GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Veröffentlichung der Ergebnisse von Forschungsvorhaben im BMBF-Programm

Pflanzenbiotechnologie - Verbundvorhaben: 'Erforschung von Triticeae-Genomen per Hochdurchsatz - Sequenzierung (TRITEX)' - Teilprojekt C

Förderkennzeichen: 0315954C

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

Ausführende Stelle: Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH) - Genomik und Systembiologie der Pflanzen (PGSB), Ingolstädter Landstr. 1, 85764 Oberschleißheim

Projektleitung: Herr Dr. Klaus Mayer

Projektlaufzeit: 01.07.2011 bis 30.11.2014

"Das diesem Bericht zugrundeliegende BMBF-Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 0315954C gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor".

Funding Activity	„Plant Biotechnology for the Future“
Project Title (+ Acronym)	TRITEX: exploring Triticeae genomes on the basis of an advanced draft of the barley genome
Identific. Number (FKZ)	0315954C
Grant holder (institution/company/…)	Klaus Mayer PGSB Plant Genome and Systems Biology, Helmholtz Center Munich, German Research Center for Environmental Health (GmbH)
Principle Investigator	Klaus Mayer
Project Partners (if appl.)	N.Stein (IKP, GED), U.Scholz (IPK, BIT), T.Schnurbusch (IPK, PBP), M. Platzer (FLI)
Subcontractors (if appl.)	-
Project Coordinator (if consortium)	Nils Stein (IPK)

Übersicht

I. Kurzbeschreibung der Projektfragestellungen

Weizen und Gerste, zwei agronomisch bedeutsame Mitglieder der *Triticeae*, stehen nach Mais und Reis an dritter Stelle in der weltweiten Nahrungsmittelproduktion (faostat3.fao.org). Aber bedingt durch ihre sehr großen und entsprechend hochrepetitiven, komplexen Genome liegt die Erforschung ihrer genomischen Resourcen und damit verbunden eine effizientere Ausnutzung ihres Züchtungspotentials im Vergleich zu Mais und Reis deutlich zurück. Um diesen Mangel aufzuholen sollten im TRITEX Projekt stufenweise neuartige Sequenz- und Positions-Daten hergestellt und zusammen mit den Daten externer Kooperationspartner mittels bioinformatischen Methoden so weit wie möglich zu Genomgerüsten verbunden werden. Ein wichtiges Hilfsmittel dazu waren die für Gerste und Weizen über durchflußzytometrische Sortierung pro Chromosomen Arm erzeugten Sequenz Datensätze, mit denen das Assemblierungsproblem auf kleinere Maßstäbe heruntergebrochen werden konnte. Bei der Datenintegration wurde auch das bei den Gräsern vorhandenen Phänomen der konservierten Gen Reihenfolge (=Syntenie) ausgenutzt, mit dem sich die großräumige Struktur von bereits vollständig assemblierten kleineren Genomen, wie Reis oder *Brachypodium* auf noch nicht assemblierte Genome übertragen läßt.

Nach höherer Genauigkeit und entsprechend steigendem Kosten- und Arbeitsaufwand sortiert, lassen sich drei Stufen der Assemblierung von *Triticeae* Genomen unterscheiden:

1. der GenomZipper, ein indirekter Syntenie basierter Ansatz zur Bestimmung chromosomaler Genreihenfolgen, der nur eine genetische Karte und 'low pass' Sequenzen benötigt.
2. die 'Gen-dekorierte' physikalische Karte, ein genzentrischer Ansatz, der die Mehrzahl der Gene und andere 'low copy' Sequenzen, soweit verfügbar, bereits in ihren physikalischen Kontext stellt, aber noch keine Anordnug der repetitiven Transposon Bereiche liefert.
3. die vollständige BAC bei BAC Sequenzierung des minimal überspannenden Pfades der physikalischen Karte. Dieser sehr aufwändige Ansatz sollte neben den 'high copy' Regionen auch alle unter 2. noch fehlenden Gene erfassen.

Die bioinformatischen Fragestellungen des vorliegenden Projektes konzentrierten sich mit GenomZipper und 'Gen-dekorierter' physikalischer Karte auf die ersten beiden Assemblierungsstufen. Am Ende des Projektes wurden als Vorläuferstudie auch noch erste Analysen zu einem vollständig BAC bei BAC sequenzierten Gerstenchromosom durchgeführt.

II. Ziele

1. Geplante Ziele

Als Bioinformatik Partner (P5, Klaus Mayer, HMGU) waren wir im wesentlichen für drei Aufgabenbereiche zuständig:

- die strukturellen Annotation (Gene, Transposons u.a. Repeats) der im Projekt anfallenden genomischen Sequenzdaten und ihrer vergleichenden Analyse, sowie die Bewertung unterschiedlicher Assemblierungs Strategien.
- die Integration von hoch-heterogenen Sequenz und genetischen Marker Datensätzen zu chromosomalen Sequenzgerüsten mit einer Genanordnung entlang der Chromosomen, der Methoden und Parameter Validierung und der kontinuierlichen Verbesserung durch technologisch neuartige Datensätze.
- die Entwicklung und Umsetzung von Web Darstellungskonzepten, um die vielfältigen TRITEX Daten und Ergebnisse in übersichtlicher Weise der wissenschaftlichen Gemeinschaft als Grundlage für weitergehenden Forschungen zur Verfügung zu stellen.

Die folgenden vier Ziele waren innerhalb der zwei Arbeitspakete WP3.1 (task 3) und WP4 (task 3,4,5) im Antrag für die bioinformatischen Arbeiten von P5 definiert worden.

1. Architektur des Gerstengenoms: strukturelle Annotation und chromosomale Anordnung der Gene (WP 4, task 4) und explorative BAC Annotation (erweiterte Ziele der kostenneutralen Verlängerung)
2. Virtuelle Genkarten für alle 21 Weizenchromosomen (WP 4, task 3)
3. Physikalische Karte des Weizen Chromosoms 6A (WP 3.1, task 3)
4. Webinterfaces zur Daten Darstellung und Weitergabe (WP 4, task 5)

2. Erreichte Ziele

2.1. Architektur des Gerstengenoms: strukturelle Annotation und chromosomale Anordnung der Gene (WP4, task 4) und explorative BAC Annotation (erweiterte Ziele der kosten neutralen Verlängerung)

Mit der Etablierung der neuen Hochdurchsatz-Sequenziertechnologien, die zu Beginn des Projektes zur Verfügung standen, war das Gerstengenom, mit 5Gbp etwa doppelt so groß wie das menschliche Genom, erstmalig in seiner Gesamtheit (WGS, whole genome sequencing) sequenzierbar geworden. Allerdings ließen sich die erhaltenen kurzen Sequenzbruchstücke aufgrund der enormen, durch 80% Transposongehalt verursachten, Repetitivität (Tab 1) nicht ohne zusätzliche Positionsinformationen in ihren chromosomalen Kontext einordnen. Um diese Aufgabe zu bewältigen wurden im TRITEX Projekt bioinformatische Methoden konzipiert, entwickelt und validiert, die es letztendlich ermöglichten die hoch-heterogenen Datensätze in Form von (1) verschiedenen Sequenztypen (WGS Contigs, BAC Enden, BAC Klone), (2) physikalischen Karten (BAC fingerprint Assemblierungen)

und (3) mehreren sich ergänzenden genetischen Marker Karten miteinander zu verbinden (Abb 1). Das dabei entwickelte hochauflösende gen-zentrische Grundgerüst des Gerstengenoms bildete zusammen mit tiefer gehenden Genexpressions Analysen einen wesentlichen Bestandteil des vom "International Barley Sequencing Consortium" (IBSC) in Nature veröffentlichten Artikels "A physical, genetic and functional sequence assembly of the barley genome" [32].

Die der IBSC Publikation zugrunde liegende physikalische Karte von Gerste ist einer eigenen Publikation genauer beschrieben worden [14]. Sie besteht aus 9.265 'finger printed' BAC contigs, mit einer kumulativen Länge von 4,98 Gb, was einem Abdeckungsgrad von 95% des 5,1 Gb großen Genoms entspricht. 3.9 Gb davon konnten über genetische Marker (3,241 SNP und ~ 500,000 GBS Marker) linear entlang der 7 Chromosomen angeordnet werden, weitere 0.7 Gb ließen sich zumindest einem der 14 Chromosomenarme zuordnen. An direkt integrierten BAC Sequenzen standen letztendlich 6.278 überwiegend gentragende Klone und 304.523 BAC End Sequenzpaare mit zusammen ~ 1,1 Gb zur Verfügung. Zusätzlich konnten noch 112.989 WGS contigs (0.3 Gb des 1.9 Gb großen WGS Assemblies) über Sequenz Homologie an spezifischen Positionen der physikalischen Karte und weitere 1.881 and Chromosomenarmen verankert werden.

Das Transkript- und Geninventar von Gerste wurde durch Mapping von umfangreichen RNSseq Datensätzen (1.67 * 106 reads, 167 Gb) und 28.000 volllängen cDNAs auf den 1.9 Gb WGS Contigs identifiziert. Von den 79.379 gefundenen Transkript Clustern sind 26,159 aufgrund ihrer guten Homologie zu bekannten Genen als 'high confidence (hc)' klassifiziert worden. Viele der restlichen 53.220 'low confidence' Transkript Loci bestehen aus kurzen stark degenerierten Genfragmenten, die durch die Aktivität von Transposons überall im Genom verteilt worden sind. Eine Vollständigkeitsanalyse der 26.000 hc Gene ergab 86%, was zu einer Gesamtabschätzung von 30.400 klassischen proteinkodierenden Genen für Gerste führen würde. Insgesamt ließen sich 24.154 (92%) der hc Gene an dem Genomgerüst verankern: 15.719 mit direkter linearer Abfolge, die restlichen über Chromosomenarmzugehörigkeit oder das Syntenie Konservierungs Modell.

Eine sorgfältige Analyse der umfassenden gewebe- und entwicklungsspezifischen RNAseq Expressionsdaten, zeigte eindrucksvoll, daß nicht allein die Transkriptionsraten für die Menge an funktionalen Genprodukten verantwortlich sind. Es gibt eine bisher unterschätzte zusätzliche regulatorische Ebene, die über alternatives splicing und vorzeitige stop codons die Syntheseraten der Protein Endprodukte in Abhängikeit von raumzeitlichen Bedingungen differentiell steuert. Für die Züchtungsforschung läßt sich daraus die wichtige Erkenntnis ableiten, daß auch agronomisch interessante Merkmalsausprägungen möglicherweise nicht aussschließlich an proteinkodierenden Sequenzen und deren Mutationen festzumachen sind. Das Vorhandensein und die Menge an funktionsfähigen Genprodukten wird darüberhinaus noch von bisher wenig verstandenen hochkomplexen Regulationsmechanismen beeinflußt, deren weitere Erforschung aufschlußreiche Erkenntnisse verspricht.

Abbildung 2 zeigt die chromosomale Architektur von Gerste mit gendichteren Chromosomenenden (d) und stark ausgedehnten, LTR-Retrotransposon dominierten, pericentromerischem Heterochromatin (e). Obwohl die Gendichte dort geringer ausfält, befindet sich allein schon wegen der Größenausdehnung ein nicht unbeträchtlicher Teil der Gersten Gene in Bereichen mit sehr niedriger Rekombinationsrate und ist damit für die durch Rekombination verursachten allelischen Genaustauschereignisse nicht zugänglich. Assoziationsanalysen zur Identifizierung von

merkmalsbestimmenden Genen und züchterischen Methoden, die auf Rekombination beruhen sind für diese "eingeschlossenen" Gene wenig erfolgsversprechend und sollten generell bei allen Getreidearten durch alternative Ansätze ergänzt werden.

Mit der im TRITEX Projekt entwickelten neuartigen Assemblierungs-Datenresource des gen-zentrischen Gersten "Gene-omes" liegen jetzt erstmals die chromosomalen Positionen für den überweigenden Teil aller Gersten Gene vor. Gerste ist mit 5,1 Gb im Moment noch das größte Genom von dem eine solch detaillierte physikalisch verankerte Genkarte erstellt werden konnte. Obwohl es sich "nur" um einen vorläufigen Entwurf handelt, ist die von uns im TRITEX Projekt erreichte gen-zentrischen Assemblierung des Gerstengenoms für viele wissenschaftliche Fragestellungen und züchtungsrelevanten Anwendungen schon ähnlich brauchbar wie die im klassischen Sinn in ihrer Gesamtsequenz fast vollständig assemblierten kleineren Genome. Die jetzt vorliegende hochauflösende und physikalisch verankerte Genkarte von Gerste wurde auch als nützliche Referenz für das nah verwandte und hochkomplexe hexaploide Weizengenom (17 Gbp) verwendet [u.a. 6, 33]. Die unter 2.4 beschriebenen Gersten Webinterfaces und die über FTP bereitgestellten Daten werden von der Triticeae Forschergemeinschaft gut genutzt und haben auch schon zu ersten Erfolgen im Aufspüren von merkmalsassoziierten Genen geführt [z.B. 27]. Ein weiteres Zeichen für die Nützlichkeit der im Projekt entwickelten "Gene-ome" Resourcen ist die Tatsache, daß das die Gersten Genom Publikation [32] mittlerweile bereits 173 mal zitiert worden ist, dabei diente das kartierte Gersten Genkomplement häufig als Grundlage für weitergehende vergleichende Analysen.

Neben den beschriebenen Hauptergebnissen, wurden von unserer Gruppe im Rahmen von TRITEX eine Reihe von nützlichen, universell einsetzbaren bioinformatischen "Helfer"-Applikationen entwickelt bzw verbessert:

- "Sequence Content Checker" zum Aufspüren von Kontaminationen (z.B. durch humane oder bakterielle DNA) und zum Überprüfen der Assemblierungsvollständigkeit (Abb 3)
- "CarmA" (chromosome arm assignment): nutzt chromosomal vorsortierte Sequenz Resourcen aus aus, um schnell und zuverlässig das Herkunfts Chromosom von Sequenzen unbekannter Zuordnung zu bestimmen, hat einen hohen Datendurchsatz und verringert Zuweisungs-Fehlerraten (Abb 4)
- "chromoWIZ" [4]: Visualisierung von Eigenschaften entlang der Chromosomen (z.B. Gendichte), Bestandteil des GenomZippers (Abb 9) zur Identifizierung syntenischer Bereiche
- Erweiterung der Transposon Datenbank "PGSB-REdat" [30] (pgsb.helmholtz-muenchen.de/plant/recat/RecatTree.jsp) mit de novo detektierten *Triticeae* LTR-Retrotransposons und Implementierung einer hochdurchsatzfähigen Repeat Maskierungs Routine zur Sequenzvorbehandlung. Eine effektive Elimination der hoch-repetitiven Bereiche reduziert Rechenzeiten, verringert Mapping Ambiguitäten und die Vorhersage von 'falschen', Transposon Genmodellen.

Aufgrund der großen Menge an hoch-repetitiven Transposon Sequenzen (Tab 1) bedarf die lückenlose Anordung sämtlicher Gersten Sequenzbereiche noch weiterer Anstrengungen. Die dazu

nötigen vom IBSC koordinierten BAC bei BAC Sequenzierungen aller 7 Gersten Chromosomen sind mittlerweile abgeschlossen. In einem nächsten Schritt müssen die gewaltigen Datenmengen jetzt im Rahmen von anderen Projekten in ihrer Gesamtheit eingehend annotiert und vergleichend analysiert werden. Allerdings konnten erste explorative Vorstudien an den zu diesem Zeitpunkt bereits verfügbaren BAC Assemblies des Gersten Chromosoms 3 (3H) noch in der kostenneutralen Verlängerunsphase durchgeführt werden. Dabei lag das Hauptaugenmerk auf der Fragestellung welche konkreten Vorteile die vollständigere Sequenzabdeckung und bessere Assembly Qualität mit sich bringt.

