FACULTEIT
BIO-INGENIEURSWETENSCHAPPEN

DE ROL VAN PREDICTIEMODELLEN IN GENETISCHE VOORUITGANG IN DE PLANTENTEELT

Jonas Vanhaecke

Promotor: prof. dr. Bernard De Baets

Tutor: ir. David Vanavermaete

Masterproef voorgelegd voor het behalen van de graad Master of Science in de industriële wetenschappen: biochemie

Academiejaar: 2020–2021

UNIVERSITEIT
GENT

De auteur en promotor geven de toelating deze scriptie voor consultatie beschikbaar te stellen en delen ervan te kopiëren voor persoonlijk gebruik. Elk ander gebruik valt onder de beperkingen van het auteursrecht, in het bijzonder met betrekking tot de verplichting uitdrukkelijk de bron te vermelden bij het aanhalen van resultaten uit deze scriptie.

The author and promoter give the permission to use this thesis for consultation and to copy parts of it for personal use. Every other use is subject to the copyright laws, more specifically the source must be extensively specified when using results from this thesis.

Gent, 21 mei 2021

De auteur,

Jonas Vanhaecke

De promotor en tutor,

prof. dr. Bernard De Baets

ir. David Vanavermaete

Woord vooraf

Het woord vooraf, iets waarmee u als lezer start en waarmee ik als schrijver van deze thesis eindig. Een ideaal moment dus om even kort terug te blikken op het gehele proces van het schrijven van deze thesis. Het maken van deze thesis was, vermoedelijk net zoals bij vele andere thesisstudenten, een proces van vallen en opstaan. Uiteindelijk kan ik wel stellen dat het een zeer leerrijk proces is geweest waardoor ik zaken heb bijgeleerd die mij later zeker nog van pas kunnen komen. In het begin was het maken van deze thesis als een volledig nieuwe wereld die openging. Van de eerste papers die ik las in verband met het onderwerp begreep ik niet veel en pas na verloop van tijd werd meer en meer duidelijk. Ik vond het een zeer interessant onderwerp en het voorkomen van meerdere domeinen binnen het onderzoek (genetica, wiskunde, informatica...) was ook zeker een meerwaarde vond ik.

Ik zou graag enkele mensen willen bedanken die deze thesis mee mogelijk hebben gemaakt. In eerste instantie wil ik graag mijn tutor, ir. David Vanavermaete, bedanken voor al het werk dat hij gestoken heeft in mijn thesis. Zowel voor het grondig nalezen van mijn thesis als voor alle feedback en uitleg die hij gegeven heeft, ben ik hem zeer dankbaar. Ik kon hem op elk moment bereiken en ik kon steeds terecht bij hem met al mijn vragen waarvoor hij uitgebreid de tijd nam om deze te beantwoorden. Ik had mij geen betere tutor kunnen wensen. Daarnaast wil ik ook mijn promotor, prof. dr. Bernard De Baets, bedanken voor alle feedback die hij gegeven heeft en voor het grondig nalezen van mijn thesis. Ik zou ook graag dr. Steven Maenhout willen bedanken voor het interview dat ik met hem kon doen en voor de inzichten die hij mij heeft gegeven in verband met de praktische kant van het onderzoek. Ten slotte wil ik ook mijn vrienden en familie bedanken voor de steun die ze mij hebben gegeven.

INHOUDSOPGAVE

Woord vooraf i
Inhoudsopgave iii
Abstract Nederlands v
Abstract Engels vi
Lijst met afkortingen vii
Inleiding 1
1 Literatuuronderzoek 3
1.1 Een inleiding tot klassieke veredeling 3
1.2 Stamboomanalyse 4
1.3 Marker-assisted selection 5
1.4 Genomic selection 8
1.4.1 Genomic best linear unbiased prediction 9
1.4.2 Genomic estimated breeding values 9
1.4.3 Trainingspopulatie 10
1.4.4 Truncatieselectie 12
1.4.5 Scoping methode 15
1.4.6 Voordelen van genomic selection 16
1.5 Predictiemodellen 18
1.5.1 Brede waaier aan mogelijkheden 18
1.5.2 Predictieaccuraatheid van een model 18
1.6 Economisch aspect 20
2 Experimenteel werk 21
2.1 Materiaal en methoden 21
2.1.1 Algemene proefopzet 21
2.1.2 Datasets en veredelingsprogramma 22
2.1.3 Truncatieselectie 22
2.1.4 Scoping methode 23
2.1.5 Trainingspopulatie 23
2.1.6 Simulator 24

2.1.7 Parameterwaarden 24
2.1.8 Gemiddelde genetische waarde van de top-10 individuen 25
2.1.9 Aantal veredelingscycli 26
2.1.10 Predictiemodellen 27
2.1.11 Lineair regressiemodel 28
2.1.12 Statistische analyse 29
2.2 Resultaten en discussie 31
2.2.1 RR-GBLUP 31
2.2.2 Vergelijking tussen verschillende modellen 52
2.2.3 Praktisch aspect 54
2.3 Algemeen besluit 56
Bibliografie 58
Bijlagen 1–12 63

Abstract

Door de snel aangroeiende wereldpopulatie en de toenemende vraag naar voedsel die hiermee gepaard gaat, dienen opbrengsten van gewassen voortdurend verhoogd te worden. Door *genomic selection* (GS) konden deze reeds worden verhoogd, maar er is nog steeds nood aan nieuwe innovatieve methoden binnen GS om de opbrengsten verder te kunnen laten toenemen. Een dergelijke nieuwe methode binnen GS is de scoping methode. Deze ouderlijke selectiemethode tracht het behoud van genetische variatie te maximaliseren. Dit staat haaks op truncatieselectie, een vaak gebruikte selectiemethode die de genetische opbrengst op korte termijn maximaliseert, maar ook resulteert in het verlies van grote delen van de genetische variatie wat kan leiden tot vroegtijdige convergentie van de genetische waarde. Desondanks wordt truncatieselectie nog steeds gebruikt. De prestatie van een veredelingspopulatie wordt beïnvloed door verschillende parameters zoals de ouderlijke selectiegrootte en het aantal nakomelingen per ouderpaar. Deze parameters moeten zorgvuldig bestudeerd worden om een evenwicht te vinden tussen enerzijds de opbrengst en anderzijds de kost. Sommige parameters blijken een grote invloed te hebben (zoals de ouderlijke selectiegrootte) terwijl de invloed van andere parameters echter beperkt is (zoals die van het gebruikte model). Voor de verschillende parameterwaarden worden op lange termijn steeds (veel) hogere genetische waarden behaald voor de scoping methode dan voor truncatieselectie.

Trefwoorden

Genomic selection, plantenveredeling, scoping methode, truncatieselectie, RR-GBLUP, Bayesiaanse modellen

Abstract

The human population is constantly growing, increasing the global food demand. Therefore, the yield of major crops should be increased. Genomic selection (GS) has already increased the genetic progress in several crops, but new strategies are still required to maintain the genetic progress of a breeding population. The scoping method is a parental selection method that maximizes the preservation of genetic variation and thus maximizing the long-term gains. This is in contrast with truncation selection, a greedy selection method resulting in high short-term gains. Unfortunately, truncation selection is also associated with the loss of genetic variation, resulting in a premature convergence of the long-term gains. Nevertheless, truncation selection is still often used in breeding programs. The genetic progress of a breeding population is influenced by several parameters like the parental selection size and the number of individuals per cross. These parameters should be studied carefully to find an equilibrium between yield maximization and cost minimization. Certain parameters turned out to have a large effect on the genetic progress, whereas other parameters only had a small non-significant influence. Across the different parameter values that were tested, higher long-term genetic gains were obtained when the scoping method was used instead of truncation selection.

Keywords

Genomic selection, plant breeding, scoping method, truncation selection, RR-GBLUP, Bayesian models

Lijst met afkortingen

ANOVA	Analysis of variance
BLUP	Best linear unbiased prediction
BRR	Bayesian ridge regression
CD	Coefficient of determination
DNA	Deoxyribonucleic acid
EBV	Estimated breeding value
GBLUP	Genomic best linear unbiased prediction
GEBV	Genomic estimated breeding value
GS	Genomic selection
GV	Genetic value
LD	Linkage disequilibrium
MAS	Marker-assisted selection
OSG	Ouderlijke selectiegrootte
PEV	Prediction error variance
QQ-plot	Quantile-quantile plot
QTL	Quantitative trait locus
RR	Ridge regression
RR-GBLUP	Ridge regression genomic best linear unbiased prediction
SNP	Single nucleotide polymorphism
SR	Scoping rate
SSD	Single-seed descent

Inleiding

In dit onderzoek worden predictiemodellen gebruikt om genetische vooruitgang te bewerkstelligen. Dit gebeurt aan de hand van simulaties waarbij gebruikgemaakt wordt van datasets van gerst. Er wordt gewerkt met datasets van planten omdat deze enerzijds groter zijn en anderzijds ook ruimer beschikbaar zijn dan datasets van dieren. Door de sterk groeiende wereldpopulatie en de bijhorende sterke stijging van de vraag naar voedsel is genetische vooruitgang in de plantenteelt essentieel. Opbrengsten van gewassen dienen snel verhoogd te worden om steeds voldoende voedsel te kunnen blijven voorzien. Door *genomic selection* (GS) konden opbrengsten van gewassen reeds worden verhoogd, maar er is nog steeds nood aan nieuwe innovatieve methoden binnen GS om deze verder te kunnen blijven verhogen.

Veredelaars maken voor de ouderlijke selectie vaak gebruik van truncatieselectie. Hierbij worden steeds de beste individuen met elkaar gekruist met als doel het verkrijgen van een zo gunstig mogelijk nakomelingschap. De langetermijnopbrengsten zullen door gebruik van deze ouderlijke selectiemethode echter snel convergeren naar een lokaal optimum, wat er in de toekomst voor zal zorgen dat er onvoldoende voedsel kan worden voorzien. De scoping methode is een nieuwe ouderlijke selectiemethode die vooral op lange termijn de genetische waarde verhoogt in vergelijking met truncatieselectie. Omdat voor de scoping methode op korte termijn gelijkaardige genetische waarden worden bekomen als voor truncatieselectie, zijn veredelaars niet geneigd om snel over te stappen naar een nieuwe en onbekende methode. In deze thesis wordt de invloed van verschillende parameters voor beide ouderlijke selectiemethoden onderzocht. Hierdoor wordt de parametrische ruimte van de scoping methode beter in kaart gebracht wat aanleiding zou kunnen geven tot een verdere optimalisatie van de methode en een veiligere introductie van de methode in huidige veredelingsprogramma's. Er zal worden nagegaan in welke mate de scoping methode beter is dan truncatieselectie en onder welke omstandigheden het verschil tussen beide selectiemethoden het grootst is.

Er zal worden onderzocht wat de invloed is van de ouderlijke selectiegrootte op de genetische waarde die behaald wordt. Daarnaast zal ook worden gezocht naar de gunstigste ouderlijke selectiegrootte en dit zowel voor truncatieselectie als voor de scoping methode. Voor verschillende ouderlijke selectiegroottes wordt het verschil in genetische waarde tussen de scoping methode en truncatieselectie beschouwd. Er is hierbij ook aandacht voor hoe de verschillende ouderlijke selectiemethodes zich gedragen tegenover elkaar gedurende meerdere veredelingscycli om op die manier een beeld te krijgen van zowel de korte als de lange termijn.

Een tweede parameter waarvan de invloed op de genetische waarde zal worden onderzocht, is het aantal nakomelingen per ouderpaar. Er wordt gekeken naar hoeveel nakomelingen per ouderpaar optimaal zijn om de beste individuen te bekomen op enerzijds korte termijn en anderzijds lange termijn. Daarnaast wordt ook onderzocht voor welke waarde van deze parameter de balans tussen de korte en de lange termijn optimaal is.

Vervolgens wordt nagegaan of een verschillende genomische constructie invloed heeft op enerzijds de genetische waarde die behaald kan worden en anderzijds de performantie van de scoping methode. De invloed van het aantal *quantitative trait loci* (QTL's) wordt hiervoor bekeken alsook de invloed van de heritabiliteit.

Er zal ook worden onderzocht of het gebruikte predictiemodel een invloed heeft op de genetische waarden die worden behaald en op de performantie van de scoping methode. Hiervoor wordt het model *ridge regression genomic best linear unbiased prediction* (RR-GBLUP) in verschillende situaties vergeleken met vier Bayesiaanse modellen, meer bepaald BayesA, BayesB, BayesC en *Bayesian ridge regression* (BRR). Er wordt nagegaan welk model het best wordt gebruikt.

Ten slotte wordt ook eens gekeken naar hoe het er in de praktijk aan toegaat. Er wordt onder andere nagegaan welk model in de praktijk vooral wordt gebruikt, of de scoping methode al wordt gebruikt in de praktijk en hoe veredelaars handelen. Hiervoor wordt onder andere een interview gedaan met dr. Steven Maenhout, zaakvoerder van Progeno.

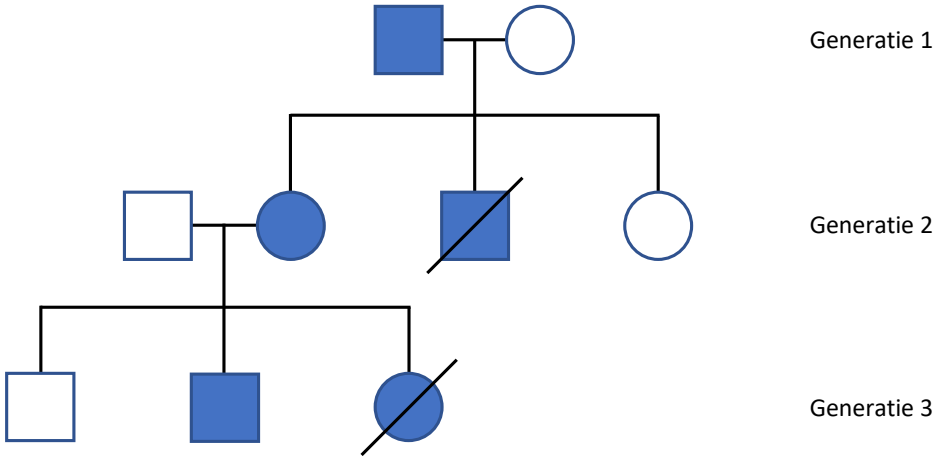
HOOFDSTUK 1

LITERATUURONDERZOEK

1.1 Een inleiding tot klassieke veredeling

Al sinds de prehistorie doet de mens aan landbouw. Tussen 10.000 en 5.000 jaar geleden deed landbouw zijn intrede in tal van regio's van waaruit landbouw zich steeds verder verspreidde over de menselijke samenleving (Weisdorf, 2005). Vóór de introductie van landbouw leidde de mens een nomadisch bestaan bestaande uit jagen en het verzamelen van voedsel uit zijn omgeving. De overgang van deze jacht- en verzamelcultuur naar een samenleving waarin aan landbouw wordt gedaan, wordt de neolithische revolutie genoemd (Weisdorf, 2005). Dit was één van de belangrijkste transities in de geschiedenis van de menselijke samenleving (Smith, 2015). In deze vroege transitie bestond landbouw enerzijds uit akkerbouw en anderzijds uit het houden van dieren (veeteelt). Tal van planten- en diersoorten werden vanaf dit moment gedomesticeerd en veredeld, wat inhoudt dat planten of dieren met gewenste eigenschappen worden geselecteerd en met elkaar worden gekruist met als doel het doorgeven van deze gewenste eigenschappen aan de nakomelingen. Deze selectie gebeurt op basis van het fenotype, dit is de verzameling van de fysische karakteristieken van het organisme. Enkele voorbeelden van dergelijke fysische karakteristieken zijn de lengte van een organisme, de massa, de kleur, de vruchtopbrengst, enzovoort.

Klassieke veredeling heeft echter wel tal van nadelen. Zo kost dit veel tijd en geld. Fenotypische selectie is slechts mogelijk nadat het individu volledig ontwikkeld is en de fysische eigenschappen zichtbaar worden. Wanneer bijvoorbeeld gewerkt wordt met aardbeiplanten, zal de teler de vruchtopbrengst pas kunnen meten tijdens het tweede levensjaar van de plant omdat deze planten dan pas vruchten dragen. Klassieke veredeling heeft ook als nadeel dat de omgevingsfactoren het fenotype van individuen kunnen beïnvloeden. Wanneer een plant bijvoorbeeld meer zonlicht krijgt dan een andere plant, kan dit een gunstiger meetresultaat opleveren van een welbepaalde eigenschap ten gevolge van dit verschil in hoeveelheid zonlicht en dus niet ten gevolge van de aanwezigheid van gunstige genen. Hier wordt later nog meer in detail op ingegaan (zie Sectie 1.3). Er wordt binnen deze studie gefocust op planten, werken met dieren is minder geschikt voor dit onderzoek. De reden hiervoor is dat



Figuur 1.1: Een voorbeeld van een stamboom, de lijnen geven verwantschappen weer, de stamboom bestaat uit drie generaties. Vierkant: mannelijk individu, cirkel: vrouwelijk individu, ingekleurd: drager van de welbepaalde ziekte of eigenschap, doorstreept: gestorven (Hart, 2001).

datasets van planten groter zijn en ook ruimer beschikbaar. Data van planten kunnen namelijk eenvoudiger worden bekomen dan data van dieren. Het is beter om over een dataset te kunnen beschikken die meer data bevat om de resultaten die verkregen zullen worden, correct te kunnen interpreteren.

Klassieke veredeling is dus zeker niet zo ideaal als op het eerste zicht zou lijken, er dient dus gezocht te worden naar een beter alternatief.

1.2 Stamboomanalyse

Eén van de mogelijke verbeteringen van klassieke veredeling is het gebruik van stambomen (het gebruik van zogenaamde *pedigrees*). Een stamboom is een genetische voorstelling van een bepaalde populatie waarvan de individuen tot dezelfde familie behoren; de overerving van een bepaalde eigenschap of ziekte over de verschillende generaties heen wordt hierin weergegeven (Hart, 2001). Een stamboom geeft weer in welke mate twee individuen met elkaar verwant zijn en toont welke individuen een welbepaalde eigenschap hebben of er drager van zijn (Hart, 2001). Een voorbeeld van een dergelijke stamboom is weergegeven in Figuur 1.1.

Deze stamboom bestaat uit drie generaties. De verschillende generaties zijn rechts van de figuur weergegeven. De lijnen die cirkels en vierkanten verbinden, geven de verwantschap weer tussen de verschillende individuen. Wanneer twee individuen horizontaal met elkaar zijn verbonden door middel van een lijn, betekent dit dat deze

twee individuen werden gekruist. De lijn die uit deze verbindingslijn vertrekt, splitst al dan niet op en geeft weer welke individuen zijn voortgekomen uit deze kruising. In deze stamboom staat een vierkant voor een mannelijk individu en een cirkel voor een vrouwelijk individu. Wanneer dit vierkant of deze cirkel is ingekleurd, is het desbetreffende individu drager van de beschouwde eigenschap of ziekte en wanneer deze doorstreept is, is het individu gestorven.

Aan de hand van deze stamboom kan de kans berekend worden dat een bepaald individu een eigenschap (of een ziekte) zal hebben of de kans dat dit individu deze eigenschap of ziekte zal doorgeven aan het nakomelingschap (Hart, 2001). Stamboomanalyse wordt vaak gebruikt wanneer slechts een beperkt aantal individuen per generatie beschikbaar is of voor het bestuderen van een populatie met een lange generatietijd (McClean, 2000).

1.3 Marker-assisted selection

Marker-assisted selection (MAS) is een selectietechniek die is kunnen ontstaan uit de snelle vooruitgang binnen de moleculaire genetica (Lande and Thompson, 1990). Bij MAS is de selectie van de individuen gebaseerd op zogenaamde merkers. Een merker is een DNA-fractie die informatie bezit over een nabijgelegen gen. De merkers die interessant zijn voor een bepaalde toepassing zijn deze die zeer dicht gelegen zijn tegen genen die bijdragen aan de expressie van het kenmerk waarvoor interesse is. Deze genen worden *quantitative trait loci* (QTL's) genoemd.

De expressie van een welbepaald kenmerk is afhankelijk van de bijdrage van honderden genen (of QTL's). Deze genen worden verondersteld een additief effect te hebben. Binnen dit onderzoek worden dominantie en epistasie (*epistasis*) verondersteld afwezig te zijn. Met een additief effect wordt bedoeld dat verschillende genen onafhankelijk van elkaar en van de allelsamenstelling bijdragen aan een kenmerk. Dergelijke kenmerken of eigenschappen worden kwantitatief genoemd. Kwantitatieve kenmerken worden beïnvloed door meerdere en vaak honderden genen maar ook omgevingsfactoren kunnen deze eigenschap beïnvloeden. Wanneer de omgeving een belangrijke invloed heeft op de expressie van het kenmerk, kan eenzelfde genotype resulteren in zeer veel verschillende fenotypes. Wanneer de variatie echter vooral een gevolg is van de invloed van genen, is er een sterkere correlatie tussen het genotype en het fenotype. De mate waarin een eigenschap beïnvloed wordt door zijn omgeving, wordt gedefinieerd door de heritabiliteit (zie verder).

De fenotypische eigenschappen van planten zijn vaak kwantitatieve eigenschappen. Er wordt getracht de effecten van alle QTL's die een rol spelen in de expressie van

deze eigenschap zo correct mogelijk in te schatten. Deze eigenschap kan dan zodoende gekwantificeerd worden door de som te nemen van de aanwezige QTL effecten. Deze som van de aanwezige QTL effecten zal resulteren in de genetische waarde.

Als een merker zeer dicht gelegen is tegen een QTL, is de kans zeer klein dat deze niet samen zullen worden overgeërfd. Doordat deze namelijk op eenzelfde chromosoom vlakbij elkaar zijn gelegen, is de kans dat er een crossing-over plaatsvindt precies in het gebied tussen de merker en de QTL, zeer klein. Wanneer er sprake is van een dergelijke kleine afstand tussen merker en QTL, wordt gezegd dat de merker in *linkage disequilibrium* (LD) is met de QTL. Aangezien de QTL effecten en de QTL locaties vaak niet gekend zijn, worden merkers gebruikt die in LD zijn met één of meerdere QTL's om de effecten van die QTL's op de genetische waarde te kunnen voorspellen. De merkers die worden gebruikt, zijn namelijk wel eenvoudig te identificeren.

Als merkers worden vaak *single nucleotide polymorphisms* (SNP's) gebruikt. Een SNP is een polymorfisme waarbij slechts één nucleotide is gewijzigd. Er wordt gebruikgemaakt van bi-allelische SNP-merkers die aangeven welk allel aanwezig is. Hierdoor kan het DNA-basepaar eenvoudig worden gerepresenteerd door een getal, meer bepaald het getal 1, 0 of -1. Hierbij betekent een omzetting naar 1 dat het eerste allel aanwezig is op beide chromosomen, -1 betekent dat het tweede allel op beide chromosomen aanwezig is en 0 betekent dat het eerste allel aanwezig is op één van beide chromosomen en het tweede allel aanwezig is op het andere chromosoom. Het maakt niet uit welk allel wordt gekozen als referentie, deze keuze heeft namelijk geen invloed op de grootte van het geschatte merkereffect, enkel op het teken ervan (Martini et al., 2017). Het is ook mogelijk om andere coderingen te gebruiken voor de SNP-merkers, hiervoor hoeven niet noodzakelijk de getallen 1, 0 en -1 gebruikt te worden. Omdat in dit onderzoek wordt verondersteld dat zowel dominantie als epistasie afwezig is, kunnen bijvoorbeeld ook de getallen 0, 1 en 2 gebruikt worden voor de codering van de SNP-merkers zonder dat dit een invloed heeft op het resultaat (Martini et al., 2019).

Zoals eerder al vermeld, wordt een eigenschap van een individu bepaald door enerzijds het genotype en anderzijds de omgeving. Deze eigenschap wordt de fenotypische waarde genoemd. Het fenotype van een individu kan dus beschouwd worden als het resultaat van de wisselwerking tussen het genotype en de omgeving waarin het individu voorkomt. De fenotypische waarde is datgene wat uiteindelijk kan worden waargenomen of datgene wat uiteindelijk kan worden bepaald door een meting, bijvoorbeeld de graanopbrengst. Dit kan worden weergegeven aan de hand van de volgende formule:

$$\sigma_t^2 = \sigma_g^2 + \sigma_e^2. \tag{1.1}$$

Hierbij staat σ_t^2 voor de totale fenotypische variantie, σ_g^2 voor de genotypische variantie en σ_e^2 voor de omgevingsvariantie of dus de variantie die toe te schrijven is aan de omgeving waarin het individu leeft (Audenaert, 2018). Er kan ook sprake zijn van interactie tussen het genotype van het organisme en het milieu waarin het organisme voorkomt, wat betekent dat het effect van het milieu verschilt naargelang het genotype (Audenaert, 2018). In de formule dient dan een extra factor in rekening gebracht te worden:

$$\sigma_t^2 = \sigma_g^2 + \sigma_e^2 + \sigma_{ge}^2, \tag{1.2}$$

met σ_{ge}^2 een factor die corrigeert voor de interactie tussen het genotype en de omgeving.

