Sachbericht zum Verwendungsnachweis

Vorhabenbezeichnung:

„Nutzung von Big Data in Weizen zur Präzisionszüchtung (BigData)“

Teilprojekt 3

Förderkennzeichen: 2818408C18

Zuwendungsempfänger:
Saaten-Union Biotec GmbH (SU BIOTEC)
Hovedisser Str. 92
33818 Leopoldshöhe

Ausführende Stelle
Saaten-Union Biotec GmbH
Betriebsstätte Gatersleben
Am Schwabeplan 6
06466 Gatersleben

Laufzeit des Verbundprojektes
01.02.2020 – 31.01.25

"Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Ernährung und Landwirtschaft unter dem Förderkennzeichen 2818408C18 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin/beim Autor".

Sachbericht 2818408C18 Teil I

„Nutzung von Big Data in Weizen zur Präzisionszüchtung (BigData)“ Teilprojekt 3

Im Jahr 2021 wurde das BigData Projekt von der German Seed Alliance GmbH (GSA) an die Saaten Union BIOTEC GmbH (SU BIOTEC) übertragen

I. Kurze Darstellung

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

Im Rahmen des Zuchtprozesses von Sorten sowie in zahlreichen Projekten werden umfangreiche Phäno- und Genotypdaten sowie Metadaten erhoben. Diese Informationen können durch Daten aus anderen Quellen wie Bodenkarten oder Klimadaten ergänzt werden. Das Hauptziel des Forschungsvorhabens BigData ist, durch Bündelung dieser Informationen, das Potenzial von Big Data für die Züchtung von leistungsfähigen Sorten angesichts des Klimawandels zu erschließen. 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 kommerziellen Weizenzuchtprogrammen gebündelt werden. Aufbauend auf diesen Daten sollten biometrische Analyseverfahren implementiert werden, um mit den phänotypischen Daten genotypische Werte zu schätzen.

2. Ablauf des Vorhabens

Die Arbeiten des Verbundprojektes gliederten sich in drei Arbeitspakete (AP). In AP1 implementierte der Partner IPK 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. Die phänotypischen und genomischen Daten wurden integriert und verrechnet. SU BIOTEC hat dafür geno- und phänotypische sowie Metadaten aus den Weizenzuchtprogrammen der DSV, Nordsaat Saatzucht GmbH und WvB in einer eigenen Datenbank zusammengeführt. Dafür wurde das SU BIOTEC Datenbanksystem kontinuierlich und entsprechend den Anforderungen angepasst und weiter ausgebaut.

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. Während der Speicherung erfolgt eine Überprüfung der Daten gemäß der im Projekt definierten Richtlinien und Vorgaben. Ergebnis ist ein Statusreport mit detaillierten Fehlermeldungen. Das AP2 bildete den zentralen Hub für die Übertragung der Daten der Projektpartner und war Grundlage für die Aktivitäten in AP1, welche dann wiederum die Basis für die Arbeiten in AP3 bildeten. Die SU BIOTEC hat die aufbereiteten Daten aus den Weizenzuchtprogrammen der DSV, Nordsaat und WvB aus der eigenen Datenbank in das Datenmanagementsystem des IPKs transferiert.

Im Rahmen von AP3 implementierte und entwickelte der Partner IPK biometrische Modelle für genomweite Vorhersagen und Assoziationskartierung für die umfassenden Daten. Verbesserungen

sollten vor allem aufgrund der Erweiterung der Trainingssatzgröße im Verhältnis zur genetischen Vielfalt ermöglicht werden.

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 den 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

Durch die großen Datenmengen konnte die genomische Vorhersagegenauigkeit im Vergleich zu den bisher durchgeführten firmenspezifischen Vorhersagen signifikant gesteigert werden. Auf das Gesamtprojekt bezogen, war dieses Ergebnis 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, wurde seitens des Partners IPK eine Zusammenarbeit mit dem NFDI-Konsortium FAIRagro begonnen.

Öffentlichkeitswirksame Projektdarstellung (in deutscher und englischer Sprache)

Ziel des Projekts war die Nutzung von Big Data zur Züchtung resilienter Weizensorten durch die Integration phänotypischer und genomischer Daten über verschiedene Züchtungsprogramme hinweg. Hierfür wurde beim ZE das eigene Datenbanksystem kontinuierlich und entsprechend den Anforderungen angepasst und weiter ausgebaut. Die Daten wurden des Weiteren in das vom Partner IPK entwickelte Weizen-Informationssystem integriert, so dass langfristig der Zugriff auf die Daten ermöglicht wird. Die Ergebnisse zeigten, dass Big Data die Vorhersagegenauigkeit signifikant steigern und somit Potenzial für den genetischen Fortschritt in der Züchtung bieten kann.