Für die 3H Gen Modell Vorhersagen mußte die Annotations Pipeline über die Integration von nicht-3H Hintergrundsequenzen speziell angepaßt werden, um eine Übervorhersage durch unspezifische Mappings zu vermindern (Abb 5). Bei den ~12.000 annotierten Gen Loci zeichnet sich schon jetzt ein deutlicher Vorteil der BAC Sequenzen gegenüber den WGS Contigs ab: die Gene werden um bis zu 20 % vollständiger (Abb 6) erfaßt, es gibt 10% weniger mono-Exon Gene, die durchschnittliche Transkript Länge erhöht sich um 14% an, die durchschnittliche Locus Länge sogar um 40%, was auf einer verbesserten Erfassung repeat-haltiger Introns beruht (Tab 2).

Noch deutlicher tritt der Vorteil der BAC Sequenzen bei der *de novo* Detektion von Voll-Längen LTR-Retrotransposons auf, deren identische solo-LTR Komponenten im WGS Assembly vollständig kollabieren. Während im WGS Assembly nur 33 Voll-Längen LTR-Retrotransposons zu finden waren, konnten in den 3H BAC Sequenzen 9.463 strukturell vollständige LTR-Retrotransposons identifiziert und in ihrer Untergruppen- und Alterszusammensetzung charakterisiert werden (Abb 7). Die BAC Sequenzen bilden damit eine gute Quelle, um die Transposon Datenbank mit neuen Elementen aufzufüllen. Außerdem ist es mit dieser wesentlich verbesserten Sequenz Rekonstruktion zukünftig möglich den Einfluß unterschiedlicher Transposon Nachbarschaften auf Gene zu untersuchen.

Da sich erst nachträglich herausgestellt hatte, daß die 3H BAC Sequenzen aufgrund der BAC Überlappungen noch Redundanzen enthalten kann man davon ausgehen, daß die tatsächlichen Elementzahlen um etwa 2/3 niedriger als die genannten Werte ausfallen und somit im Bereich von 8.000 für die Gene und 6.300 für die Voll-Längen LTR-Retrotransposons liegen.

Ein weiteres in der Verlängerungsphase bearbeitetes Thema betraf die Pseudogen Detektion in Gerste. Pseudogene sind defekte Kopien von Genen, die ihre ursprüngliche Funktion nicht mehr erfüllen können, aber teilweise noch transkribiert werden. Es ist noch unbekannt, ob sie bei Pflanzen überwiegend 'nur' als Nebenprodukt der Evolution anzusehen sind oder in welchem Ausmaß sie auch neue Aufgaben z.B. als regulatorische RNA übernehmen, wie eine steigende Zahl von Beispielen aus der Tier-Genomik nahelegt. Generell werden Pseudogene in der Annotation von Pflanzen Genomen bisher nicht routinemäßig erfaßt und sind dementsprechend noch wenig untersucht. Tabelle 3 faßt die vorläufigen Eckdaten für Gerste aus der von uns entwickelten Pseudogene Detektions- und Charakterisierungs-Pipeline zusammen. Auf dem Gersten WGS Contig Assembly wurden ~45.000 Pseudogene über eine Homologie Suche zu dem publizierten Set von ~23.000 Gersten Genen [32] gefunden. Mit einer durchschnittlichen und medianen Länge von nur ~220 bp bzw ~120 bp bestehen die meisten Pseudogene nur aus einem Fragment ihres Eltern-Gens (¢ 1.108 bp). Von etwa der Hälfte der 23.000 Gene finden sich ein oder mehrere Pseudogene Gegenstücke im Genom. Die Einteilung in die beiden herkunftsbestimmenden Pseudogen

Untergruppen 'duplicated' und 'retroposed' zeigt, daß trotz des hohen Retrotransposon Gehaltes in Pflanzen nur ein sehr kleiner Anteil der Pseudogene von weniger als 2% eindeutig auf die reverse Transkribierung von mRNA zurückgeführt werden kann. Dieser erste Einblick in das Pseudogen Komplement einer Triticeeae Spezies macht deutlich, daß die Anzahl der Pseudogene die der Gene bei weitem übertrifft und eröffnet gleichzeitig spannende Forschungsansätze zur bisher vernachläßigten Rolle von Pseudogenen in der Evolution und Merkmalsausprägung von Pflanzen.

Wie das in Abb 8 dargestellte Schema beispielhaft skizziert, läßt sich zusammenfassend feststellen, daß die mit TRITEX aufgebauten Datenresourcen des Gersten "Gene-omes" einen essentiellen und vielfach genutzten Grundstock für Folgearbeiten darstellen.

2.2. Virtuelle Genkarten für alle 21 Weizenchromosomen (WP 4, task 3)

Der bereits im Vorläuferprojekt BARLEX konzipierte bioinformatische GenomZipper Ansatz macht sich die zwischen Gräsern stark konservierten Gen Reihenfolgen (Syntenie) zu Nutze, um schon mit relativ geringen Sequenzmengen und nur einer genetischen Karte als Grundgerüst virtuelle Gen-Karten zu erstellen. Diese chromosomalen Genabfolgen sind ein kostengünstiger und vorläufiger, aber trotzdem vielseitig verwendbarer und mittlerweile etablierter Ersatz für vollständig assemblierte Genome. Sie erlauben einen ersten detaillierten Einblick in die Struktur großer Pflanzengenome (>5 Gb), von denen bisher trotz der rasanten technologischen Fortschritte im Bereich der Genomsequenzierung noch keines in seiner Gesamtsequenz-Abfolge vollständig vorliegt.

Zu Beginn des Projektes wurde die GenomZipper Implementierung verbessert, stärker automatisiert und an unterschiedliche Arten von Eingabe Sequenzen (kurze reads oder längere Contigs) angepaßt. Eine von der IPK Gruppe (P2) durchgeführte experimentelle Validierung der Gersten Zipper Daten ergab eine sehr gute Spezifität von 95% [21]. Der GenomZipper Workflow [22] besteht aus drei Hauptschritten (Abb 9): (1) Maskierung von repetitiven Transposon Bereichen, um mehrdeutige Zuordnungen zu vermeiden und die Rechenzeit zu verkürzen, (2) Identifizierung von syntenisch konservierten Regionen zwischen Ziel- und Referenz-Genom und (3) Aufbau der virtuellen Genkarte entlang der genetischen Marker Karte in Form eines mehrschichtigen Datengerüsts.

Bei dem endgültigen GenomeZipper für alle 21 Weizen Chromosome handelte es sich bereits um die 5. Version (Tab 4). Die Updates waren jeweils nötig, um bereinigte Fehler (z.B. Sequenz Verunreinigungen und Assemblierungsprobleme) zu berücksichtigen, die meistens erst im Rahmen der bioinformatischen Analysen aufgedeckt worden sind. Ein Vergleich der aus der physikalischen Karte von 6A gewonnenen Genreihenfolge (WP 3.1) mit der entsprechenden virtuellen Anordnung des GenomeZippers zeigt eine insgesamt gute Übereinstimmung (Abb 10) und beweist, dass die mit dem GenomeZipper generierten Daten trotz des indirekten Syntenie basierten Ansatzes realitätsnahe Ergebnisse liefern.

Der in Science publizierte Weizengenom Artikel des IWGSC (International Wheat Genome Sequencing Consortium) „A chromsome-based draft sequence oft the hexaploid wheat genome“ [6] enthält als einen wichtigen Eckpunkt den vollständigen GenomeZipper (Tab 4) mit 67,351 verankerten Gen Loci (das entspricht ~ 60% aller ‚high confidence‘ HC-Gene) und eine Reihe davon abgeleiteter Analysen, die die Genreihenfolge benötigen um bestimmte Gen Attribute entlang

der Chromosomen zu beschreiben und vergleichend darzustellen (z.B. Abb 11, 12, 13, 14).

Zusammenfassend handelt es sich bei dem GenomZipper Ansatz um einen Syntenie basierten und damit indirekten Entwurf einer Genreihenfolge, der noch mit den entsprechenden Unsicherheiten behaftet ist. Sein großer Erfolg bei Triticeae Genomen mit zehn in der TRITEX Förderperiode entstandenen Anwendungs Publikationen [1, 6, 15, 19, 23, 26, 27, 29, 32, 33] ist neben seiner experimentell validierten Genauigkeit [21] auch darin begründet, daß man mit einer vergleichsweise geringen Menge an Daten - nur genetische Karte plus niedrig abgedeckte NGS-Sequenzen - und damit kostengüstig schon eine sehr nützliche Genanordnug seines Zielorganismus bekommt. So sind z.B. großräumige chromosomale Translokationen [1, 20] gut erkennbar. Darüberhinaus ermöglicht die Verknüpfung der herkömmlichen genetischen Centimorgan Distanzen von QTLs (Region eines quantitativen Merkmals) mit funktionell charakterisierten Genen eine direkte Identifizierung von merkmalstragenden Genkandidaten. Dies eröffnet auch für noch unvöllständig sequenzierte Getreidearten die Möglichkeiten einer gezielt merkmalsgeleiteten und damit wesentlich effizienteren Pflanzenzüchtung.

2.3. Physikalische Karte des Weizen Chromosoms 6A (WP 3.1, task 3)

Als deutscher Beitrag zu dem langfristigen Ziel des IWGSC (International Wheat Genome Sequencing Consortium) sämtliche 21 Chromosomen des hexaploiden 17 Gb Brotweizens vollständig BAC bei BAC zu sequenzieren, wurde im TRITEX Projekt für eines der Chromosomen (6A) die dazu benötigte BAC Reihenfolge entlang des Chromosoms ausgearbeitet und bereits publiziert [11].

Abweichend vom ursprünglichen Antrag wurde die physikalische Karte von 6A (Gesamtlänge 705 Mb) nicht über eine hochauflösende Fingerprint Karte (HICF), sondern mit der neueren Technologie des WGP (whole genome profiling) angefertigt. WGP hat den entscheidenden Vorteil, dass es im Gegensatz zu reinen Restriktions-fragment Fingerprints auch kurze Sequenztags (über Illumina NGS) liefert, die wiederum zur Verankerung von weiteren Sequenz Ressourcen und damit letztendlich auch Genen genutzt werden kann. Somit konnte durch das Umschwenken auf WGP zusätzlich zum ursprünglichen Ziel einer BAC Anordnung auch noch eine physikalische Anordung der Gene entlang des Weizenchromosom 6A erreicht werden.

Das Weizen 6A Teilprojekt war anfangs verzögert worden, da bei der Erstellung des ersten Datensets das Ausgangsmaterial im Labor verwechselt worden war und statt des kurzen Arms von Subgenom A (6AS) der homologe kurze Arm von Subgenom D (6DS) verwendet worden war. Der Fehler war erst am Ende des gesamten Prozesses bei unserer gründlichen bioinformatischen Analyse entdeckt worden. Der beschriebene und auch in anderen Projekten vorgekommene Verwechslungs- und Kontaminationsvorfälle unterstreichen die Bedeutung einer sorgfältigen und kritischen bioinformatischen Analytik.

Die WGP Daten für den kurzen (6AS) und langen Arm (6AL) des Weizen Chromosoms 6A wurden von Partner P3 (T. Schnurbusch) erstellt und mit zwei unterschiedlichen Methoden (FPC und LTC) assembliert. Nach eingehenden Evaluierungen hat sich die LTC Assemblierung mit 2.330 LT-contigs und einer kumulativen Gesamtlänge von 1064 Mb als höherwertig (im Sinne von weniger chimärischen Verbindungen) erwiesen und wurde als Grundlage für die weiteren bioinformatischen

Verankerungsschritte verwendet.

Da die WGP Tags mit 50-100 bp zu kurz für eine eindeutige und möglichst vollständige Verankerung an den genetischen Markern sind wurde ein Verfahren entwickelt sie mit aus anderen Projekten verfügbaren Weizen Sequenzen zu verlängern. Dazu wurden in einem ersten Schritt die 6AS und 6AL spezifischen Sequenz Contigs des IWGSC (International Wheat Genome Sequencing Consortium) stringent auf die WGP Tags der 6AS und 6AL BACs kartiert, in einer 2. Runde mit genomischen Sequenzdaten von Triticum urartu und Aegilops tauschii (zwei öffentlich zugängliche Weizen Vorgänger Ressourcen) und in einer letzten Runde nochmals mit den IWGSC Contigs verbunden (Abb 15). Die auf diese Weise stark mit Sequenzdaten angereicherte physikalische Karte wurde anschließend an eine kürzlich publizierte genetische Karte vom neuartigen GBS (genotyping by sequencing) Typ durch Sequenzhomologie mit einer Gesamtlänge von 661 Mb (682 LT-contigs) verankert. Diese jetzt chromosomal angeordneten physikalischen LT-contigs enthalten 132 Mb T.aestivum, 303 Mb T. urartu und 129 Mb Ae.tauschii Sequenzdaten (Abb 16). Mit einem nachfolgenden syntenischen Stratifizierungsschritt gegen das vorläufige Gerstengenom gelang es weitere 296 LT-contigs (170 Mb) in den chromosomalen Kontext einzubinden. Die Zusammenführung der IWGSC Weizen Gen Annotation, die in unserer Gruppe (P5) erstellt und eingehend analysiert worden ist, mit der jetzt genetisch verankerten physikalischen Karte von 6A ermöglichte es schließlich sogar 3.355 Gene (das sind 67% aller 6A Gene) entlang des Weizen Chromosoms 6A zu positionieren, davon liegen 1.673 auf 6AS und 1.682 auf 6AL. Wie von anderen großen Genomen bekannt sind auch hier die Gendichten an den Enden der Chromosomen am höchsten (Abb 17).

Letztendlich ließ sich mit der WGP Technologie aus Chromosomen Arm sortierten BAC Bibliotheken und dem für komplexe Genome besser geeigneten LTC (lineare Topologie) Assembly unter geschickter Ausnutzung bereits anderweitig vorhandener *Triticeae* Sequenz und Daten Ressourcen, sowohl eine genetisch verankerte physikalische Karte, als auch eine Anordnung von ca. 2/3 aller 6A Gene erstellen. Ein solches Ergebnis wäre bis vor kurzem nur durch eine wesentlich aufwändigere BAC End Sequenzierung möglich gewesen.

