

Endbericht

Zuwendungsempfänger:

Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK)

Gemeinschaft zur Förderung von Pflanzeninnovation GFPI e.V.

KWS LOCHOW GmbH

SAATEN-UNION BIOTEC GmbH

Förderkennzeichen:

2818408B18

Vorhabensbezeichnung: Verbundprojekt: Nutzung von Big Data in Weizen zur Präzisionszucht (BigData)

Laufzeit des Vorhabens:

01.02.2020-31.01.2025

Berichtszeitraum:

Endbericht

Inhaltsverzeichnis	Seite
I. Kurze Darstellung	
1. Ursprüngliche Aufgabenstellung sowie der wissenschaftliche und technische Stand an den angeknüpft wurde	01
2. Ablauf des Vorhabens	01
3. Wesentliche Ergebnisse sowie ggf. Zusammenarbeit mit anderen Stellen	02
II. Eingehende Darstellung	
1. Verwendung der Zuwendung	03
2. Wichtigsten Positionen des zahlenmäßigen Nachweises	13
3. Notwendigkeit und Angemessenheit der geleisteten Arbeit	13
4. Voraussichtlichen Nutzen im Sinne des Verwertungsplans	13
5. Bekanntgewordener Fortschritt bei anderen Stellen	13
6. Veröffentlichungen der Ergebnisse	14
III. Erfolgskontrollbericht	
1. Wissenschaftlich-technische Ergebnis des Vorhabens, die erreichten Nebenergebnisse und die gesammelten wesentlichen Erfahrungen	15
2. Fortschreibung des Verwertungsplans	17
3. Arbeiten, die zu keiner Lösung geführt haben	18
4. Einhaltung der Kosten- und Zeitplanung	18

I. Kurze Darstellung

1. Ursprüngliche Aufgabenstellung sowie der wissenschaftliche und technische Stand an den angeknüpft wurde

Für die wissenschaftsbasierte Züchtung von resilienten Weizensorten sind umfangreiche phänotypische und genomische Daten eine wesentliche Voraussetzung. Das Ziel des Projektes war es, das Potenzial von Big Data für die Züchtung leistungsfähiger und resilienter Sorten zu erschließen. In vielen Forschungs- und Züchtungsvorhaben werden Daten von Weizenpopulationen mit einigen hundert bis wenigen tausend Genotypen erhoben, die zu verschiedenen Aspekten der Züchtung beitragen. Um jedoch die Dimension von Big Data zu erreichen, war es erforderlich, diese unterschiedlichen, projektspezifischen Datensätze zu integrieren.

Im BigData-Projekt sollte daher ein Weizen-Informationssystem entwickelt werden, das einen langfristigen Zugriff auf Daten gemäß den FAIR-Prinzipien („Findable, Accessible, Interoperable, and Reusable“) ermöglicht. In diesem System sollten phänotypische und genomische Daten aus öffentlich geförderten Projekten sowie aus Weizenzuchtprogrammen gebündelt werden. Aufbauend auf diesen Daten sollten biometrische Analyseverfahren implementiert werden, um mit den heterogenen phänotypischen Daten genotypische Werte zu schätzen.

Die verfügbare Weizengenomsequenz ermöglichte es, heterogene genomische Daten aus verschiedenen Projekten zu vereinheitlichen. Weiterhin sollten Verfahren zur genomweiten Vorhersage komplexer Merkmale mittels Big Data entwickelt und validiert werden. Für die Assoziationsgenetische Analyse der genetischen Architektur von Merkmalen sollte die Informationsfülle projektübergreifender Datenressourcen genutzt werden. Basierend auf diesen umfassenden Daten sollten die neu entwickelten Methoden validiert werden.

2. Ablauf des Vorhabens

Die Arbeiten des Verbundprojektes gliederten sich in drei Arbeitspakete (AP). In AP1 implementierten wir eine umfassende Pipeline für die Kuration der phänotypischen und genotypischen Daten. Diese Daten stammten aus abgeschlossenen Drittmittelprojekten, den laufenden Zuchtprogrammen von vier Unternehmen und den Bundessortenversuchen. Dazu wurden Übergabeprotokolle für die Daten definiert und Validierungsregeln erstellt und implementiert. Dabei wurden Standards für Phänotypisierungsdaten wie MIAPPE berücksichtigt. Ebenfalls für die genomischen Daten wurden Übergabeprotokolle definiert. Im Nachgang wurden die einzelnen Datensätze kuratiert und die Interoperabilität zwischen den Teildaten hergestellt. Die phänotypischen und genomischen Daten wurde integriert verrechnet. Die Ergebnisse belegen die hohe Qualität der kuratierten Daten.

Im AP2 wurde ein Weizen-Informationssystem weiterentwickelt. Das System ist so konzipiert, dass die Projektpartner ihre Daten aus verschiedenen Anbaujahren in ein gesichertes Datenmanagementsystem, welches am IPK in Gatersleben betrieben wird, speichern. Hierbei wird eine datenbankgestützte Authentifizierungs- und Autorisierungsinfrastruktur verwendet. Damit

ist es möglich, dass verschiedene Personen eines Projektpartners, in einem für diesen Projektpartner zugriffsgeschützten Bereich einlagern. Während der Speicherung erfolgt eine Überprüfung der Daten gemäß der im Projekt definierten Richtlinien und Vorgaben. Ergebnis ist ein Statusreport mit detaillierten Fehlermeldungen. Weiterhin erfolgt eine Versionierung der gespeicherten Dateien, um bereits mit (Teil-)Daten zu arbeiten, die im Laufe des Projektes weiter ergänzt wurden. Das AP2 bildet den zentralen Hub für die Übertragung der Daten der Projektpartner und ist Grundlage für die Aktivitäten in AP1, welche dann wiederum die Basis für die Arbeiten in AP3 bilden.

Im Rahmen von AP3 implementierten und entwickelten wir biometrische Modelle für genomweite Vorhersagen und Assoziationskartierung für die umfassenden Daten. Unsere Studien zeigten, dass die Nutzung von Big Data die Vorhersagegenauigkeit für den Kornertrag substantiell steigern kann und individuelle Trainingssätze der einzelnen Firmen deutlich übertroffen werden. Diese Verbesserung ist vor allem auf die Erweiterung der Trainingsatzgröße im Verhältnis zur genetischen Vielfalt zurückzuführen. Big Data bietet somit ein erhebliches Potenzial zur Beschleunigung des genetischen Fortschritts in der Winterweizen-Züchtung.

