure 6.4: Simulation results of the original and adaptive scoping methods (using $t = 10$, 20 and 50) for a heritability of 0.2 and 0.8 using 100 QTLs (top) and for a heritability of 0.5 using 50 and 200 QTLs (bottom). The impact of both methods on the genetic value and on the maximum reachable genetic value is reported. In each case, shortly after t breeding cycles, the adaptive scoping method results in the highest genetic value throughout a certain number of breeding cycles.

6.4 Discussion

6.4.1 The effect of a variable scoping rate on the genetic gain

The loss in genetic variation and the resulting risk associated with truncation selection are well known (Jannink, 2010; Meuwissen et al., 2001). In response to this, parental selection techniques such as the scoping method were developed to better preserve genetic variation and thus maximize genetic gain in the long term. To achieve this, individuals with a lower GEBV can also be considered as a parent when they contribute to the genetic diversity. To avoid an adverse effect on the rate of genetic progress, a fraction of individuals with the highest GEBVs in the breeding population is pre-selected. This fraction is controlled by the scoping rate. The (original) scoping method relied on a fixed scoping rate throughout the different breeding cycles. For a scoping rate of 0.3, the maximum reachable genetic value decreases significantly over the first breeding cycles, indicating that several favorable QTL alleles are lost from the population at the early stages of the breeding program. In principle, this could be avoided by increasing the scoping rate (and thus pre-selecting more parents), but this would unavoidably slow down the genetic progress and, hence, require a very large number of breeding cycles to outperform the truncation selection method.

By introducing a variable scoping rate, the trade-off between genetic gain and genetic variation can be controlled during the breeding process itself. Over the first breeding cycles, a high value for the scoping rate prevents the loss of favorable QTL alleles. The scoping rate is decreased linearly, gradually prioritizing genetic progress over preserving genetic variation. This leads to a slower, but more accurate fixation of the QTL alleles, translating into lower short-term, but higher long-term genetic gains. The parameter t represents the number of breeding cycles over which the scoping rate is varied, and can thus be used to control the time frame over which the genetic value is to be optimized. After t breeding cycles, the adaptive scoping method fully prioritizes the increase of the genetic gain, and a rapid fixation of QTL alleles is observed.

6.4.2 Optimizing the breeding population within a predefined time frame

The key advantage of the adaptive scoping method is that it can be used to optimize the genetic gain of a breeding population within a predefined time frame. Depending on the goals of the breeder, an appropriate choice for t can be made:

a low value of t will provide fair genetic values in the short term, whereas higher values of t will lead to higher genetic values in the long term. Once t breeding cycles have been completed, genetic gain is fully prioritized and the highest genetic values will quickly be reached during the next few breeding cycles. Irrespective of the choice of t, the breeder can expect the adaptive scoping method to yield superior genetic values during a short time window that follows breeding cycle t. This is shown in Figure 6.5 where at each breeding cycle, the results of the method that yields the highest genetic values are shown.

The adaptive scoping method proves robust even when the prediction accuracy is low. This was demonstrated in Figure 6.4 by decreasing the heritability or increasing the number of QTLs. In both cases, selecting the best parents based on the GEBVs becomes more tedious.

6.4.3 Comparison of the scoping and adaptive scoping methods

Compared to the (original) scoping method which uses a fixed scoping rate, the adaptive scoping method has two important advantages.

First, during the initial breeding cycles, the adaptive scoping method uses a higher scoping rate and thus better prevents the loss of favorable QTL alleles. The effect of the loss of favorable QTL alleles is clearly observed in Figures 6.3 and 6.4: during the first few breeding cycles, the maximum reachable genetic value of the scoping method decreases significantly whereas this is less pronounced for the adaptive scoping method.

Second, after t breeding cycles have been completed, the adaptive scoping method relies on a low scoping rate to efficiently convert the remaining genetic variation into genetic gain. From breeding cycle t onward, the scoping rate reaches its minimum value and the pre-selection procedure yields the same parental population as truncation selection (i.e., the individuals with the highest GEBVs). However, whereas truncation selection relies on a random crossing of parents, the adaptive scoping method constructs the crossing block using an identical procedure as the scoping method. The latter was demonstrated to result in an overall higher genetic gain (see Figure 5.5). As such, the adaptive scoping method allows for a better and more accurate exploitation of the remaining genetic variation toward the end of the pre-defined time window.

Except for the case where the optimization of a breeding population in a very short period of time is desired, the adaptive scoping method outperforms the (original) scoping method. In turn, the scoping method was demonstrated to outperform parental selection methods such as truncation selection, the population merit method and the maximum variance total method in simulation studies (see

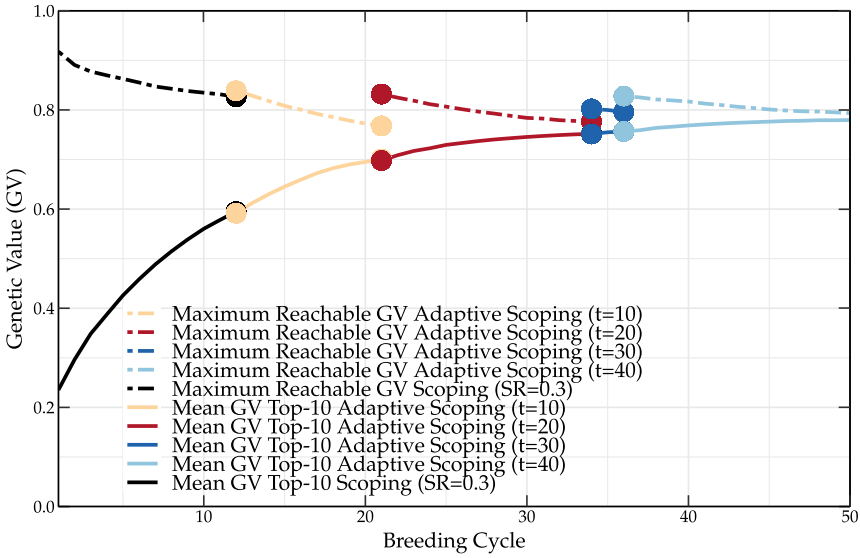

Figure 6.5: Simulation results of the scoping and adaptive scoping methods. At each breeding cycle, the mean genetic value of the top-10 individuals and the maximum reachable genetic value is depicted for the method and/or value of t that yields the highest genetic value. The adaptive scoping method yields superior genetic values during a short time window that follows breeding cycle t.

Figure 5.7). As such, the adaptive scoping method appears to be an attractive parental selection method.

6.5 Conclusion

We proposed the adaptive scoping method as an enhanced version of the original scoping method. By dynamically balancing genetic progress and genetic variation, we demonstrated its ability to maximize the genetic gain of a breeding population within a specific, predefined time frame of interest. This unique ability enables breeders to balance between exploration and exploitation of their breeding population: they can obtain fair genetic values in a relatively short term, or they can aim for the highest genetic values in the longer term. Regardless of this choice of time frame, the adaptive scoping method was shown to outperform the original scoping method.

7

The deep scoping method

Genomic prediction is often combined with truncation selection to identify superior parental individuals that can pass on favorable quantitative trait locus (QTL) alleles to their offspring. However, truncation selection reduces genetic variation within the breeding population, causing a premature convergence to a sub-optimal genetic value. In order to also increase genetic gain in the long term, different methods have been proposed that better preserve genetic variation. However, when the genetic variation of the breeding population has already been reduced as a result of prior intensive selection, even those methods will not be able to avert such premature convergence. Pre-breeding provides a solution for this problem by reintroducing genetic variation into the breeding population. Unfortunately, as pre-breeding often relies on a separate breeding population to increase the genetic value of wild specimens before introducing them in the elite population, it comes with an increased financial cost. In this chapter, on the basis of a simulation study, we propose a new method that reintroduces genetic variation in the breeding population on a continuous basis without the need for a separate pre-breeding program or a larger population size. This way, we are able to introduce favorable QTL alleles into an elite population and maximize the genetic gain in the short as well as in the long term without increasing the financial cost.

The material of this chapter is based on the following publication:
Vanavermaete, D., Fostier, J., Maenhout, S., De Baets, B., 2021. Deep scoping: a breeding strategy to preserve, reintroduce and exploit genetic variation. Theor Appl Genet, 1–17.

7.1 Introduction

Truncation selection is often used in genomic selection to rapidly increase the short-term genetic gain of a breeding population. By selecting individuals with the highest genomic estimated breeding values (GEBVs), breeders hope to maximally pass favorable properties to their offspring. The underlying idea is easy to understand and matches the gut feeling of most breeders, making it one of the most popular strategies in plant breeding. Unfortunately, truncation selection is also associated with a loss in genetic variation (Jannink, 2010). Besides entailing the loss of favorable QTL alleles from the breeding population, truncation selection causes a premature convergence of the genetic value, reducing the long-term genetic gain (see Figure 5.2). Therefore, truncation selection can only promise a temporary, short-term increase of the genetic gain. To ensure a continuous increase of the genetic value, new selection methods are needed that maximize both the short-term and the long-term genetic gains.

Different variants of truncation selection that try to remedy the loss in genetic variation have already been proposed in the literature. One way to achieve this is by weighting the marker effects of favorable or low-frequency marker alleles and thus reducing the risk of eliminating important QTL alleles during breeding (Jannink, 2010; Liu et al., 2015). The genetic variation can also be preserved by avoiding the selection of closely related individuals (as in the *population merit* method (Lindgren and Mullin, 1997)) or by penalizing the GEBV when two parents with high coancestry are selected (as in the *maximum variance total* method (Cervantes et al., 2016)). The latter was further improved upon by also minimizing the rate of inbreeding, thus controlling the allele heterozygosity as well as the allele diversity (Brisbane and Gibson, 1995; Akdemir and Sánchez, 2016). In another strategy, the GEBV was replaced by the *criterion of usefulness* (UC), which not only takes into account the mean predicted genetic value of the offspring, but also the selection intensity, prediction accuracy and genetic variation of the offspring (Lehermeier et al., 2017). The *scoping* method combines pre-selection with a score function to avoid the selection of individuals with a too low GEBV while preserving genetic variation of the breeding population, thus maximizing the long-term genetic gain (as observed in Figure 5.5). Whereas the GEBV is based on the total sum of the additive marker effects, the *optimal haploid value* (OHV) scores individuals based on their haplotypes, and can therefore better preserve favorable QTL alleles in the breeding population, increasing the long-term genetic

gain (Daetwyler et al., 2015). Müller et al. (2018) propose the expected maximum haploid breeding value (EMBV) to evaluate the potential of a candidate by measuring a limited number of gametes of each parent. The optimal cross selection (OCS) scores a crossing block based on the mean predicted genetic value of the offspring, but also constrains the loss in genetic diversity of the offspring (Akdemir and Sánchez, 2016; Gorjanc et al., 2018).

Unfortunately, the aforementioned methods are generally tested on breeding populations that demonstrate a broad genetic variation. In reality, however, the genetic variation present in most breeding populations has been eroded to some extent by years of consecutive truncation selection. In such cases, the options to further increase the genetic value in the breeding population are strongly reduced. To demonstrate this, we simulate three breeding populations that suffer, to a varying degree, from reduced genetic variation by applying respectively 0, 5, and 20 breeding cycles of truncation selection. Next, using these three breeding populations as a starting point, the performances of the population merit method (Lindgren and Mullin, 1997) and the scoping method (see Chapter 5) are compared. When these methods are initiated at a later point, the maximum reachable genetic value of the breeding population is lower, indicating that during truncation selection, favorable QTL alleles have been eliminated from the breeding population (see Figure 7.1). Both methods will only be able to preserve a fraction of the genetic variation that is still present in the breeding population. Therefore, the added value of these methods is dramatically reduced when the genetic variation in the breeding population is limited.

When the genetic variation has already been substantially reduced, a gene bank could be used to (re)introduce alleles and haplotypes into the breeding population, resulting in an increase of the maximum reachable genetic value. A gene bank is an (inter)national collection of different plants ranging from wild specimens to different crop varieties at different stages of selection. To optimally reintroduce genetic variation into a breeding population and thus increase the genetic gain in the long term, the gene bank must show a broad genetic variation (Simmonds, 1993; Salhuana and Pollak, 2006). The introduction of gene bank accessions into the breeding population generally implies a reduction in short-term genetic gain. Depending on the available germplasm collection of the gene bank, different methods have been proposed to introduce such individuals into an elite breeding population. When a phenotypic trait is controlled by only a few genes with large effects, the favorable genes can be introgressed in the breeding population using marker-assisted backcrossing (Han et al., 2017; Smith and Beavis, 1996). However, this proved unsuccessful when the phenotypic trait is controlled by many genes of small effect, which is the case for quantitative traits such as grain yield (Bouchez et al., 2002). In this setting, genomic selection (GS) can be used to rapidly introduce (new) QTL alleles from a gene bank into the breeding population (Bernardo, 2009). Different mating designs use multi-parental crosses

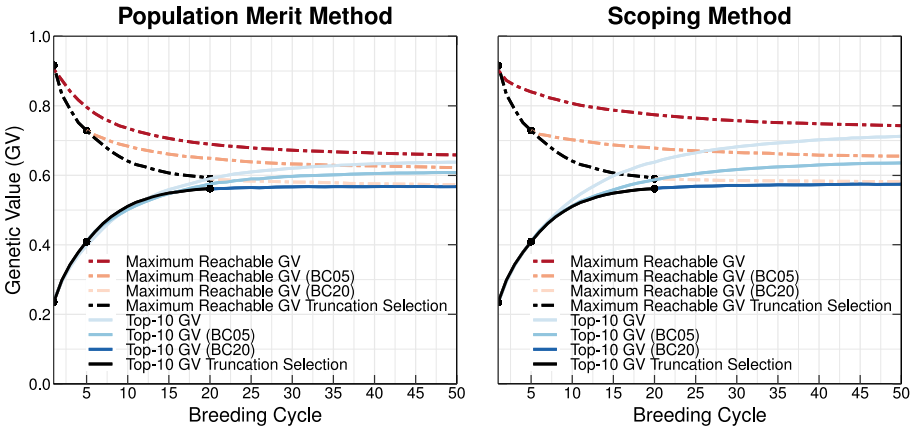

Figure 7.1: Mean genetic value (GV) of the top-10 individuals in the breeding population using the population merit method (left) and the scoping method (right) after first applying 0, 5, or 20 breeding cycles (BC) of truncation selection (black line). When the genetic variation of the breeding population is already reduced by means of truncation selection, both the population merit method and the scoping method result in a lower genetic value.

to combine elite individuals with donor individuals selected from a gene bank (Allier et al., 2019; Schopp et al., 2017). Gene bank accessions are first intercrossed to increase the frequency of favorable alleles before they are introduced into the breeding population. Cramer and Kannenberg (1992) proposed a five-year open-ended hierarchical breeding program (HOPE) to introduce new wild specimens into the breeding population using three consecutive gene pools. The HOPE method allows to effectively pass on favorable QTL alleles from the gene bank to the elite breeding population, but the need for additional pre-breeding populations drives up the total cost of the breeding program.

Allier et al. (2020a) recently proposed a new selection method, combining the *haploid estimated breeding value* (HEBV) and the UC to select and cross elite individuals with donor individuals. However, the calculation of the UC requires the construction of a covariance matrix, which considerably increases the computational requirements of the simulation while the bridging population can only reintroduce a fraction of the genetic variation into the breeding population. The parental selection can also be guided using *genotyping-by-sequencing* or related techniques, in which the relatedness of germplasm collections and elite individuals in the breeding population can be quantified and used to preserve the genetic variation in the breeding population (Glaubitz et al., 2014; Gouesnard et al., 2017). The genetic variation of a breeding population can also be increased by using exotic material, but a higher investment is needed to successfully incorporate those alleles in an elite breeding population (Salhuana and Pollak, 2006; Wu et al., 2016).

We propose a new method that incorporates the use of a gene bank to reintroduce genetic variation into the breeding population, maximizing the long-term genetic

gain without reducing the short-term genetic gain. By using a fraction of the breeding population for pre-breeding, the sizes of both the breeding population and the parental population remain unchanged, avoiding additional costs. This method, coined *deep scoping*, divides the breeding population into an elite population and different layers of pre-breeding individuals. The elite population contains the accessions that have the highest GEBVs and delivers high short-term genetic gain. In the first layer (Layer 0), individuals from a gene bank are crossed with individuals of the elite population, reintroducing genetic variation in the breeding population. Next, different layers are added, allowing for a gradual flow of favorable QTL alleles from the first layer to the elite population. Over each layer, the genetic variation is exploited, increasing the genetic gain and maximizing the transition of pre-bred individuals into the elite population.

7.2 Materials and methods

We adopt the base population and breeding scheme of Neyhart et al. (2017). The base population consists of two datasets of North American barley (Hordeum *vulgare*) from the University of Minnesota (UMN) and the University of North Dakota (NDSU), counting respectively 384 and 380 six-row spring inbred lines with 1590 biallelic SNP loci. The same base population was also used in Chapter 5, ensuring that the performance of the deep scoping method can be compared with that of the scoping method. The parental selection methods are compared using four base populations that differ in their available genetic variation for a single trait of interest. These four base populations (referred to as Population BC05, Population BC10, Population BC15 and Population BC20) are created by reducing the genetic variation using truncation selection in a recurrent breeding scheme for respectively 5, 10, 15, and 20 breeding cycles.

7.2.1 Breeding scheme

The recurrent breeding scheme depicted in Figure 7.2 has been described in Chapter 4. In this chapter, minor modifications are made to this scheme. Over the first breeding cycles, the recurrent breeding scheme is used to decrease the genetic variation of the breeding population. Starting at breeding cycle 0, based on phenotypic data, the top-50 individuals of the NDSU dataset are crossed with the top-50 individuals of the UMN dataset. In the subsequent breeding cycles, the parental selection is completely based on GEBVs, reducing the financial cost of phenotyping. The GEBVs are predicted based on a linear mixed effects model (see Subsection 7.2.6). In the recurrent breeding scheme, each parental couple is crossed 20 times, creating in total 1000 F1-hybrids. The F3-individuals are obtained after two

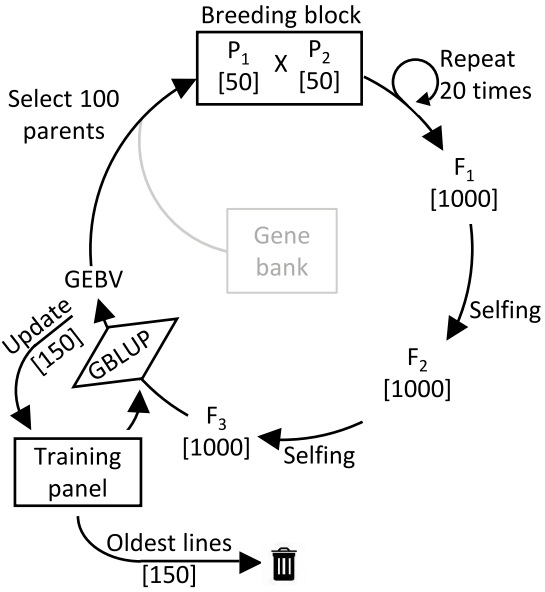

Figure 7.2: Overview of the recurrent breeding scheme. First, 50 couples of parents (P_1, P_2) each produce 20 offspring yielding a total of 1000 F1-hybrids. Then, after two generations of single-seed descent, 1000 F3-individuals are obtained. From those F3-individuals, new parental lines are selected. Three different parental selection methods are considered: i) Truncation selection selects 100 parents with the highest GEBVs and crosses them randomly; ii) The deep scoping method introduces new genetic information into the breeding population while maximizing the short- and the long-term genetic gain; iii) The HUC method with bridging introduces new genetic information into the breeding population by means of a bridging population.

cycles of single-seed descent. The recurrent breeding scheme is used to reduce the genetic variation of the breeding population by using truncation selection over 5, 10, 15 or 20 breeding cycles, selecting 100 parents with the highest GEBVs and crossing them at random. In the subsequent breeding cycles, the parents can be selected according to the deep scoping method or the HUC method with bridging. Additionally, both methods will also be able to select parents from a gene bank. Each simulation consists of 50 breeding cycles and all results are averaged over 100 simulation runs.

7.2.2 Truncation selection

Truncation selection selects 100 individuals with the highest GEBVs and couples them randomly. Breeders have been using truncation selection for centuries in the hope to pass favorable properties to the next generation. Unfortunately, this method also causes a strong reduction of the genetic variation. Therefore, truncation selection is an ideal and realistic method to simulate the loss of genetic variation in a breeding population as a result of selection.

7.2.3 Haploid estimated breeding values

In plant breeding, GEBVs are commonly used to select the parental population. Daetwyler et al. (2015) proposed the OHV as an alternative selection metric in which the highest genetic value of each haplotype segment is used instead of the marker effects. In theory, a haplotype segment contains several alleles and markers that are always inherited together, but in the OHV approach, each chromosome is divided into different haplotype segments containing an equal number of markers. A diploid individual contains n_H different haplotype segments and will have two haplotype values per segment representing the sum of the additive marker effects that are present in that segment on each homologous chromosome. The OHV is obtained by taking the sum of the highest haploid values per segment. In contrast to the GEBV, the OHV is better able to capture the potential benefits of heterozygous states in the breeding population. The HEBV proposed by Allier et al. (2020a) is similar to the OHV but allows for an overlap between the different haplotype segments. In this simulation study, the genotype is split into different haplotype segments containing 20 markers (window size) with an overlap of five markers (step size) (see Figure 7.3). The same simulation parameters were adopted as reported by Allier et al. (2020a) and remained unchanged during the whole simulation study to allow for a fair comparison between the different methods. A matrix $\mathbf{M}$ of size $k \times n_H$, with k the number of markers, is constructed to keep track of the selected markers per haplotype segment, such that $M_{ij} = 1$ if marker

i is part of the j-th haplotype segment and $M_{ij} = 0$ otherwise. Mathematically, the HEBV matrix $\mathbf{H}$ can be written as:

$$\mathbf{H} = (\mathbf{X} \circ \mathbf{1}_{2n}\boldsymbol{\beta}^T)\mathbf{M}\,, \tag{7.1}$$

with $\mathbf{X}$ a matrix of size $2n \times k$ containing the haplotype of n different individuals and k different markers coded as 0 and 1 (such that the haplotype of individual i is represented at rows $2i-1$ and $2i$), $\circ$ the Hadamard product operator, $\mathbf{1}_{2n}$ a vector of size $2n$ containing ones and $\boldsymbol{\beta}$ a vector of size k with estimated marker effects. Similar to the OHV, the HEBV between two individuals i and j is calculated as:

$$\text{HEBV}(i,j) = \lambda \sum_{h=1}^{n_H} \max\left(\mathbf{H}_{2i-1,h}, \mathbf{H}_{2i,h}, \mathbf{H}_{2j-1,h}, \mathbf{H}_{2j,h}\right), \tag{7.2}$$

with λ a scaling parameter defined as the ratio between the step size and the window size. If the step size and window size are equal, then $\lambda = 1$ and the HEBV reduces to the OHV.

In a breeding population, an elite subpopulation (denoted E), containing individuals with high GEBVs, can be distinguished. The H-score $H(i)$ of an individual i represents the maximal HEBV between this individual and any member of the elite subpopulation E (Allier et al., 2020a):

$$H(i) = \max_{j \in E} \text{HEBV}(i,j)\,. \tag{7.3}$$

In other words, an individual with a high H-score contains different favorable haplotype segments that are not available in the elite subpopulation (E) and should thus be selected as a parent.

7.2.4 The deep scoping method

The deep scoping method combines truncation selection with the (re)introduction of (new) QTL alleles in the breeding population with the aim of maximizing both the short- and long-term genetic gain. To introduce new QTL alleles, a gene bank is used, containing a population with a high genetic variation, but lower mean genetic value. When individuals of the gene bank are introduced into the breeding population, their lower genetic value prevents them from being selected during truncation selection. This will create a gap between the genetic value of the elite individuals and the rest of the breeding population, isolating them from one another. Although both QTL alleles will still be present in the breeding population, the QTL alleles of the individuals in the elite population will still be fixed causing a premature convergence of the genetic value. Therefore, a three-step selection

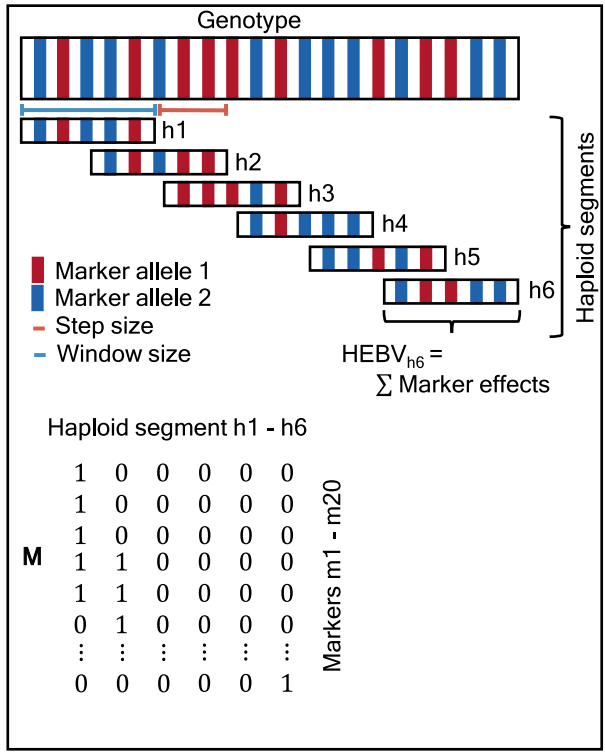

Figure 7.3: The haplotype is split into different haplotype segments containing an equal number of markers. Next, the HEBV is calculated per segment by summing up the marker effects of each segment separately.

procedure was designed to not only introduce QTL alleles into the breeding population but also in the elite population. To do so, the breeding population is divided into two subpopulations: the elite population and the pre-breeding population (see Figure 7.4). Individuals of the elite population are selected based on the highest GEBVs and are crossed to maximize short-term genetic gain. The selection of the pre-breeding population is divided into two steps: the selection for Layer 0 and the selection for Layers 1–4. For Layer 0, elite individuals are crossed with individuals from the gene bank to maximally introduce QTL alleles into the breeding population. The parental selection for the subsequent layers maximizes the flow of individuals between the pre-breeding population and the elite population, exploiting the genetic variation such that (new) favorable QTL alleles can be introduced into the elite population. Loosely inspired by deep learning (Ivakhnenko, 1971), the deep scoping method uses different layers in which individuals flow from one layer to the next, in the hope that the information that was once present in the first layer can be useful in the future and thus be transferred to the elite population.

The breeding population consists of an elite (sub)population containing 500 individuals and a pre-breeding (sub)population containing five different layers with each 100 individuals. In order to create the elite population, 50 individuals with the highest GEBVs are selected. In contrast to truncation selection, the parents are not crossed at random. The individual with the highest GEBV is selected as the P1 parent and is coupled with a P2 parent that minimizes the genetic relationship between both parents. Other crossing block designs have been considered as well, such as crossing the two individuals with the highest GEBVs with each other or crossing the top-50 individuals with the top 51-100 individuals, but both designs resulted in a significantly lower long-term genetic gain.