2.4. Webinterface zur Daten Darstellung und Weitergabe (WP 4, task 5)

Die im TRITEX Projekt erarbeiteten Gerste und Weizen Daten von sind unter der speziell hervorgehobenen Gruppe der Triticeae Genome in das allgemeine Framework unserer PGSB-PlantsDB Webseiten eingebunden (Abb 18) und werden auch weiterhin verfügbar sein (pgsb.helmholtz-muenchen.de/plant/triticeae/genomes/index.jsp). Die entsprechenden Darstellungen der Sequenz, Genannotations- und Zipper-Daten von Gerste und Weizen wurden projektbegleitend fortlaufend aktualisiert und mit zusätzlichen Funktionalitäten ausgestattet.. Eine gesonderte Datenbank Veröffentlichung [30] beschreibt u.a. auch die speziell für die TRITEX Daten entwickelten Webinterfaces, Daten Resourcen und Zugriffsmöglichkeiten. Zwei weitere im Projektzusammenhang entwickelte Web Applikationen sind ebenfalls publiziert worden. Dabei handelt es sich zum einen um chromoWIZ (pgsb.helmholtz-muenchen.de/cgi-bin/db2/chromowiz/index.cgi), ein Tool, welches mit Hilfe von Heatmaps die chromosomale Verteilung von Benutzer definierbaren Genen oder Sequenzen anzeigt [4]. Das zweite Tool, der generische RNASeqExpressionBrowser [5] ist als Gersten Instanz dargestellt (pgsb.helmholtz-muenchen.de/

plant/RNASeqExpressionBrowser/projects.jsp). Er dient zur übersichtlichen Visualisierung von Expressionsdaten in Verbindung mit funktionellen Genannotationen und erleichtert ihre Auswertung und Interpretation.

Für die Triticeae Genome sind zusätzlich noch folgende Komponenten unserer PGSB Web Präsenz als besonders nützlich und entsprechend nachgefragt hervorzuheben:

- GenomeZipper Daten Browser mit Suchfunktion und FTP download für die 7 Gersten und die 21 Weizenchromosomen (Abb 19)
- CrowsNest, ein Synteny Viewer mit verschiedenen Makro- und Mikro-Ebenen (Abb 21), inklusive der physikalischen Karte von Gerste (Abb 20)
- Repeat Sequenz Datenbank mit ~ 60,000 Einträgen und Repeat/Transposon Klassifizierung (pgsb.helmholtz-muenchen.de/plant/recat/RecatTree.jsp, Abb 22)

2.5 Im Projektzusammenhang entstandene Publikationen und Abschlußarbeiten

Insgesamt sind im Projektzusammenhang bisher 33 Publikationen unter der Beteiligung der HMGU Gruppe (Klaus Mayer, P5) veröffentlicht worden. Neben den beiden Hauptpublikationen zum Gersten [32] und Weizen Genom [6] gibt es 13 weitere direkt aus dem TRITEX Projekt heraus entstandene Veröffentlichungen, aufgeteilt in vier Ergebnis beschreibende [11, 14, 21, 24], sechs methodische [2, 4, 5, 16, 17, 30] und drei strategische [13, 22, 31] Manuskripte. Die übrigen 18 Publikationen sind Folgearbeiten, die TRITEX Daten Resourcen für weitergehende Analysen verwenden. Darunter befinden sich acht Arbeiten mit GenomZipper Daten [1, 15, 19, 23, 26, 27, 29, 33] und 10 weitere [3, 7, 8, 9, 10, 12, 18, 20, 25, 28], häufig mit Expressionsanalysen auf TRITEX generierten Gendaten.

Zusätzlich wurden in unserer Gruppe vier Dissertationen, eine Master und eine Bachelor Arbeit erfolgreich zum Abschluß gebracht, die sich in ihren Teilbereichen entweder mit der Integration der in TRITEX erzeugten Rohdaten [35, 36, 39] oder mit Folgeanalysen [34, 37, 38] auf den aufgebauten Datenresourcen befassten.

2.5.1 Publikationsliste

Die Mitarbeiter der vorliegenden Berichtsgruppe (Klaus Mayer, PGSB, HMGU) sind in der Authorenauflistung zur besseren Übersicht hervorgehoben. Die Publikationsliste ist absteigend nach Datum sortiert.

[1] Helguera M, Rivarola M, Clavijo B, **Martis MM**, Vanzetti LS, González S, Garbus I, Leroy P, Šimková H, Valárik M, Caccamo M, Doležel J, **Mayer KF**, Feuillet C, Tranquilli G, Paniego N, Echenique V **New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing.** *Plant Sci*, 2015 Apr, 233:200-12

[2] Cviková K, Cattonaro F, Alaux M, Stein N, **Mayer KF**, Doležel J, Bartoš J **High-throughput physical map anchoring via BAC-pool sequencing.** *BMC Plant Biol,* 2015 Jan, 15(1):99

[3] Dey S, Wenig M, Langen G, **Sharma S, Kugler KG**, Knappe C, Hause B, Bichlmeier M, Babaeizad V, Imani J, Janzik I, Stempfl T, Hückelhoven R, Kogel KH, **Mayer KF**, Vlot AC **Bacteria-triggered systemic immunity in barley is associated with WRKY and ETHYLENE RESPONSIVE FACTORs but not with salicylic acid.** *Plant Physiol,* 2014 Dec, 166(4):2133-51

[4] **Nussbaumer T, Kugler KG**, Schweiger W, **Bader KC**, **Gundlach H, Spannagl M,** Poursarebani N, **Pfeifer M, Mayer KF chromoWIZ: a web tool to query and visualize chromosome-anchored genes from cereal and model genomes.** *BMC Plant Biol*, 2014 Nov, 14:348

[5] **Nussbaumer T, Kugler KG, Bader KC, Sharma S, Seidel M, Mayer KF RNASeqExpressionBrowser-a web interface to browse and visualize high-throughput expression data.** *Bioinformatics*, 2014 Sep, 30(17):2519-20

[6] International Wheat Genome Sequencing Consortium (IWGSC) **A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome.** *Science*, 2014 Jul, 345(6194):1251788

[7] Marcussen T, Sandve SR, Heier L, **Spannagl M, Pfeifer M,** International Wheat Genome Sequencing Consortium,, Jakobsen KS, Wulff BB, Steuernagel B, **Mayer KF,** Olsen OA **Ancient hybridizations among the ancestral genomes of bread wheat.** *Science*, 2014 Jul, 345(6194):1250092

[8] **Pfeifer M, Kugler KG**, Sandve SR, Zhan B, Rudi H, Hvidsten TR, International Wheat Genome Sequencing Consortium, **Mayer KF**, Olsen OA **Genome interplay in the grain transcriptome of hexaploid bread wheat.** *Science*, 2014 Jul, 345(6194):1250091

[9] Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury JM, **Mayer K,** Berges H, Quesneville H, Wincker P, Feuillet C **Structural and functional partitioning of bread wheat chromosome 3B.** *Science*, 2014 Jul, 345(6194):1249721

[10] Eversole K, Feuillet C, **Mayer KF**, Rogers J **Slicing the wheat genome. Introduction** *Science*, 2014 Jul, 345(6194):285-7

[11] Poursarebani N, **Nussbaumer T,** Simková H, Safář J, Witsenboer H, van Oeveren J, Doležel J, **Mayer KF**, Stein N, Schnurbusch T **Whole-genome profiling and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A.** *Plant J*, 2014 Jul, 79(2):334-47

[12] Keilwagen J, Kilian B, Özkan H, Babben S, Perovic D, **Mayer KF**, Walther A, Poskar CH, Ordon F, Eversole K, Börner A, Ganal M, Knüpffer H, Graner A, Friedel S **Separating the wheat from the chaff - a strategy to utilize plant genetic resources from ex situ genebanks.** *Sci Rep*, 2014 May, 4:5231

[13] Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, **Mayer KF** **Plant genome sequencing - applications for crop improvement.** *Curr Opin Biotechnol*, 2014 Apr, 26:31-7

[14] Ariyadasa R, Mascher M, **Nussbaumer** T, Schulte D, Frenkel Z, Poursarebani N, Zhou R, Steuernagel B, **Gundlach H**, Taudien S, Felder M, Platzer M, Himmelbach A, Schmutzer T, Hedley PE, Muehlbauer GJ, Scholz U, Korol A, **Mayer KF**, Waugh R, Langridge P, Graner A, Stein N **A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms.** *Plant Physiol*, 2014 Jan, 164(1):412-23

[15] Kopecký D, **Martis M**, Číhalíková J, Hřibová E, Vrána J, Bartoš J, Kopecká J, Cattonaro F, Stočes Š, Novák P, Neumann P, Macas J, Šimková H, Studer B, Asp T, Baird JH, Navrátil P, Karafiátová M, Kubaláková M, Šafář J, **Mayer K**, Doležel J **Flow sorting and sequencing meadow fescue chromosome 4F**. *Plant Physiol*, 2013 Nov, 163(3):1323-37

[16] Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, Muñoz-Amatriaín M, Close TJ, Wise RP, Schulman AH, Himmelbach A, **Mayer KF**, Scholz U, Poland JA, Stein N, Waugh R **Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ).** *Plant J,* 2013 Nov, 76(4):718-27

[17] Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, Ayling S, Steuernagel B, **Pfeifer M**, D'Ascenzo M, Akhunov ED, Hedley PE, Gonzales AM, Morrell PL, Kilian B, Blattner FR, Scholz U, **Mayer KF**, Flavell AJ, Muehlbauer GJ, Waugh R, Jeddeloh JA, Stein N **Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond.** *Plant J*, 2013 Nov, 76(3):494-505

[18] **Kugler KG**, Siegwart G, **Nussbaumer T**, Ametz C, **Spannagl M,** Steiner B, Lemmens M, **Mayer KF**, Buerstmayr H, Schweiger W **Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (Triticum aestivum L.).** *BMC Genomics*, 2013 Oct, 14:728

[19] Lüpken T, Stein N, Perovic D, Habekuß A, Serfling A, Krämer I, Hähnel U, Steuernagel B, Scholz U, Ariyadasa R, **Martis M, Mayer K**, Niks RE, Collins NC, Friedt W, Ordon F **High-resolution mapping of the barley Ryd3 locus controlling tolerance to BYDV.** *Molecular Breeding*, 2013 Oct, 1-12

[20] Ma J, Stiller J, Berkman PJ, Wei Y, Rogers J, Feuillet C, Dolezel J, **Mayer KF**, Eversole K, Zheng YL, Liu C **Sequence-based analysis of translocations and inversions in bread wheat (Triticum aestivum L.).** *PLoS One,* 2013 Sep, 8(11):e79329

[21] Poursarebani N, Ariyadasa R, Zhou R, Schulte D, Steuernagel B, **Martis MM**, Graner A, Schweizer P, Scholz U, **Mayer K**, Stein N **Conserved synteny-based anchoring of the barley genome physical map.** *Funct Integr Genomics*, 2013 Aug, 13(3):339-50

[22] **Spannagl M, Martis MM, Pfeifer M, Nussbaumer T, Mayer KF** **Analysing complex Triticeae genomes - concepts and strategies.** *Plant Methods*, 2013 Jul, 9(1):35

[23] Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier C, Simková H, Safář J, Bellec A, Vautrin S, Frenkel Z, Cattonaro F, Magni F, Scalabrin S, **Martis MM, Mayer KF**, Korol A, Bergès H, Doležel J, Feuillet C **A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat.** *Genome Biol*, 2013 Jun, 14(6):R64

[24] Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, **Spannagl M, Nussbaumer T, Mayer** KF, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N **Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome.** *Genome Biol,* 2013 Jun, 14(6):R58

[25] Silvar C, Perovic D, **Nussbaumer T, Spannagl M**, Usadel B, Casas A, Igartua E, Ordon F **Towards positional isolation of three quantitative trait loci conferring resistance to powdery mildew in two Spanish barley landraces.** *PLoS One*, 2013 May, 8(6):e67336

[26] Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen CM, Zhang Y, McGuire PE, Pasternak S, Stein JC, Ware D, Kramer M, McCombie WR, Kianian SF, **Martis MM, Mayer KF,** Sehgal SK, Li W, Gill BS, Bevan MW, Simková H, Dolezel J, Weining S, Lazo GR, Anderson OD, Dvorak J **A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor.** *Proc Natl Acad Sci U S A*, 2013 May, 110(19):7940-5

[27] Lüpken T, Stein N, Perovic D, Habekuss A, Krämer I, Hähnel U, Steuernagel B, Scholz U, Zhou R, Ariyadasa R, Taudien S, Platzer M, **Martis M, Mayer K**, Friedt W, Ordon F **Genomics-based high-resolution mapping of the BaMMV/BaYMV resistance gene rym11 in barley (Hordeum vulgare L.).** *Theor Appl Genet*, 2013 May, 126(5):1201-12

[28] Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, **Pfeifer M**, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, **Spannagl M, Mayer KF**, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Xu J, Gao J, Middleton C, Quan Z, Liu G, Wang J, International Wheat Genome Sequencing Consortium, Yang H, Liu X, He Z, Mao L, Wang J **Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation.** *Nature,* 2013 Apr, 496(7443):91-5

[29] **Pfeifer M, Martis M, Asp T, Mayer KF**, Lübberstedt T, Byrne S, Frei U, Studer B **The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics.** *Plant Physiol,* 2013 Feb, 161(2):571-82

[30] **Nussbaumer T, Martis MM**, Roessner SK, **Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M** **MIPS PlantsDB: a database framework for comparative plant genome research** *Nucleic Acids Res*, 2013 Jan, 41(Database issue):D1144-51

[31] Feuillet C, Stein N, Rossini L, Praud S, **Mayer K**, Schulman A, Eversole K, Appels R **Integrating cereal genomics to support innovation in the Triticeae.** *Funct Integr Genomics*, 2012 Nov, 12(4):573-83

[32] International Barley Genome Sequencing Consortium, **Mayer KF**, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N **A physical, genetic and functional sequence assembly of the barley genome.** *Nature,* 2012 Nov, 491(7426):711-6

[33] Hernandez P, **Martis M**, Dorado G, **Pfeifer M**, Gálvez S, Schaaf S, Jouve N, Šimková H, Valárik M, Doležel J, **Mayer KF** **Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content.** *Plant J*, 2012 Feb, 69(3):377-86

2.5.2 Dissertationen und Abschlußarbeiten

[34] **Manuel Spannagl**, **The genomic repertoire of complex and polyploid cereal genomes.** *Dissertation*, TU München (2015)

[35] **Mihaela-Maria Martis**, **GenomeZipper - a bioinforamtic approach to unravel highly complex cereal genomes.** *Dissertation*, TU München (2014)

[36] Thomas Nussbaumer, **Genomassemblierung komplexer Gräsergenome aus hoch-heterogenen Datensets**, *Dissertation*, TU München (2014)

[37] **Matthias Pfeifer**, **The genome and transcriptome of Triticeae genomes and the impact of polyploidization**. *Dissertation*, TU München (2014)

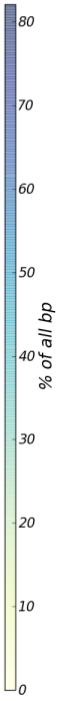
[38] **Verena Prade**, **Closing the Gap: Assessment of Pseudogenes and Successive Gene Decay in Plants.** *Masterarbeit*, TU München (2014)

[39] **Verena Prade**, **Eine Validierung von CarmA: "Read based Chromosome arm Assignment",** *Bachelorarbeit,* TU München (2011)

2.6 Abbildungen und Tabellen

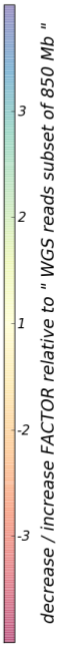
A)