Im Rahmen von AP3 wurde weiterhin untersucht, ob Assoziationsstudien auf Basis von Big Data zur zuverlässigeren Identifikation von QTLs und präziseren Schätzung ihrer genetischen Effekte beitragen können. Entgegen der Erwartungen wurden in den größeren, kombinierten Datensätzen weniger Assoziationen zwischen Markern und Merkmalen gefunden als in den einzelnen Versuchsserien. Dennoch zeigte sich, dass die Vorhersagegenauigkeit auf Basis der Marker-Merkmal-Assoziationen des integrierten Datensatzes höher war, was die Integration von Big Data als vielversprechenden Ansatz zur Verbesserung der QTL-Identifikation unterstützt.

3. Wesentliche Ergebnisse sowie ggf. Zusammenarbeit mit anderen Stellen

Unsere Ergebnisse belegen das große Potenzial, Daten über verschiedene Züchtungsunternehmen hinweg zu nutzen: Wir konnten die genomische Vorhersagegenauigkeit im Vergleich zu den internen Vorhersagen der Unternehmen signifikant steigern. Dieses zentrale Ergebnis war ein ausschlaggebender Faktor für die Entwicklung eines Konzepts für ein Datenökosystem in der Pflanzenzüchtung im Rahmen des *BreedFides* Projekts. Um die Informationsdichte der Beschreibung der Umwelten zu erhöhen, haben wir mit dem NFDI-Konsortium FAIRagro zusammengearbeitet.

II. Eingehende Darstellung

1. Verwendung der Zuwendung

Qualitätsprüfung der phänotypischen und genomischen Daten

Für das BigData-Projekt flossen phänotypische Daten zum Blühzeitpunkt, zur Wuchshöhe und zum Kornertrag sowie Markerdaten ein. Diese Daten stammten aus abgeschlossenen Drittmittelprojekten, den laufenden Zuchtprogrammen von vier Unternehmen und den Bundessortenversuchen. Zu Beginn des Projekts definierte die Arbeitsgruppe QG (IPK-QG) in Zusammenarbeit mit den beteiligten Unternehmen Übergabeprotokolle für die Daten. Im Anschluss wurden Validierungsregeln erstellt und implementiert, um die Datenqualität sowohl auf syntaktischer als auch auf Wertebereichsebene sicherzustellen. Dabei wurden Standards für Phänotypisierungsdaten wie MIAPPE (<http://www.miappe.org>) berücksichtigt.

Die genomischen Daten wurden von den beteiligten Züchtern an den Partner IPK-BIT übergeben, ohne dass zuvor etwaige Fehlwerte durch Schätzungen ersetzt wurden. Zusätzlich wurden Metadaten zum Ausleseformat der SNP-Arrays sowie zu den Oligonukleotidsequenzen, die für das Design der Arrays verwendet wurden, bereitgestellt. Für das weitere Handling der Genotypdaten wurde ein Datenmodell entwickelt, das in Abbildung 1 dargestellt ist. Für die heterogenen Daten aus den SNP-Arrays implementierten wir einen Algorithmus, der die Harmonisierung und Integration der genomischen Daten ermöglichte (Abbildung 2).

Nach der Integration der genotypischen Daten wurde eine Hauptkoordinatenanalyse (PCoA) der integrierten genotypischen Daten durchgeführt (Abbildung 3). Die Ergebnisse der PCoA zeigen, dass es keine ausgeprägte Populationsstruktur zwischen den Genotypen der vier verschiedenen Unternehmen (Exp-5-8) gibt.

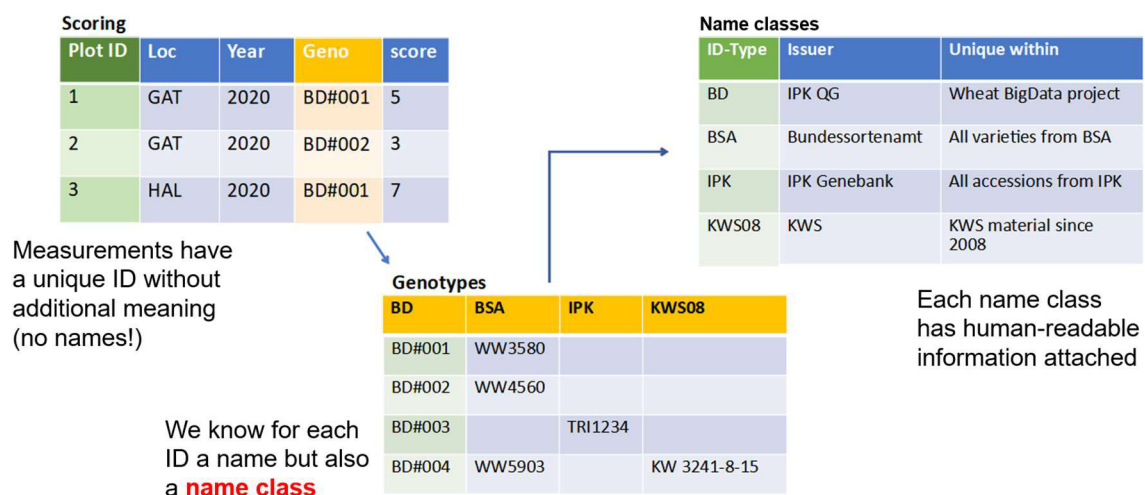


Abbildung 1: Datenmodell für Genotyp-Bezeichner mit kleinem Beispiel-Datensatz. Die drei rechteckigen Bereiche stellen drei Tabellen (Relationen) dar. Diese sind, von oben links, gegen den Uhrzeigersinn: Boniturdaten: Schlüsselspalte Parzellen-ID (grün), der Genotyp-ID (gelb) und weiteren Beispieldaten. Genotypen: Zuordnung von Genotyp-IDs aus den Boniturdaten zu einem oder mehreren Namen, die selbst zu einer Namens-Klasse gehören. Die Namen, die in der gleichen Spalte angegeben sind, gehören zur selben Klasse, deren Name im Spaltenkopf (außer Spalte 1) gezeigt ist. In der Datenbankpraxis ist die Klasse ein einzelnes Attribut und jeder Name ist ein Eintrag, die Darstellung hier ist vereinfachend.

Namensklassen: Jede Namensklasse, identifiziert durch ihren Namen in der ersten Spalte, hat eine menschenlesbare Dokumentation.