De heritabiliteit geeft aan hoe sterk een eigenschap bepaald wordt door het genotype en dus ook hoe sterk de omgeving deze eigenschap beïnvloedt. De heritabiliteit neemt een waarde aan tussen 0 en 1. Wanneer de heritabiliteit gelijk is aan 1, wil dit zeggen dat de eigenschap volledig bepaald wordt door het genotype en de omgeving dus geen invloed heeft op de eigenschap. Wanneer de heritabiliteit daarentegen 0 bedraagt, wil dit zeggen dat de eigenschap volledig bepaald wordt door de omgeving en het genotype geen invloed heeft op de eigenschap. Een hogere heritabiliteit zal er dus voor zorgen dat het fenotype beter voorspeld kan worden op basis van de aanwezige genen.

Er worden twee soorten heritabiliteiten onderscheiden: enerzijds de heritabiliteit in brede zin en anderzijds de heritabiliteit in enge zin. De heritabiliteit in brede zin (H^2) is de verhouding van de genotypische variantie tot de totale fenotypische variantie:

$$H^2 = \frac{\sigma_g^2}{\sigma_t^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2 + \sigma_{ge}^2}. \tag{1.3}$$

De heritabiliteit in enge zin (h^2) daarentegen, ook wel de heritabiliteit in nauwe zin genoemd, is de verhouding van de variantie ten gevolge van de additiviteit tot de totale fenotypische variantie (Audenaert, 2018). Dit kan ook als volgt worden voorgesteld:

$$h^2 = \frac{\sigma_A^2}{\sigma_t^2}, \tag{1.4}$$

met σ_A^2 de variantie die toe te schrijven is aan de additiviteit. Het is namelijk interessant om de additiviteit te beschouwen aangezien dit de enige constante is over de verschillende generaties heen. Op die manier is de additiviteit soms een betere maatstaf (Audenaert, 2018).

Het is aangetoond dat MAS een grotere selectiewinst oplevert in vergelijking met selectie op basis van het fenotype voor normaal verdeelde kwantitatieve kenmerken

(Knapp, 1998). MAS kan doeltreffender zijn dan selectie op basis van het fenotype wanneer er sprake is van vrij grote populaties en wanneer er sprake is van eigenschappen met een relatief lage heritabiliteit. MAS kan echter minder doeltreffend zijn op lange termijn in vergelijking met selectie op basis van het fenotype (Hospital et al., 1997). De ouderplanten kunnen bij MAS in een vroeger stadium geselecteerd worden in vergelijking met selectie op basis van het fenotype. Hierdoor zal de kost per veredelingscyclus dalen en ook de tijd die gemiddeld kruipt in één veredelingscyclus zal hierdoor afnemen. De selectie bij MAS gebeurt dus aan de hand van merkers. Uit deze merkers wordt informatie verkregen over de verschillende QTL's die bijdragen tot een welbepaald kwantitatief kenmerk. Er is een lineair verband tussen de merkers en het fenotype; dit lineaire verband kan worden ingeschat door gebruik te maken van welbepaalde modellen, meer bepaald lineaire gemengde modellen (Meuwissen et al., 2001; Henderson, 1984). *Best linear unbiased prediction* (BLUP) is een algoritme dat helpt bij het bouwen van een dergelijk lineair gemengd model. BLUP zal, op basis van genotype en fenotype, de *estimated breeding values* (EBV's) zo goed mogelijk proberen te voorspellen (Meuwissen et al., 2016).

MAS bestaat uit twee stappen: ten eerste het detecteren en het in kaart brengen van die genen die bijdragen aan het welbepaalde kenmerk van interesse (QTL's) en ten tweede de informatie die verkregen wordt uit deze QTL's, introduceren in een BLUP-EBV (Meuwissen et al., 2016). Een groot nadeel van deze selectietechniek is echter dat de herhaalbaarheid van het in kaart brengen van deze QTL's laag is. Een oorzaak hiervan is dat de meerderheid van de QTL's slechts een zeer klein effect heeft op het kwantitatieve kenmerk en wanneer er dan zeer grote hoeveelheden merkers getest worden, is er sprake van een zogenaamd Beavis effect waarbij het geschatte effect van gewichtige merkers wordt overschat (Meuwissen et al., 2016).

1.4 Genomic selection

Het gebruik van merkers voor het voorspellen van genotypische waarden was een revolutionaire methode die de veredelingstijd en kost kon verlagen terwijl de genetische winst gemaximaliseerd kon worden. Toch kon er via MAS geen hoge genetische winst behaald worden (Jannink et al., 2010; Bernardo, 2008; Xu and Crouch, 2008). Daarom werd door Meuwissen et al. (2001) *genomic selection* (GS) als een alternatieve methode voorgesteld. In GS wordt een groter aantal merkers gebruikt die uniform verspreid zijn over het genoom. Hierdoor zal niet elke merker in LD zijn met één of meerdere QTL's. Bij MAS wordt enkel een kleine selectie van merkers gebruikt die volgens voorspellingen in LD zouden zijn met één of meerdere QTL's. Bij GS daarentegen zullen de merkers die niet in LD zijn ook nog steeds worden gebruikt;

deze helpen de algemene variatie vast te leggen. Dit zorgt ervoor dat GS een betere selectiemethode is dan MAS.

1.4.1 Genomic best linear unbiased prediction

Het meest populaire algoritme binnen GS is *genomic best linear unbiased prediction* (GBLUP). Dit algoritme is zeer gelijkaardig aan BLUP al is er wel één groot verschil tussen beide algoritmes: daar waar bij BLUP gebruikgemaakt wordt van onderlinge relaties op basis van verwantschap, wordt bij GBLUP gebruikgemaakt van onderlinge relaties op basis van het genotype (Meuwissen et al., 2016). Deze onderlinge relatie tussen twee individuen op basis van het genotype, wordt berekend als de Pearson correlatie tussen de variabele die alle SNP's bevat van het ene individu (het genotype van dit individu wordt voorgesteld als X_{j1}) en de variabele die alle SNP's bevat van het andere individu (het genotype van dit individu wordt voorgesteld als X_{j2}) (Meuwissen et al., 2016). Hierbij stelt j de index voor van de merker en de cijfers 1 en 2 slaan op de individuen voor welke de onderlinge relatie op basis van het genotype bepaald wordt. Omdat verbanden op basis van het genotype accurater zijn dan verbanden die gebaseerd zijn op verwantschap, zal GBLUP een hogere predictieaccuraatheid hebben dan BLUP (Meuwissen et al., 2016).

Het GBLUP algoritme helpt bij het bouwen van lineaire gemengde modellen van de vorm:

$$\mathbf{y} = \mathbf{1}_n\beta + \mathbf{Zu} + \boldsymbol{\epsilon}\,, \tag{1.5}$$

met $\mathbf{y}$ een $[n \times 1]$ vector met de fenotypische waarden van n verschillende individuen, $\mathbf{1}_n$ een vector van lengte n met waarde 1, β de gemiddelde fenotypische waarde binnen de populatie, $\mathbf{Z}$ een $[n \times m]$ matrix die de merkerinformatie bevat (gecodeerd als $-1, 0, 1$) van n verschillende individuen over m verschillende merkers, $\mathbf{u}$ een $[m \times 1]$ vector met de additieve merkereffecten en $\boldsymbol{\epsilon}$ een $[n \times 1]$ vector met de residuele fouten.

Een dergelijk lineair gemengd model is een predictiemodel. In dit onderzoek zullen tal van predictiemodellen worden gebruikt. Deze modellen worden later meer in detail besproken (zie Subsectie 2.1.10).

1.4.2 Genomic estimated breeding values

Bij GS zal het effect van de merkers eerst geschat worden in een trainingspopulatie en vervolgens zal dit worden gebruikt om de genetische waarde van de populatie te

voorspellen (Sonesson and Meuwissen, 2009). Deze voorspelde genetische waarden worden *genomic estimated breeding values* (GEBV's) genoemd.

De GEBV is de som van de effecten van de verschillende merkers. Het berekenen van de GEBV van welbepaalde individuen kan worden gedaan aan de hand van de volgende formule:

$$\mathbf{GEBV} = \mathbf{Z}\hat{\mathbf{u}}, \tag{1.6}$$

met **GEBV** een $[n \times 1]$ vector die de *genomic estimated breeding value* voorstelt van n verschillende organismen en $\hat{\mathbf{u}}$ een $[m \times 1]$ vector die de geschatte merkereffecten voorstelt.

De genetische waarde kan op een gelijkaardige manier worden bekomen als de GEBV. De genetische waarde van welbepaalde organismen kan namelijk worden bekomen aan de hand van de volgende formule:

$$\mathbf{GV} = \mathbf{Z}\mathbf{u}, \tag{1.7}$$

met **GV** (*genetic value*) een $[n \times 1]$ vector met de genetische waarden van n verschillende individuen.

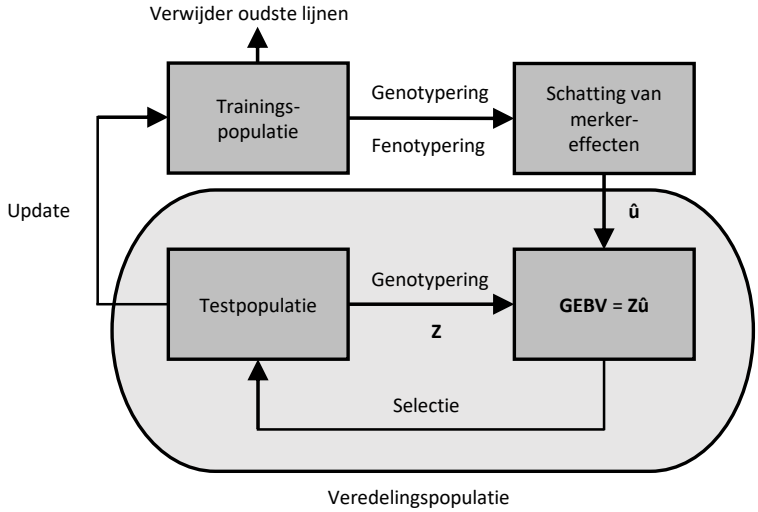
De GEBV is een voorspelling van de GV. Daar waar bij de berekening van de GEBV's gebruikgemaakt wordt van een schatting van de merkereffecten, wordt voor de berekening van de GV's gebruikgemaakt van de effectieve merkereffecten. Hoe kleiner het verschil tussen de GEBV's en de GV's, hoe groter de predictieaccuraatheid van het desbetreffende model. Deze predictieaccuraatheid wordt later meer in detail behandeld.

Een schematisch overzicht van de verschillende stappen waaruit GS bestaat, is weergegeven in Figuur 1.2.

1.4.3 Trainingspopulatie

Bij GS worden merkereffecten ingeschat met behulp van een trainingspopulatie die bestaat uit individuen die zowel zijn gegenotypeerd als gefenotypeerd. Vervolgens kunnen de geschatte merkereffecten gebruikt worden om de genetische waarde van de populatie te schatten. Er is dan enkel genotypische informatie van de populatie nodig (Meuwissen et al., 2016).

De informatie die verkregen wordt uit de trainingspopulatie dient voortdurend te worden geüpdatet. In het onderzoek van Neyhart et al. (2017) werd nagegaan wat de beste methode is om een trainingspopulatie te updaten. Het updaten van een trai-



Figuur 1.2: Schema dat de verschillende stappen weergeeft van GS. De trainingspopulatie zal zowel gegenotypeerd als gefenotypeerd worden, aan de hand hiervan zullen de merkereffecten worden geschat. De testpopulatie zal enkel worden gegenotypeerd. Vervolgens wordt de **GEBV** berekend van de individuen die behoren tot de testpopulatie. Dit zal gebeuren op basis van enerzijds de merkereffecten die werden geschat ($\hat{\mathbf{u}}$) en anderzijds het genotype van de individuen zelf ($\mathbf{Z}$). Het is verder ook belangrijk dat de trainingspopulatie voortdurend wordt geüpdatet, er worden individuen toegevoegd vanuit de testpopulatie die worden geselecteerd op basis van de **GEBV**. Uit de trainingspopulatie zullen de oudste lijnen worden verwijderd. De populatie die gebruikt wordt voor de veredeling zelf wordt weergegeven in een licht grijs ovaal (Robertsen et al., 2019).

ningspopulatie is namelijk noodzakelijk aangezien, als gevolg van recombinatie, LD's tussen QTL's en merkers vervallen. Wanneer geen nieuwe data worden toegevoegd aan de trainingspopulatie, zal geen rekening gehouden worden met deze verandering in LD's, wat resulteert in een afname van de accuraatheid van de voorspelling (Neyhart et al., 2017). Dit laatste dient uiteraard vermeden te worden. Het updaten van de trainingspopulatie kan op tal van manieren gebeuren. Neyhart et al. (2017) vergeleken zeven verschillende methoden voor het updaten van de trainingspopulatie. Deze zeven methoden zijn de volgende: *Top*, *Bottom*, *Random*, *PEVmean*, *CDmean*, *Tails* en *No Change*. De methode *Top* houdt in dat de selectiekandidaten met de hoogste GEBV's worden toegevoegd aan de trainingspopulatie. De methode *Bottom* doet net hetzelfde, maar dan met de selectiekandidaten die de laagste GEBV's hebben. Bij de methode *Random* wordt een willekeurige selectie van individuen gemaakt. *PEVmean* is een methode die gebaseerd is op het minimaliseren van de gemiddelde *prediction error variance* (PEV) (Rincent et al., 2012; Neyhart et al., 2017). Deze PEV geeft een beeld van hoe groot de fout op de voorspelling mag verwacht worden. CDmean is een methode die gebaseerd is op het maximaliseren van de verwachte betrouwbaarheid van de voorspellingen (*coefficient of determination*, CD) (Rincent et al., 2012; Neyhart et al., 2017). De methode *Tails* zal een gelijk aantal individuen selecteren uit beide staarten van de normaal verdeelde populatie; dit zijn enerzijds de individuen met de hoogste GEBV's en anderzijds de individuen met de laagste GEBV's. Ten slotte zal bij de methode *No Change* de trainingspopulatie niet worden geüpdatet. Van al deze mogelijke manieren voor het updaten van de trainingspopulatie, lijken de methoden *Top* en *Tails* iets interessanter om te gebruiken dan de andere methoden. De methode *Top* heeft namelijk een hoge predictieaccuraatheid en ook de eenvoud en de gebruiksvriendelijkheid van deze methode zijn een voordeel (Neyhart et al., 2017). Voor de methode *Tails* werden de hoogste genetische waarden bekomen (Neyhart et al., 2017; Vanavermaete et al., 2020). Er kon wel geen significant verschil vastgesteld worden tussen de verschillende voorgestelde updatemethoden. Er is echter wel een groot verschil tussen deze updatemethoden en de methode *No Change*. Het belangrijkste is dus dat de trainingspopulatie in het algemeen dient geüpdatet te worden (Neyhart et al., 2017).

1.4.4 Truncatieselectie

De methode waarbij steeds de beste individuen worden geselecteerd om verder mee te kruisen, wordt gulzige selectie (*greedy selection*) genoemd. Indien deze selectiemethode gecombineerd wordt met het willekeurig koppelen van geselecteerde ouders, wordt gesproken over truncatieselectie (*truncation selection*). Truncatieselectie lijkt een eenvoudige manier om welbepaalde eigenschappen van de individuen bin-

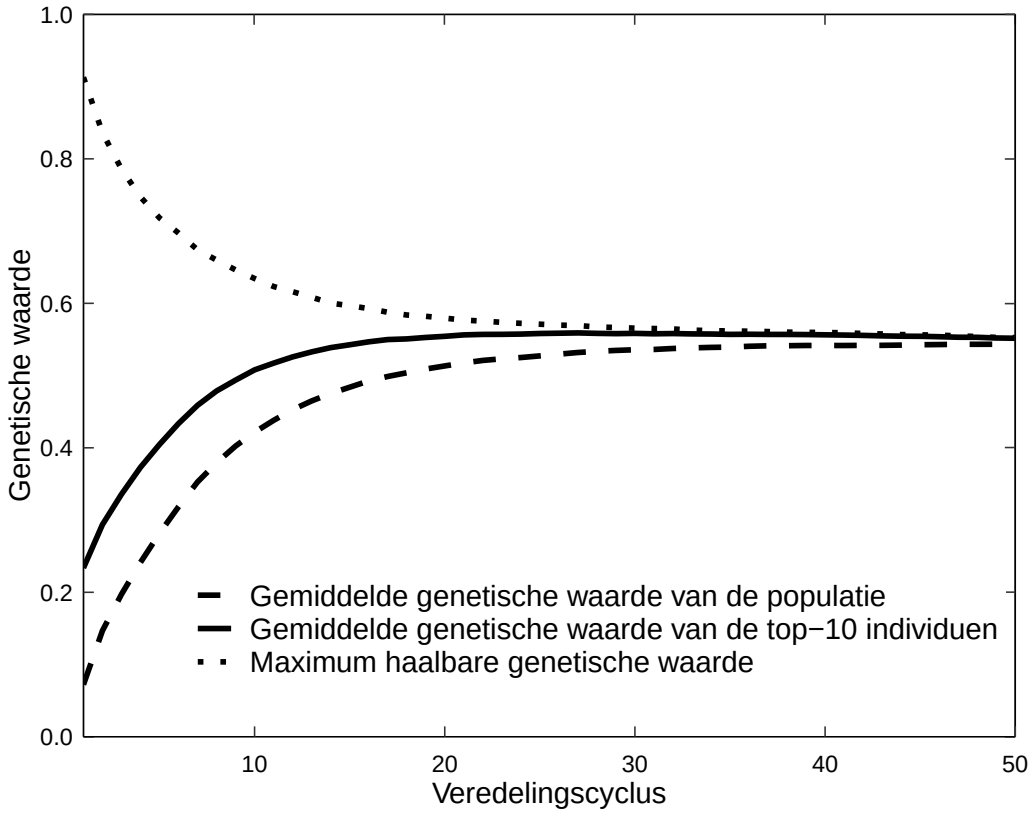
nen een populatie snel te verbeteren. Deze selectiemethode is echter niet zo goed als op het eerste zicht lijkt, er zijn tal van nadelen aan deze methode die voornamelijk betrekking hebben op de lange termijn. Door truncatieselectie toe te passen, daalt de genetische variatie in de populatie. Doordat de beste individuen met elkaar worden gekruist, zullen namelijk allelen sneller fixeren wat ervoor zorgt dat tal van allelen verloren gaan. Eens een welbepaald allel tijdens de selectie verdwijnt uit de populatie, kan dit allel enkel terug in de populatie komen door wilde planten te herintroduceren in deze populatie.

Een groot deel van de gefixeerde allelen zal een negatieve impact hebben op de genetische waarde. Dat bij truncatieselectie de resultaten op lange termijn slecht zijn, wordt geïllustreerd in Figuur 1.3.

Het fixeren van QTL allelen die een negatief effect hebben op de genetische waarde, en in het algemeen het verloren gaan van genetische variatie, zorgt ervoor dat er zo goed als geen vooruitgang meer kan worden geboekt op lange termijn. Daarnaast is het behouden van genetische variatie ook essentieel omdat een populatie hierdoor beter in staat is te reageren op veranderingen in de leefomgeving. Wanneer bijvoorbeeld een bepaalde ziekte uitbreekt onder een populatie planten of wanneer het leefmilieu waarin deze planten leven plots drastisch verandert (bijvoorbeeld extreme droogte), is er een hogere kans dat al deze planten op eenzelfde manier zullen reageren doordat ze genetisch niet zoveel van elkaar verschillen. Allelen die dan bijvoorbeeld resistentie tegen deze ziekte beïnvloeden of bepaalde allelen die de planten in staat zouden stellen te overleven in deze barre leefomgeving, kunnen dan tijdens de selectie uit de populatie verdwenen zijn. Dit kan ervoor zorgen dat de planten minder goed presteren (Markgraf, 2018; Lacy, 1997). Om de maximum haalbare genetische waarde te verhogen (en dus om betere genetische waarden te kunnen bekomen op lange termijn), kunnen allelen worden geherintroduceerd in de populatie. Hier kan het dan een goed idee zijn om wilde genen (dit zijn genen die afkomstig zijn van wilde planten) te introduceren in de plantenpopulatie. Studies toonden al aan dat er een duidelijk positief effect is wanneer genen van buiten de populatie worden geïntroduceerd in de populatie (Allier et al., 2020).

Nog een nadeel van truncatieselectie is dat er meer negatieve mutaties kunnen optreden doordat de individuen onderling steeds sterkere genetische gelijkenissen vertonen en er op die manier sprake is van een afname van genetische diversiteit (Miller, 2015).

Truncatieselectie lijkt dus niet de meest geschikte selectiemethode te zijn. Bij deze selectiemethode zullen niet enkel gunstige allelen snel gefixeerd worden, maar ook ongunstige allelen zullen snel worden gefixeerd. Hierdoor zal de maximum haalbare



Figuur 1.3: Deze grafiek geeft een simulatie weer van de genetische waarden wanneer truncatieselectie gebruikt wordt voor 50 opeenvolgende veredelingscycli. De x-as geeft de veredelingscycli weer en de y-as de genetische waarde. Er is een snelle toename zichtbaar van de gemiddelde genetische waarde van de gehele populatie op korte termijn (gedurende de eerste veredelingscycli). Op lange termijn daarentegen is te zien dat er nog amper betere genetische waarden worden behaald. Dit is een gevolg van het verlies van tal van QTL allelen door de daling van de genetische variatie binnen de populatie. Het valt verder op dat de gemiddelde genetische waarde van de tien beste individuen van de populatie na verloop van tijd zeer dicht komt te liggen bij de gemiddelde genetische waarde van de gehele populatie. Het verschil tussen de beste individuen van de populatie en de minder goede individuen wordt dus steeds kleiner. Ook dit is toe te schrijven aan de snelle afname van de genetische variatie. Ten slotte zal de maximum haalbare genetische waarde, dit is de hoogste genetische waarde die nog kan worden bereikt door de populatie, quasi volledig gelijk zijn aan de gemiddelde genetische waarde van de gehele populatie na het voltooien van 50 veredelingscycli. Dit wijst er dus op dat er al snel bijna geen ruimte meer zal zijn voor verbetering van de genetische waarde binnen de populatie (Vanavermaete et al., 2020).

genetische waarde, dit is de hoogste genetische waarde die nog kan worden bereikt door de populatie, snel dalen.

1.4.5 Scoping methode

Om te voorkomen dat de genetische waarden convergeren op lange termijn, kan een andere selectieprocedure gehanteerd worden. In plaats van die ouders te selecteren met de hoogste GEBV's, wordt een ouder met een hoge GEBV gecombineerd met een ouder op zo een manier dat de genetische variatie van deze twee geselecteerde ouders gemaximaliseerd wordt over elke merker (Vanavermaete et al., 2020). Deze manier van selectie van ouderlijke individuen wordt de *scoping* methode genoemd. Het doel van deze nieuwe methode is voorkomen dat de genetische waarde van de populatie convergeert naar een lokaal optimum ten gevolge van het verlies van gunstige QTL allelen of dus de fixatie van ongunstige QTL allelen. Door truncatieselectie wordt de kans kleiner dat het globale optimum op lange termijn kan bereikt worden. De scoping methode zorgt ervoor dat, door het behouden van de genetische variatie, hogere genetische waarden worden bekomen op lange termijn terwijl op korte termijn maar iets lagere genetische waarden worden gehaald door het gebruik van de scoping methode in vergelijking met truncatieselectie (Vanavermaete et al., 2020).

