Lehr- und Forschungszentrum für Landwirtschaft
Raumberg-Gumpenstein

62. Tagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs 2011, 3 – 6
ISBN: 978-3-902559-74-6

Genom-basierte Vorhersage der Testkreuzungsleistung bei Mais
Genome-based prediction of testcross performance in maize

Theresa Albrecht und Chris-Carolin Schön[1*]

Abstract

Most traits of agronomic importance follow a quantitative distribution. The assessment of these quantitative traits in performance tests is time and resource consuming. Thus, prediction of the genetic potential of individuals based on their DNA profile is highly desirable. This study comprised 1380 doubled haploid (DH) lines genotyped with 1152 single nucleotide polymorphism (SNP) markers. A subset of 759 DH lines was genotyped with additional 56110 SNP markers. Testcross progenies of the DH lines were phenotyped for the two traits grain yield and grain dry matter content in seven environments. We used best linear unbiased prediction and stratified 5-fold cross-validation to compare the performance of different statistical models with respect to genome-based prediction of testcross performance in maize. Models including genomic information outperformed the pedigree-based prediction model. Average predictive abilities were high for grain yield when the prediction was performed within families (0.66-0.68). Predictions across distantly related families lead to a decrease in predictive abilities (0.43-0.44). Predictive abilities were still relatively high when the population size was halved. For grain yield, an increase in predictive ability from 0.59 to 0.62 was achieved when the number of markers was increased from 1152 to 56110. Our results look promising for the implementation of genome-based prediction into breeding programs.

Keywords

Genome-based prediction, genomic selection, testcross performance, *Zea mays*

Einleitung

In der Pflanzenzüchtung basiert die genetische Verbesserung komplexer Merkmale weitestgehend auf der phänotypischen Evaluierung einer großen Zahl von Selektionskandidaten in ressourcenintensiven, mehrortigen und mehrjährigen Feldversuchen. Die Anwendung der von MEUWISSEN et al. (2001) für die Rinderzucht vorgeschlagenen Methode der genomischen Selektion könnte jedoch dazu führen, dass neben der im Feld gemessenen phänotypischen Leistung auch der genomische Wert eines Individuums in der Selektion Berücksichtigung finden wird.

Methodisch ist die genomische Selektion als eine Weiterentwicklung der marker-gestützten Selektion zu sehen. Die dichte Abdeckung des Genoms mit molekularen Markern, wie den Einzelbasenmutationen (engl. *single nucleotide polymorphisms*, SNPs), erlaubt den genetischen Wert eines Individuums auf der Basis des Markerprofils des Gesamtgenoms zu schätzen. Während in der klassischen Kopplungs- oder Assoziationskartierung einzelne Genomregionen mit signifikantem Effekt auf das Zielmerkmal identifiziert und zur Selektion genutzt werden, ist es durch eine massive Erhöhung der Markerdichte möglich, über Kopplungsungleichgewichte (engl. *linkage disequilibrium*; LD) alle Gene, die für ein quantitatives Merkmal kodieren (engl. *quantitative trait loci*; QTL), zu erfassen und in die Vorhersage des genetischen Werts einzubeziehen.

Zunächst müssen an einer möglichst großen Referenzpopulation sowohl genotypische als auch phänotypische Daten erhoben werden. Mittels statistischer Modelle werden auf der Basis dieser Daten Vorhersagemodelle entwickelt, die dazu dienen, die Leistung der nicht phänotypisierten Selektionskandidaten vorherzusagen und die besten Kandidaten auf der Basis ihres DNA Profils zu selektieren.