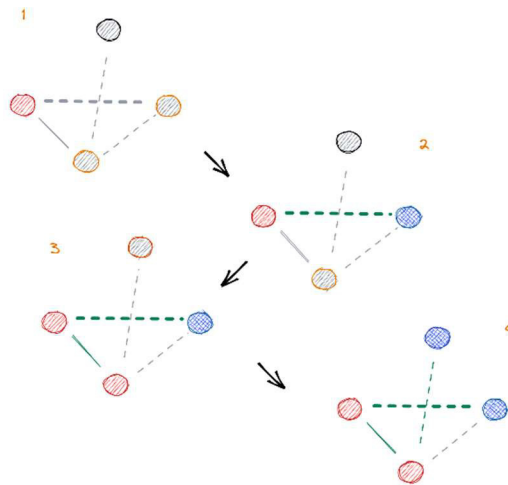


Abbildung 2: Schematischer Ablauf des implementierten Algorithmus zur Integration von heterogenen SNP-Array Daten. Die Knoten repräsentieren die SNP-Arrays, die Farben rot und blau zeigen Arrays, für die entschieden wurde, ob die Daten invertiert werden oder nicht. Grau = noch unentschieden. Orange = zur Auswahl für den nächsten Schritt, ausgewählt wird nach maximalem Kantengewicht. Kanten: Gemeinsame Genotypen zwischen den Arrays existieren. Die Gewichtung der Kanten ist im Schema nicht gezeigt. Wenn die Daten der gemeinsamen Genotypen zeigen, dass einer der beiden Arrays invertiert werden muss, ist die Kante gestrichelt. Der Verlauf des Algorithmus ist anhand der Pfeile von oben links nach unten rechts gezeigt.

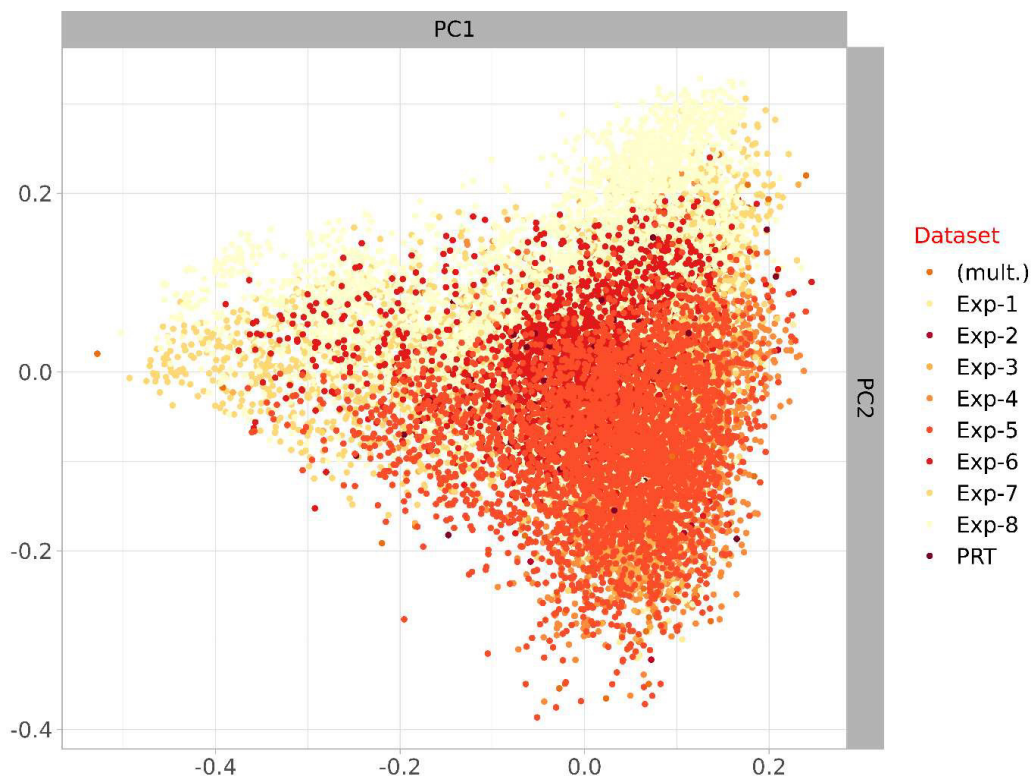


Abbildung 3: Die Hauptkoordinatenanalyse basierend auf den integrierten SNP-Daten.

Für die phänotypischen Daten wurde für jede einzelne Umwelt eine Plausibilitätsprüfung der Spannweiten und Mittelwerte der Beobachtungen durchgeführt. Heritabilitäten wurden für die

einzelnen Umwelten geschätzt, insbesondere im Fall von wiederholten Feldversuchen. Zusätzlich wurden die genomischen und phänotypischen Daten jeder Umwelt genutzt, um kreuzvalidierte Vorhersagegenauigkeiten mittels GBLUP-Modellen zu schätzen. Sowohl die Heritabilitäten in den einzelnen Umwelten als auch die Genauigkeiten der genomweiten Vorhersagen dienen als Qualitätsfilter. Die phänotypischen Daten von Serien (d.h. derselbe Satz an Genotypen, der in mehreren Umwelten geprüft wurde) wurden für Korrelationsanalysen verwendet, um mögliche Ausreißer zwischen den Umwelten zu identifizieren. Über Wetterdaten wurde dann geprüft, ob die Ausreißer auf starke Genotyp-Umwelt-Interaktionseffekte zurückzuführen sind oder ob Versuchsfehler wahrscheinlicher erscheinen.