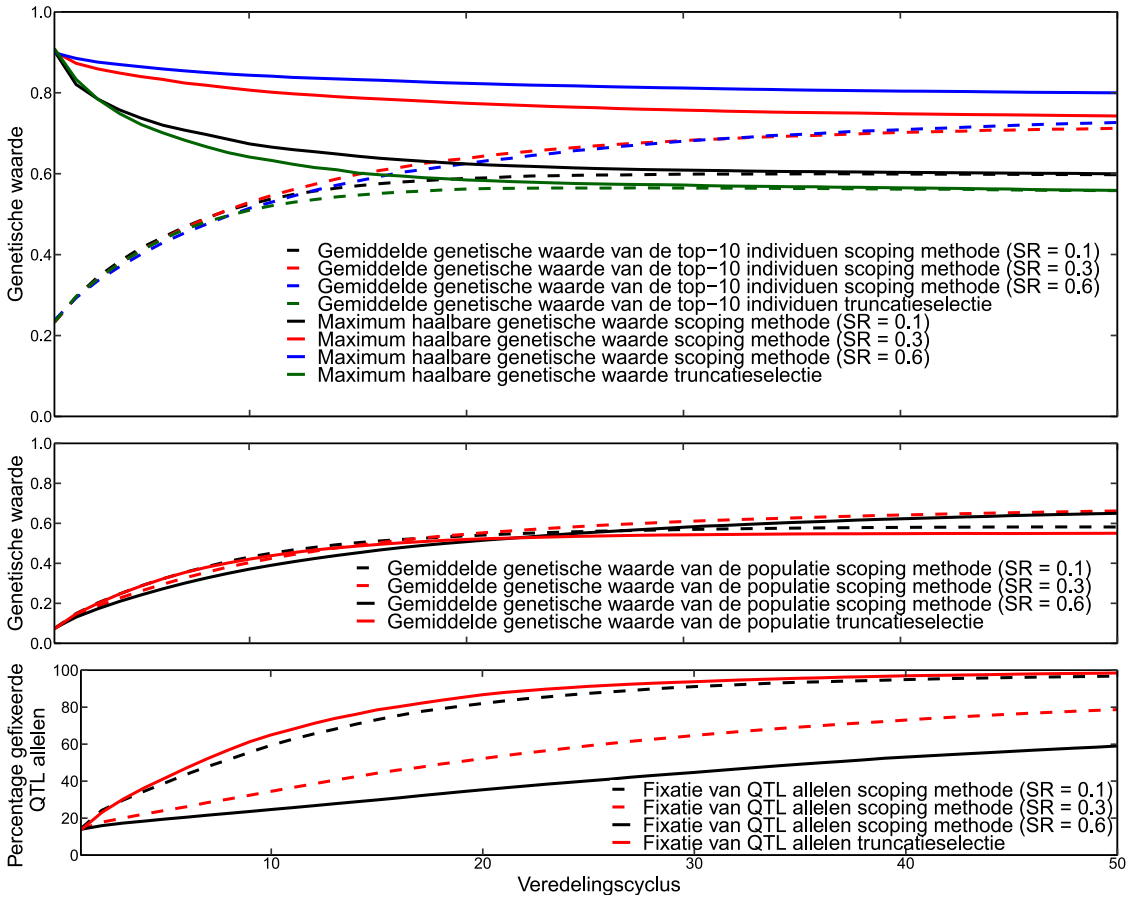
In Figuur 1.4 worden de scoping methode en truncatieselectie met elkaar vergeleken wat betreft de genetische waarde van de beste tien individuen van de populatie, de maximum haalbare genetische waarde, de gemiddelde genetische waarde van de gehele populatie en het percentage gefixeerde QTL allelen. De evolutie van elk van deze genoemde kenmerken zal worden beschouwd over 50 veredelingscycli. Voor de scoping methode worden drie curves weergegeven, elk met een andere zogenaamde *scoping rate* (SR). De SR is de fractie van de populatie die wordt voorgeselecteerd door middel van truncatieselectie, deze voorgeselecteerde individuen zijn dus die individuen met de hoogste GEBV's binnen de populatie. De SR kan een waarde aannemen tussen 0,1 en 1 (Vanavermaete et al., 2020). De drie waarden die de SR hier aanneemt, zijn 0,1, 0,3 en 0,6. Op de eerste grafiek van deze figuur is te zien dat de gemiddelde genetische waarde van de tien beste individuen van de populatie, na het doorlopen van 50 veredelingscycli, duidelijk het laagst is voor truncatieselectie. Voor de scoping methode met een SR van 0,1 is deze gemiddelde genetische waarde iets hoger dan voor truncatieselectie. Voor de scoping methode met een SR van 0,3 en 0,6 wordt echter wel een gemiddelde genetische waarde behaald na 50 veredelingscycli die veel hoger is dan deze van truncatieselectie. Op deze grafiek is eenzelfde trend zichtbaar voor de maximum haalbare genetische waarde, dit is de hoogste genetische waarde die nog bereikt kan worden door de populatie. Deze waarde is ook

hier het laagst voor truncatieselectie gevolgd door de scoping methode met een SR van 0,1. Voor de scoping methode met een SR van 0,3 wordt een hogere maximum haalbare genetische waarde bereikt en voor een SR van 0,6 is deze waarde nog hoger. Op de tweede grafiek van deze figuur zijn opnieuw dezelfde vier curves weergegeven maar nu voor de gemiddelde genetische waarde van de gehele populatie. Bij de scoping methode worden ook hier hogere waarden gehaald na 50 veredelingscycli dan bij truncatieselectie, al zijn de verschillen tussen de curves onderling nu wel kleiner in vergelijking met de vorige grafiek. De laatste grafiek geeft het percentage gefixeerde QTL allelen weer voor beide methoden. Bij truncatieselectie is het percentage gefixeerde QTL allelen steeds hoger dan bij de scoping methode. Er is echter wel niet zo een groot verschil tussen het percentage gefixeerde QTL allelen bij truncatieselectie en bij de scoping methode met een SR van 0,1. Dit percentage is wel veel minder hoog wanneer gebruikgemaakt wordt van de scoping methode met een SR van 0,3 of 0,6. Er worden dus steeds betere resultaten behaald bij de scoping methode dan bij truncatieselectie. Voor de scoping methode worden in het algemeen betere resultaten behaald bij een hogere waarde van de SR.

De scoping methode zal een zeer belangrijke rol spelen in dit onderzoek.

1.4.6 Voordelen van genomic selection

GS heeft tal van voordelen. Zo is GS een geschikte selectietechniek wanneer de fenotypering van de individuen zeer tijdrovend is of wanneer het duur is om de organismen te fenotyperen. Daarnaast dient er bij GS ook geen rekening gehouden te worden met omgevingsfouten. Met omgevingsfouten worden invloedsfactoren van de omgeving bedoeld die niet in rekening worden gebracht. Een voorbeeld hiervan is een plant die kleiner is (en dus doorgaans een minder gunstig fenotype heeft) omdat deze tijdens zijn ontwikkeling werd overschaduwd en zodoende dus niet dezelfde gunstige omgeving kende als andere planten die tot dezelfde populatie behoren. Of bijvoorbeeld wanneer een bepaalde plant grotere vruchten heeft dan een andere plant uit dezelfde populatie, maar deze plant wel voorzien werd van extra voedingsstoffen. GS is in het bijzonder zeer geschikt wanneer er gewerkt wordt met organismen die een lange generatietijd hebben, voor kenmerken die een lage heritabiliteit vertonen en wanneer selectie gebeurt op basis van een kenmerk dat pas laat in de ontwikkeling van het organisme tot expressie komt (Harfouche et al., 2012). Een voorbeeld van een organisme waarvoor bepaalde kenmerken pas laat in de ontwikkeling tot expressie komen, is de oliepalm. Deze planten kunnen nakomelingen geven vanaf een leeftijd van drie jaar, maar bepaalde fenotypische kenmerken komen pas tot expressie na 13 tot 15 jaar. Door gebruik te maken van GS in plaats van selectie op basis van fenotype zal de



Figuur 1.4: De grafieken in deze figuur vergelijken de scoping methode met truncatieselectie. Er worden drie verschillende curves van de scoping methode onderscheiden, afhankelijk van de waarde van de SR. In de eerste grafiek worden beide methoden met elkaar vergeleken voor de gemiddelde genetische waarde van de tien beste individuen van de populatie en voor de maximum haalbare genetische waarde, in de tweede grafiek voor de gemiddelde genetische waarde van de gehele populatie en in de derde grafiek voor het percentage gefixeerde QTL allelen. De evolutie wordt steeds bekeken over 50 veredelingscycli (Vanavermaete et al., 2020).

tijd die één veredelingscyclus in beslag neemt hier dan gereduceerd kunnen worden van 15 jaar naar drie jaar (Cros et al., 2018).

1.5 Predictiemodellen

1.5.1 Brede waaier aan mogelijkheden

Er zijn reeds veel verschillende predictiemodellen ontwikkeld. De modellen die gebruikt zullen worden in dit onderzoek zijn RR-GBLUP en vier Bayesiaanse modellen (BayesA, BayesB, BayesC en *Bayesian ridge regression* (BRR)). De verschillende modellen worden uitgebreid besproken in Subsectie 2.1.10.

1.5.2 Predictieaccuraatheid van een model

De predictieaccuraatheid van een model is een maat voor hoe goed een bepaald model de data kan voorspellen. Op basis van de predictieaccuraatheid kunnen verschillende modellen ook onderling worden vergeleken. De predictieaccuraatheid van een GS model geeft weer in welke mate de GEBV's die voorspeld worden door middel van het model, overeenkomen met de genetische waarden. Hoe hoger de predictieaccuraatheid van een model in een welbepaalde situatie, hoe beter dit model in deze situatie de genetische waarden kan voorspellen.

Bij GS is deze predictieaccuraatheid de Pearson correlatie tussen de GEBV's en de genetische waarden. De Pearson correlatiecoëfficiënt r_{XY} meet de lineaire afhankelijkheid tussen twee variabelen **X** en **Y**. Deze correlatiecoëfficiënt is één van de meest gebruikte manieren om de relatie tussen twee variabelen te meten (Zhou et al., 2016). De variabelen **X** en **Y** stellen dus vectoren voor van respectievelijk de GEBV's en de genetische waarden van de verschillende individuen. De Pearson correlatiecoëfficiënt is gelijk aan de covariantie van deze twee variabelen gedeeld door het product van de standaardafwijkingen van deze twee variabelen:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\sum_{i=1}^{n}(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \tag{1.8}$$

met

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.9}$$

het gemiddelde van X en

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{1.10}$$

het gemiddelde van Y (Zhou et al., 2016). Hierbij is X_i de i-de waarde van $\mathbf{X}$ en Y_i de i-de waarde van $\mathbf{Y}$. De index i wijst erop dat er gesommeerd wordt over elke waarde tussen 1 en n, waarbij n het aantal individuen voorstelt. Bij GS zullen dus alle verschillende GEBV's worden vergeleken met de corresponderende genetische waarden. Van al deze paarsgewijze vergelijkingen zal dan een som gemaakt worden die de Pearson correlatiecoëfficiënt voorstelt van deze twee variabelen. Hoe groter de Pearson correlatiecoëfficiënt van twee variabelen is, hoe groter de overeenkomst tussen de GEBV's en de genetische waarden en dus hoe hoger de predictieaccuraatheid van het model. De Pearson correlatiecoëfficiënt kan waarden aannemen tussen -1 en 1, voor deze toepassing zijn echter enkel de waarden tussen 0 en 1 relevant. Een nadeel van het gebruik van de Pearson correlatiecoëfficiënt is de zeer hoge gevoeligheid van deze correlatiecoëfficiënt voor uitschieters. Indien extreme datapunten aanwezig zijn, kan het daarom aangeraden zijn een alternatief voor de Pearson correlatiecoëfficiënt te gebruiken. Voorbeelden van dergelijke alternatieven zijn de rangcorrelatiecoëfficiënt van Spearman en het gebruik van de gemiddelde genetische waarde van de populatie. De rangcorrelatiecoëfficiënt van Spearman is een niet-parametrische techniek om na te gaan in welke mate er een correlatie is tussen twee variabelen (Gauthier, 2001). Het grote verschil tussen de rangcorrelatiecoëfficiënt van Spearman en de Pearson correlatiecoëfficiënt is dat bij de eerste gebruikgemaakt wordt van de rang van de data daar waar bij de tweede ruwe data worden gebruikt. De rangcorrelatiecoëfficiënt van Spearman is distributie-onafhankelijk en is relatief ongevoelig voor uitschieters (Gauthier, 2001). Dit is het grote voordeel van de rangcorrelatiecoëfficiënt van Spearman ten opzichte van de Pearson correlatiecoëfficiënt. Het nadeel is dat er informatie verloren gaat wanneer gewerkt wordt met de rang van de data in plaats van met de ruwe data en als de data normaal verdeeld zijn, is deze rangcorrelatiecoëfficiënt minder krachtig dan de Pearson correlatiecoëfficiënt (Gauthier, 2001). Een tweede alternatief voor de Pearson correlatiecoëfficiënt is het gebruik van de gemiddelde genetische waarde van de populatie. Dit alternatief zal later nog meer in detail worden besproken.

Er dient ook opgelet te worden voor overfitting van de trainingspopulatie. Bij overfitting wordt de ruis op datapunten door parameters opgenomen in het model. Wanneer dit gebeurt, wordt een model verkregen dat de trainingspopulatie quasi perfect voorspelt, maar geen betrouwbare voorspellingen kan maken op nieuwe datapunten. De predictieaccuraatheid op nieuwe datapunten is met andere woorden zeer laag wat te allen tijde vermeden moet worden.

1.6 Economisch aspect

Ook het economisch aspect van dit gehele onderzoek mag zeker niet over het hoofd worden gezien. Wanneer er gezocht wordt naar nieuwe methoden om bijvoorbeeld de vruchtopbrengst te verhogen, dient zeker rekening gehouden te worden met de betaalbaarheid van deze nieuwe methoden. Wanneer de kosten voor de implementatie van de nieuwe techniek of de kosten die dienen gemaakt te worden om over voldoende data te kunnen beschikken of andere dergelijke kosten die hierbij kunnen komen kijken, te hoog zijn, zullen deze verbeteringen slechts voor een beperkt aantal telers te verwezenlijken zijn. De telers moeten hun producten ook kunnen verkopen aan een concurrentiële prijs. Het is dus belangrijk een goed evenwicht te vinden tussen enerzijds de opbrengst en anderzijds de kosten.

Het moet uiteraard ook nog steeds de bedoeling zijn dat deze nieuwe, verbeterde technieken op een zo groot mogelijke schaal kunnen worden geïmplementeerd. Het is namelijk noodzakelijk dat er snel grote vooruitgang kan worden geboekt op vlak van bijvoorbeeld vruchtopbrengst, dit om tal van redenen. Denk hierbij maar aan de aangroeiende wereldpopulatie waardoor grotere voedselreserves dienen aangelegd te worden. De constante groei van de wereldbevolking en van de consumptie zullen er namelijk voor zorgen dat de globale vraag naar voedsel nog zal blijven toenemen voor minstens 40 jaar (Godfray et al., 2010). De huidige verbeteringen in opbrengst van planten zijn mogelijks niet voldoende om op lange termijn steeds voldoende voedsel te kunnen voorzien voor iedereen (Allier et al., 2020). Daarom is er nog steeds nood aan nieuwe innovatieve methodes binnen GS.

HOOFDSTUK 2

EXPERIMENTEEL WERK

2.1 Materiaal en methoden

2.1.1 Algemene proefopzet

Bij de verschillende experimenten die tijdens dit onderzoek uitgevoerd werden, werd de genetische waarde beschouwd over meerdere veredelingscycli. Er werd in eerste instantie één welbepaald model beschouwd, namelijk RR-GBLUP. Voor dit model werden twee verschillende methoden beschouwd; enerzijds truncatieselectie en anderzijds de scoping methode. Wat deze methoden precies inhouden, werd eerder al vermeld in de literatuurstudie en zal in volgende secties nog verder worden toegelicht. Voor beide methoden werd onderzocht wat de invloed is van verschillende ouderlijke selectiegroottes op de genetische waarde. Met ouderlijke selectiegrootte wordt het aantal ouders bedoeld dat wordt geselecteerd aan het begin van een nieuwe veredelingscyclus. Er werd nagegaan of de verschillen in genetische waarde tussen verschillende ouderlijke selectiegroottes significant waren en hoe groot deze verschillen gemiddeld waren. Daarnaast werd er ook gezocht vanaf welke waarde voor de selectiegrootte de genetische waarde ongevoelig werd voor verdere verhogingen van die waarde. Nadien werden ook verschillende waarden onderzocht voor het aantal nakomelingen per kruising.

In voorgaande experimenten werd steeds verondersteld dat het beschouwde kenmerk beïnvloed werd door 100 QTL's en dat er sprake was van een heritabiliteit van 0,5. Dit is in de praktijk echter niet altijd het geval. Er werden vervolgens situaties beschouwd waarin deze parameters andere waarden aannamen. Eerst werd het aantal QTL's gevarieerd en de heritabiliteit constant gehouden en vervolgens werden verschillende heritabiliteiten beschouwd bij een constant aantal QTL's. Voor deze verschillende situaties werden steeds drie verschillende ouderlijke selectiegroottes beschouwd. Voor elk experiment werden de ouders zowel eens geselecteerd aan de hand van truncatieselectie als aan de hand van de scoping methode.

In tweede instantie werden vijf verschillende modellen beschouwd. Het model RR-GBLUP, dat tot nu toe steeds werd beschouwd, en daarnaast vier Bayesiaanse modellen. Deze vijf modellen werden vergeleken in verschillende situaties. Drie verschillende waarden voor de ouderlijke selectiegrootte, het aantal QTL's en de heritabiliteit werden beschouwd. Hierbij nam telkens slechts één parameter een waarde aan verschillend van de standaardwaarde. Er werd gekozen voor het onderzoeken van verschillende ouderlijke selectiegroottes in plaats van verschillende aantallen nakomelingen per kruising omdat de modellen gevoeliger zijn voor de ouderlijke selectiegrootte dan voor het aantal nakomelingen per kruising. Door het aantal nakomelingen per kruising te variëren, blijft de genenpoel (*gene pool*) namelijk hetzelfde terwijl door het variëren van de ouderlijke selectiegrootte de genenpoel wel verandert. Door de beperkte tijd die voorhanden was voor dit onderzoek diende er een keuze gemaakt te worden tussen beide parameters.

Voor elk van de modellen werden experimenten uitgevoerd waarbij de ouders enerzijds eens werden geselecteerd aan de hand van truncatieselectie en anderzijds eens aan de hand van de scoping methode.

2.1.2 Datasets en veredelingsprogramma

De populatie en het veredelingsprogramma die gebruikt werden in dit onderzoek, werden overgenomen uit het onderzoek van Neyhart et al. (2017). Deze populatie bestond uit twee datasets van gerst (*Hordeum vulgare*, zesrijige zomergerst) die afkomstig waren van twee Amerikaanse universiteiten, meer bepaald van de universiteiten van Minnesota (*University of Minnesota*) en Noord-Dakota (*North Dakota State University*) (Neyhart et al., 2017; Vanavermaete et al., 2020). Er werd gestart met het selecteren en koppelen van 100 individuen waarbij 50 individuen van de ene dataset werden gekoppeld aan 50 individuen van de andere dataset. Per kruising werden 20 nakomelingen verkregen, op die manier werden 1000 F1 hybriden bekomen die na twee generaties *single-seed descent* (SSD) resulteren in opnieuw 1000 F3 individuen. Uit de nieuwe veredelingspopulatie die zodoende werd verkregen, werden opnieuw 100 individuen geselecteerd die een nieuwe veredelingscyclus startten. De manier waarop deze 100 individuen werden geselecteerd, was uiteraard afhankelijk van de gebruikte selectiemethode (truncatieselectie of scoping methode).

2.1.3 Truncatieselectie

Doorheen dit onderzoek stonden twee methoden centraal; enerzijds de scoping methode (zie verder) en anderzijds truncatieselectie. In het literatuuronderzoek werden

beide methoden reeds uitvoerig besproken. Zoals eerder vermeld, kan truncatieselectie gezien worden als een gulzige selectie waarbij de geselecteerde ouders willekeurig met elkaar worden gekoppeld. Truncatieselectie wordt vaak gebruikt als referentiemethode in wetenschappelijk onderzoek (Vanavermaete et al., 2020; Gorjanc et al., 2018; Cowling et al., 2016). In het voorafgaand literatuuronderzoek werd truncatieselectie ook gebruikt als referentie, de scoping methode werd namelijk vergeleken met deze selectiemethode (zie Figuur 1.4).

2.1.4 Scoping methode

Zoals reeds vermeld werd in het literatuuronderzoek speelt de scoping methode een belangrijke rol in het onderzoek. De methode van ouderlijke selectie waarbij een ouder met een hoge GEBV gecombineerd wordt met een ouder op zo een manier dat de genetische variatie van deze twee geselecteerde ouders gemaximaliseerd wordt over elke merker, wordt de *scoping* methode genoemd (Vanavermaete et al., 2020). Deze methode van selectie van ouderlijke individuen gebeurt door het maximaliseren van de S-score tussen individu i en individu j:

$$S(i,j) = \sum_{m=1}^{k} \text{var}\{Z_{im}, Z_{jm}\} p_m \,, \tag{2.1}$$

met k het aantal merkers, Z_{im} en Z_{jm} het genotype van respectievelijk individu i en j voor de m-de merker en p_m een Boolese waarde voor de m-de merker (Vanavermaete et al., 2020). p_m is initieel gelijk aan 1 voor elke merker. Wanneer echter beide allelen van merker m aanwezig zijn in de geselecteerde populatie, zal p_m de waarde 0 krijgen alvorens het volgende ouderpaar wordt geselecteerd. Op die manier zal de S-score de variatie van het genotype maximaliseren over elke merker waarvoor beide allelen nog niet aanwezig zijn in de geselecteerde populatie. Zo wordt er vermeden dat merkerallelen met een lage frequentie verloren gaan. Als p_m gelijk is aan 0 voor alle merkers zal p_m terug veranderd worden naar 1 om op die manier opnieuw de variatie over alle merkers te maximaliseren. Op dat moment zijn namelijk alle beschikbare merkerallelen van de populatie aanwezig in de geselecteerde populatie (Vanavermaete et al., 2020).

2.1.5 Trainingspopulatie

Zoals in het literatuuronderzoek al gesteld werd (zie Subsectie 1.4.3), was het belangrijk dat de trainingspopulatie werd geüpdatet. Dit werd gedaan door 150 individuen van de veredelingspopulatie toe te voegen aan de trainingspopulatie na elke

veredelingscyclus, deze werden geselecteerd aan de hand van de *Tails* methode (zie Subsectie 1.4.3) (Neyhart et al., 2017). Om de grootte van de trainingspopulatie niet te wijzigen, werden de 150 oudste individuen verwijderd.

2.1.6 Simulator

Alle experimenten werden uitgevoerd in R (versie 3.6.3). De simulator die gebruikt werd, was gebouwd op het werk van Neyhart et al. (2017). Honderd QTL's werden willekeurig geselecteerd uit de 1590 beschikbare biallelische SNP loci. De QTL effecten werden berekend volgens een meetkundige reeks waarbij voor de k-de QTL de waarde van de gunstige homozygoot a^k was, voor de heterozygoot 0 en voor de ongunstige homozygoot $-a^k$ met $a = (L-1)/(L+1)$ en met $L = 100$ (het aantal QTL's dat werd geselecteerd uit de 1590 SNP loci). Er werd verondersteld dat er geen sprake was van dominantie en epistasie. De genetische waarde werd berekend als de som van alle aanwezige QTL allelen.

Het lineair gemengd model werd gebouwd via RR-GBLUP gebruikmakend van het *package* `rrBLUP` (Endelman, 2011) of via de Bayesiaanse modellen gebruikmakend van het *package* `BGLR` (Pérez and de los Campos, 2014).

Bij elke simulatie werd gewerkt met de gemiddelde waarde van 100 herhalingen van het experiment. Hierbij werden 100 verschillende genomen gebruikt.

2.1.7 Parameterwaarden

Er zijn heel wat parameters die werden gevarieerd tijdens het onderzoek. Zo werden verschillende ouderlijke selectiegroottes beschouwd alsook verschillende aantallen nakomelingen per ouderpaar. Daarnaast werd het aantal QTL's alsook de heritabiliteit gevarieerd. Om het effect van elke parameter correct in kaart te brengen, werd telkens slechts één parameter gevarieerd. Elke parameter had een zogenaamde standaardwaarde (*default value*), dit is een waarde die de parameter aannam wanneer deze niet werd gevarieerd.

In Tabel 2.1 wordt een overzicht gegeven van de verschillende parameters die werden gevarieerd tijdens dit onderzoek. Daarnaast is ook de standaardwaarde van elke parameter weergegeven, alsook alle waarden die door de parameter werden aangenomen tijdens het onderzoek. De standaardwaarden die werden gebruikt in dit onderzoek, komen ook in de literatuur vaak terug (Vanavermaete et al., 2020; Neyhart et al., 2017; Bernardo, 2003; VanRaden, 2008; Knapp, 1998; Li and Dungey, 2018; Calus

et al., 2008). De waarden die in de tabel zijn weergegeven voor de ouderlijke selectiegrootte werden niet allemaal aangenomen voor zowel truncatieselectie als voor de scoping methode.

Tabel 2.1: In deze tabel wordt een overzicht gegeven van de verschillende parameters die werden gevarieerd tijdens dit onderzoek. In de tweede kolom is telkens de standaardwaarde weergegeven van deze parameter en in de derde kolom is te zien welke waarden deze parameter allemaal heeft aangenomen tijdens het onderzoek. Niet alle waarden die in de tabel zijn weergegeven voor de ouderlijke selectiegrootte werden zowel voor truncatieselectie als voor de scoping methode aangenomen.