The pre-breeding population tries to introduce favorable marker alleles into the breeding population and ultimately in the elite population. To select the first parents for Layer 0, the HEBVs for the individuals of the gene bank are calculated. Next, the H-score is calculated for each individual of the gene bank. The five individuals with the highest H-score and thus containing the most favorable haplotype segments are selected as P1 parents. The five P2 parents are selected from the elite population to maximize the genetic value of the offspring. To maximize the genetic variation of the offspring, the scoping method is used instead of truncation selection. The scoping method has been proposed in Chapter 5 and consists of two important steps: the pre-selection and the parental selection. The pre-selection will select a fraction of the breeding population containing individuals with the highest GEBVs. Next, each selected P1 parent is crossed with a pre-selected individual that maximizes the S-score between both parents. The S-score between two individuals i and j is computed as:

$$S(i,j) = \sum_{m=1}^{k} \text{var}\{Z_{im}, Z_{jm}\} p_m \,, \tag{7.4}$$

with k the number of markers, $\mathbf{Z}$ a matrix of size $n \times k$ containing the genotype of n selected individuals and k different markers coded as -1, 0, or 1 and $\mathbf{p}$ a vector of size k with $p_m = 0$ if both alleles of marker m have been selected in the parental population or $p_m = 1$ otherwise. An individual with a high S-score contains different marker alleles that are not yet present in the parental population and should thus be selected as a parent. In contrast to the F_{score} that maximizes the genetic variation of a parental population, the S-score maximizes the genetic variation between two parents. It is possible that an individual of Layer 0 is selected as an elite P2 parent as long as it maximizes the genetic variation of the offspring.

The subsequent layers of the pre-breeding population gradually increase the genetic value of the Layer 0 individuals, while the genetic variation is slowly decreased such that favorable QTL alleles can be passed to the elite population. To allow for a continuous flow of favorable QTL alleles into the elite population, four additional layers are used. The effect of using a different number of layers will be discussed later (see Subsection 7.4.5). In the subsequent layers, the P1 parents are selected from the previous layer, selecting individuals with the highest H-score. This ensures that individuals with favorable haplotype segments can flow to the next layer. The P2 parents are selected such that the genetic value of the offspring is maximized while preserving the genetic variation as much as possible. Individuals of previous layers are not considered as potential parents because they could reduce the genetic value of the offspring and thus interrupt the flow of QTL alleles in the breeding population. Both pre-selection and the S-score are used to select the P2 parent. First, based on the GEBV, candidate parents are pre-selected. Next, P2 parents are selected such that the S-score is maximized between both parents. In the parental selection for Layer 1, the top-400 individuals are pre-selected and can thus be used to select the P2 parents. In the parental selection for the subsequent layers, the number of individuals that are pre-selected decreases over each layer to increase the genetic gain. The parental selection for Layer 2 only pre-selects 300 individuals, followed by 200 and 100 individuals for the selection for Layer 3 and Layer 4, respectively. Again, it is possible that an elite parent is also selected as a pre-breeding parent as long as it maximizes the genetic variation of the offspring. The use of the scoping method during the parental selection helps to preserve the genetic variation, allowing for a slower but more accurate fixation of the QTL alleles. Individuals of the fourth and last layer should have the highest genetic values and could therefore be selected during truncation selection, finally introducing favorable QTL alleles into the elite population. Note that the elite population selects the individuals with the highest GEBVs over the entire breeding population, making it possible to select individuals of any layer into the elite population as long as the GEBV is high enough.

The implementation of the deep scoping method will require several breeding cycles. Starting with a truncation-selected breeding population, when the deep scoping method is used for the first time, the parental selection for Layer 0 crosses in-

dividuals of the elite population with individuals of the gene bank, but the parents of the subsequent layers will still be selected from truncation-selected individuals. In the next breeding cycle, the parental selection for Layer 1 crosses individuals of Layer 0 with individuals of the elite population, but the parents for Layers 2–4 will be selected from the offspring of truncation-selected individuals. For each layer that is used in the deep scoping method, one additional breeding cycle will be required before the deep scoping method becomes fully operational.

7.2.5 The HUC method with bridging

The HUC method combines the HEBV and the UC to select the parental population. A full description of the HUC method has been reported by Allier et al. (2020a). The HUC method combines an elite population (PopE) with a second donor population (PopD). The donor population is selected from a gene bank containing 500 different individuals. First, individuals with the highest GEBVs are selected as elite parents. Next, the individuals in the donor population with the highest H-scores are selected as donor parents. Next, a crossing block between the selected parents from the elite population and the donor population is built by maximizing the UC, which is calculated as:

$$U = \hat{\mu}_p + i\rho\hat{\sigma}_p\,, \tag{7.5}$$

with U the UC, $\hat{\mu}_p$ the predicted mean genetic value of the progeny, i the selection intensity, ρ the model performance and $\hat{\sigma}_p$ the predicted genetic variance of the progeny. Both parameters i and ρ are kept constant during the entire simulation. The UC was calculated using the implementation and parameter settings as published by Allier et al. (2020a) with $i = 2.06$ representing a selection intensity of 5% and $\rho = 1$.

In our simulation study, the genetic value of the individuals of the gene bank is low. In such case, to allow for a fair comparison between the HUC method and the deep scoping method, the HUC method should be extended with a bridging population to assist the introduction of the individuals of the gene bank into the elite population (Allier et al., 2020b). This means that the breeding population is split into two parts: an elite population and a pre-breeding population. According to Allier et al. (2020b), 75% of the parental population is used to select the elite population, while the remaining 25% is used to select the pre-breeding individuals. Because the recurrent breeding scheme used in our simulation study requires the selection of an even number of parents, 80% of the parental population is used to select the elite population and the remaining 20% is used to select the pre-breeding population. In the elite population, 80 individuals with the highest GEBVs are selected and crossed using truncation selection as described in the deep scoping method. In the pre-breeding population, 10 elite individuals are crossed with 10 individuals of the gene bank (donors) according to the HUC method.

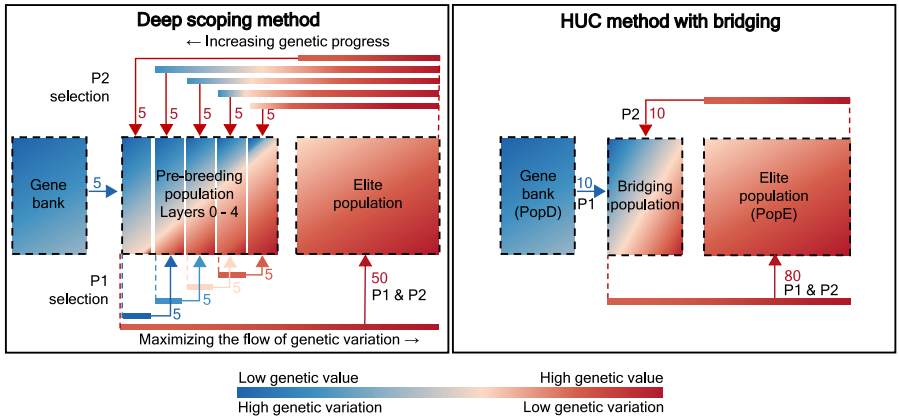

Figure 7.4: Left panel: overview of the deep scoping method. First, the individuals with the highest GEBVs are selected as elite parents. Next, for each layer, individuals from the elite population and the previous layer are selected. For Layer 0 elite individuals are combined with individuals of the gene bank. For each of the subsequent layers, the individuals can mature in the breeding population, increasing their genetic value, while the genetic variation is gradually decreased. Right panel: overview of the HUC method with bridging. First, the individuals with the highest GEBVs are selected as elite parents. Next, individuals of the gene bank with the highest H-scores are selected and crossed with individuals of the elite population containing the highest GEBVs.

7.2.6 Prediction model

The GEBVs are predicted by fitting a linear mixed effects model:

$$\mathbf{y} = \mathbf{1}_n\beta + \mathbf{Zu} + \boldsymbol{\epsilon}\,, \tag{7.6}$$

with $\mathbf{y}$ a vector of phenotypic values, $\mathbf{1}_n$ a vector of size n containing ones, n the number of individuals in the training panel, β the fixed effect (phenotypic mean), $\mathbf{Z}$ the incidence matrix of the training panel with marker information, $\mathbf{u}$ the marker effects following a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{G})$ with $\mathbf{G} = \sigma_u^2\mathbf{I}_k$ (with $\mathbf{I}_k$ the identity matrix of dimension k), k the number of markers and $\boldsymbol{\epsilon}$ the residual effects following a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \sigma_e^2\mathbf{I}_n$. Both variance components σ_u^2 and σ_e^2 are estimated by means of Restricted Maximum Likelihood using the rrBLUP package (Endelman, 2011). The GEBVs of the individuals are calculated as:

$$\hat{\mathbf{g}} = \mathbf{Z}\hat{\mathbf{u}}\,, \tag{7.7}$$

with $\hat{\mathbf{g}}$ the GEBVs, $\mathbf{Z}$ the marker information and $\hat{\mathbf{u}}$ the predicted marker effects.

In the first breeding cycle, the complete base population is used as a training panel. In the subsequent breeding cycles, 150 individuals are phenotyped and added to the training panel according to the *tails* method, selecting 75 individuals based on the tails of the normally distributed GEBVs (Neyhart et al., 2017). Ac-

cording to Neyhart et al. (2017), the tails method delivers a non-significant higher genetic gain compared to other update methods. In the case of the deep scoping method, the tails method builds a training panel with elite individuals and pre-breeding individuals improving the prediction of GEBVs of the whole breeding population without the need for two separate prediction models. Each time the training panel is updated, 150 individuals that have been longest in the training panel are removed from the training panel to reduce computational time without reducing the prediction accuracy (Neyhart et al., 2017). To calculate the UC, the Markov chain Monte Carlo (MCMC) samples of the marker effects are required. This matrix is obtained by estimating the GEBVs that are used in the HUC method via the BGLR package using a Gibbs sampler with Gaussian prior (BRR) (Allier et al., 2020a; Pérez and de los Campos, 2014).

7.2.7 Simulation of the population

The simulation is built upon the work of Neyhart et al. (2017), using the packages `GSSimTPUpdate` and `hypred` in R (version 3.6.3). The dataset contains 1590 biallelic SNP markers from which 100 are selected as QTLs ($L = 100$) and 1490 are used as markers to predict the genetic value. The true phenotypic value of the i-th individual (y_i) is calculated over three different environments:

$$y_i = \frac{1}{3}\sum_{j=1}^{3} g_i + e_j + \epsilon_{ij}\,, \tag{7.8}$$

with g_i the genetic value of the i-th individual, e_j the j-th environmental effect, and ϵ_{ij} the residual effect of the i-th individual and the j-th environment. The genetic value is calculated by taking the sum of the QTL effects. The QTL effects are sampled from a geometric series such that at the k-th QTL, the favorable homozygote has a value of a^k, the unfavorable homozygote has a value of $-a^k$ and the heterozygote has a value of zero with $a = (L-1)/(L+1)$. Both the environmental and residual effects are drawn from a normal distribution with mean 0 and a variance component σ_E^2 and σ_e^2, respectively. The variance component of the environmental effect is defined as eight times the genetic variance, while the variance component of the residual effect is scaled to simulate a heritability of 0.5 (Bernardo, 2014).

The simulation of the different breeding cycles is described in Chapter 4. In this chapter, a gene bank is added to the simulation. The gene bank is created by crossing individuals of the UMN population with individuals of the NDSU population. First, individuals of the UMN dataset are selected at random. For each parent, an individual of the NDSU dataset is selected that maximizes the S-score between both parents. The size of the gene bank is set at 500 individuals delivering a good balance between the preservation of the genetic variation of the base population

and keeping the simulation time low.

7.3 Results

7.3.1 Truncation selection

To simulate a realistic initial breeding population, truncation selection is used to reduce the genetic variation. At breeding cycle zero, both alleles are present in the breeding population at 92% of the marker sites (see Figure 7.5). Using truncation selection, the average genetic value of the breeding population increases while QTL alleles get fixed. The maximum reachable genetic value represents the sum of the QTL effects that are fixed (both favorable and unfavorable) and the sum of the favorable QTL effects that are not yet fixed. In other words, it represents the maximum genetic value that could still be reached, taking into account the fixation of unfavorable QTL alleles that has already occurred. A decrease in the maximum reachable genetic value, as observed in Figure 7.5, indicates that favorable QTL alleles are eliminated from the breeding population. This causes a convergence to sub-optimal genetic values.

To assess different selection strategies, we consider the mean genetic value of only the top-10 individuals in the breeding population. This reflects the genetic value of the elite individuals that are candidates for commercialization. It allows for a better comparison between the different methods because the mean genetic value of the entire breeding population will be negatively influenced when individuals of a gene bank are introduced even if the genetic value of the elite individuals remains unchanged.

7.3.2 The deep scoping method

The deep scoping method relies on a gene bank to introduce new genetic material into the breeding population. Next, the genetic value of the pre-breeding individuals is increased facilitating their transition into the elite population. In the first scenario, the deep scoping method is used after five breeding cycles of truncation selection (Population BC05). Once the gene bank is available, the newly-introduced marker alleles lead to an increase of the maximum reachable genetic value (see Figure 7.6). In the short term, the deep scoping method reaches the same genetic values as truncation selection. However, truncation selection result in a premature convergence, while the genetic value of the deep scoping method continues to increase, resulting in an 18% points higher genetic value in the long term (see

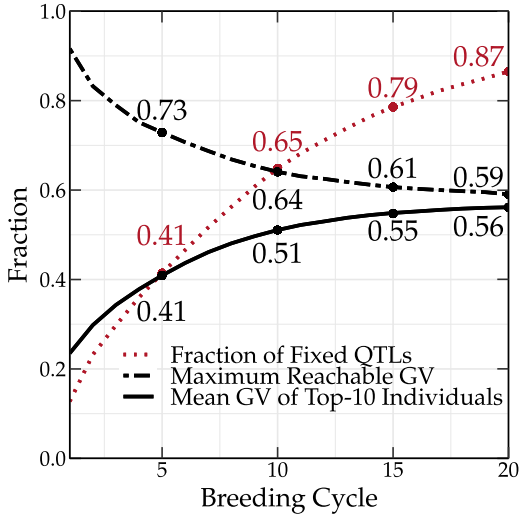

Figure 7.5: Simulation results of truncation selection over 20 breeding cycles. When truncation selection is used, the mean genetic gain of the top-10 individuals increases rapidly. Unfortunately, truncation selection also causes fixation of unfavorable QTL alleles, leading to a decrease in maximum reachable genetic value and causing a premature convergence of the mean genetic value of the top-10 individuals.

Tables 10.1 and 10.8). In the other scenarios, truncation selection is used for 10, 15, or even 20 breeding cycles, resulting in a breeding population with a higher number of fixed QTL alleles. Again, when the deep scoping method is used, the maximum reachable genetic value increases rapidly. The genetic value differs according to the starting point at which the deep scoping method is first invoked. The longer truncation selection is used, the longer it takes before a breeding population reaches a certain genetic value. Certainly, when 20 breeding cycles of truncation selection have been used, several breeding cycles of the deep scoping method are needed before the genetic value can escape from the local optimum. In the (very) long term, the four different scenarios will converge to the same value.

In Figure 7.7, the flow of individuals between the different layers of the deep scoping method after five initial breeding cycles of truncation selection is illustrated. The genetic value of the individuals over the first two layers is still too low, limiting their selection into the elite population. The individuals of Layers 3–4 have a higher genetic value, allowing for the transition of approximately one to two individuals into the elite population over each breeding cycle.

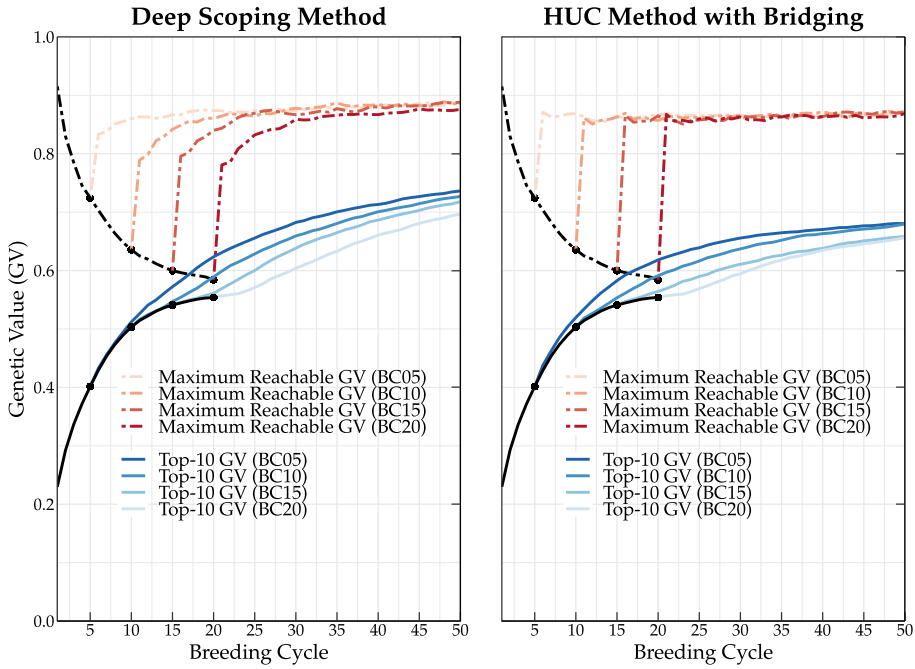

Figure 7.6: Simulation results of the deep scoping method and the HUC method with bridging starting at breeding cycles 5, 10, 15, and 20. Prior to the selection methods, truncation selection is used to reduce the genetic variation of the breeding population (black line). By deploying a gene bank, new QTL alleles are introduced into the breeding population, increasing the maximum reachable genetic value and avoiding a premature convergence of the genetic value. Compared to the HUC method with bridging, the deep scoping method can introduce more QTL alleles into the breeding population leading to a higher maximum reachable genetic value. The deep scoping method reaches a higher mean genetic values of the top-10 individuals in the long term compared to the HUC method with bridging.

7.3.3 The HUC method with bridging

The HUC method with bridging combines two different parental selection schemes. On the one hand, elite individuals with the highest GEBVs are selected and crossed with each other, maximizing the short-term genetic gain. On the other hand, individuals of the gene bank are crossed with elite individuals to introduce QTL alleles into the breeding population. By crossing the individuals of the gene bank with an elite individual, the GEBVs of the offspring increase such that they can be selected as an elite parent in the next breeding cycle, maximizing the long-term genetic gain. However, as individuals of the gene bank generally have a lower mean genetic value compared to the elite individuals, crossing individuals of the gene bank with elite individuals will mostly yield offspring with mediocre genetic values. This will prevent the selection of these pre-breeding individuals into the elite population, disrupting the introduction of genetic variation, thus resulting in a premature convergence of the genetic value of the elite population. Increasing the number of initial breeding cycles using truncation selection (e.g. Population BC20) increases

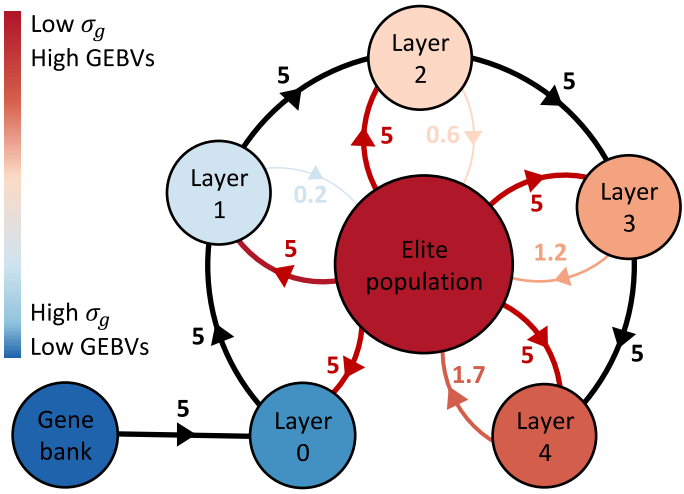

Figure 7.7: Overview of how individuals flow between the different subpopulations when the deep scoping method is applied on a breeding population after five initial breeding cycles of truncation selection. A color scheme is used to indicate the change in genetic value and genetic variation of the different subpopulations. The black and dark red arrows represent the selection of respectively the first and second parent. From the first two layers, very few individuals are selected into the elite population. After progressing over three to four layers, an average of one to two parents per breeding cycle is accepted into the elite population.

the gap between the genetic value of the breeding population and the gene bank, resulting in the convergence to even lower genetic values (see Figure 7.6).

The mean genetic values of the top-10 individuals of the proposed methods are reported in Tables 10.1 and 10.8.

7.3.4 Robustness of the deep scoping method

The deep scoping method has been tested in different simulation settings. Each experiment consists of testing 100 different genomes such that the effects of using the deep scoping method or the HUC method with bridging can be studied using different QTL and marker positions. The effect of the heritability and of the number of QTLs on the genetic gain of both methods have also been tested and are shown in Figure 7.8. Simulation studies were performed using a heritability of 0.2, 0.5, and 0.8 using 100 different QTLs, and a heritability of 0.5 using 50, 100, and 200 QTLs. In all six cases, the deep scoping method resulted in higher genetic values in the long term compared with the HUC method with bridging. Regardless of the heritability, number of QTLs, or the QTL and marker positions, the deep scoping method outperformed the HUC method with bridging in the long term.

7.4 Discussion

7.4.1 Introducing and preserving the genetic variation in the breeding population

The primary goal of the deep scoping method is to preserve newly-introduced marker alleles in the breeding population by allowing for a gradual flow of genetic material from the pre-breeding population into the elite population. To achieve this, the scoping method has been redesigned. In the original scoping method, the first parent was greedily selected to prioritize the genetic progress, while the second parent was selected by maximizing the S-score between both parents, preserving the genetic variation of the breeding population. In the deep scoping method, the individuals with the highest H-scores are selected as P1 parents, prioritizing the preservation of favorable marker alleles in the breeding population. Similar to the scoping method, the second parent is selected by maximizing the S-score between both parents. Since both the H-score and the S-score quantify genetic variation, the second parent is taken from a pre-selected population that contains individuals with the highest GEBVs. This way, genetic progress is maximized as well, facilitating the transition of pre-breeding individuals into the elite

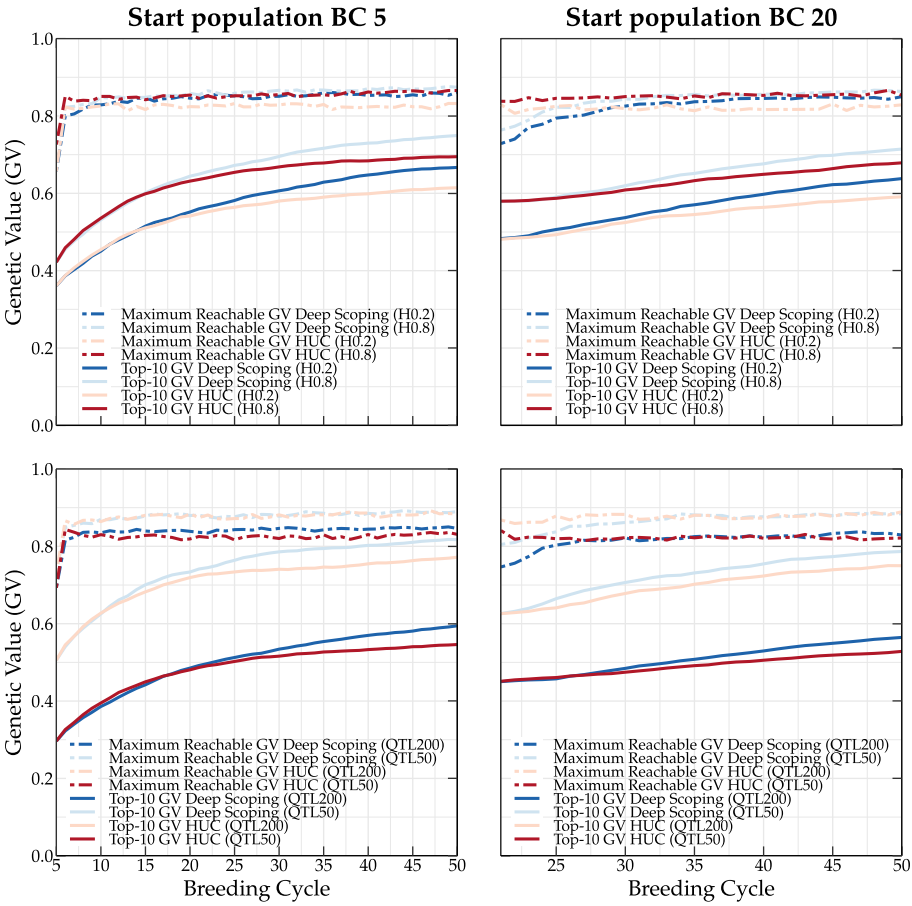

Figure 7.8: Simulation results of the deep scoping method and the HUC method with bridging for a heritability of 0.2 and 0.8 using 100 QTLs (top) and for a heritability of 0.5 using 50 and 200 QTLs (bottom line). The impact of both methods on the genetic value and on the maximum reachable genetic value is shown after 5 (left) and 20 (right) breeding cycles of truncation selection. In all six cases, the deep scoping method resulted in higher genetic values in the long term compared with the HUC method with bridging.

population. Moreover, by using the S-score, both alleles of each marker will be preserved in the breeding population to the extent possible, thus minimizing the loss of (favorable) QTL alleles. By decreasing the size of the pre-selection fraction over each layer, the genetic variation will gradually decrease over each layer while the average genetic value increases.

The deep scoping method continuously introduces genetic material into the breeding population and therefore, the loss of genetic variation of the elite population is not an issue. In the case of the scoping method, when an unfavorable QTL allele was fixed in the breeding population, the maximum reachable genetic value was reduced causing a lower genetic value in the long term. In the case of the deep scoping method, the different QTL alleles are still preserved in the pre-breeding population or the gene bank, and when an unfavorable QTL allele is fixed in the elite population, it is thus still possible to introduce the corresponding favorable QTL allele into the next breeding cycles. Nevertheless, enforcing the preservation of genetic variation remains important to maximize the genetic gain over each breeding cycle.

7.4.2 Population size and required resources

The deep scoping method does not require a separate pre-breeding program and is able to improve the individuals of the gene bank with the same resources as a breeding program without pre-breeding, minimizing the cost of the deep scoping method. This means that with the same population size (and resources) as truncation selection, the deep scoping method reintroduces genetic variation into the breeding population and maximizes the genetic gain thereof. Only the gene bank could be seen as an additional investment. The gene bank contains 500 individuals that are created by crossing individuals of both the UMN and the NDSU dataset by maximizing the genetic variation between both crosses. This means that the individuals of the gene bank will have different heterozygous markers and low genetic values. This way, the ability to include low-GEBV individuals of both the deep scoping method and the HUC method with bridging was studied. Including more exotic germplasm could require more layers, whereas the inclusion of individuals with a higher genetic value could also be done with a lower number of layers.

7.4.3 GEBVs versus HEBVs

In genomic selection, GEBVs are often used to select the parental population. To do so, a mixed effects model is used to predict the marker effects, which are used to calculate the GEBVs according to Eq. (4.6). The genotype of an individual is

often represented by bi-allelic markers coded as -1, 0, or 1. Assuming a positive marker effect u_m, the m-th marker of an homozygous individual i will yield a positive contribution if the individual carries the reference allele ($Z_{im} = 1$) or a negative contribution if the alternative allele ($Z_{im} = -1$) is present. If the individual is heterozygous ($Z_{im} = 0$), meaning that both alleles are present once, the marker does not affect the GEBV despite the fact that it still has a 50% probability to pass the favorable marker allele to the next generation. In other words, the GEBVs penalize heterozygous markers despite the fact that they contain favorable marker alleles. Taking into account that the genetic information of thousands of markers is reduced to a single value, unfavorable QTL alleles could be fixed into the breeding population when their negative marker effect is masked by many other positive QTL effects, or favorable QTL alleles could be eliminated from the breeding population when their QTL effects are masked by many other negative QTL effects.