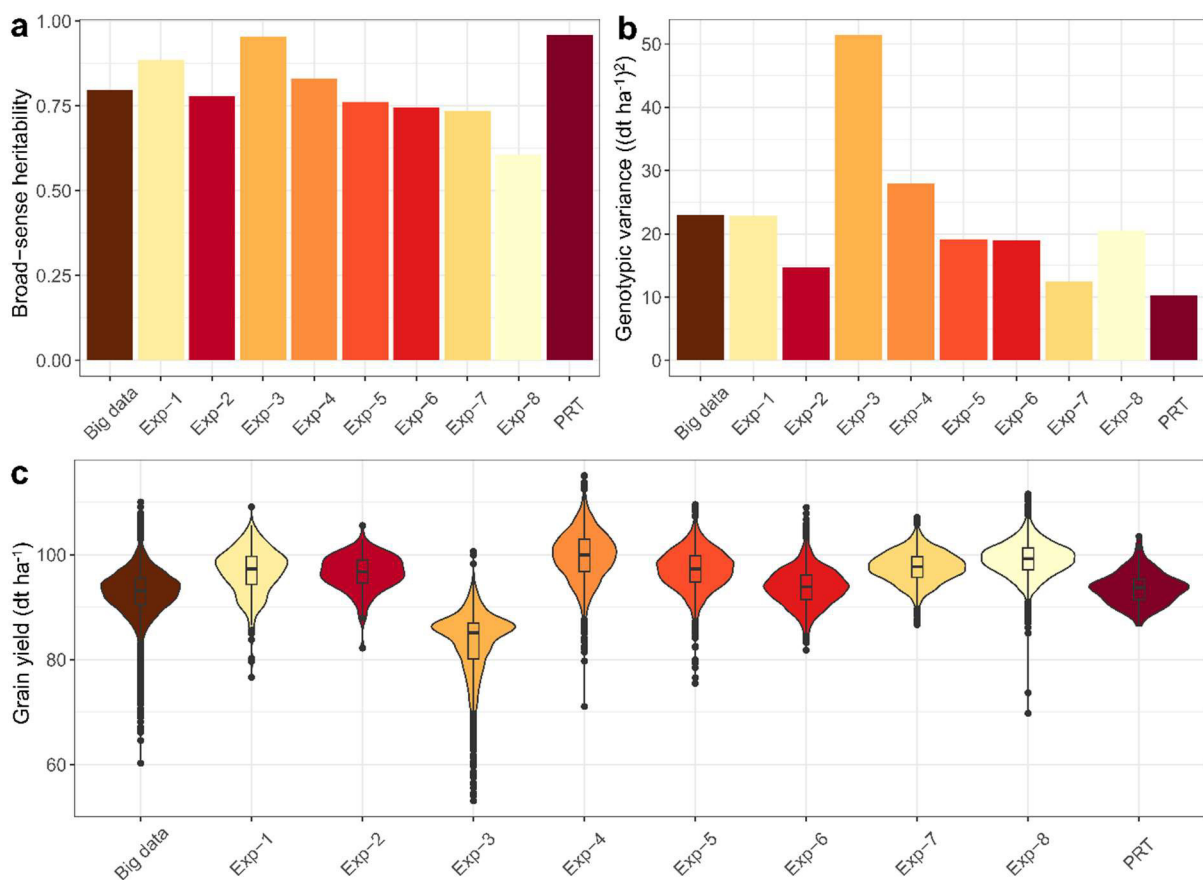


Abbildung 4: Qualität der phänotypischen Daten für Korntrag innerhalb und zwischen den Versuchsreihen (Exp-5-8 sind Daten aus den 4 Zuchtprogrammen; PRT sind Daten aus dem Bundessortenversuchen, Exp-1-3 sind Daten aus abgeschlossenen Drittmittelprojekten). (a) Heritabilität, (b) Genotypische Varianz und (c) Verteilung der adjustierten Mittelwerte.

Für die oben beschriebenen Analysen wurden von IPK-QG Funktionen in der Statistiksoftware R entwickelt. Aufgrund der hohen Spezifität der Datensätze konnten diese jedoch nicht vollständig automatisiert werden. Der Einsatz der entwickelten Funktionen beschleunigte und erleichterte jedoch nach einer intensiven Testphase die Qualitätsüberprüfung.

Die Qualität der kuratierten phänotypischen Daten war ausgezeichnet. Dies wird in Abbildung 4 exemplarisch für das Merkmal Korntrag dargestellt. Die Heritabilitäten waren hoch. Die

genotypische Varianz variierte und war am höchsten in einem Datensatz, der genetische Ressourcen einbezog. Abgesehen von diesem Datensatz beobachteten wir durchweg hohe mittlere Kornerträge. Dies unterstreicht die hervorragende Arbeit der Züchtungsunternehmen.

Weiterentwicklung des Weizen-Informationssystems

Das System wurde durch den Projektpartner IPK-BIT mit einer webbasierten graphischen Benutzungsoberfläche („Wheat Big Data Portal“) implementiert und ermöglicht den Projektpartnern phänotypische sowie genotypische Daten aus verschiedenen Anbaujahren in einem zentralen Datenmanagementsystem am IPK in Gatersleben zu speichern. Die Verwendung des Systems ist durch eine Benutzeranmeldung geschützt und wurde durch eine am IPK etablierte, datenbankgestützte Authentifizierungs- und Autorisierungs-Infrastruktur umgesetzt. User mit Administrationsrechten können neue Nutzer sowie Nutzergruppen anlegen und zuordnen. Somit ist es möglich, dass mehrere Personen eines Projektpartners die Daten dieses Projektpartners mit gleichen Berechtigungen verwalten können. Die Daten der einzelnen Projektpartner sind hierbei für den Zugriff durch andere Projektpartner geschützt. Für den Zugriff auf alle Informationen sind ausschließlich die Datenauswerter autorisiert.

Gemeinsam mit dem Partner IPK-QG wurden umfangreiche Templates sowohl für phänotypische als auch für genotypische Daten entworfen und über die gesamte Projektlaufzeit verbessert sowie weiterentwickelt. Somit existieren die Templates in verschiedenen Entwicklungsstufen bzw. Versionen, die direkt im System inklusive eines vollständigen Änderungsprotokolls dokumentiert und verwaltet werden.

Damit die Projektpartner Ihre phänotypischen und genotypischen Daten im System einlagern können, müssen alle Daten gemäß den Vorgaben in den definierten Templates formatiert und als Microsoft-EXCEL-Files gespeichert werden. Das System selbst stellt wie beschrieben die Templates in verschiedenen Versionen zur Verfügung und bietet eine umfangreiche Hilfe an. Konkret sind dies die vollständige Beschreibung der geforderten Informationen einschließlich Spaltenname, Spaltenbeschreibung sowie Datentyp. Die Informationen werden ähnlich wie in der Excel-Datei in einer Registerkarte angezeigt. Die Abbildung 5 illustriert die Benutzungsoberfläche für die phänotypischen Daten inklusive Version und umfangreiche Beschreibungen.

Phenotypic Template

The easiest way to get started uploading your data is using the latest version of our phenotypic data importation template.

Current Version: 2022.2

Available since: 26/01/2022

Hash: 25E8F885B6AAB80F50AC747A496B6D0362470B9A4D25E1A74398EAD69F486A39

[Download Template](#)

Template Data Entries

Column	Header	Datatype	Required	Description
A	Schlag	string	No	Arable land name.
B	Ort	string	Yes	Location name where experiment was performed.
C	Breitengrad	string	No	Latitude coordinates in the form 'degree N/S'.
D	Längengrad	string	No	Longitude coordinates in the form 'degree O/W'.
E	Höhe	integer (any between -500 and 8000) or NA	No	Altitude in meters above sea level.
F	Ackerzahl	integer (any between 1 and 120) or NA	No	Field number.
G	GPS-Genauigkeit	integer (any between 1 and 1000000) or NA	No	GPS precision.
H	Bodentyp	string from domain list or NA	No	Soil type according to German soil systematics.
I	Bodentextur	string from domain list or NA	No	Soil texture (sand, silt, clay, loam ...)

Abbildung 5: Screenshot vom Leitfaden für phänotypische Daten inklusive der Versionsinformation des Templates.

Analog für die phänotypischen Daten existiert in der Weboberfläche des Systems eine Unterseite für die genotypischen Daten, die in Abbildung 6 dargestellt sind .

Genotypic Template

The easiest way to get started uploading your data is using the latest version of our genotypic data importation template.