% OF ALL BP	WGS reads subset of 850 Mb	BES 374 Mb	random BACs 63Mb	gene bearing BACs 379 Mb	Bowman WGS assembly 1.8 Gb	Morex WGS assembly 1.9Gb	Barke WGS assembly 2.0 Gb
MDR-l1: unique/low copy (20mer <10x)	30.6	29.3	30.2	34.0	57.8	56.9	50.2
MDR-l2: medium/high copy (>=10 to <1000x)	33.9	37.1	36.1	33.6	33.3	31.4	35.4
MDR-l3: very high copy (20mer >= 1000x)	35.5	33.5	33.7	32.4	8.9	11.7	14.4
Mobile Element	81.6	81.8	82.1	74.1	61.0	58.9	58.5
Class I: Retroelement (RXX)	75.3	76.9	75.5	67.1	55.0	52.8	52.7
LTR Retrotransposon (RLX)	75.1	76.6	75.2	66.4	54.4	52.3	52.2
Copia (RLC)	15.3	12.5	13.7	14.5	8.7	8.5	8.5
Gypsy (RLG)	22.3	21.8	20.8	17.0	19.1	18.2	18.0
Gypsy/Copia ratio	1.45	1.74	1.53	1.17	2.2	2.1	2.1
unclassified LTR	37.5	42.3	40.7	35.0	26.6	25.6	25.8
non-LTR Retrotransposon	0.22	0.29	0.31	0.63	0.58	0.58	0.53
LINE (RIX)	0.22	0.28	0.29	0.59	0.55	0.55	0.5
SINE (RSX)	0.005	0.01	0.027	0.041	0.033	0.029	0.028
Class II: DNA Transposon (DXX)	5.6	4.6	6.2	6.4	5.2	5.2	5.0
Retro-TE/DNA-TE ratio	13.4	16.8	12.2	10.5	10.5	10.1	10.6
DNA Transposon Superfamily (DTX)	5.4	4.3	5.9	5.8	4.6	4.6	4.4
CACTA superfamily (DTC)	5.2	4.1	5.5	5.2	4.0	4.0	3.9
hAT superfamily (DTA)	0.019	0.014	0.029	0.033	0.023	0.024	0.023
Mutator superfamily (DTM)	0.12	0.13	0.21	0.29	0.27	0.28	0.26
Tc1/Mariner superfamily (DTT)	0.021	0.025	0.029	0.055	0.061	0.061	0.056
PIF/Harbinger (DTH)	0.058	0.091	0.089	0.18	0.21	0.2	0.17
unclassified	0.016	0.015	0.024	0.049	0.052	0.051	0.045
MITE (DXX)	0.18	0.2	0.23	0.47	0.53	0.52	0.48
Helitron (DHH)	0.012	0.026	0.026	0.062	0.034	0.04	0.036
unclassified DNA transposon	0.021	0.025	0.041	0.044	0.037	0.066	0.036
Unclassified Element (XXX)	0.65	0.3	0.43	0.63	0.83	0.81	0.74
Simple Sequence Repeat	0.32	0.21	0.35	0.096	0.22	0.21	0.22
rRNA gene	0.29	0.53	0.11	0.12	0.014	0.059	0.12



B)

increase/decrease FACTOR relative to " WGS reads subset of 850 Mb " % OF ALL BP	BES 374 Mb	random BACs 63Mb	gene bearing BACs 379 Mb	Bowman WGS assembly 1.8 Gb	Morex WGS assembly 1.9Gb	Barke WGS assembly 2.0 Gb
MDR-l1: unique/low copy (20mer <10x)	-1.04	-1.02	1.11	1.89	1.86	1.64
MDR-l2: medium/high copy (>=10 to <1000x)	1.1	1.07	-1.01	-1.02	-1.08	1.04
MDR-l3: very high copy (20mer >= 1000x)	-1.06	-1.05	-1.1	-4.0	-3.0	-2.5
Mobile Element	1.0	1.01	-1.1	-1.34	-1.39	-1.4
Class I: Retroelement (RXX)	1.02	1.0	-1.12	-1.37	-1.43	-1.43
LTR Retrotransposon (RLX)	1.02	1.0	-1.13	-1.38	-1.44	-1.44
Copia (RLC)	-1.23	-1.12	-1.06	-1.75	-1.81	-1.81
Gypsy (RLG)	-1.02	-1.07	-1.31	-1.17	-1.23	-1.24
unclassified LTR	1.13	1.08	-1.07	-1.41	-1.46	-1.45
non-LTR Retrotransposon	1.28	1.4	2.8	2.6	2.6	2.4
LINE (RIX)	1.26	1.31	2.7	2.5	2.5	2.3
SINE (RSX)	1.96	5.1	7.7	6.3	5.5	5.4
Class II: DNA Transposon (DXX)	-1.22	1.11	1.14	-1.07	-1.07	-1.12
DNA Transposon Superfamily (DTX)	-1.24	1.1	1.07	-1.17	-1.16	-1.21
CACTA superfamily (DTC)	-1.27	1.07	1.0	-1.29	-1.29	-1.33
hAT superfamily (DTA)	-1.32	1.56	1.75	1.24	1.31	1.25
Mutator superfamily (DTM)	1.11	1.8	2.5	2.3	2.3	2.2
Tc1/Mariner superfamily (DTT)	1.18	1.37	2.6	2.9	2.8	2.6
PIF/Harbinger (DTH)	1.58	1.54	3.2	3.6	3.5	3.0
unclassified	-1.08	1.51	3.0	3.2	3.2	2.8
MITE (DXX)	1.08	1.26	2.6	2.9	2.8	2.6
Helitron (DHH)	2.2	2.2	5.3	3.0	3.4	3.1
unclassified DNA transposon	1.23	1.98	2.1	1.78	3.2	1.73
Unclassified Element (XXX)	-2.1	-1.51	-1.03	1.29	1.26	1.14
Simple Sequence Repeat	-1.5	1.12	-3.3	-1.43	-1.54	-1.43
rRNA gene	1.79	-2.7	-2.4	-20.9	-5.0	-2.4



Tab 1: Charakterisierung der repetitiven Bereiche des Gersten Genoms

A) Prozentuale Wiederfindungsrate in unterschiedlichen Sequenzarten

B) An- und Abreicherungsraten im Vergleich zu den realen Verhältnissen

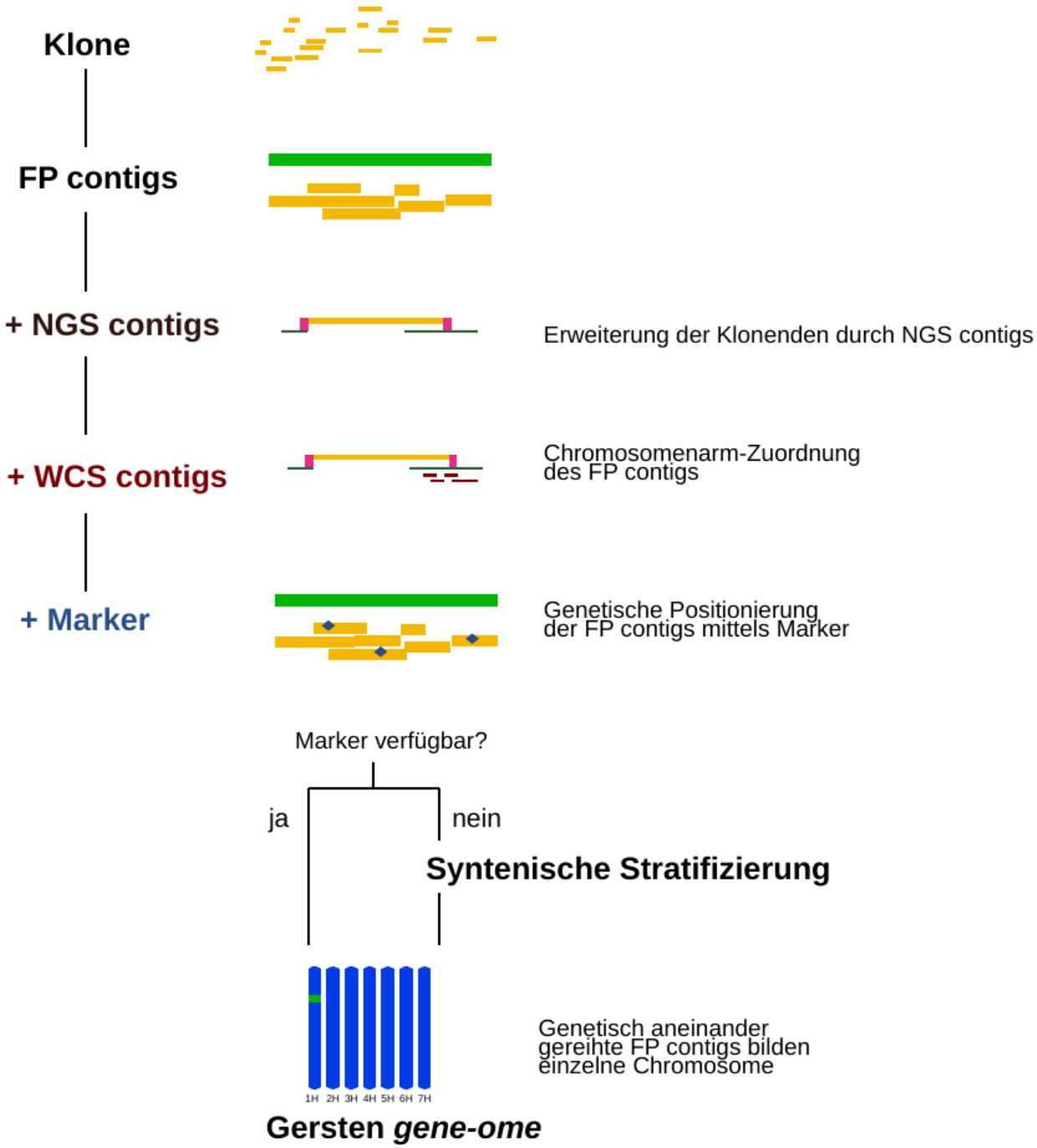
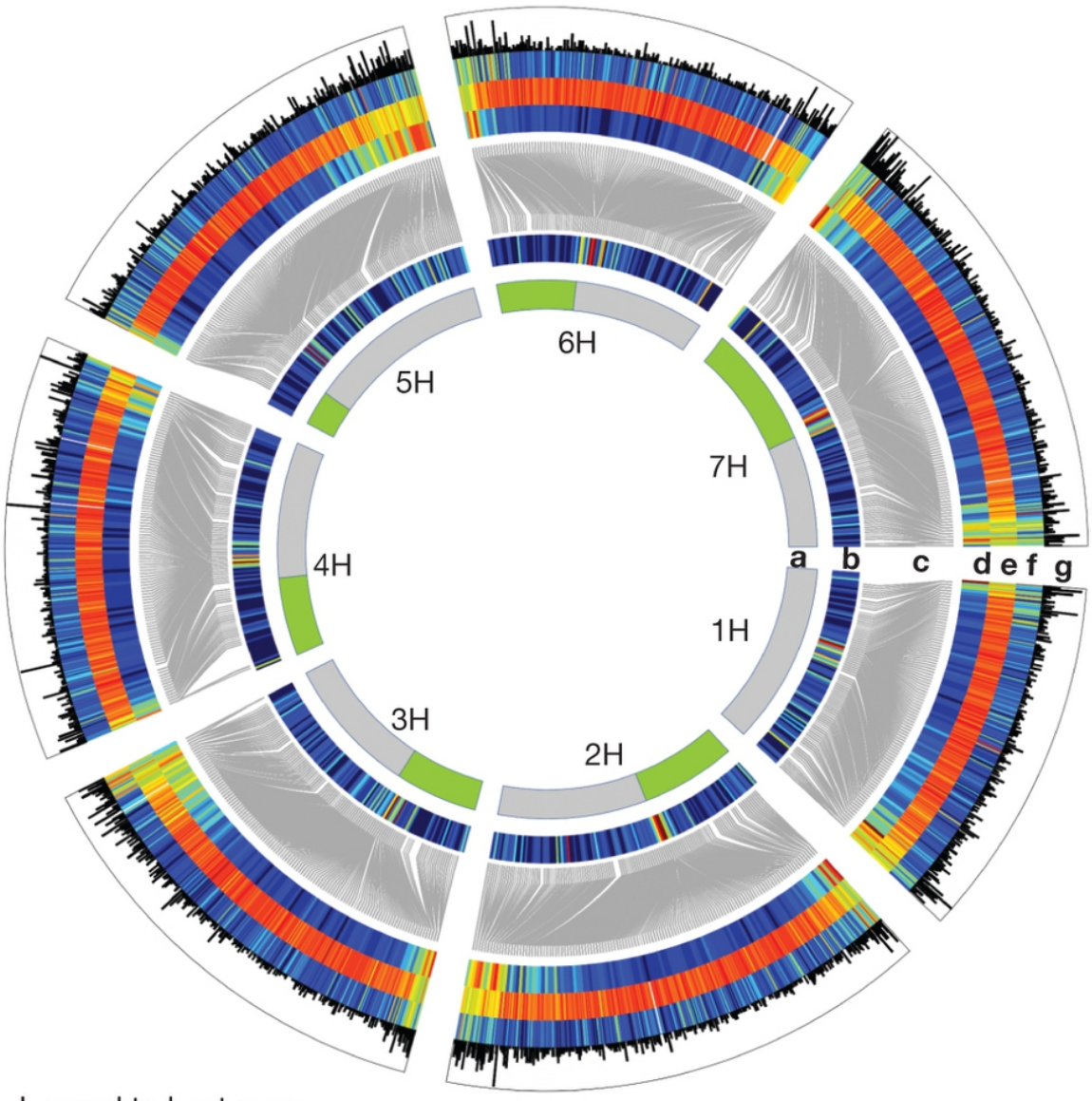


Abb 1: Datenintegrations Schema zur genetisch verankerten physikalischen karte von Gerste



Legend to heat map:

Track **b** genes per cM	min: 0, max: 80
Track **d** genes per Mb	min: 0, max: 20
Track **e** LTR retroelements (%)	min: 0, max: 100
Track **f** DNA transposons (%)	min: 0, max: 20
Track **g** sequenced BAC clones per Mb	0 ◺ 20

max

min

Abb 2: Architektur des Gerstengenoms

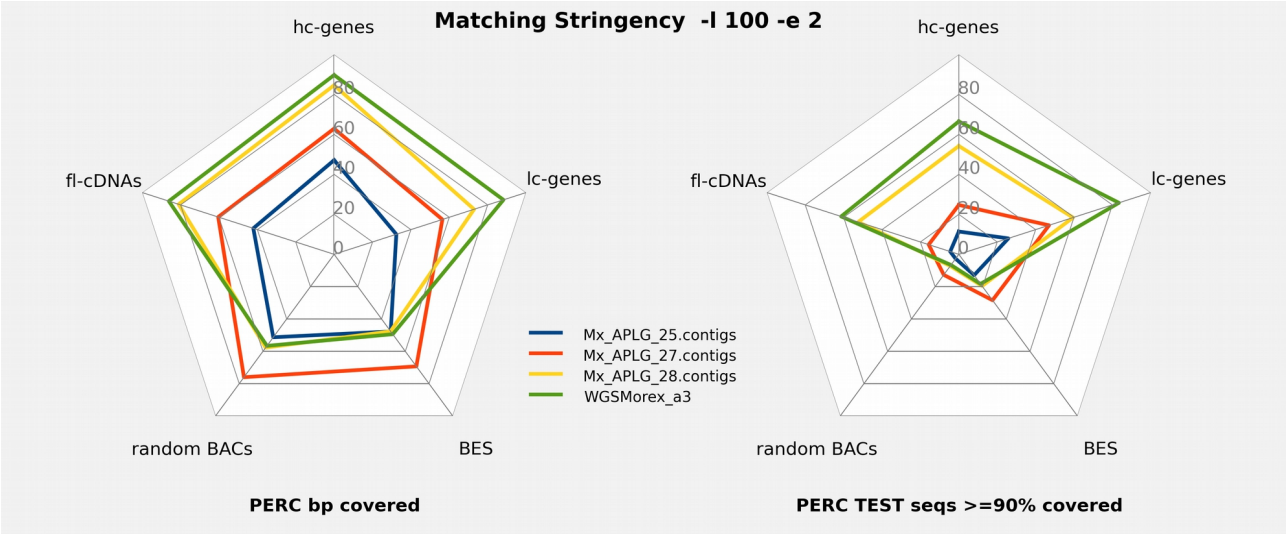


Abb 3: Multifaktorielle Vollständigkeits Analysen für verschiedene Sequenz Assembly Versionen