A parental selection based on GEBVs could also lead to the selection of closely related individuals, fixating several favorable and unfavorable QTL alleles in the breeding population. This could be avoided by penalizing the GEBVs to minimize the rate of inbreeding and reducing the loss in genetic variation (Cervantes et al., 2016; Akdemir and Sánchez, 2016). Nevertheless, reducing the genetic information of thousands of markers into a single value remains a major disadvantage of the GEBVs and therefore, the HEBVs are used instead. In contrast to GEBVs, HEBVs split the genotype into different haplotypes and each haplotype is compared among the selected individuals. The available genotypic information is no longer reduced into a single value, lowering the probability to mask certain QTL effects and thus avoiding the elimination of one or more favorable QTL alleles. The HEBV score is calculated by taking the maximum genetic value between the haplotype segments of different individuals, which means that heterozygous alleles could have the same contribution as the favorable homozygous marker, preserving all the favorable QTL alleles in the breeding population. The H-score is an extension of the HEBV and scores the ability of an individual to bring new favorable haplotype segments into the elite population, maximizing the genetic preservation by both the HUC method with bridging and the deep scoping method.

Replacing the GEBVs with the HEBVs in the deep scoping method made it possible to reach higher long-term genetic gains. Until now, only GEBVs were used to preselect the individuals in the deep scoping method. We could assume that the HEBVs could further maximize the long-term genetic gain of the scoping method. Unfortunately, as depicted in Figure 7.9, using the HEBVs reduces the genetic gain while the maximum reachable genetic value slightly increases. In the deep scoping method, the HEBVs were only used to preserve the genetic gain of the different layers. The elite population was still selected based on GEBVs, resulting in these high genetic gains. Using the HEBVs in the scoping method will allow for the preselection of individuals with a lower genetic value, decreasing the genetic progress

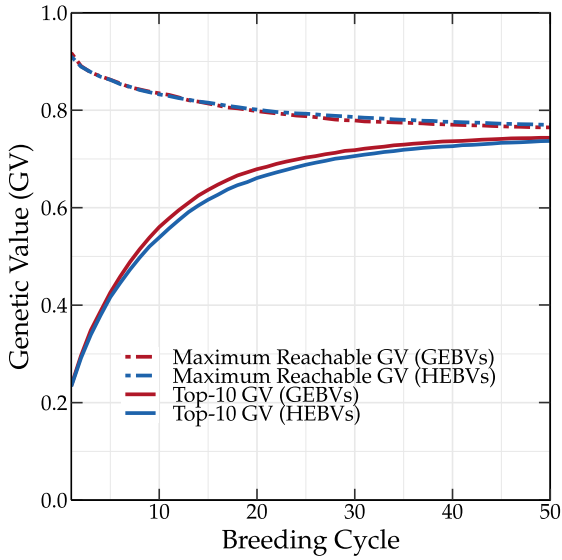

Figure 7.9: Genetic values of the scoping method (SR=0.3) using the GEBVs and the HEBVs as a selection criterion. Replacing the GEBVs with HEBVs resulted in a lower genetic value in the short as well as in the long term.

of the population. We can conclude that the HEBVs allow for a better preservation of the genetic variation, but with a lower genetic progress. Combining the HEBVs and GEBVs therefore seems the best solution, maximizing both the genetic gain and the genetic variation of the breeding population.

7.4.4 Comparison of the HUC method and the deep scoping method

As both the HUC method and the deep scoping method rely on the use of a gene bank, it is possible to make a fair comparison between these methods. The HUC method with bridging is a combination of, on the one hand, the HUC method and, on the other hand, a breeding scheme with a bridging population (Allier et al., 2020a,b). Originally, the HUC method selects individuals based on the H-score in order to cross elite individuals with a donor population. The donor population contains all the individuals of the breeding population that are not selected in the elite population and the individuals of the gene bank. The elite population often contains closely related individuals with low genetic variation; therefore, individuals of the gene bank will have a high H-score and will be selected as a parent, reducing the genetic value of the offspring. Originally, the HUC method was designed to introduce individuals with a similar or slightly lower genetic value than

the elite population. However, when a gene bank is used containing individuals with low GEBVs, the HUC method fails to increase the genetic value of the breeding population. For this case, Allier et al. (2020b) designed a new breeding scheme that incorporates a bridging population. However, this breeding scheme uses the GEBVs to select the parental population. By using the HUC method in a breeding scheme with bridging, we were able to compare the deep scoping method with the HUC method, both using the HEBVs as selection criterion.

Both the HUC method with bridging and the deep scoping method use truncation selection to maximize the short-term genetic gain. The deep scoping method consists of two different populations: the elite population and the pre-breeding population, which is built up out of five different layers. The size of the elite population will be smaller compared to that of the HUC method with bridging, which only contains two populations: the elite population and the bridging population. In the HUC method with bridging, 10 individuals of the gene bank and 10 individuals of the elite population are crossed. The deep scoping method only selects five individuals of the gene bank and crosses them with the elite individuals. Therefore, the HUC method with bridging will be able to reintroduce more genetic variation into the breeding population after one breeding cycle. The deep scoping method will need several breeding cycles before reaching the same maximum reachable genetic value as the HUC method with bridging. As long as pre-breeding individuals are not selected in the elite population, the same individuals of the gene bank will be selected into Layer 0, therefore, two breeding cycles of deep scoping will not result in the same amount of genetic variation as observed after one breeding cycle using the HUC method with bridging. This also explains why, when the deep scoping method starts at breeding cycle five, a higher increase in the maximum reachable genetic value is observed for the HUC method with bridging. Increasing the size of Layer 0 in the deep scoping method may increase the maximum reachable genetic value in a similar way as in the HUC method with bridging, but because it does not influence the genetic gain in the short or the long term, increasing the size of Layer 0 is not necessary.

The pre-breeding is similar in both methods, however, the HUC method with bridging builds a crossing block by maximizing the UC, whereas the deep scoping method does this by using the S-score. In other words, the HUC method pairs up parents to maximize the genetic gain of the offspring, while the deep scoping method pairs up parents to maximize the genetic variation of the offspring. The deep scoping method also introduces four additional layers that are used to guide the development of pre-breeding individuals into elite individuals and thus to facilitate the flow of (favorable) QTL alleles into the elite population. In the HUC method with bridging, the selected individuals of the gene bank only have one breeding cycle to be selected as elite individuals. Therefore, when the difference in genetic value between the gene bank and the breeding population increases, the transition of pre-breeding individuals into the elite population degrades, causing a

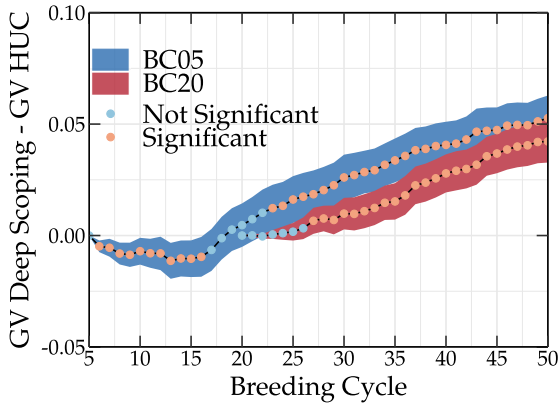

Figure 7.10: Mean genetic value of the top-10 individuals is compared between the deep scoping method and the HUC method with bridging using a paired sampled t-test, for two scenarios with respectively 5 and 20 initial breeding cycles of truncation selection. The difference in mean genetic value with a 95% confidence interval is reported. The color of each dot indicates whether the difference in genetic value between both methods is significant ($p < 0.05$).

premature convergence of the genetic value.

Although the deep scoping method results in a 4–6% points higher long-term genetic gain compared with the HUC method with bridging, in the short term, a slightly lower genetic value is observed (see Figure 7.10). The elite population of the HUC method represents 80% of the breeding population, whereas in the deep scoping method only 50% of the breeding population is used. Therefore, when both methods are used after five breeding cycles of truncation selection, the HUC method will be able to select more elite individuals to convert the remaining genetic variation of the breeding population into genetic gain resulting in significantly higher genetic values in the short term compared with the deep scoping method. The HUC method with bridging also selects more individuals from the gene bank, increasing the genetic variation in the breeding population, which also contributes to the maximization of the short-term genetic gain. However, when both the HUC method with bridging and the deep scoping method are used after 20 breeding cycles of truncation selection, the genetic variation of the breeding population has been reduced and the HUC method with bridging will be unable to gain higher short-term genetic values compared with the deep scoping method.

7.4.5 Flow from the pre-breeding population into the elite population

The deep scoping method uses different layers to guide the genetic progress of the pre-breeding individuals and to facilitate their transition into the elite popula-

tion. Each layer selects parents from the previous layer and couples them with an individual of the elite population, maximizing the genetic gain of the offspring. Prior to the deep scoping method, truncation selection is used to simulate a realistic breeding population. When the deep scoping method is introduced in the breeding program, an additional breeding cycle per layer will be required before the deep scoping method becomes fully operational. Once that is done, the number of layers represents the number of breeding cycles a pre-breeding individual is allowed to be crossed with an elite individual to increase the genetic value of their offspring and thus to pass on their genotypic information into the elite population. Individuals that are not selected in Layer 4 will be eliminated from the breeding population.

Figure 7.11 illustrates the flow of pre-breeding individuals into the elite population. In the first layer, individuals are rarely accepted into the elite population, except for the first breeding cycle after the introduction of the deep scoping method. This indicates that the use of a single layer (like in the HUC method with bridging) is insufficient to properly introduce pre-breeding individuals into the elite population. In the second layer, the transition of individuals into the elite population is still limited and it is only after progressing over three or four layers that a more substantial flow of pre-breeding individuals into the elite population can be observed. Each layer that is added to the deep scoping method allows for a better development of the pre-breeding individuals, which will facilitate their transition into the elite population. However, to avoid additional financial costs, the parental population size is kept constant, meaning that increasing the number of layers in a breeding population will also decrease the number of parents per layer and could therefore reduce the flow of QTL alleles from the pre-breeding population into the elite population. When the number of individuals per layer decreases, the probability to develop potentially interesting individuals also decreases causing lower genetic gains in the long term. This can be avoided by increasing the size of the parental population, but that will increase the total financial cost. Therefore, we recommend the use of four layers, allowing for enough time to develop pre-breeding individuals in the breeding population.

In Figure 7.12, the genetic value is shown for a population size of 500, 1000, and 2000 individuals using a different number of layers. Increasing the number of layers in a breeding population will decrease the number of individuals (and parents) per layer, but will allow for more time to increase the genetic value of the pre-breeding individuals increasing the flow of individuals from the gene bank to the elite population. Therefore, increasing the number of layers will often result in higher long-term genetic gains. Nevertheless, when the number of individuals per layer becomes too small, the probability to develop potentially interesting individuals is reduced, causing a lower long-term genetic gain. This can be avoided by increasing the size of the parental population, but that will also increase the total financial cost.

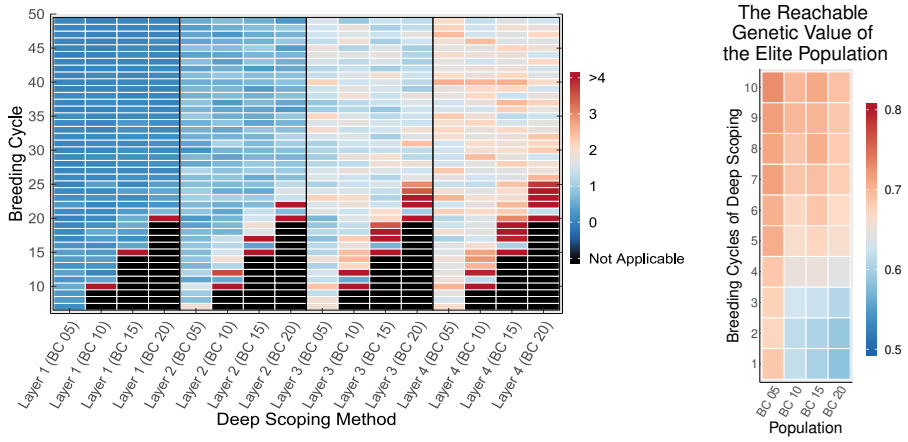

Figure 7.11: Left: an overview of the mean number of individuals that are selected in the elite population for each subpopulation. Right: the maximum reachable genetic value of the elite population for the first ten breeding cycles of the deep scoping method after 5, 10, 15 and 20 breeding cycles of truncation selection.

7.4.6 Applying the deep scoping method

Regardless of the fact whether the breeding population underwent 5, 10, 15, or 20 breeding cycles of truncation selection, when the deep scoping method is used, for each layer, a similar flow of individuals to the elite population is observed (see Figure 7.11). The longer truncation selection is used prior to the deep scoping method, the longer it will take to reach the same long-term genetic gain (see Figure 7.14). When the deep scoping method is initiated after only five breeding cycles of truncation selection, the different layers, still filled with the offspring of truncation-selected individuals, will be able to maximize the genetic value in the short term until the pre-breeding individuals are introduced to maximize the long-term genetic gain in the elite population. However, when the genetic variation of the breeding population is reduced after e.g. 20 breeding cycles of truncation selection, the genetic value of the base population has already converged. At that point, at least five breeding cycles will be required before individuals of the pre-breeding population will be accepted in the elite population, reintroducing genetic variation in the elite population that will allow the breeding population to escape from the local optimum (see Figure 7.11 right). If the deep scoping method would have been invoked after only five breeding cycles of truncation selection, the mean genetic value of the top-10 individuals would have been 8% points higher at breeding cycle 25 (see Figure 7.14). In other words, the sooner the deep scoping method is adopted in a breeding program, the sooner the breeding population will produce higher genetic gains compared with the same breeding population using truncation selection (see Figure 7.14).

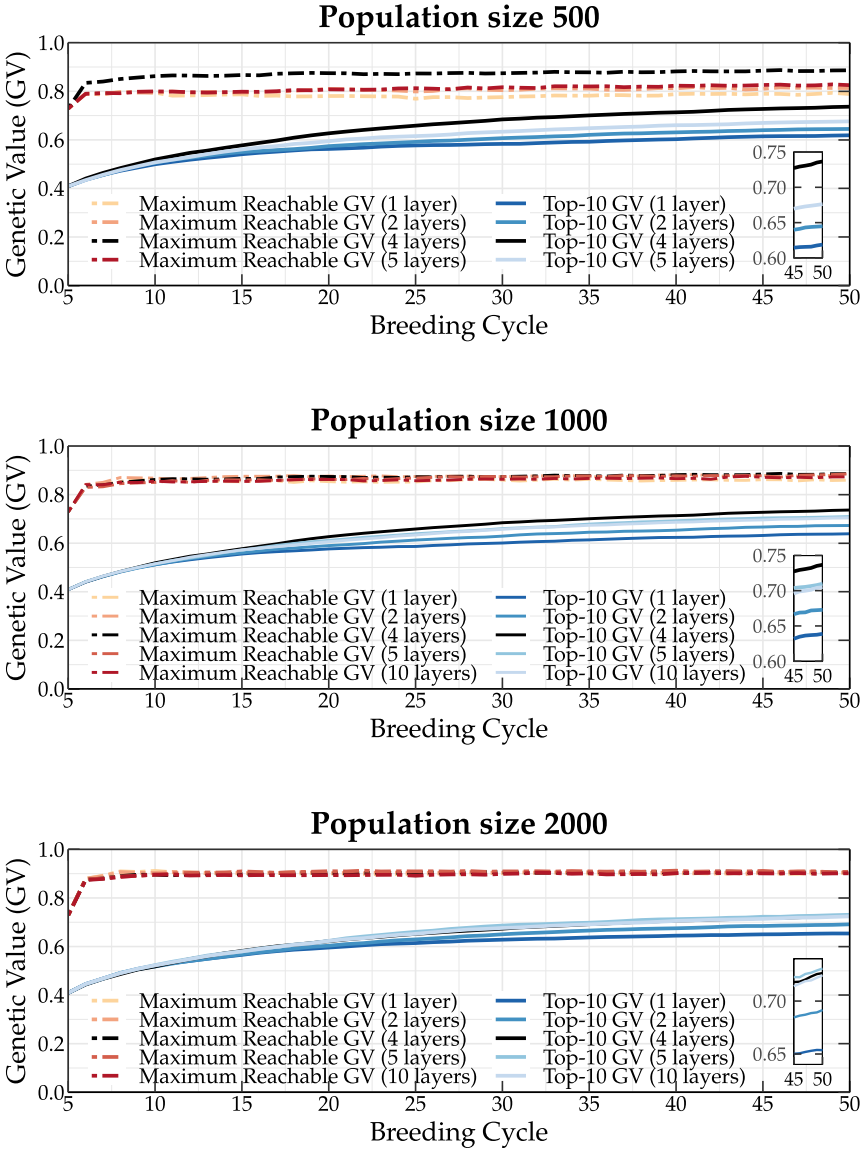

Figure 7.12: Simulation results of the deep scoping method using 1 to 10 different layers for a population size of 500, 1000, and 2000 individuals. Increasing the number of layers often results in higher long-term genetic gains. However, if the number of individuals per layer is too low, the long-term genetic gain is reduced.

Even compared with the scoping method, using the deep scoping method after five breeding cycles of truncation selection will result in a 3% points higher genetic value in the long term (see see Table 10.1). While the scoping method is able to reach high long-term genetic gains, the method is not perfect and allows for the fixation of a few unfavorable QTL alleles. Therefore, methods like deep scoping that can reintroduce genetic variation into the breeding population are important to maximize the genetic gain and avoid a premature convergence of the genetic value.

7.4.7 Designing the elite population

To maximize the genetic gain of the elite population, truncation selection is used. However, different crossing block designs were considered. Crossing individuals with the highest GEBVs resulted in a lower genetic gain compared with random crossing. Individuals with a high GEBV are often closely related, limiting the genetic variation that can be passed to the offspring and resulting in a rapid fixation of the QTL alleles and a premature convergence of the genetic value (see Figure 7.15). Minimizing the genetic relationship between both parents resulted in higher genetic gains compared with random crossing. This design avoids the coupling of closely related parents, minimizing the loss in genetic variation while preserving the genetic progress and thus maximizing the short- and long-term genetic gain.

7.4.8 The size of the gene bank

Both the deep scoping method and the HUC method with bridging use a gene bank to reintroduce genetic variation into the breeding population. Increasing the size of the gene bank had almost no effect on the mean genetic value of the top-10 individuals. Only when the size of the gene bank was reduced to 200 individuals, a reduction of the genetic value in the long term was observed. At that point, the gene bank was too small to contain both alleles of each QTL, reducing the introduction of genetic variation into the breeding population. In this simulation setting, a gene bank is essential to maximize the genetic gain in the long term, but it is not necessary to collect thousands of different individuals. As long as all the QTL alleles are present in the gene bank, the deep scoping method will be able to maximize the long-term genetic gain.

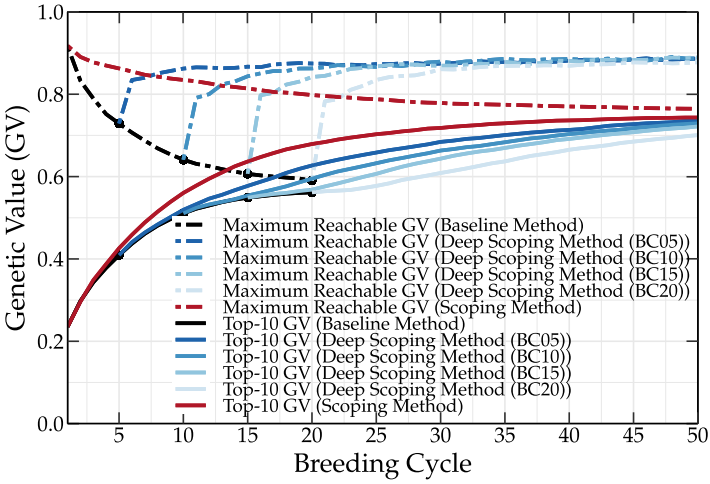

Figure 7.13: Simulation results of the deep scoping method and the scoping method over 50 breeding cycles. When the scoping method is used from the first breeding cycle, an overall higher genetic gain is observed in the short and long term.

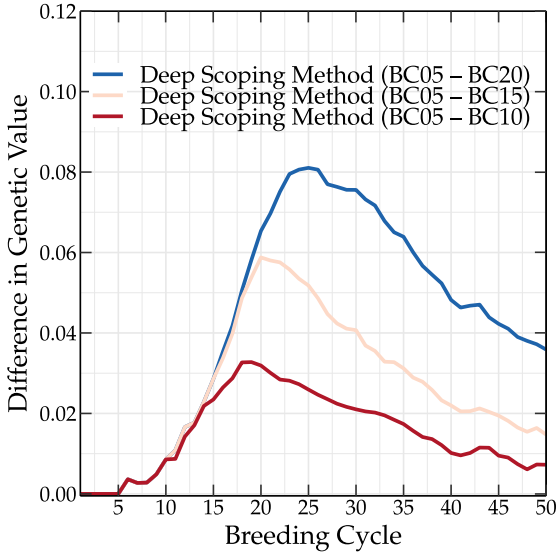

Figure 7.14: Difference in genetic value between the deep scoping method using two different breeding populations. The sooner truncation selection is replaced by the deep scoping method, the sooner the genetic gain of the top-10 individuals will increase until the breeding value converges to the same value in the long term.

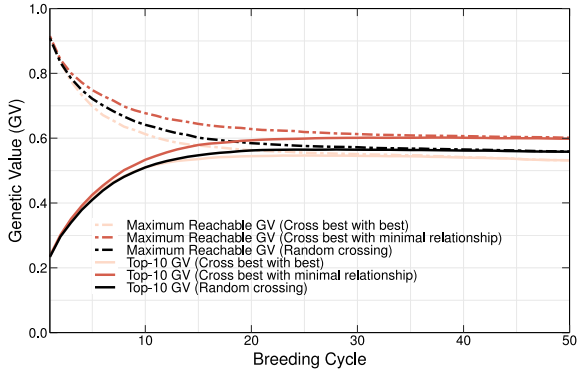

Figure 7.15: Simulation results of truncation selection using different crossing block designs. Three crossing block designs were considered: i) crossing the individuals with the highest GEBVs, ii) crossing the individuals with the highest GEBV with an individual that minimizes the genetic relationship between both parents and iii) random crossing.

7.4.9 Updating the training population

The deep scoping method uses different layers, each containing individuals at a different stage of development. Similar to Chapter 5, different methods to update the training population are considered. Normally, as long as the training population is updated, the genetic value will converge to the same genetic value (Neyhart et al., 2017). This is not the case when the deep scoping method is used. The *top* update method results in the lowest long-term genetic gains whereas the tails method results in the highest long-term genetic gains (see Figure 7.16). To understand this phenomenon, the Pearson correlation between the GEBVs and the genetic values is calculated for the three subpopulations (Layer 0, Layers 1–4 and the elite population). The results are shown in Figure 7.17. The Pearson correlation of the elite population is relatively lower compared to the Pearson correlation of Layer 0 or Layers 1–4. Because the elite population only selects individuals with a high GEBV, the genetic variation of these individuals will be low, resulting in a low prediction accuracy (or Pearson correlation).

As expected, the top update method results in the highest prediction accuracy for the elite population but it fails to predict the GEBVs of the pre-breeding population (Layer 0) and Layers 1–4. The *bottom* update method results in a low prediction accuracy of the elite population but is able to accurately predict the GEBVs of Layer 0 and Layers 1–4. Because the top update method results in a lower long-term genetic gain, it is clearly more important to accurately predict pre-breeding individuals than to predict the genetic value of elite individuals. Moreover, the top update method only select individuals with a high GEBV that are often closely related, resulting in a training population that only covers a small fraction of genetic variation. The bottom update method, on the other hand, will contain individu-

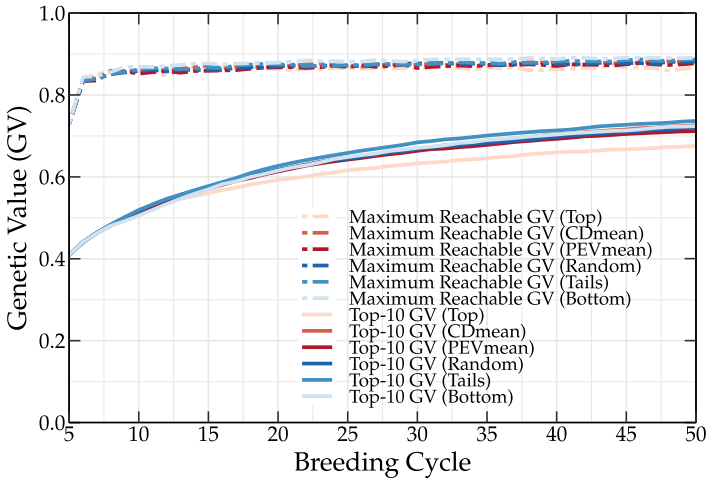

Figure 7.16: Simulation results of truncation selection using the top, bottom, random, tails, CDmean and PEVmean methods to update the training panel. Each update method results in a similar long-term genetic gain with the exception of the top update method that results in a lower long-term genetic gain.

als with a broad genetic variation in the training population resulting in a more accurate estimate of the marker effects.

The tails update method is a combination of the top and bottom update methods, selecting individuals with the highest and lowest GEBVs to update the training population. This update method results in the highest Pearson correlation for the three subpopulations. The *random*, *PEVmean* and *CDmean* update methods also result in a similar, but slightly lower Pearson correlation for the three subpopulations. Each of these methods is able to select individuals in the TP that maximizes the prediction accuracy.

7.4.10 Influence of the prediction model

In Subsection 5.4.6 we demonstrated that removing low-frequency marker alleles decreases the genetic value in the long term. The removal of low-frequency alleles could in certain cases increase the prediction accuracy of a trait (Edriss et al., 2013), but according to VanRaden et al. (2009), removing these markers is unnecessary. It could lead to the loss of important information resulting in a lower short-term genetic gain. Therefore, all the available markers are used to fit the different prediction models. Only a small non-significant ($p < 0.05$) difference is observed between the Bayesian models and rr-gBLUP (see Figure 7.18).

Both rr-gBLUP and BRR uses a similar strategy to estimate the marker effects and variance components of a linear mixed effects model. Therefore, the estimate

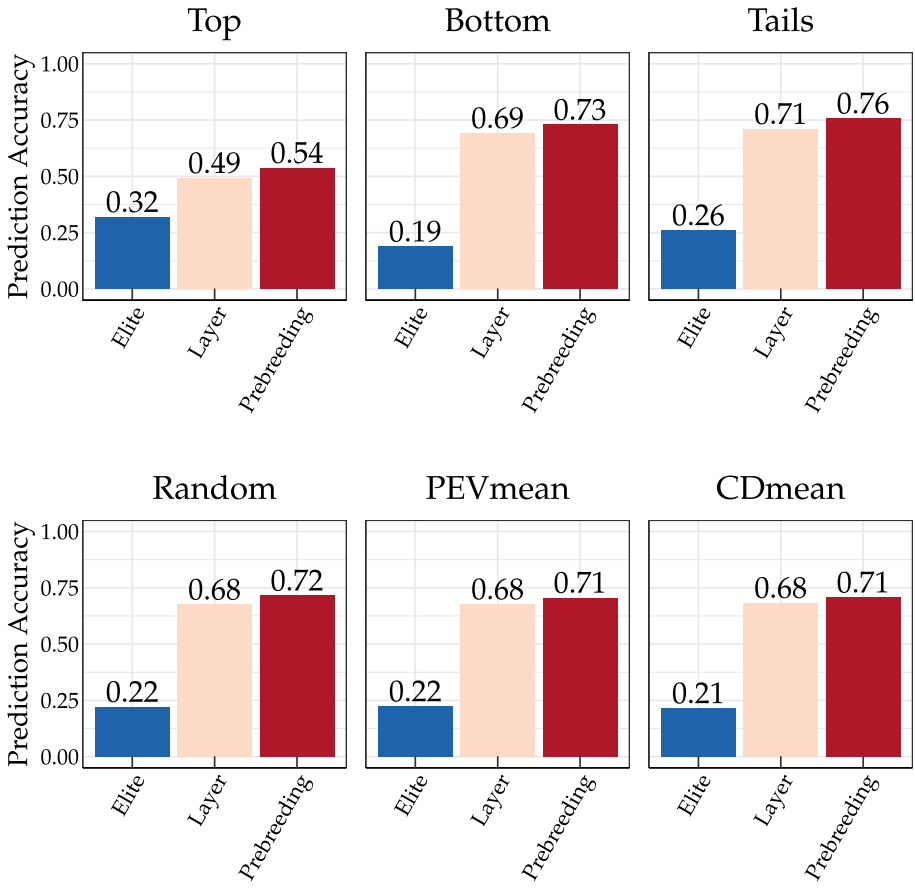

Figure 7.17: Overview of the prediction accuracy of the elite population, the different layers (Layer 1–4) and the pre-breeding population (Layer 0) using truncation selection. The training panel is updated according to the top, bottom, tails, random, CDmean or PEVmean method.