The aim of the project was to use big data to breed resilient wheat varieties by integrating phenotypic and genomic data across different breeding programmes. To this end, ZE's own database system was continuously adapted and expanded in line with requirements. The data was also integrated into the wheat information system developed by the partner IPK to enable long-term access to the data. The results showed that big data can significantly increase the accuracy of predictions and thus offer potential for genetic progress in breeding.

Sachbericht 2818408C18 Teil II

„Nutzung von Big Data in Weizen zur Präzisionszüchtung (BigData)“ Teilprojekt 3

II. Eingehende Darstellung

1. Verwendung der Zuwendung

Der ZE war an den Arbeitspaketen AP1 und AP2 beteilig.

AP1 Qualitätsprüfung der phänotypischen und genomischen Daten

Um die Prüfung der übermittelten Rohdaten zu ermöglichen, haben sich die Projektpartner auf gemeinsame Standards für die phänotypischen Daten geeinigt. Nur so konnten Ergebnisse aus verschiedenen Herkünften, z.B. Zuchtprogrammen, miteinander verglichen werden. Eine zentrale Rolle stellte hierbei das gemeinsam entwickelte, standardisierte Übertragungsformat (Excel-Template), in welchem alle Roh- und Metadaten zu den Versuchen übermittelt werden. Darüber hinaus werden vom Projektpartner IPK weitere Plausibilitätsprüfungen an den übermittelten Daten durchgeführt, beispielsweise über die ermittelte Spannweite von Merkmalen oder durch Mittelwertvergleiche.

Im Vorfeld abgesprochene phänotypische und genotypische Daten von geprüften Stämmen aus verschiedenen Zuchtstufen der Züchtungsunternehmen DSV, Nordsaat und WvB aus den Erntejahren 2020 bis 2024 wurden um Ausreißer korrigiert und anschließend zusammen mit den verfügbaren Metadaten in die eigene Datenbank der SU BIOTEC geladen. Phänotypische Daten beinhalteten u.a. den Blühzeitpunkt, die Wuchshöhe und den Kornertrag. Aus der Saison 2019/20 wurden insgesamt 42.290 phänotypische Datenpunkte basierend auf 2.800 Stämmen aus verschiedenen Zuchtstufen von DSV, Nordsaat und WvB aufgearbeitet, geprüft und an den Projektpartner IPK übermittelt. In den Tabelle 1-3 sind beispielhaft die phänotypischen Datenpunkte der Erntejahre 2021 - 2023 zusammengefasst.

Tab.1-3: Zusammenfassung der phänotypischen Datenpunkte des Erntejahres 2021 - 2023

Tab. 1: Anzahl ins Projekt übertragener phänotypischer Datenpunkte (DP) in 2021

Zucht-stufe	Ernte-jahr	Anz. GT	DP Ährensch.	DP Wuchshöhe	DP Lager vor Ernte	DP Ertrag	Summe
WP-1	2021	332	3.890	4.273	4.203	8.940	21.306
WP-2	2021	1.313	5.915	7.981	7.084	15.693	36.673
WP-3	2021	5.455	3.325	7.005	6.204	10.664	27.198
Summe	2021	7.100	13.130	19.259	17.491	35.297	85.177

Tab. 2: Anzahl ins Projekt übertragener phänotypischer Datenpunkte (DP) in 2022

Zucht-stufe	Ernte-jahr	Anz. GT	DP Ährensch.	DP Wuchshöhe	DP Lager vor Ernte	DP Ertrag	Summe
WP-1	2022	144	1.577	1.953	150	4.235	7.915
WP-2	2022	712	1.971	2.896	900	5.759	11.526
WP-3	2022	3.136	4.172	800		5.769	10.741
Summe	2022	3.992	7.720	5.649	1.050	15.763	30.182

Tab. 3: Anzahl ins Projekt übertragener phänotypischer Datenpunkte (DP) in 2023

Zucht-stufe	Ernte-jahr	Anz. GT	DP Ährensch.	DP Wuchshöhe	DP Lager vor Ernte	DP Ertrag	DP Gesamt
WP-1	2023	244	3.112	3.598	855	5.935	13.744
WP-2	2023	1.320	5.849	8.746	1.400	12.083	29.398
WP-3	2023	5.145	8.432	6.629	800	13.742	34.748
Gesamt	**2023**	**6.709**	**17.393**	**18.973**	**3.055**	**31.760**	**77.890**

Genotypische Daten wurden mittels Arrayanalyse bei TraitGenetic GmbH (SGS) unter Einsatz des 15K MultiCrop-, der 25K/26K Weizen- und eines 22,5K Multicrop-Arrays durchgeführt. Beispielhaft sind in den Tabellen 4 und 5 Übersichten über die genotypischen Datenpunkte der Erntejahre 2021, 2022 und 2024 aufgeführt.