In der Tierzüchtung finden diese Methoden bereits breite Anwendung bei der genomischen Zuchtwertschätzung, z.B. von ungeprüften Bullen (VANRADEN et al. 2009) und erste experimentelle Studien zeigen, dass die genom-basierte Leistungsvorhersage auch für die Pflanzenzüchtung von Bedeutung sein kann (ALBRECHT et al. 2011, HEFFNER et al. 2011). Zur Abschätzung des züchterischen Potentials der genom-basierten Leistungsvorhersage muss ihre relative Effizienz im Vergleich zur phänotypischen Selektion bewertet werden. Eine wichtige Komponente ist hierbei die Genauigkeit der genomischen Vorhersage, d.h. die Höhe der Korrelation des vorhergesagten mit dem wahren genetischen Wert der Selektionskandidaten. In der Hybridmaiszüchtung steht dabei vor allem die Genauigkeit der Vorhersage der Testkreuzungsleistung von Inzuchtlinien im Mittelpunkt der Betrachtung.

In dieser Studie zeigen wir an einem experimentellen Datensatz verschiedene Methoden und Ergebnisse zur genom-basierten Vorhersage der Testkreuzungsleistung bei Mais. Mittels Kreuzvalidierung bestimmen wir (1) die Vorhersagegenauigkeit verschiedener Modelle, basierend auf der erwarteten oder realisierten Verwandtschaft zwischen DH-Linien, (2) die Bedeutung des Verwandtschaftsgrades für die Vorhersagegenauigkeit, (3) den Einfluss der Populationsgröße auf die Vorhersagegenauigkeit und (4) das Potential der Vorhersage basierend auf dem MaizeSNP50 BeadChip von Illumina (GANAL et al. 2011).

[1] Lehrstuhl für Pflanzenzüchtung, TU München, Emil-Ramann-Straße 4, D-85354 FREISING
[*] Ansprechpartner: Chris-Carolin SCHÖN, chris.schoen@wzw.tum.de

Material und Methoden

Die vorliegende Studie an Mais (*Zea mays* L.) umfasst insgesamt 1380 doppelhaploide (DH) Linien aus 36 Kreuzungen mit je 14 bis 60 Nachkommen. Die Abstammung der DH-Linien kann bis zu drei Generationen zurückverfolgt werden (ALBRECHT et al. 2011). Die Phänotypisierung fand in Form von Testkreuzungen an sieben europäischen Standorten statt. In diesen Körnermaisversuchen wurden zwei Merkmale, Korntrockenmasseertrag (KE, dt·ha^{-1}) und Korntrockenmassegehalt (KTMG, %), erfasst. An jedem Standort bestand das experimentelle Design aus 15 unwiederholten 10×10 Gitteranlagen jeweils bestehend aus 92 DH-Linien und vier wiederholten Standards. Adjustierte Mittelwerte am Einzelort wurden mittels der wiederholten Standards berechnet.

Für die Genotypisierung der 1380 DH-Linien wurde ein VeraCode-Assay mit 1152 SNPs verwendet. Drei DH-Linien mussten aufgrund nicht ausreichender Qualität der genotypischen Daten verworfen werden. Für die weiteren Analysen wurden 732 SNPs mit weniger als 10% fehlenden Werten und einer Frequenz des weniger häufigen Allels (engl. *minor allele frequency*, MAF) größer 1% verwendet.

Ein Teil der DH-Linien (N=759) wurde zusätzlich mit dem MaizeSNP50 BeadChip von Illumina (GANAL et al. 2011) mit insgesamt 56110 SNP-Markern genotypisiert. Nach einer Qualitätskontrolle (GenTrain-Score >0,9; MAF>0,01; fehlende Werte <0,10) gingen 20742 SNP-Marker in die weiteren Analysen ein. Die 20742 SNP-Marker sind gleichmäßig über alle zehn Chromosomen des Maisgenoms verteilt und haben einen durchschnittlichen Abstand von 0,11 Mb. Fehlende Werte wurden anhand der Familienstruktur ersetzt. Eine Übersicht über die in dieser Studie verwendeten Datensätze ist in *Tabelle 1* gegeben.