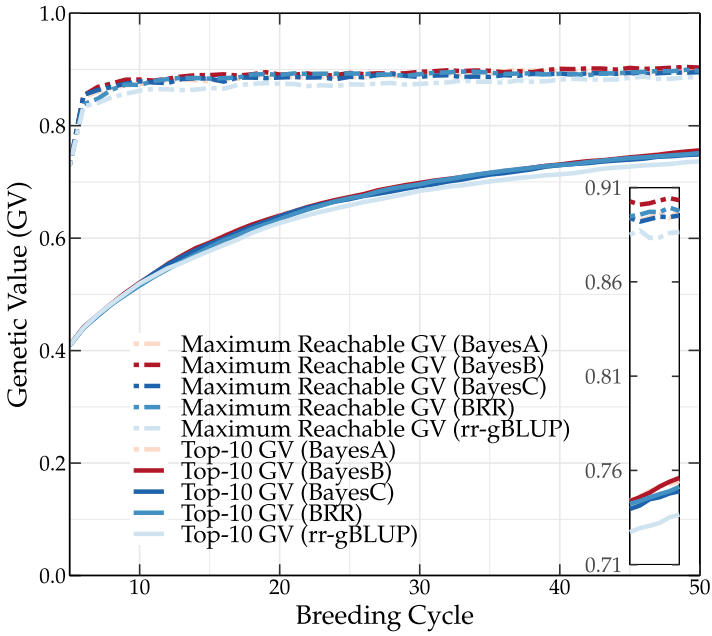

Figure 7.18: Effect of rr-gBLUP and Bayesian models on the genetic gain of a breeding population using the deep scoping method. A slightly lower, but non significant long-term genetic gain is observed for the rr-gBLUP model.

of both models should converge to the same values. The estimated values of both variance components and the Pearson correlation for rr-gBLUP and BRR are shown in Figure 7.19. Two scenarios were considered. In the first scenario, all the 1490 available biallelic markers were used, whereas, in the second scenario, the markers with a low allele frequency were removed. Compared to BRR, rr-gBLUP results in a lower value for σ_e^2 and a slightly higher prediction accuracy (Pearson correlation). The BRR model estimates the variance components using a Gibbs sampler. Compared to BRR, the rr-gblup results in a slightly lower genetic value. This result was also reported by Sayfzadeh et al. (2013).

In contrast to the deep scoping method, when truncation selection is used, a similar long-term genetic gain was observed for both rr-gBLUP and BRR. The variance components and prediction accuracy for both scenarios and both models are shown in Figure 7.20. Over the different breeding cycles, the predicted variance component of the residual errors quickly stabilizes to a fixed value. This indicates that over each breeding cycle, both variance components are accurately estimated. As discussed before, by removing the low-frequency markers, a lower prediction accuracy and a slightly different estimate of the variance components are observed.

When truncation selection is used, the estimate of the variance component $\mathrm{var}(u)$ will become unstable in the long term, resulting in a high standard deviation between the different experiments. This is caused by the fixation of most markers, resulting in a problem of rank deficiency. Because the deep scoping method introduces genetic variation in the breeding population, instability of the predicted variance components is avoided. Although rr-gBLUP still succeeds to estimate a value for both variance components, BRR often estimates an infinite variance component when the low-frequency markers are removed from the training population. Therefore, all the available markers should be used to fit the mixed effects model.

7.4.11 Cost analysis

Compared to the scoping method, the deep scoping method uses a gene bank to introduce genetic variation into the breeding population. Therefore, if such a gene bank is not yet available, a certain investment will be required. Although the size of the gene bank remains unchanged during the whole simulation study, the germinative power of the seeds that are stored in a gene bank decreases and should be renewed after a certain timestamp. The deep scoping method was developed to combine the pre-breeding and elite population into a single breeding population using the same resources as a classical breeding program. Therefore, using the deep scoping method will not increase the costs or required resources of the breeding program. Moreover, based on the results illustrated in Figure 7.12,

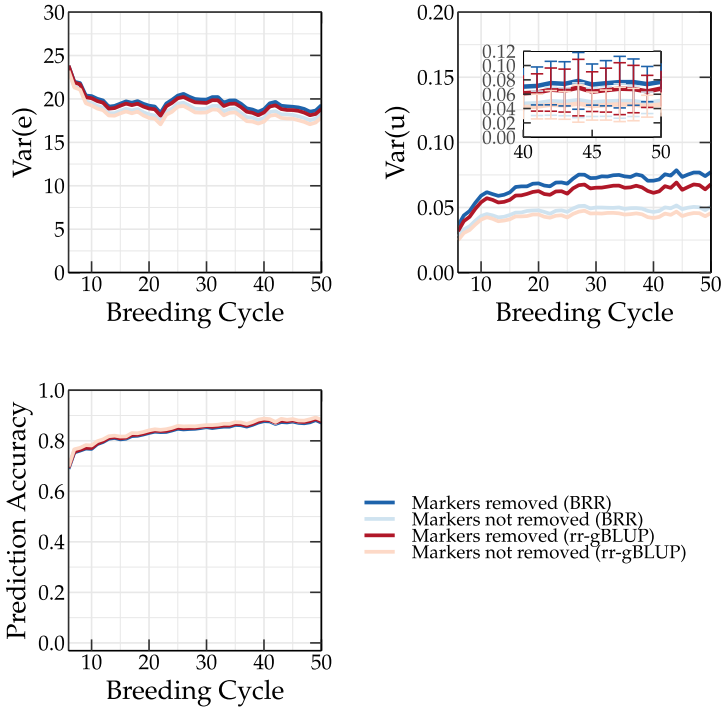

Figure 7.19: Simulation results of the deep scoping method using rr-gBLUP and BRR. Top left, the predicted variance component of the residuals errors. Right top, the predicted variance component of the additive marker effects, and left bottom, the prediction accuracy. The model is fitted using the available markers or after first removing the low-frequency marker alleles.

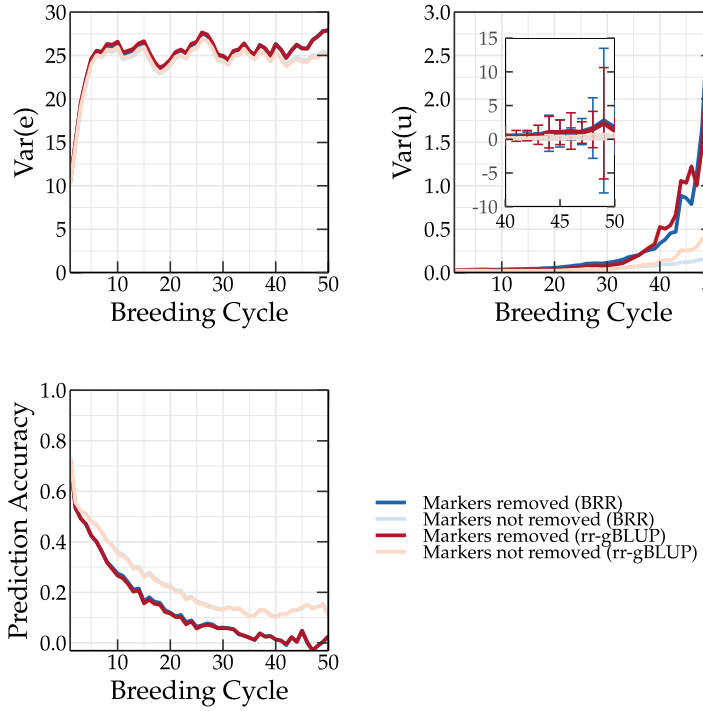

Figure 7.20: Simulation results of truncation selection using rr-gBLUP and BRR. Top left, the predicted variance component of the residuals errors. Right top, the predicted variance component of the additive marker effects, and left bottom, the prediction accuracy. The model is fitted using the available markers or after first removing the low-frequency marker alleles.

the population size could be reduced, without reducing the long- or short-term genetic gain. This could thus even decrease the cost and required resources of a breeding program.

Once the deep scoping method is used, the genetic gain is rapidly increased compared to a breeding population in which truncation selection is still used. In other words, when the deep scoping method is applied, the breeding population will deliver individuals with a superior phenotype compared to competitive breeding programs that still use truncation selection. Depending on the strategy of the breeding company, these lines could increase the market share of that breeding company. The costs to maintain a gene bank will probably only be a fraction of the gained profits that could be achieved using the deep scoping method.

According to Figure 7.14, when the genetic variation is low due to years of intensive selection, the earlier the deep scoping method is deployed, the faster high genetic gains will be obtained. Therefore, using the truncation selection instead of the deep scoping method will only increase the marginal cost.

7.5 Conclusion

Truncation selection often reduces the genetic variation of the breeding population in its striving for a maximal genetic value of the breeding population, to the extent that the adoption of variation-preserving methods such as the population merit and scoping methods is rather pointless. To increase the genetic variation of the breeding population, a gene bank containing a broad genetic variation is needed. However, when the gap between the genetic value of the breeding population and the gene bank increases, the transition of pre-breeding individuals into the elite population becomes difficult. The deep scoping method uses different layers of pre-breeding individuals to maximize the flow of pre-breeding individuals into the elite population, thereby introducing (favorable) QTL alleles into the elite population. Replacing the frequently used GEBVs by HEBVs allows for a more accurate selection of individuals containing favorable QTL alleles, maximizing the long-term genetic gain. In summary, the deep scoping method combines an elite population with different layers using the HEBVs, H-score, and S-score, resulting in higher genetic gains compared to truncation selection and the HUC method with bridging in the short as well as in the long term without the need of a separate pre-breeding population.

8

Oracle methods

8.1 Introduction

Different methods to maximize the long-term genetic gain have been proposed in the previous chapters. In theory, according to Eq. (2.6), the genetic gain can be maximized by reducing the time per breeding cycle and by increasing the selection intensity, the prediction accuracy, and the genetic variation. The genetic gain represents the change in genetic value that can be achieved in a breeding population per unit of time. By reducing the time per breeding cycle, the same change in genetic value can be obtained in a shorter time frame, increasing the genetic gain. This can be achieved for example, by means of *rapid generation advancement* (RGA) methods in which homozygous lines are rapidly generated by using specific semi-controlled greenhouse conditions. These conditions result in an early flower induction, and accelerate the breeding process (Srinivasan et al., 2020). The reduction in breeding time does not rely on a specific selection strategy. Therefore, this approach is beyond the scope of this dissertation.

In a second approach, the genetic gain is maximized by increasing the selection intensity, for example, by only selecting the most superior lines as a parent (like

truncation selection). This often results in the selection of closely related individuals, reducing the genetic variation of the offspring and causing a lower genetic gain in subsequent breeding cycles. To avoid the loss of genetic variation, the *scoping* and *deep scoping* methods were proposed. Compared to truncation selection, by slightly decreasing the selection intensity, both methods were able to better preserve the genetic variation (see Figure 5.10), increase the prediction accuracy (see Figure 5.11) and maximize the long-term genetic gain (see Figure 5.5). Although the scoping and deep scoping methods were able to outperform different existing methods (see Chapters 5 and 7), it is difficult to measure to which extent these breeding methods really maximize the long-term genetic gain.

An oracle method is a theoretical concept in which a value of interest (e.g. the genetic gain) is maximized using the ground truth. It can be used to select an optimal parental population or an optimal training population. This way, the characteristics of such an optimized population can be studied and could be used to develop new (non-oracle) selection methods. The *true selection* method selects individuals based on their QTL effects to rapidly fixate the favorable QTL alleles in the breeding population. By comparing the true selection method with the scoping and deep scoping methods, we can assess to which extent these methods really maximize the genetic gain and whether further optimization of these methods is required. The *forward selection* method and *stepwise selection* method use an iterative algorithm to select an optimal training population that maximizes the prediction accuracy (Pearson correlation) of the breeding population. Different approaches to update a training population have been proposed but according to Neyhart et al. (2017), all these methods result in the same long-term genetic gain. The idea is that by studying the training population of both the forward and stepwise selection methods, a new approach to update the training population can be found that maximizes both the prediction accuracy and the long-term genetic gain of a breeding population.

8.2 Material and methods

A complete overview of the breeding population is described in Chapter 4.

8.2.1 The true selection method

The true selection method is a theoretical concept that reveals the full potential of the parental selection. The scoping method is constrained to use the information of molecular markers that are linked to the QTL that underlie the trait of interest. The true selection method, in contrast, is allowed to use the actual QTL effects to guide

the parental selection process. It attributes a score to each individual, expressing the number of favorable QTL alleles. Intuitively, the method selects individuals with the highest number of favorable QTL alleles, giving priority to QTL positions that have not yet been selected in the parental population, thus preventing the loss of these rare favorable QTL alleles. Formally, each individual has a score between 0 and L (number of QTLs), representing the number of favorable QTL alleles that are present in its genome. The individual with the highest score is selected as a first parent. The remaining individuals are scored again, this time only taking into account QTL positions whose favorable alleles have not yet been selected in the parental population. After recalculating the score, the individual with the highest score is selected. This is repeated until all the required parents are selected. As soon as all the favorable QTL alleles are present in at least one of the selected parents, the score is again calculated over all the QTLs. The parents are then randomly crossed with each other using a recurrent breeding scheme (see Chapter 4). The true selection method should maximize the genetic progress while avoiding the loss of favorable QTL alleles.

8.2.2 Selection of an optimal training population

We propose two selection methods to construct an optimal training population that maximizes the prediction accuracy of the genomic prediction model. Originally, we proposed three selection methods: namely the forward selection method, the *backward selection* method and the stepwise selection method. The backward selection method requires fitting a linear mixed effects model using a large training dataset in an iterative scheme, resulting in a high computation time. Therefore, the backward selection method was no longer considered as a selection method but was instead integrated into the stepwise selection method. Similar to the true selection method, the forward and stepwise selection methods are theoretical concepts to study the characteristics of an optimal training population. In the next step, these characteristics can be used to develop a new (non-oracle) selection method.

The oracle selection methods use truncation selection (baseline method) to simulate a breeding population over 15 breeding cycles. The training population is constructed by randomly selecting 100 individuals from the base population. In contrast to previous methods, only part of the base population is used as a training population to limit the simulation time of both selection methods. Over each breeding cycle, with the exception of the first breeding cycle, the 50 oldest individuals are removed from the training population. Next, the training population is updated by selecting 50 new individuals from the breeding population according to one of the candidate selection methods. The size of the training population remains constant at 150 individuals and limits the simulation time while avoiding

the problem of rank deficiency of the matrices involved in fitting the mixed model equations by means of the `rrBLUP` package in `R` (Endelman, 2011). In each selection strategy, the Pearson correlation between the GEBVs and the true genetic values is used to select individuals in the training population. Once an optimal training population is obtained, 100 individuals with the highest GEBVs are selected as a parent and crossed randomly.

Both the forward and stepwise selection methods are compared to the *top*, *bottom*, *random*, *tails*, *CDmean*, and *PEVmean* update methods (Neyhart et al., 2017; Rincent et al., 2012). A full description of these methods was already given in Subsection 4.6. To allow for a fair comparison between the different update methods, only 100 individuals are selected from the base population and a training population size of 150 individuals is used, removing and adding 50 individuals per breeding cycle with the exception of the first breeding cycle. By comparing these methods with the oracle methods, we can assess to which extent current methods are able to select an optimal training population. Because these methods do not use the true genetic values, we do not benchmark them against the oracle methods.

8.2.3 The forward selection method

The forward selection method updates the training population by adding 50 new individuals using a greedy algorithm. Starting with an existing training population, the contribution of each individual to the training population is evaluated. This is done by adding each individual separately to the training population. The contribution of each individual can then be evaluated by recalculating the prediction accuracy (Pearson correlation) after refitting the linear mixed effects model using the new training population. The individual that maximizes the prediction accuracy after its addition to the training population is accepted. The remaining individuals are reevaluated before selecting the next individual. In total, 50 individuals will be selected in the training population unless the addition of an individual cannot increase the prediction accuracy anymore.

8.2.4 The backward selection method

Although the backward selection method is not considered as a selection method, a short description of the method is given to facilitate the introduction of the stepwise selection method in the next subsection.

In contrast to the forward selection method, the whole breeding population is used as training population. The backward selection method then removes individuals

from the training population that have a negative impact on the prediction accuracy. Starting with an existing training population, the impact of removing each individual separately from the training population is assessed and the individual that maximizes the prediction accuracy after its removal is eliminated from the training population. This is repeated until the removal of any individual does not increase the prediction accuracy anymore. Similar to the forward selection method, after each iteration, each individual needs to be reevaluated before removing the next individual from the training population.

8.2.5 The stepwise selection method

The stepwise selection method is a combination of the forward and backward selection methods. Individuals are added and removed from the training population iteratively. Therefore, it is not necessary to remove 50 individuals from the training population at the start of each breeding cycle (as described in Subsection 8.2.2). First, an individual of the breeding population is added to the training population using the forward selection method. Next, each individual of the training population is evaluated using the backward selection method. Both the forward and backward selection methods will always reevaluate the individuals based on the last update of the training population, taking into account all the previous changes. In total, 50 individuals can be added and removed from the training population. An individual will only be added to or removed from the training population if that action maximizes the prediction accuracy. It is, however, possible that each individual in the training population has a positive contribution towards the prediction accuracy and that no individual is removed. Therefore, the size of the training population could vary, depending on the addition and removal of individuals.

8.2.6 Prediction model

For the true selection method, at the first breeding cycle, the complete base population is used as training population. In the subsequent breeding cycles, 150 individuals are phenotyped and added to the training population according to the tails method, selecting 75 individuals with the highest and 75 individuals with the lowest GEBVs (Neyhart et al., 2017). In the case of the forward and stepwise selection methods, only 100 randomly selected individuals of the base population are used as training population. In the subsequent breeding cycles, 50 individuals are added to and removed from the training population as described in Subsection 8.2.2.

8.3 Results

8.3.1 The true selection method

The true selection method is a hypothetical concept that uses the knowledge of QTL positions and QTL effects to demonstrate the effect of an almost perfect parental selection on the genetic value. In the initial population, 10% of the QTLs are already fixed for one of the two possible alleles. Some of these alleles have a negative contribution to the genetic value. This explains why the maximum reachable genetic value is slightly lower than 1 for the initial population (see Figure 8.1). Over the subsequent breeding cycles, the maximum reachable genetic value remains constant, indicating that no favorable QTL alleles are eliminated during breeding. Over the different breeding cycles, the frequency of favorable QTL alleles in the breeding population increases, leading to a strong increase in the mean genetic value. Finally, unfavorable QTL alleles are lost from the breeding population, leading to the fixation of favorable QTL alleles. The mean genetic value of the top-10 individuals, the maximum reachable genetic value, and the fixed genetic value are listed in Table 10.9.

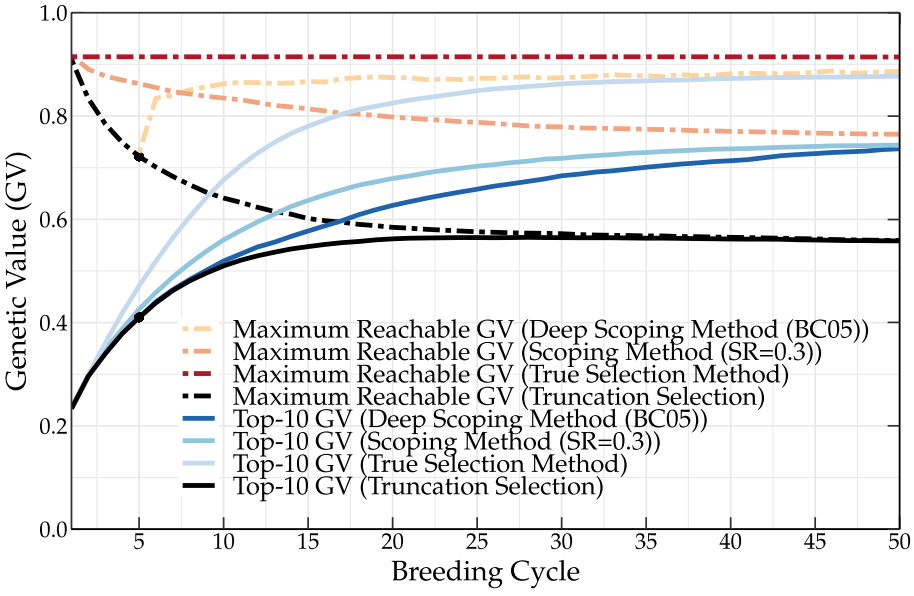

Figure 8.1: Mean genetic value of the top-10 individuals and maximum reachable genetic value of a breeding population using the true selection method, scoping method (SR = 0.3), deep scoping method (BC05), and truncation selection over 50 breeding cycles. The true selection method leads to a high increase of the mean genetic value over the first breeding cycles while the maximum reachable genetic value remains constant, indicating that no favorable QTL alleles are lost. The difference in genetic value between the true selection method and the other methods indicates that further improvements of the parental selection methods could increase the genetic value up to 14 percentage points.

The true selection method reaches higher genetic values in the short as well as in the long term. The scoping method (SR = 0.3) and the deep scoping method (BC05) are able to increase the long-term genetic gain but according to the true selection method, the genetic value could still be increased up to 14 percentage points in the long term. This indicates that even with the scoping and deep scoping methods, further improvements of the parental selection could significantly increase the short- as well as the long-term genetic gain.

8.3.2 Optimizing the training population

The forward and stepwise selection methods are compared to different update methods proposed by Neyhart et al. (2017) and Rincent et al. (2012). The mean genetic values of the top-10 individuals and the maximum reachable genetic values are shown in Figure 8.2 for the different update methods. At the start of the simulation, each method results in the same mean genetic value of the top-10 individuals. After two breeding cycles, the forward and stepwise selection methods result in a higher long-term genetic gain compared to the other update methods. At breeding cycle 15, a difference of 19 percentage points is observed between the stepwise selection method and the top method. The different methods proposed by Neyhart et al. (2017) and Rincent et al. (2012) converge to approximately the same genetic value. It is clear that the forward and stepwise selection methods result in a better prediction of the GEBVs, leading to higher long-term genetic gains. The maximum reachable genetic value and the mean genetic value of the top-10 individuals using different update methods are listed in Tables 10.10 and 10.11, respectively.

While both the forward and stepwise selection methods aim to optimize the training population, they generally do not converge to the same result. The forward selection starts with 50 randomly chosen individuals. Over the next iterations, the forward selection method will maximize the prediction accuracy of the population. However, the forward selection method is not able to remove individuals from the training population. This is in contrast to the stepwise selection method, which can further optimize the training population by removing superfluous individuals. Nevertheless, the forward and stepwise selection methods resulted in a similar breeding value in the long term indicating that removing superfluous individuals only has a minor effect on the genetic value of the breeding population.

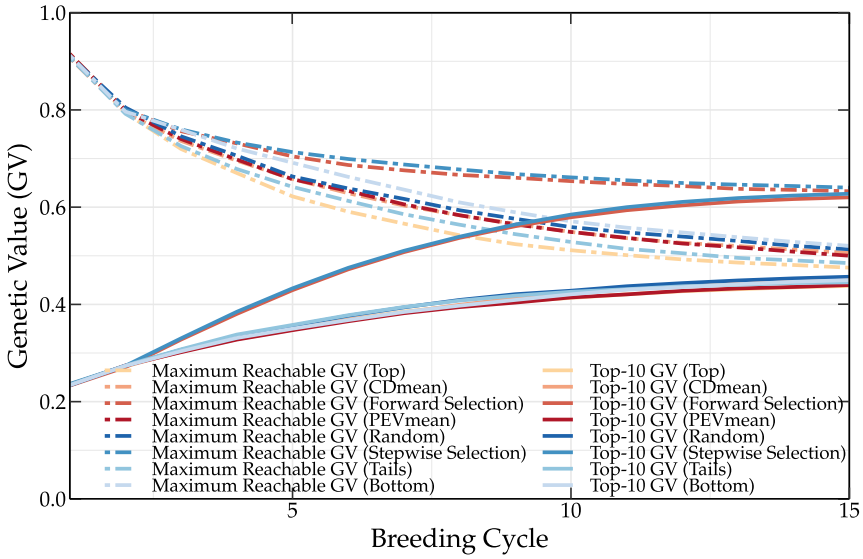

Figure 8.2: Simulation results of the forward and stepwise selection methods compared to truncation selection. Both the forward and stepwise selection methods have a higher genetic value in the long term compared to truncation selection using the top, bottom, tails, random, PEVmean or CDmean update methods.

8.4 Discussion

8.4.1 The greedy selection of QTL alleles

The true selection method was developed to study the effects of using a modified truncation selection scheme in which the frequency of all favorable QTL alleles is maximized. The true selection method assumes knowledge of the actual QTL effects and is therefore only of theoretical interest; in vivo, only genetic markers are available to guide parental selection. Although the true selection method is able to maximize the genetic gain by greedily selecting the favorable QTL alleles in the parental population, when the parental selection process relies on genetic markers that are putatively linked to the casual QTL effects, greedily selecting individuals (as observed for truncation selection) often result in a premature convergence of the genetic value. In other words, by only preserving the marker alleles that have a positive estimated marker effect, the loss of favorable QTL alleles cannot be prevented. Preserving both marker alleles in the breeding population prevents the elimination of poorly estimated QTL alleles resulting in a higher long-term genetic gain compared to a greedy strategy like truncation selection.

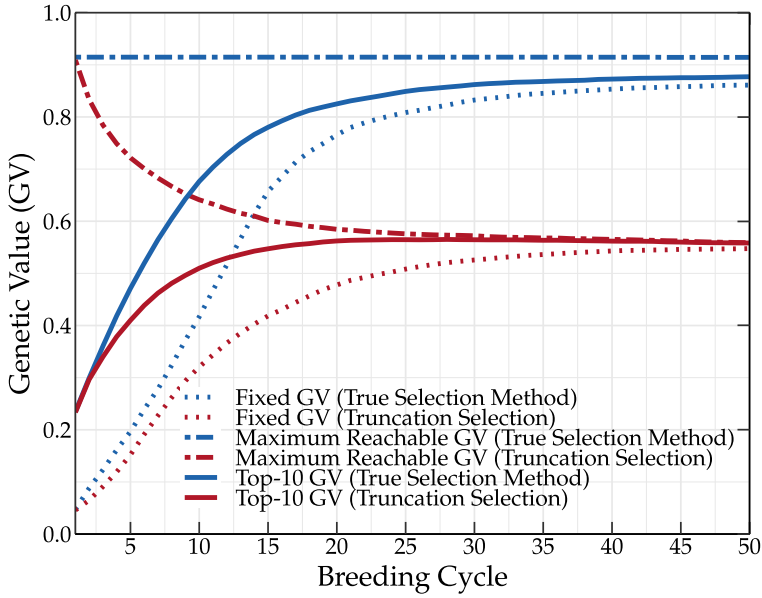

Figure 8.3: Simulation results using the true selection method over 50 breeding cycles. The true selection method leads to a high increase of the mean genetic value over the first breeding cycles. The maximum reachable genetic value remains constant, indicating that no favorable QTL alleles are lost. Due to selection, the frequency of the favorable QTL alleles increases, finally leading to the loss of unfavorable QTL alleles. This eventually results in high genetic values.