Parameter	**Standaardwaarde**	**Beschouwde waarden**
Ouderlijke selectiegrootte	100	50, 76, 100, 150, 200, 300, 350, 400, 500, 600
Aantal nakomelingen per kruising	20	10, 20, 30, 40, 50
Aantal QTL's	100	50, 100, 200
Heritabiliteit	0,5	0,2; 0,5; 0,8

Zoals in Tabel 2.1 te zien is, werden overheen de experimenten ook verschillende waarden beschouwd voor de heritabiliteit. Er bestaan echter twee heritabiliteiten; enerzijds die in brede zin en anderzijds die in enge zin. Het is belangrijk te vermelden dat de heritabiliteit die gebruikt wordt als parameter in de experimenten de heritabiliteit in enge zin voorstelt (h^2). Voor deze heritabiliteit werden drie verschillende waarden beschouwd. Enerzijds de standaardwaarde 0,5 en anderzijds werden ook heritabiliteiten van 0,2 en 0,8 onderzocht. Er werd gekozen voor de heritabiliteiten 0,2 en 0,8 omdat deze waarden beschouwd kunnen worden als extreme waarden van de heritabiliteit. Bijna alle heritabiliteiten die gevonden worden in de praktijk zijn namelijk gelegen tussen 0,2 en 0,8. Een heritabiliteit die kleiner is dan 0,2 of groter dan 0,8 is eerder uitzonderlijk. Zo werd in het onderzoek van Bargougui (2016) voor gerst een heritabiliteit geschat van 0,49 voor het aantal aren; 0,24 voor het aantal graankorrels per aar en 0,25 voor het gewicht van duizend graankorrels. In het onderzoek van Eshghi and Akhundova (2010) werden voor enkele belangrijke landbouwkundige eigenschappen van gerst volgende heritabiliteiten geschat: 0,72–0,79 (planthoogte); 0,92–0,93 (aantal dagen tot de plant volwassen is); 0,6–0,85 (aantal scheuten); 0,2–0,4 (aantal graankorrels per aar) en 0,21–0,37 (graanopbrengst per plant). Er is te zien dat de meeste heritabiliteiten gelegen zijn tussen 0,2 en 0,8, al zijn er wel uitzonderingen (bijvoorbeeld voor het aantal dagen tot een gerstplant volwassen is).

2.1.8 Gemiddelde genetische waarde van de top-10 individuen

Voor het beoordelen van de verschillende modellen en methoden, kon de Pearson correlatiecoëfficiënt niet worden gebruikt door de hoge gevoeligheid van deze corre-

latiecoëfficiënt voor uitschieters (zie Subsectie 1.5.2). Bij de scoping methode kunnen namelijk extreme GEBV's geselecteerd worden; enerzijds worden individuen geselecteerd met hoge GEBV's en anderzijds individuen met lagere GEBV's met als doel de genetische waarde te maximaliseren op lange termijn. Door deze grote verschillen in GEBV's van de geselecteerde individuen zou de Pearson correlatiecoëfficiënt een vertekend beeld geven. Bij truncatieselectie zou dit veel minder het geval zijn omdat bij deze selectiemethode individuen met hoge GEBV's worden geselecteerd zonder dat ook individuen met lagere GEBV's worden geselecteerd. Voor het beoordelen en vergelijken van verschillende modellen en methoden werd in de plaats van de Pearson correlatiecoëfficiënt gebruikgemaakt van de gemiddelde genetische waarde van de populatie.

Zoals tijdens de literatuurstudie reeds vermeld werd, kunnen meerdere genetische waarden worden opgevolgd over de verschillende veredelingscycli heen. Enkele voorbeelden hiervan zijn de gemiddelde genetische waarde van de gehele populatie, de gemiddelde genetische waarde van de top-10 individuen van de populatie en de maximum haalbare genetische waarde (zie Figuur 1.3). Tijdens dit onderzoek zal ervoor gekozen worden om de gemiddelde genetische waarde van de top-10 individuen van de populatie te beschouwen. Deze genetische waarde wordt verkozen omdat er vooral interesse is in de beste individuen van de populatie. Deze individuen zullen namelijk op de markt worden gebracht. Daarnaast zou de gemiddelde genetische waarde van de populatie ook een verkeerd beeld geven aangezien de scoping methode individuen selecteert met een lagere genetische waarde om de genetische variatie in de populatie te behouden. Zodoende wordt de gemiddelde genetische waarde van de populatie sterk beïnvloed door deze individuen met een lagere genetische waarde terwijl er echter interesse is in de beste individuen van de populatie.

2.1.9 Aantal veredelingscycli

De gemiddelde genetische waarde van de top-10 individuen werd beschouwd over verschillende veredelingscycli. Aanvankelijk werden deze genetische waarden beschouwd over 50 veredelingscycli, maar er werd na verloop van tijd overgeschakeld naar 25 veredelingscycli. Dit werd gedaan omdat tijdens de eerste experimenten duidelijk werd dat er reeds voldoende informatie verkregen was na het doorlopen van 25 veredelingscycli. Het reduceren van het aantal veredelingscycli verminderde de nodige simulatietijd aanzienlijk. De experimenten die gedaan werden op basis van het model RR-GBLUP en waarbij de heritabiliteit 0,5 bedroeg en het aantal QTL's 100, werden beschouwd over 50 veredelingscycli. De andere experimenten die werden gedaan, werden beschouwd over 25 veredelingscycli.

2.1.10 Predictiemodellen

Ridge regression genomic best linear unbiased prediction

In het literatuuronderzoek kwam GBLUP reeds uitgebreid aan bod (zie Subsectie 1.4.1). *Ridge regression genomic best linear unbiased prediction* (RR-GBLUP) is ontstaan uit GBLUP, het grote verschil tussen beide is dat er bij RR-GBLUP assumpties worden gemaakt die leiden tot zogenaamde *ridge regression* (RR). Door deze aanname namen alle merkereffecten gelijk af naar nul, maar is het nog steeds mogelijk dat merkers verschillende effecten hadden (Heffner et al., 2011). Bij RR wordt de som van alle parameters die aanwezig zijn in het model geminimaliseerd waardoor de parameters geen extreme waarden kunnen aannemen. Dit heeft als doel overfitting te voorkomen.

In dit onderzoek werd steeds de term RR-GBLUP gebruikt. In de literatuur wordt echter dikwijls de term RR-BLUP gebruikt hiervoor. Het is hierbij wel belangrijk om op te merken dat deze twee termen niet altijd hetzelfde zijn. Enkel als in de literatuur gesproken wordt over RR-BLUP waarbij relaties op basis van het genotype zijn gedefinieerd, houden beide termen hetzelfde in. De term RR-GBLUP werd in dit onderzoek gebruikt om te benadrukken dat enerzijds onderlinge relaties waren gebaseerd op genetische merkers in plaats van op verwantschap (Meuwissen et al., 2016) en dat er anderzijds sprake was van *ridge regression*.

Bayesiaanse modellen

Naast RR-GBLUP worden vier Bayesiaanse modellen beschouwd, meer bepaald BayesA, BayesB, BayesC en *Bayesian ridge regression* (BRR). Bij BayesB wordt verondersteld dat het effect van een merker (SNP) nul is met een waarschijnlijkheid van $\pi \in]0, 1[$ en het merkereffect wordt verondersteld normaal verdeeld te zijn met een waarschijnlijkheid van $(1 - \pi)$ (Habier et al., 2011). Deze π-waarde is steeds groter dan nul waardoor bij BayesB kan worden gesteld dat tal van merkers geen effect hebben (Habier et al., 2011). BayesA lijkt in grote mate op BayesB maar het grote verschil tussen beide is dat bij BayesA alle merkers verondersteld worden een effect te hebben daar waar dit bij BayesB niet het geval is. Indien bij BayesA of BayesB het merkereffect niet nul is (wat dus geldt voor alle merkers bij BayesA en voor sommige merkers bij BayesB), wordt dit merkereffect verondersteld een normale verdeling te volgen met een gemiddelde van 0 en een variantie die specifiek is voor de locus. Bij BayesA en BayesB verschillen de varianties van merker tot merker. Deze locusspecifieke variantie kan worden vervangen door een variantie die gelijk is voor alle

merkers (Habier et al., 2011). In dit geval wordt gesproken van BayesC. Net zoals bij BayesB is er bij BayesC ook sprake van een π-waarde die steeds groter is dan 0.

Ten slotte is BRR een Bayesiaans model waarbij er, net zoals bij RR-GBLUP, sprake is van *ridge regression*. De merkereffecten worden bij BRR, net zoals bij BayesC, verondersteld een normale verdeling te volgen (Desta and Ortiz, 2014). Een groot verschil tussen BRR en BayesC is dat bij BRR alle merkers verondersteld worden een effect te hebben terwijl dit bij BayesC niet het geval is.

In het algemeen kan worden gesteld dat bij BayesA en BayesB de merkers afkomstig zijn van een geschaalde t-verdeling daar waar deze bij BayesC en BRR afkomstig zijn van een normale verdeling. Bij BayesB en BayesC heeft een welbepaalde fractie van de merkers geen effect daar waar bij BayesA en BRR elke merker een effect heeft.

Zoals eerder al werd aangehaald, werd gebruikgemaakt van het *package* BGLR (Pérez and de los Campos, 2014). Hierin is de π-waarde echter anders gedefinieerd dan in het onderzoek van Habier et al. (2011). De fractie van de merkers die verondersteld wordt geen effect te hebben, wordt namelijk weergegeven als $(1-\pi)$ terwijl dit bij Habier et al. (2011) weergegeven werd als π.

De π-waarde waarmee standaard wordt gewerkt in het *package* BGLR, werd ook gebruikt in dit onderzoek (enkel van toepassing voor BayesB en BayesC). π volgde op die manier steeds een β-verdeling $\beta(p_0, \pi_0)$ met $p_0 = 10$ en $\pi_0 = 0,5$ (Pérez and de los Campos, 2014).

Alle beschouwde Bayesiaanse modellen kunnen als volgt geschreven worden:

$$\mathbf{y} = \mathbf{1}_n\beta + \mathbf{Xu} + \boldsymbol{\epsilon}\,, \tag{2.2}$$

met $\mathbf{X}$ een $[n \times m]$ matrix die het merker-gecentreerde genotype bevat waarbij elke merker werd gecentreerd door het gemiddelde af te trekken en werd gestandaardiseerd door vervolgens te delen door de standaarddeviatie van de steekproef (Gapare et al., 2018).

2.1.11 Lineair regressiemodel

Zoals al aan bod kwam in het literatuuronderzoek werd gewerkt met lineaire gemengde modellen. Aan de voorwaarden van een lineair regressiemodel dient dus steeds voldaan te zijn. De eerste voorwaarde van een lineair regressiemodel is de lineariteitsvoorwaarde. Deze voorwaarde kan gecontroleerd worden met behulp van een residuplot. Er kan lineariteit worden verondersteld wanneer de residuen, die zichtbaar zijn op de residuplot, willekeurig verspreid liggen rond de x-as (Luca, 2019).

Dit is steeds het geval in het onderzoek. Een tweede voorwaarde is normaliteit van de residuen. Dit kan worden geverifieerd met behulp van een normale kwantielplot (QQ-plot, *quantile-quantile plot*) van de residuen (Luca, 2019). Aan deze voorwaarde is echter steeds voldaan aangezien $\boldsymbol{\epsilon}$ normaal verdeeld is (zie Formule (1.5)). Ten derde is er de homoscedasticiteitsvoorwaarde. Aan deze voorwaarde is voldaan aangezien de residuen uit dezelfde normale verdeling komen (Luca, 2019). Ten slotte dienen de residuen onderling onafhankelijk te zijn. Dit is moeilijk te controleren aan de hand van een residuplot maar kan wel in beperkte mate door na te gaan of er geen patroon te vinden is in de opeenvolgende residuen (Luca, 2019). Er kan worden aangenomen dat voldaan is aan deze voorwaarde omdat enerzijds de experimenten onafhankelijk van elkaar gebeurden en anderzijds omdat in de residuplots nooit dergelijke patronen konden worden vastgesteld. Op die manier is voldaan aan alle voorwaarden van een lineair regressiemodel.

2.1.12 Statistische analyse

In dit onderzoek gebeurden de statistische analyses aan de hand van variantieanalyses of ANOVA-analyses (*analysis of variance*). Aan de hand van een ANOVA kan worden onderzocht of er een significant verschil is tussen de gemiddelden van twee of meerdere groepen van een onafhankelijke variabele (Skidmore and Thompson, 2013). De reden waarom gekozen wordt voor ANOVA is omdat er direct meerdere groepen van data onderling kunnen worden vergeleken.

De onafhankelijke variabelen hebben een categorische indeling wat betekent dat elk datapunt kan ingedeeld worden in één welbepaalde categorie binnen elke onafhankelijke variabele. In dit onderzoek is de afhankelijke variabele de genetische waarde, meer bepaald de gemiddelde genetische waarde van de top-10 individuen van de populatie, en de onafhankelijke variabelen zijn de ouderlijke selectiegrootte en het aantal nakomelingen per kruising.

Er bestaan meerdere soorten ANOVA's. In dit onderzoek werd de zogenaamde éénwegs ANOVA (*one-way ANOVA*) gebruikt. Hierbij wordt slechts één onafhankelijke variabele per keer beschouwd en wordt gekeken naar wat de invloed is van de verschillende groepen binnen deze variabele op de genetische waarde (Mackenzie, 2018). Er wordt onderzocht of de verschillende groepsgemiddelden significant verschillend zijn van elkaar ($p<0,05$). De nulhypothese bij een éénwegs ANOVA houdt in dat de gemiddelden van alle groepen niet significant verschillend zijn van elkaar. Wanneer de nulhypothese wordt verworpen, en de alternatieve hypothese wordt aangenomen, zijn minstens twee groepsgemiddelden significant verschillend. Via een post-hoc analyse kan dan worden onderzocht welke groepsgemiddelden significant verschillend

zijn van elkaar (Brown, 2005). De methode van Tukey-Kramer werd in dit onderzoek gebruikt voor de post-hoc analyse. Dit is een methode die vaak gebruikt wordt om paarsgewijze verschillen te onderzoeken (Luca, 2019).

Wanneer een statistische analyse wordt uitgevoerd, is het steeds van groot belang dat voldaan is aan alle voorwaarden van de statistische test. Een eerste voorwaarde waaraan voldaan moet zijn om een ANOVA-analyse te kunnen uitvoeren, is homoscedasticiteit. Dit betekent dat de verschillende groepen die onderling worden vergeleken gelijke varianties dienen te hebben. Deze voorwaarde werd gecontroleerd aan de hand van de test van Levene. De nulhypothese van deze test houdt homoscedasticiteit in. Wanneer bij deze test dus een p-waarde wordt verkregen groter dan 0,05 betekent dit dat er geen voldoende bewijs kan gevonden worden om de nulhypothese te verwerpen. Alle p-waarden die werden bekomen aan de hand van deze test waren groter dan 0,05, er kan dus steeds verondersteld worden dat er aan de homoscedasticiteitsvoorwaarde is voldaan. Ten tweede is er de normaliteitsvoorwaarde. Deze voorwaarde werd gecontroleerd door boxplots, histogrammen en de Shapiro-Wilk test te gebruiken. De Shapiro-Wilk test kan op eenzelfde manier worden geïnterpreteerd als de test van Levene, enkel wordt bij deze test een normale verdeling verondersteld in de nulhypothese. Er waren echter vaak afwijkingen ten opzichte van de normaliteit, het ging wel meestal over relatief lichte afwijkingen waardoor volgens de centrale limietstelling toch normaliteit kon worden verondersteld. De centrale limietstelling stelt namelijk dat er bij lichte afwijkingen ten opzichte van de normaliteit toch normaliteit mag worden verondersteld op voorwaarde dat er minstens 30 geobserveerde gemiddelden zijn. In elk experiment van dit onderzoek werd gewerkt met minstens 30 geobserveerde gemiddelden waardoor deze stelling mag worden toegepast en dus mag worden verondersteld dat ook aan de normaliteitsvoorwaarde is voldaan. Een derde en laatste voorwaarde is onafhankelijkheid. Ondanks dat de verschillende experimenten allemaal gebaseerd zijn op dezelfde data kan toch worden gesteld dat de verschillende experimenten volledig onafhankelijk van elkaar gebeurden aangezien de QTL's willekeurig werden geselecteerd uit de 1590 beschikbare biallelische SNP loci. Er kan dus worden gesteld dat ook aan deze voorwaarde is voldaan. Op die manier is voldaan aan alle voorwaarden en kunnen ANOVA-analyses worden uitgevoerd.

2.2 Resultaten en discussie

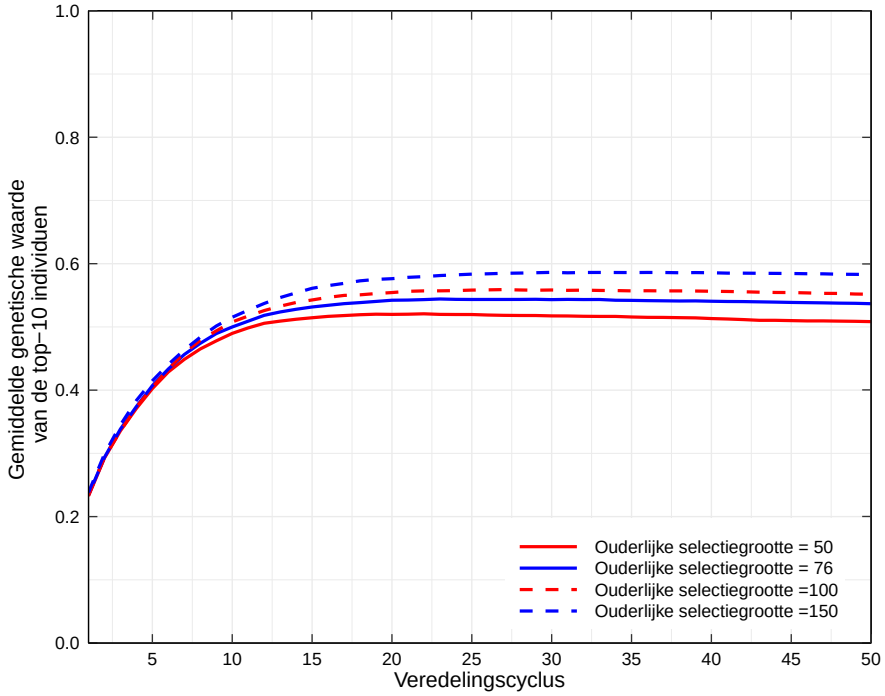
2.2.1 RR-GBLUP

Eerst wordt er gefocust op het model RR-GBLUP. De invloed van verschillende parameters wordt onderzocht wanneer truncatieselectie of de scoping methode gebruikt wordt als ouderlijke selectiemethode. In eerste instantie zullen verschillende ouderlijke selectiegroottes worden beschouwd en in tweede instantie zal het aantal nakomelingen per ouderpaar worden onderzocht. Ten slotte zullen verschillende waarden voor het aantal QTL's en de heritabiliteit worden beschouwd.

Ouderlijke selectiegrootte

De invloed van de ouderlijke selectiegrootte op de gemiddelde genetische waarde van de top-10 individuen wordt onderzocht. Hiervoor worden experimenten gedaan met verschillende ouderlijke selectiegroottes. Eerst wordt de invloed nagegaan wanneer de ouders worden geselecteerd aan de hand van truncatieselectie. In Figuur 2.1 zijn de gemiddelde resultaten weergegeven van de top-10 kandidaten voor verschillende ouderlijke selectiegroottes. Er is te zien dat een toename van het aantal geselecteerde ouders ervoor zorgt dat er hogere genetische waarden worden behaald. Niet tussen elke twee ouderlijke selectiegroottes is er na 50 veredelingscycli echter sprake van een significant verschil (Tabel A1 in Bijlagen). Er is een significant verschil tussen de ouderlijke selectiegroottes 50 en 100 ($p=0,017$), 50 en 150 ($p<0,001$) en 76 en 150 ($p=0,018$). Tussen de andere ouderlijke selectiegroottes zijn de verschillen niet significant. De genetische waarde die na 50 veredelingscycli wordt behaald voor een ouderlijke selectiegrootte van 76 is 2,8%-punt hoger dan voor een ouderlijke selectiegrootte van 50 en die voor een ouderlijke selectiegrootte van 100 is 1,5%-punt hoger dan die voor een ouderlijke selectiegrootte van 76 (Tabel A2 in Bijlagen). Ten slotte wordt voor een ouderlijke selectiegrootte van 150 een genetische waarde behaald die 3,1%-punt hoger is dan die voor een ouderlijke selectiegrootte van 100.

Dat er hogere genetische waarden worden behaald naarmate meer ouders worden geselecteerd, ligt in lijn met de verwachtingen. Wanneer meer ouders worden geselecteerd, kan de genetische variatie namelijk beter worden behouden (Figuur A1 in Bijlagen) wat ervoor zorgt dat er een hogere genetische waarde kan worden behaald op lange termijn. Er is te zien dat de genetische waarden vrij snel convergeren wat betekent dat er na verloop van tijd nog amper hogere genetische waarden worden behaald wanneer truncatieselectie wordt toegepast. Dit komt doordat de genetische variatie snel afneemt wat gepaard gaat met een snelle fixatie van QTL allelen. Dit

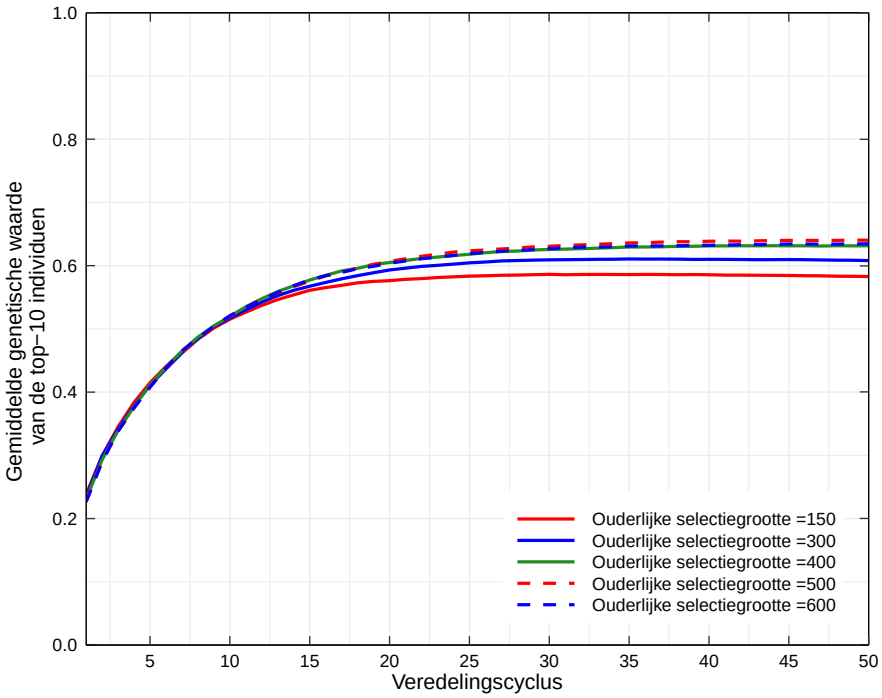


Figuur 2.1: Op deze grafiek zijn vier verschillende ouderlijke selectiegroottes voor truncatieselectie weergegeven. Er is te zien dat de genetische waarde die wordt bekomen op lange termijn groter is naarmate meer ouders worden geselecteerd.

kwam ook in de literatuurstudie al uitgebreid aan bod. Er is ook te zien dat naarmate er meer ouders worden geselecteerd, er pas later (dus na meer veredelingscycli) convergentie optreedt. De verklaring hiervoor kan opnieuw gevonden worden in de genetische variatie die beter behouden wordt wanneer meer ouders worden geselecteerd.

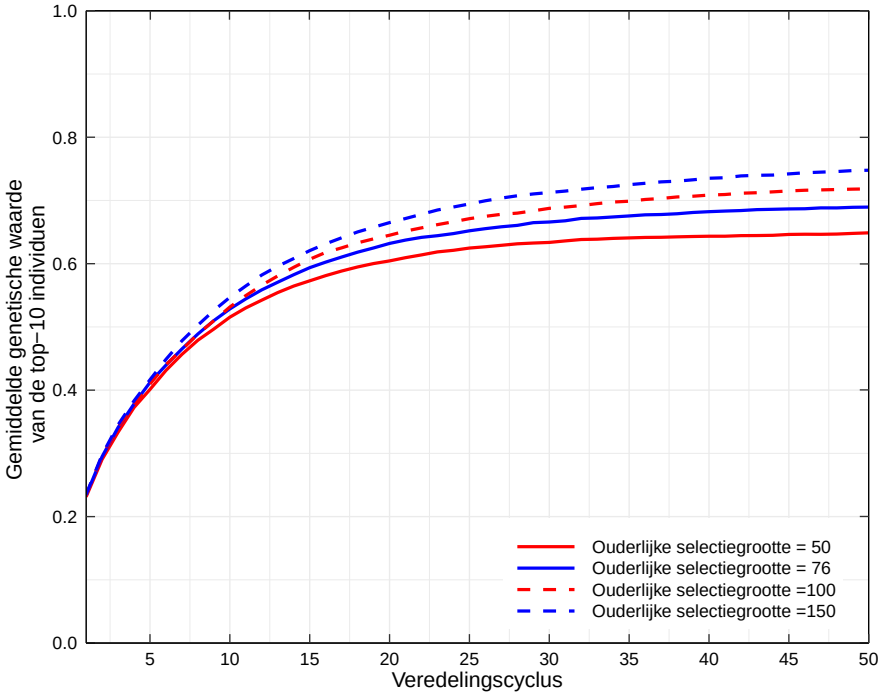
Wanneer de financiële middelen voor het verhogen van de ouderlijke selectiegrootte voorhanden zijn, lijkt het voordelig dit ook effectief te doen. Er wordt onderzocht of een verhoging van de ouderlijke selectiegrootte zal blijven zorgen voor het behalen van hogere genetische waarden of dat er een bepaalde ouderlijke selectiegrootte bestaat vanaf wanneer quasi geen hogere genetische waarden meer kunnen worden behaald indien deze nog verder zou toenemen. Om dit te onderzoeken worden tal van experimenten gedaan waarbij steeds meer ouders worden geselecteerd. De resultaten die horen bij deze experimenten zijn weergegeven in Figuur 2.2.