Tabelle 1: **Anzahl DH-Linien und SNP-Marker für drei in dieser Studie verwendete Datensätze**

Table 1: **Number of DH lines and SNP markers for three data sets used in this study**

Datensatz	Anzahl DHs (N)	Anzahl SNPs (M)
VC1	1377	732
VC2	759	654
50k	759	20742

Vorhersage der Testkreuzungsleistung

Die Vorhersage der Testkreuzungsleistung der DH-Linien wurde mit linearen gemischten Modellen durchgeführt, wobei die zufälligen Effekte der Genotypen sogenannte BLUPs (engl. *best linear unbiased predictors*; HENDERSON 1984) sind. Die Vorhersagegenauigkeit folgender drei Modelle, die sich in der Modellierung der zufälligen Effekte unterscheiden, wurde untersucht:

PBLUP $\quad \mathbf{y} = \boldsymbol{\mu} + \mathbf{Zt} + \mathbf{e}$

GBLUP $\quad \mathbf{y} = \boldsymbol{\mu} + \mathbf{Zs} + \mathbf{e}$

P+GBLUP $\quad \mathbf{y} = \boldsymbol{\mu} + \mathbf{Zt} + \mathbf{Zs} + \mathbf{e}$

In allen drei Modellen enthält der Vektor $\mathbf{y}$ die N adjustierten Mittelwerte aus der Auswertung der phänotypischen Daten über die sieben Standorte. In allen drei Modellen steht der Vektor $\boldsymbol{\mu}$ für das Gesamtmittel und der Residuenvektor $\mathbf{e}$ folgt einer Normalverteilung mit $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2)$. Die Designmatrix $\mathbf{Z}$ ordnet die Phänotypen den zufälligen Effekten zu.

Die drei Modelle unterscheiden sich durch die den zufälligen Effekten zugrunde liegende Varianz-Kovarianz-Struktur. Im Modell PBLUP (Pedigree-BLUP) ist der Vektor $\mathbf{t}$ normalverteilt mit $\mathbf{t} \sim N(0, \mathbf{K}\sigma^2_t)$. Die Varianz-Kovarianz-Matrix $\mathbf{K}$ entspricht hier der Verwandtschaftsmatrix basierend auf drei Generationen Abstammungsinformation und σ^2_t ist die für das Modell PBLUP spezifische genetische Varianz der Testkreuzungen.

Im GBLUP (Genomic-BLUP) ist der Vektor $\mathbf{s}$ ebenfalls normalverteilt mit $\mathbf{s} \sim N(0, \mathbf{S}\sigma^2_s)$ und σ^2_s ist die für das Modell GBLUP spezifische Varianz der Testkreuzungen. Die marker-basierte Verwandtschaftsmatrix $\mathbf{S}$ wird über den Simple-Matching-Koeffizienten (SNEATH and SOKAL 1973) berechnet, der für alle möglichen paarweisen Linienkombinationen den Anteil übereinstimmender Markergenotypen relativ zur Gesamtzahl der Marker quantifiziert. Dabei ergibt sich für die Matrix $\mathbf{S}_{SM}$ mit paarweisen Simple-Matching-Koeffizienten folgende Formel:

$$\mathbf{S}_{SM} = \frac{(\mathbf{W} - \mathbf{J}_{N\times M})(\mathbf{W} - \mathbf{J}_{N\times M})' + M\mathbf{J}_{N\times N}}{2M}$$

wobei die Matrix $\mathbf{W}$ die Marker-Genotypen der DHs enthält (0 oder 2 Kopien des selteneren Alleles), $\mathbf{J}_{N\times M}$ und $\mathbf{J}_{N\times N}$ Matrizen mit Einsen als Elemente und den Dimensionen $N\times M$ und $N\times N$ sind und M die Anzahl der Marker ist. Um daraus die Matrix $\mathbf{S}$, die im Modell GBLUP die Varianz-Kovarianz-Struktur der zufälligen Effekte beschreibt, zu erhalten, wird jedes Element der Matrix mit dem Minimum s_{min} der Matrix $\mathbf{S}_{SM}$ korrigiert und $\mathbf{S}$ ergibt sich als:

$$\mathbf{S} = \frac{(\mathbf{W} - \mathbf{J}_{N\times M})(\mathbf{W} - \mathbf{J}_{N\times M})' + M\mathbf{J}_{N\times N} - 2Ms_{\min}\mathbf{J}_{N\times N}}{2M(1 - s_{\min})}$$

Im Modell P+GBLUP wird sowohl die abstammungsbasierte als auch die marker-basierte Verwandtschaft berücksichtigt.