Current Version: 2023.2

Available since: 12/12/2023

Hash: AC5E66D5889BFF880BB62E78F7F31FDD88B603083159EF0EE2D06757476AB371

[Download Template](#)

Template Data Entries

Column	Header	Datatype	Required	Description
A	Marker Name	string	Yes	A string that identifies the marker. Each value here must also be found in Marker Info sheet.
B ... XFD	Identifier ⓘ	string	No	SNP Data of the identifier.

ⓘ Identifier is a string for each genotype that you can choose freely but it has to be exactly the same in each sheet of the file. It will connect information from different sheets within the file. To identify the genotype using external names, provide information in the sheet Genotype Info.

Abbildung 6: Screenshot vom Leitfaden für genotypische Daten inklusive der Versionsinformation des Templates.

Die Daten wurden nach dem Hochladen vor der Speicherung auf ihre syntaktische Korrektheit und Vollständigkeit überprüft. Hierbei wurden die in der Aktivität „Qualitätsprüfung der phänotypischen und genomischen Daten“ ausgeführten Ansätze angewendet. Ergebnis ist immer ein Statusreport für die User, der die Erfüllung der Vorgaben mit detaillierten Fehlermeldungen beinhaltet. Weiterhin wurde eine Versionierung der gespeicherten Datendateien implementiert, die es ermöglicht, neue Versionen von Dokumenten mit zusätzlichen Informationen abzuspeichern. Dies ermöglicht es, mit (Teil-)Daten zu arbeiten, die im Laufe des Projekts weiter ergänzt werden können.

Im Upload-System ist eine funktionale Erweiterung integriert, die ohne aktive Benutzerinteraktion arbeitet. Dieses Feature ermöglicht die automatische Identifizierung des hochgeladenen Datei-Typs. Das System unterscheidet aktuell Files für „P – phänotypische“ oder „G – genotypische“ Daten (siehe Abbildung 7).

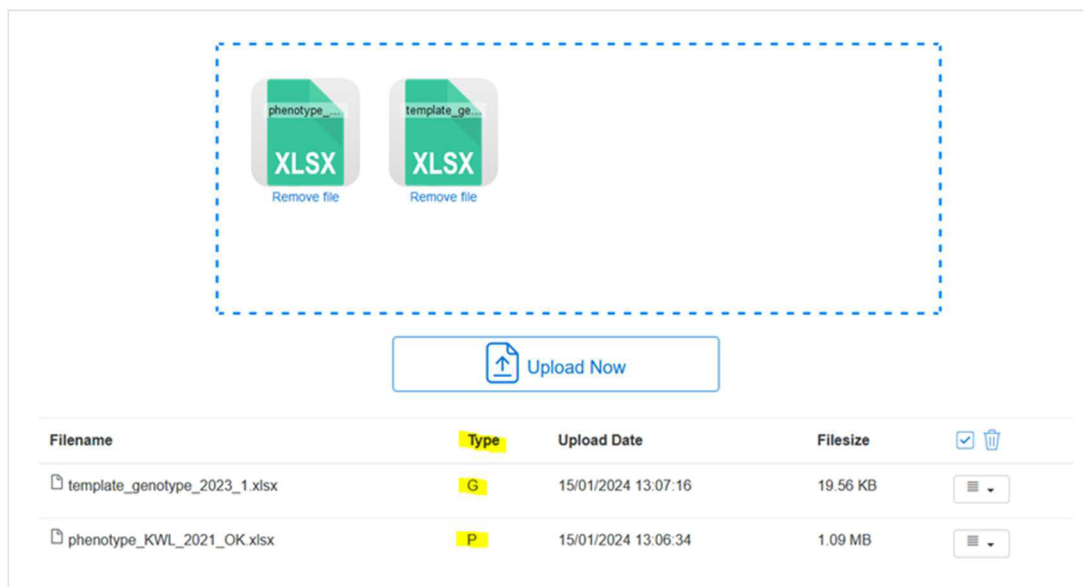


Abbildung 7: Screenshot der automatischen Erkennung des Dateninhalts der Upload-Datei (genotypisch oder phänotypisch).

Die automatisierte Validierung für hochgeladene Files mit phänotypischen und genotypischen Dateien im Wheat Big Data Portal wurde erfolgreich implementiert und gewährleistet die Konsistenz, Vollständigkeit und Leistungsfähigkeit der Daten in allen Dateielementen. Dies umfasst detaillierte Überprüfungen zur Sicherstellung der Datenintegrität sowie -qualität in den verschiedenen Tabellenblättern jeder hochgeladenen Datei und bildet eine Basis für die weiteren Kurationschritte in den anschließenden Auswertungen.

Biometrische Modelle für genomweite Vorhersagen und Assoziationskartierung

Die konsolidierten Daten wurden in Trainings- und Testdatensätze aufgeteilt. Der Trainingsdatensatz wurde für die Kalibrierung der genomischen Vorhersagemodelle verwendet, während der Testsatz ausschließlich zur Validierung der Vorhersagen diente. Die Genauigkeit der

genomischen Vorhersage wurde durch Berechnung der Korrelation zwischen den vorhergesagten und den beobachteten Werten in den Testdatensätzen bestimmt. Als Referenzpunkt dient das klassische GBUP-Modell für die genomischen Vorhersagen:

$$y_i = \mu + a_i + r_i,$$

y_i ist der adjustierte Mittelwert des Genotyps i , a_i der additive Effekt und r_i der Restfehler. Die Elemente des Vektors $a = (a_1, a_2, \dots, a_n)$ folgen einer Normalverteilung $a \sim N(0, K_a \sigma_a^2)$. σ_a^2 ist hierbei die additive genetische Varianz. Neben dem klassischen GBUP-Modell haben wir auch weitere Modelle implementiert, die epistatische Interaktionen berücksichtigen.



Abbildung 8: Mittlere Vorhersagegenauigkeit (Y-Achse) von fünffachen Kreuzvalidierungen, die für Daten jeden Jahres (X-Achse) für alle Merkmale (Zeilen) der vier Unternehmen (Spalten) durchgeführt wurden. p_{year} ist das Jahr für das die Werte vorhergesagt wurden.

Es wurden verschiedene Szenarien zum Testen der Vorhersagemodelle angewendet. Auf Züchterebene führten wir fünffache Kreuzvalidierungen durch. Zusätzlich wurden für jeden Züchterdatensatz einjährige Kreuzvalidierungen vorgenommen, bei denen jeweils ein Jahr als Testdatensatz und der Rest als Trainingsdatensatz verwendet wurde, bis alle Jahre als Testdatensätze eingesetzt wurden. Das vierte Szenario stellt eine Erweiterung der einjährigen Kreuzvalidierungen dar, bei der die Daten der anderen Projektpartner zum Trainingsdatensatz

hinzugefügt wurden. Die Größe des Trainingsdatensatzes wurde auf zwei Arten erhöht: Erstens wurde jeder Datensatz, der in Big Data verfügbar ist, in aufsteigender Reihenfolge der Größe als Trainingsdatensatz verwendet. Nachdem der größte Datensatz zum Einsatz kam, wurde jeder Trainingsdatensatz erweitert, indem die Datensätze kumuliert wurden. Die zweite Methode bestand darin, Zufallsstichproben aus dem gesamten Datensatz zu ziehen.