Er is te zien op de grafiek dat de genetische waarde die hoort bij een ouderlijke selectiegrootte van 300 op lange termijn duidelijk groter is dan de genetische waarde die hoort bij een ouderlijke selectiegrootte van 150, dit is ook het geval voor de genetische waarde die hoort bij een ouderlijke selectiegrootte van 400 ten opzichte van een ouderlijke selectiegrootte van 300. Wanneer er echter 500 ouders worden geselecteerd, convergeert de gemiddelde genetische waarde van de top-10 individuen



Figuur 2.2: Op de grafiek zijn hogere waarden van de ouderlijke selectiegrootte voor truncatieselectie voorgesteld. Er kan uit de grafiek worden afgeleid dat het niet interessant is om meer dan 400 ouders te selecteren wanneer truncatieselectie wordt gebruikt, er worden namelijk amper hogere genetische waarden bekomen wanneer nog meer ouders worden geselecteerd. Er is wel een duidelijk verschil te zien tussen het selecteren van 400 ouders en het selecteren van een lager aantal ouders.

gelijkaardig als die voor een selectiegrootte van 400. Het verder verhogen van de selectiegrootte naar 600 resulteert opnieuw in de convergentie van de genetische waarde naar een gelijkaardige waarde als die voor een selectiegrootte van 500 of 400. Aan de hand van deze resultaten zou kunnen worden verondersteld, specifiek voor de data waarmee in het onderzoek gewerkt wordt, dat het vanaf een ouderlijke selectiegrootte van 400 minder interessant is nog meer ouders te selecteren. De bijhorende ANOVA-analyse toont echter dat er geen significante verschillen zijn tussen al deze beschouwde ouderlijke selectiegroottes (p=0,126). Er kan dus worden gesteld dat er wel duidelijke verschillen zijn tussen de verschillende beschouwde ouderlijke selectiegroottes maar dat deze verschillen niet significant zijn. Zo bedraagt de genetische waarde wanneer 300 ouders worden geselecteerd 2,4%-punt meer dan wanneer 150 ouders worden geselecteerd en het verschil tussen ouderlijke selectiegroottes van 300 en 400 bedraagt 2,5%-punt (zie Tabel A3 in Bijlagen). Voor een ouderlijke selectiegrootte van 500 wordt daarentegen een genetische waarde bekomen die slechts 0,5%-punt hoger is dan de genetische waarde die bekomen wordt wanneer 400 ouders worden geselecteerd. Wanneer 600 ouders worden geselecteerd, wordt zelfs een genetische waarde bekomen die lager is dan de genetische waarde wanneer 500 ouders worden geselecteerd, het verschil in genetische waarde tussen beide ouderlijke selectiegroottes bedraagt 0,4%-punt. Zowel het verschil in genetische waarde tussen ouderlijke selectiegroottes van 400 en 500 als tussen ouderlijke selectiegroottes van 500 en 600 is vermoedelijk te wijten aan willekeurige effecten. Naast de genetische waarde worden ook willekeurige effecten in rekening gebracht in de fenotypische waarde. Deze willekeurige effecten zijn enerzijds milieueffecten en anderzijds residuele fouten, beide afkomstig uit een normale verdeling. Indien deze willekeurige effecten sterk verschillend zijn voor beide experimenten kan voor eenzelfde genetische waarde een (licht) verschillende fenotypische waarde worden bekomen. Omdat het fenotype een belangrijke rol speelt bij het fitten van het lineair gemengd model, kan hierdoor een verschillende predictie worden verkregen die ervoor kan zorgen dat er andere ouders worden geselecteerd. Dat deze willekeurige effecten aan de basis liggen van deze verschillen in genetische waarde zou met meer zekerheid kunnen worden gesteld indien meer herhalingen van de experimenten zouden worden beschouwd. Het is echter niet mogelijk dergelijke experimenten te beschouwen wegens de beperkte hoeveelheid tijd die voorhanden is voor dit onderzoek. Daarnaast dient ook opgemerkt te worden dat voor zowel het verschil tussen selectiegroottes van 400 en 500 als voor het verschil tussen selectiegroottes van 500 en 600 de standaarddeviatie van dit verschil groter is dan het verschil zelf (zie Tabel A3 in Bijlagen). Hieruit kan worden gesteld dat de kans groot is dat deze verschillen wel degelijk te wijten zijn aan deze willekeurige effecten.

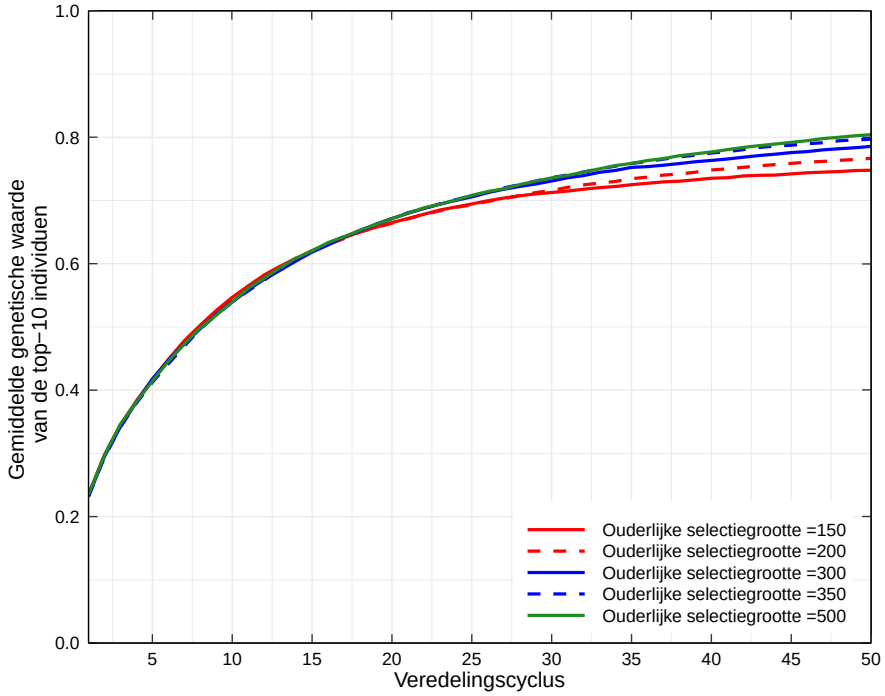


Figuur 2.3: Op deze figuur worden de genetische waarden weergegeven voor verschillende ouderlijke selectiegroottes waarbij de ouders worden geselecteerd aan de hand van de scoping methode. Er is te zien dat, net zoals bij truncatieselectie, op lange termijn hogere genetische waarden worden bekomen wanneer meer ouders worden geselecteerd. Het valt verder op dat convergentie veel minder snel optreedt dan bij truncatieselectie.

Er kan dus worden gesteld voor de data die gebruikt worden in het onderzoek dat het vanaf een ouderlijke selectiegrootte van 400 minder interessant is om nog meer ouders te selecteren. Vanaf dit punt zal het verhogen van de ouderlijke selectiegrootte enkel zorgen voor een extra kost, er kunnen namelijk nog amper hogere genetische waarden worden gehaald door het verder verhogen van de ouderlijke selectiegrootte. Het is hierbij wel belangrijk om op te merken dat steeds de afweging dient gemaakt te worden tussen enerzijds de opbrengst die gepaard gaat met een hogere genetische waarde en anderzijds de kost die het verhogen van de ouderlijke selectiegrootte met zich meebrengt. Opnieuw valt op dat de gemiddelde genetische waarde van de top-10 individuen snel convergeert. Deze convergentie vindt in het algemeen sneller (dus na minder veredelingscycli) plaats naarmate minder ouders worden geselecteerd.

In Figuur 2.3 wordt de gemiddelde genetische waarde van de top-10 individuen weergegeven voor verschillende ouderlijke selectiegroottes waarbij de ouders worden geselecteerd aan de hand van de scoping methode.

Er is te zien dat ook hier hogere genetische waarden worden verkregen naarmate meer ouders worden geselecteerd. Aan de hand van een ANOVA-analyse worden significante verschillen gevonden tussen de ouderlijke selectiegroottes 50 en 76 (p=



Figuur 2.4: Relatief hoge waarden van de ouderlijke selectiegrootte voor de scoping methode zijn voorgesteld op deze grafiek. Het is moeilijk om aan de hand van deze resultaten de ouderlijke selectiegrootte te bepalen vanaf wanneer het minder interessant wordt om meer ouders te selecteren. Wel kan over deze waarde worden gesteld dat deze vermoedelijk niet groter is dan 350 aangezien de genetische waarden die behaald worden voor ouderlijke selectiegroottes van 350 en 500 convergeren naar een gelijkaardig punt.

0,008), 50 en 100 (p<0,001), 50 en 150 (p<0,001) en 76 en 150 (p<0,001) (zie Tabel A4 in Bijlagen). Tussen de andere ouderlijke selectiegroottes zijn er ook merkbare verschillen maar deze zijn dus niet significant. Voor een ouderlijke selectiegrootte van 76 worden genetische waarden bekomen die 4,0%-punt hoger zijn dan die wanneer 50 ouders worden geselecteerd (zie Tabel A5 in Bijlagen). De verhoging in genetische waarde wanneer 100 in plaats van 76 ouders worden geselecteerd, bedraagt 2,9%-punt. Ook wanneer 150 in plaats van 100 ouders worden geselecteerd, is het verschil in genetische waarde gelijk aan 2,9%-punt. Net zoals bij truncatieselectie convergeren de genetische waarden minder snel wanneer meer ouders worden geselecteerd. De verklaring hiervoor is analoog als bij truncatieselectie, meer ouders selecteren zorgt voor het behoud van meer genetische variatie (zie Figuur A2 in Bijlagen) waardoor op langere termijn hogere genetische waarden kunnen worden behaald.

In Figuur 2.4 zijn genetische waarden weergegeven die behaald worden voor hogere ouderlijke selectiegroottes. Ook voor de scoping methode wordt getracht de ouderlijke selectiegrootte te bepalen vanaf wanneer het minder interessant is om meer ouders te selecteren.

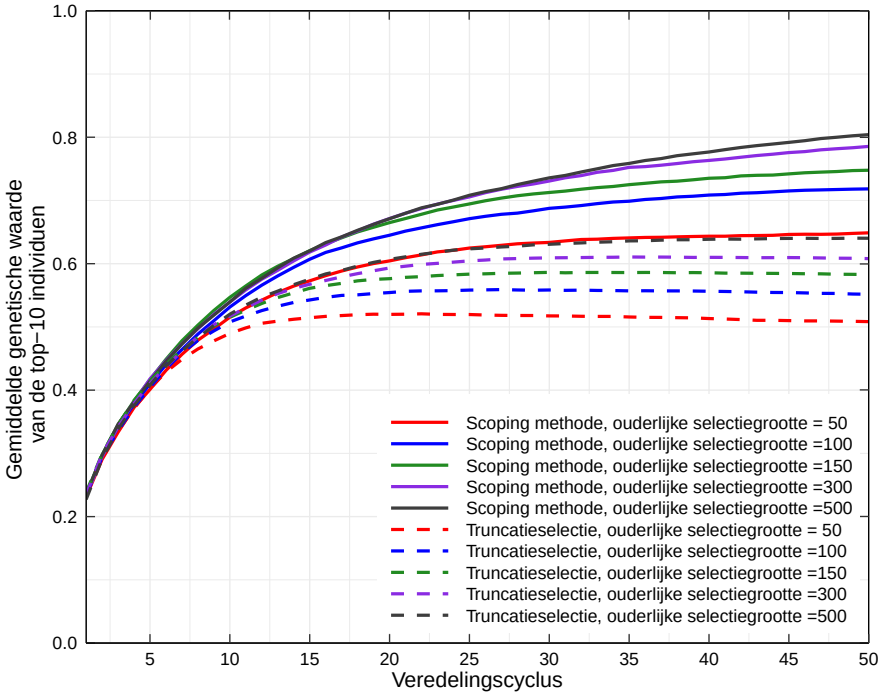
Op de grafiek is te zien dat er een duidelijk verschil is tussen ouderlijke selectiegroottes van 150 en 200. Het verschil in genetische waarde tussen beide bedraagt 1,9%-punt (zie Tabel A6 in Bijlagen). Ook tussen ouderlijke selectiegroottes van 200 en 300 is er een duidelijk verschil zichtbaar, het verschil in genetische waarde bedraagt 1,9%-punt. Het verschil in genetische waarde tussen ouderlijke selectiegroottes van 300 en 350 bedraagt 1,1%-punt en tussen ouderlijke selectiegroottes van 350 en 500 bedraagt dit verschil 0,7%-punt. Aan de hand van deze resultaten is het moeilijk om de ouderlijke selectiegrootte te bepalen vanaf wanneer het minder interessant wordt om meer ouders te selecteren. Er kan wel worden gesteld dat deze waarde hoogstwaarschijnlijk niet hoger is dan 350 aangezien de genetische waarden voor ouderlijke selectiegroottes van 350 en 500 naar een gelijkaardige waarde convergeren. Tussen ouderlijke selectiegroottes van 150 en 200 is er een duidelijk verschil in genetische waarde en is de stapgrootte relatief klein waardoor deze waarde hoogstwaarschijnlijk groter zal zijn dan 200. Tussen de genetische waarden die behaald worden voor selectiegroottes van 200 en 300 is er ook een duidelijk verschil maar dit verschil geeft een vertekend beeld aangezien de stapgrootte hier 100 bedraagt. Aangezien de stapgrootte tussen 300 en 350 slechts 50 bedraagt, is het moeilijk om het verschil in genetische waarde tussen deze twee selectiegroottes in te schatten ten opzichte van het verschil in genetische waarde tussen selectiegroottes van 200 en 300. Er dienen meer experimenten uitgevoerd te worden met een ouderlijke selectiegrootte tussen 200 en 350 om de waarde vanaf wanneer het minder interessant is om meer ouders te selecteren, specifiek voor de gebruikte data, nauwkeuriger te kunnen bepalen. Door de beperkte tijd voor dit onderzoek werden vaste waarden per parameter vastgelegd waardoor geen experimenten werden uitgevoerd voor selectiegroottes tussen 200 en 300. Het dient verder ook opgemerkt te worden dat de corresponderende ANOVA-analyse aantoont dat er echter wel geen significante verschillen zijn tussen deze relatief hoge ouderlijke selectiegroottes (p=0,571).

Verder valt het op dat ook op lange termijn nog steeds hogere genetische waarden kunnen worden bekomen daar waar dit bij truncatieselectie amper het geval is. De verklaring hiervoor kan ook gevonden worden in de genetische variatie die veel beter behouden wordt wanneer de scoping methode wordt gebruikt voor het selecteren en kruisen van de ouders. Door het behouden van de genetische variatie zal de scoping methode in staat zijn gunstige maar slecht geschatte QTL allelen in de populatie te behouden. Hierdoor kunnen op langere termijn hogere genetische waarden worden behaald. Bij truncatieselectie daarentegen zal een slechte schatting van de QTL effecten vaak leiden tot de fixatie van die verkeerde allelen waardoor de gemiddelde genetische waarde vroegtijdig convergeert.

Een tweede opvallend verschil tussen beide selectiemethoden is dat bij de scoping methode de ouderlijke selectiegrootte vanaf wanneer het minder interessant is om meer ouders te selecteren kleiner is dan bij truncatieselectie. De verklaring hiervoor kan opnieuw gevonden worden in de genetische variatie, doordat deze beter wordt behouden bij de scoping methode zal een relatief lage ouderlijke selectiegrootte al in staat zijn een groot deel van de genetische variatie te behouden. Bij de scoping methode kan met andere woorden uit een lager aantal ouders sneller alle beschikbare informatie worden gehaald en worden doorgegeven naar de volgende generatie. Hierdoor zal de ouderlijke selectiegrootte vanaf wanneer een verhoging van het aantal geselecteerde ouders resulteert in een zeer beperkte toename van de genetische variatie groter zijn bij truncatieselectie aangezien de genetische variatie daar in veel mindere mate behouden wordt. Een toename van het aantal geselecteerde ouders zal er zodoende langer voor kunnen blijven zorgen dat de genetische variatie relatief sterk toeneemt.

Op de grafieken is te zien dat er hogere genetische waarden worden bekomen voor de scoping methode dan voor truncatieselectie. Het verschil tussen beide selectiemethoden wordt nog eens duidelijk weergegeven in Figuur 2.5. Op deze figuur worden beide selectiemethoden met elkaar vergeleken voor verschillende ouderlijke selectiegroottes. Er is te zien dat voor elke ouderlijke selectiegrootte duidelijk hogere genetische waarden worden bekomen voor de scoping methode. Op basis van een ANOVA-analyse kan worden gesteld dat elk verschil tussen de genetische waarde bij de scoping methode en die bij truncatieselectie voor eenzelfde ouderlijke selectiegrootte significant is. De verschillen in genetische waarde tussen beide selectiemethoden voor eenzelfde ouderlijke selectiegrootte alsook de corresponderende p-waarden kunnen worden teruggevonden in Tabel 2.2. Er kan dus worden gesteld dat, ongeacht de ouderlijke selectiegrootte, het aangewezen is om de scoping methode altijd te gebruiken in de plaats van truncatieselectie.

Dat hogere genetische waarden worden behaald voor de scoping methode in vergelijking met truncatieselectie kan verklaard worden door de genetische variatie. In Figuur 2.6 is te zien dat voor elke ouderlijke selectiegrootte de genetische variatie veel beter behouden wordt voor de scoping methode dan voor truncatieselectie. Hierdoor kunnen fouten die gemaakt worden in de voorspellingen beter worden gecorrigeerd bij de scoping methode aangezien zoveel mogelijk allelen worden behouden in de populatie zodat deze na verloop van tijd nog steeds beschikbaar zijn. Bij truncatieselectie daarentegen zorgen enerzijds predictiefouten en anderzijds het verlies aan informatie door GEBV's te gebruiken (de GEBV is een som van vele variabelen) ervoor dat vaak goede allelen uit de populatie worden verwijderd. Hierdoor zal de maximum



Figuur 2.5: De genetische waarden die worden behaald door middel van beide selectiemethoden worden vergeleken voor tal van ouderlijke selectiegroottes. Voor de scoping methode worden hogere genetische waarden behaald. Een tweede opvallend verschil tussen beide selectiemethoden is dat er veel sneller convergentie optreedt bij truncatieselectie dan bij de scoping methode.

haalbare genetische waarde sneller dalen voor truncatieselectie dan voor de scoping methode (zie Figuur A3 in Bijlagen).

Er is in het algemeen te zien dat het verschil in genetische waarde tussen beide selectiemethoden voor eenzelfde ouderlijke selectiegrootte toeneemt naarmate meer ouders worden geselecteerd. Dit verschil is minder groot bij een ouderlijke selectiegrootte van 150 ten opzichte van een ouderlijke selectiegrootte van 100. Wanneer 500 ouders worden geselecteerd, is te zien dat dit verschil kleiner is dan bij ouderlijke selectiegroottes van 100, 150 en 300. Dat in het algemeen het verschil in genetische waarde tussen beide methoden voor eenzelfde selectiegrootte stijgt naarmate meer ouders worden geselecteerd, kan worden verklaard aan de hand van de genetische variatie. Het verschil in genetische variatie tussen beide selectiemethoden voor eenzelfde aantal geselecteerde ouders neemt toe naarmate meer ouders worden geselecteerd (zie Figuur 2.6 en Tabel 2.2). Ook bij een ouderlijke selectiegrootte van 500 is dit verschil groter dan wanneer minder ouders worden geselecteerd. Het afwijkend verschil in genetische waarde tussen beide selectiemethoden voor een ouderlijke selectiegrootte van 500 kan met andere woorden niet op deze manier verklaard worden. Er is te zien dat de genetische waarde bij een ouderlijke selectiegrootte van 500 voor de scoping methode nog niet is geconvergeerd na 50 veredelingscycli

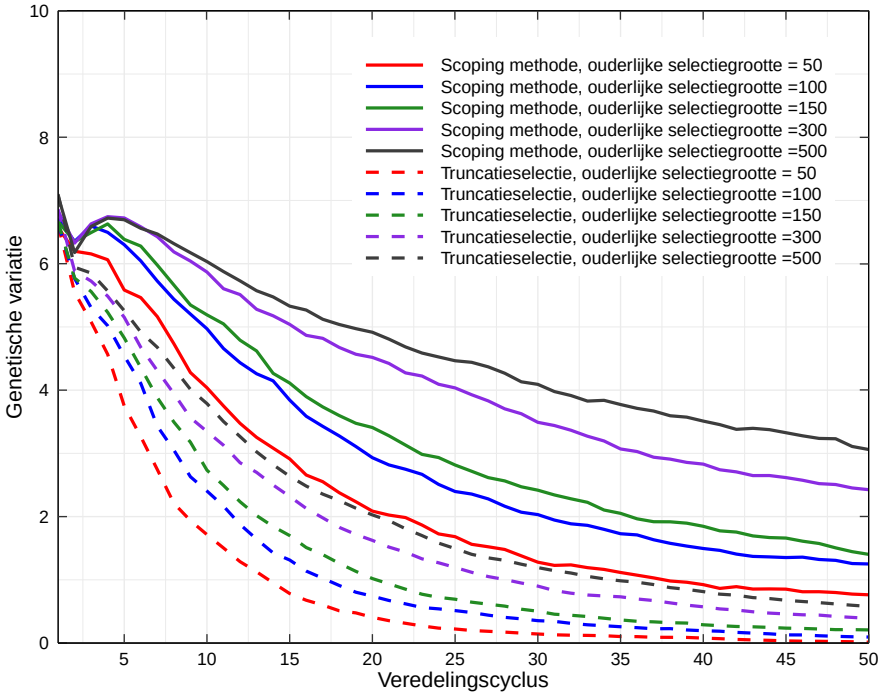
Tabel 2.2: In deze tabel is in de tweede kolom het verschil in genetische waarde weergegeven tussen de scoping methode en truncatieselectie voor verschillende ouderlijke selectiegroottes (OSG) en in de derde kolom de p-waarde die correspondeert met dit verschil. In de vierde kolom is het verschil in genetische variatie tussen beide selectiemethoden weergegeven. Het verschil in genetische waarde tussen beide selectiemethoden neemt over het algemeen toe naarmate er meer ouders worden geselecteerd. Ook het verschil in genetische variatie tussen beide methoden neemt toe naarmate meer ouders worden geselecteerd. De p-waarden wijzen erop dat er voor elke beschouwde ouderlijke selectiegrootte een significant verschil is tussen de genetische waarde die behaald wordt door middel van de scoping methode en de genetische waarde die behaald wordt door middel van truncatieselectie. Zowel de verschillen in genetische waarde als de verschillen in genetische variatie tussen beide selectiemethoden zijn berekend na 50 veredelingscycli.

OSG	**Verschil in gen. waarde**	**p**	**Verschil in gen. var.**
50	14,2 ± 0,7%-punt	<0,001	0,749 ± 0,072
100	16,9 ± 0,6%-punt	<0,001	1,155 ± 0,066
150	16,6 ± 0,8%-punt	<0,001	1,185 ± 0,057
300	17,7 ± 0,7%-punt	<0,001	1,989 ± 0,073
500	16,3 ± 1,0%-punt	<0,001	2,478 ± 0,076

(Figuur 2.5). Dit is echter wel het geval bij dezelfde ouderlijke selectiegrootte voor truncatieselectie, het verschil in genetische waarde tussen beide selectiemethoden zal dus nog toenemen na 50 veredelingscycli. Dit is veel minder het geval wanneer minder ouders worden geselecteerd, er is namelijk te zien dat bij lagere ouderlijke selectiegroottes voor de scoping methode op lange termijn meer convergentie optreedt dan bij een ouderlijke selectiegrootte van 500. Er is te zien dat tussen 35 en 50 veredelingscycli de genetische waarde die hoort bij een selectiegrootte van 500 voor de scoping methode divergeert van de genetische waarden die horen bij lagere selectiegroottes voor de scoping methode. Voor het bekomen van een maximaal verschil in genetische waarde tussen de scoping methode en truncatieselectie zijn meer veredelingscycli nodig wanneer meer ouders worden geselecteerd.

De genetische variatie kan gemakkelijker worden behouden wanneer meer ouders worden geselecteerd waardoor het iets interessanter is om de scoping methode te verkiezen boven truncatieselectie bij lagere ouderlijke selectiegroottes. Het grotere verschil in genetische waarde tussen beide selectiemethoden voor een ouderlijke selectiegrootte van 100 dan voor een ouderlijke selectiegrootte van 150 kan vermoedelijk opnieuw worden toegeschreven aan de milieueffecten en de residuele effecten die naast de genetische waarde in rekening worden gebracht in de fenotypische waarde. Zoals eerder al werd gesteld, kan voor eenzelfde genetische waarde toch een verschillende fenotypische waarde worden bekomen indien deze willekeurige effecten van beide experimenten sterk verschillen. Dit kan zorgen voor een verschillende schatting van de merkereffecten.