Kreuzvalidierung

Die Kreuzvalidierung (engl. *cross-validation*; CV) ist eine Methode zur Bestimmung der Vorhersagegenauigkeit statistischer Modelle und zum Vergleich verschiedener Modelle. Dabei werden die Daten zufällig auf k gleich große Teildatensätze aufgeteilt. Aus k-1 Teildatensätzen wird ein Trainingsdatensatz (engl. *estimation set*; ES) gebildet, anhand dessen das Vorhersagemodell entwickelt wird. Für den verbleibenden Teildatensatz (engl. *test set*; TS) wird die Testkreuzungsleistung der DH-Linien vorhergesagt. Anhand der Korrelation zwischen der vorhergesagten Testkreuzungsleistung der DH-Linien im TS und ihrer tatsächlich beobachteten Leistung kann die Vorhersagegenauigkeit des Modells bewertet werden.

In dieser Studie haben wir eine 5-fach CV unter Berücksichtigung der Verwandtschaftsstruktur der DH-Linien durchgeführt (LEGARRA et al. 2008). Dabei wird der Datensatz in fünf gleich große Teildatensätze geteilt. Jeder der fünf Teildatensätze bildet ein TS mit zugehörigem Trainingsdatensatz bestehend aus den anderen vier Teildatensätzen. Wird die Aufteilung der DH-Linien auf die Teildatensätze zehnmal neu randomisiert, ergeben sich insgesamt 50 CV-Läufe. Um die Abhängigkeit der Vorhersagegenauigkeit vom Grad der Verwandtschaft zwischen ES und TS zu untersuchen, wurde

bei der Aufteilung der DH-Linien auf die Teildatensätze die Familienstruktur berücksichtigt. Wird jede Familie in fünf Teile geteilt (engl. *CV within*; CV-W), ist eine engere Verwandtschaft zwischen ES und TS zu erwarten, als wenn ganze Familien entweder dem ES oder dem TS zugeteilt werden (engl. *CV across*; CV-A).

Um den Effekt der Populationsgröße auf die Vorhersagegenauigkeit zu ermitteln, wurde der Datensatz VC1 in zwei, vier und acht Teildatensätze geteilt. Durch wiederholte, neu randomisierte Zuordnung der DH-Linien zu den Teildatensätzen entstanden für N=688, 344 und 172 jeweils 32 unterschiedliche Teildatensätze. In jedem dieser Teildatensätze wurde die Vorhersagegenauigkeit mittels 5-fach CV mit zufälliger Aufteilung der DH-Linien auf ES und TS für die drei beschriebenen Modelle bewertet.

Ergebnisse und Diskussion

Die Genauigkeiten der Vorhersage der Testkreuzungsleistung für die beiden Merkmale Kornertrag und Korntrockenmassegehalt sind für die VC1-Daten mit 1377 DH-Linien und 732 SNPs in *Tabelle 2* dargestellt. Die Modelle, die genomische Information nutzen (GBLUP und P+GBLUP), erzielten signifikant höhere Vorhersagegenauigkeiten als das PBLUP Modell. Für Kornertrag lag die Vorhersagegenauigkeit für GBLUP bei 0,66, wenn bei der Kreuzvalidierung die Aufteilung in ES und TS innerhalb der Familien vorgenommen wurde (CV-W) und bei 0,44 wenn Familien entweder dem ES oder dem TS zugeteilt wurden (CV-A). Bei PBLUP wurde für CV-W eine Korrelation zwischen vorhergesagter und beobachteter Testkreuzungsleistung von 0,51 erreicht. Wie erwartet war die Vorhersagegenauigkeit für PBLUP und CV-A mit 0,11 sehr gering. Für Korntrockenmassegehalt lagen die Genauigkeiten insgesamt höher als für Kornertrag mit 0,72 für GBLUP in CV-W und 0,59 in CV-A.