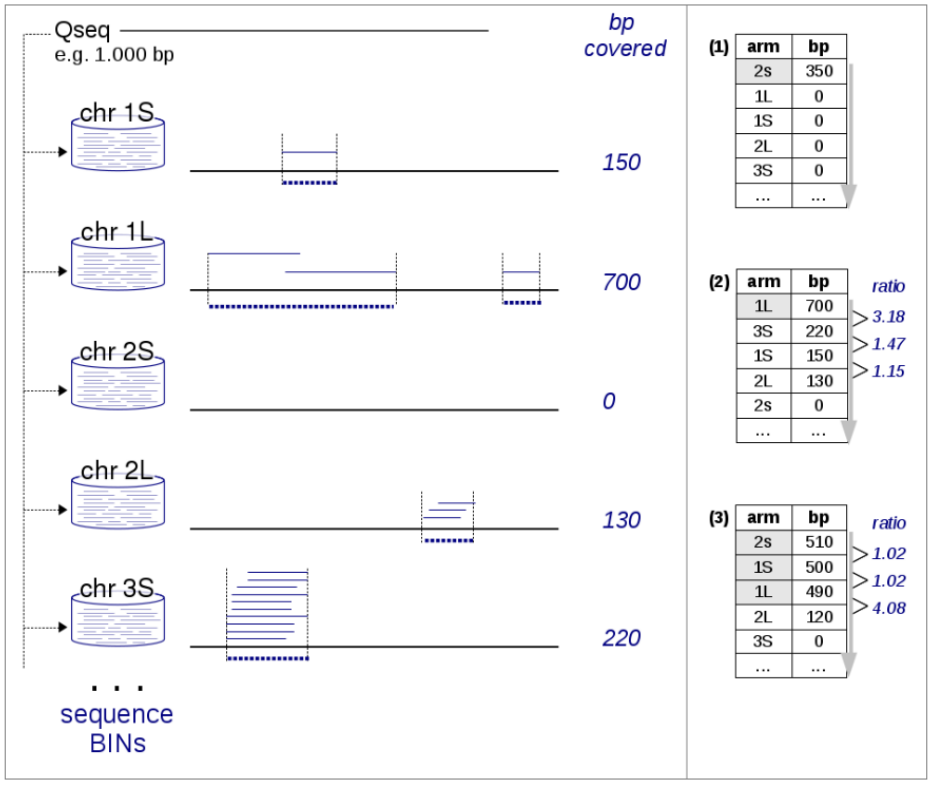


Abb 4: Prinzip der Chromosomen Arm Zuordnung (CarmA) durch Maskierung mit armspezifischen Sequenz Resourcen.

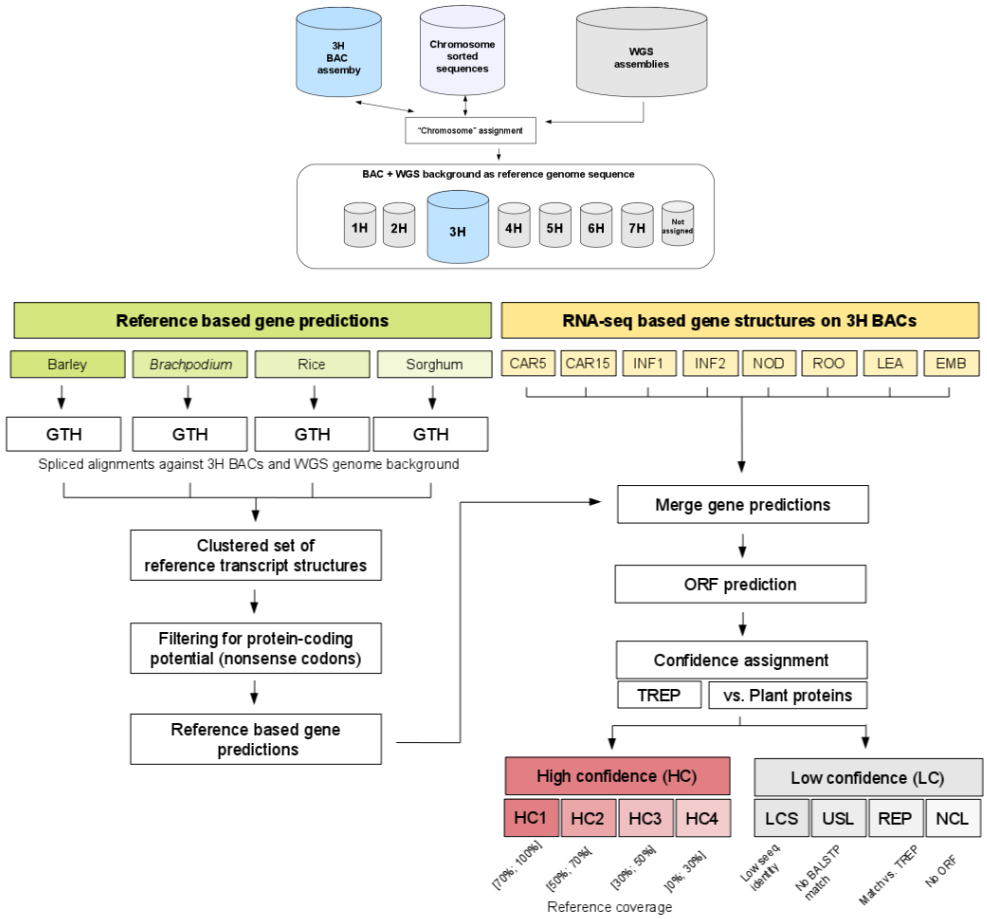


Abb 5: Anpassung der Genvorhersage Pipeline für die 3H BAC Sequenzen von Gerste

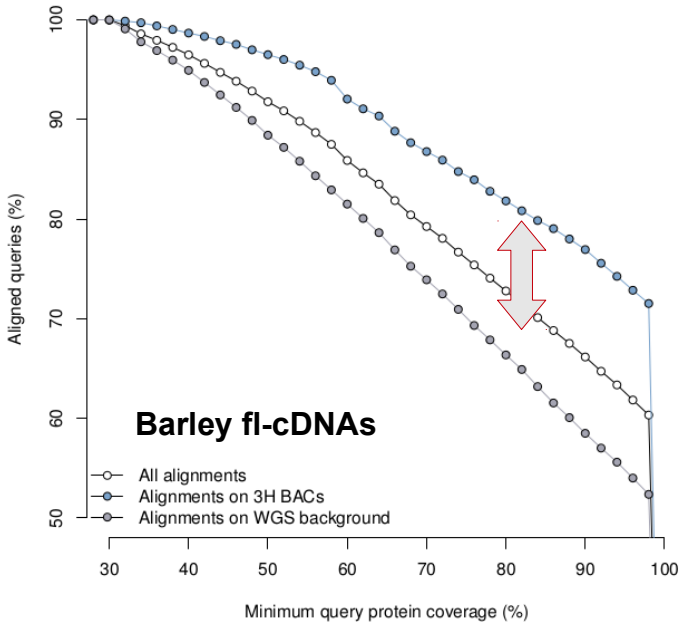


Abb 6: Vollständigere Genmodelle auf den 3H BAC Sequenzen

	On 3H BACs	On WGS genome background	Σ
Loci	**11,871**	**21,907**	**33,778**
Single exon loci	3,552 (30%)	8,856 (40%)	12408 (37%)
Multi exon loci	8,319 (70%)	13,051 (60%)	21370 (63%)
Alternative transcripts at locus	3,431 (29%)	4,555 (21%)	7986 (24%)
Mean locus length (bp)	2,283	1,630	1,859
Median locus length (bp)	1,419	1,028	1,189
Transcripts	**17,075**	**28,426**	**45,501**
Mean / median per locus	1.44 / 1	1.30 / 1	1.35 / 1
Mean transcript length (bp)	1,043	909	959
Median transcript length (bp)	918	711	783
Distinct exons	**57,386**	**87,897**	**145,283**
Mean / median per locus	4.83 / 3	4.01 / 2	4.30 / 2
Mean / median per transcript	4.37	3.87	4.06 / 3
Mean exon length (bp)	260	251	255
Median exon length (bp)	141	143	142

Tab 2: Verbesserung der Gen Metriken auf den 3H BAC Sequenzen

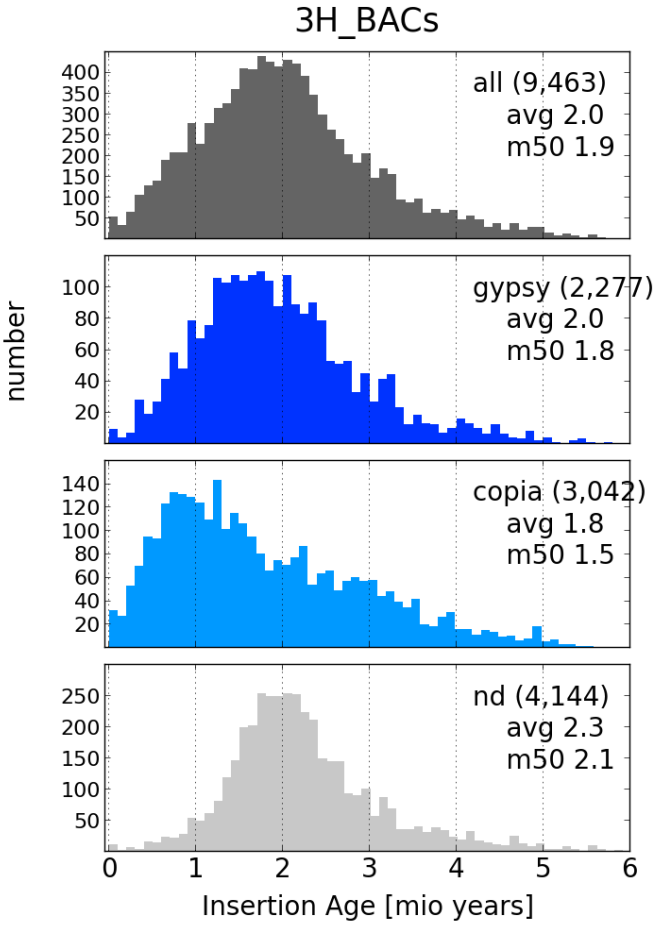


Abb 7: Altersverteilung der Gersten 3H Voll-Längen LTR-Retrotransposons und ihrer Untergruppen

	Pseudogenes					Pseudogene Parents			Genes	
	#	# %	avg length	# per N-free Mb	pseudo-genes per parent	#	avg length	% of genes	#	avg length
all	45.382	100	218	25,0	4,0	11.249	1.179	48,6	23.145	1.108
duplicated	7.893	17,4	340	4,3	2,1	3.676	1.331	15,9		
processed	742	1,6	248	0,4	1,6	474	1.393	2,0		
mono-exonic parent	18.723	41,3	222	10,3	4,9	3.838	838	16,6		
exon-junction less	17.678	39,0	152	9,7	3,0	5.919	1.395	25,6		
chimeric	346	0,8	489	0,2	1,2	278	1.597	1,2		

Tab 3: Pseudogen Detektions Metriken für Gerste (Morex WGS Assembly)

