GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Veröffentlichung der Ergebnisse von Forschungsvorhaben im BMBF-Programm
11.5 NKBF98

Biologische Innovation und Ökonomie

Förderkennzeichen: 0314000C

Forschungsvorhaben: Verbundprojekt: 'Aufklärung des Gerstegenoms: Verankerung der physikalischen karte des Gerstegenoms mittels explorativer Genomsequenzierung' (Teilprojekt C)

Zuwendungsempfänger: Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt, GmbH (HMGU),
Postfach 11 29, 85758 Oberschleißheim

Projektleitung: Herr Dr. Mayer

Laufzeit: 01.07.2007 bis 31.03.2011

"Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 0314000C gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor".

Project Title (+ Acronym)	GABI-BARLEX - Aufbau einer genetisch verankerten physischen Karte des Gerstengenoms als Plattform für gezielte Genisolierung in den Triticeae Getreidespezies und als Grundlage für die Genomsequenzierung in Gerste, Teilvorhaben C
Identific. Number (FKZ)	**0314000C**
Grant holder (+ Principle Investigator)	Dr. Klaus Mayer
Project Partners (if appl.)	Dr. Nils Stein (IPK Gatersleben) Dr. Uwe Scholz (IPK Gatersleben) Dr. Matthias Platzer (FLI, Jena) DR. Frank Ordon (JKI, Quedlinburg)
Project Koordinator	Dr. Nils Stein
Subcontractor(s)	n.a.

I. Kurzbeschreibung der Projektfragestellungen

Ziel des gesamten Projektes war die Erstellung einer vollständigen physischen Karte des Gerstengenoms und deren Verankerung auf die genetische Karte unter Einbindung aller wesentlicher Sequenzresourcen und Markerdaten. Gleichzeitig sollte das Genkomplement von Gerste identifziert und soweit wie möglich entlang der Chromosomen angeordnet werden. Die Grundlage der physischen Karte bildete eine BAC Bibliothek aus 571,000 Klonen (ca 13 fache Genomabdeckung), die mittels High Information Content Fingerprinting (HICF) charakterisiert und ebenfalls im Rahmen des BARLEX Projektes vom IPK zu fingerprinted contigs (FPcontigs) assembliert worden sind.

Bei den zu integrierenden Sequenzresourcen handelte es sich überwiegend um neuartige Daten, die innerhalb von Barlex mit den 'next generation' Sequenziertechniken generiert worden sind, wie z.B. 3,500 BACs, 'whole genome shutgun' assemblies und 454 reads von sortierten Chromosomarmen. Darüberhinaus sollten auch wichtige ergänzende Datenquellen von Kollaborationspartnern des IBSC (International Barley Sequencing Consortium) integriert werden, um ein möglichst vollständiges Bild über das Gerstengenom zu erhalten.

Die Ergebnisse von BARLEX sollten öffentlich zur Verfügung gestellt werden und u.a. die effiziente Isolierung jeden Gens aus Gerste, sowie auch aus verwandten Arten wie Weizen und Roggen, wesentlich erleichtern. Desweiteren bilden die in BARLEX gewonnenen Resourcen und Erfahrungen eine wertvolle Grundlage für den nächsten anvisierten Schritt: der vollständigen Sequenzierung und Anordnung des mit 5 Gb bisher größten Genoms.

II. Ziele

1. Geplante Ziele

Das vorliegende bioinformatische Teilprojekt (TP5) hatte folgende spezifischen Aufgaben:

(1) Virtuelle Genkarten: wie gut läßt sich das Gen Inventar von Gerste aus kurzen genomischen Sequenzen mit Hilfe von Syntenie basierten Ansätzen zusammenstellen?

(2) Strukturelle Genom Annotation: Entwicklung und Anwendung von effizienten Gen- und Transpsoson Annotationspipelines

(3) Arbeiten zur Integration der physischen und genetischen Karte von Gerste

(4) Webdarstellung der Daten und Ergebnisse: Ausarbeitung und Umsetzung von Konzepten zur Präsentation der komplexen Datenstrukturen, öffentliche Zugänglichkeit der Daten über die IBIS Webseiten

2. Erreichte Ziele

2.1 Virtuelle Genkarten

Im Rahmen des Projektes wurde ein neuartiges bioinformatisches Analysetool zur Erstellung von virtuellen Genkarten aus genomischen Sequenzreads durchfluss-zytometrisch sortierter Chromosomen (-arme) entwickelt. Das GenomeZipper genannte Verfahren nutzt die gut erhaltenen syntänen Beziehungen zwischen den Grasgenomen aus und verzahnt wie ein Reißverschluss diejenigen Gene der Referenzgenome miteinander, die eine gute Homologie zu Gersten genmischen Sequenzen aufweisen. Als Referengenome wurden Reis, Hirse und *Brachypodium* verwendet. Bei Hirse und Brachypodium war die MIPS Arbeitsgruppe maßgeblich an der Gen und Transposon Annotation beteiligt. Die entsprechende "*in house*" Expertise und Datensammlung hat das Projekt entscheidend vorangebracht.