Ein Grund für das bessere Abschneiden der marker-basierten Modelle liegt in der genaueren Vorhersage der Testkreuzungsleistung innerhalb von Familien. Alle Nachkommen einer Kreuzung erhalten basierend auf Stammbauminformationen denselben Abstammungskoeffizienten und somit kann mit PBLUP bei der Vorhersage der Testkreuzungsleistung nicht zwischen DH-Linien, die auf dieselbe Kreuzung zurückgehen, differenziert werden. Bei GBLUP hingegen wird die realisierte Verwandtschaft anhand der genom-weiten Markerdaten berechnet und somit kann bei der Leistungsvorhersage auch zwischen DH-Linien, die auf dieselbe Kreuzung zurückgehen, unterschieden werden. In *Abbildung 1* ist die vorhergesagte Testkreuzungsleistung für Kornertrag von 264 DH-Linien eines zufällig ausgewählten TS aus CV-W für die Modelle PBLUP und GBLUP gegen die aus dem Gesamtdatensatz geschätzte mittlere Familienleistung gezeigt. Während für PBLUP nur ein gemeinsamer Wert für die Testkreuzungsleistung aller DH-Linien einer Familie angegeben ist, differenziert GBLUP auch zwischen DH-Linien derselben Kreuzung.

Der Zusammenhang zwischen der Größe des Trainingsdatensatzes und der Vorhersagegenauigkeit für die drei Modelle ist in *Abbildung 2* dargestellt. Bei Halbierung des Datensatzes von 1377 DHs auf 688 DHs fällt die Vorhersagegenauigkeit aller Modelle leicht ab. Bei Reduktion der Populationsgröße von 344 auf 172 Linien fällt bei allen drei Modellen die Vorhersagegenauigkeit stark ab.

Die Erhöhung der Markerdichte von 654 SNPs (VC2, N=759) auf 20742 SNPs (50k, N=759) führte für beide Merkmale zu einer signifikant höheren Vorhersagegenauigkeit (*Abbildung 3*). Für Kornertrag erhöhte sich die Vorhersagegenauigkeit von 0,59 auf 0,62 mit CV-W und von 0,35 auf 0,39 mit CV-A. Beim direkten Vergleich der mit GBLUP erzielten Vorhersagegenauigkeit der Datensätze VC1 und 50k wird jedoch erkennbar, dass die Verbesserung der Vorhersagegenauigkeit durch mehr Marker den durch die geringere Populationsgröße entstandenen Verlust an Genauigkeit nicht kompensieren konnte.

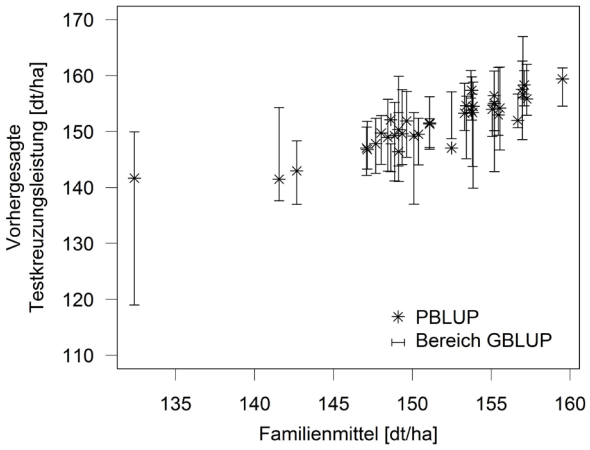

Abbildung 1: **Vorhergesagte Testkreuzungsleistung mit PBLUP sowie der mit GBLUP vorhergesagte Wertebereich eines TS aus CV-W gegen das jeweilige Familienmittel für das Merkmal Kornertrag im Datensatz VC1**