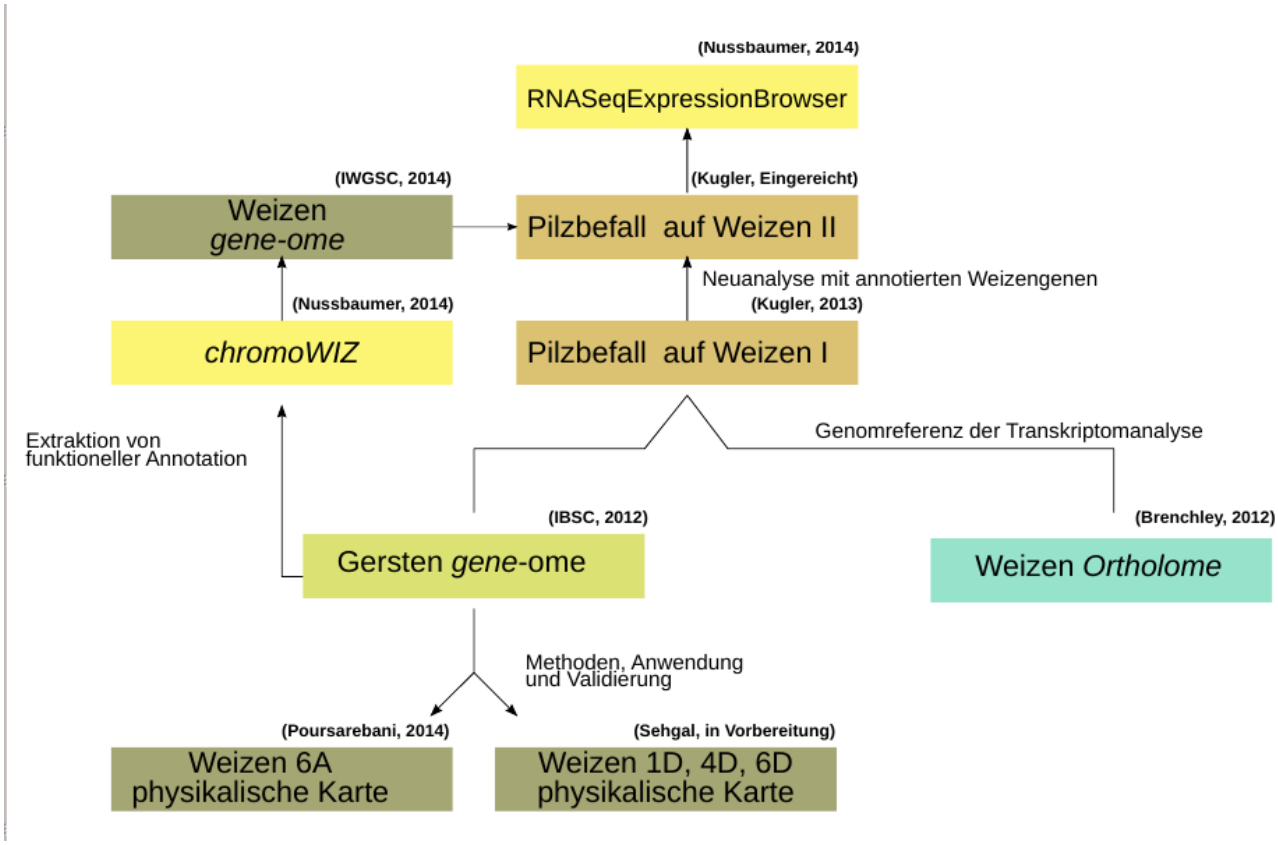


Abb 8: Das Gerstengenom als wichtiger Grundstock für weiterführende Arbeiten

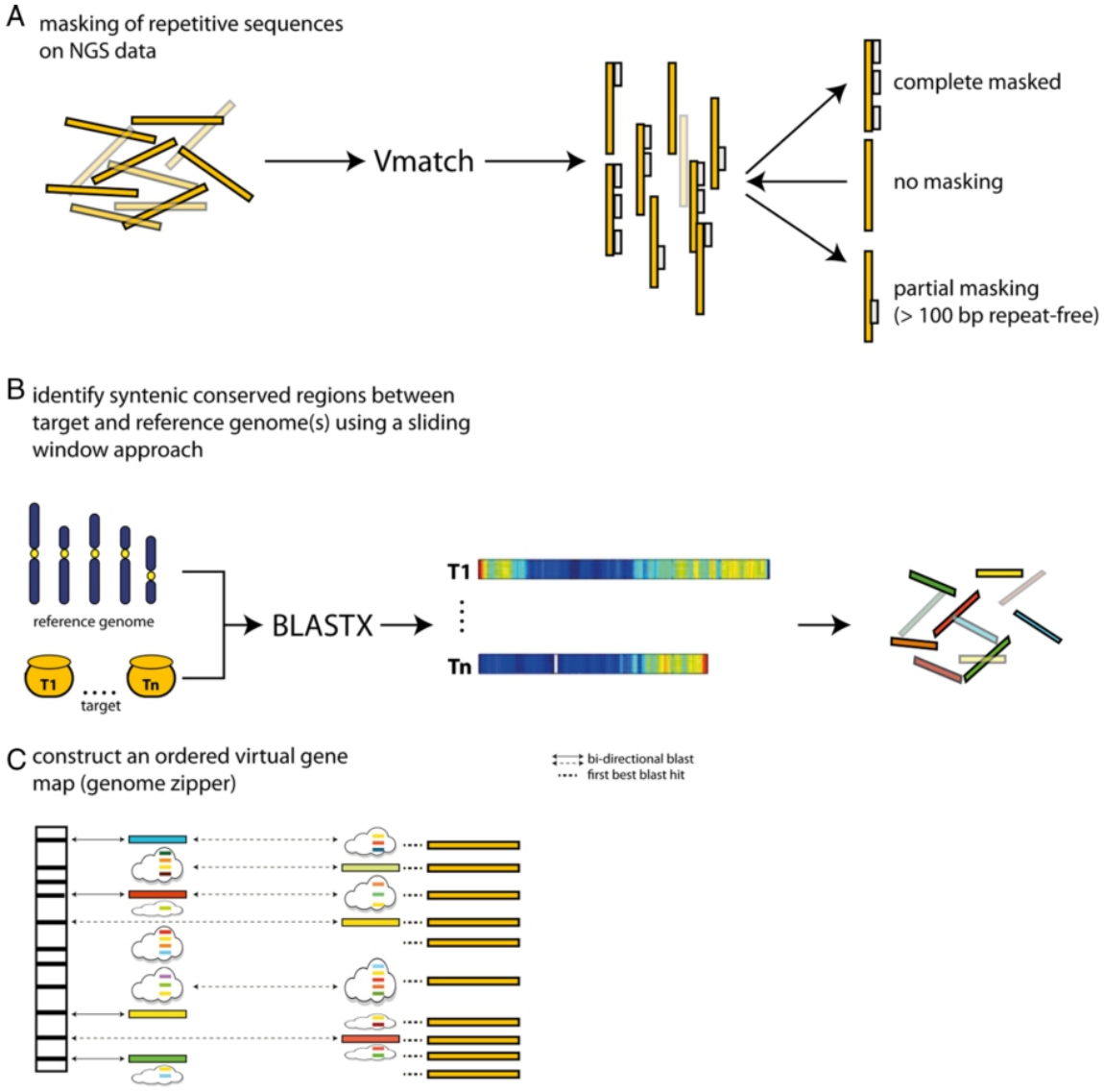


Abb 9: Workflow des GenomZippers

A) Maskierung von repetitiven Bereichen, B) Identifizierung von syntenisch konservierten Regionen zwischen Ziel- und Referenz-Genom, C) Aufbau der virtuellen Genkarte entlang des genetischen Marker Gerüsts.

	1A	2A	3A	4A	5A	6A	7A	∑
Markers	527	417	438	367	459	460	447	**3,115**
Contigs	2,687	4,172	3,623	3,461	3,069	2,233	3,700	**22,945**
Total anchored gene loci	2,460	3,294	2,793	2,725	2,572	2,073	2,618	**18,535**
	1B	**2B**	**3B**	**4B**	**5B**	**6B**	**7B**	∑
Markers	426	706	500	191	571	462	338	**3,194**
Contigs	2,703	4,938	5,011	2,527	3,872	2,440	3,062	**24,553**
Total anchored gene loci	2,291	3,652	3,490	2,062	3,165	1,905	2,041	**18,607**
	1D	**2D**	**3D**	**4D**	**5D**	**6D**	**7D**	∑
Markers	911	1,196	1,012	767	969	708	1094	**6,657**
Contigs	3,765	8,827	6,645	4,405	4,326	2,896	5,089	**35,953**
Total anchored gene loci	2,524	4,017	2,974	2,494	3,080	2,105	3,015	**20,209**

Tab 4: GenomeZipper Statistiken für das gesamte Weizen Genom mit je 7 Chromosomen pro A,B und D Subgenom.

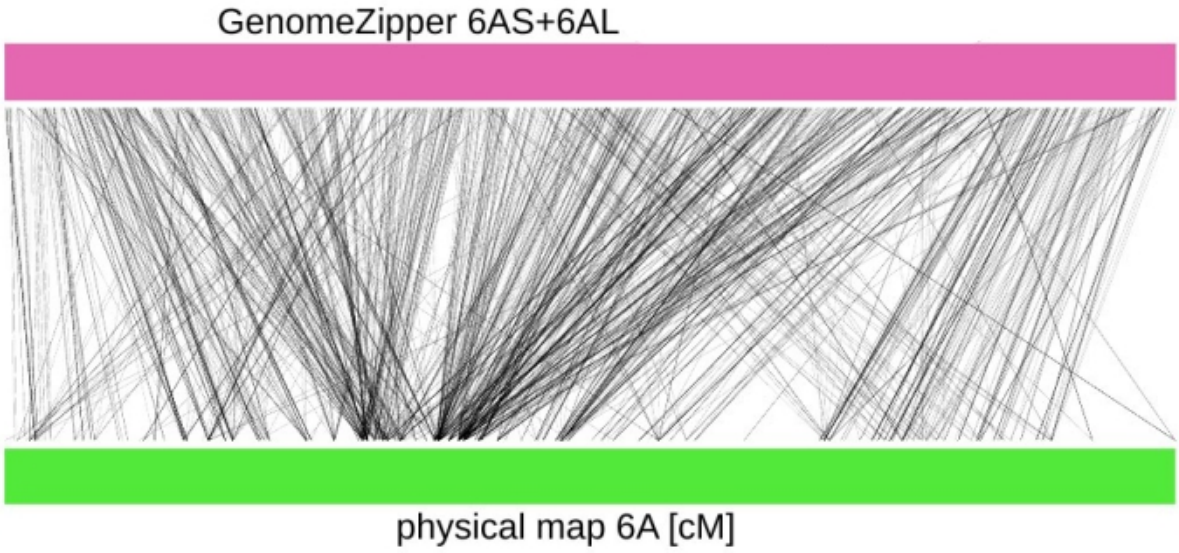


Abb 10: Vergleich der virtuellen GenomeZipper Gen Reihenfolge mit der physikalischen Karte des Weizen Chromosoms 6A.

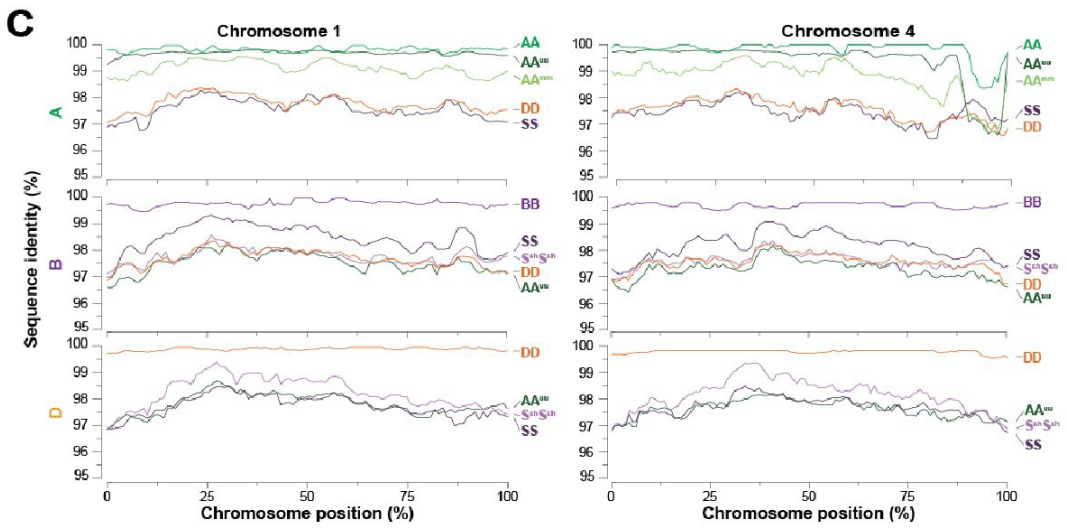


Abb 11: Chromosomale Verteilung von Sequenzähnlichkeiten zwischen Weizen Genen und den homologen Genen der di- und tetraploiden Weizen Vorfahren.

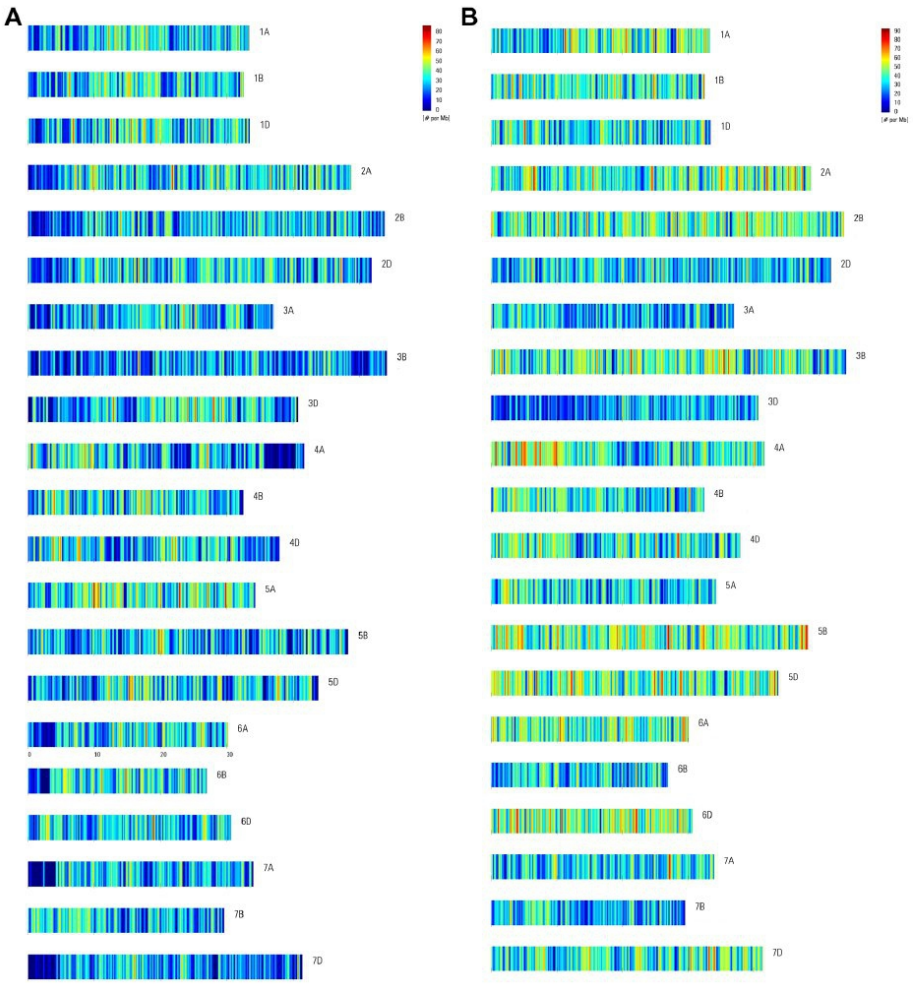


Abb 12: Chromosomale Verteilung von HC Weizen Genen.
A) Core Gene, B) Einzelkopie Gene

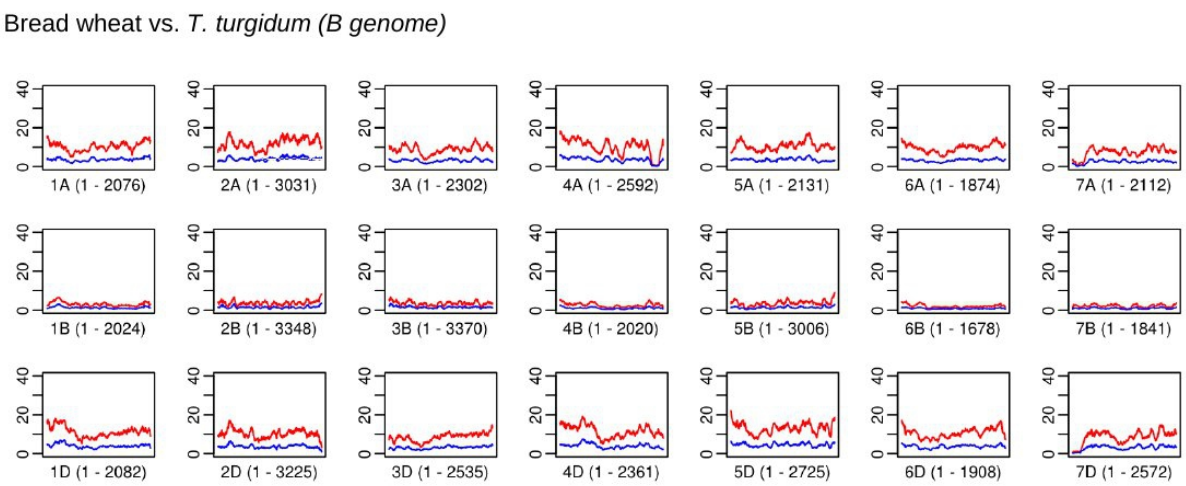


Abb 13: Eine Beispiel für die chromosomale Verteilung von Nukleotid (rot) und Aminosäure (blau) Substitutionen zwischen Weizen und seinen taxonomisch verwandten Vorgängern.