Verder is in Figuur 2.5 te zien dat er veel sneller convergentie optreedt bij truncatieselectie in vergelijking met de scoping methode. Dit is een gevolg van de snelle

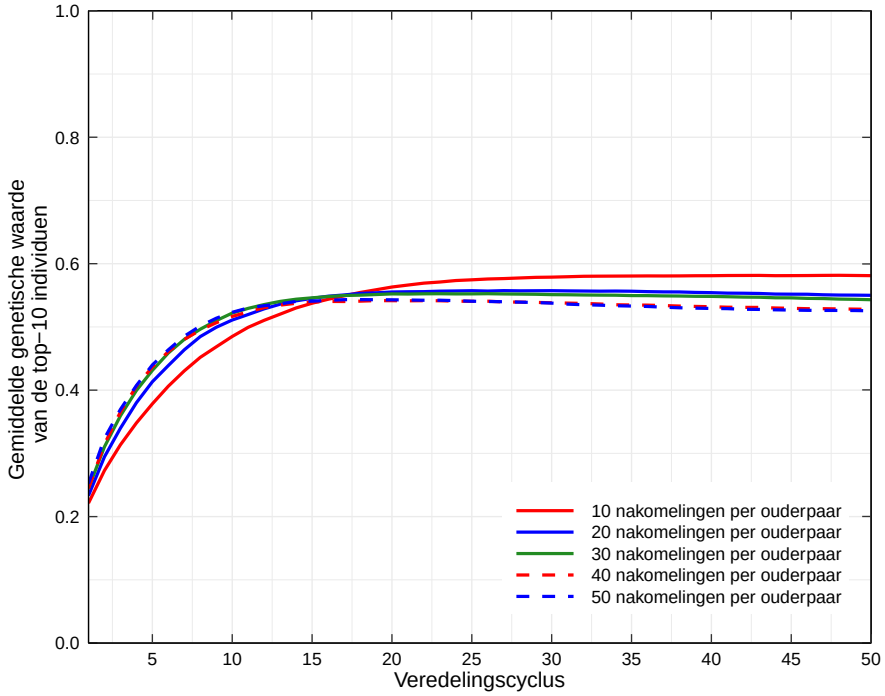


Figuur 2.6: De genetische variatie voor beide selectiemethoden wordt op bovenstaande figuur vergeleken voor verschillende ouderlijke selectiegroottes. Er is te zien dat de genetische variatie steeds (veel) beter behouden wordt bij de scoping methode dan bij truncatieselectie.

afname van de genetische variatie bij truncatieselectie terwijl bij de scoping methode de genetische variatie veel beter wordt behouden (zie Figuur 2.6). Daarnaast valt het op dat bij truncatieselectie de genetische waarde voor sommige ouderlijke selectiegroottes (vooral bij een selectiegrootte van 50) afneemt op lange termijn. Dit komt doordat het model onnauwkeurig wordt; niet langer de beste individuen zullen worden geselecteerd maar een willekeurige selectie van individuen zal plaatsvinden. De individuen waarvan voorspeld wordt dat deze de hoogste genetische waarden hebben, hebben in werkelijkheid niet de hoogste genetische waarden. Door deze willekeurige selectie van individuen zullen ook de allelen willekeurig gefixeerd worden wat zorgt voor een daling van de genetische waarde. Bij hogere ouderlijke selectiegroottes is dit ook het geval, maar in mindere mate doordat meer genetische variatie behouden wordt wanneer meer ouders worden geselecteerd (zie Figuur 2.6).

In Figuur 2.6 is te zien dat er voor de scoping methode initieel een sterke daling is van de genetische variatie en vervolgens direct terug een stijging. Deze stijging is zichtbaar vanaf veredelingscyclus twee. Dit komt doordat bij de scoping methode initieel truncatieselectie wordt gebruikt en ouders pas worden geselecteerd door middel van de scoping methode vanaf veredelingscyclus één.

Ten slotte is op Figuur 2.5 te zien dat er voor de scoping methode pas hogere genetische waarden worden bekomen op langere termijn. Tijdens de eerste veredelingscycli

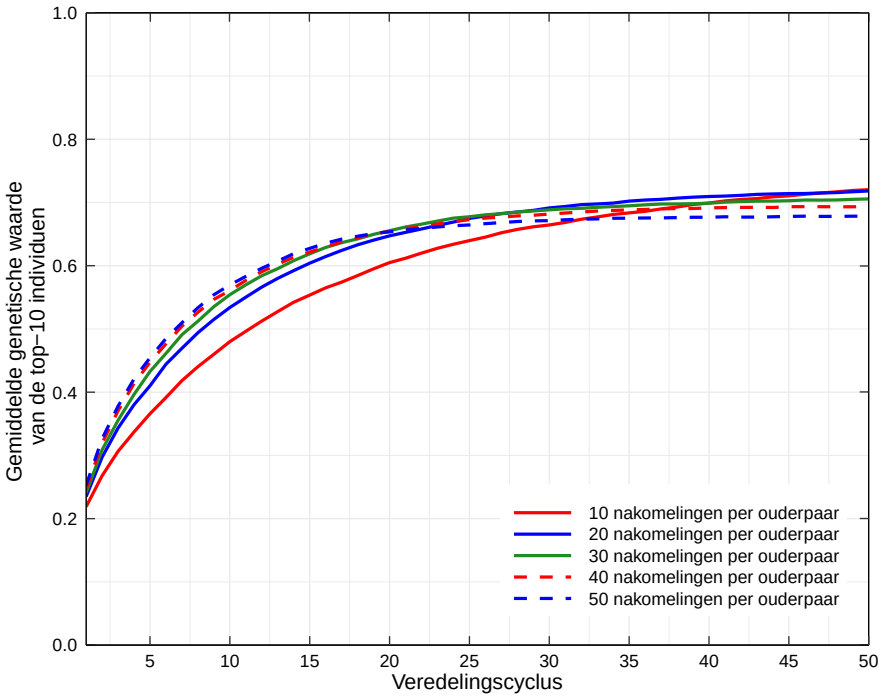


Figuur 2.7: Op bovenstaande grafiek zijn verschillende waarden voorgesteld van het aantal nakomelingen per ouderpaar voor truncatieselectie. Op korte termijn is te zien dat een hogere genetische waarde wordt behaald wanneer er meer nakomelingen zijn per ouderpaar. Op lange termijn daarentegen neemt de genetische waarde toe naarmate het aantal nakomelingen per ouderpaar afneemt.

is het verschil in genetische waarde tussen truncatieselectie en de scoping methode verwaarloosbaar. Er is te zien dat gedurende de eerste zeven veredelingscycli gelijkaardige genetische waarden worden behaald voor beide selectiemethoden. Nadat ongeveer zeven cycli doorlopen zijn, wordt het verschil in genetische waarde tussen beide selectiemethoden duidelijk groter. De scoping methode heeft met andere woorden tijd nodig om zich te kunnen onderscheiden van truncatieselectie. Uit de figuur kan ook worden afgeleid dat, ongeacht de ouderlijke selectiegrootte, de scoping methode zich steeds rond hetzelfde punt (na eenzelfde aantal veredelingscycli) onderscheidt van truncatieselectie.

Aantal nakomelingen per kruising

Een tweede parameter die werd gevarieerd binnen het model RR-GBLUP is het aantal nakomelingen per ouderpaar. Ook de invloed van deze parameter op de genetische waarde, meer bepaald op de gemiddelde genetische waarde van de top-10 individuen van de populatie, wordt onderzocht. In Figuren 2.7 en 2.8 zijn de genetische waarden weergegeven die behaald worden voor verschillende waarden van deze parameter voor respectievelijk truncatieselectie en de scoping methode.



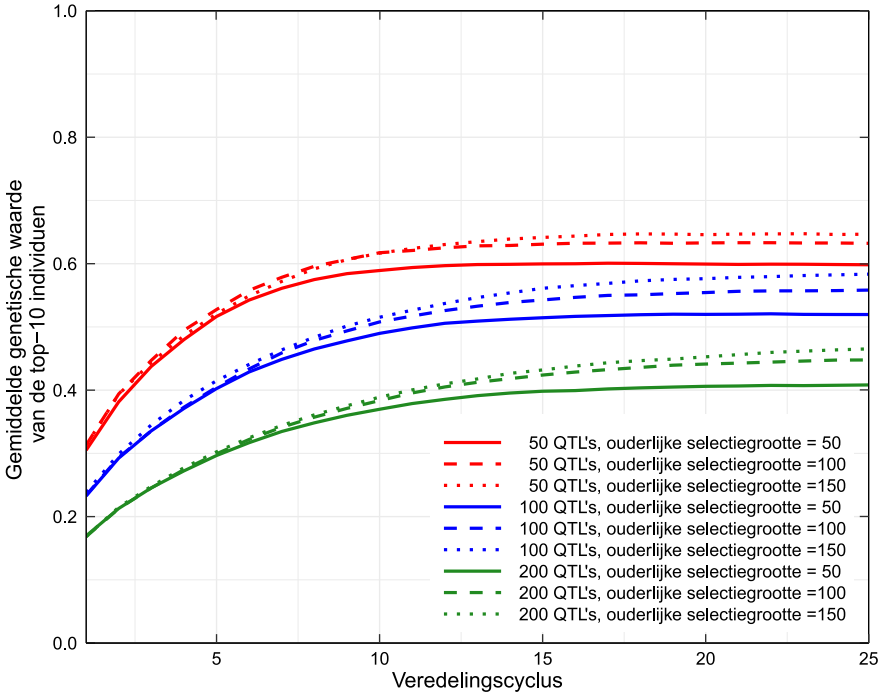
Figuur 2.8: De genetische waarden die behaald worden voor verschillende aantallen nakomelingen per ouderpaar voor de scoping methode zijn weergegeven op bovenstaande grafiek. Net zoals bij truncatieselectie worden hogere genetische waarden behaald voor meer nakomelingen per ouderpaar op korte termijn terwijl dit op lange termijn net omgekeerd is. Het punt vanaf wanneer een hogere genetische waarde wordt behaald voor minder nakomelingen per ouderpaar ligt duidelijk later (na meer veredelingscycli) bij de scoping methode dan bij truncatieselectie.

Voor beide selectiemethoden is te zien dat na het doorlopen van 50 veredelingscycli de genetische waarde hoger is naarmate het aantal nakomelingen per ouderpaar kleiner is. Op korte termijn is net het omgekeerde te zien, de genetische waarde die behaald wordt is groter naarmate er meer nakomelingen voortkomen uit één ouderpaar. Beide waarnemingen zijn in lijn met de verwachtingen, wanneer er minder nakomelingen zijn per ouderpaar is de kans groter dat de nakomelingen die worden geselecteerd voor de volgende cyclus uit verschillende ouderparen komen waardoor de genetische variatie groter is (zie Figuren A4 en A5 in Bijlagen) en op lange termijn hogere genetische waarden kunnen worden bekomen voor minder nakomelingen per ouderpaar. Wanneer daarentegen het aantal nakomelingen per ouderpaar groot is, kan er op korte termijn een hogere genetische waarde worden bekomen doordat meer individuen van de allerbeste ouderparen kunnen worden geselecteerd voor de volgende veredelingscyclus. De beste ouders worden meerdere malen met elkaar gekruist waardoor de kans groter is dat de gewenste allelen van beide ouders aan de nakomelingen worden doorgegeven.

Het aantal veredelingscycli dat dient doorlopen te worden om hogere genetische waarden te bekomen voor een lager aantal nakomelingen per ouderpaar is wel sterk verschillend voor beide methoden. Bij truncatieselectie is dit keerpunt er veel eerder dan bij de scoping methode. Wanneer er meer nakomelingen zijn per ouderpaar worden bij truncatieselectie QTL allelen sneller gefixeerd door het snelle verlies aan genetische variatie (zie Figuur A4 in Bijlagen). Bij truncatieselectie worden de ouders namelijk geselecteerd op basis van de GEBV's. De GEBV's van individuen die dezelfde ouders en een gelijkaardig genotype hebben, zijn ook gelijkaardig. Bij truncatieselectie worden op die manier vaak individuen geselecteerd die met elkaar gerelateerd zijn waardoor de genetische variatie snel afneemt en verschillende QTL's snel gefixeerd worden. Bij de scoping methode daarentegen wordt de genetische variatie over elk ouderpaar gemaximaliseerd. Zodoende wordt vermeden dat twee sterk gerelateerde individuen met elkaar worden gekruist. Hierdoor gebeurt de fixatie van QTL allelen trager en accurater.

Aantal QTL's

Tot hiertoe werd steeds verondersteld dat 100 QTL's een invloed hebben op de welbepaalde eigenschap die onderzocht wordt. Nu wordt nagegaan wat het effect is van het aantal QTL's op de genetische waarde. Verschillende aantallen QTL's worden beschouwd, naast 100 QTL's wordt ook gekeken naar situaties waarbij het aantal QTL's 50 en 200 bedraagt. Ook zal worden onderzocht of het verschil tussen enerzijds verschillende ouderlijke selectiegroottes en anderzijds beide selectiemethoden zich op

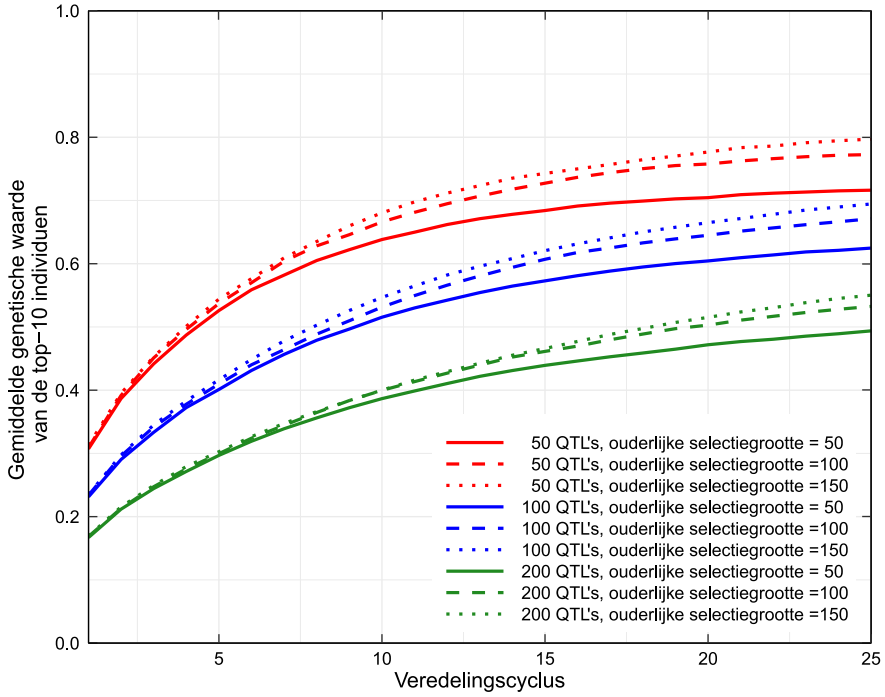


Figuur 2.9: Op deze figuur worden voor truncatieselectie drie verschillende aantallen QTL's beschouwd waarvoor telkens drie verschillende ouderlijke selectiegroottes worden onderzocht. Er worden duidelijk hogere genetische waarden behaald wanneer er minder QTL's zijn. Zoals ook eerder werd gesteld, stijgt de genetische waarde naarmate meer ouders worden geselecteerd. Dit is zowel het geval voor 100 QTL's als voor 50 en 200 QTL's.

eenzelfde manier gedraagt wanneer het aantal QTL's wijzigt. Wegens de beperkte tijd die voorhanden is voor dit onderzoek kan slechts één parameter worden onderzocht. Er wordt gekozen voor de ouderlijke selectiegrootte omdat een wijziging van het aantal nakomelingen per ouderpaar er niet voor zorgt dat de genenpoel verandert. Er is een lagere gevoeligheid van het model voor een dergelijke wijziging. Wanneer het aantal geselecteerde ouders wordt gewijzigd, zal de genenpoel wel veranderen waardoor het interessanter is om deze parameter te beschouwen. Voor elk aantal QTL's dat wordt beschouwd, worden drie ouderlijke selectiegroottes onderzocht; 50, 100 en 150. Verder is het ook belangrijk te vermelden dat voortaan gewerkt wordt met 25 veredelingscycli in plaats van met 50 veredelingscycli zoals eerder het geval was.

In Figuren 2.9 en 2.10 zijn de genetische waarden weergegeven voor telkens drie verschillende ouderlijke selectiegroottes bij 50, 100 en 200 QTL's voor respectievelijk truncatieselectie en de scoping methode.

Er is voor beide selectiemethoden te zien dat er hogere genetische waarden worden bekomen wanneer er minder QTL's zijn. Wanneer minder QTL effecten dienen geschat te worden bij eenzelfde aantal merkers, kunnen de QTL effecten beter worden bena-



Figuur 2.10: In bovenstaande figuur zijn de genetische waarden weergegeven die behaald worden voor verschillende aantallen QTL's en ouderlijke selectiegroottes voor de scoping methode. Er worden drie verschillende aantallen QTL's beschouwd en voor elk aantal QTL's worden drie verschillende ouderlijke selectiegroottes bekeken. Ook hier kan worden gezien dat er hogere genetische waarden worden bekomen voor minder QTL's. Voor elk beschouwd aantal QTL's worden hogere genetische waarden behaald wanneer meer ouders worden geselecteerd.

derd en zullen minder negatieve QTL allelen gefixeerd worden. Zodoende kunnen hogere genetische waarden worden bekomen bij minder QTL's. Het aantal merkers is echter wel niet exact hetzelfde voor alle beschouwde aantallen QTL's aangezien de QTL's willekeurig gekozen worden uit 1590 biallelische SNP loci en de overige loci gebruikt worden als merkers. Wanneer er bijvoorbeeld 50 QTL's zijn in plaats van 100, zijn er 50 extra merkers beschikbaar. Dit is een extra factor die ervoor zorgt dat er hogere genetische waarden worden behaald wanneer er minder QTL's zijn.

Daarnaast is ook zowel voor truncatieselectie als voor de scoping methode te zien dat de genetische waarde binnen eenzelfde aantal QTL's toeneemt naarmate meer ouders worden geselecteerd. De verschillen in genetische waarde tussen de beschouwde ouderlijke selectiegroottes binnen eenzelfde aantal QTL's zijn voor zowel truncatieselectie als de scoping methode weergegeven in Tabel 2.3. De verschillen in genetische waarde tussen ouderlijke selectiegroottes nemen sneller af naarmate er meer QTL's zijn voor de scoping methode dan voor truncatieselectie. Op basis hiervan kan afgeleid worden dat het verschil tussen de scoping methode en truncatieselectie kleiner wordt naarmate er meer QTL's zijn. Het lijkt daarom interessant om de verschillen in genetische waarde tussen de scoping methode en truncatieselectie te bekijken voor verschillende aantallen QTL's en ouderlijke selectiegroottes (zie Tabel 2.4).

Tabel 2.3: In deze tabel worden de verschillen in genetische waarde tussen ouderlijke selectiegroottes (OSG's) binnen eenzelfde aantal QTL's weergegeven. Dit gebeurt zowel voor de scoping methode als voor truncatieselectie. Deze verschillen worden beschouwd na 25 veredelingscycli.

Aantal QTL's	**Selectiemethode**	**OSG's**	**Verschil in gen. waarde**
50	Truncatie	50 en 100	3,5 ± 1,0%-punt
50	Truncatie	100 en 150	1,4 ± 1,0%-punt
50	Scoping	50 en 100	5,6 ± 0,8%-punt
50	Scoping	100 en 150	2,4 ± 0,8%-punt
100	Truncatie	50 en 100	4,1 ± 0,7%-punt
100	Truncatie	100 en 150	2,6 ± 0,6%-punt
100	Scoping	50 en 100	4,5 ± 0,6%-punt
100	Scoping	100 en 150	2,3 ± 0,6%-punt
200	Truncatie	50 en 100	4,0 ± 0,6%-punt
200	Truncatie	100 en 150	1,8 ± 0,5%-punt
200	Scoping	50 en 100	3,8 ± 0,5%-punt
200	Scoping	100 en 150	1,9 ± 0,5%-punt

Er is te zien dat het verschil in genetische waarde tussen de scoping methode en truncatieselectie wel degelijk afneemt naarmate er meer QTL's zijn en dit voor alle drie de beschouwde ouderlijke selectiegroottes. Er kan met andere woorden worden gesteld dat de scoping methode performanter wordt naarmate het aantal QTL's afneemt, al worden voor de scoping methode bij een hoger aantal QTL's wel nog steeds (veel) hogere genetische waarden behaald op lange termijn dan voor truncatieselec-

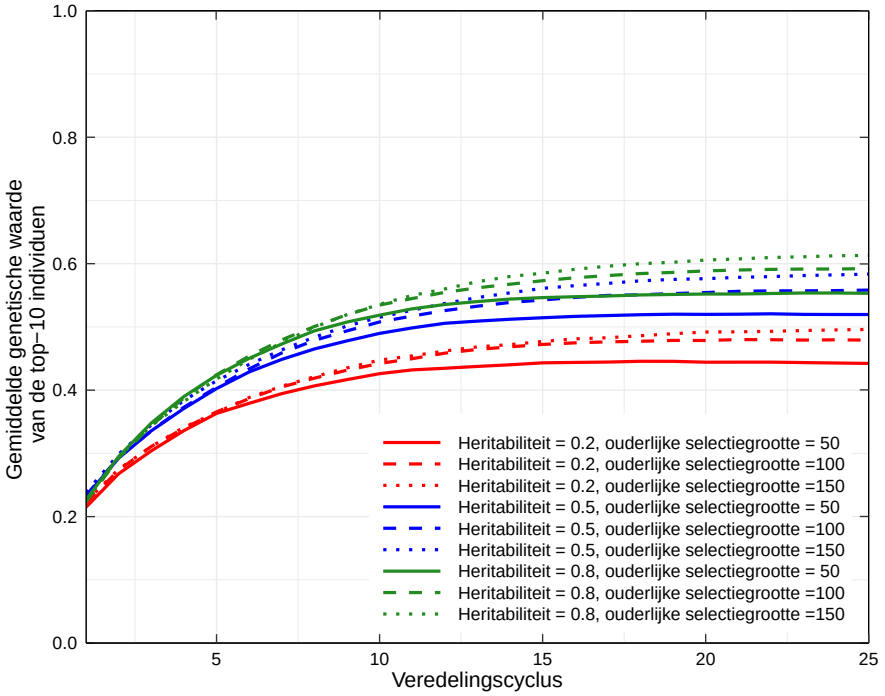
Tabel 2.4: Deze tabel geeft de verschillen weer in genetische waarde tussen de scoping methode en truncatieselectie voor eenzelfde ouderlijke selectiegrootte (OSG) en eenzelfde aantal QTL's. Drie verschillende ouderlijke selectiegroottes en drie verschillende aantallen QTL's worden beschouwd. Deze verschillen worden steeds berekend na 25 veredelingscycli.

Aantal QTL's	**OSG**	**Verschil in gen. waarde**
50	50	11,9 ± 0,9%-punt
50	100	14,0 ± 1,0%-punt
50	150	15,0 ± 0,9%-punt
100	50	10,6 ± 0,7%-punt
100	100	11,5 ± 0,6%-punt
100	150	11,2 ± 0,7%-punt
200	50	8,6 ± 0,5%-punt
200	100	8,3 ± 0,6%-punt
200	150	8,4 ± 0,6%-punt

tie. Het is hierbij wel belangrijk te vermelden dat de verschillen tussen beide selectiemethoden steeds beschouwd worden na 25 veredelingscycli. De genetische waarde die behaald wordt voor truncatieselectie is na 25 veredelingscycli reeds in grote mate geconvergeerd. Dit is echter veel minder het geval voor de scoping methode. Vermoedelijk zal de genetische waarde die behaald wordt voor de scoping methode op lange termijn nog verder toenemen waardoor ook het verschil tussen de scoping methode en truncatieselectie vermoedelijk nog verder zal toenemen. Dit is vooral het geval bij een hoger aantal QTL's aangezien de genetische waarde voor de scoping methode bij meer QTL's minder convergentie vertoont dan de genetische waarde voor de scoping methode bij minder QTL's. Dat het verschil in genetische waarde tussen beide selectiemethoden op lange termijn vermoedelijk nog zal toenemen voornamelijk bij meer QTL's zou eenvoudig achterhaald kunnen worden door de simulatietijd op te voeren. Door de beperkte tijd die beschikbaar is voor dit onderzoek, is dit niet uitgevoerd kunnen worden.