Für jedes Züchtungsunternehmen führten die fünffachen Kreuzvalidierungen innerhalb eines Jahres für alle drei Merkmale in den meisten Fällen zu moderaten bis hohen Vorhersagegenauigkeiten (Abbildung 8). Dies deutet auf gute Genotyp-Phänotyp-Übereinstimmungen für alle Merkmale in jedem Jahr jedes Unternehmens hin. Die Vorhersagegenauigkeiten der Leave-One-Year-Out-Kreuzvalidierung waren niedriger im Vergleich zu den fünffachen Kreuzvalidierungen (Abbildung 9). Dies lässt sich durch die Interaktionseffekte zwischen Genotypen und Jahren erklären und unterstreicht die Relevanz, umfassende Trainingspopulationen zusammenzustellen.

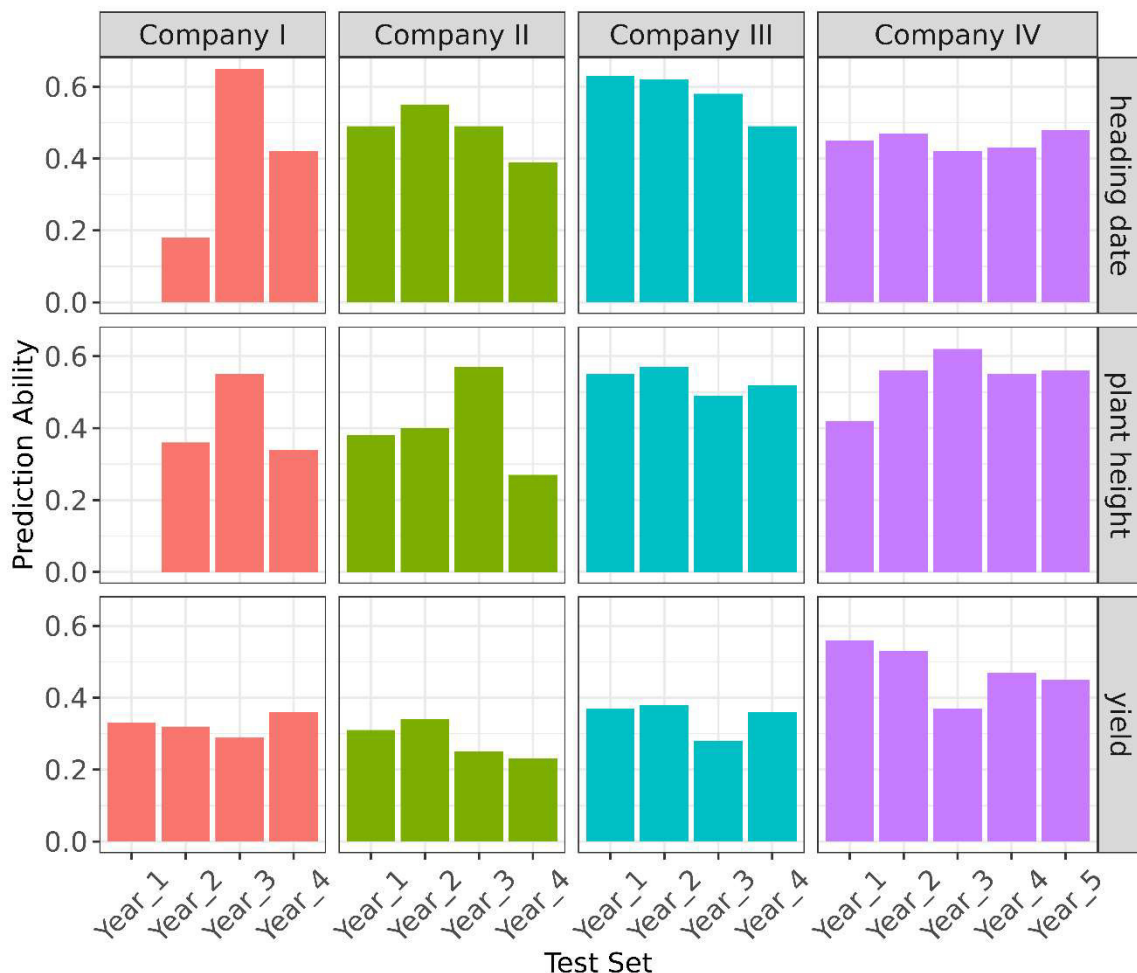


Abbildung 9: Kreuzvalidierungsergebnisse innerhalb von Unternehmen für Leave-One-Year-Out-Kreuzvalidierung.

Bei den Analysen über die verschiedenen Züchter hinweg wurde deutlich, dass die Vergrößerung der Trainingsdatensätze die Vorhersagefähigkeit erhöht (Abbildung 10). Dies belegt den

Nutzen von Big Data für die genomische Selektion. Die Herausforderung besteht jedoch darin, dass Genomvorhersagen bei größeren Trainingsdatensätzen mehr Rechenleistung und Zeit erfordern. Wird andererseits die Größe des Trainingsdatensatzes mithilfe von Zufallsstichproben erhöht, zeigt sich eine erhebliche Variation in der Vorhersagegenauigkeit, die mit zunehmender Größe des Trainingsdatensatzes abnimmt. Diese Ergebnisse verdeutlichen, wie wichtig es ist, trotz großer Datenmengen geeignete Methoden zur Optimierung der Trainingspopulation einzusetzen.