Figure 1: **Predicted testcross performance of DH lines with PBLUP and range of predicted performance with GBLUP for one TS randomly chosen from CV-W plotted against the respective family mean for grain yield in the data set VC1**

Tabelle 2: **Durchschnittliche Vorhersagegenauigkeit drei verschiedener Modelle für die Merkmale Kornertrag (KE) und Korntrockenmassegehalt (KTMG), basierend auf 50 Kreuzvalidierungen innerhalb (CV-W) und zwischen Familien (CV-A) des VC1 Datensatzes**

Table 2: **Average predictive abilities of three models for the traits grain dry matter yield (KE) and grain dry matter content (KTMG) evaluated with 50 cross-validation runs based on sampling within (CV-W) and across families (CV-A) of the VC1 data set**

	KE		KTMG	
Modell	CV-W	CV-A	CV-W	CV-A
PBLUP	0,51	0,11	0,50	0,31
GBLUP	0,66	0,44	0,72	0,59
P+GBLUP	0,68	0,43	0,72	0,59

Schlussfolgerung

Die Ergebnisse dieser Studie sind ermutigend für die Implementierung der genom-basierten Leistungsvorhersage in Zuchtprogrammen bei Mais. Für die beiden Merkmale Kornertrag und Korntrockenmassegehalt konnten mit

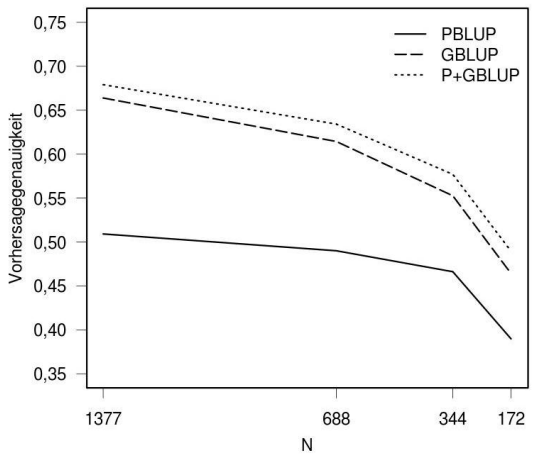

Abbildung 2: **Vorhersagegenauigkeit für Kornertrag der Modelle PBLUP, GBLUP und P+GLUP mit zufälliger Kreuzvalidierung in Abhängigkeit von der Populationsgröße *N* für den Datensatz VC1**

Figure 2: **Predictive ability for grain yield of models PBLUP, GBLUP and P+GLUP derived from random cross-validation as a function of population size *N* for the data set VC1**

marker-basierten Vorhersagemodellen und Kreuzvalidierung innerhalb Familien hohe Korrelationen zwischen vorhergesagter und beobachteter Testkreuzungsleistung erzielt werden. Jedoch hatte der Verwandtschaftsgrad zwischen den DH-Linien in der Trainingspopulation und den DH-Linien, deren Leistung vorhergesagt werden sollte, einen großen Einfluss auf die Vorhersagegenauigkeit. Auch die Populationsgröße wirkte sich deutlich auf die Vorhersagegenauigkeit der Modelle aus. Bis zu einer Größe von *N*=688 erzielte GBLUP jedoch immer noch gute Ergebnisse. Ein Anstieg der Vorhersagegenauigkeit konnte mit Zunahme der Markerdichte auf 20742 SNPs erreicht werden.

Um zu einer soliden Bewertung der relativen Effizienz der genom-basierten Vorhersage zu kommen, bedarf es weiterer experimenteller Ergebnisse zur Genauigkeit der Vorhersage über Materialgruppen, Jahre, Tester und Selektionszyklen. Es ist jedoch auf der Basis der vorliegenden Ergebnisse zu erwarten, dass die marker-basierte Leistungsvorhersage bestehende, phänotyp-basierte Zuchtprogramme sinnvoll ergänzen kann.