Heritabiliteit

Vervolgens wordt op een analoge wijze als bij het aantal QTL's onderzocht wat het effect is van verschillende heritabiliteiten op de genetische waarde. Er wordt onderzocht wat het effect is van de heritabiliteit enerzijds op het verschil in genetische waarde tussen ouderlijke selectiegroottes en anderzijds op het verschil in genetische waarde tussen beide selectiemethoden. De verschillen in genetische waarde worden steeds berekend na het doorlopen van 25 veredelingscycli. Tot hiertoe werd steeds verondersteld dat de heritabiliteit gelijk is aan 0,5. Nu zullen ook heritabiliteiten van 0,2 en 0,8 worden onderzocht. Voor elke heritabiliteit worden dezelfde drie ouderlijke selectiegroottes beschouwd als voor de verschillende aantallen QTL's, namelijk 50, 100 en 150. De genetische waarden die behaald worden voor deze heritabiliteiten en

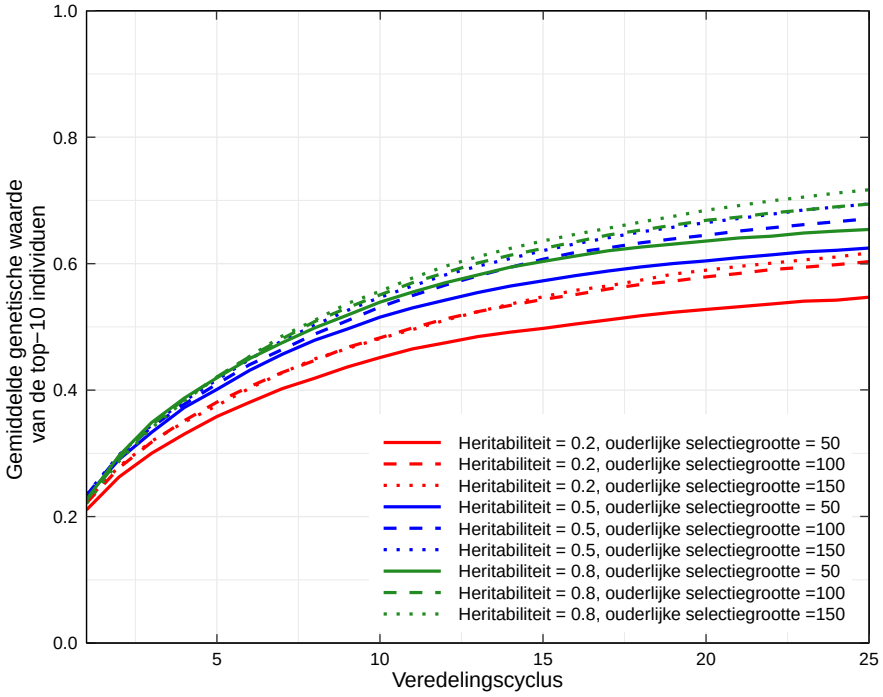


Figuur 2.11: Deze grafiek geeft de genetische waarden weer die behaald worden voor verschillende heritabiliteiten voor truncatieselectie. Binnen elke heritabiliteit worden drie ouderlijke selectiegroottes beschouwd. Er is te zien dat de genetische waarden die worden behaald groter zijn wanneer er een hogere heritabiliteit is. Daarnaast is de genetische waarde binnen elke beschouwde heritabiliteit groter wanneer meer ouders worden geselecteerd.

ouderlijke selectiegroottes zijn weergegeven in Figuren 2.11 en 2.12 voor respectievelijk truncatieselectie en de scoping methode.

Voor beide selectiemethoden is te zien dat de genetische waarden hoger zijn wanneer de heritabiliteit (h^2) toeneemt. Deze heritabiliteit is de verhouding van de variantie ten gevolge van de additiviteit tot de totale fenotypische variantie (zie Formule (1.4)) (Audenaert, 2018). Als de heritabiliteit met andere woorden groter is, wordt meer van de totale fenotypische variatie verklaard door merkers en is de invloed van de omgeving minder groot. Hierdoor kan het fenotype beter worden voorspeld door de merkers wat resulteert in hogere genetische waarden.

In Tabel 2.5 zijn de verschillen in genetische waarde tussen de beschouwde ouderlijke selectiegroottes weergegeven voor verschillende heritabiliteiten. Een duidelijke trend in deze verschillen over de beschouwde heritabiliteiten heen is er niet te zien, al lijken de verschillen in genetische waarde tussen de ouderlijke selectiegroottes wel af te nemen voor de scoping methode wanneer de heritabiliteit groter wordt. Daarom lijkt het interessant om ook eens de verschillen in genetische waarde tussen beide selectiemethoden te beschouwen voor verschillende heritabiliteiten. Deze verschillen zijn te zien in Tabel 2.6. De waarden in deze tabel tonen dat het verschil tussen beide



Figuur 2.12: Verschillende heritabiliteiten worden op deze figuur weergegeven waarbij de ouders worden geselecteerd door middel van de scoping methode. Voor elke heritabiliteit worden drie ouderlijke selectiegroottes beschouwd. Bij een grotere heritabiliteit worden hogere genetische waarden bekomen. Naarmate meer ouders worden geselecteerd, zijn de genetische waarden groter. Dit is voor alle drie de beschouwde heritabiliteiten het geval.

selectiemethoden kleiner wordt wanneer de heritabiliteit toeneemt. Er kan op die manier worden gesteld dat de hoogste genetische waarden voor de scoping methode in vergelijking met truncatieselectie worden bekomen bij een lagere heritabiliteit.

Tabel 2.5: Deze tabel geeft de verschillen weer in genetische waarde tussen ouderlijke selectiegroottes (OSG's) voor eenzelfde heritabiliteit. Drie heritabiliteiten en drie ouderlijke selectiegroottes worden onderzocht. Deze verschillen worden berekend zowel voor truncatieselectie als voor de scoping methode. Deze verschillen worden beschouwd na 25 veredelingscycli.

Heritabiliteit	**Selectiemethode**	**OSG's**	**Verschil in gen. waarde**
0,2	Truncatie	50 en 100	3,7 ± 0,8%-punt
0,2	Truncatie	100 en 150	1,5 ± 0,8%-punt
0,2	Scoping	50 en 100	5,6 ± 0,7%-punt
0,2	Scoping	100 en 150	1,4 ± 0,7%-punt
0,5	Truncatie	50 en 100	4,1 ± 0,7%-punt
0,5	Truncatie	100 en 150	2,6 ± 0,6%-punt
0,5	Scoping	50 en 100	4,5 ± 0,6%-punt
0,5	Scoping	100 en 150	2,3 ± 0,6%-punt
0,8	Truncatie	50 en 100	4,0 ± 0,7%-punt
0,8	Truncatie	100 en 150	2,0 ± 0,7%-punt
0,8	Scoping	50 en 100	4,0 ± 0,6%-punt
0,8	Scoping	100 en 150	2,3 ± 0,6%-punt

Tabel 2.6: Het verschil in genetische waarde tussen de scoping methode en truncatieselectie voor verschillende heritabiliteiten en verschillende waarden van de ouderlijke selectiegrootte (OSG) is weergegeven in deze tabel. Dit verschil wordt beschouwd nadat 25 veredelingscycli zijn afgerond.

Heritabiliteit	**OSG**	**Verschil in gen. waarde**
0,2	50	10,5 ± 0,8%-punt
0,2	100	12,4 ± 0,8%-punt
0,2	150	12,1 ± 0,8%-punt
0,5	50	10,6 ± 0,7%-punt
0,5	100	11,5 ± 0,6%-punt
0,5	150	11,2 ± 0,7%-punt
0,8	50	9,9 ± 0,7%-punt
0,8	100	10,2 ± 0,7%-punt
0,8	150	10,4 ± 0,7%-punt

Bij een lagere heritabiliteit is het moeilijker om de merkereffecten in te schatten doordat er meer ruis is (bijvoorbeeld wanneer de heritabiliteit 0,2 bedraagt, kan 80% van de fenotypische variatie niet verklaard worden door de merkers). Bij meer ruis wijken de GEBV's sterker af van de genetische waarden waardoor de kans groter is dat er foute individuen worden geselecteerd. Dergelijke fouten zijn makkelijker te corrigeren bij de scoping methode aangezien beide allelen van de merkers behouden worden waardoor deze later indien nodig kunnen worden geïntroduceerd in de beste kandidaten. Bij truncatieselectie kunnen dergelijke fouten minder goed worden gecorrigeerd aangezien de beste individuen worden geselecteerd zonder aandacht te hebben voor het behoud van zoveel mogelijk allelen. Hierdoor gaan snel veel allelen verloren.

Op die manier is het verschil in genetische waarde tussen de scoping methode en truncatieselectie groter wanneer de heritabiliteit lager is.

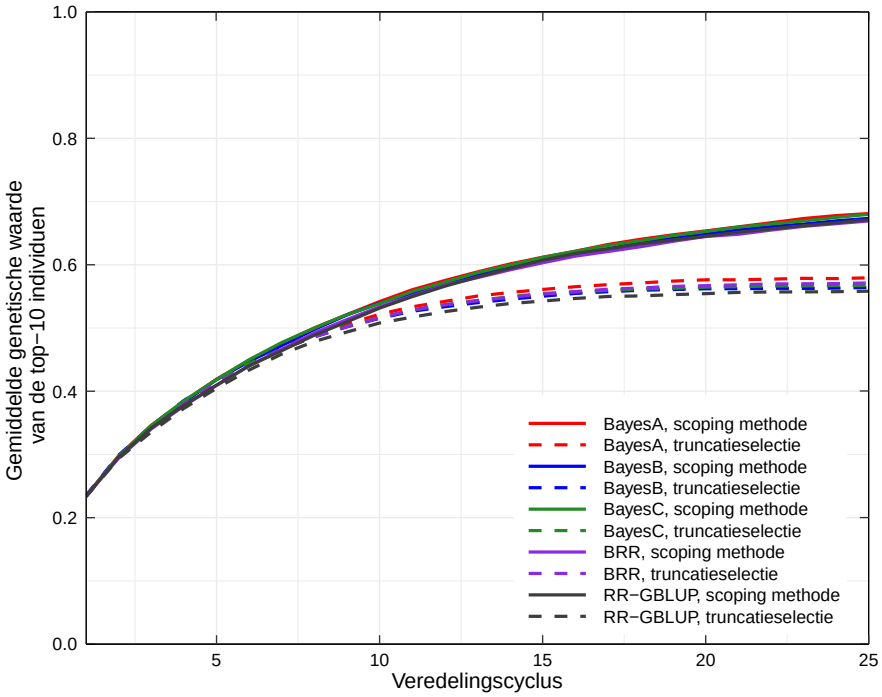
2.2.2 Vergelijking tussen verschillende modellen

Tot hiertoe werd steeds gewerkt met het model RR-GBLUP. Er bestaan echter nog tal van andere predictiemodellen, zoals Bayesiaanse modellen. Vier verschillende Bayesiaanse modellen worden vergeleken met het model RR-GBLUP in verschillende situaties. Er worden verschillende ouderlijke selectiegroottes beschouwd alsook verschillende aantallen QTL's en verschillende heritabiliteiten. De vier Bayesiaanse modellen die worden beschouwd, zijn BayesA, BayesB, BayesC en BRR. Deze modellen werden reeds uitgebreid besproken in Subsectie 2.1.10. De verschillende modellen worden steeds vergeleken na 25 veredelingscycli. Voor elk model worden de ouders geselecteerd aan de hand van truncatieselectie of aan de hand van de scoping methode. Eerst wordt de invloed van elk model bestudeerd onder standaardwaarden van de simulatieparameters (zie Tabel 2.1). In Figuur 2.13 zijn de genetische waarden weergegeven voor de verschillende modellen waarbij de ouderlijke selectie gebeurt door middel van truncatieselectie of de scoping methode.

Er is te zien dat de genetische waarden die behaald worden voor de verschillende modellen voor eenzelfde ouderlijke selectiemethode zeer dicht bij elkaar zijn gelegen. Er is met andere woorden, onder standaardwaarden van de parameters, amper verschil in de genetische waarden die behaald worden voor de verschillende modellen. Dit is zowel het geval voor truncatieselectie als voor de scoping methode. Er is wel een duidelijk verschil tussen de genetische waarden die behaald worden wanneer de ouders worden geselecteerd via truncatieselectie of de scoping methode. Aangezien de curves in Figuur 2.13 dicht bij elkaar zijn gelegen, is het soms moeilijk om de genetische waarden die horen bij de verschillende modellen van elkaar te onderscheiden. In Tabel 2.7 zijn daarom de genetische waarden weergegeven die worden bekomen voor de vijf verschillende modellen en telkens twee ouderlijke selectiemethoden na het doorlopen van 25 veredelingscycli.

Aan de hand van ANOVA-analyses worden geen significante verschillen gevonden tussen de modellen wanneer truncatieselectie ($p=0,528$) of de scoping methode ($p=0,623$) wordt gebruikt als ouderlijke selectiemethode.

Vervolgens worden de verschillende modellen ook vergeleken in tal van andere situaties. Ouderlijke selectiegroottes van 50 en 150 worden beschouwd alsook situaties waarin het aantal QTL's gelijk is aan 50 en 200. Ook heritabiliteiten van 0,2 en 0,8 worden onderzocht. Hierbij neemt telkens slechts één parameter een waarde aan



Figuur 2.13: Op deze figuur zijn de genetische waarden voorgesteld die worden behaald voor de vijf verschillende modellen over 25 veredelingscycli. Voor elk model worden telkens beide ouderlijke selectiemethoden eens gebruikt. Op de grafiek is te zien dat de genetische waarden die behaald worden voor de verschillende modellen voor eenzelfde ouderlijke selectiemethode niet veel van elkaar verschillen. Er is echter wel voor elk model een groot verschil tussen de genetische waarden die behaald worden aan de hand van truncatieselectie en de genetische waarden die behaald worden aan de hand van de scoping methode.

Tabel 2.7: In deze tabel wordt de genetische waarde weergegeven die wordt bekomen na 25 veredelingscycli voor de vijf verschillende modellen. In de bovenste helft van de tabel worden voor alle modellen de ouders geselecteerd aan de hand van truncatieselectie, in de onderste helft aan de hand van de scoping methode. Deze modellen werden steeds beschouwd onder standaardwaarden van de simulatieparameters (ouderlijke selectiegrootte van 100, 100 QTL's, heritabiliteit 0,5). Er is te zien dat er binnen elke selectiemethode niet veel verschil is tussen de genetische waarden die behaald worden voor de verschillende modellen. Voor alle modellen is er echter wel een duidelijk verschil tussen de genetische waarde die behaald wordt wanneer ouders worden geselecteerd aan de hand van truncatieselectie en wanneer ouders worden geselecteerd aan de hand van de scoping methode.

Model	Selectiemethode	Genetische waarde
BayesA	Truncatie	57,9 ± 8,3%-punt
BayesB	Truncatie	56,4 ± 8,5%-punt
BayesC	Truncatie	56,7 ± 7,9%-punt
BRR	Truncatie	57,1 ± 8,1%-punt
RR-GBLUP	Truncatie	55,8 ± 8,0%-punt
BayesA	Scoping	68,1 ± 6,9%-punt
BayesB	Scoping	67,3 ± 7,1%-punt
BayesC	Scoping	68,0 ± 7,0%-punt
BRR	Scoping	66,9 ± 6,8%-punt
RR-GBLUP	Scoping	67,1 ± 7,7%-punt

die verschillend is van de standaardwaarde. In elk van deze situaties worden geen significante verschillen gevonden tussen de modellen. Dit is steeds het geval voor beide ouderlijke selectiemethoden. De bijhorende p-waarden zijn terug te vinden in Tabel A7 in Bijlagen.

Er kan dus in het algemeen worden gesteld dat de verschillen in de gemiddelde genetische waarde van de top-10 individuen tussen de verschillende modellen minimaal zijn. Hieruit kan worden geconcludeerd dat in de beschouwde situaties de genetische vooruitgang van een veredelingspopulatie niet afhangt van de keuze tussen RR-GBLUP of één van de vier beschouwde Bayesiaanse modellen. Ook in eerder onderzoek werd al gesteld dat er geen significante verschillen zijn tussen enerzijds RR-GBLUP en anderzijds Bayesiaanse modellen (de Azevedo Peixoto et al., 2017).

RR-GBLUP is echter vaak stabieler en heeft een minder lange rekentijd in vergelijking met de Bayesiaanse modellen (Wang et al., 2018; de Azevedo Peixoto et al., 2017). Daardoor lijkt het gebruik van RR-GBLUP toch iets interessanter te zijn dan het gebruik van de Bayesiaanse modellen.

2.2.3 Praktisch aspect

In het laatste deel van dit onderzoek wordt eens gekeken naar hoe het er in de praktijk aan toegaat. Hiervoor wordt ook een interview gedaan met dr. Steven Maenhout, oprichter van Progeno. Progeno is een spin-off bedrijf van de Universiteit van Gent dat

veredelaars adviseert. Dr. Maenhout vertelde dat de scoping methode, die centraal staat in dit onderzoek, niet wordt gebruikt in de praktijk doordat het enerzijds een heel recent ontwikkelde methode is en anderzijds (en vooral) omdat veredelaars zelden op lange termijn denken. De scoping methode biedt vooral voordeel ten opzichte van truncatieselectie op lange termijn. Op korte termijn is het verschil tussen beide selectiemethoden gering en daarom is de scoping methode minder interessant voor veredelaars.

Uit dit onderzoek bleek dat het verschil in genetische waarde tussen RR-GBLUP en de Bayesiaanse modellen niet significant is. Beide modellen zouden dus kunnen worden aangewend, al leek het wel iets interessanter om RR-GBLUP te gebruiken omwille van enerzijds de hogere stabiliteit en anderzijds de minder lange rekentijd in vergelijking met de Bayesiaanse modellen. Ook in de praktijk blijkt de voorkeur naar RR-GBLUP te gaan. Bij Progeno wordt namelijk ook meestal gebruikgemaakt van dit model. Hiervoor worden dezelfde twee redenen gegeven: de beperkte rekentijd en de stabiliteit van het resultaat. Bij Progeno wordt naar schatting in 95% van de gevallen gebruikgemaakt van RR-GBLUP.

Ten slotte is ook de trainingspopulatie een belangrijk onderdeel van GS. De grootte ervan werd niet gevarieerd tijdens dit onderzoek maar er kan wel worden gesteld dat een toename van de grootte van de trainingspopulatie zorgt voor een hogere predictieaccuraatheid maar daarnaast ook voor een hogere kost, de balans tussen beide dient dus gemaakt te worden. Dr. Maenhout adviseert om in de praktijk te werken met een trainingspopulatie waarvan de grootte minstens 400 bedraagt.

Het belangrijkste voor veredelaars is dus de korte termijn. Er wordt gekeken naar hoe de beste individuen op korte termijn kunnen worden bekomen. Het doorlopen van veredelingscycli neemt namelijk heel wat tijd en geld in beslag. Ondanks dat de genetische waarden die op lange termijn worden bekomen voor de scoping methode veel hoger zijn dan voor truncatieselectie zal een bedrijf niet geneigd zijn te kiezen voor de scoping methode omdat de winst op korte termijn door gebruik van de scoping methode niet hoger zal zijn. Kiezen voor een (iets) lagere winst op korte termijn om op die manier deze op lange termijn te verhogen, is iets wat een (veredelings)bedrijf niet snel zal doen; de winst op korte termijn primeert.

Ondanks dat voor truncatieselectie duidelijk hogere genetische waarden worden bekomen op lange termijn wanneer gekozen wordt voor minder nakomelingen per ouderpaar lijkt het toch interessanter voor veredelaars om te kiezen voor meer nakomelingen per ouderpaar omdat op korte termijn hiervoor duidelijk hogere genetische waarden worden bekomen (zie Figuur 2.7). Kiezen voor 30 nakomelingen per ouderpaar lijkt een goeie balans tussen beide te zijn. Er is namelijk op korte termijn te zien

dat er niet veel verschil is tussen 30, 40 en 50 nakomelingen per ouderpaar terwijl er wel een groter verschil is met 20 nakomelingen per ouderpaar en zeker met 10 nakomelingen per ouderpaar. Op lange termijn daarentegen is er wel een duidelijk verschil tussen enerzijds 40 en 50 nakomelingen per ouderpaar en anderzijds 20 en 30 nakomelingen per ouderpaar. Ook voor de scoping methode lijkt het het meest interessant om te kiezen voor 30 nakomelingen per ouderpaar omdat er voor deze waarde relatief hoge genetische waarden worden verkregen op korte termijn alsook op lange termijn (zie Figuur 2.8).

Zoals eerder werd aangetoond, onderscheidt de scoping methode zich sneller (na minder veredelingscycli) van truncatieselectie bij een lager aantal QTL's alsook bij een lagere heritabiliteit. Hierdoor lijkt het voor veredelaars iets interessanter om de scoping methode te verkiezen boven truncatieselectie wanneer de eigenschap waarvoor interesse is, bepaald wordt door minder QTL's of wanneer er sprake is van een lagere heritabiliteit.

2.3 Algemeen besluit

In elke situatie die werd onderzocht in dit onderzoek worden op lange termijn hogere genetische waarden bekomen wanneer de ouders worden geselecteerd aan de hand van de scoping methode dan wanneer deze worden geselecteerd aan de hand van truncatieselectie. Op korte termijn daarentegen is het verschil in genetische waarde tussen beide selectiemethoden gering. In de praktijk wordt nog steeds truncatieselectie toegepast, voor veredelaars primeert de winst op korte termijn. Deze selectiemethode zou toch moeten vervangen worden door betere alternatieven zoals de scoping methode.

Een verhoging van de ouderlijke selectiegrootte zorgt ervoor dat hogere genetische waarden kunnen worden behaald op lange termijn. Vanaf een welbepaalde ouderlijke selectiegrootte convergeert de genetische waarde op lange termijn naar hetzelfde lokale optimum. Het verder laten toenemen van de ouderlijke selectiegrootte heeft geen invloed meer op de genetische waarde, maar verhoogt wel de totale kost van de veredelingspopulatie. Voor truncatieselectie bedraagt deze ouderlijke selectiegrootte 400, terwijl deze voor de scoping methode gelegen is tussen 200 en 350. Het is hierbij belangrijk om op te merken dat deze waarden van de ouderlijke selectiegrootte specifiek zijn voor de gebruikte data.

Op lange termijn worden hogere genetische waarden bekomen wanneer er minder nakomelingen zijn per ouderpaar terwijl op korte termijn net het omgekeerde wordt bekomen. Veredelaars zullen dus geneigd zijn meer nakomelingen per ouderpaar

te verkiezen. Het lijkt echter interessanter om ook hier in te zetten op het behoud van genetische variatie om op die manier ook hoge genetische waarden te kunnen verkrijgen op lange termijn. Kiezen voor 30 nakomelingen per ouderpaar lijkt een goede balans te zijn tussen beide.

Wanneer een eigenschap van een individu bepaald wordt door een klein aantal QTL's kunnen hogere genetische waarden worden behaald dan wanneer een eigenschap bepaald wordt door een groter aantal QTL's. Er kunnen ook hogere genetische waarden worden behaald wanneer de heritabiliteit groter is. Het verschil in genetische waarde tussen de scoping methode en truncatieselectie is groter bij een lager aantal QTL's en bij een lagere heritabiliteit. Het is met andere woorden nog net iets interessanter om de scoping methode te verkiezen boven truncatieselectie wanneer een eigenschap bepaald wordt door een klein aantal QTL's of wanneer de heritabiliteit laag is. Hierbij dient wel te worden benadrukt dat ook wanneer de heritabiliteit hoog is of wanneer het aantal QTL's groot is, voor de scoping methode nog steeds veel hogere genetische waarden worden bekomen dan voor truncatieselectie. De scoping methode zou steeds verkozen moeten worden boven truncatieselectie.

Er is niet veel verschil tussen de genetische waarden die worden behaald wanneer RR-GBLUP wordt gebruikt en deze die worden behaald wanneer één van de vier beschouwde Bayesiaanse modellen wordt gebruikt (BayesA, BayesB, BayesC of BRR). Toch lijkt RR-GBLUP het meest interessante model om te gebruiken omwille van enerzijds de beperkte rekentijd en anderzijds de hoge stabiliteit van RR-GBLUP.

BIBLIOGRAFIE

A. Allier, S. Teyssèdre, C. Lehermeier, L. Moreau, and A. Charcosset. Optimized breeding strategies to harness genetic resources with different performance levels. *BMC Genomics*, 21(349):1–16, 2020.

K. Audenaert. *Cursus Genetica*. Universiteit Gent, Gent, Belgium, 2018.

M.A. Bargougui. Genetic analysis of barley (Hordeum vulgare L.) grain yield components. *Journal of New Sciences*, 31(8):1794–1799, 2016.

R. Bernardo. Parental selection, number of breeding populations, and size of each population in inbred development. *Theoretical and Applied Genetics*, 107:1252–1256, 2003.

R. Bernardo. Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Science*, 48(5):1649–1664, 2008.

A.M. Brown. A new software for carrying out one-way ANOVA post hoc tests. *Computer Methods and Programs in Biomedicine*, 79(1):89–95, 2005.

M.P.L. Calus, T.H.E. Meuwissen, de Roos A.P.W., and R.F. Veerkamp. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, 178(1): 553–561, 2008.

W.A. Cowling, L. Li, K.H.M. Siddique, M. Henryon, P. Berg, R.G. Banks, and B.P. Kinghorn. Evolving gene banks: improving diverse populations of crop and exotic germplasm with optimal contribution selection. *Journal of Experimental Botany*, 68(8): 1927–1939, 2016.

D. Cros, B. Tchounke, and L. Nkageu-Nkamba. Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. *Molecular Breeding*, 38(89):1–12, 2018.