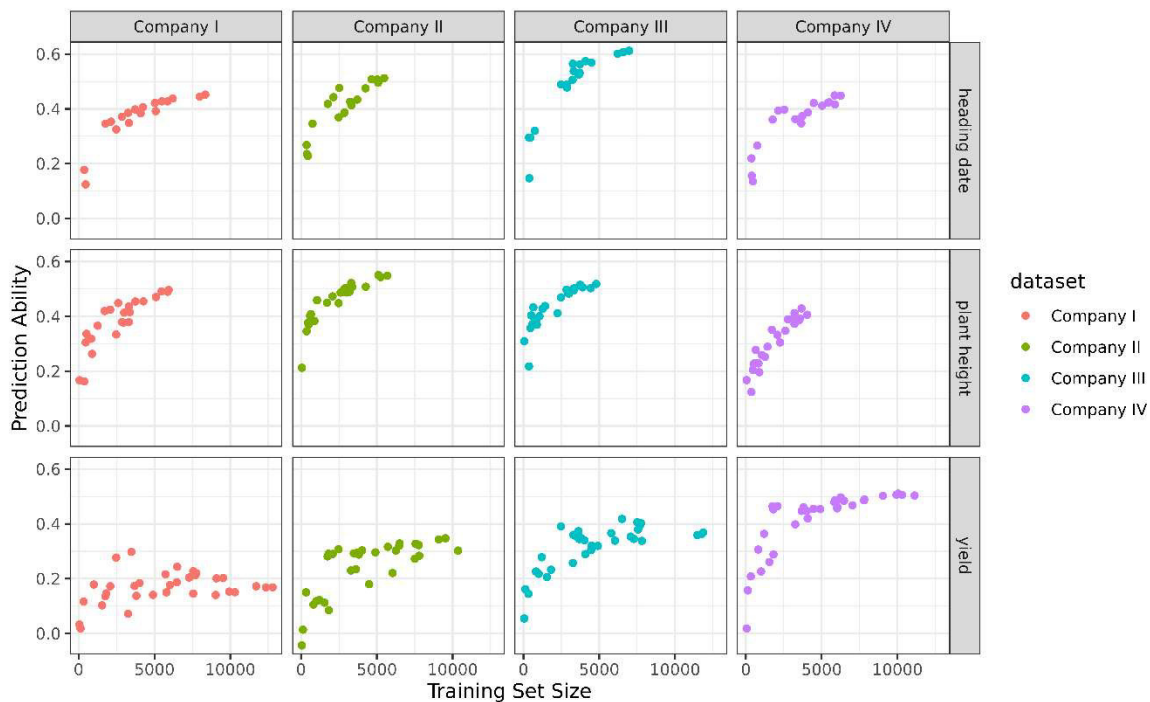


Abbildung 10: Erhöhung der Trainingsdatensatzgröße durch Zusammenführen der einzelnen Datensätze und ihre Auswirkung auf die Vorhersagegenauigkeit.

Als Ansatz des maschinellen Lernens wurde ein sogenanntes Convolutional Neural Network (CNN) ausgewählt. Die Vorhersagegenauigkeit des dynamischen CNN-Modells wurde mit EGBLUP verglichen. Als Testdatensatz dienten die Daten aus den Bundessortenversuchen. Auf den ersten Blick zeigten die Ergebnisse (Abbildung 11), dass CNNs EGBLUP bei der Vorhersage des PRT-Datensatzes nicht übertreffen, mit maximalen Vorhersagegenauigkeiten von 0,6 bzw. 0,63. Das Vorhersageplateau für den Bundessortenversuchsdatsatz wurde bei EGBLUP beginnend mit einer Trainingsdatensatzgröße von etwa 12.000 Genotypen erreicht. Ähnlich erreichen CNNs diese Vorhersagekapazität bei ungefähr der gleichen Trainingsgröße. Bei genauerer Betrachtung fällt aber auf, dass die Korrelationen zwischen der Trainingsdatensatzgröße und der Vorhersagegenauigkeit der CNNs einen höheren Korrelationskoeffizienten (0,74) aufweisen im Vergleich zu EGBLUP (0,62). Dies deutet auf das Potenzial hin, die Vorhersagegenauigkeiten über CNNs weiter zu steigern. Eine weitere Erhöhung der Populationsgröße erscheint dabei von entscheidender Bedeutung zu sein.

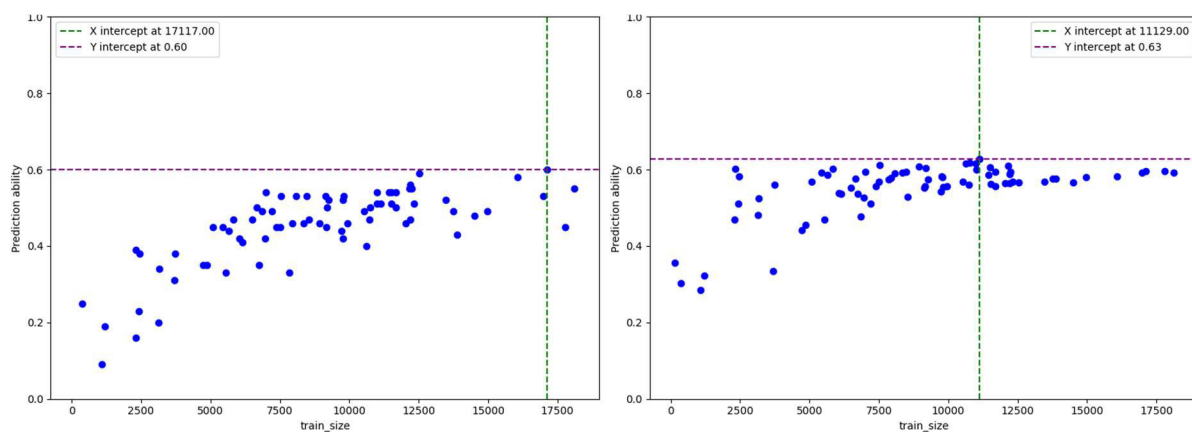


Abbildung 1: Vorhersagegenauigkeit in Abhängigkeit der Größe der Trainingspopulation für Convolutional Neural Network (links) und eines traditionellen EGBLUP Modells (rechts).

Im Rahmen des BigData-Projekts haben wir die Hypothese getestet, dass genomweite Assoziationsstudien (GWAS), die auf Big Data basieren, dazu beitragen, QTLs verlässlicher zu identifizieren und ihre zugrunde liegenden genetischen Effekte präziser zu schätzen. Als Merkmal haben wir uns hierbei auf den Blühzeitpunkt fokussiert. Entgegen den Erwartungen wurden in den größeren, kombinierten Daten weniger Assoziationen zwischen Markern und Merkmalen gefunden als in den einzelnen Versuchsserien. Allerdings war die Vorhersagegenauigkeit auf Grundlage der Marker-Merkmal-Assoziationen des integrierten Datensatzes über die Datensätze hinweg höher (Abbildung 12). Diese Ergebnisse zeigen, dass die Integration von mittelgroßen zu Big Data einen vielversprechenden Ansatz darstellt, um die Trennschärfe zur Identifikation von QTLs mittels GWAS zu erhöhen.