8.4.2 Reaching the theoretical maximum genetic value

The favorable QTL alleles are not always abundantly present in the initial population and many breeding cycles may be needed before fixation occurs. This explains the relatively slower increase of the fixed genetic value compared to the mean genetic value of the top-10 individuals (see Figure 8.3). In a standard setting, the fixed genetic value represents the overall effect of all the QTL alleles that are fixed in the breeding population. As the true selection method avoids the fixation of all QTL alleles, favorable and unfavorable, the fixed genetic value can, in this case, be used to monitor the fixation of favorable QTL alleles. After almost 30 breeding cycles, the genetic value and the fixed genetic value converge to a slightly lower value than the maximum reachable genetic value. The true selection method, which was designed to prevent the loss of favorable QTL alleles, should make it possible to reach the maximum genetic value of 1. However, when two or more QTL alleles are in strong linkage disequilibrium (LD) w.r.t. one another, linking a favorable QTL allele to an unfavorable QTL allele, fixation of both QTL alleles becomes difficult. This was the case for five percent of the QTLs, preventing the genetic value from reaching its absolute maximum and explaining why the genetic value and the fixed genetic value did not converge to the same value. The oracle method demonstrates that, even in an ideal situation, at least 30 breeding

cycles are needed to obtain the maximum reachable genetic value in the breeding population for the base population and simulation settings used in this study.

8.4.3 Genetic values, phenotypic values, and genomic estimated breeding values

The main goal of breeding is to maximize the genetic value of various traits of interest and this both in the short and long term. Unfortunately, the genetic value cannot be measured, and thus selection is often based on the phenotype. Because the phenotype is influenced by the environment, compared to genetic values, using the phenotype to select the parental population will result in a lower genetic gain. This is shown in Figure 8.4. Measuring the phenotype is a time-consuming and expensive process, therefore, GEBVs are often used instead, predicting the genetic values using a linear mixed effects model. The selection of superior individuals using GEBVs hinges on the prediction accuracy of the underlying genomic prediction model. The construction of a genomic prediction model requires a training population for which phenotypic and genotypic data of the breeding population is required. Due to prediction errors, selecting parents based on the GEBVs results in a lower genetic gain compared to a parental selection based on phenotypic or genetic values (see Figure 8.4). The difference in the genetic value obtained by selecting the parents based on the GEBVs and phenotypic values could be reduced by using a more accurate prediction model and a better training population design. However, in genomic selection, the linear mixed effects model in combination with rr-gBLUP often results in a high performance (Moser et al., 2009), and according to Neyhart et al. (2017), as long as the training population is updated, the genetic value converges to the same long-term value.

8.4.4 The size of the training population

The stepwise selection method can select and remove 50 individuals from the training population. An individual can only be added to or removed from the training population if a higher prediction accuracy can be obtained. At the start of the simulation, the training population is constructed from a random selection of 100 individuals from the base population. At the first breeding cycle, the stepwise selection method adds 50 individuals to and removes 50 individuals from the breeding population, replacing on by on the randomly chosen individuals in the training population (see Figure 8.5). In the subsequent breeding cycles, the number of individuals that are removed from the training population is reduced, allowing for an increase in the size of the training population. Over time, more

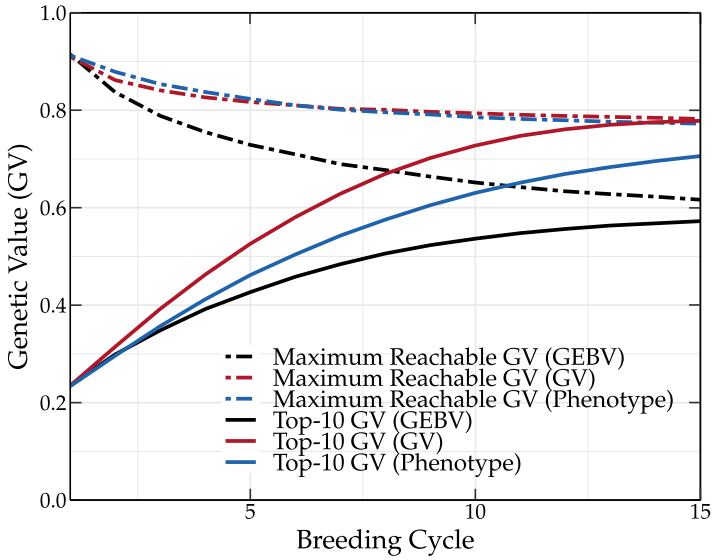

Figure 8.4: Mean genetic value of a breeding population using truncation selection. The parents are selected based on the GEBV, GV, or phenotypic values. Selecting parents based on the GV will result in high genetic values, followed by phenotypic values and GEBVs (using the tails method).

individuals will be removed from the training population, while the number of individuals that are added to the training population will decrease. This means that by removing superfluous individuals, a lower number of individuals should be added to the training population to maximize the prediction accuracy.

The size of the training population starts to converge at breeding cycle 15. At that point, less than 50 individuals are added to the training population. This observation seems to indicate that using huge datasets to fit a prediction model may not be the best strategy.

We could expect that if the number of individuals that can be added to the training population was not limited to 50, a higher number of individuals would have been selected over the first breeding cycles, increasing the prediction accuracy and the genetic gain. Unfortunately, this would also increase the simulation time.

8.4.5 The genetic relationship between the training population and the breeding population

A training population should contain individuals that represent the genetic diversity of the current breeding population, allowing for an accurate prediction of each

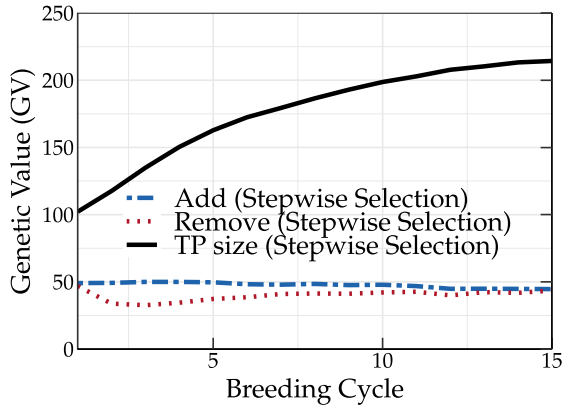

Figure 8.5: Overview of the number of individuals that are added to and removed from the training population using the stepwise selection method. Over the first breeding cycles, a lower number of individuals are removed from the breeding population, allowing for an increase in the size of the training population.

individual. The genetic relationship can be calculated based on the genotype using Eq. (4.3). A high genetic relationship between two individuals indicates that both individuals are closely related. It is expected that if the training population is a good representation of the genetic diversity of the current breeding population, the mean genetic relationship between the breeding population and the training population will be maximized. This mean genetic relationship for the different selection methods is shown in Figure 8.6.

The top and tails update methods result in a training population with a high genetic relationship. Both update methods select individuals with a high GEBV that are also selected as a parent. Therefore, the training population will be a good representation of the individuals in the subsequent breeding cycles. The tails update method also selects individuals with a low GEBV, hence the lower genetic relationship compared to the top update method over the first breeding cycles. At breeding cycle 10, due to the loss of genetic variation, the prediction accuracy is almost reduced to zero. At that point, the GEBVs do not represent the genetic value of each individual correctly, resulting in an almost random selection. Therefore, a training population that was selected by means of the tails update method will result in a higher mean genetic relationship with the current breeding population compared to the top update method. The bottom update method only selects individuals with the lowest GEBVs, and will thus have a low genetic relationship towards the breeding population. However, when the prediction accuracy and genetic variation decrease, the bottom update method results in a higher genetic relationship.

The forward and stepwise selection methods result in a similar low mean genetic relationship between breeding cycles 5 and 10. Both selection methods were developed to maximize the prediction accuracy, therefore, based on the results

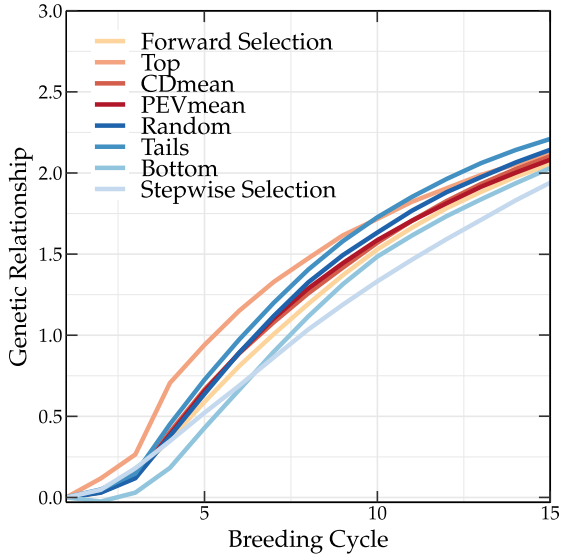

Figure 8.6: Mean genetic relationship between individuals of the training population and the individuals of the breeding population. The top and tails update methods result in a training population with a high mean genetic relationship, whereas the forward selection, stepwise selection, and bottom update methods result in a lower mean genetic relationship.

presented in Figure 8.6, maximizing the genetic relationship between the training population and the breeding population does not seem to be the best strategy. According to the oracle methods, a selection algorithm to update the training population should aim to minimize the genetic relationship between the breeding population and the training population.

8.4.6 Studying the optimal training population

To understand how the oracle methods optimize a training population, the mean marker variance, mean genetic value of the training population, and the mean absolute residual error of the individuals in the training population have been analyzed. The results are shown in Figure 8.7.

When a training population is updated with individuals with a high or low GEBV, the mean genetic value of the training population will be respectively higher or lower compared to the breeding population. This is observed for the top and bottom update methods. In the case of the forward and stepwise selection methods, the same mean genetic value is observed between the training population and the breeding population indicating that the selection is not based on the genetic value of the individuals.

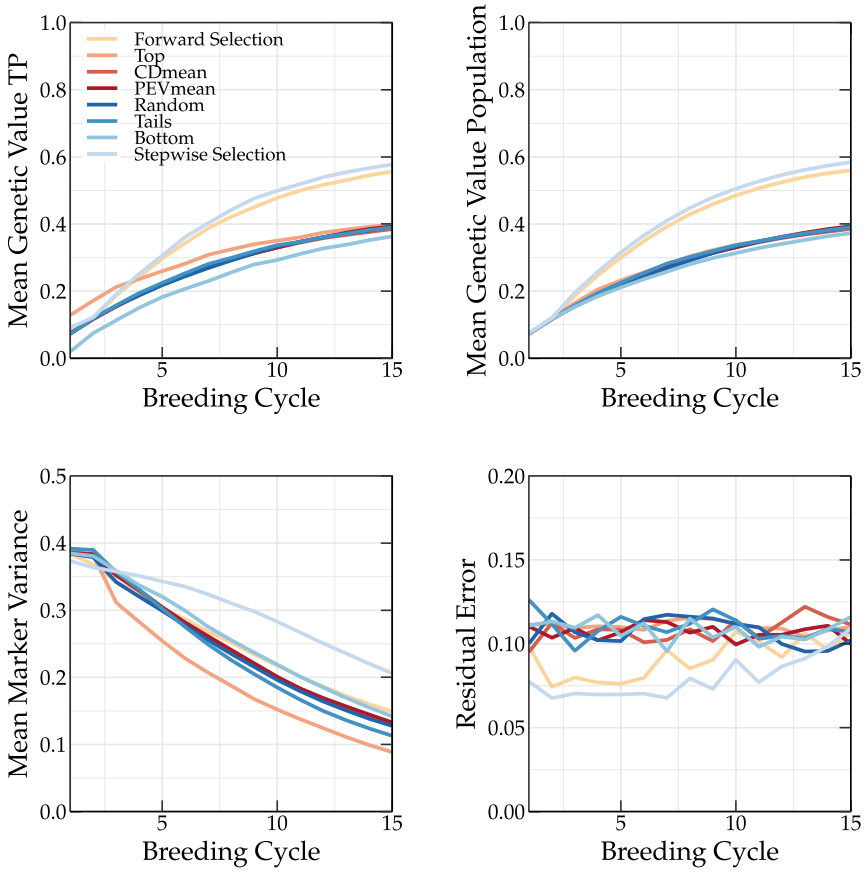

Figure 8.7: Top left, the mean genetic value of the training population, top right, the mean genetic value of the breeding population, bottom left, the mean marker variance, and bottom right the absolute residual error. Both the forward and stepwise selection methods select individuals that minimize the residual error, preserves the genetic variation in the training population, while preserving the same mean genetic value as observed in the breeding population.

Similar to the F_{score}, the mean marker variance summarizes the variance of the marker alleles in the training population. Both oracle methods generally demonstrate higher mean marker variances compared to the other methods although the bottom method seems to favor marker variance as well. The bottom method selects individuals with the lowest GEBVs. These individuals are often unrelated and contain a high level of heterozygosity, maximizing the genetic variation of the training population. Taking into account that by using truncation selection, after each breeding cycle, the genetic variation will decrease in the breeding population. By consequence, the genetic variation will also be reduced in the training population, as observed in Figure 8.6 for each update method. Because the stepwise selection method eliminates individuals in a more efficient manner, a slower decrease in the mean marker variance is observed compared to other update methods.

The residual error is the absolute difference between the genetic value and the phenotypic value of an individual. Over the first breeding cycles, both the forward and stepwise selection methods select individuals with a lower residual error. In vivo, the residual error is unknown, and thus selecting individuals that minimize the residual error cannot be achieved. The PEVmean update method (Rincent et al., 2012) selects individuals by minimizing the prediction error variance using the mixed model equations, but according to Neyhart et al. (2017) it was not able to outperform other update methods in the long term. Therefore, non-oracle update methods will probably not be able to reach the same long-term genetic gain as the oracle method as long as the residual error cannot be measured or predicted more accurately.

8.5 Conclusion

The results obtained by means of the oracle methods indicate that current methods to update the parental population or the training population are far from optimal. Although the scoping method increased the long-term genetic gain considerably compared to truncation selection, the optimal breeding population is not yet reached, opening the door for future work. The training population update methods were also not able to maximize the genetic gain compared to the oracle methods. Unfortunately, these oracle methods selected individuals with a lower residual error, which cannot be achieved in vivo. Therefore, more research is needed to assess to which extent an optimal breeding population can maximize the genetic gain when the residual error is taken out of the equation. According to both the forward and stepwise selection methods, maximizing the genetic variation of the training population could result in an optimal training population and should be further investigated.

PART III

CONCLUSIONS AND PERSPECTIVES

9

Discussion and future work

9.1 Genomic estimated breeding values and their limitations

GEBVs are generally calculated as the sum of the estimated effects of genetic markers across the genome of an individual, and are often used in GS as a selection criterion. Because the GEBVs reduce the genotypic information of an individual into a single value, important information that was available in the genotype may be lost. *Truncation selection* only selects individuals with the highest GEBVs as a parent. This often results in the selection of closely related individuals, causing a rapid fixation of the QTL alleles. Unfortunately, unfavorable QTL alleles are also getting fixed in the breeding population, reducing the long-term genetic gain. Selecting parents solely based on GEBVs cannot ensure that all the favorable QTL alleles will be passed to the next generation.

The loss of favorable QTL alleles could be reduced by preserving the genetic variation of the breeding population. The preservation of the genetic variation will,

however, also preserve unfavorable QTL alleles in the population. The parental selection should, therefore, also try to maximize the GEBVs, allowing for a slower, yet more accurate fixation of the favorable QTL alleles. To do so, different methods to preserve genetic variation have been discussed in this dissertation. The *population merit* method (Lindgren and Mullin, 1997) and the *maximum variance total* method (Cervantes et al., 2016) penalize the selection of closely related parents by using respectively the genetic relationship and the averaged inbreeding coefficient. The population merit method is able to preserve the genetic variation and to reduce the loss of favorable QTL alleles. This results in a higher long-term genetic gain compared to truncation selection, indicating that the use of genotypic information is crucial to optimize the parental selection. Nevertheless, both methods only use the genetic relationship matrix to penalize the GEBVs. However, this does not enforce the preservation of each marker (and QTL) allele.

The *scoping* method uses the genotypic information to ensure the selection of each marker allele to the extent possible. To do so, the F_{score} uses a Boolean variable that tracks if both alleles of a marker have already been selected in the parental population. The scoping method also uses a pre-selection that selects the fraction of the breeding population having the highest GEBVs. This fraction, controlled by the scoping rate, avoids the selection of individuals with a lower GEBV as a parent, which could decrease the genetic progress of the breeding population. This also means that if certain marker alleles are not present in the pre-selected population, these marker alleles will be eliminated from the breeding population. Even if all the marker alleles are present in the pre-selected population, the number of parents that can be selected may also be insufficient to select both alleles at each marker locus. Especially over the first breeding cycles, the scoping method was not able to preserve all the marker alleles by selecting only 100 parents, resulting in the loss of favorable QTL alleles. Hence, we observe a decrease in the maximum reachable genetic value. This could be avoided by using a different strategy to pre-select individuals (e.g. *adaptive scoping* method), by using a different crossing block design (see Section 9.3), or by selecting more parents (see Section 9.2). Nevertheless, the scoping method is still able to maximize the long-term genetic gain, outperforming truncation selection, the population merit method, and the maximum variance total method.

The HEBVs have been proposed as an alternative selection criterion to replace the GEBVs (Allier et al., 2020a). The HEBVs are based on the OHVs (Daetwyler et al., 2015), in which the marker effects of the best haploid segments are used to score an individual. If a heterozygous individual contains the favorable marker alleles on the same haploid segment, that individual will have a high HEBV and could thus be selected as a parent whereas his GEBV will have a lower value due to the presence of these heterozygous markers. The H-score evaluates individuals by calculating the HEBV between the haploid segments of that individual and an elite population. During the design of the *deep scoping* method, both the H-score and

the GEBVs were considered, but because the H-score resulted in the highest long-term genetic gains, the HEBVs were used to select the parents for the different layers. The H-score has also been considered in the scoping method to pre-select individuals, but a reduction of the genetic value was observed in the short as well as in the long term. The H-score results in the selection of heterozygous individuals with a lower GEBV, decreasing the genetic progress of the breeding population. Therefore, the H-score should only be used in combination with a greedy component. Allier et al. (2020a) combined the H-score with the selection of elite individuals based on GEBVs. In the deep scoping method, the H-score is only used to select the parents for the different layers, maximizing the preservation of the genetic variation while elite individuals are selected based on the GEBVs, ensuring the genetic progress.

9.2 Capacity required to preserve the genetic variation

As stated in the previous section, the loss of favorable QTL alleles over the first breeding cycles as observed for the scoping method could be reduced by selecting more parents. The effect of the parental selection intensity (PSI) is shown in Figure 9.1 for both truncation selection and the scoping method. Increasing the size of the parental population will result in higher genetic values and a higher maximum reachable genetic value, indicating that the favorable QTL alleles are better preserved in the breeding population. Further increasing the size of the parental population will only result in a smaller increase of the genetic value. Therefore, we could assume that for a certain parental population size, including more parents will no longer affect the long-term genetic gain. A breeder should, however, never aim for that point as it would increase the financial costs and required resources. The size of the breeding population is in most breeding programs limited by the available funds and resources, making it difficult to increase the population size. Based on the results illustrated in Figure 9.1, we can conclude that increasing the size of the breeding population will increase the long-term genetic gain. However, the breeder should find a good balance between the costs and profits of increasing the population size.

Increasing the population size will also allow for a better preservation of the genetic variation when truncation selection is used (baseline method). Because the scoping method can preserve the genetic variation more efficiently, according to Figure 9.1, truncation selection requires 10 times more parents to reach the same long-term genetic value as the scoping method (PSI = 50). This means that by using the scoping method with only 50 parents (and a population size of 500 individuals) the same long-term genetic gains can be obtained as observed for trunca-

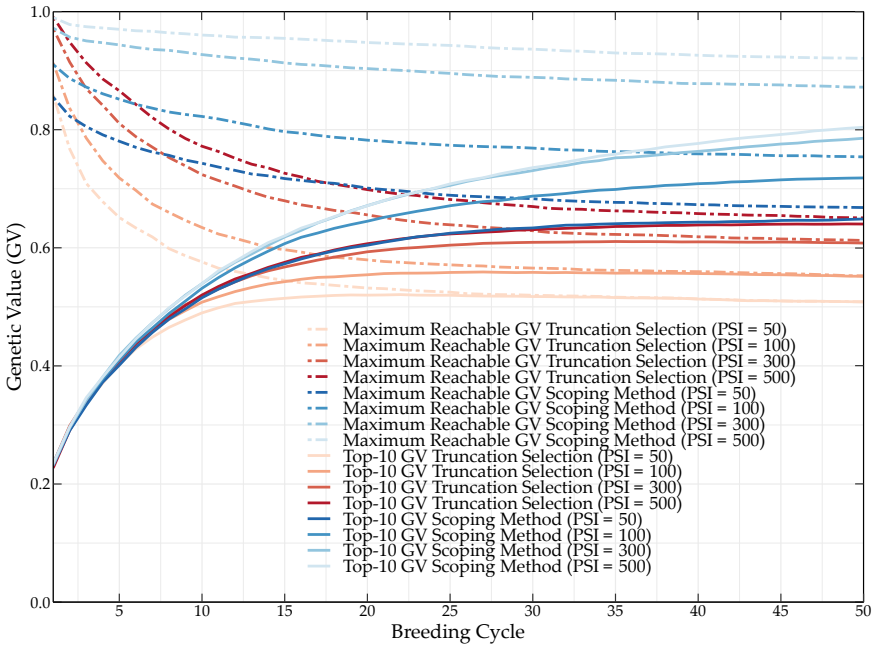

Figure 9.1: Simulation overview of truncation selection and the scoping method for different parental selection intensities (PSI).

tion selection using 500 parents (and a population size of 5000 individuals). This only emphasizes the importance of replacing truncation selection by the scoping method (or the deep scoping method).

The deep scoping method introduces genetic variation in the breeding population and will thus be less affected by the size of the parental population (see Figure 7.12). When a smaller population size is used, the number of layers should be reconsidered, but the long-term genetic gain will remain the same. Normally, when the number of parents is reduced, the genetic variation that can be passed to the next generation is limited, resulting in a rapid fixation of the QTL alleles. When genetic variation is introduced in the breeding cycle using the deep scoping method, the flow of favorable QTL alleles from the gene bank to the elite population is ensured. As long as the number of parents allows for a continuous flow of the genetic variation from the gene bank into the elite population, the parental population size will not affect the long-term genetic gain.

9.3 An alternative crossing block design

The loss of favorable QTL alleles could be reduced by further modifying the crossing block design. Although the scoping method was able to preserve a certain fraction of the genetic variation, by selecting only 50 parents based on the F_{score} (and 50 based on the GEBVs), not all the available marker alleles of the pre-selected population can be selected in the parental population. Therefore, new crossing block designs are needed.

In an attempt to maximize the genetic gain, the crossing block of the adaptive scoping method was redesigned. Therefore, the crossing block was split into two parts: the first part is selected according to the scoping method, whereas the second part is selected by maximizing the F_{score} between each parent and the already selected parents. In other words, in the second part, the P1 parent is not selected based on its GEBV, but is also selected to maximize the genetic variation of the parental population. When both parents are selected solely on their genetic variation (F_{score}), a low genetic gain would be observed in the next generation. Therefore, similar to the scoping rate, the fraction of the crossing block that is selected solely based on the F_{score} is decreased from 1 to 0 in t breeding cycles. At breeding cycle 1, all parents are selected based on the F_{score}. In the subsequent breeding cycles, a fraction of the parents will be selected according to the scoping method. This will increase the genetic progress at the expense of preservation of the genetic variation. At breeding cycle t and all the subsequent breeding cycles, all the parents are selected according to the scoping method, maximizing the genetic progress of the breeding population. The initial results are shown in Figure 9.2. By using an alternative crossing block design, the loss in favorable QTL alleles was avoided and high long-term genetic gains were observed, reaching values up to 0.80. Therefore, we believe that methods like this should be further investigated as they hold great promise for maximising genetic gain.

9.4 Using a high marker density in the scoping method

The number of available markers also plays a crucial role in the preservation of genetic variation. A low number of markers cannot always accurately grasp the QTL effects, resulting in poorly estimated GEBVs. Increasing the number of markers allows for a more accurate estimation of the QTL effects which in turn facilitates the identification and preservation of favorable QTL alleles in the population. However, a high number of markers might make it more challenging for the scoping method to preserve both alleles of each marker locus in the breeding population. Dedicated selection strategies are likely required to preserve the full genetic variation

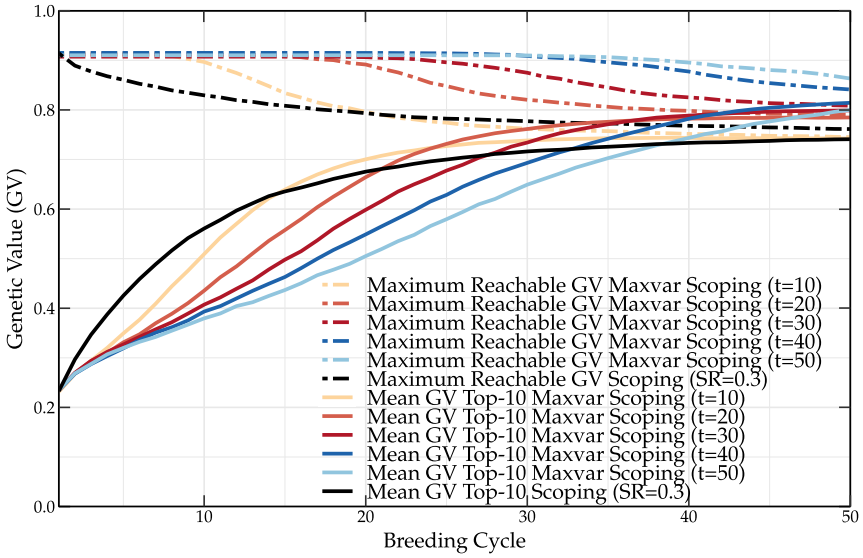

Figure 9.2: Mean genetic values of the top-10 individuals and maximum reachable genetic values for the scoping method and the adaptive scoping method using an alternative crossing block design (Maxvar Scoping).

in this setting. One way to achieve this is for example by optimizing the crossing block, allowing for a better preservation of rare marker alleles in the breeding population. This strategy could open new important opportunities for the use of the scoping method in different plant breeding applications.

9.5 Preserving the genetic variation in the long term

The deep scoping method was proposed as a new method that could replace the scoping method when the genetic variation of the breeding population has been reduced due to years of consecutive breeding. The deep scoping method was evaluated after 5, 10, 15, and 20 breeding cycles of truncation selection in which it consistently outperformed the scoping method. However, if the breeding population still contains a broad genetic variation as observed at the first breeding cycle, the scoping method can obtain a higher genetic gain. In Figure 9.3 the results of the scoping and deep scoping methods are shown after 1 and 3 breeding cycles of truncation selection. After three breeding cycles of truncation selection, the genetic variation is reduced to such an extent that the scoping method yields a lower long-term genetic gain compared to the deep scoping method. Although

the scoping method is still able to obtain higher short-term genetic gains, the point at which the deep scoping method will outperform the scoping method will be observed at an earlier point in time when the genetic variation of the base population is further reduced.

The loss of genetic variation has led to different bottlenecks in plant breeding. Therefore, breeders should make considerable efforts to reintroduce and preserve genetic variation in their breeding pools. Unfortunately, this remains difficult in crops with a low margin profit. Different techniques such as gene editing can help to increase the genetic variation in commercial breeding lines, but today's policies complicate the application of such methods (Louwaars, 2018).

9.6 Balance between the short- and long-term genetic gain

Different methods have been proposed to maximize the long-term genetic variation of the breeding population. The scoping method uses the scoping rate, a parameter that controls the number of individuals that will be pre-selected, and thus be available as a potential parent. The scoping rate can thus be used as a parameter to find the desired balance between the short- and long-term genetic gain. A scoping rate of 1, will use the whole breeding population to maximize the F_{score}, resulting in a high long-term, but a lower short-term genetic gain. By decreasing the scoping rate, the pre-selection of an individual with a lower GEBV is avoided. On the one hand, a lower scoping rate will result in higher short-term genetic gains while on the other hand, it will also expedite the loss in genetic variation, resulting in lower long-term genetic gains. In theory, the scoping rate could thus be used to maximize the genetic gain at a certain point in time.