L. de Azevedo Peixoto, T.C. Moellers, J. Zhang, A.J. Lorenz, L.L. Bhering, W.D. Beavis, and A.K. Singh. Leveraging genomic prediction to scan germplasm collection for crop improvement. *PLOS ONE*, 12(6), 2017.

Z.A. Desta and R. Ortiz. Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9):592–601, 2014.

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

R. Eshghi and E. Akhundova. Inheritance of some important agronomic traits in hulless barley. *International Journal of Agriculture and Biology*, 12(1):73–76, 2010.

W. Gapare, S. Liu, W. Conaty, Q. Zhu, V. Gillespie, D. Llewellyn, W. Stiller, and I. Wilson. Historical datasets support genomic selection models for the prediction of cotton fiber quality phenotypes across multiple environments. *G3: Genes, Genomes, Genetics*, 8(5):1721–1732, 2018.

T.D. Gauthier. Detecting trends using Spearman's rank correlation coefficient. *Environmental Forensics*, 2(4):359–362, 2001.

H. Godfray, J. Beddington, I. Crute, L. Haddad, D. Lawrence, J. Muir, J. Pretty, S. Robinson, S. Thomas, and C. Toulmin. Food security: the challenge of feeding 9 billion people. *Science*, 327:812–818, 2010.

G. Gorjanc, R.C. Gaynor, and J.M. Hickey. Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, 131:1953–1966, 2018.

D. Habier, R.L. Fernando, Kizilkaya K., and D.J. Garrick. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12(186), 2011.

A. Harfouche, R. Meilan, M. Kirst, M. Morgante, W. Boerjan, M. Sabatti, and G. Mugnozza. Accelerating the domestication of forest trees in a changing world. *Trends in Plant Science*, 17(2):1360–1385, 2012.

S. Hart. Pedigree. `https://www.genome.gov/genetics-glossary/Pedigree`, 2001. Bezocht op: "20-11-2020".

E.L. Heffner, J. Jannink, Iwata H., E. Souza, and M.E. Sorrells. Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Science*, 51: 2597–2606, 2011.

C.R. Henderson. *Applications of linear models in animal breeding*. University of Guelph, Guelph, Ontario, Canada, 1984.

F. Hospital, L. Moreau, F. Lacoudre, A. Charcosset, and A. Gallais. More on the efficiency of marker-assisted selection. *Theoretical and Applied Genetics*, 95:1181–1189, 1997.

J. Jannink, A. Lorenz, and H. Iwata. Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, 9(2):166–177, 2010.

S. Knapp. Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. *Crop Science*, 38:1164–1174, 1998.

R.C. Lacy. Importance of genetic variation to the viability of mammalian populations. *Journal of Mammalogy*, 78(2):320–335, 1997.

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

Y. Li and H.S. Dungey. Expected benefit of genomic selection over forward selection in conifer breeding and deployment. *PLOS ONE*, 13(12), 2018.

S. Luca. *Cursus Statistische Data-analyse*. Universiteit Gent, Gent, Belgium, 2019.

R.J. Mackenzie. One-way vs two-way anova: differences, assumptions and hypotheses. `https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553`, 2018. Bezocht op: "04-03-2021".

B. Markgraf. Advantages & disadvantages of selective breedings. `https://sciencing.com/advantages-disadvantages-selective-breedings-8132953.html`, 2018. Bezocht op: "16-11-2020".

J.W.R. Martini, N. Gao, D.F. Cardoso, V. Wimmer, M. Erbe, R.J.C. Cantet, and H. Simianer. Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics*, 18(3):1–16, 2017.

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

P. McClean. Mendelian genetics: pedigree analysis. `https://www.ndsu.edu/pubweb/ mcclean/plsc431/mendel/mendel9.htm`, 2000. Bezocht op: "02-11-2020".

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

T. Meuwissen, B. Hayes, and M. Goddard. Genomic selection: a paradigm shift in animal breeding. *Animal Frontiers*, 6(1):6–14, 2016.

B. Miller. 16 main advantages and disadvantages of selective breeding. `https://greengarageblog.org/16-main-advantages-and-disadvantages-of-selective-breeding`, 2015. Bezocht op: "09-11-2020".

J. Neyhart, T. Tiede, A Lorenz, and K. Smith. Evaluating methods of updating training data in long-term genomewide selection. *G3: Genes, Genomes, Genetics*, 7:1499–1510, 2017.

P. Pérez and G. de los Campos. Genome-wide regression and prediction with the bglr statistical package. *Genetics*, 198(2):483–495, 2014.

R. Rincent, D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V.M. Rodríguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C-C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset, and L. Moreau. 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, 2012.

C. Robertsen, R. Hjortshøj, and L. Janss. Genomic selection in cereal breeding. *Agronomy*, 9(2):1–16, 2019.

S.T. Skidmore and B. Thompson. Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods*, (45):536–546, 2013.

B. Smith. A comparison of niche construction theory and diet breadth models as explanatory frameworks for the initial domestication of plants and animals. *Journal of Archaeological Research*, 23:215–262, 2015.

A. Sonesson and T. Meuwissen. Testing strategies for genomic selection in aquaculture breeding programs. *Genetics Selection Evolution*, 41(1):1–9, 2009.

D. Vanavermaete, J. Fostier, S. Maenhout, and B. De Baets. Preservation of genetic variation in a breeding population for long-term genetic gain. *G3: Genes, Genomes, Genetics*, 10(8):2753–2762, 2020.

P.M. VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008.

J. Wang, Z. Zhou, Z. Zhang, H. Li, D. Liu, Q. Zhang, P.J. Bradbury, E.S. Buckler, and Z. Zhang. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity*, 121:648–662, 2018.

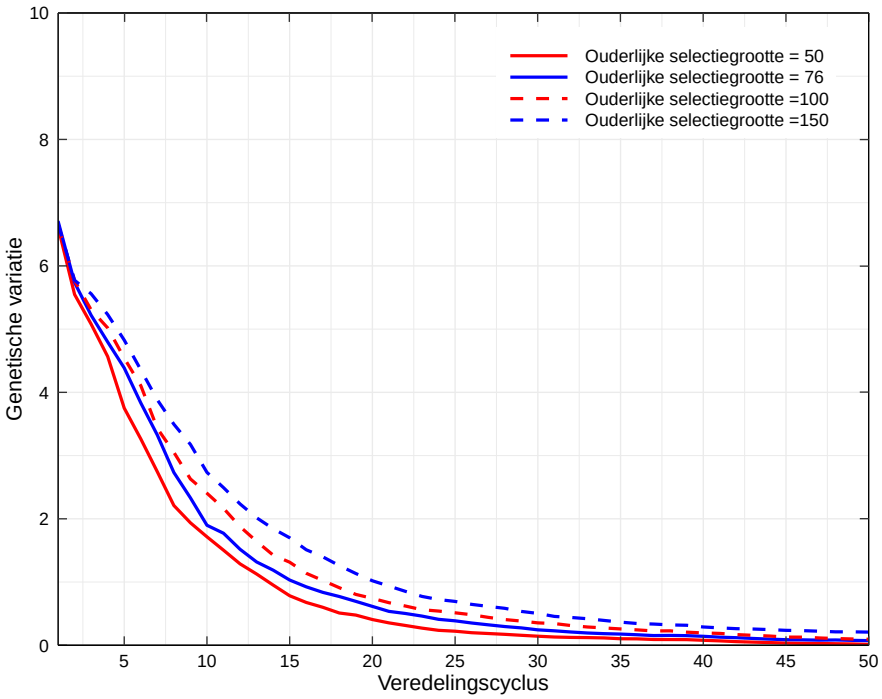
J. Weisdorf. From foraging to farming: explaining the neolithic revolution. *Journal of Economic Surveys*, 19(4):561–586, 2005.

Y. Xu and J.H. Crouch. Marker-assisted selection in plant breeding: from publications to practice. *Crop Science*, 48(2):391–407, 2008.

H. Zhou, Z. Deng, Y. Xia, and M. Fu. A new sampling method in particle filter based on Pearson correlation coefficient. *Neurocomputing*, 216:208–215, 2016.

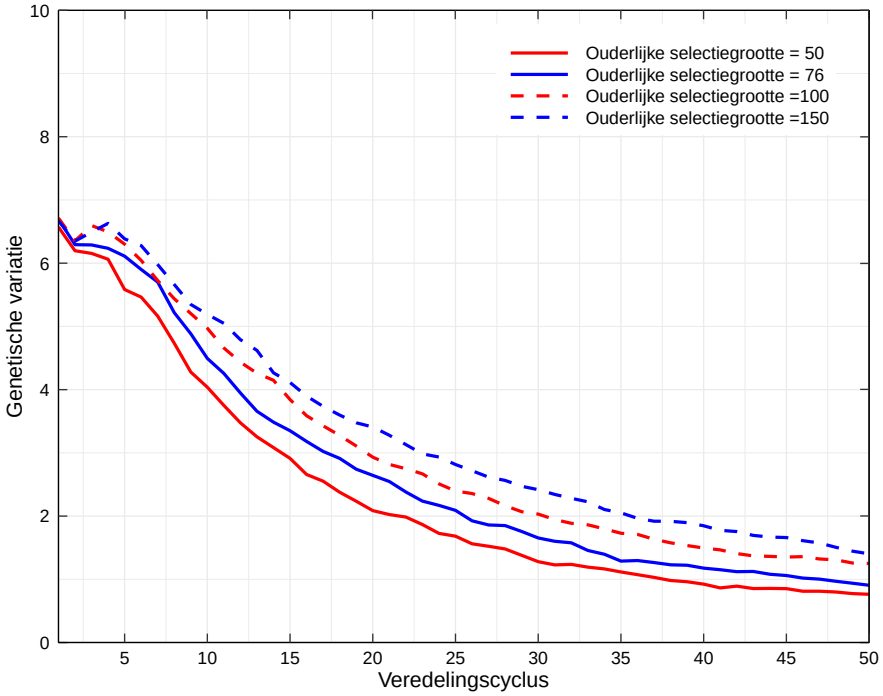
BIJLAGEN 1–12

Bijlage 1: Genetische variatie van verschillende relatief lage ouderlijke selectiegroottes voor truncatieselectie



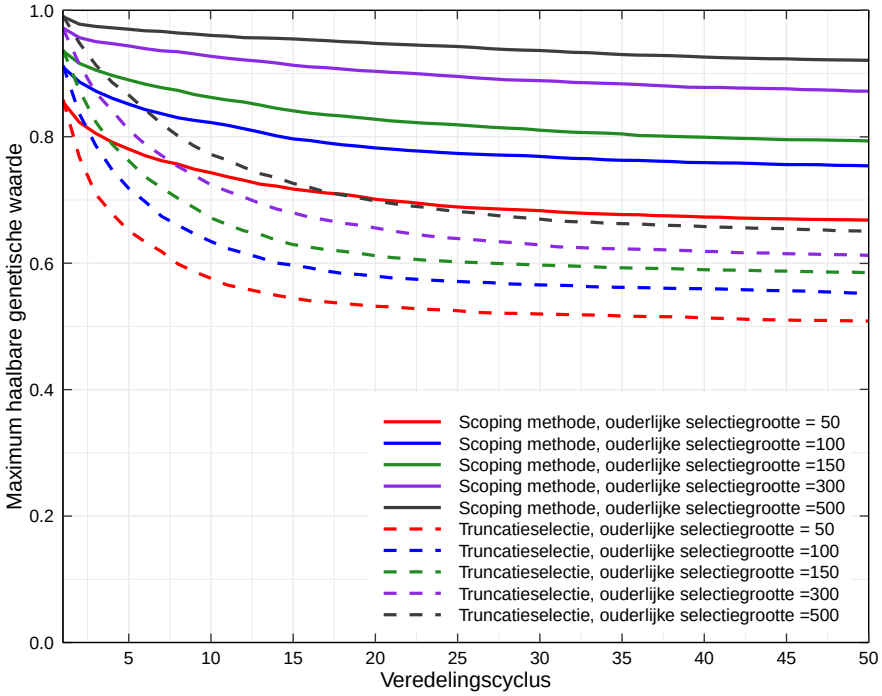
Figuur A1: Op bovenstaande figuur wordt de genetische variatie weergegeven voor verschillende ouderlijke selectiegroottes in het geval van truncatieselectie. De genetische variatie wordt beter behouden over de veredelingscycli heen wanneer meer ouders worden geselecteerd. Het valt op dat de genetische variatie bij truncatieselectie snel afneemt.

Bijlage 2: Genetische variatie van verschillende relatief lage ouderlijke selectiegroottes voor de scoping methode



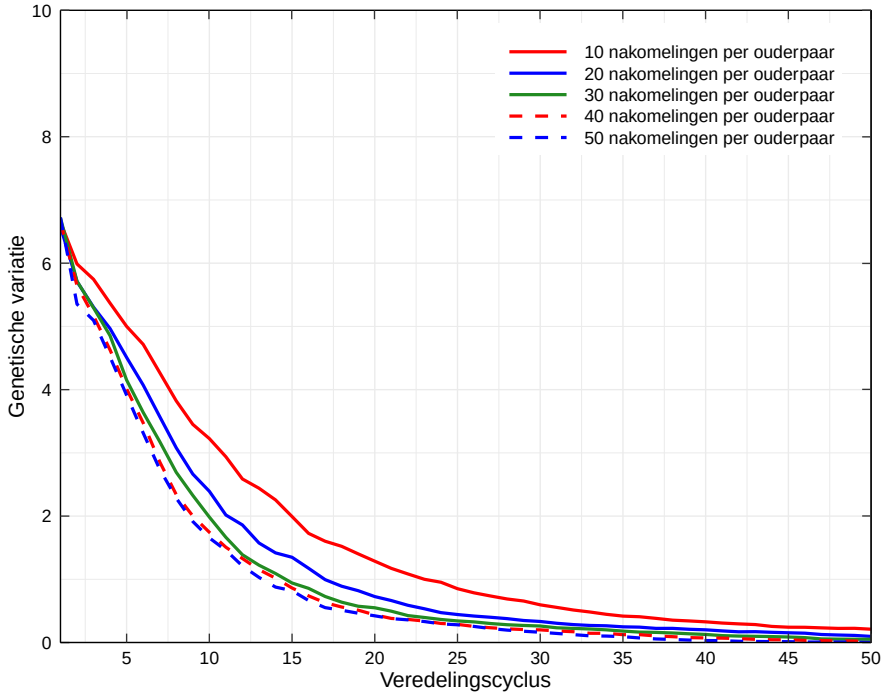
Figuur A2: Bovenstaande grafiek geeft de genetische variatie weer voor verschillende ouderlijke selectiegroottes voor de scoping methode. Net zoals bij truncatieselectie wordt de genetische variatie beter behouden wanneer meer ouders worden geselecteerd. De genetische variatie neemt bij de scoping methode duidelijk veel minder snel af in vergelijking met truncatieselectie.

Bijlage 3: Maximum haalbare genetische waarde voor verschillende ouderlijke selectiegroottes voor zowel de scoping methode als truncatieselectie



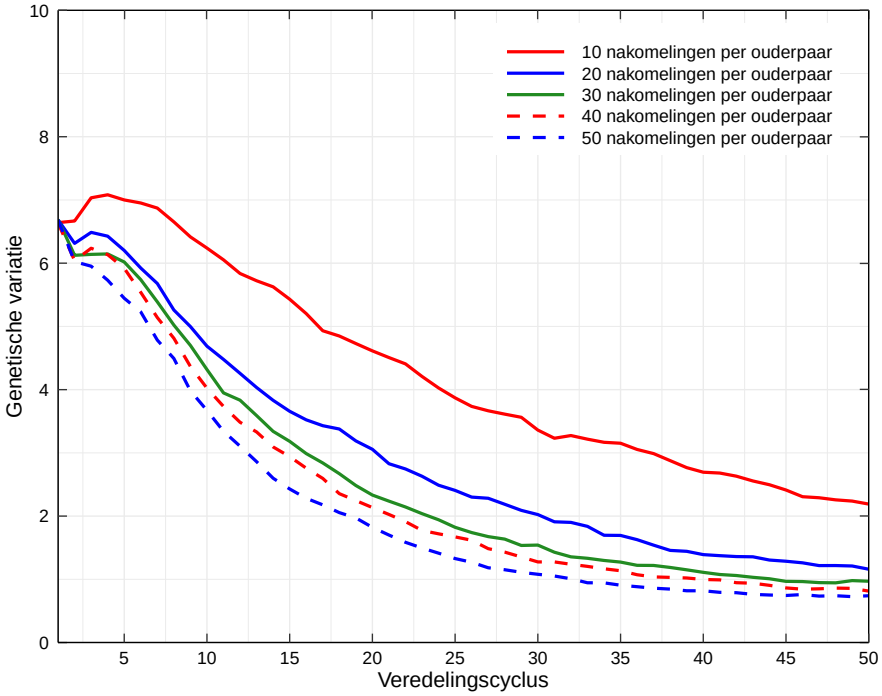
Figuur A3: Op deze figuur wordt de maximum haalbare genetische waarde voorgesteld voor ouderlijke selectiegroottes van 50, 100, 150, 300 en 500 en dit zowel voor de scoping methode als voor truncatieselectie. Er is te zien dat de maximum haalbare genetische waarde sneller afneemt voor truncatieselectie dan voor de scoping methode. Deze waarde neemt ook sneller af wanneer de ouderlijke selectiegrootte lager is.

Bijlage 4: Genetische variatie van verschillende aantallen nakomelingen per ouderpaar voor truncatieselectie



Figuur A4: De genetische variatie voor verschillende aantallen nakomelingen per ouderpaar voor truncatieselectie is weergegeven op bovenstaande grafiek. Er is te zien dat de genetische variatie hoger is naarmate het aantal nakomelingen per ouderpaar kleiner is. Er is ook te zien dat de genetische variatie snel afneemt bij truncatieselectie.

Bijlage 5: Genetische variatie van verschillende aantallen nakomelingen per ouderpaar voor de scoping methode



Figuur A5: Deze grafiek geeft de genetische variatie weer voor verschillende aantallen nakomelingen per ouderpaar voor de scoping methode. Er is net zoals bij truncatie-selectie te zien dat de genetische variatie beter wordt behouden wanneer er minder nakomelingen zijn per ouderpaar. Daarnaast valt het op dat de genetische variatie minder snel verloren gaat bij de scoping methode dan bij truncatieselectie.

Bijlage 6: Significantie van het verschil in genetische waarde tussen relatief lage ouderlijke selectiegroottes voor truncatieselectie

Tabel A1: In deze tabel wordt de significantie nagegaan van het verschil in genetische waarde tussen verschillende relatief lage ouderlijke selectiegroottes (OSG's) voor truncatieselectie. In de eerste kolom wordt weergegeven welke twee ouderlijke selectiegroottes worden vergeleken en in de tweede kolom is de corresponderende p-waarde weergegeven. De nulhypothese houdt steeds in dat beide groepsgemiddelden gelijk zijn. Indien de p-waarde minder dan 0,05 bedraagt, kan worden gesteld dat er sprake is van een significant verschil in genetische waarde tussen beide ouderlijke selectiegroottes. Deze verschillen worden steeds beschouwd na het doorlopen van 50 veredelingscycli. Er werd eerst een ANOVA-analyse uitgevoerd waarbij een p-waarde werd bekomen van 2,65e-5. Aangezien deze waarde kleiner is dan 0,05 kon worden overgegaan naar een Tukey-Kramer post-hoc analyse.

Beschouwde OSG's	**p-waarde**
50 en 76	0,221
50 en 100	0,017
50 en 150	<0,001
76 en 100	0,719
76 en 150	0,018
100 en 150	0,239

Bijlage 7: Verschil in genetische waarde tussen relatief lage ouderlijke selectiegroottes voor truncatieselectie

Tabel A2: Deze tabel toont de verschillen in genetische waarde tussen opeenvolgende relatief lage ouderlijke selectiegroottes (OSG's) voor truncatieselectie. De verschillen worden beschouwd na 50 veredelingscycli.

Beschouwde OSG's	Verschil in genetische waarde
50 en 76	2,8 ± 0,7%-punt
76 en 100	1,5 ± 0,7%-punt
100 en 150	3,1 ± 0,7%-punt

Bijlage 8: Verschil in genetische waarde tussen relatief hoge ouderlijke selectiegroottes voor truncatieselectie

Tabel A3: Het verschil in genetische waarde tussen opeenvolgende beschouwde ouderlijke selectiegroottes (OSG's) voor truncatieselectie is weergegeven in deze tabel. Het verschil werd steeds berekend na 50 veredelingscycli. *Het minteken voor deze waarde wijst erop dat er voor een ouderlijke selectiegrootte van 600 een lagere genetische waarde wordt bekomen dan voor een ouderlijke selectiegrootte van 500.

Beschouwde OSG's	Verschil in genetische waarde
150 en 300	2,4 ± 0,7%-punt
300 en 400	2,5 ± 0,7%-punt
400 en 500	0,5 ± 0,7%-punt
500 en 600	−0,4 ± 0,8%-punt *

Bijlage 9: Significantie van het verschil in genetische waarde tussen relatief lage ouderlijke selectiegroottes voor de scoping methode

Tabel A4: De significantie van het verschil in genetische waarde tussen verschillende relatief lage ouderlijke selectiegroottes (OSG's) voor de scoping methode wordt nagegaan in deze tabel. De eerste kolom toont aan welke twee ouderlijke selectiegroottes worden vergeleken en in de tweede kolom wordt de corresponderende p-waarde weergegeven. De nulhypothese formuleert steeds dat beide groepsgemiddelden gelijk zijn. Indien de p-waarde kleiner is dan 0,05 is er sprake van een significant verschil in genetische waarde tussen beide ouderlijke selectiegroottes. Deze verschillen worden steeds beschouwd na 50 veredelingscycli. Eerst werd een ANOVA-analyse uitgevoerd waarbij een p-waarde werd verkregen van 3,95e-11. Aangezien deze waarde kleiner is dan 0,05 kon een Tukey-Kramer post-hoc analyse worden gedaan.

Beschouwde OSG's	**p-waarde**
50 en 76	0,008
50 en 100	<0,001
50 en 150	<0,001
76 en 100	0,285
76 en 150	<0,001
100 en 150	0,135

Bijlage 10: Verschil in genetische waarde tussen relatief lage ouderlijke selectiegroottes voor de scoping methode

Tabel A5: In deze tabel is het verschil weergegeven in genetische waarde tussen opeenvolgende relatief lage ouderlijke selectiegroottes (OSG's) voor de scoping methode. Dit verschil wordt berekend nadat 50 veredelingscycli zijn afgewerkt.

Beschouwde OSG's	Verschil in genetische waarde
50 en 76	4,0 ± 0,6%-punt
76 en 100	2,9 ± 0,6%-punt
100 en 150	2,9 ± 0,6%-punt

Bijlage 11: Verschil in genetische waarde tussen relatief hoge ouderlijke selectiegroottes voor de scoping methode

Tabel A6: In deze tabel zijn de verschillen in genetische waarde tussen opeenvolgende relatief hoge ouderlijke selectiegroottes (OSG's) weergegeven voor de scoping methode. Deze verschillen werden beschouwd na 50 veredelingscycli.

Beschouwde OSG's	Verschil in genetische waarde
150 en 200	1,9 ± 0,6%-punt
200 en 300	1,9 ± 0,6%-punt
300 en 350	1,1 ± 0,5%-punt
350 en 500	0,7 ± 0,8%-punt

Bijlage 12: Significantie van het verschil in genetische waarde tussen de modellen voor alle beschouwde situaties

Tabel A7: Aan de hand van ANOVA-analyses wordt onderzocht of er significante verschillen zijn tussen de genetische waarden die behaald worden voor de beschouwde modellen voor tal van situaties. Een ouderlijke selectiegrootte (OSG) van 50 en 150 wordt beschouwd alsook 50 en 200 QTL's en een heritabiliteit van 0,2 en 0,8. Hierbij neemt telkens slechts één parameter per keer een waarde aan verschillend van de standaardwaarde. De modellen worden enerzijds eens vergelijken wanneer de ouders worden geselecteerd door middel van truncatieselectie en anderzijds eens wanneer deze worden geselecteerd door middel van de scoping methode. De verschillende modellen worden steeds vergeleken nadat 25 veredelingscycli zijn doorlopen. De p-waarden die horen bij de verschillende ANOVA-analyses zijn weergegeven in de laatste kolom. Aangezien elke p-waarde groter is dan 0,05 kan voor elke beschouwde situatie worden gesteld dat er geen significante verschillen zijn tussen de modellen wanneer eenzelfde ouderlijke selectiemethode wordt gebruikt.

Parameter	Waarde	Selectiemethode	p-waarde
OSG	50	Truncatie	0,597
OSG	50	Scoping	0,088
OSG	150	Truncatie	0,862
OSG	150	Scoping	0,939
Aantal QTL's	50	Truncatie	0,697
Aantal QTL's	50	Scoping	0,727
Aantal QTL's	200	Truncatie	0,867
Aantal QTL's	200	Scoping	0,961
Heritabiliteit	0,2	Truncatie	0,502
Heritabiliteit	0,2	Scoping	0,834
Heritabiliteit	0,8	Truncatie	0,196
Heritabiliteit	0,8	Scoping	0,962