Durch die Einbindung in ein Gerüst aus ~3000 genetischen Markern von Gerste lassen sich die Gersten Sequenzreads zusammen mit ihren gut charakterisierten Genpartnern aus den Referenzgenomen entlang der einzelnen Gersten Chromosomenarme anordnen (Abb1). Zusätzlich wurden die entsprechenden Volllängen cDNAs aus Gerste (Kooperation mit NIAS, Japan) dazu assoziiert. Die “proof of principle” Analysen für den GenomeZipper wurden mit den zuerst vorliegenden Sequenz Daten von Chromosome 1 aus Gerste erfolgreich durchgeführt und im Oktober 2009 publiziert (Mayer et. al, Plant Phys 2009).

Im folgenden wurden der GenomeZipper weiterentwickelt und für alle sieben Gersten-chromosome berechnet (Abb.2, Tab.1). Die Weiterentwicklung betraf vor allem die Einbindung zusätzlicher Gersten Sequenz Resourcen, wie BAC Sequenzen (im Projekt generiert) und Arrayhybridisierungsdaten (vom Kollaborationspartner SCRI). Die gleichzeitige Verwendung dieser Datentypen lässt es nun zu nahezu 22.000 Gene entlang

des Gerstengenoms exakt zu positionieren und etwa 20.000 Gersten Volllängen cDNAs anzuordnen. Diese hochauflösende Anordnung von rund 2/3 der geschätzten Gerstengene erlaubt es genaue vergleichende Analysen gegen Modellgenome (Reis, Sorghum, Brachypodium) durchzuführen. Zahlreiche Analysen die Vorhandensein/Nichtvorhandensein, Kolinearität, Grad der Konservierung und evolutionäre Muster umfassen wurden durchgeführt und geben nun ein Bild des Gerstengenoms dass in seiner Genauflösung ähnlich einer vollständigen Genom Sequenz ist (Abb.3, Abb.4). Mit dieser detaillierten Charakterisierung des Genkomplements von Gerste wurden die bei Antragsstellung erwarteten Ergebnisse deutlich übertroffen. Die entsprechenden Ergebnisse wurden 2011 mit dem Titel "Unlocking the barley genome by chromosomal and comparative genomics" in Plant Cell publiziert (Mayer KF et al., Plant Cell. 2011).

Zu Ende des Projektes wurden noch Teile des 30-fachen Gersten Whole Genome Shotgun Assemblies (1.6 Gb, 2.3 Mio contigs, 0.3x Abdeckung) und 2,000 der 3,500 sequenzierten BACs in den GenomeZipper eingebunden.

Im Laufe des Projektes hat sich herausgestellt, daß der GenomZipper ein allgemein anwendbares Prinzip zur Herstellung von virtuellen Genkarten für die bisher nicht vollständig sequenzierten großen Grasgenome darstellt. Die an definierten Chromsomenpositionen verankerten Gene liefern auch für die Pflanzenzüchtung wertvolle Informationen zur Verbindung von Phänotypen mit konkreten Gen Loci.

Erste Versuche mit dem 16Gb Weizengenom haben gezeigt, dass der GenomeZipper auch hier sehr gut funktioniert (Abb.5).

2.2 Strukturelle Genom Annotation

Von grundlegender Wichtigkeit war die Fertigstellung des *Brachypodium distachyon* Genoms im Projektzeitraum mit 2,3 Millionen assozierten EST reads und einer strukturellen Genannotation die in ihrer Qualität nach verschiedenen Abschätzungen und Evaluierungen die Qualität des diploiden Models *Arabidopsis thaliana* nach einem Jahrzehnt manueller Kuratierung erreicht (!). Dieses hochqualitative *Poideae* Referenzgenom ist eine essentielle Basis, um die homologie- und evidenzbasierende strukturelle Genannotation in Gerste mit hoher Spezifität durchzuführen.