In Chapter 6, the scoping rate is varied from SR_{max} to SR_{min} over t breeding cycles. Over the first breeding cycles, the preservation of genetic variation is prioritized, whereas over the later breeding cycles, when the scoping rate is decreased to SR_{min}, the genetic progress is maximized. The parameter t allows to maximize the genetic value after a certain number of breeding cycles, allowing the breeder to choose between short- or long-term genetic gains. For the deep scoping method, the short- and long-term genetic gains can be controlled by the size of the different layers and the size of the elite population, as well as the size of the pre-selection.

The effect of using a specific value for the scoping rate or t on the genetic value in the short as well as in the long term will depend on the characteristics of the current breeding population. To that end, a simulation study is required. By using phenotypic and genotypic data of the last breeding cycles, the effects of a parental selection method on the genetic gain of a specific breeding population

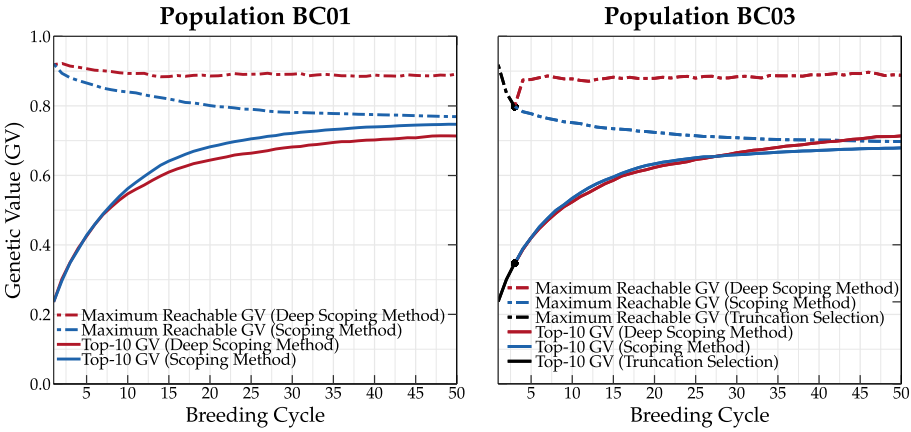

Figure 9.3: Simulation results of the deep scoping method and the scoping method starting after 1 (left), and 3 (right) breeding cycles of truncation selection. The scoping method can gain the highest long-term genetic values. However, if truncation selection is used prior to the scoping method, the long-term genetic gain decreases, and the deep scoping method can gain higher genetic gains.

can be assessed in the short and the long term. Unfortunately, the simulation of such a breeding population is difficult as it requires QTL information to predict the genetic values of future generations. In my opinion, further research is needed to study to which extent a simulator can be used to study the possible future outcomes of a specific breeding population. Taking into account that crossing overs occur at random, it will not be possible to perfectly predict the offspring of heterozygous parents. This implies that at each consecutive breeding cycle, one or more corrective measures should be considered. Nevertheless, such a simulator could advance the current genetic progress in different breeding programs.

9.7 Combining truncation selection and the scoping method

To maximize the short-term genetic gain, we have considered using truncation selection and the scoping method together to combine the advantages of both methods. To do so, the breeding population was split into two subpopulations: a greedy population and a scoping population. The idea was that the scoping method would preserve the genetic variation, ensuring high long-term genetic gains, whereas truncation selection would maximize the short-term genetic gains. Unfortunately, by combining truncation selection with the scoping method, a lower genetic gain was observed. Using truncation selection to select a fraction of the parental population resulted in a high selection intensity, decreasing the genetic variation of its offspring. In the next generation, the offspring of the truncated selected parents had a high GEBV and were thus pre-selected by the scoping method, reducing

the genetic variation in the pre-selected population. In other words, the scoping method was not able to preserve the genetic variation because truncation selection already reduced the genetic variation of the individuals that would be pre-selected. By selecting only 10 out of the 100 parents with truncation selection, the long-term genetic gain was significantly reduced without increasing the short-term genetic gain. This clearly indicates that the preservation of genetic variation is important to reach high long-term genetic gains.

In the deep scoping method, a variant of the scoping method was successfully combined with truncation selection. The deep scoping method uses different layers, restraining the individuals that can be selected as a parent. By also using a gene bank, genetic variation is constantly reintroduced in the breeding population, avoiding the permanent loss of favorable QTL alleles.

9.8 Pearson correlation versus mean genetic value of the top-10 individuals

In genomic selection, the Pearson correlation is often used to evaluate the prediction accuracy of the GEBVs. In Chapter 5, we proposed to only use the mean genetic value of the top-10 individuals to evaluate the performance of the different selection methods. The Pearson correlation is sensitive to outliers and can be misleading when a breeding population contains individuals with low and high genetic values. Starting at the first breeding cycle, the scoping and deep scoping methods are able to reach high long-term genetic gains (see Figure 9.4). As the deep scoping method constantly reintroduces individuals with a lower GEBV from the gene bank into the first layer, a high Pearson correlation is observed between the GEBVs and the genetic values of the whole breeding population. In the case of the scoping method, only the available genetic variation is preserved, and thus over time, the Pearson correlation degrades. When the deep scoping method is used, the Pearson correlation is overestimated, giving the impression that the GEBVs are better estimated compared to the scoping method while, at that point, the scoping method yields higher genetic values. On the one hand, in the deep scoping method, a cross between an elite individual and an individual of the gene bank can result in an outlier causing an overestimation of the prediction accuracy. On the other hand, the deep scoping method contains a high amount of genetic variation, resulting in a higher Pearson correlation compared to a similar breeding population that contains a lower amount of genetic variation (Glass and Hopkins, 1996; Goodwin and Leech, 2006). This is also shown in Figure 7.17 in which the Pearson correlation for each subpopulation of the deep scoping method is lower compared to the Pearson correlation of the whole breeding population. Therefore, we advise not to use the Pearson correlation when the breeding population con-

tains individuals with extreme breeding values or to compare the performance of two breeding populations that differentiate in terms of genetic variation.

9.9 Limitations of the simulation study

The simulation of a breeding population played a central role in this dissertation and made it possible to study the evolution of a breeding population under different circumstances. The development of a simulator can be time-consuming, especially when it is built from scratch. First, the biological aspects of a breeding population need to be translated into different mathematical equations that can, in the next step, be implemented into an application. Different assumptions and design choices (e.g. the distributions of the QTLs) are required. Finally, the simulation results should be validated using different datasets.

The number of simulators that are available to simulate a breeding population is rather limited. Faux et al. (2016) proposed Alphasim, a software package that can simulate a breeding population over different breeding cycles. Unfortunately, due to the closed-source nature of the software at that time, it was not possible to customize each aspect of the simulator, limiting its applicability in this dissertation. Moreover, as Alphasim was only made available as a pre-built binary, it was not possible to determine the exact implementation of various steps in the simulation process, making it difficult to fully understand the obtained results. Recently, the source code of Alphasim has been made publicly available, opening new perspectives for research studies that require simulated breeding data.

The simulator that was used in this dissertation was built upon the work of Neyhart et al. (2017). This simulator was written in R, allowing for the modification of each aspect of the implementation. In vivo, there can be so many interactions between the genes and the environment which makes it more difficult, if not impossible, to study the effect of a single, isolated parameter. In silico, however, one can choose to simplify various aspects of the underlying genetic model, for example, assuming the absence of dominance, epistatis and mutation. This simplifies the simulation, but still grasps the fundamentals of genetics. In a later stage, it can be interesting to also study the different parental selection methods under more realistic conditions. Therefore, each of these interactions should be implemented in the simulator.

The simulator can be extended to account for heterotic effects, such that the scoping and deep scoping methods can be evaluated in hybrid breeding. The simulation of the gametes can be improved by adding recombination hot spots and mutations and the QTL effects could be simulated based on a dataset, better reflecting the behavior of an actual breeding population.

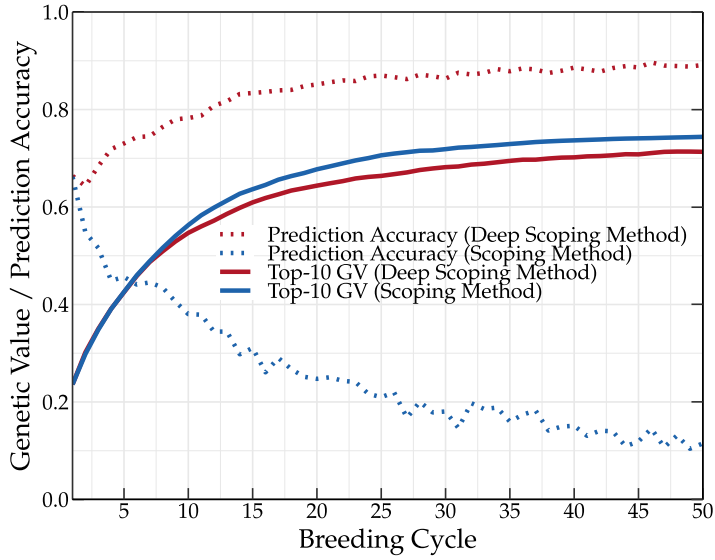

Figure 9.4: Mean genetic value of the top-10 individuals and Pearson correlation of a breeding population using the scoping and deep scoping methods. Both methods are used to select the parents over 50 breeding cycles.

9.10 Future work

In this dissertation, different parental selection methods were proposed. These methods were only studied in a recurrent breeding scheme. It could be interesting to alter our methods such that they can be used in other settings as well. For hybrid breeding, after taking into account the heterosis effect, the scoping and deep scoping methods could be used on the one hand to preserve the genetic variation of both heterotic groups and on the other hand, both methods could be used to guide the parental selection of the hybrids. Both selection methods could require a different approach and therefore require further investigation.

In Subsection 9.3, an alternative method was proposed. Although the initial results were promising, the robustness of the method should be further investigated, but this alternative method could in time replace the adaptive scoping method.

In this dissertation, we have demonstrated the importance of preserving genetic variation within a breeding population to maximize the long-term genetic gain of a trait. Often, crops are selected on more than one trait. Each trait is then controlled by a different number of QTLs. One QTL can control more than one trait, and thus if a QTL has a positive effect on one trait and a negative effect on another trait, it is less obvious which parents should be selected to maximize all the traits. Once again, a method using the core idea of the scoping method could assist the parental selection. The genetic gain can be maximized by finding which genotype

optimizes all the traits. By preserving both QTL alleles in the population, each QTL allele can still be reintroduced in an elite individual if necessary.

Hopefully, our proposed methods can also be evaluated in field tests and be used to improve current breeding populations. Both the scoping and deep scoping methods have been developed to maximize the long-term genetic gain while preserving the short-term genetic gain such that both methods can be used in a competitive setting.

10

Conclusion

Truncation selection is often used in GS and plant breeding to maximize the short-term genetic gain, but it also reduces the maximum reachable genetic value, causing a premature convergence of the genetic value to a local optimum. Although a greedy component is important to ensure genetic progress, the preservation of genetic variation is paramount to avoid a premature convergence of the genetic value. Both the scoping method and deep scoping method confirm our hypothesis: preserving genetic variation increases the long-term genetic gain. The F_{score} combines the genetic variation of each marker into a score that can be used to maximize the genetic variation of the offspring. The scoping method successfully selects parents based on the F_{score} and GEBVs, maximizing the genetic gain in the long term. The scoping rate, a parameter that controls the fraction of the breeding population that will be pre-selected, can be used to prioritize the preservation of genetic variation or to maximize the genetic progress. When the genetic variation of the breeding population has already been reduced due to years of consecutive breeding, a gene bank should be used to reintroduce the genetic variation in the breeding population. The deep scoping method uses different layers to increase the genetic values of the pre-breeding individuals before introducing them into the elite population. By doing so, the genetic gain is maximized in the short as well as in the long term.

Bibliography

Adeniyi, I. A., Yahya, W. B., 2020. Bayesian generalized linear mixed effects models using normal-independent distributions: Formulation and applications.

Akdemir, D., Sánchez, J. I., 2016. Efficient breeding by genomic mating. Front. Genet. 7, 1–12.

Akdemir, D., Sanchez, J. I., Jannink, J. L., 2015. Optimization of genomic selection training populations with a genetic algorithm. Genet. Sel. Evol. 47 (1), 1–10.

Allard, R. W., 2021. `https://www.britannica.com/science/plant-breeding/Breeding-self-pollinated-species`. Accessed: 2021-08-07.

Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., Lehermeier, C., 2019. Usefulness criterion and post-selection parental contributions in multi-parental crosses: Application to polygenic trait introgression. G3 9 (5), 1469–1479.

Allier, A., Teyssèdre, S., Lehermeier, C., Charcosset, A., Moreau, L., 2020a. Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. Theor. Appl. Genet. 133 (1), 201–215.

Allier, A., Teyssèdre, S., Lehermeier, C., Moreau, L., Charcosset, A., 2020b. Optimized breeding strategies to harness genetic resources with different performance levels. BMC Genet. 21 (1), 1–16.

Appleby, N., Edwards, D., Batley, J., 2009. New technologies for ultra-high throughput genotyping in plants. Plant Genomics, 19–39.

Ashkani, S., Rafii, M. Y., Rusli, I., Sariah, M., Abdullah, S. N. A., Rahim, H. A., Latif, M. A., 2012. Ssrs for marker-assisted selection for blast resistance in rice (oryza sativa l.). Plant Mol. Biol. Rep. 30 (1), 79–86.

Bargougui, M. A., 2016. Genetic analysis of barley (hordeum vulgare l.) grain yield components. J. New Sci. 31 (8), 1794–1799.

Bateson, W., 1909. Heredity and variation in modern lights. in: Seward, a.c. ed., darwin and modern science, cambridge university press, cambridge, 85-101.

Beckett, T. J., Morales, A. J., Koehler, K. L., Rocheford, T. R., 2017. Genetic relatedness of previously plant-variety-protected commercial maize inbreds. PLoS ONE 12 (12), 1–23.

Bennewitz, J., Meuwissen, T. H. E., 2005. A novel method for the estimation of the relative importance of breeds in order to conserve the total genetic variance. Genet. Sel. Evol. 37 (3), 315–337.

Bernardo, R., 2009. Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Sci. 49, 419–425.

Bernardo, R., 2014. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. Crop Sci. 54 (6), 2586–2595.

Bernardo, R., Yu, J., 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47 (3), 1082–1090.

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., Sehabiague, P., Makumbi, D., Magorokosho, C., Oikeh, S., Gakunga, J., Vargas, M., Olsen, M., Prasanna, B. M., Banziger, M., Crossa, J., 2015. Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. Crop Sci. 55 (1), 154–163.

Boerma, H. R., Cooper, R. L., 1975. Comparison of three selection procedures for yield in soybeans 1. Crop Sci. 15 (2), 225–229.

Boeven, P. H. G., Longin, C. F. H., Würschum, T., 2016. A unified framework for hybrid breeding and the establishment of heterotic groups in wheat. Theor. Appl. Genet. 129 (6), 1231–1245.

Bouchez, A., Hospital, F., Causse, M., Gallais, A., Charcosset, A., 2002. Marker-assisted introgression of favorable alleles at quantitative trait loci between maize elite lines. Genetics 162 (4), 1945–1959.

Brisbane, J. R., Gibson, J. P., 1995. Balancing selection response and inbreeding by including predicted stabilised genetic contributions in selection decisions. Genet. Sel. Evol. 27 (6), 541–549.

Bulmer, M. G., 1971. The effect of selection on genetic variability. Am. Nat. 105, 201–211.

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W., Veerkamp, R. F., 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178 (1), 553–561.

Camerarius, R. J., 1694. De Sexu Plantarum Epistola. Rommey, Tübingen.

Capettini, F., 2009. Plant Breeding and Farmer Participation. Food and Agriculture Organization of the United Nations, Ch. Selection methods Part 2: Pedigree method, pp. 223–228.

Cervantes, I., Gutiérrez, J. P., Meuwissen, T. H. E., 2016. Response to selection while maximizing genetic variance in small populations. Genet. Sel. Evol. 48 (1), 1–9.

Cervantes, I., Meuwissen, T. H. E., 2011. Maximization of total genetic variance in breed conservation programmes. J. Anim. Breed. Genet. 128 (6), 465–472.

Chagné, D., Batley, J., Edwards, D., Forster, J. W., 2007. Single nucleotide polymorphism genotyping in plants. In: Association mapping in plants. Springer, pp. 77–94.

Chang, L. Y., Toghiani, S., Ling, A., Aggrey, S. E., Rekaya, R., 2018. High density marker panels, SNPs prioritizing and accuracy of genomic selection. BMC Genet. 19 (1), 1–10.

Clark, S. A., Hickey, J. M., Van Der Werf, J. H. J., 2011. Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43 (1), 1–9.

Comstock, R. E., Robinson, H. F., Harvey, P. H., 1949. breeding procedure designed to make maximum use of both general and specific combining ability. Agron. J. 41 (8), 360–367.

Cox, R., Mirkin, S. M., 1997. Characteristic enrichment of dna repeats in different genomes. Proc. Natl. Acad. Sci. U.S.A. 94 (10), 5237–5242.

Cramer, M. M., Kannenberg, L. W., 1992. Five years of HOPE: the hierarchical open-ended corn breeding system. Crop Sci. 32 (5), 1163–1171.

Cros, D., Tchounke, B., Nkague-Nkamba, L., 2018. Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. Mol. Plant Breed. 38 (7), 1–12.

Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., Braun, H. J., 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186 (2), 713–724.

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., Hayes, B. J., 2015. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics 200 (4), 1341–1348.

De Coninck, A., Fostier, J., Maenhout, S., De Baets, B., 2014. DAIRRy-BLUP: A high performance computing approach to genomic prediction. Genetics 197, 813–822.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., Goddard, M. E., 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179 (3), 1503–1512.

Deery, D., Jimenez-Berni, J., Jones, H., Sirault, X., Furbank, R., 2014. Proximal remote sensing buggies and potential applications for field-based phenotyping. Agron. 4 (3), 349–379.

Dekkers, J. C. M., Hospital, F., 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3 (1), 22–32.

Devlin, S. J., Gnanadesikan, R., Kettenring, J. R., 12 1975. Robust estimation and outlier detection with correlation coefficients. Biometrika 62 (3), 531–545.

Du, B., Liu, L., Wang, Q., Sun, G., Ren, X., Li, C., Sun, D., 2019. Identification of qtl underlying the leaf length and area of different leaves in barley. Sci. Rep. (9), 4431–4439.

Duvick, D. N., 2005. The contribution of breeding to yield advances in maize (zea mays l.). Vol. 86 of Advances in Agronomy. Academic Press, pp. 83–145.

Edriss, V., Guldbrandtsen, B., Lund, M. S., Su, G., 2013. Effect of marker-data editing on the accuracy of genomic prediction. J. Anim. Breed. Genet. 130 (2), 128–135.

Endelman, J. B., 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4 (3), 250–255.

Eynard, S. E., Croiseau, P., Laloë, D., Fritz, S., Calus, M. P. L., Restoux, G., 2017. Which individuals to choose to update the reference population? Minimizing the loss of genetic diversity in animal genomic selection programs. G3 8 (1), 113–121.

Falconer, D. S., Mackay, T. F. C., 2000. Anecdotal, historical and critical commentaries on genetics. Genetics 154 (4), 1419–1426.

Faux, A. M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., Hearne, S. J., Gonen, S., Hickey, J. M., 2016. Alphasim: software for breeding program simulation. Plant Genome 9 (3), 1–14.

Fisher, R. A., 1918. The correlation between relatives on the supposition of mendelian inheritance. Trans. R. Soc. Edinb. 52, 399–433.

Furbank, R. T., Tester, M., 2011. Phenomics–technologies to relieve the phenotyping bottleneck. Trends Plant Sci. 16 (12), 635–644.

Gajghate, R., Tyagi, V., Rana, M., 2018. Single seed descent (ssd) method. Reader Shelf 14, 25–27.

Garrard, A., 1999. Charting the emergence of cereal and pulse domestication in south-west asia. Environ. Archaeol. 4, 67–86.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. (6), 721–741.

Glass, G., Hopkins, K., 1996. Statistical methods in education and psychology. Psyccritiques 41 (12).

Glaubitz, Y. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., Buckler, E. S., 2014. Tassel-gbs: A high capacity genotyping by sequencing analysis pipeline. PLOS ONE 9 (2), 1–11.

Goddard, M. E., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257.

Goddard, M. E., Hayes, B. J., 2002. Optimisation of response using molecular data. In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production. Vol. 33. pp. 3–10.

Goddard, M. E., Hayes, B. J., E., M. T. H., 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128, 409–421.

Goodwin, L. D., Leech, N. L., 2006. Understanding correlation: Factors that affect the size of r. Int. J. Exp. Educ. 74 (3), 249–266.

Gorjanc, G., Gaynor, R. C., Hickey, J. M., 2018. Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor. Appl. Genet. 131 (9), 1953–1966.

Gouesnard, B., Negro, S., Laffray, A., Glaubitz, J., Melchinger, A., Revilla, P., Moreno-Gonzalez, J., Madur, D., Combes, V., Tollon-Cordet, C., Laborde, J., Kermarrec, D., Bauland, C., Moreau, L., Charcosset, A., Nicolas, S., 2017. Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. Theor. Appl. Genet. 130 (10), 2165–2189.

Habier, D., Fernando, R. L., Kizilkaya, K., Garrick, D. J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12, 1–12.

Haldane, J. S., Thompson, D. W., Mitchell, P. C., Hobhouse, L. T., 1918. Symposium: Are physical, biological and psychological categories irreducible? Proc. Aristot. Soc. 1, 11–74.

Hallauer, A., 12 2007. History, contribution, and future of quantitative genetics in plant breeding: Lessons from maize. Crop Sci. 47, 1–16.

Han, Y., Cameron, J. N., Wang, L., Beavis, W. D., 2017. The predicted cross value for genetic introgression of multiple alleles. Genetics 205, 1409–1423.

Hayes, B., Goddard, M., P., V., 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91, 47–60.

Heffner, E. L., Lorenz, A. J., Jannink, J. L., Sorrells, M. E., 2010. Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 50 (5), 1681–1690.

Heffner, E. L., Sorrells, M. E., Jannink, J. L., 2009. Genomic selection for crop improvement. Crop Sci. 49 (1), 1–12.

Henderson, C. R., 1963. Selection index and expected genetic advance. Statistical Genetics and Plant Breeding 982, 141–163.

Hillman, G., Hedges, R., Moore, A., Colledge, S., Pettitt, P., 2001. New evidence of lateglacial cereals cultivation at abu hurayra on the euphrates. The Holocene 4 (11), 383–393.

Holland, J. B., 2004. Implementation of molecular markers for quantitative traits in breeding programs: challenges and opportunities. p. 26. In T. Fischer et al. (ed.) New Directions for a Diverse Planet: Proc. for the 4th Int. Crop Science Congress, Brisbane, Australia. 26 Sept. - 1 Oct. 2004. Regional Institute, Gosford, Australia.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T., Lan, T., Welch, A. J., Juárez, M. J. A., Simpson, J., et al., 2013. Architecture and evolution of a minute plant genome. Nature 498 (7452), 94–98.

Illumina, 2021. Trusted bead-based technology: A fundamentally different approach to high-density arrays.

Isidro, J., Jannink, J. L., Akdemir, D., Poland, J., Heslot, N., Sorrells, M. E., 2015. Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128 (1), 145–158.

Ivakhnenko, A. G., 1971. Polynomial theory of complex systems. IEEE Trans. Syst. Man Cybern. Syst. (4), 364–378.

Jannink, J. L., 2010. Dynamics of long-term genomic selection. Genet. Sel. Evol. 42 (1), 1–11.

Jannink, J. L., Bink, M., C., J. R., 2001. Using complex plant pedigrees to map valuable genes. Trands. Plant Sci. 6, 337–342.

Jannink, J. L., Lorenz, J., Iwata, H., 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9 (2), 166–177.

Johannsen, W., 1903. Ueber Erblichkeit in Populationen und in reinen Linien: ein Beitrag zur Beleuchtung schwebender Selektionsfragen. Fischer, Jena.

Juliana, P., Singh, R. P., Poland, J., Mondal, S., Crossa, J., Montesinos-López, O. A., Dreisigacker, S., Pérez-Rodríguez, P., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., 2018. Prospects and challenges of applied genomic selection - a new paradigm in breeding for grain yield in bread wheat. Plant Genome 11 (3), 1–17.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., Eskin, E., 2008. Efficient control of population structure in model organism association mapping. Genetics 178 (3), 1709–1723.

King, R. C., Stansfield, W. D., 1997. A Dictionary of Genetics, 5th Edition. Oxford University Press, New York.

Kumar, P., 2007. Green revolution and its impact on environment. Int. j. res. soc. sci. 5 (3), 54–57.

Laloë, D., 1993. Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25 (6), 557–576.

Lande, R., Thompson, R., 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124, 743–756.

Lange, K., 1978. Central limit theorems for pedigrees. J. Math. Biol. 6, 59–66.

Lange, K., 1997. Mathematical and Statistical Methods for Genetic Analysis. Springer New York, Ch. The Polygenic Model, pp. 123–141.

Lehermeier, C., Teyssèdre, S., Schön, C. C., 2017. Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. Genetics 207 (4), 1651–1661.

Leister, D., Varotto, C., Pesaresi, P., Niwergall, A., Salamini, F., 1999. Large-scale evaluation of plant growth in arabidopsis thaliana by non-invasive image analysis. Plant Physiol. Biochem. 37 (9), 671–678.

Lenz, P. R. N., Nadeau, S., Mottet, M., Perron, M., Isabel, N., Beaulieu, J., Bousquet, J., 2020. Multi-trait genomic selection for weevil resistance, growth, and wood quality in norway spruce. Evolutionary applications 13 (1), 76–94.

Lindgren, D., Mullin, T. J., 1997. Balancing gain and relatedness in selection. Silvae Genet. 3 (2), 124–129.

Liu, H., Meuwissen, T. H. E., Sørensen, A. C., Berg, P., 2015. Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. Genet. Sel. Evol. 47 (1), 1–14.

Lorenz, A. J., Hamblin, M. T., Jannink, J. L., 2010. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PLoS ONE 5 (11), 1–11.

Louwaars, N. P., 2018. Plant breeding and diversity: A troubled relationship? Euphytica 214 (7), 1–9.

Maenhout, S., De Baets, B., Haesaert, G., 2010. Graph-based data selection for the construction of genomic prediction models. Genetics 185 (4), 1463–1475.

Martini, J. W. R., Rosales, F., Ha, N. T., Heise, J., Wimmer, V., Kneib, T., 2019. Lost in translation: On the problem of data coding in penalized whole genome regression with interactions. G3 9 (4), 1117–1129.

Mason, A. S., 2015. Ssr genotyping. In: Plant genotyping. Springer, pp. 77–89.

Melchinger, A. E., Gumber, R. K., 1998a. Concepts and Breeding of Heterosis in Crop Plants. John Wiley & Sons, Ltd, Ch. Overview of Heterosis and Heterotic Groups in Agronomic Crops, pp. 29–44.

Melchinger, A. E., Gumber, R. K., 1998b. Overview of Heterosis and Heterotic Groups in Agronomic Crops. Vol. 25. Wiley, pp. 29–44.

Mendel, G., 1866. Experiments in plant hybridization. Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb.org/Mendel.html (accessed on 1 April 2021).

Meuwissen, T. H. E., 1997. Maximizing the respond of selection with a predifined rate of inbreeding. J. Anim. Sci. 75, 934–940.

Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Miedaner, T., Korzun, V., 2012. Marker-assisted selection for disease resistance in wheat and barley breeding. Phytopathology 102 (6), 560–566.