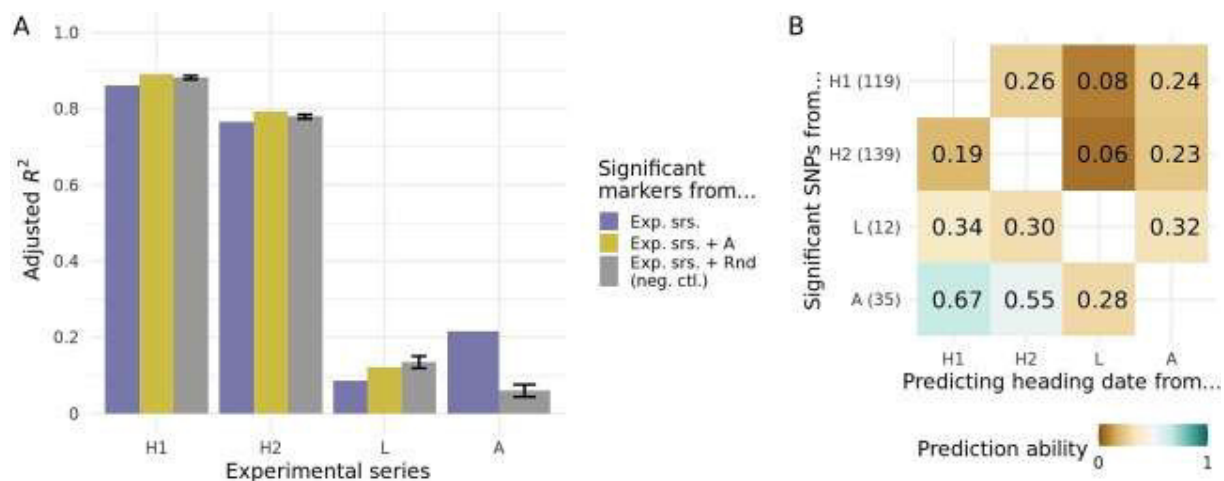


Abbildung 12: Erklärte Varianz und Vorhersagefähigkeit der signifikanten Marker. (A) Phänotypische Varianz, die durch QTL erklärt wird (R^2), mit und ohne Marker, die ausschließlich im gemeinsamen Datensatz A als signifikant befunden wurden. Blau zeigt die erklärte Varianz unter Verwendung der Marker mit signifikanten Marker-Merkmal-Assoziationen, die nur mit Daten der jeweiligen Versuchsserie (Exp. srs.) gefunden wurden. Gelb zeigt die erklärte Varianz unter Berücksichtigung der Marker, die zusätzlich im gemeinsamen Datensatz A als signifikant identifiziert wurden. Grau stellt die negative Kontrolle dar, Fehlerbalken zeigen die Standardabweichung aus 50 Stichprobenläufen. (B) Vorhersagefähigkeit des Blühzeitpunkts unter Verwendung der Effektschätzungen signifikanter Marker. Der Trainingsatz ist auf der vertikalen Achse und der Testdatensatz auf der horizontalen Achse dargestellt.

2. Wichtigsten Positionen des zahlenmäßigen Nachweises

Beschreibung	Plan in EUR	Aufwand in EUR
0812 Wissenschaftler	685.959,56	658.563,65
0817 Angestellte		
0822 Beschäftigungsentgelte	32.397,40	31.747,81
0838 Verbrauchsmaterial	39.365,88	39.553,88
0844 Dienstreisen	3.216,04	607,54
0850 Investitionen mehr als 400 €		
0861 Summe	760.938,88	730.472,88

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Big Data birgt ein enormes Potenzial, um die Genauigkeit der datengestützten Pflanzenzüchtung zu erhöhen und somit den Herausforderungen einer nachhaltigen Ertragssteigerung im Kontext des Klimawandels zu begegnen. Einzelne Unternehmen stoßen jedoch hinsichtlich der Größe ihrer Datensätze an Grenzen, weshalb ein Umdenken hin zu einer gemeinsamen Nutzung von Daten erforderlich ist. Die ersten Schritte in diese Richtung wären ohne die Unterstützung des BMEL/BLE nur schwer realisierbar gewesen. Im Rahmen des Projekts BigData konnte der Mehrwert einer gemeinsamen Nutzung von Daten erfolgreich demonstriert werden. Dabei wurden in den Analyseansätzen neue Wege beschritten. Die bisherigen Publikationen belegen die Angemessenheit und den Erfolg der geleisteten Arbeit.

4. Voraussichtlichen Nutzen im Sinne des Verwertungsplans

Im Projekt BigData wurden Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler im Bereich der Züchtungsinformatik ausgebildet, die ihre Expertise künftig in Unternehmen einbringen werden. Die im Rahmen des Projekts bearbeiteten Fragestellungen haben bereits zu wissenschaftlichen Veröffentlichungen in renommierten, peer-reviewed Fachzeitschriften geführt. Weitere Manuskripte, die auf der einzigartigen Datenbasis und Analyseplattform basieren, wurden erstellt und befinden sich entweder kurz vor der Einreichung oder sind bereits zur Begutachtung eingereicht. Durch die umfangreiche Förderung konnten die beteiligten Gruppen ihre wissenschaftliche und wirtschaftliche Spitzenstellung in den Bereichen Statistische Genomik, Züchtungsinformatik und Züchtungsinformatik weiter ausbauen. So bildete das Konsortium den Nukleus für ein Forschungsprojekt, das seit November 2024 vom BMBF gefördert wird.

5. Bekanntgewordener Fortschritt bei anderen Stellen

Das BigData-Konsortium arbeitete an vorderster Front der Datenwissenschaften in der Pflanzenzüchtung und konnte daher nur begrenzt auf relevante Fortschritte anderer Stellen zurückgreifen.

6. Veröffentlichungen der Ergebnisse

Bereits publiziert:

Lell, M., Y. Zhao, and Jochen C. Reif. 2024. Leveraging the potential of big genomic and phenotypic data for genome-wide association mapping in wheat. *The Crop Journal* <https://doi.org/10.1016/j.cj.2024.03.005>.

Lell, M., J.C. Reif, and Y. Zhao. 2021. Optimizing the setup of multi-environmental hybrid wheat yield trials for boosting the selection capability. *The Plant Genome* 19:e20150. <https://doi.org/10.1002/tpg2.20150>.

Im Begutachtungsprozess:

Thomsen, L.E., R.R. Gundala, Y. Zhao, U. Avenhaus and J.C. Reif. Integrating genomic predictions into an applied Central European wheat breeding program. (In review).

Lell, M., U. Avenhaus, J. Dörnte, W.M. Eckhoff, T. Eschholz, M. Gils, M. Kirchhoff, M. Koch, S. Kollers, N. Pfeiffer, M. Rapp, V. Wimmer, M. Wolf, J.C. Reif, and Y. Zhao. Breaking down data silos across companies to train genome-wide predictions – a feasibility study in wheat. (In review).

Gundala, R.R., U. Avenhaus, J. Doernte, W.M. Eckhoff, J. Foerster⁵, M. Gils, M. Kirchhoff, M. Koch, S. Kollers, N. Pfeiffer, M. Rapp, M. Spiller, V. Wimmer, M. Wolf, Y. Zhao, J.C. Reif. Harnessing big data for enhanced genome-wide prediction in winter wheat breeding (In review)