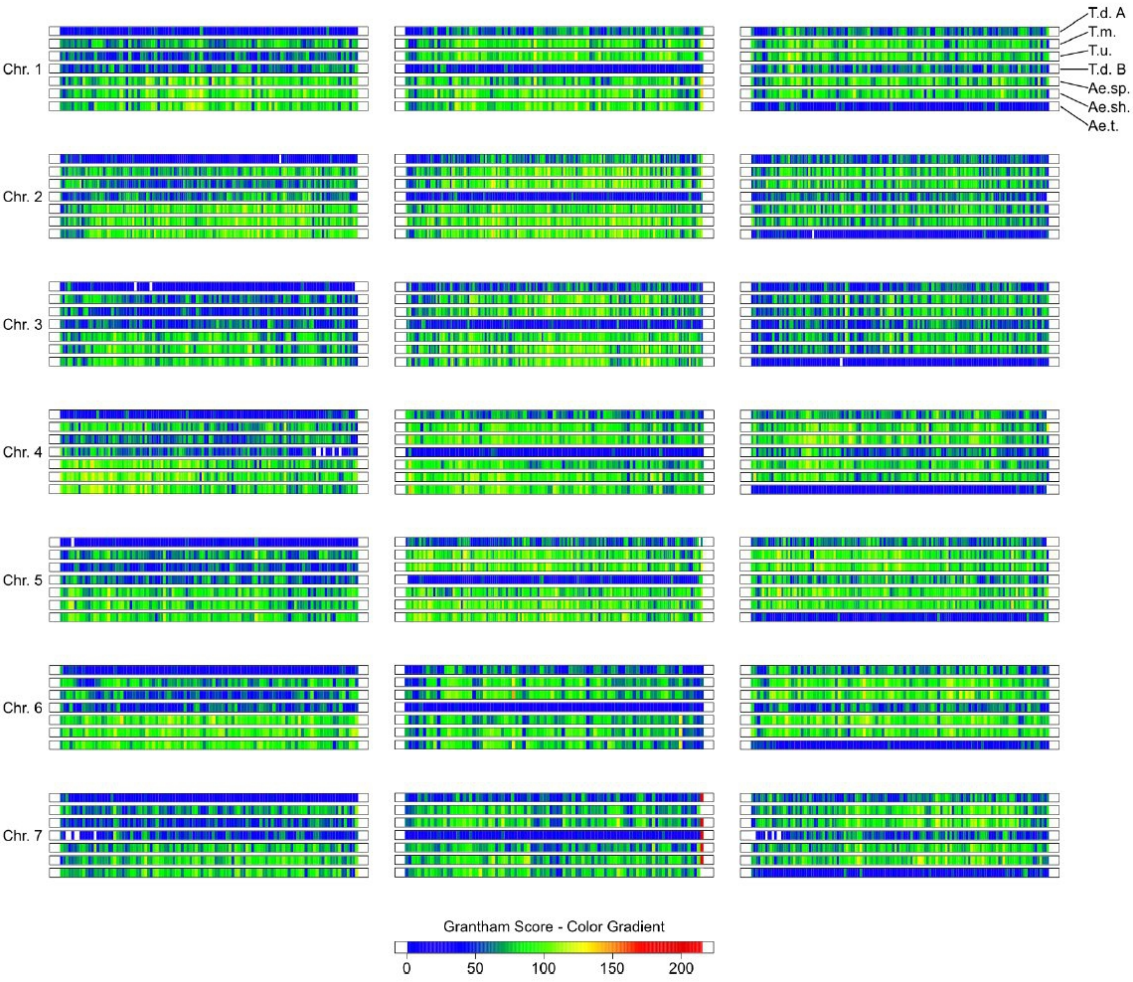


Abb 14:Grantham Score Verteilung von chromosomal positionierten Weizen Genen versus ihren homologen Vertretern in den taxonomisch verwandten Weizen Vorgänger Arten.

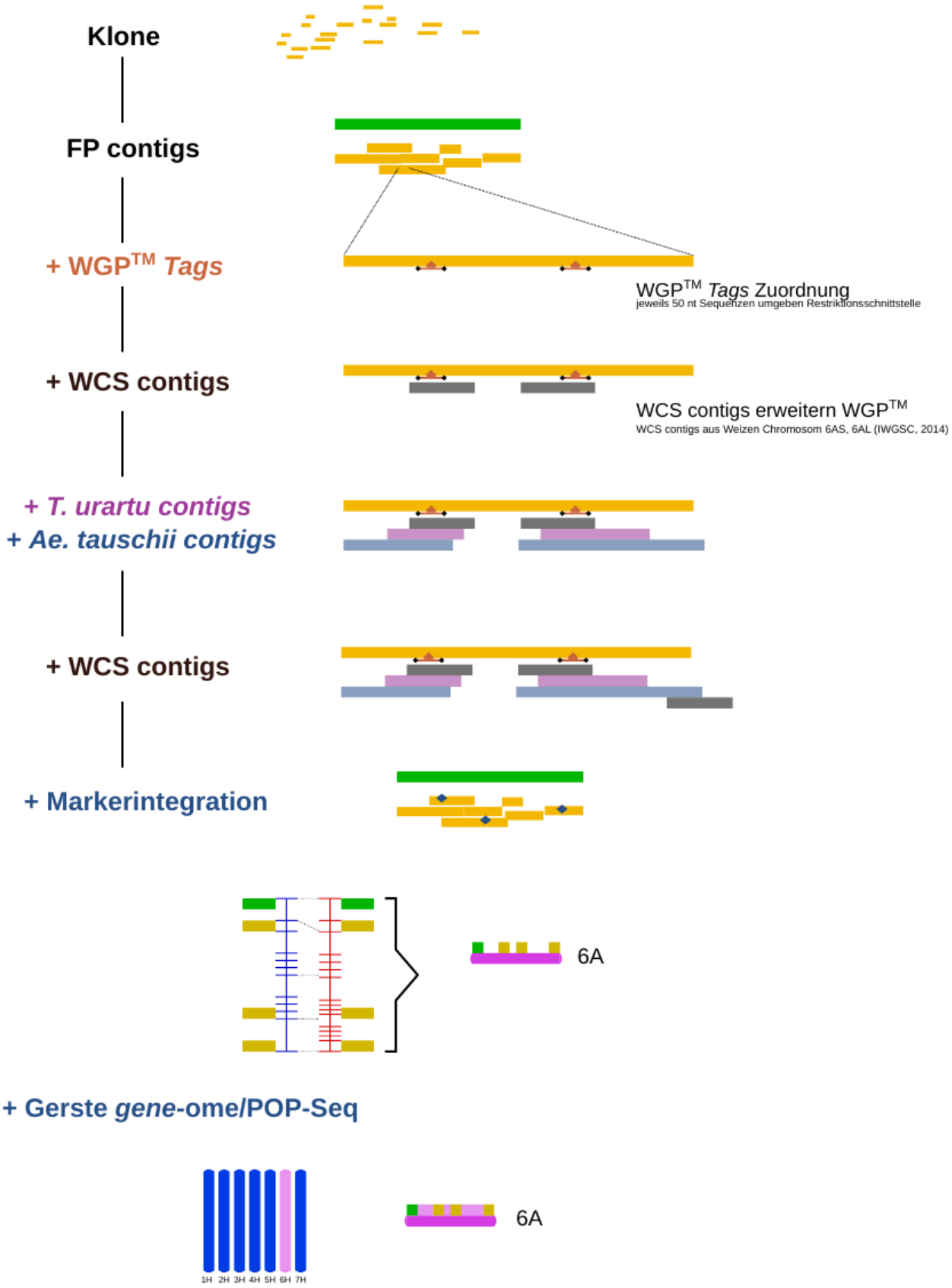


Abb 15: Iterative Strategie zur Integration von Sequenz Resourcen an den WGP Tags der physikalischen Karte von Weizen Chromosom 6A.

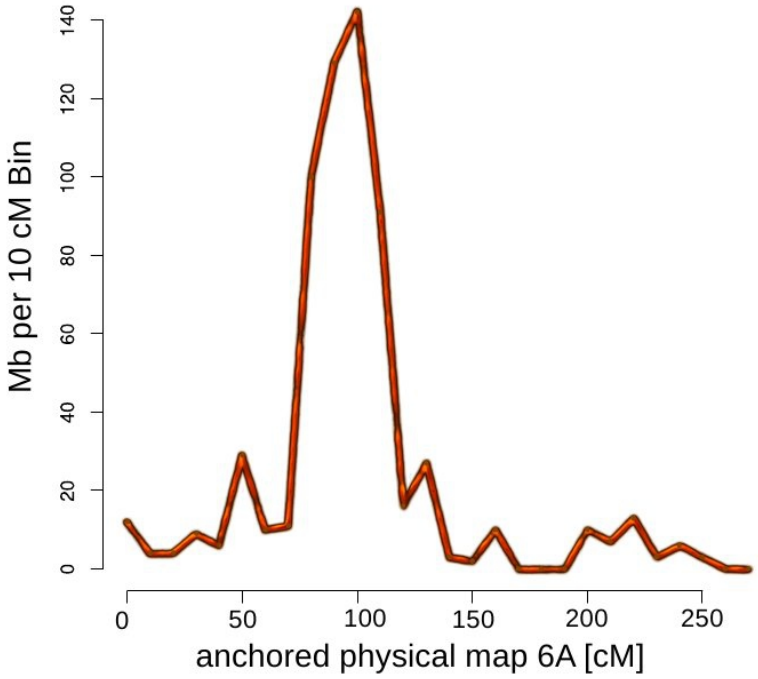


Abb 16: Verhältnis von physikalischen (Mb) zu genetischen (cM) Distanzen. Das Maximum bei 100cM repräsentiert das genetische Centromer mit geringen Rekombinationsraten aber großen Mb Distanzen.

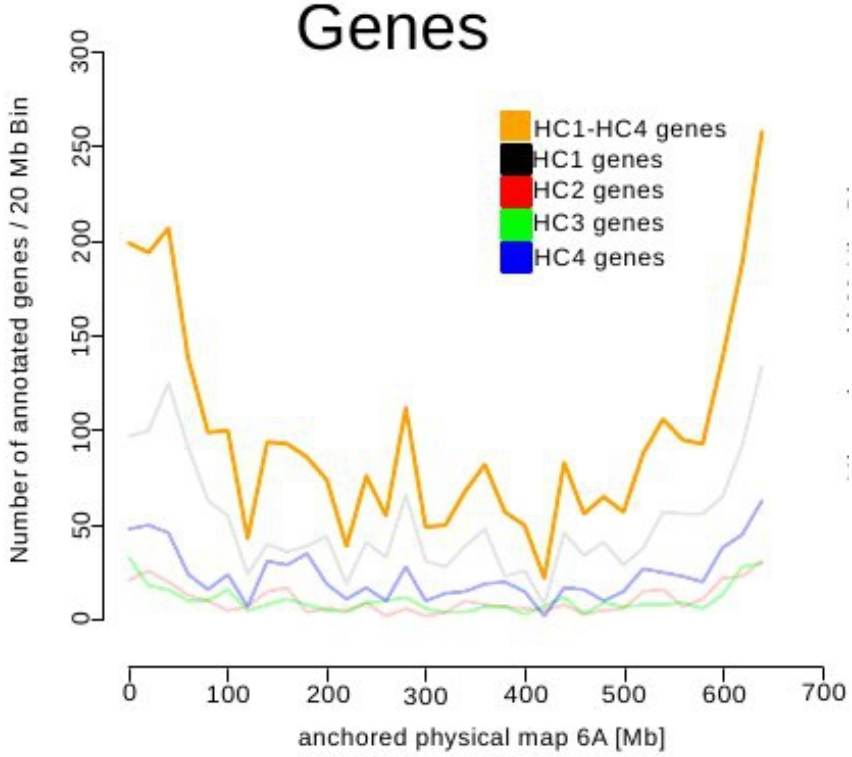


Abb 17: Gendichte entlang des Weizen Chromosoms 6A. Bei HSC1 bis HCS4 handelt es sich um eine Einteilung der Gene in unterschiedliche Konfidenzklassen.