Moore, A. M. T., Hillman, G. C., Legge, A. J., 2000. Village on the Euphrates: from Foraging to Farming at Abu Hureyra. Oxford University Press, New York.

Morgante, M., Rafalski, A., Biddle, P., Tingey, S., Olivieri, A., 1994. Genetic mapping and variability of seven soybean simple sequence repeat loci. Genome 37 (5), 763–769.

Moser, G., Tier, B., E., C. R., Khatkar, M. S., Raadsma, H. W., 2009. A comparison of five methods to predict genomic breeding values of diary bulls from genome-wide snp markers. Genet. Sel. Evol. 41, 56.

Müller, D., Schopp, P., Melchinger, A. E., 2018. Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. G3 8 (4), 1173–1181.

Neves, H. H. R., Carvalheiro, R., Queiroz, S. A., 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genet. 13, 1–17.

Neyhart, J. L., Tiede, T., Lorenz, A. J., Smith, K. P., 2017. Evaluating methods of updating training data in long-term genomewide selection. G3 7 (5), 1499–1510.

Orf, J. H., 2008. Breeding, Genetics, and Production of Soybeans. Soybeans: Chemistry, Production, Processing, and Utilization, 33–65.

Patterson, H. D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58 (3), 545–554.

Pejic, I., Ajmone-Marsan, P., Morgante, M., Kozumplick, V., Castiglioni, P., Taramino, G., Motto, M., 1998. Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. Theor. Appl. Genet. 97 (8), 1248–1255.

Pellicer, J., Hidalgo, O., Dodsworth, S., Leitch, I. J., 2018. Genome size diversity and its impact on the evolution of land plants. Genes 9 (2), 88.

Pérez, P., de los Campos, G., 2014. Genome-wide regression and prediction with the bglr statistical package. Genetics 198 (2), 483–495.

Piepho, H. P., Möhring, J., Melchinger, A. E., Büchse, A., 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161 (1), 209–228.

Pincot, D. D. A., Hardigan, M. A., Cole, G. S., Famula, R. A., Henry, P. M., Gordon, T. R., Knapp, S. J., 2020. Accuracy of genomic selection and long-term genetic gain for resistance to verticillium wilt in strawberry. The Plant Genome 13 (3), 1–19.

Pryce, J. E., Mekonnen, H., 2020. Symposium review: Genomic selection for reducing environmental impact and adapting to climate change. Int. J. Dairy Sci. 103 (6), 5366–5375.

Pszczola, M., Strabel, T., Mulder, H. A., Calus, M. P. L., 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95 (1), 389–400.

Pyeritz, R. E., Korf, B. R., Grody, W. W., 2018. Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics: Foundations. Academic Press.

Qi-Lun, Y., Ping, F., Ke-Cheng, K., Guang-Tang, P., 2008. Genetic diversity based on ssr markers in maize (zea mays l.) landraces from wiling mountain region in china. J. Genet. 87 (3), 287–291.

Rabier, C. E., Barre, P., Asp, T., Charmet, G., Mangin, B., 2016. On the accuracy of genomic selection. PloS one 11 (6), 1–16.

Rédei, G. P., 2008. Encyclopedia of Genetics, Genomics, Proteomics, and Informatics. Springer Science & Business Media.

Rendel, J. M., Robertson, A., 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. J. Genet. 50, 1–8.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V. M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C. C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., Moreau, L., 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics 192 (2), 715–728.

Roychowdhury, R., Taoutaou, A., Hakeem, K. R., Gawwad, M. R. A., Tah, J., 2013. Crop Improvement in the Era of Climate Change. Springer.

Salhuana, W., Pollak, L. M., 01 2006. Latin american maize project (lamp) and germplasm enhancement of maize (gem) project: generating useful breeding germplasm. Maydica 51, 339–355.

Sayfzadeh, S., Honarvar, M., Taheri, F., Afshari, K. P., Nourbakhsh, H., 2013. Analyses and comparison of genomic accuracy of random regression blup and bayesian ridge regression. AGC. Res. 57 (3), 32–37.

Schön, C., Utz, S., Groh, B., Truberg, S., Openshaw, S., Melchiner, A., 2004. Qtl mapping based on resampling in a vast maize testcross experiment confi rms the infi nitesimal model of quantitative genetics for complex traits. Genetics 167, 485–498.

Schopp, P., Müller, D., Wientjes, Y. C. J., Melchinger, A. E., 2017. Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops. G3 7, 3571–3586.

Shull, G. H., Gowen, J. W., 1952. Heterosis. Iowa State College Press, Ch. Beginnings of the Heterosis Concept.

Simmonds, N. W., 1993. Introgression and incorporation. strategies for the use of crop genetic resources. Biol. Rev. 68, 539–562.

Smith, S., Beavis, W., 1996. The impact of plant molecular genetics. Birkhäuser Boston, Boston, MA, Ch. Molecular marker assisted breeding in a company environment, pp. 259–272.

Sneller, C. H., Mather, D. E., Crepieux, S., 2009. Analytical approaches and population types for finding and utilizing QTL in complex plant populations. Crop Sci. 49, 363–380.

Sonesson, A. K., Woolliams, J. A., Meuwissen, T. H. E., 2012. Genomic selection requires genomic control of inbreeding. Genet. Sel. Evol. 44 (27), 1–10.

Song, Z. P., Xu, X., Wang, B., Chen, J. K., Lu, B. R., 2003. Genetic diversity in the northernmost oryza rufipogon populations estimated by ssr markers. Theor. Appl. Genet. 107, 1492–1499.

Sprague, G. F., Tatum, L. A., 1942. General vs. specific combining ability in single crosses of corn. J. Amer. Soc. Agron. 34 (39), 923–932.

Srinivasan, S., Madhuparni, S., S., S. B., G., P. M., 2020. Rapid generation advance (RGA) in chickpea to produce up to seven generations per year and enable speed breeding. Crop J. 8, 164–169.

Strand, M., Prolla, T. A., Liskay, R. M., Petes, T. D., 1993. Destabilization of tracts of simple repetitive dna in yeast by mutations affecting dna mismatch repair. Nature 365 (6443), 274–276.

Stuiver, M., Reimer, P. J., Bard, E., Beck, J. W., Burr, G. S., Hughen, K. A., Kromer, B., McCormac, G., van der Plicht, J., Spurk, M., 1998. Intcal98 radiocarbon age calibration, 24,000-0 cal bp. Radiocarbon 40 (3), 1041–1083.

Sun, J., Poland, J. A., Mondal, S., Crossa, J., Juliana, P., Singh, R. P., Rutkoski, J. E., Jannink, J. L., Crespo-Herrera, L., Velu, G., Huerta-Espino, J., Sorrells, M. E., 2019. High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. Theor. Appl. Genet. 132 (6), 1705–1720.

Suontama, M., Klápště, J., Telfer, E., Graham, N., Stovold, T., Low, C., McKinley, R., Dungey, H., 2019. Efficiency of genomic prediction across two Eucalyptus nitens seed orchards with different selection histories. Heredity 122 (3), 370–379.

Sweeney, D. W., Rooney, T. E., Sorrells, M. E., 2021. Gain from genomic selection for a selection index in two-row spring barley. The Plant Genome, 1–14.

Teich, A. H., 1984. Heritability of grain yield, plant height and test weight of a population of winter wheat adapted to southwestern ontario. Theor. Appl. Genet. 68, 21–23.

Tester, M., Landridge, P., 2010. Breeding Technologies to Increase Crop Production in a Changing World. Science 327, 818–822.

Thompson, R., Cullis, B., Smith, A., Gilmour, A., 2003. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. Aust. N. Z. J. Stat. 45 (4), 445–459.

VanRaden, P. M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91 (11), 4414–4423.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., Schenkel, F. S., 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92 (1), 16–24.

Vilmorin, 2021. Investors presentation. `https://www.vilmorincie.com/wp-content/uploads/2021/02/02_Investors-presentation_Fevrier-2021.pdf`, accessed: 2021-08-07.

Voss-Fels, K. P., Cooper, M., Hayes, B. J., 2018. Accelerating crop genetic gains with genomic selection. Theor. Appl. Genet. 132 (3), 669–686.

Walter, A., Liebisch, F., Hund, A., 2015. Plant phenotyping: from bean weighing to image analysis. Plant methods 11 (1), 1–11.

Wang, C. S., Rutledge, J. J., Gianola, D., 1993. Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. 25 (1), 41–62.

Wang, Q., Yu, Y., Zhang, Q., Zhang, X., Huang, H., Xiang, J., Li, F., 2019. Evaluation on the genomic selection in Litopenaeus vannamei for the resistance against Vibrio parahaemolyticus. Aquaculture 505 (2), 212–216.

Weber, J. L., May, P. E., 1989. Abundant class of human dna polymorphisms which can be typed using the polymerase chain reaction. American journal of human genetics 44 (3), 388.

White, T. L., 2004. Tree breeding, principles. Elsevier, Ch. Breeding Theory and Genetic Testing, pp. 1551–1561.

Whitford, R., Fleury, D., Reif, J. C., Garcia, M., Okada, T., Korzun, V., Langridge, P., 2013. Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. J. Exp. Bot. 64 (18), 5411–5428.

Winter, P., Kahl, G., 1995. Molecular marker technology for plant improvement. World journal of Microbiology and Biotechnology 11, 438–448.

Woolliams, J., Bijma, P., Villanueva, B., 2000. Expected genetic contribution and their impact on gene flow and genetic gain. Genetics 153, 1009–1020.

Woolliams, J. A., Berg, P., Dagnachew, B. S., Meuwissen, T. H. E., 2015. Genetic contributions and their optimization. J. Anim. Breed. Genet. 132 (2), 89–99.

Woolliams, J. A., Wray, N. R., Thompson, R., 1993. Prediction of long-term contributions and inbreeding in populations undergoing mass selection. Genet. Res. 62, 231–242.

Wray, N. R., Goddard, M. E., 1994. Increasing long-term response to selection. Genet. Sel. Evol. 26, 431–451.

Wray, N. R., Thompson, R., 1990. Prediction of rates of inbreeding in selected populations. Genet. Res. 55, 41–54.

Wright, S., 1931. Evolution in mendelian populations. Genetics 16 (2), 97–159.

Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., Sudha, N., Olsen, M., Prasanna, B. M., Zhang, X., Babu, R., 2016. Molecular characterization of cimmyt maize inbred lines with genotyping-by-sequencing snps. Theor. Appl. Genet. 129, 753–765.

Xu, X., Sharma, R., Tondelli, A., Russell, J., Comadran, J., Schnaithmann, F., Pillen, K., Kilian, B., Cattivelli, L., Thomas, W. T. B., 2018. Genome-wide association analysis of grain yield-associated traits in a pan-european barley cultivar collection. Plant Genome 11 (1), 170073.

Xu, Y., 2010. Molecular plant breeding. Cabi.

Xu, Y., Crouch, J. H., 2008. Marker-assisted selection in plant breeding: from publications to practice. Crop Sci. 48, 391–407.

Yang, J., B., B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., Visscher, P. M., 2010. Common snps explain a large proportion of heritability forhuman height. National Genetics 42 (7), 565–569.

Yule, G. U., 1907. On the Theory of Inheritance of Quantitative Compound Characters on the Basis of Mendel's Laws: a Preliminary Note. Spottiswoode & Company, Limited.

Zhao, H., Li, Y., Petkowski, J., Kant, S., Hayden, M. J., Daetwyler, H. D., 2020. Genomic prediction and genomic heritability of grain yield and its related traits in a safflower genebank collection. Plant Genome, 1–15.

Appendix

Table 10.1: The mean genetic value and the standard deviation of the top-10 individuals for each parental selection method over 250 experiments.

Breeding Cycle	Baseline Method	Scoping Method (SR = 0.1)	Scoping Method (SR = 0.3)	Scoping Method (SR = 0.6)	Population Merit Method ($c = 20$)	MVT Method
1	0.23 ± 0.08	0.23 ± 0.08	0.23 ± 0.08	0.23 ± 0.08	0.23 ± 0.08	0.23 ± 0.08
3	0.34 ± 0.08	0.34 ± 0.08	0.34 ± 0.07	0.34 ± 0.08	0.34 ± 0.08	0.34 ± 0.08
5	0.41 ± 0.09	0.42 ± 0.07	0.41 ± 0.07	0.40 ± 0.08	0.40 ± 0.08	0.40 ± 0.08
7	0.46 ± 0.08	0.47 ± 0.08	0.47 ± 0.07	0.45 ± 0.07	0.45 ± 0.08	0.45 ± 0.08
9	0.50 ± 0.08	0.51 ± 0.08	0.51 ± 0.07	0.50 ± 0.07	0.49 ± 0.07	0.48 ± 0.08
11	0.52 ± 0.08	0.54 ± 0.08	0.55 ± 0.07	0.53 ± 0.07	0.52 ± 0.08	0.51 ± 0.08
13	0.54 ± 0.08	0.56 ± 0.08	0.57 ± 0.07	0.56 ± 0.07	0.54 ± 0.07	0.53 ± 0.08
15	0.55 ± 0.08	0.57 ± 0.07	0.60 ± 0.07	0.58 ± 0.06	0.56 ± 0.07	0.54 ± 0.08
17	0.55 ± 0.08	0.58 ± 0.08	0.62 ± 0.07	0.60 ± 0.07	0.57 ± 0.07	0.55 ± 0.08
19	0.56 ± 0.08	0.59 ± 0.08	0.63 ± 0.07	0.62 ± 0.07	0.58 ± 0.07	0.56 ± 0.08
21	0.56 ± 0.08	0.59 ± 0.08	0.65 ± 0.07	0.63 ± 0.07	0.60 ± 0.07	0.56 ± 0.08
23	0.56 ± 0.08	0.59 ± 0.08	0.66 ± 0.07	0.65 ± 0.07	0.60 ± 0.07	0.57 ± 0.08
25	0.56 ± 0.08	0.60 ± 0.08	0.66 ± 0.07	0.66 ± 0.07	0.61 ± 0.07	0.57 ± 0.08
27	0.56 ± 0.08	0.60 ± 0.08	0.67 ± 0.07	0.67 ± 0.07	0.61 ± 0.07	0.57 ± 0.08
29	0.56 ± 0.08	0.60 ± 0.08	0.68 ± 0.07	0.68 ± 0.06	0.62 ± 0.07	0.58 ± 0.08
31	0.56 ± 0.08	0.60 ± 0.08	0.68 ± 0.07	0.68 ± 0.06	0.62 ± 0.07	0.58 ± 0.08
33	0.56 ± 0.08	0.60 ± 0.08	0.69 ± 0.07	0.69 ± 0.06	0.63 ± 0.07	0.58 ± 0.08
35	0.56 ± 0.08	0.60 ± 0.08	0.69 ± 0.07	0.70 ± 0.06	0.63 ± 0.07	0.58 ± 0.08
37	0.56 ± 0.08	0.60 ± 0.08	0.70 ± 0.07	0.70 ± 0.06	0.63 ± 0.07	0.58 ± 0.08
39	0.56 ± 0.08	0.60 ± 0.08	0.70 ± 0.07	0.71 ± 0.06	0.63 ± 0.07	0.58 ± 0.08
41	0.56 ± 0.08	0.60 ± 0.08	0.70 ± 0.07	0.71 ± 0.06	0.63 ± 0.07	0.58 ± 0.08
43	0.56 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.71 ± 0.06	0.63 ± 0.07	0.58 ± 0.08
45	0.56 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.72 ± 0.06	0.64 ± 0.07	0.58 ± 0.08
47	0.56 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.72 ± 0.06	0.64 ± 0.07	0.58 ± 0.08
49	0.56 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.73 ± 0.06	0.64 ± 0.07	0.58 ± 0.08
50	0.56 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.73 ± 0.06	0.64 ± 0.07	0.58 ± 0.08

Table 10.2: The mean genetic value and the standard deviation of the breeding population for each parental selection method over 250 experiments.

Breeding Cycle	Baseline Method	Scoping Method (SR = 0.1)	Scoping Method (SR = 0.3)	Scoping Method (SR = 0.6)	Population Merit Method ($c = 20$)	MVT Method
1	0.07 ± 0.09	0.07 ± 0.09	0.07 ± 0.09	0.07 ± 0.09	0.07 ± 0.09	0.07 ± 0.09
3	0.20 ± 0.09	0.20 ± 0.08	0.19 ± 0.08	0.17 ± 0.09	0.19 ± 0.08	0.19 ± 0.08
5	0.29 ± 0.09	0.29 ± 0.08	0.26 ± 0.08	0.24 ± 0.08	0.26 ± 0.08	0.27 ± 0.08
7	0.35 ± 0.09	0.35 ± 0.08	0.33 ± 0.08	0.30 ± 0.08	0.32 ± 0.08	0.33 ± 0.08
9	0.40 ± 0.09	0.41 ± 0.08	0.38 ± 0.08	0.35 ± 0.08	0.37 ± 0.08	0.42 ± 0.08
11	0.44 ± 0.08	0.45 ± 0.08	0.42 ± 0.07	0.39 ± 0.07	0.40 ± 0.08	0.42 ± 0.08
13	0.47 ± 0.08	0.48 ± 0.08	0.46 ± 0.07	0.42 ± 0.07	0.44 ± 0.08	0.45 ± 0.08
15	0.49 ± 0.08	0.50 ± 0.08	0.49 ± 0.07	0.45 ± 0.07	0.46 ± 0.08	0.47 ± 0.08
17	0.50 ± 0.08	0.52 ± 0.08	0.52 ± 0.07	0.48 ± 0.07	0.48 ± 0.08	0.49 ± 0.08
19	0.52 ± 0.08	0.53 ± 0.08	0.54 ± 0.07	0.50 ± 0.07	0.50 ± 0.08	0.50 ± 0.08
21	0.52 ± 0.08	0.54 ± 0.08	0.55 ± 0.07	0.52 ± 0.07	0.51 ± 0.08	0.51 ± 0.08
23	0.53 ± 0.08	0.55 ± 0.08	0.57 ± 0.07	0.53 ± 0.07	0.53 ± 0.08	0.52 ± 0.08
25	0.53 ± 0.08	0.56 ± 0.08	0.58 ± 0.07	0.55 ± 0.07	0.54 ± 0.08	0.53 ± 0.08
27	0.54 ± 0.08	0.56 ± 0.08	0.59 ± 0.07	0.56 ± 0.07	0.55 ± 0.08	0.53 ± 0.08
29	0.54 ± 0.08	0.57 ± 0.08	0.60 ± 0.07	0.57 ± 0.07	0.55 ± 0.08	0.54 ± 0.08
31	0.54 ± 0.08	0.57 ± 0.08	0.61 ± 0.07	0.59 ± 0.07	0.56 ± 0.08	0.54 ± 0.08
33	0.55 ± 0.08	0.57 ± 0.08	0.62 ± 0.07	0.59 ± 0.07	0.56 ± 0.08	0.55 ± 0.08
35	0.55 ± 0.08	0.58 ± 0.08	0.63 ± 0.07	0.60 ± 0.07	0.57 ± 0.08	0.55 ± 0.08
37	0.55 ± 0.08	0.58 ± 0.08	0.63 ± 0.07	0.61 ± 0.07	0.57 ± 0.08	0.55 ± 0.08
39	0.55 ± 0.08	0.58 ± 0.08	0.64 ± 0.07	0.62 ± 0.07	0.58 ± 0.08	0.55 ± 0.08
41	0.55 ± 0.08	0.58 ± 0.08	0.64 ± 0.07	0.63 ± 0.07	0.58 ± 0.08	0.56 ± 0.08
43	0.55 ± 0.08	0.58 ± 0.08	0.65 ± 0.07	0.63 ± 0.07	0.58 ± 0.08	0.56 ± 0.08
45	0.55 ± 0.08	0.58 ± 0.08	0.65 ± 0.08	0.64 ± 0.07	0.59 ± 0.08	0.56 ± 0.08
47	0.55 ± 0.08	0.58 ± 0.08	0.66 ± 0.07	0.64 ± 0.07	0.59 ± 0.08	0.56 ± 0.08
49	0.55 ± 0.08	0.58 ± 0.08	0.66 ± 0.07	0.65 ± 0.07	0.59 ± 0.08	0.56 ± 0.08
50	0.55 ± 0.08	0.58 ± 0.08	0.66 ± 0.07	0.65 ± 0.07	0.59 ± 0.08	0.56 ± 0.08

Table 10.3: The maximum reachable genetic value and the standard deviation of the breeding population for each parental selection method over 250 experiments.

Breeding Cycle	Baseline Method	Scoping Method (SR = 0.1)	Scoping Method (SR = 0.3)	Scoping Method (SR = 0.6)	Population Merit Method ($c = 20$)	MVT Method
1	0.91 ± 0.05	0.90 ± 0.05	0.90 ± 0.05	0.90 ± 0.05	0.91 ± 0.05	0.91 ± 0.05
3	0.78 ± 0.07	0.78 ± 0.08	0.86 ± 0.06	0.88 ± 0.05	0.84 ± 0.07	0.80 ± 0.07
5	0.72 ± 0.08	0.74 ± 0.08	0.84 ± 0.06	0.86 ± 0.06	0.80 ± 0.07	0.75 ± 0.08
7	0.68 ± 0.08	0.71 ± 0.08	0.82 ± 0.07	0.85 ± 0.06	0.76 ± 0.07	0.71 ± 0.08
9	0.65 ± 0.09	0.69 ± 0.09	0.81 ± 0.07	0.85 ± 0.06	0.74 ± 0.07	0.68 ± 0.08
11	0.63 ± 0.08	0.67 ± 0.09	0.80 ± 0.07	0.84 ± 0.06	0.73 ± 0.07	0.66 ± 0.08
13	0.62 ± 0.09	0.65 ± 0.08	0.79 ± 0.07	0.84 ± 0.06	0.71 ± 0.07	0.64 ± 0.08
15	0.60 ± 0.09	0.64 ± 0.08	0.79 ± 0.07	0.83 ± 0.06	0.70 ± 0.07	0.63 ± 0.09
17	0.59 ± 0.08	0.63 ± 0.08	0.78 ± 0.07	0.83 ± 0.06	0.70 ± 0.07	0.62 ± 0.08
19	0.59 ± 0.08	0.63 ± 0.08	0.78 ± 0.07	0.82 ± 0.06	0.69 ± 0.07	0.61 ± 0.09
21	0.58 ± 0.08	0.62 ± 0.08	0.77 ± 0.07	0.82 ± 0.06	0.68 ± 0.07	0.61 ± 0.08
23	0.58 ± 0.08	0.62 ± 0.08	0.77 ± 0.07	0.82 ± 0.06	0.68 ± 0.07	0.60 ± 0.08
25	0.58 ± 0.09	0.61 ± 0.08	0.76 ± 0.07	0.82 ± 0.06	0.68 ± 0.07	0.60 ± 0.08
27	0.57 ± 0.09	0.61 ± 0.08	0.76 ± 0.07	0.81 ± 0.06	0.68 ± 0.07	0.60 ± 0.08
29	0.57 ± 0.09	0.61 ± 0.08	0.76 ± 0.07	0.81 ± 0.06	0.67 ± 0.07	0.59 ± 0.08
31	0.57 ± 0.09	0.61 ± 0.08	0.76 ± 0.07	0.81 ± 0.06	0.67 ± 0.07	0.59 ± 0.08
33	0.57 ± 0.09	0.61 ± 0.08	0.75 ± 0.07	0.81 ± 0.06	0.67 ± 0.07	0.59 ± 0.08
35	0.57 ± 0.09	0.61 ± 0.08	0.75 ± 0.07	0.81 ± 0.06	0.67 ± 0.07	0.59 ± 0.08
37	0.57 ± 0.09	0.60 ± 0.08	0.75 ± 0.07	0.81 ± 0.06	0.67 ± 0.07	0.58 ± 0.08
39	0.57 ± 0.08	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06	0.67 ± 0.07	0.58 ± 0.08
41	0.56 ± 0.08	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06	0.66 ± 0.07	0.58 ± 0.08
43	0.56 ± 0.09	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06	0.66 ± 0.07	0.58 ± 0.08
45	0.56 ± 0.09	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06	0.66 ± 0.07	0.58 ± 0.08
47	0.56 ± 0.08	0.60 ± 0.08	0.74 ± 0.07	0.80 ± 0.06	0.66 ± 0.07	0.58 ± 0.08
49	0.56 ± 0.08	0.60 ± 0.08	0.74 ± 0.07	0.80 ± 0.06	0.66 ± 0.07	0.58 ± 0.08
50	0.56 ± 0.08	0.60 ± 0.08	0.74 ± 0.07	0.80 ± 0.06	0.66 ± 0.07	0.58 ± 0.08

Table 10.4: The mean genetic value and the standard deviation of the top-10 individuals for the backcrossing and combined methods over 250 experiments.

BC	Backcrossing Method	Combined Method (SR = 0.1)	Combined Method (SR = 0.2)	Combined Method (SR = 0.3)
1	0.23 ± 0.09	0.23 ± 0.08	0.23 ± 0.08	0.23 ± 0.08
3	0.34 ± 0.08	0.34 ± 0.08	0.34 ± 0.07	0.34 ± 0.08
5	0.41 ± 0.08	0.42 ± 0.07	0.41 ± 0.07	0.40 ± 0.08
7	0.46 ± 0.08	0.47 ± 0.08	0.47 ± 0.07	0.45 ± 0.07
9	0.49 ± 0.08	0.51 ± 0.08	0.51 ± 0.07	0.50 ± 0.07
11	0.45 ± 0.08	0.54 ± 0.08	0.55 ± 0.07	0.53 ± 0.07
13	0.50 ± 0.08	0.56 ± 0.08	0.57 ± 0.07	0.56 ± 0.07
15	0.53 ± 0.08	0.57 ± 0.08	0.60 ± 0.07	0.58 ± 0.06
17	0.55 ± 0.08	0.58 ± 0.08	0.62 ± 0.07	0.60 ± 0.07
19	0.56 ± 0.08	0.59 ± 0.08	0.63 ± 0.07	0.62 ± 0.07
21	0.49 ± 0.08	0.59 ± 0.08	0.65 ± 0.07	0.63 ± 0.07
23	0.55 ± 0.08	0.59 ± 0.08	0.66 ± 0.07	0.65 ± 0.07
25	0.57 ± 0.09	0.60 ± 0.08	0.66 ± 0.07	0.66 ± 0.07
27	0.59 ± 0.08	0.60 ± 0.08	0.67 ± 0.07	0.67 ± 0.07
29	0.59 ± 0.08	0.60 ± 0.08	0.68 ± 0.07	0.68 ± 0.06
31	0.51 ± 0.08	0.60 ± 0.08	0.68 ± 0.07	0.68 ± 0.06
33	0.57 ± 0.08	0.60 ± 0.08	0.69 ± 0.07	0.69 ± 0.06
35	0.60 ± 0.09	0.60 ± 0.08	0.69 ± 0.07	0.70 ± 0.06
37	0.61 ± 0.09	0.60 ± 0.08	0.70 ± 0.07	0.70 ± 0.06
39	0.61 ± 0.09	0.60 ± 0.08	0.70 ± 0.07	0.71 ± 0.06
41	0.53 ± 0.08	0.60 ± 0.08	0.70 ± 0.07	0.71 ± 0.06
43	0.59 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.71 ± 0.06
45	0.61 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.72 ± 0.06
47	0.62 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.72 ± 0.06
49	0.63 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.73 ± 0.06
50	0.63 ± 0.08	0.60 ± 0.08	0.71 ± 0.07	0.73 ± 0.06

Table 10.5: The maximum reachable genetic value and the standard deviation of the breeding population for the backcrossing and combined methods over 250 experiments.