Während des Projektes wurde eine leistungsfähige Hochdurchsatz Gen Annotationspipeline weiterentwickelt, die von den neuen Erfahrungen und Erkenntnissen im Bereich der strukturellen Grasgenom Annotation profitiert hatte. Die Pipeline enthält genomspezifische Trainingsschritte und verbindet die Ergebnisse verschiedener Gendetektions Programme in einem kombinatorischen Ansatz. Die strukturellen Annotationsdaten wurden durch Vergleiche gegen Referenzgenome und gegen EST Bibliotheken überprüft sowie indirekten Assoziationen mit *shotgun read* Daten unterzogen. Die ermittelten Daten belegen hochwertige strukturelle Gencalls und erlauben bisher nicht verankerte BAC Klone individuellen Chromosomen zuzuordnen. Jeder der 3.500 aus 454 Sequenz-*reads*

assemblierten Gersten BACs (Steurnagel et al., BMC Genomics. 2009) wurde einer Gen Annotation unterzogen, die dabei charakterisierten 10.155 Gene, sind auch auf der MIPS Webseite öffentlich zugänglich (mips.helmholtz-muenchen.de/plant/barley/bacs/index.jsp).

Da über 80% des Gerstengenoms aus Transposons und deren Überresten bestehen, musste auch ein besonderer Augenmerk auf die Repeat Detektion gelegt werden. Sämtliche Analysen, die Gene (z.B GenomeZipper, Gendetektion) oder Verankerungsfragen betreffen wurden auf den zuvor repeatmaskierten Sequenzen durchgeführt, um falsche Gen-Vorhersagen, Ambiguitäten und überlange Rechenzeiten zu vermeiden. Im Projektverlauf wurde die eigene repeat Datenbank (mipsREdat) auch mit neu detektieren Gersten LTR-Retrotransposons (1.123 Elemente aus den 454 sequenzierten BACs) und anderen Poaceae spezifischen Transposons kontinuierlich aufgefüllt. Die aktuelle Version 8.6 von mipsREdat enthält zur Zeit 28.090 Elemente mit einer Gesamtgröße 255 Mb. Gleichzeitig wurde eine effiziente Homologie basierte Repeat Maskierungs und Annotations Pipeline aufgebaut, die die große Datenfülle der "next generation" Sequenzen gut bewältigen kann.

Die im Projekt gewonnen detailreichen Ergebnisse der strukturellen Gen und Repeat Annotation und ihrer Korrelationen fließen in eine internationale Publikation über das gesamte Gerstengenom ein, die sich zur Zeit in der Vorbereitung befindet.

2.3 Integration der physischen und genetischen Karte von Gerste

Die vom IPK erstellten "finger printed contigs" (FPcontigs) bilden das Grundgerüst der physischen Karte. Sie basieren auf ca. 500.000 fingerprinted BAC Klonen (14 fache Genomabdeckung), die zu rund 9.000 contigs assembliert worden sind. Zusätzlich gibt es unterschiedliche BAC basierte Sequenz und Mapping Datensätze, die sowohl aus dem vorliegenden (wie z.B. 3,500 sequenzierten BAC Klone und fingerprinted BACs) als auch von Kollaborationspartnern (z.B. BAC end Sequenzen (BES), ILL-SNP Marker BAC Zuordnung, Harvest35 EST assemblies) stammen. Um all diese sehr heterogenen Daten miteinander vernetzen zu koennen wurden Parser und ein gemeinsames Datenbankschema entwickelt und implementiert. Insgesamt wurden etwa 30 verschiedene Datensatz Dateien mit Hilfe der Datenbank in einem gemeinsamen Datenformat vereinigt und für weitere Analysen zugänglich gemacht.

Die integrierten Daten (Abb.6) ermöglichten es die anfangs völlig sequenzfreien FPcontigs mit Sequenzinformationen aus BES, 3500 sequenzierten BAC Klonen, Harvest35 cDNAs und WGS Assemblies zu „dekorieren“. Die Verknüpfung der einzelnen FPcontigs zur genetischen Karte erfolgt über BAC-Klone, auf denen sich ILL-SNP Marker befinden. Als zusätzliche Validierung der Zuordnungen wurde ein bioinformatisches Verfahren entwickelt, das es erlaubt die chromosomenarmspezifische Herkunft von Gerstensequenzen zu bestimmen. Die Methode beruht auf der Sequenzhomologie zu den nicht repetitiven Bereichen der chromosomensortierten *read* Sequenzen.