Danksagung

Diese Studie wurde durchgeführt in enger Zusammenarbeit zwischen den Autoren und Hans-Jürgen Auinger, Malena Erbe, Carsten Knaak, Milena Ouzunova, Henner Simianer und Valentin Wimmer. Diese Studie wurde mit Mitteln des Bundeministeriums für Bildung und Forschung (BMBF) durch das AgroClustEr *Synbreed - Synergistische Pflanzen- und Tierzüchtung* (FKZ: 0315528A) gefördert.

Literatur

ALBRECHT T, WIMMER V, AUINGER HJ, ERBE M, KNAAK C, OUZUNOVA M, SIMIANER H, SCHÖN CC. 2011: Genome-based prediction of testcross values in maize. Theor Appl Genet 123: 339-350.

GANAL MW, DURSTEWITZ G, POLLEY A, BÉRARD A, BUCKLER ES, CHARCOSSET A, CLARKE JD, GRANER EM, HANSEN M, JOETS J, LE PASLIER MC, MCMULLEN MD, MONTALENT P, ROSE M, SCHÖN CC, SUN Q, WALTER H, MARTIN OC, FALQUE M. 2011: A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS One 6(12): e28334. (DOI:10.1371/journal.pone.0028334)

HEFFNER ELJ, JANNINK JL, SORRELLS JL, 2011: Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4: 65-75.

HENDERSON CR, 1984: Applications of linear models in animal breeding. University of Guelph, Guelph.

LEGARRA A, ROBERT-GRANIE C, MANFREDI E, ELSEN JM, 2008: Performance of genomic selection in mice. Genetics 180: 611-618.

MEUWISSEN THE, HAYES BJ, GODDARD ME. 2001: Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

SNEATH PH, SOKAL RR, 1973: Numerical taxonomy: the principles and practice of numerical classification. Freemann, San Francisco, CA.

VANRADEN PM, TASSELL CV, WIGGANS GR, SONSTEGARD TS, SCHNABEL RD, TAYLOR JF, SCHENKEL FS. 2009: Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92: 16-24.

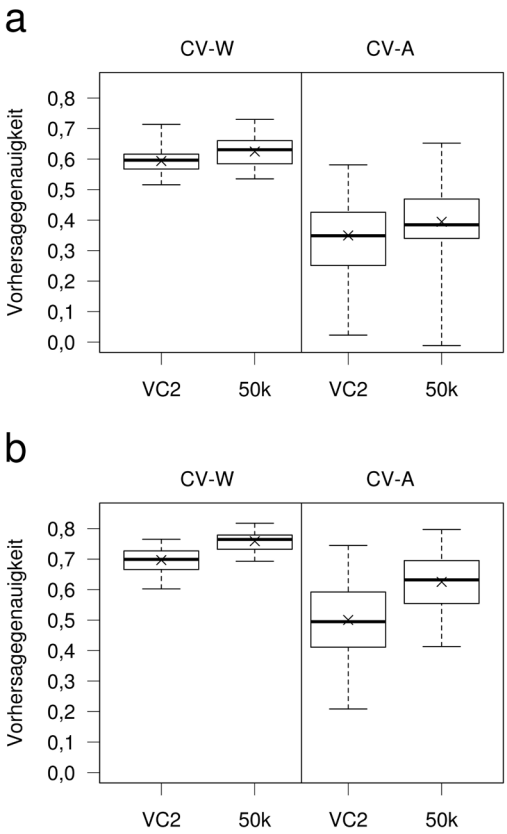

Abbildung 3: **Vergleich der Vorhersagegenauigkeit von GBLUP für die Datensätze VC2 (*M*=654, *N*=759) und 50k (*M*=20742, *N*=759) mittels Kreuzvalidierung innerhalb (CV-W) und zwischen Familien (CV-A) für die Merkmale Kornertrag (a) und Korntrockenmassegehalt (b).**

Figure 3: **Comparison of the predictive ability of GBLUP with data sets VC2 (*M*=654, *N*=759) and 50k (*M*=20742, *N*=759) determined from cross-validation within (CV-W) and across families (CV-A) for grain yield (a) and grain dry matter content (b).**