BC	Backcrossing Method	Combined Method (SR = 0.1)	Combined Method (SR = 0.2)	Combined Method (SR = 0.3)
1	0.90 ± 0.05	0.90 ± 0.05	0.90 ± 0.05	0.90 ± 0.05
3	0.77 ± 0.08	0.78 ± 0.08	0.86 ± 0.06	0.88 ± 0.05
5	0.71 ± 0.09	0.74 ± 0.08	0.84 ± 0.06	0.86 ± 0.06
7	0.67 ± 0.09	0.71 ± 0.08	0.82 ± 0.07	0.85 ± 0.06
9	0.64 ± 0.09	0.69 ± 0.09	0.81 ± 0.07	0.85 ± 0.06
11	0.79 ± 0.08	0.67 ± 0.09	0.80 ± 0.07	0.84 ± 0.06
13	0.71 ± 0.08	0.65 ± 0.08	0.79 ± 0.07	0.84 ± 0.06
15	0.66 ± 0.09	0.64 ± 0.08	0.79 ± 0.07	0.84 ± 0.06
17	0.64 ± 0.09	0.63 ± 0.08	0.78 ± 0.07	0.83 ± 0.06
19	0.63 ± 0.09	0.63 ± 0.08	0.78 ± 0.07	0.83 ± 0.06
21	0.80 ± 0.07	0.62 ± 0.08	0.77 ± 0.07	0.82 ± 0.06
23	0.71 ± 0.08	0.62 ± 0.08	0.77 ± 0.07	0.82 ± 0.06
25	0.67 ± 0.08	0.61 ± 0.08	0.76 ± 0.07	0.82 ± 0.06
27	0.65 ± 0.08	0.61 ± 0.08	0.76 ± 0.07	0.82 ± 0.06
29	0.64 ± 0.09	0.61 ± 0.08	0.76 ± 0.07	0.81 ± 0.06
31	0.80 ± 0.08	0.61 ± 0.08	0.76 ± 0.07	0.81 ± 0.06
33	0.73 ± 0.08	0.61 ± 0.08	0.75 ± 0.07	0.81 ± 0.06
35	0.69 ± 0.08	0.61 ± 0.08	0.75 ± 0.07	0.81 ± 0.06
37	0.67 ± 0.08	0.60 ± 0.08	0.75 ± 0.07	0.81 ± 0.06
39	0.66 ± 0.08	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06
41	0.80 ± 0.07	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06
43	0.73 ± 0.08	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06
45	0.70 ± 0.08	0.60 ± 0.08	0.75 ± 0.07	0.80 ± 0.06
47	0.68 ± 0.08	0.60 ± 0.08	0.74 ± 0.07	0.80 ± 0.06
49	0.67 ± 0.08	0.60 ± 0.08	0.74 ± 0.07	0.80 ± 0.06
50	0.66 ± 0.08	0.60 ± 0.08	0.74 ± 0.07	0.80 ± 0.06

Table 10.6: The mean genetic value and the standard deviation of the top-10 individuals using the scoping and adaptive scoping methods, averaged over 100 runs.

Breeding	Scoping	Adaptive Scoping				
Cycle	SR = 0.3	$t = 10$	$t = 20$	$t = 30$	$t = 40$	$t = 50$
1	0.23 ± 0.09	0.23 ± 0.09	0.23 ± 0.09	0.23 ± 0.09	0.23 ± 0.09	0.23 ± 0.09
3	0.35 ± 0.09	0.32 ± 0.08	0.32 ± 0.08	0.32 ± 0.08	0.32 ± 0.08	0.32 ± 0.08
5	0.43 ± 0.08	0.39 ± 0.08	0.38 ± 0.08	0.38 ± 0.08	0.38 ± 0.08	0.37 ± 0.08
7	0.49 ± 0.08	0.45 ± 0.08	0.44 ± 0.08	0.43 ± 0.08	0.43 ± 0.08	0.42 ± 0.08
9	0.54 ± 0.08	0.51 ± 0.07	0.49 ± 0.08	0.48 ± 0.08	0.48 ± 0.07	0.47 ± 0.08
11	0.58 ± 0.08	0.57 ± 0.07	0.53 ± 0.08	0.52 ± 0.08	0.52 ± 0.07	0.51 ± 0.07
13	0.61 ± 0.08	0.61 ± 0.07	0.58 ± 0.08	0.56 ± 0.08	0.56 ± 0.07	0.54 ± 0.07
15	0.64 ± 0.08	0.65 ± 0.07	0.61 ± 0.07	0.59 ± 0.07	0.59 ± 0.07	0.57 ± 0.07
17	0.66 ± 0.08	0.67 ± 0.07	0.65 ± 0.07	0.62 ± 0.08	0.62 ± 0.07	0.60 ± 0.07
19	0.67 ± 0.08	0.69 ± 0.07	0.67 ± 0.07	0.65 ± 0.08	0.64 ± 0.07	0.63 ± 0.07
21	0.68 ± 0.08	0.70 ± 0.07	0.70 ± 0.07	0.67 ± 0.08	0.66 ± 0.07	0.65 ± 0.07
23	0.69 ± 0.08	0.71 ± 0.07	0.72 ± 0.06	0.69 ± 0.08	0.68 ± 0.07	0.67 ± 0.07
25	0.70 ± 0.08	0.72 ± 0.07	0.73 ± 0.06	0.70 ± 0.08	0.70 ± 0.07	0.69 ± 0.07
27	0.71 ± 0.08	0.72 ± 0.07	0.74 ± 0.06	0.72 ± 0.07	0.71 ± 0.07	0.70 ± 0.07
29	0.72 ± 0.08	0.72 ± 0.07	0.74 ± 0.06	0.73 ± 0.07	0.72 ± 0.07	0.71 ± 0.07
31	0.72 ± 0.08	0.72 ± 0.07	0.75 ± 0.06	0.74 ± 0.07	0.73 ± 0.07	0.72 ± 0.07
33	0.73 ± 0.08	0.72 ± 0.07	0.75 ± 0.06	0.75 ± 0.07	0.74 ± 0.07	0.73 ± 0.07
35	0.73 ± 0.08	0.72 ± 0.07	0.75 ± 0.06	0.75 ± 0.07	0.75 ± 0.06	0.74 ± 0.07
37	0.73 ± 0.08	0.73 ± 0.07	0.75 ± 0.06	0.76 ± 0.07	0.76 ± 0.06	0.75 ± 0.07
39	0.74 ± 0.08	0.73 ± 0.07	0.75 ± 0.06	0.76 ± 0.07	0.77 ± 0.06	0.75 ± 0.07
41	0.74 ± 0.08	0.73 ± 0.07	0.76 ± 0.06	0.76 ± 0.07	0.77 ± 0.06	0.76 ± 0.07
43	0.74 ± 0.08	0.73 ± 0.07	0.76 ± 0.06	0.76 ± 0.07	0.77 ± 0.06	0.76 ± 0.07
45	0.74 ± 0.08	0.72 ± 0.07	0.76 ± 0.06	0.76 ± 0.07	0.78 ± 0.06	0.77 ± 0.07
47	0.74 ± 0.08	0.72 ± 0.07	0.76 ± 0.06	0.77 ± 0.07	0.78 ± 0.06	0.77 ± 0.07
49	0.74 ± 0.08	0.72 ± 0.07	0.76 ± 0.06	0.77 ± 0.07	0.78 ± 0.06	0.78 ± 0.07
50	0.74 ± 0.08	0.72 ± 0.07	0.76 ± 0.06	0.77 ± 0.07	0.78 ± 0.06	0.78 ± 0.07

Table 10.7: The maximum reachable genetic value and the standard deviation of a breeding population using the scoping and the adaptive scoping methods, averaged over 100 runs.

Breeding Cycle	Scoping	Adaptive Scoping				
	SR = 0.3	$t = 10$	$t = 20$	$t = 30$	$t = 40$	$t = 50$
1	0.92 ± 0.05	0.91 ± 0.05	0.92 ± 0.05	0.92 ± 0.05	0.92 ± 0.05	0.91 ± 0.05
3	0.88 ± 0.07	0.89 ± 0.06	0.90 ± 0.05	0.90 ± 0.05	0.90 ± 0.05	0.90 ± 0.05
5	0.86 ± 0.07	0.89 ± 0.06	0.89 ± 0.05	0.89 ± 0.05	0.89 ± 0.05	0.89 ± 0.05
7	0.85 ± 0.07	0.88 ± 0.06	0.88 ± 0.06	0.88 ± 0.06	0.88 ± 0.05	0.89 ± 0.05
9	0.84 ± 0.07	0.87 ± 0.06	0.88 ± 0.06	0.88 ± 0.06	0.88 ± 0.05	0.88 ± 0.05
11	0.83 ± 0.08	0.85 ± 0.06	0.88 ± 0.06	0.87 ± 0.06	0.88 ± 0.05	0.88 ± 0.06
13	0.82 ± 0.08	0.83 ± 0.07	0.87 ± 0.06	0.87 ± 0.06	0.87 ± 0.05	0.87 ± 0.06
15	0.81 ± 0.08	0.81 ± 0.07	0.86 ± 0.06	0.87 ± 0.06	0.87 ± 0.05	0.87 ± 0.06
17	0.81 ± 0.08	0.79 ± 0.07	0.86 ± 0.05	0.86 ± 0.06	0.87 ± 0.06	0.86 ± 0.06
19	0.80 ± 0.08	0.78 ± 0.07	0.85 ± 0.06	0.86 ± 0.06	0.86 ± 0.06	0.86 ± 0.06
21	0.80 ± 0.08	0.77 ± 0.07	0.83 ± 0.06	0.85 ± 0.06	0.86 ± 0.06	0.86 ± 0.06
23	0.79 ± 0.08	0.76 ± 0.07	0.82 ± 0.05	0.85 ± 0.06	0.85 ± 0.06	0.86 ± 0.06
25	0.79 ± 0.08	0.75 ± 0.07	0.81 ± 0.06	0.84 ± 0.06	0.85 ± 0.06	0.85 ± 0.06
27	0.78 ± 0.08	0.75 ± 0.07	0.80 ± 0.06	0.83 ± 0.06	0.85 ± 0.06	0.85 ± 0.06
29	0.78 ± 0.08	0.74 ± 0.07	0.79 ± 0.06	0.83 ± 0.06	0.84 ± 0.06	0.85 ± 0.06
31	0.78 ± 0.08	0.74 ± 0.07	0.78 ± 0.06	0.82 ± 0.06	0.84 ± 0.06	0.84 ± 0.06
33	0.78 ± 0.08	0.74 ± 0.07	0.78 ± 0.06	0.81 ± 0.07	0.83 ± 0.06	0.84 ± 0.06
35	0.77 ± 0.08	0.73 ± 0.07	0.77 ± 0.06	0.80 ± 0.07	0.83 ± 0.06	0.84 ± 0.06
37	0.77 ± 0.08	0.73 ± 0.07	0.77 ± 0.05	0.79 ± 0.07	0.83 ± 0.06	0.83 ± 0.06
39	0.77 ± 0.08	0.73 ± 0.07	0.77 ± 0.05	0.79 ± 0.07	0.82 ± 0.06	0.83 ± 0.06
41	0.77 ± 0.08	0.73 ± 0.07	0.77 ± 0.06	0.78 ± 0.07	0.81 ± 0.06	0.83 ± 0.06
43	0.77 ± 0.08	0.73 ± 0.07	0.76 ± 0.06	0.78 ± 0.07	0.81 ± 0.06	0.83 ± 0.06
45	0.77 ± 0.08	0.73 ± 0.07	0.76 ± 0.06	0.78 ± 0.07	0.80 ± 0.06	0.82 ± 0.06
47	0.77 ± 0.08	0.73 ± 0.07	0.76 ± 0.06	0.77 ± 0.07	0.80 ± 0.06	0.82 ± 0.06
49	0.76 ± 0.08	0.73 ± 0.07	0.76 ± 0.06	0.77 ± 0.07	0.80 ± 0.06	0.81 ± 0.06
50	0.76 ± 0.08	0.73 ± 0.07	0.76 ± 0.06	0.77 ± 0.07	0.79 ± 0.06	0.81 ± 0.06

Table 10.8: The mean genetic value and the standard deviation of the top-10 individuals for the deep scoping method and the HUC method with bridging over 100 experiments.

BC	Deep scoping (BC05)	HUC (BC05)	Deep scoping (BC10)	HUC (BC10)	Deep scoping (BC15)	HUC (BC15)	Deep scoping (BC20)	HUC (BC20)
1	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
5	0.40 ± 0.09	0.40 ± 0.09	-	-	-	-	-	-
7	0.46 ± 0.09	0.46 ± 0.09	-	-	-	-	-	-
9	0.49 ± 0.08	0.50 ± 0.08	-	-	-	-	-	-
11	0.53 ± 0.08	0.53 ± 0.08	0.52 ± 0.09	0.52 ± 0.09	-	-	-	-
13	0.55 ± 0.08	0.56 ± 0.08	0.53 ± 0.09	0.54 ± 0.09	-	-	-	-
15	0.57 ± 0.08	0.58 ± 0.08	0.55 ± 0.08	0.55 ± 0.09	0.54 ± 0.09	0.54 ± 0.09	-	-
17	0.59 ± 0.08	0.60 ± 0.07	0.56 ± 0.08	0.57 ± 0.09	0.55 ± 0.09	0.55 ± 0.09	-	-
19	0.61 ± 0.08	0.61 ± 0.07	0.58 ± 0.08	0.59 ± 0.09	0.56 ± 0.09	0.56 ± 0.09	-	-
21	0.63 ± 0.08	0.62 ± 0.07	0.60 ± 0.08	0.60 ± 0.09	0.57 ± 0.09	0.57 ± 0.09	0.56 ± 0.09	0.56 ± 0.09
23	0.64 ± 0.08	0.63 ± 0.07	0.61 ± 0.08	0.61 ± 0.09	0.58 ± 0.08	0.58 ± 0.09	0.56 ± 0.09	0.56 ± 0.09
25	0.66 ± 0.08	0.64 ± 0.07	0.63 ± 0.08	0.62 ± 0.09	0.60 ± 0.08	0.59 ± 0.09	0.57 ± 0.09	0.57 ± 0.09
27	0.67 ± 0.08	0.65 ± 0.07	0.64 ± 0.08	0.63 ± 0.08	0.62 ± 0.08	0.60 ± 0.09	0.59 ± 0.08	0.58 ± 0.09
29	0.68 ± 0.07	0.65 ± 0.07	0.65 ± 0.08	0.63 ± 0.08	0.63 ± 0.08	0.61 ± 0.08	0.60 ± 0.08	0.59 ± 0.09
31	0.69 ± 0.07	0.66 ± 0.07	0.66 ± 0.08	0.64 ± 0.08	0.65 ± 0.08	0.61 ± 0.09	0.61 ± 0.08	0.60 ± 0.09
33	0.69 ± 0.07	0.66 ± 0.07	0.67 ± 0.08	0.65 ± 0.08	0.66 ± 0.08	0.62 ± 0.08	0.62 ± 0.08	0.61 ± 0.08
35	0.70 ± 0.07	0.66 ± 0.07	0.68 ± 0.07	0.65 ± 0.08	0.66 ± 0.08	0.63 ± 0.08	0.63 ± 0.08	0.62 ± 0.08
37	0.71 ± 0.07	0.67 ± 0.07	0.69 ± 0.08	0.66 ± 0.08	0.67 ± 0.08	0.63 ± 0.08	0.65 ± 0.08	0.62 ± 0.08
39	0.71 ± 0.07	0.67 ± 0.07	0.70 ± 0.08	0.66 ± 0.08	0.68 ± 0.08	0.64 ± 0.08	0.66 ± 0.08	0.63 ± 0.08
41	0.72 ± 0.07	0.67 ± 0.07	0.70 ± 0.08	0.66 ± 0.08	0.69 ± 0.08	0.64 ± 0.08	0.66 ± 0.08	0.64 ± 0.08
43	0.72 ± 0.07	0.67 ± 0.07	0.71 ± 0.07	0.67 ± 0.08	0.70 ± 0.08	0.65 ± 0.08	0.67 ± 0.08	0.64 ± 0.08
45	0.73 ± 0.07	0.68 ± 0.07	0.71 ± 0.07	0.67 ± 0.08	0.70 ± 0.08	0.65 ± 0.08	0.68 ± 0.08	0.65 ± 0.08
47	0.73 ± 0.07	0.68 ± 0.06	0.72 ± 0.07	0.67 ± 0.08	0.71 ± 0.08	0.65 ± 0.08	0.69 ± 0.08	0.65 ± 0.08
49	0.73 ± 0.06	0.68 ± 0.07	0.72 ± 0.07	0.68 ± 0.08	0.71 ± 0.08	0.66 ± 0.08	0.69 ± 0.08	0.65 ± 0.08
50	0.74 ± 0.07	0.68 ± 0.07	0.73 ± 0.07	0.68 ± 0.08	0.72 ± 0.08	0.66 ± 0.08	0.70 ± 0.08	0.66 ± 0.08

Table 10.9: The maximum reachable genetic value, the genetic value of the fixed QTL alleles, the mean genetic value of the top-10 individuals and the standard deviation using the true selection method. The genetic values are averaged over 100 different experiments.

BC	Maximum Reachable GV	Fixed GV	Top-10 GV
1	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
3	0.91 ± 0.05	0.09 ± 0.05	0.30 ± 0.09
5	0.91 ± 0.05	0.16 ± 0.06	0.42 ± 0.09
7	0.91 ± 0.05	0.24 ± 0.07	0.52 ± 0.09
9	0.91 ± 0.05	0.32 ± 0.08	0.61 ± 0.08
11	0.91 ± 0.05	0.42 ± 0.08	0.68 ± 0.08
13	0.91 ± 0.05	0.52 ± 0.09	0.73 ± 0.07
15	0.91 ± 0.05	0.61 ± 0.08	0.77 ± 0.07
17	0.91 ± 0.05	0.69 ± 0.08	0.79 ± 0.06
19	0.91 ± 0.05	0.73 ± 0.07	0.81 ± 0.06
21	0.91 ± 0.05	0.77 ± 0.07	0.83 ± 0.06
23	0.91 ± 0.05	0.79 ± 0.06	0.84 ± 0.06
25	0.91 ± 0.05	0.80 ± 0.06	0.84 ± 0.06
27	0.91 ± 0.05	0.81 ± 0.07	0.85 ± 0.06
29	0.91 ± 0.05	0.82 ± 0.07	0.86 ± 0.06
31	0.91 ± 0.05	0.83 ± 0.06	0.86 ± 0.06
33	0.91 ± 0.05	0.84 ± 0.06	0.87 ± 0.06
35	0.91 ± 0.05	0.84 ± 0.06	0.87 ± 0.06
37	0.91 ± 0.05	0.85 ± 0.06	0.87 ± 0.06
39	0.91 ± 0.05	0.85 ± 0.06	0.87 ± 0.06
41	0.91 ± 0.05	0.85 ± 0.06	0.87 ± 0.06
43	0.91 ± 0.05	0.86 ± 0.06	0.87 ± 0.06
45	0.91 ± 0.05	0.86 ± 0.06	0.87 ± 0.06
47	0.91 ± 0.05	0.86 ± 0.06	0.88 ± 0.06
49	0.91 ± 0.05	0.86 ± 0.06	0.88 ± 0.06
50	0.91 ± 0.05	0.86 ± 0.06	0.88 ± 0.06

Table 10.10: The maximum reachable genetic value and standard deviation of the breeding population using different training panel update methods. The maximum reachable genetic value is averaged over 100 experiments.

BC	Forward	Stepwise	Best	CDmean	PEVmean	Random	Tails	Worst
1	0.92 ± 0.06	0.91 ± 0.05	0.91 ± 0.05	0.91 ± 0.05	0.91 ± 0.06	0.91 ± 0.05	0.91 ± 0.05	0.91 ± 0.05
2	0.80 ± 0.08	0.80 ± 0.07	0.79 ± 0.08	0.79 ± 0.08	0.80 ± 0.09	0.80 ± 0.07	0.79 ± 0.08	0.80 ± 0.08
3	0.76 ± 0.09	0.76 ± 0.08	0.72 ± 0.09	0.74 ± 0.09	0.74 ± 0.10	0.75 ± 0.09	0.73 ± 0.09	0.76 ± 0.08
4	0.73 ± 0.08	0.73 ± 0.08	0.67 ± 0.09	0.70 ± 0.09	0.70 ± 0.11	0.71 ± 0.09	0.68 ± 0.10	0.72 ± 0.08
5	0.70 ± 0.08	0.71 ± 0.08	0.62 ± 0.10	0.66 ± 0.09	0.66 ± 0.11	0.66 ± 0.09	0.64 ± 0.10	0.69 ± 0.08
6	0.69 ± 0.08	0.70 ± 0.08	0.59 ± 0.10	0.63 ± 0.10	0.63 ± 0.12	0.64 ± 0.09	0.61 ± 0.10	0.66 ± 0.08
7	0.68 ± 0.08	0.69 ± 0.08	0.57 ± 0.10	0.60 ± 0.10	0.61 ± 0.12	0.62 ± 0.10	0.59 ± 0.10	0.64 ± 0.08
8	0.67 ± 0.08	0.68 ± 0.08	0.54 ± 0.09	0.58 ± 0.09	0.58 ± 0.12	0.59 ± 0.10	0.56 ± 0.09	0.61 ± 0.08
9	0.66 ± 0.08	0.67 ± 0.08	0.52 ± 0.10	0.57 ± 0.09	0.56 ± 0.12	0.58 ± 0.10	0.54 ± 0.09	0.59 ± 0.08
10	0.65 ± 0.08	0.66 ± 0.08	0.51 ± 0.10	0.55 ± 0.09	0.55 ± 0.12	0.56 ± 0.10	0.53 ± 0.10	0.57 ± 0.09
11	0.65 ± 0.08	0.66 ± 0.08	0.50 ± 0.10	0.54 ± 0.09	0.54 ± 0.12	0.55 ± 0.10	0.51 ± 0.10	0.56 ± 0.09
12	0.64 ± 0.08	0.65 ± 0.08	0.49 ± 0.10	0.53 ± 0.09	0.53 ± 0.12	0.54 ± 0.10	0.51 ± 0.10	0.55 ± 0.09
13	0.64 ± 0.08	0.65 ± 0.08	0.49 ± 0.10	0.52 ± 0.09	0.52 ± 0.12	0.53 ± 0.09	0.50 ± 0.10	0.54 ± 0.09
14	0.64 ± 0.08	0.64 ± 0.08	0.48 ± 0.10	0.51 ± 0.09	0.51 ± 0.12	0.52 ± 0.09	0.49 ± 0.09	0.53 ± 0.09
15	0.63 ± 0.08	0.64 ± 0.08	0.48 ± 0.09	0.51 ± 0.09	0.50 ± 0.12	0.51 ± 0.10	0.48 ± 0.09	0.52 ± 0.09

Table 10.11: The mean genetic value and standard deviation of the top-10 individuals using different training panel update methods. The genetic value is averaged over 100 experiments.

BC	Forward	Stepwise	Best	CDmean	PEVmean	Random	Tails	Worst
1	0.23 ± 0.09	0.24 ± 0.09	0.24 ± 0.09	0.24 ± 0.09	0.23 ± 0.09	0.23 ± 0.09	0.23 ± 0.09	0.23 ± 0.09
2	0.27 ± 0.09	0.27 ± 0.09	0.27 ± 0.09	0.27 ± 0.09	0.27 ± 0.09	0.27 ± 0.09	0.27 ± 0.09	0.27 ± 0.09
3	0.33 ± 0.09	0.33 ± 0.08	0.31 ± 0.09	0.30 ± 0.09	0.30 ± 0.09	0.31 ± 0.09	0.31 ± 0.09	0.30 ± 0.09
4	0.38 ± 0.08	0.38 ± 0.08	0.34 ± 0.09	0.33 ± 0.09	0.33 ± 0.09	0.33 ± 0.09	0.34 ± 0.09	0.33 ± 0.09
5	0.43 ± 0.08	0.43 ± 0.08	0.35 ± 0.09	0.35 ± 0.09	0.35 ± 0.09	0.36 ± 0.09	0.36 ± 0.09	0.35 ± 0.09
6	0.47 ± 0.08	0.48 ± 0.08	0.37 ± 0.09	0.37 ± 0.09	0.37 ± 0.09	0.38 ± 0.09	0.38 ± 0.09	0.37 ± 0.09
7	0.51 ± 0.07	0.51 ± 0.08	0.39 ± 0.09	0.38 ± 0.09	0.38 ± 0.09	0.39 ± 0.09	0.39 ± 0.09	0.38 ± 0.09
8	0.54 ± 0.07	0.54 ± 0.08	0.40 ± 0.09	0.40 ± 0.09	0.39 ± 0.09	0.41 ± 0.09	0.41 ± 0.09	0.40 ± 0.09
9	0.56 ± 0.07	0.56 ± 0.08	0.41 ± 0.09	0.41 ± 0.09	0.40 ± 0.09	0.42 ± 0.09	0.42 ± 0.09	0.41 ± 0.09
10	0.58 ± 0.07	0.58 ± 0.08	0.41 ± 0.09	0.42 ± 0.09	0.41 ± 0.10	0.43 ± 0.09	0.43 ± 0.09	0.42 ± 0.09
11	0.59 ± 0.07	0.60 ± 0.07	0.42 ± 0.09	0.43 ± 0.09	0.42 ± 0.10	0.44 ± 0.09	0.43 ± 0.09	0.43 ± 0.09
12	0.60 ± 0.07	0.61 ± 0.08	0.43 ± 0.09	0.44 ± 0.09	0.43 ± 0.10	0.44 ± 0.09	0.44 ± 0.09	0.44 ± 0.09
13	0.61 ± 0.07	0.62 ± 0.07	0.43 ± 0.09	0.44 ± 0.09	0.43 ± 0.10	0.45 ± 0.09	0.44 ± 0.09	0.44 ± 0.08
14	0.62 ± 0.07	0.62 ± 0.08	0.44 ± 0.09	0.45 ± 0.09	0.44 ± 0.10	0.45 ± 0.09	0.44 ± 0.09	0.45 ± 0.09
15	0.62 ± 0.08	0.63 ± 0.08	0.44 ± 0.09	0.45 ± 0.09	0.44 ± 0.10	0.46 ± 0.09	0.45 ± 0.09	0.45 ± 0.09

Curriculum vitae

Personalia

Name	David Marcel K. Vanavermaete
Date of birth	21/10/1993
Place of birth	Ronse, Belgium
Nationality	Belgian
E-mail	David.Vanavermaete@gmail.com

Education

University

2015 - 2017

M.Sc. in Bioscience Engineering: Chemistry and Bioprocess Technology, University of Ghent, Ghent, Belgium

2012 - 2015

B.Sc in Bioscience Engineering (Chemistry and Food Technology), University of Ghent, Ghent, Belgium

Secondary education

2006 - 2012

Biotechnology, Vrij Landelijk Instituut Oudenaarde, Oudenaarde, Belgium

Employment

September 2017 - August 2021

Full-time researcher (Ph.D.) at the Research Unit Knowledge-Based Systems, Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University

Tutorship of master and bachelor theses

De rol van predictiemodellen in genetische vooruitgang in de plantenteelt. Master thesis, academic year 2020-2021 (Jonas Vanhaecke).

Scientific output

Publications in international journals (ISI-papers)

D.M. Vanavermaete, J. Fostier, S. Maenhout, B. De Baets (2020). Preservation of genetic variation in a breeding population for long-term genetic gain. G3: Genes, Genomes, Genetics 10.8: 2753-2762.

D.M. Vanavermaete, J. Fostier, S. Maenhout and B. De Baets (2021). Deep scoping: a breeding strategy to preserve, reintroduce and exploit genetic variation. Theoretical and Applied Genetics, 1-17.

Submitted manuscripts

D.M. Vanavermaete, J. Fostier, S. Maenhout and B. De Baets. Adaptive scoping: balancing short- and long-term genetic gain in plant breeding (submitted 2021).

Conference Contribution

D.M. Vanavermaete, J. Fostier, S. Maenhout, B. De Baets (2017). Algorithms for training panel construct in agricultural settings. Eucarpia biometric, September 2018, Ghent, Belgium.

Peer review

Peer review for Heredity, September 2021, Springer Nature, London, GB.