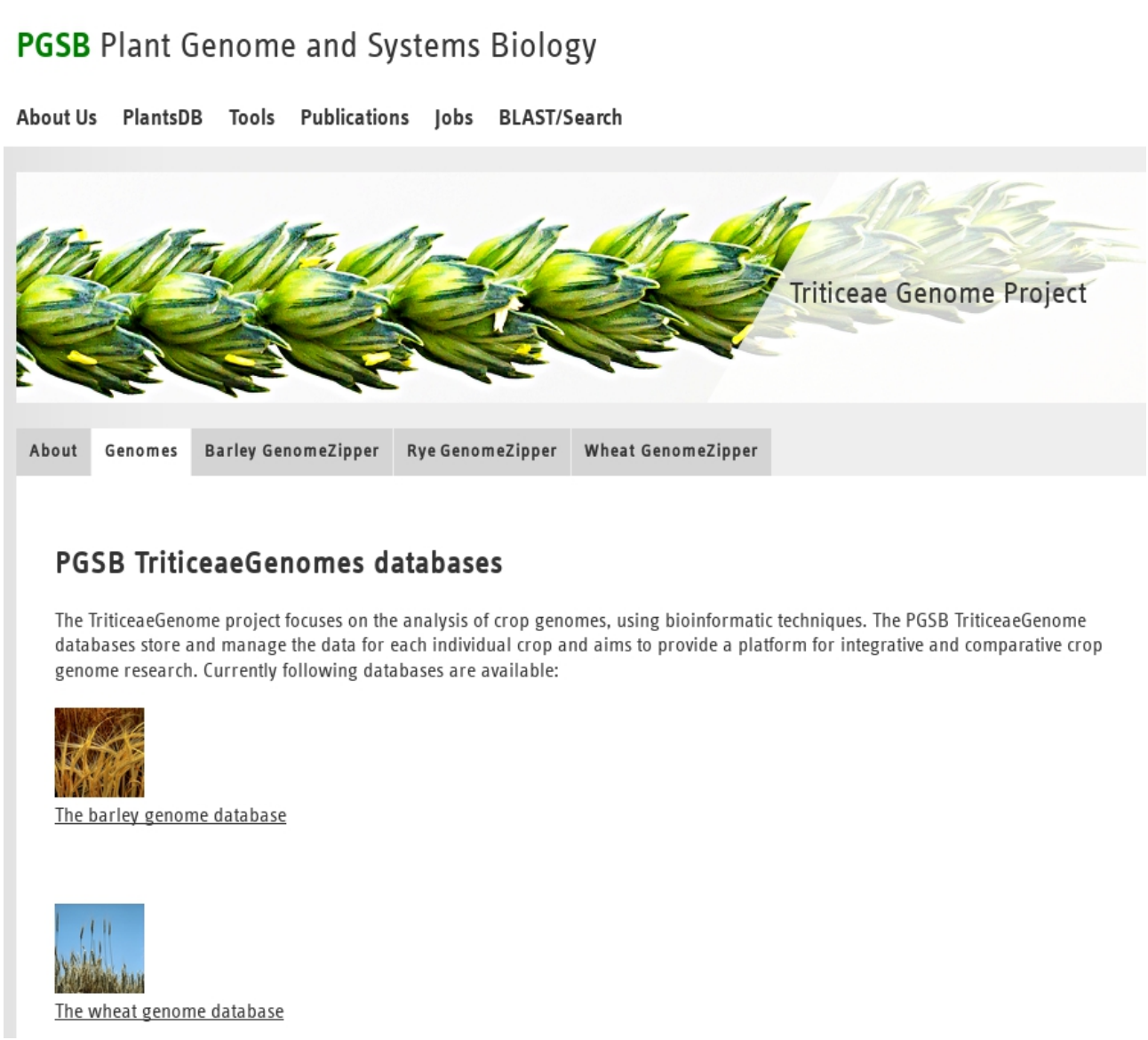


Abb 18: Einstiegsseite der Triticeae Webseiten

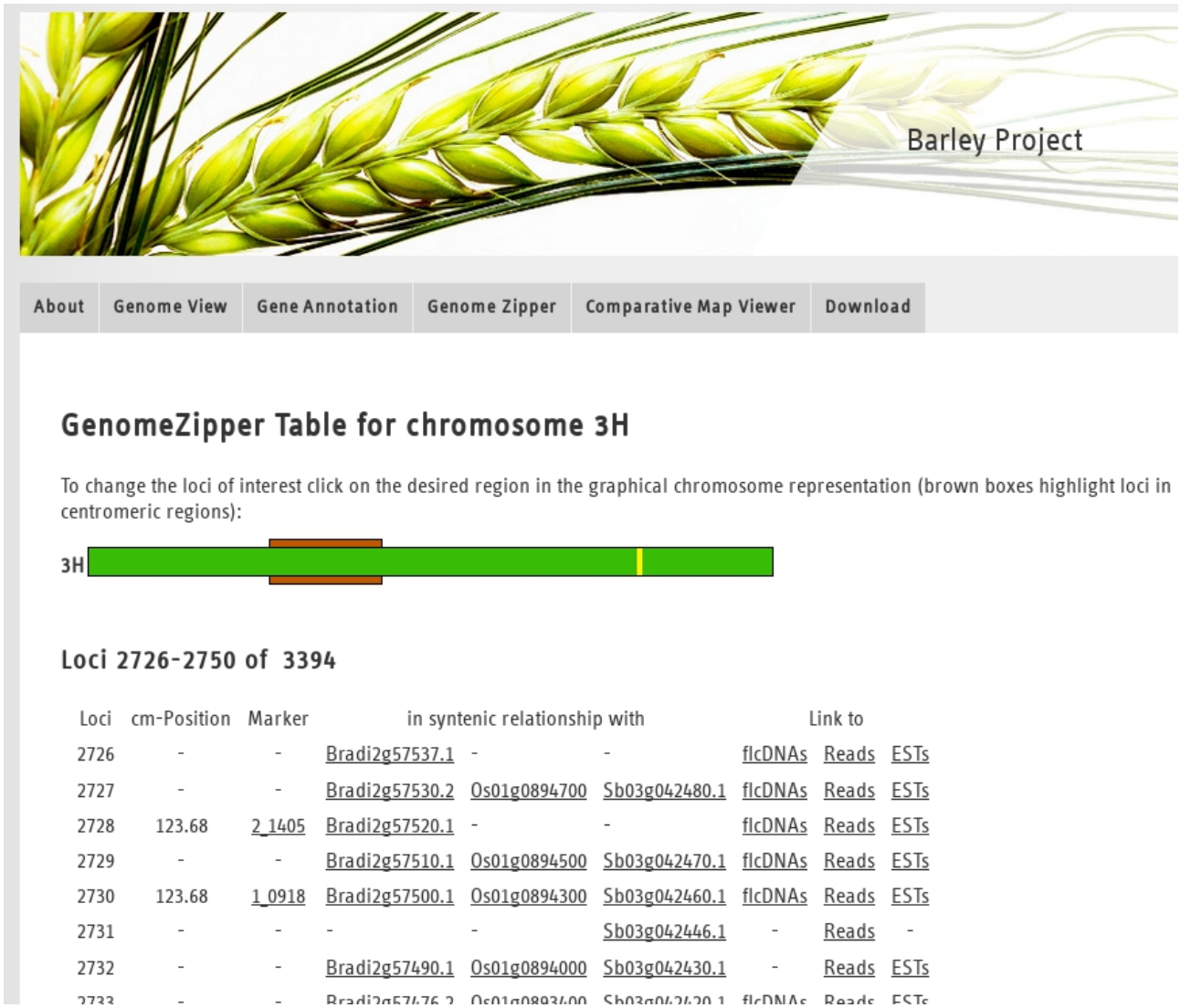


Abb 19: Interaktive Web Präsentation des GenomZippers

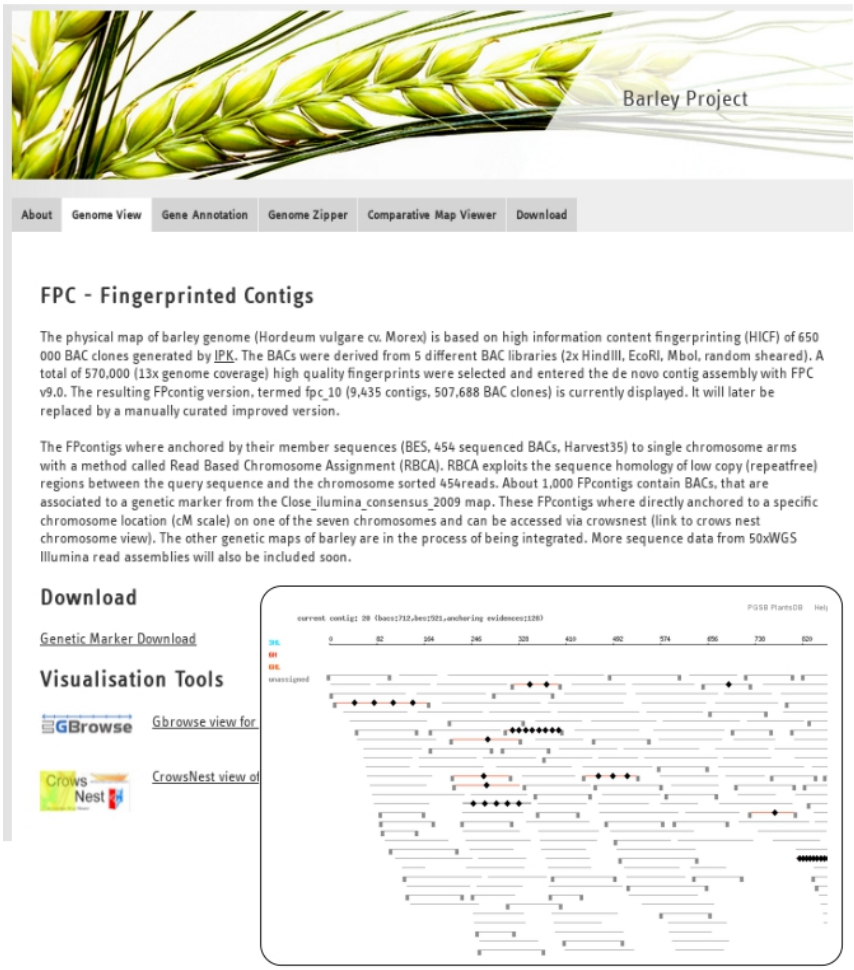


Abb 20: Web Präsentation der physikalischen Karte von Gerste

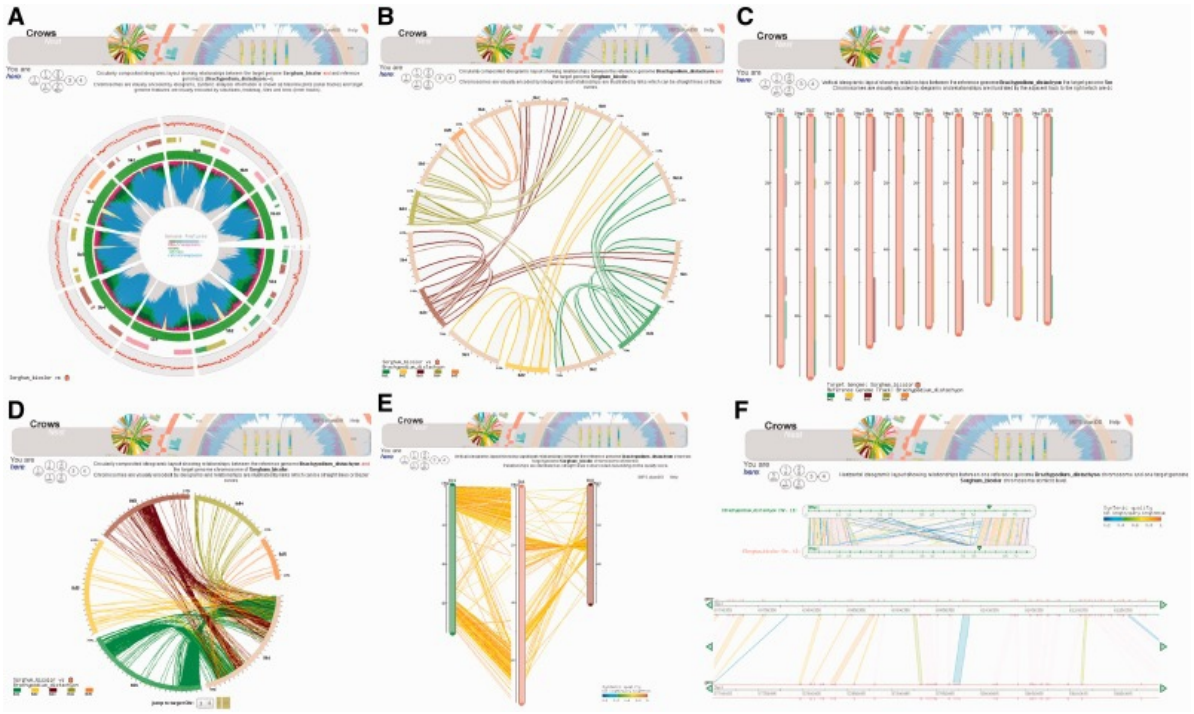


Abb 21: Visualisierungs Ebenen des CrowsNest Synteny Viewers

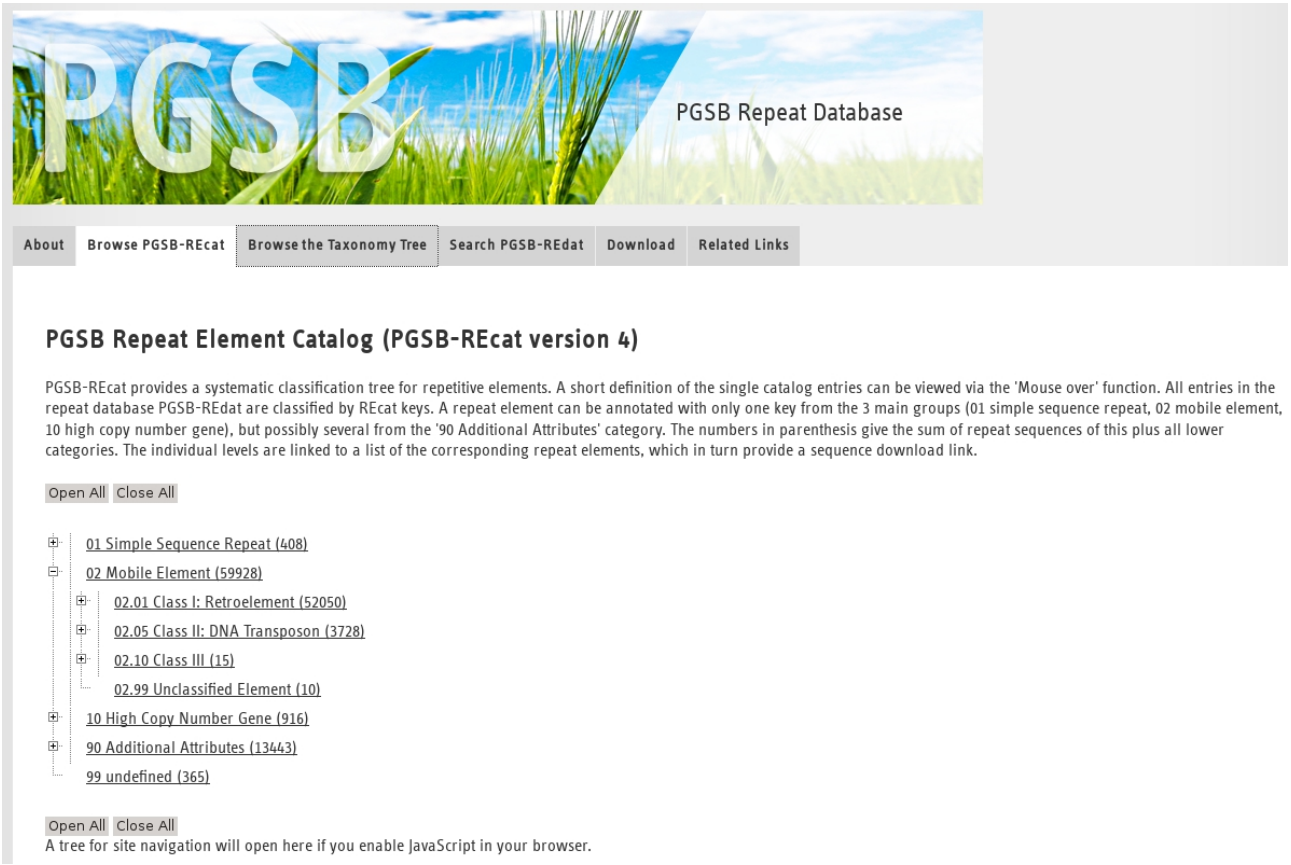


Abb 22: PGSB Repeat Datenbank: Einstieg über die Repeat Klassifizierung

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Abschlussbericht

Pflanzenbiotechnologie - Verbundvorhaben: 'Erforschung von Triticeae-Genomen per Hochdurchsatz - Sequenzierung (TRITEX)' - Teilprojekt C

4. Autor(en) [Name(n), Vorname(n)]	5. Abschlussdatum des Vorhabens 30.11.15
Klaus Mayer, Heidrun Gundlach	6. Veröffentlichungsdatum
	7. Form der Publikation Schlussbericht
8. Durchführende Institution(en) (Name, Adresse)	9. Ber. Nr. Durchführende Institution
Dr. Klaus Mayer, Genomik und Systembiologie der Pflanzen (PGSB), Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstr. 1, 85764 Oberschleißheim	10. Förderkennzeichen 0315954C
	11. Seitenzahl 32
12. Fördernde Institution (Name, Adresse)	13. Literaturangaben 33
Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	14. Tabellen 4
	15. Abbildungen 22

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)

18. Kurzfassung

Ziel des Projektes war es für die sehr großen, komplexen und hoch repetitiven Genome der beiden Getreidearten Gerste (5 Gb) und Weizen (17 Gb) eine weitestgehende Genanordnung und strukturelle Charakterisierung zu erhalten. Als Grundlage dienten chromosomenarm sortierte Sequenzen in Verbindung mit neuen Hochdurchsatz Sequenzierungs- und genetischen Marker Technologien. Aus den entsprechend hoch-heterogenen Datensätzen wurden mit innovativen bioinformatischen Ansätzen mehrschichtige Gene tragende Chromosomen Gerüste und erweiterte physikalische Karten aufgebaut. Obwohl es sich dabei noch um vorläufige Anordnungen handelt, sind die im Projekt entwickelten Genkarten für viele Forschungs und Züchtungsfragestellungen schon genauso gut einsetzbar wie sequenzvollständige Assemblierungen, die für diese Genomgrößen bisher noch nicht erreicht worden sind. Eine Reihe von publizierten Anwendungsbeispielen belegen die Nützlichkeit der im Projekt erzeugten Daten Resourcen.

19. Schlagwörter

Bioinformatik, Genomik, Sequenzierung, genetische Marker, Genom Assemblierung, Annotation, Syntenie, GenomZipper, Pflanzenzüchtung, Gerste, Weizen

20. Verlag	21. Preis