Tab. 4: Eingebrachte Daten genotypisierter Weizen-Stämme (Stand 2022)

Zuchtstufe	Erntejahr	Anz. GT	Array	genotyp. DP
WP-1	2022	282	25 K (24.145 SNP)	6.808.890
WP-3	2021	2.162	7 K (6.731 SNP)	14.552.422
WP-2	2022	731	7 K (6.731 SNP)	4.920.361
WP-3	2022	2.875	7 K (6.731 SNP)	19.351.625
Summe		6.050		45.633.298

Tab.5: Erhobene genotypische Datenpunkte mittels Arrayanalyse im Erntejahr 2024

Zuordnung	Versuch	Erntejahr	Anz. GT	Array	genotyp. DP
BigData	WP-1	2024	282	25 K	6.808.890
BigData	WP-2	2024	461	7.5 K	3.410.939
In-kind	WP-2	2024	404	7.0 K	2.719.324
BigData	WP-3	2024	3.299	7.0 K & 7.5 K	24.072.629
In-kind	WP-3	2024	942	7.0 K & 7.5 K	6.834.922
Summe			**5.388**		**43.846.704**

Sowohl bei den phänotypischen als auch bei den genotypischen Daten wurde die Anzahl gegenüber dem Projektantrag übertroffen, jeweils durch in-kind Leistungen der beteiligten Züchter des ZE mittels Einbeziehung weiterer Feldexperimente und den zugehörigen Genotypisierungsdaten und darüber hinaus durch Aushandeln aktualisierter Arraypreise.

Die eigene Datenbank wurde für einen reibungslosen Import / Export und spezifische Datenabfragen fortlaufend hinsichtlich ihrer Infrastruktur und den entsprechenden Datenmanagementressourcen weiterentwickelt und erweitert, um somit eine nachhaltige Nutzung der Daten gewährleisten zu können.

AP2 Weiterentwicklung des Weizen-Informationssystems
Mittels des BigData-Templates vom Partner IPK wurden die o.g. Daten in das Datenmanagementsystem des IPKs transferiert und von dem Partner qualitativen Prüfungen unterzogen.

Insgesamt wurden über 300.000 phänotypische Datenpunkte basierend auf Stämmen aus verschiedenen Zuchtstufen von DSV, Nordsaat und WvB an den Projektpartner IPK übermittelt. Damit eine nachfolgende statistische Analyse der Feldversuche erfolgen konnte, wurden die phänotypischen Daten möglichst aller Stämme eines Feldexperiments, z.B. 8 x 8 Gitter, geliefert, auch wenn nicht alle Stämme für die Genotypisierung vorgesehen waren.

Nicht von allen Versuchen konnten voll umfänglich Metadaten hinsichtlich der Klimadaten und der geographischen Lage zur Verfügung gestellt werden. In diesen Fällen kann der Projektpartner IPK ggfls. auf Daten der in der Nähe des Versuchsstandorts gelegenen offiziellen Wetterstationen zurückgreifen.

2. Darstellung der wichtigsten Positionen des zahlenmäßigen Nachweises

Position	Gesamtvorkalkulation (€)	Gesamtnachkalkulation (€)
0823 Fremdleistungen	263.751,77	266.857,92
0837 Personalkosten	394.280,36	398.013,82
0838 Reisekosten	800,00	200,28
0855 **Summe** unmittelbare Vorhabenkosten	658.832,13	665.072,02
0881 **gesamte Selbstkosten des Vorhabens** (Summe Pos. 0855 – 0860)	329.416,07	335.655,96

3. Darstellung der Notwendigkeit und Angemessenheit der geleisteten Projektarbeit,
Die Notwendigkeit und Angemessenheit der geleisteten Projektarbeit war gegeben. Wegen der großen Zahlen an zu bearbeitendem Material, der Verarbeitung aller geno- und phänotypischen Daten in den entsprechenden Templates für ihre Einbindung in die beiden Datenbanken (SU BIOTEC, IPK) sind solche Projekte nur mit staatlicher Förderung durchführbar.

4. Darstellung des voraussichtlichen Nutzens, insbesondere der Verwertbarkeit des Ergebnisses – auch konkrete Planungen für die nähere Zukunft – im Sinne des fortgeschriebenen Verwertungsplans,
Die Grundidee und die erhobenen Daten werden in das Projekt DRIVE „Data-driven and genome-edited breeding of locally-adapted wheat varieties to enhance agricultural biodiversity, sustainable climate resilience, and resource efficiency“ einfließen mit dem Ziel, Innovationen in den Bereichen Datenwissenschaft, Präzisionszüchtung und moderne Züchtungsinstrumente für die Züchtung von klima- und standortangepassten, widerstandsfähigen Weizengenotypen nutzbar zu machen.

5. des während der Durchführung des Vorhabens dem ZE bekannt gewordenen Fortschritts auf dem Gebiet des Vorhabens 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. der erfolgten oder geplanten Veröffentlichungen des Ergebnisses nach Nr. 5 der NABF
Die Veröffentlichung der Ergebnisse übernimmt der Projektpartner IPK.